ASTIN BULLETIN Volume 33, No. 2 November 2003 PEETERS

ISSN 0515-0361

Volume 33, No 2 November 2003

ASTIN BULLETIN

The Journal of the ASTIN and AFIR Sections of the International Actuarial Association

CONTENTS

continued on next page

PEETERS

EDITORIAL POLICY

ASTIN BULLETIN started in 1958 as a journal providing an outlet for actuarial studies in non-life insurance. Since then a well-established non-life methodology has resulted, which is also applicable to other fields of insurance. For that reason *ASTIN BULLETIN* has always published papers written from any quantitative point of view – whether actuarial, econometric, engineering, mathematical, statistical, etc. – attacking theoretical and applied problems in any field faced with elements of insurance and risk. Since the foundation of the AFIR section of IAA, i.e. since 1988, *ASTIN BULLETIN* has opened its editorial policy to include any papers dealing with financial risk.

We especially welcome papers opening up new areas of interest to the international actuarial profession.

ASTIN BULLETIN appears twice a year (May and November).

Details concerning submission of manuscripts are given on the inside back cover.

MEMBERSHIP

ASTIN and AFIR are sections of the International Actuarial Association (IAA). Membership is open automatically to all IAA members and under certain conditions to non-members also. Applications for membership can be made through the National Correspondent or, in the case of countries not represented by a national correspondent, through a member of the Committee of ASTIN.

Members of ASTIN and AFIR receive *ASTIN BULLETIN* as part of their annual subscription.

SUBSCRIPTION AND BACK ISSUES

Subscription price: 80 Euro.

Payments should be sent to Peeters, Bondgenotenlaan 153, B-3000 Leuven, Belgium.

To place your order or make an inquiry please contact: Peeters, Journals Department, Bondgenotenlaan 153, B-3000 Leuven, Belgium or e-mail: peeters@peeters-leuven.be

Orders are regarded as binding and payments are not refundable.

Subscriptions are accepted and entered by the volume. Claims cannot be made 4 months after publication or date of order, whichever is later.

INDEX TO VOLUMES 1-27

The Cumulative Index to Volumes 1-27 is also published for ASTIN by Peeters at the above address and is available for the price of 80 Euro.

EDITORIAL

The Combination of Theory and Practice as well as Finance and Insurance is a successful formula

Astin meetings are held in high regard and affection by the attendees; both academics and practitioners. There is a history of regular attendance, which indicates the value the participants place on the meetings. The Astin Bulletin has an excellent reputation in the insurance industry and its articles are regularly cited and continuously referred to by other journals. However there is a school of thought that it is all too theoretical and is therefore of limited interest to the practising actuary. As a practising actuary, I have not found this to be the case and without fail find practical ideas and applications from each Astin meeting I attend and also each edition of the Astin Bulletin I read.

THE IMPORTANCE OF THEORY

Actuaries need to have a sound theoretical base to undertake their work. This is particularly true of the general insurance and investment fields. It is very easy, in both these areas, to believe that one is undertaking sound analyses while actually making serious mistakes or performing sub-optimally if one is not up to date with all the theoretical developments. It is important in this context to realise that 'You don't know what you don't know'. Unfortunately this is all too common among actuaries who pride themselves on their practicality.

DFA is a case in point. It is regarded as a very practical subject, but it is surprising the number of DFA practitioners who have not kept up with recent developments. The use of copulas to handle tail dependency rather than naively assuming independence, has a major impact on capital allocation. This theory is not new, though the ability to use it computationally is a phenomenon of the reduction in the cost of computing power. Consequently, it has recently become much more useful to the practitioner. There have been a number of Astin papers that have analysed the application of copulas further and also adapted them to modern computing technology including covering many of the practical issues of choosing which copula to use.

Some of the developments in the theory of risk measures are also very relevant to capital allocation and DFA. The theory can seem very abstract but the use of non-coherent risk measures can dramatically distort the results and cause wrong decisions to be made. A good example is the use of Var as a risk measure. It is widely used in the banking world for market risk and where work on capital allocation preceded that in the insurance industry. Consequently there is a temptation in diversified financial groups to use Var as a capital allocation tool for insurance risks. When that is done, it is usually with disastrous results as there will often be an understatement of the capital requirements for the insurance risks. Without understanding the theoretical reasons as to why this occurs is a recipe for disaster. Var works for banks for market risk because the risks are largely symmetrical. Under these circumstances Var will order risks in the same way as a number of coherent risk measures. However this is not the case with skew risks, especially if there are tail dependencies. This is usually the situation for insurance companies and also for operational risk. Some very large financial institutions have made some serious mistakes as a result of not being aware of the relevant theory. For the individual actuary there is the real risk of a career ending in tears by not being up to speed with the theory. There are also very serious implications for the profession as a whole.

Another development is pricing methodology. The work by Wang, Panjer et al is of major importance in providing consistent ways of pricing across markets and perhaps even more practically for measuring the relative attractiveness of different markets or of measuring the underwriting cycle. A major practical problem is how to maintain consistency within an organisation. The Wang transform provides an approach that allows practical and readily implementable methodology to solve an insurance company management problem that would otherwise be an extremely complex organisational issue.

Often the practical actuary can utilise what superficially seem to be very theoretical papers. For example, practising actuaries are not usually interested in probabilities of ruin and hence may be tempted to dismiss the many papers dealing with approximations to it to be merely of interest to the academics. However these formulae also provide an approximation to the claim frequency on an excess of loss contract. Simple appreciation of this fact allows the practising actuary to utilise the work in this area and thus streamline the pricing. I have found judicial reading of such articles to be a fruitful source of ideas.

ASTIN WELCOMES PRACTICAL PAPERS

The Astin Bulletin would welcome more papers by practising actuaries. Indeed this would create a virtuous circle of creating more interest in the application of theoretical developments which in turn would encourage more academic work in this area and also more practising actuaries to participate in Astin.

It would also encourage academicians to solve problems that were complex but for which practically important. Assumptions that are mathematically convenient or elegant are not helpful to the practitioner. Solutions to problems that are very complex are invaluable. The academics can identify fruitful areas for them to research by understanding where the problems lie for the practitioner. In any event, actuarial science is an applied science and therefore by definition must have its roots in the real world and produce relevant applications.

However it is important that Astin does not let its academic standards slip, especially in the Bulletin. It is important that it has a high reputation for academic rigour not only to maintain its reputation and attract academics to publish articles. The two are not incompatible and also demonstrate to the practitioner the importance of correct theory. However the discussions at meetings cover many more of the practical aspects. This is not surprising and a forum for transmitting these to a wider audience would be of value to the whole profession.

Convergence of financial markets

The Astin Bulletin now incorporates the Journal of the Afir Section of the IAA. This is not just an administrative convenience or a consequence of banking and insurance converging. It is also of immense use to the practising and academic actuary alike as it allows ideas in one area to be more easily utilised in the other. Furthermore an analysis of the different techniques provides some insights into the differences of the two areas. For example, much of the pricing methodology in general insurance is based on the assumption that the risk will be held on the balance sheet of the insurer and not traded(there is no real market place or natural short sellers for most insurance risks). However the banking approach assumes a market based approach and the corresponding pricing methodologies. This requires markets to exist and the ability to diversify risk. This is not possible if the risks are too large for the number of participants in the market place. Hence the 'insurance pricing models' will be required with corresponding charges for risk. However for something that can be readily diversified, then it is likely that a market based risk charge can apply. This also has implications for the relative pricing of reinsurance versus insurance.

Diversification by risk and also by risk class is something that the actuary can learn from the different approaches. Financial risks usually have significant correlations and/or tail dependencies that diversification is used to mitigate. However this is often not the case between physical risks. Thus earthquake risk is only correlated across zones and not world wide whereas there are world wide correlations on credit risks and hence the problems that some insurers ran into in recent years when they did not realise this.

Another area is risk. Many insurance company actuaries (and indeed many nor-actuaries) intuitively believe that high risk should be handled by high discount rates. They do not understand that this gives the wrong answers and the need for risk neutral probabilities, martingales etc. This is an issue not just of theory but of being exposed to different techniques. It is also an area where incorrect methodology gives rise to wrong decisions and is heavily biased towards long term projects.

The Future

I believe that the combination of academic and practitioner, of finance and insurance is a powerful one. Provided we can all co-operate and communicate together the future is bright and that we will all benefit from each other and that the sum of the parts is much greater than the whole. Failure to keep up with modern developments even if they seem abstruse, mathematically complicated and difficult to apply in practice is not a sign of a lack of practicality but more of a high risk strategy to both the individual and the profession.

NEW ECON FOR LIFE ACTUARIES

BY

KNUT K. AASE AND SVEIN-ARNE PERSSON*

ABSTRACT

In an editorial in *ASTIN BULLETIN*, Hans Bühlmann (2002) suggests it is time to change the teaching of life insurance theory towards the real life challenges of that industry. The following note is a response to this editorial. In Bergen we have partially taught the NUMAT, or the NUMeraire based Actuarial Teaching since the beginning of the 90's at the Norwegian School of Economics and Business Administration (NHH). In this short note we point out that there may be some practical problems when these principles are to be implemented.

1. ACTUARIAL MATHEMATICS VS FINANCIAL ECONOMICS

As recognized by Bühlmann the model used in Life Insurance Mathematics is built on the two elements: (i) mortality, and (ii) time value of money. This is, however, not sufficient to comprise a consistent pricing theory of a financial product, such as a private life insurance contract, a pension or an annuity. It is rather remarkable that mathematicians have, for more that 200 years, arrogantly (or more precisely, ignorantly) disregarded any economic principles in pricing such products (or any other insurance products for that matter). It should not come as a surprise that it is rather natural to use the economic theory of contracts to study — insurance contracts.

Financial pricing of life insurance contracts often starts by assuming the existence of a market of zero coupon bonds. The market price at time zero $B_0(t)$ of a default free *unit discount bond* maturing at the future time t is typically given by the formula

$$B_0(t) = E^Q\left\{e^{-\int_0^t r(s)\,ds}\right\}, \tag{1}$$

where $r(t)$ is the spot interest rate process, and Q is a risk adjusted probability measure equivalent to the originally given probability P. Standard references such as Heath, Jarrow, and Morton (1992) or Duffie (2001) show that most popular term structure models lead to this representation of the market price of a unit disount bond.

[1] In addition to the response from Hans Bühlmann, the authors appreciate the comments from Editor Andrew Cairns.

Without going into further technical details regarding such models, let us consider some standard actuarial formulae for the most common life insurance contracts. We consider first the two building blocks for life and pension insurance regarding one life: *pure endowment insurance* and *whole life insurance*. We start with the former, stating that "one unit" is to be paid to the insured if he is alive at time t. Let ${}_tp_x$ be the probability that a person of age x shall still be alive after time t. That is, if T_x represents the remaining life time of an x year old representative insurance customer at the time of initiation of an insurance contract, then ${}_tp_x = P(T_x > \mathrm{t})$. In the traditional framework, the single premium for a pure endowment insurance is

$$ {}_tE_x = e^{-\int_0^t (\delta + \mu_{x+s})\,ds} = {}_tp_x e^{-\delta t}, \tag{2} $$

where δ is the "force of interest", or technical interest rate, and μ_x is the death rate of an x year old insurance buyer. On the other hand, the above formula reads in the new language

$$ {}_tE_x^M = {}_tp_x B_0(t), \tag{3} $$

provided the mortality risk is "diversifiable", or uncorrelated with the financial risk and "unsystematic". The superscript M will be used to indicate *marked based* valuation. Notice that the difference between (2) and (3) is how we value the "unit" at the inception of the contract.

The simplest way to show relation (3) is as follows: Let $I_{(T_x > t)}$ denote the *indicator function* of the event $(T_x > t)$, i.e., $I_{(T_x > t)} = 1$ if $T_x > t$ and zero otherwise. Observe that $E(I_{(T_x > t)}) = {}_tp_x$.

By financial theory the market value of the above contract is

$$ {}_tE_x^M = E^Q\left\{e^{-\int_0^t r(s)\,ds} I_{(T_x > t)}\right\}, $$

where $E^Q\{\cdot\}$ denotes the expectation under an equivalent martingale measure Q. The expectation under the measure Q can alternatively be written

$$ E^Q\left\{e^{-\int_0^t r(s)\,ds} I_{(T_x > t)}\right\} = E\left\{\xi_t e^{-\int_0^t r(s)\,ds} I_{(T_x > t)}\right\}, $$

where ξ_t is the "density" process, i.e., $\pi_t = \xi_t e^{-\int_0^t r(s)\,ds}$ is a state price. Under the stipulated conditions the state price depends only on market variables, in this case the interest rate process, and is thus independent of the random variable T_x. By this independence we get:

$$ E\left\{\xi_t e^{-\int_0^t r(s)\,ds} I_{(T_x > t)}\right\} = E\left\{\xi_t e^{-\int_0^t r(s)\,ds}\right\} E\left(I_{(T_x > t)}\right) = E^Q\left\{e^{-\int_0^t r(s)\,ds}\right\} {}_tp_x, $$

the first equality follows from independence, the second from from properties of the probability measure Q. The result finally follows from expression (1).

Turning to the other building block in life insurance, the *whole life insurance contract*, here “a unit” is payable upon death. The single premium is denoted by $\bar{A}_x$, and is given by the formula

$$\bar{A}_x = 1 - \delta \int_0^\infty {}_t p_x e^{-\delta t} \, dt \tag{4}$$

in the traditional approach, while in the new approach it is given by

$$\bar{A}_x^M = 1 + \int_0^\infty {}_t p_x B_0'(t) dt \tag{5}$$

where $B_0'(t) = \frac{\partial B_0(t)}{\partial (t)}$. Here the difference between (4) and (5) stems from how we compute time changes in the present value of the “unit” in the two different models. Again it is the difference in how we value the “unit” in a dynamic financial market based framework that matters.

From these two contracts all the other standard contracts could easily be developed. One example which we use below is *term insurance*, i.e., “a unit” is payable upon death, but only if death occurs before a given horizon T. The single premium $A^{1}_{x:\overline{T}|}$ of the term insurance contract can be expressed as

$$A^{1}_{x:\overline{T}|} = \bar{A}_{x:\overline{T}|} - e^{\delta T} {}_T p_x, \tag{6}$$

where $\bar{A}_{x:\overline{T}|} = 1 - \delta \int_0^T {}_t p_x e^{-\delta t} \, dt$, is the single premium of the endowment insurance. In the new language this formula becomes

$$A^{1,M}_{x:\overline{T}|} = \bar{A}^{M}_{x:\overline{T}|} - B_0(T) {}_T p_x, \tag{7}$$

where $\bar{A}^{M}_{x:\overline{T}|} = 1 + \int_0^T {}_t p_x B_0'(t) dt$.

This approach would also be the starting point for valuing guarantees, and other financial derivatives that exist in this industry today. Other numeraires than the zero cupon bond would have to be considered as the contracts may be related to different portfolios of financial primitives.

The principles described above were indeed included in an elementary textbook in insurance mathematics (see[1] Aase (1996)) already in the beginning of the 90’s. At NHH this could be easily done, in the Humboltian tradition, since our program does not have any formal ties, or strings attached to the actuarial profession, and could e.g., ignore any legal aspects or accounting standards[2].

[1] This book is based on lecture notes from 1993.

[2] Some universities have, in our view, a too close connection to the professional industry, which in some cases may actually hamper the natural development of the field.

Possible problems with the new approach

There are several scientific papers on the issues raised above[3], but our aim is not to give a complete account of these here. We would, however, like to point out a few difficulties with the new approach.

First, the above price $B_0(t)$ could, according to Bühlmann (2002), "be read in today's newspaper". A quick look at the existing markets for bonds reveals that this is not possible, not even in highly liquid markets such as the UK Market, see e.g., Davis and Mataix-Pastor (2003). On the contrary, there is a serious *"missing markets"* problem, meaning that the complete term structure for maturities longer than 1 year must typically be extracted from only a small number (maybe not more than two or three) of bond prices.

The above formulae require, on the other hand, the functions $B_0(t)$ to be given for all t, and moreover, this should be possible at every instant, e.g., at every day, as time goes.

Even if this difficulty could be partially overcome technically, by smoothing the yield curve (see e.g., Adams and van Deventer (1994) or Cairns (1998)), the issuer of the insurance products would face a second problem, this time of a *pedagogical* nature: Identical and long term insurance contracts may obtain discernible different single premia on consecutive days, or even within the same day. This difference would thus be due to daily (or intra-daily!) fluctuations in the financial market, ceteris paribus. None of these issues arise in the traditional approach, which is based on a so-called technical interest rate, completely separated from real world financial market conditions.

Let us illustrate the latter problem here. We use term structure data for the Norwegian market[4] from the first Friday of each month in 2002. Daily observations of the 1 year, 3 year, 5 year, and 10 year interest rates were available. These observations were interpolated to obtain the 2 year, 4 year, and the 6-9 year interest rates. Single premiums for a 10 year pure endowment and 10 year term insurance were calculated using the Norwegian N 1963 mortality table. The benefit is normalized to 100.

Table 1 only reports monthly changes in single premiums, and thus, does not illustrate the potential problem of daily or even intra-daily price fluctuations. However, Table 1 does indicate that monthly price changes may vary from 0.47% to 3.44% for pure endowment single premiums. Actually, the average monthly change in the pure endowment single premium is 1.6%. For term insurance the monthly changes in single premiums are less, from 0.02% to 1.87%, with an average (over the 3 age groups) of the mean montly price change of 0.92%.

The volatility of a financial asset is the (annualized) square root of the instantaneous variance of the logarithmic return. We estimated the volatilities

[3] The authors have been involved e.g., in the following articles: See Persson (1998); Bacinello and Persson (2002) for pricing of life insurance under stochastic interest rates, Persson and Aase (1997); Miltersen and Persson (1999, 2003) for guarantees in life insurance, Miltersen and Persson (1999) also briefly discuss different numeraires.

[4] Found at www.norges-bank.no.

TABLE 1

Date	*Pure endowment*			*Term Insurance*		
	40 year	*60 year*	*80 year*	*40 year*	*60 year*	*80 year*
Jan 4, 02	52,26	42,36	9,90	3,09	16,00	62, 25
Feb 1, 02	52,02	42,16	9,85	3,06	15,82	61,53
Mar 1, 02	51,39	41,65	9,73	3,04	15,72	61,19
Apr 5, 02	50,38	40,84	9,54	3,02	15,63	61,02
May 3, 02	50,01	40,53	9,47	2,99	15,49	60,49
Jun 7, 02	48,94	39,67	9,27	2,95	15,25	59,67
Jul 5, 02	49,87	40,42	9,45	2,98	15,41	60,17
Aug 2, 02	51,58	41,81	9,77	3,03	15,70	61,05
Sep 6, 02	53,16	43,08	10,07	3,09	15,98	61,95
Oct 4, 02	52,86	42,84	10,01	3,08	15,96	61,93
Nov 1, 02	52,56	42,60	9,96	3,09	15,97	62,05
Dec 6, 02	53,46	43,33	10,13	3,12	16,15	62,63

Single premiums for pure endowment and term insurance contracts with benefit 100 and 10 years horizon for male insurance customer with age 40, 60, and 80 years, respectively, at the inception of the contract. Single premiums are calculated by NUMAT as follows: Equation (3) is used for the pure endowment contract. For the term insurance contract we have discretized equation (7). First observe that $A^{1,\,M}_{x\,:\,\overline{T}|} = \int_0^T f_x(t)\, B_0(t)\, dt$ where $f_x(t) = \mu_{x+t}\,{}_tp_x$ represents the probability density of an x-year old person's remaining life time. Then $A^{1,\,M}_{x\,:\,\overline{10}|}$ is discretized as $\sum_{i=1}^{10} {}_{i-1|}q_x\, B_0(i)$, where ${}_{i-1|}q_x = \Pr\left(i-1 < T_x \le i\right)$ represents the probability of an x-year old customer to die in year i after the contract is initiated. The single premiums are calculated using the prevailing term structure from the first Friday of each month in 2002 and the N 1963 mortality table.

of the same 6 contracts used as examples in Table 1, but now based on *daily* observations from 2002. The volatilities of all three pure endowment contracts are identical and equal to 6.73%, which by the very nature of this contract is the same as the volatility of the 10 year bond. The volatility of the term insurance contracts are 3.47%, 3.45%, and 2.76%, for an insurance customer of age of 40, 60, and 80 years, respectively, at the inception of the contract. These volatilities are roughly of the same magnitude as the average of the volatilities of the 1-10 year bonds, estimated to 3.26% from the data.

Also notice that 10 years is a relative short horizon for a life insurance or pension contract. Both the problem of "missing markets" and of fluctuations of the single premiums are expected to be more severe for contracts with longer horizons.

Conclusions

We have pointed out that the approach of using financial market data to price life insurance and pension contracts may lead to substantial variations in the premiums charged. The variations are due to financial market volatility, rather than any differences in idiosyncratic risk. The data period we have picked is very normal, and one can easily envision substantially more discernable effects in more volatile times, and in financial markets in other countries of the world.

References

AASE, K.K. (1996) *Anvendt sannsynlighetsteori: Forsikringsmatematikk* (in Norwegian) (English: *Applied probability theory: Insurance mathematics*). Cappelen Akademisk Forlag, Oslo, Norge.

ADAMS K. and VAN DEVENTER D. (1994) Fitting yield curves and forward rate curves with maximum smoothness. *Journal of Fixed Income*, 52-62.

BACINELLO A. and PERSSON, S.-A. (2002) Design and pricing of equity-linked life insurance under stochastic interest rates. *The Journal of Risk Finance* **3(2)**, 6-21.

BÜHLMANN, H. (2002) New math for life actuaries. *ASTIN Bulletin* **32(2)**, 209-211.

CAIRNS, A.J.G. (1998) Descriptive bond-yield and forward rate models for the british government securities' market. *British Actuarial Journal* **4(2)**, 265-321.

DAVIS, M. and MATAIX-PASTOR, V. (2003) *Finite-dimensional models of the yield curve*. Working paper, Department of Mathematics, Imperial College, London SW7 2BZ, England.

DUFFIE, D. (2001) *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton, New Jersey, USA, 3rd edition.

HEATH, D., JARROW, R. and MORTON, A.J. (1992) Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica*, **60(1)** 77-105.

MILTERSEN, K.R. and PERSSON, S.-A. (1999) Pricing rate of return guarantees in a Heath-Jarrow-Morton framework. *Insurance: Mathematics and Economics* **25**, 307-325.

MILTERSEN, K.R. and PERSSON, S.-A. (2003) Guaranteed investment contracts: Distributed and undistributed excess return. *Scandinavian Actuarial Journal*. Forthcoming.

PERSSON, S.-A. (1998) Stochastic interest rate in life insurance: The principle of equivalence revisited. *Scandinavian Actuarial Journal*, 97-112.

PERSSON, S.-A. and AASE, K. (1997) Valuation of the minimum guaranteed return embedded in life insurance contracts. *Journal of Risk and Insurance* **64**, 599-617.

KNUT K. AASE and SVEIN-ARNE PERSSON
Norwegian School of Economics and Business Administration
5045 Bergen
Norway

COMMENT ON THE DISCUSSION ARTICLE BY AASE AND PERSSON

I applaud the article as it is exactly the type of reaction to my editorial in Astin Bulletin 32(2) that I hoped to provoke. Of course Aase and Persson's contribution is much more academic in style than mine which is more journalistic.

To my understanding there are three important messages on which every reader of their article should reflect.

Message 1: The kind of **NUM**eraire based **A**ctuarial **T**eaching (NUMAT) which I have advocated has been offered since the beginning of the 90ties at the Norwegian School of Economics and Business Administration in Bergen and there exists even an elementary textbook [1] covering that subject (unfortunately only in Norwegian language).

Message 2: The structure of Zero Coupon Prices needed is not so easy to get as my editorial suggested ("You can look it up in the Financial Part of your daily newspaper"). Of course my wording was a tribute to the journalistic style. To talk as a scientist I would like to mention that in a recent Diploma Thesis at ETH [2] we have interpolated from LIBOR Forward Short Rates and from SWAP Rates to get the Zero Coupon Prices. Obviously in any real world implementation there is the necessity for modelling the market prices and the daily newspaper does not suffice.

Message 3: Prices calculated by NUMAT are indeed volatile! This lesson should be learned by every actuary. It means that the products sold by Life Insurers have substantially varying market value.

As hinted in my editorial I would see the practical role of NUMAT for calculating Embedded Value rather than for calculating Premiums. I take again an example from the above mentioned Diploma Thesis:

Take an Endowment Policy ($x = 50$, $n = 5$, sum insured 50 000 CHF). According to the Swiss Table EKM 95 the yearly premium amounts to 9375.21 CHF. The value of the initial reserve by NUMAT amounts to

- CHF 2325.45 based on the Zero Coupon Structure of May 2000,
- CHF 11.67 based on the Zero Coupon Structure of November 2002.

As the classical initial net reserve is nil you get the Economic Value by changing the sign. Hence your pretty Embedded Value in May 2000 has disappeared in November 2002.

This is again a lesson to be learned by everybody who boosts with the "wonderful Embedded Value" in her/his Life Portfolio! Embedded Value is extremely volatile and needs to be monitored continuously.

ASTIN BULLETIN, Vol. 33, No. 2, 2003, pp. 123-124

The last message is my own:

Message 4: I hope to hear from other colleagues that they are already offering Numeraire based Actuarial Teaching. So far the signal from Bergen is the only one that I have received. Clearly, there are research papers where the economic value of insurance products is discussed and explicitly calculated. The point which I tried to make in my editorial was however: NUMAT should be part of the educational curriculum of **every actuary**. We academics are challenged to get the fundamental way of thinking over to the profession!

References

[1] K.K. Aase. Anvendt sannsynlighetsteorie: Forsikring Matematik, Carrelen Akademisk Forlag, Oslo, Norge 1996.
[2] G. Baumgartner. Fair Value für Lebensversicherungen. Diploma Thesis, ETH Zürich, 2003.

Hans Bühlmann
ETH Zürich

GUARANTEED ANNUITY OPTIONS

BY

PHELIM BOYLE AND MARY HARDY[1]

ABSTRACT

Under a guaranteed annuity option, an insurer guarantees to convert a policyholder's accumulated funds to a life annuity at a fixed rate when the policy matures. If the annuity rates provided under the guarantee are more beneficial to the policyholder than the prevailing rates in the market the insurer has to make up the difference. Such guarantees are common in many US tax sheltered insurance products. These guarantees were popular in UK retirement savings contracts issued in the 1970's and 1980's when long-term interest rates were high. At that time, the options were very far out of the money and insurance companies apparently assumed that interest rates would remain high and thus that the guarantees would never become active. In the 1990's, as long-term interest rates began to fall, the value of these guarantees rose. Because of the way the guarantee was written, two other factors influenced the cost of these guarantees. First, strong stock market performance meant that the amounts to which the guarantee applied increased significantly. Second, the mortality assumption implicit in the guarantee did not anticipate the improvement in mortality which actually occurred.

The emerging liabilities under these guarantees threatened the solvency of some companies and led to the closure of Equitable Life (UK) to new business. In this paper we explore the pricing and risk management of these guarantees.

1. INTRODUCTION

1.1. An introduction to guaranteed annuity options

Insurance companies often include very long-term guarantees in their products which, in some circumstances, can turn out to be very valuable. Historically these options, issued deeply out of the money, have been viewed by some insurers as having negligible value. However for a very long dated option, with a term of perhaps 30 to 40 years, there can be significant fluctuations in economic variables, and an apparently negligible liability can become very substantial. The case of guaranteed annuity options (GAOs) in the UK provides a dramatic illustration of this phenomenon.

[1] Both authors acknowledge the support of the National Science and Engineering Research Council of Canada.

Guaranteed annuity options have proved to be a significant risk management challenge for several UK insurance companies. Bolton et al (1997) describe the origin and nature of these guarantees. They also discuss the factors which caused the liabilities associated with these guarantees to increase so dramatically in recent years. These factors include a decline in long-term interest rates and improvements in mortality. For many contracts the liability is also related to equity performance and in the UK common stocks performed very well during the last two decades of the twentieth century.

Under a guaranteed annuity the insurance company guarantees to convert the maturing policy proceeds into a life annuity at a fixed rate. Typically, these policies mature when the policyholder reaches a certain age. In the UK the most popular guaranteed rate for males, aged sixty five, was £111 annuity per annum per £1000 of cash value, or an annuity:cash value ratio of 1:9 and we use this rate in our illustrations. If the prevailing annuity rates at maturity provide an annual payment that exceeds £111 per £1000, a rational policyholder would opt for the prevailing market rate. On the other hand, if the prevailing annuity rates at maturity produce a lower amount than £111 per £1000, a rational policyholder would take the guaranteed annuity rate. As interest rates rise the annuity amount purchased by a lump sum of £1000 increases and as interest rates fall the annuity amount available per £1000 falls. Hence the guarantee corresponds to a put option on interest rates. In Sections two, three and four we discuss the option pricing approach to the valuation of GAOs.

These guarantees began to be included in some UK pension policies in the 1950's and became very popular in the 1970's and 1980's. In the UK the inclusion of these guarantees was discontinued by the end of the 1980's but, given the long-term nature of this business, these guarantees still affect a significant number of contracts. Long-term interest rates in many countries were quite high in 1970's and 1980's and the UK was no exception. During these two decades the average UK long-term interest rate was around 11% p.a. The interest rate implicit in the guaranteed annuity options depends on the mortality assumption but based on the mortality basis used in the original calculations the break-even interest rate was in the region of 5%-6% p.a. When these options were granted, they were very far out of the money and the insurance companies apparently assumed that interest rates would never fall to these low levels again[2] and thus that the guarantees would never become active. As we now know this presumption was incorrect and interest rates did fall in the 1990's.

The guaranteed annuity conversion rate is a function of the assumed interest rate and the assumed mortality rate. Bolton et al note that when many of these guarantees were written, it was considered appropriate to use a mortality table with no explicit allowance for future improvement such as a(55). This is a mortality table designed to be appropriate for immediate life annuities purchased in 1955, but was still in vogue in the 1970s. However, there was a dramatic improvement in the mortality of the class of lives on which these guarantees were written during the period 1970-2000. This improvement meant

[2] Although interest rates were of this order for part of the 1960s.

that the break-even interest rate at which the guarantee kicked in rose. For example, for a 13-year annuity-certain, a lump sum of 1000 is equivalent to an annual payment of 111 p.a. at 5.70%. If we extend the term of the annuity to sixteen years the interest rate rises to 7.72%. Hence, if mortality rates improve so that policyholders live longer, the interest rate at which the guarantee becomes effective will increase. In Section 2 we will relate these break-even rates to appropriate UK life annuity rates.

1.2. A typical contract

To show the nature of the GAO put option we use standard actuarial notation, adapted slightly. Assume we have a single premium equity-linked policy. The contract is assumed to mature at T, say, at which date the policyholder is assumed to be age 65. The premium is invested in an account with market value $S(t)$ at time t, where $S(t)$ is a random process. The market cost of a life annuity of £1 p.a. for a life age 65 is also a random process. Let $a_{65}(t)$ denote this market price.

The policy offers a guaranteed conversion rate of $g = 9$. This rate determines the guaranteed minimum annuity payment per unit of maturity proceeds of the contract; that is, £1 of the lump sum maturity value must purchase a minimum of £$1/g$ of annuity.

At maturity the proceeds of the policy are $S(T)$; if the guarantee is exercised this will be applied to purchase an annuity of $S(T)/g$, at a cost of $(S(T)/g)\,a_{65}(T)$. The excess of the annuity cost over the cash proceeds must be met by the insurer, and will be

$$\frac{S(T)}{g}a_{65}(T) - S(T)$$

If $a_{65}(T) < 9$ the guarantee will not be exercised and the cash proceeds will be annuitized without additional cost.

So, assuming the policyholder survives to maturity, the value of the guarantee at maturity is

$$S(T)\ \max\left[\left(\frac{a_{65}(T)}{g} - 1\right),\ 0\right] \tag{1}$$

The market annuity rate $a_{65}(t)$ will depend on the prevailing long-term interest rates, the mortality assumptions used and the expense assumption. We will ignore expenses and use the current long-term government bond yield as a proxy for the interest rate assumption. We see that the option will be in-the-money whenever the current annuity factor exceeds the guaranteed factor, which is $g = 9$ in the examples used in this paper.

We see from equation (1) that, for a maturing policy, the size of the option liability will be proportional to $S(T)$: the amount of proceeds to which the guarantee applies. The size of $S(T)$ will depend on the nature of the contract and also on the investment returns attributed to the policy. The procedure by which the investment returns are determined depends on the terms of the policy. Under a traditional UK with profits contract profits are assigned using reversionary bonuses and terminal bonuses. Reversionary bonuses are assigned on a regular basis as guaranteed additions to the basic maturity value and are not distributed until maturity. Terminal bonuses are declared when the policy matures such that together with the reversionary bonuses, the investment experience over the term of the contract is (more or less) fully reflected. The size of the reversionary bonuses depends both on the investment performance of the underlying investments and the smoothing convention used in setting the bonus level. The terminal bonus is not guaranteed but during periods of good investment performance it can be quite significant, sometimes of the same order as the basic maturity sum assured. Bolton et al (1997) estimate that with profits policies account for eighty percent of the total liabilities for contracts which include a guaranteed annuity option. The remaining contracts which incorporate a guaranteed annuity option were mostly unit-linked policies.

In contrast to with profits contracts, the investment gains and losses under a unit-linked (equity-linked) contract are distributed directly to the policyholder's account. Contracts of this nature are more transparent than with profits policies and they have become very popular in many countries in recent years. Under a unit-linked contract the size of the option liability, if the guarantee is operative, will depend directly on the investment performance of the assets in which the funds are invested. In the UK there is a strong tradition of investing in equities and during the twenty year period from 1980 until 2000 the rate of growth on the major UK stock market index was a staggering 18% per annum.

In this paper we consider unit-linked policies rather than with profits. Unit-linked contracts are generally well defined with little insurer discretion. With profits policies would be essentially identical to unit-linked if there were no smoothing, and assuming the asset proceeds are passed through to the policyholder, subject to reasonable and similar expense deductions. However, the discretionary element of smoothing, as well as the opaque nature of the investment policy for some with profits policies make it more difficult to analyse these contracts in general. However, the methods proposed for unit-linked contracts can be adapted for with profits given suitably well defined bonus and asset allocation strategies.

1.3. Principal factors in the GAO cost

Three principal factors contributed to the growth of the guaranteed annuity option liabilities in the UK over the last few decades. First, there was a large decline in long-term interest rates over the period. Second, there was a significant improvement in longevity that was not factored into the initial actuarial calculations. Third, the strong equity performance during the period served to increase further the magnitude of the liabilities. It would appear that these events were not considered when the guarantees were initially granted. The responsibility for long-term financial solvency of insurance companies rests with the actuarial profession. It will be instructive to examine what possible risk management strategies could have been or should have been employed to deal with this situation. It is clear now with the benefit of hindsight that it was imprudent to grant such long-term open ended guarantees of this type.

There are three main methods of dealing with the type of risks associated with writing financial guarantees. First, there is the traditional actuarial reserving method whereby the insurer sets aside additional capital to ensure that the liabilities under the guarantee will be covered with a high probability. The liabilities are estimated using a stochastic simulation approach. The basic idea is to simulate the future using a stochastic model[3] of investment returns. These simulations can be used to estimate the distribution of the cost of the guarantee. From this distribution one can compute the amount of initial reserve so that the provision will be adequate, say, 99% of the time. The second approach is to reinsure the liability with another financial institution such as a reinsurance company or an investment bank. In this case the insurance company pays a fee to the financial institution and in return the institution agrees to meet the liability under the guarantee. The third approach is for the insurance company to set up a replicating portfolio of traded securities and adjust (or dynamically hedge) this portfolio over time so that at maturity the market value of the portfolio corresponds to the liability under the guaranteed annuity option.

Implementations of these three different risk management strategies have been described in the literature. Yang (2001) and Wilkie, Waters and Yang (2003) describe the actuarial approach based on the Wilkie model. Dunbar (1999) provides an illustration of the second approach. The insurance company, Scottish Widows offset its guaranteed annuity liabilities by purchasing a structured product from Morgan Stanley. Pelsser (2003) analyzes a hedging strategy based on the purchase of long dated receiver swaptions. This is described more fully in Section 8.

In this paper we will discuss a number of the issues surrounding the valuation and risk management of these guarantees. We will also discuss the degree to which different risk management approaches would have been possible from 1980 onwards.

1.4. Outline of the paper

The layout of the rest of the paper is as follows. Section two provides background detail on the guaranteed annuity options and the relevant institutional framework. We examine the evolution of the economic and demographic variables which affect the value of the guarantee. In particular we provide a time series of the values of the guarantee at maturity for a representative contract. In Section three we use an option pricing approach to obtain the market price of the guarantee. Section Four documents the time series of market values of the guarantee. Using a simple one-factor model it is possible to estimate the

[3] One such model, the Wilkie Model was available in the UK actuarial literature as early as 1980. See MGWP(1980).

market value of the option. Section five examines a number of the conceptual and practical issues involved in dynamic hedging the interest rate risk. Sections six and seven explore the issues involved in hedging the equity risk and the mortality risk. One suggestion for dealing with these guarantees involves the insurer purchasing long dated receiver swaptions. We describe this approach in Section eight. Section nine comments on the lessons to be learned from this episode.

2. Maturity Value of the Guarantee

In this section we document the evolution of the emerging liability under the guaranteed annuity option. Specifically we examine the magnitude of the guarantee for a newly maturing policy over the last two decades. In these calculations the policy proceeds at maturity are assumed to be held constant at £100, so the cost reported is the cost % of the policy maturity cash value. We assume the annuity purchased with the policy proceeds is payable annually in arrear to a life age 65 at maturity, and has a five year guarantee period.

We consider three different mortality tables to determine the GAO cost.

- The a(55) mortality table represents the mortality assumptions being used in the 1970's to price immediate annuities. As previously mentioned, it was calculated to be appropriate to lives purchasing immediate annuities in 1955.
- The PMA80(C10) table from this series is based on UK experience for the period 1979-1982 and is projected to 2010 to reflect mortality improvements. The PMA80 series became available in 1990.
- The PMA92 series was published in 1998 and the table we use, the PMA92 (C20) table, is based on UK experience for the period 1991-1994, projected to 2020 to reflect mortality improvements.

The improvement from the a(55) table to the PMA92(C20) provides an upper bound on the mortality improvement over the period since the last named table only appeared at the end of the period and includes a significant projection for future mortality improvements. Wilkie Waters and Yang (2003) give a detailed account of the year by year relevant UK male mortality experience for the period 1984 to 2001.

The increase in longevity is quite dramatic over the period covered by these three tables. The expectation of life for a male aged 65 is 14.3 years using a(55) mortality, 16.9 years under the PMA80(C10) table and 19.8 years under the PMA92(C20) table. Thus the expected future lifetime of a male aged 65 increased by 2.6 years from the a(55) table to the PMA80(C10) table. More dramatically the expected future lifetime of a male aged 65 increased by *over five years* from the a(55) table to the PMA92(C20) table.

On the basis of the a(55) mortality table, the break-even interest rate for a life annuity. is 5.61%. That is, on this mortality basis a lump sum of 1000 will purchase an annuity of 111 at an interest rate of 5.61% p.a. effective. The guarantee will be in the money if long-term interest rates are less than 5.61%.

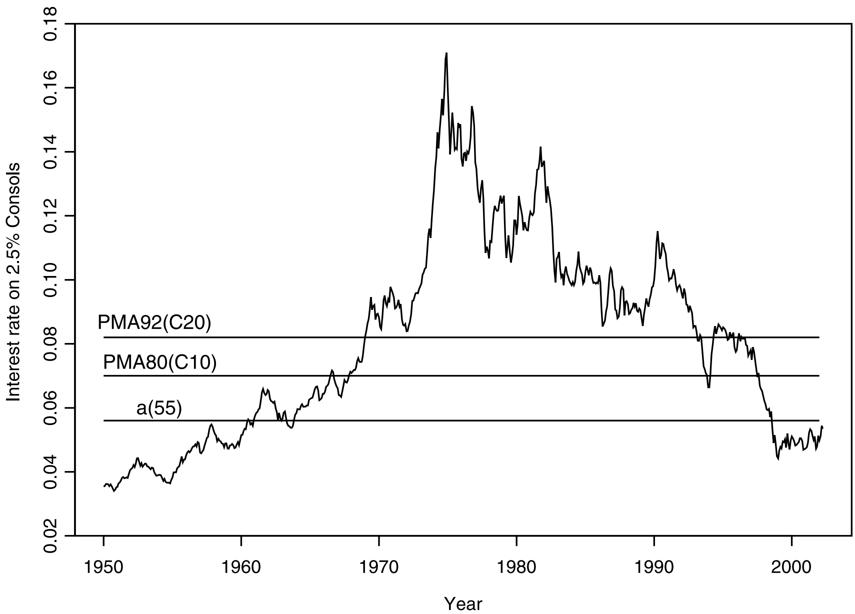


Fig 1: UK long-term interest rates 1950-2002, with interest rates levels that will trigger the guarantee for mortality tables a(55), PMA82(C10), PMA90(C20).

As a consequence of the mortality improvement the cost of immediate annuities increased significantly over this period independently of the impact of falling interest rates. Under the PMA80(C10) table the break-even rate is 7.0%, and under the PMA92(C20) table it is 8.2%.

Figure 1 illustrates the behavior of long-term interest rates in the UK since 1950. We note that rates rose through the later 1960s, remained quite high for the period 1970-1990 and started to decline in the 1990s. There was a large dip in long rates at the end of 1993 and long rates first fell below 6% in 1998 and have hovered in the 4%-6% range until the present[4] time. We also show in this figure the break-even interest rates for the GAO according to the three mortality tables.

Bolton et al (1997) provide extensive tables of the break even interest rates for different types of annuities and different mortality tables. They assume a two percent initial expense charge which we do not include. Thus in their Table 3.4 the value for the break even interest rate for a male aged 65 for an annuity of 111 payable annually in arrear with a five year guarantee is 5.9%. This is consistent with our figure of 5.6% when we include their expense assumption.

The increase in the level of the at-the-money interest rate has profound implications for the cost of a maturing guaranteed annuity option. For example, if the long-term market rate of interest is 5%, the value of the option for a

[4] Early in 2003 at the time of writing.

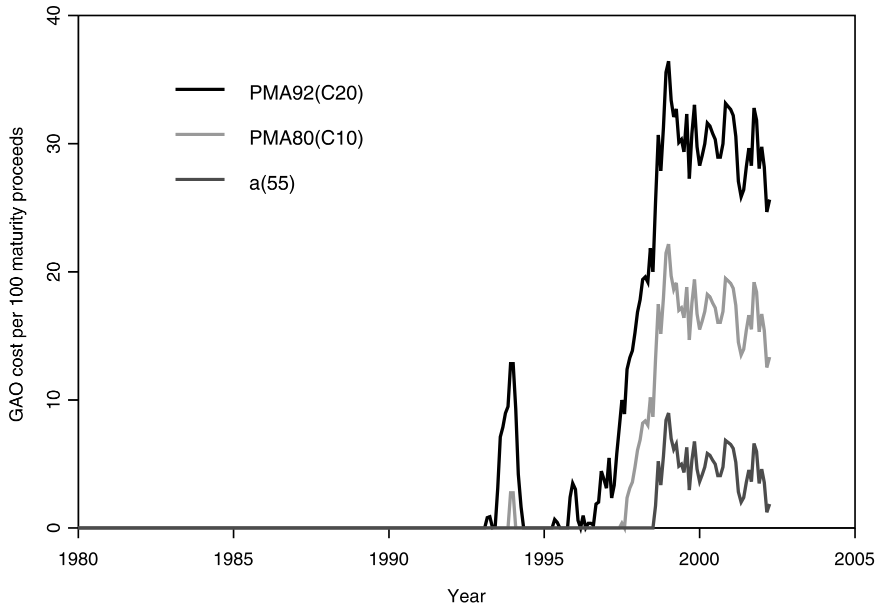


Fig 2: Value of maturing guarantee per 100 proceeds based on PMA92(C20), PMA80(C10), and a(55) mortality.

maturing policy with £100 maturity proceeds, based on a(55) is £4.45; based on PMA80(C10) it would be £16.53 and the corresponding value based on PMA92(C20) is £29.52. Note that we do not need any type of option formula to perform these calculations; we apply equation (1), with $S(T)$ fixed at 100, and $a_{65}(T)$ calculated using the appropriate mortality and an interest rate of 5% per year. Figure 2 shows the magnitude of the option liability for our benchmark contract under the three mortality assumptions using historical interest rates. Since the long-term interest rate is the main determinant of the annuity cost, we have used the yield on 2.5% consols, which are government bonds, generally considered irredeemable[5].

There is no liability on maturing contracts until the 1990's. Also note that the mortality assumption has a profound impact on the size of the liability.

We have already noted that during the period 1980-2000, UK equities performed extremely well. This resulted in increased levels of bonus to the with profits polices. For many contracts this meant that the volume of proceeds to which the guarantee applied also increased, thereby increasing the liability under the guarantee. In the case of unit-linked polices the gains are passed directly to the policyholder, apart from the various expenses. If we assume that a unit-linked contract earned the market rate of 18% minus 300 basis points this still leaves a return of 15%. At this growth rate an initial single

[5] In fact they may be redeemed at any time at the discretion of the Exchequer.

premium of £100 will accumulate to £1636.7 after twenty years. This growth would be proportionately reflected in the cost of the guarantee.

To summarize, we have discussed the evolution of the value of the liability for a sequence of maturing contracts. This analysis indicates how the three factors:

- The fall in long-term interest rates
- The improvement in mortality
- The strong equity performance

served to increase the cost of the guarantee. Note that our analysis in this section did not require any stochastic analysis or option pricing formula. We simply computed the value of the option at maturity each year of the period. In the next section we discuss the evaluation of these options prior to maturity.

3. Derivation of an Option Formula

3.1. Introduction

The guaranteed annuity option is similar to a call option on a coupon bond; the annuity payments and survival probabilities can be incorporated in the notional coupons.

First we develop the option formula without assuming a specific model for interest rates. Then, we will apply the Hull-White (1990) interest rate model (also known as extended Vasicek, from the Vasicek (1977) model) to calculate prices for the options.

We assume that the mortality risk is independent of the financial risk and that it is therefore diversifiable. In this case it is well documented (see, for example, Boyle and Schwartz (1977)) that it is appropriate to use deterministic mortality for valuing options dependent on survival or death.

3.2. The numeraire approach

Using stochastic interest rates, the price at some future date t of a zero coupon bond with unit maturity value, maturing at T is a random variable which we denote $D(t,T)$. The term structure of interest rates at t is therefore described by the function $\{D(t,T)\}_{T>t}$. This term structure is assumed to be known at time t.

So, the actuarial value at T of an immediate annuity payable to a life aged xr at T, contingent on survival, is

$$a_{xr}(T) = \sum_{j=1}^{J} {}_jp_{xr} D(T, T+j) \qquad (2)$$

where ${}_jp_{xr}$ represents the appropriate survival probability. For an annuity with an initial guarantee period of, say five years, we set the first five values of ${}_jp_{xr}$

to 1.0. The limiting age of the mortality table is denoted by ω and we set $J = (\omega - xr)$. Note that in valuing the annuity at T, the term structure is known, and there are no random variables in this expression.

Now, to value this expression at time $t < T$, we use a 'numeraire' approach. In the absence of arbitrage any market price deflated by a suitable numeraire is a martingale. We can use any traded asset which has a price that is always strictly positive as numeraire. Here we use the zero coupon bond which matures at time T as the numeraire and we denote the associated probability measure by the symbol Q_T. This is often called the 'forward measure'. If interest rates are constant, it is the same as the risk neutral measure in the standard Black-Scholes framework. See Björk (1998) for more details.

Suppose $V(s)$ is the market value at s of some payoff occurring at some time $T+j$, $j \geq 0$. We use $D(s,T)$ as the numeraire, where $s \leq T$. The martingale result means that $X_s = \frac{V(s)}{D(s,T)}$ is a martingale under Q_T, so that $E_{QT}[X_s | \mathcal{F}_t] = X_t$ for any s such that $t \leq s \leq T$. Here $| \mathcal{F}_t$ indicates that we are taking expectation of the random process at time $t + k$ given all the relevant information at t. In particular at t we know all values of $D(t,s)$, $s \geq t$.

Applying this to take expectation of the ratio $\frac{V(s)}{D(s,T)}$ at T, and using the fact that $D(T,T) = 1.0$, we have:

$$\frac{V(t)}{D(t,T)} = E_{Q_T}\left[\frac{V(T)}{D(T,T)} \middle| \mathcal{F}_t\right] = E_{Q_T}\left[V(T) | \mathcal{F}_t\right] \tag{3}$$

$$\Rightarrow V(t) = D(t,T) E_{Q_T}\left[V(T) | \mathcal{F}_t\right] \tag{4}$$

Equation (4) provides a valuation formula at t for any payoff $V(T)$. The distribution Q_T depends on the assumption made for interest rates, and we will discuss this later.

Now for the GAO we know from equation (1) that the payoff at maturity is[6]

$$V(T) = \frac{S(T)\left(a_{xr}(T) - g\right)^+}{g}$$

This is required for each policyholder surviving to time T, so to value at $t < T$ we multiply by the appropriate survival probability.

Then if $G(t)$ is the value of this benefit at t, and letting $x = xr - (T - t)$ we have:

$$G(t) = {}_{T-t}p_x D(t,T) E_{Q_T}\left[V(T) | \mathcal{F}_t\right] \tag{5}$$

$$\Rightarrow G(t) = {}_{T-t}p_x D(t,T) E_{Q_T}\left[\frac{S(T)\left(a_{xr}(T) - g\right)^+}{g} \middle| \mathcal{F}_t\right] \tag{6}$$

[6] We use $(X)^+ = \max(X, 0)$.

Initially we assume that $S(T)$ is independent of interest rates. This is a very strong assumption but it simplifies the analysis. Later we allow for correlation between equity returns and interest rates.

We have then:

$$\begin{aligned} G(t) &= \frac{{}_{T-t}p_x D(t,T) E_{Q_T}[S(T)]}{g} E_{Q_T}\left[(a_{xr}(T) - g)^+ \,\middle|\, \mathcal{F}_t \right] \\ &= \frac{{}_{T-t}p_x [S(T)]}{g} E_{Q_T}\left[(a_{xr}(T) - g)^+ \,\middle|\, \mathcal{F}_t \right] \end{aligned}$$

The last line follows from the numeraire martingale result, equation (4), because replacing $V(T)$ with $S(T)$ in that equation gives

$$\frac{S(t)}{D(t,T)} = E_{Q_T}[S(T) \mid \mathcal{F}_t]$$

Inserting the expression for $a_{xr}(T)$ from (2) we have

$$E_{Q_T}\left[(a_{xr}(T) - g)^+ \,\middle|\, \mathcal{F}_t \right] = E_{Q_T}\left[\left(\sum_{j=1}^{J} {}_j p_{xr} D(T, T+j) - g \right)^+ \,\middle|\, \mathcal{F}_t \right]$$

The expression inside the expectation on the right hand side corresponds to a call option on a coupon paying bond where the 'coupon' payment at time $(T+j)$ is ${}_j p_{xr}$. This 'coupon bond' has value at time, T:

$$\sum_{j=1}^{J} {}_j p_{xr} D(T, T+j).$$

The market value at time, t of this coupon bond is

$$P(t) = \sum_{j=1}^{J} {}_j p_{xr} D(t, T+j).$$

So $P(t)$ is the value of a deferred annuity, but without allowance for mortality during deferment. With this notation our call option has a value at time, T of $(P(T) - g)^+$. The numeraire approach is described more fully in Björk (1998).

3.3. Using Jamshidian's method for coupon bond options

Jamshidian (1989) showed that if the interest rate follows a one-factor process, then the market price of the option on the coupon bond with strike price g is equal to the price of a portfolio of options on the individual zero coupon bonds with strike prices K_j, where $\{K_j\}$ are equal to the notional zero coupon bond prices to give an annuity $a_{xr}(T)$ with market price g at T. That is, let r_T^* denote the value of the short rate for which

$$\sum_{j=1}^{J} {}_jp_{xr} D^*(T, T+j) = g \tag{7}$$

where we use the asterisk to signify that each zero coupon bond is evaluated using short rate r_T^*[7]. Then set

$$K_j = D^*(T, T+j).$$

Then the call option with strike g on the coupon bond $P(t)$ can be valued as

$$C[P(t), g, t] = \sum_{j=1}^{J} {}_jp_{xr} C\left[D(t, T+j), K_j, t\right],$$

where $C[D(t,T+j), K_j, t]$ is the price at time t of a call option on the zero coupon bond with maturity $(T+j)$ and strike price K_j.

We can use the call option $C[P(t), g, t]$ to obtain an explicit expression for the GAO value at t, $G(t)$. Recall that

$$G(t) = \frac{{}_{T-t}p_x S(T)}{g} E_{Q_T}\left[(P(T) - g)^+ \,\middle|\, \mathcal{F}_t\right].$$

From the numeraire valuation equation we have

$$\frac{C[P(t), g, t]}{D(t, T)} = E_{Q_T}\left[(P(T) - g)^+ \,\middle|\, \mathcal{F}_t\right]$$

Pulling all the pieces together we have

$$G(t) = \frac{{}_{T-t}p_x S(t)}{g} \frac{\sum_{j=1}^{J} {}_jp_{xr} C\left[D(t, T+j), K_j, t\right]}{D(t, T)}. \tag{8}$$

3.4. Applying the Hull-White interest rate model

Jamshidian's result requires a one-factor interest rate model. We use a version of the Hull-White (1990) model. This model is also known as extended Vasicek, from the Vasicek (1977) model.

The short rate of interest at t is assumed to follow the process:

$$dr(t) = \kappa(\theta(t) - r(t))\,dt + \sigma dW_t \tag{9}$$

---

7 In a one-factor model, setting the short rate determines the entire term structure.

where $\theta(t)$ is a deterministic function determined by the initial term structure of interest rates. Using the function $\theta(t)$ enables us to match the model term structure and the market term structure at the start of the projection.

Björk (1998) gives the formula for the term structure at t, using the market term structure at initial date $t = 0$, as:

$$D(t,T) = \frac{D(0,T)}{D(0,t)} \exp\left\{B(t,T)f^*(0,t) - \frac{\sigma^2}{4\kappa} B(t,T)^2(1-e^{-2\kappa t}) - B(t,T)r(t)\right\} \tag{10}$$

where $f^*(0,t)$ is the t-year continuously compounded forward rate at $t = 0$, and

$$B(t,T) = \frac{1}{\kappa}\left\{1 - e^{-\kappa(T-t)}\right\}$$

This formula can be used to identify the strike price sequence $\{K_j\}$ from equation (7), and also used to value the option for given values of $r(t)$.

The explicit formula for each individual bond option under the Hull-White model is

$$C\left[D(t,T+J),K_j,t\right] = D(t,T+j)N\left(h_1(j)\right) - K_j D(t,T)N\left(h_2(j)\right),$$

where

$$h_1(j) = \frac{\log \frac{D(t,T+j)}{D(t,T)K_j}}{\sigma_P(j)} + \frac{\sigma_P(j)}{2},$$

$$h_2(j) = \frac{\log \frac{D(t,T+j)}{D(t,T)K_j}}{\sigma_P(j)} - \frac{\sigma_P(j)}{2},$$

and

$$\sigma_P(j) = \sigma\sqrt{\frac{1-e^{-2\kappa(T-t)}}{2\kappa}}\frac{\left(1-e^{-\kappa j}\right)}{\kappa}.$$

The parameters κ and σ characterize the dynamics of the short rate of interest under the Hull-White process.

4. Valuation of the Guaranteed Annuity Option

In this section we will derive the historical time series of market values for the guarantee based on the formula derived in the last section. It would have been helpful if the UK insurance companies had computed these market values at regular intervals since they would have highlighted the emergence of the liability under the guaranteed annuity option. The technology for pricing interest rate options was in its infancy in 1980 but by 1990 the models we use were in the public domain. An estimate of the market value of the guarantee can be derived from the one-factor stochastic interest rate model. We will use the model to estimate the value of the guarantees for the period 1980-2002.

In the previous section we derived a formula for the market price of the guaranteed annuity option using the Hull-White model. A similar formula has also been derived by Ballotta and Haberman (2002). They start from the Heath Jarrow Morton model and then restrict the volatility dynamics of the forward rate process to derive tractable formulae.

We used the following parameter estimates to compute the market values of the guaranteed annuity option

Parameter	Value
κ	0.35
σ	0.025

These parameters are broadly comparable with estimates that have been obtained in the literature based on UK data for this time period. See Nowman (1997) and Yu and Phillips (2001).

Figure 3 illustrates how the market value of the option as percentage of the current fund value changes over time. We assume that the option has remaining time to maturity of ten years and that $xr = 65$ so that the age of the policyholder at the option valuation date is $x = 55$. We ignore the impact of lapses and expenses and we assume that all policyholders will take up the option at maturity if it is in their interest. The term structure at each date is obtained by assuming that the 2.5% consol yield operates for maturities of five years and longer. Yields for maturities one to five are obtained by a linear interpolation between the Treasury Bill rate and the five year rate. While this procedure does not give the precise term structure at a given time it captures the shape of the term structure at that time.

We assume two different mortality models, a(55), representing the outdated mortality in common use in the 1980s, and PMA92(C20), representing a reasonable mortality assumption for contracts on lives currently age 55.

Figure 3 shows the very substantial effect of the mortality assumption. By comparison with Figure 2, we see a very similar pattern of liability, but instead of considering the maturity value of the contract, known only at the retirement age, this figure plots the option value ten years before retirement. The similarity of these two figures shows that the simple Hull-White model may be useful for determining the market price of the option, even with the simplifying assumption of independence of stock prices and interest rates.

Figure 4 illustrates how the cost of the guaranteed annuity option for our benchmark contract varies with the volatility assumption. We used three different volatility assumptions: $\sigma = .015, .025, .035$. The market values are relatively insensitive to the volatility assumption for long periods. Indeed the only periods where we can distinguish the three separate curves corresponds to periods when the long-term interest rate is close to the strike price of the option. We know from basic Black Scholes comparative statics that the sensitivity of an option value to the volatility is highest when the underlying asset

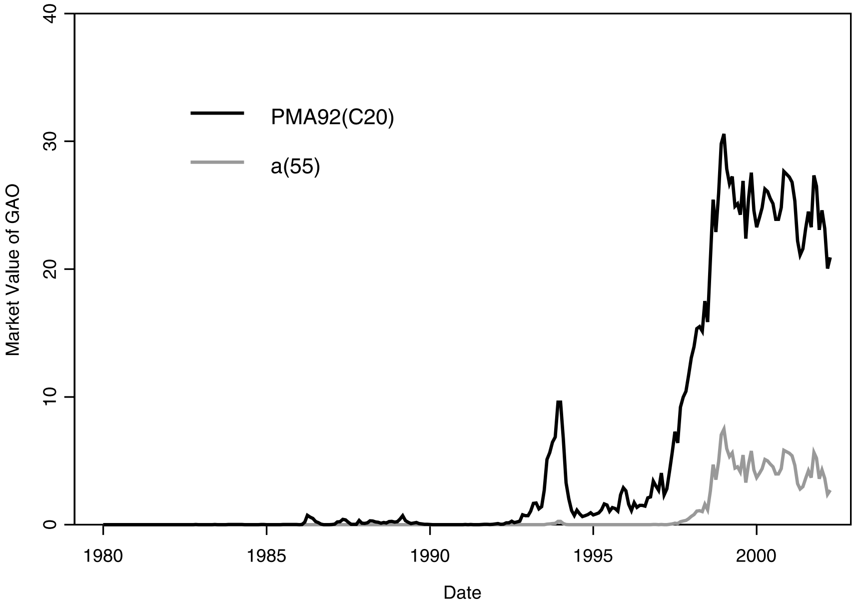


Figure 3: Historical market values of GAO with 10 years to maturity, using a(55) and PMA92(C20) mortality, per £100 of current portfolio value. Stock prices are assumed to be independent of interest rates. Hull-White model parameters $\kappa = 0.35$, $\sigma = .025$.

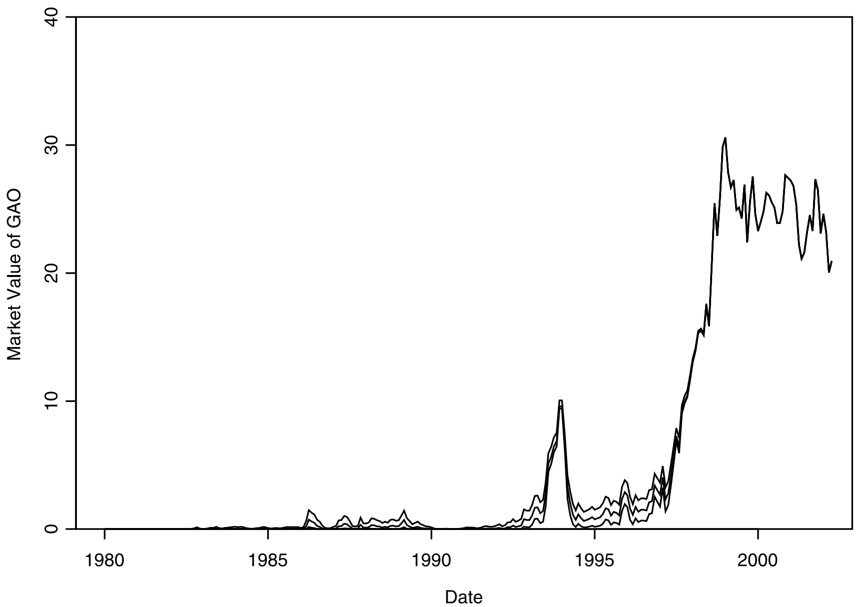


Figure 4: Market values of GAO with 10 years to maturity using $\sigma = 0.015, 0.025, 0.035$; cost per £100 current portfolio value. Stock prices are assumed to be independent of interest rates. $\kappa = 0.35$; PMA92(C20) mortality.

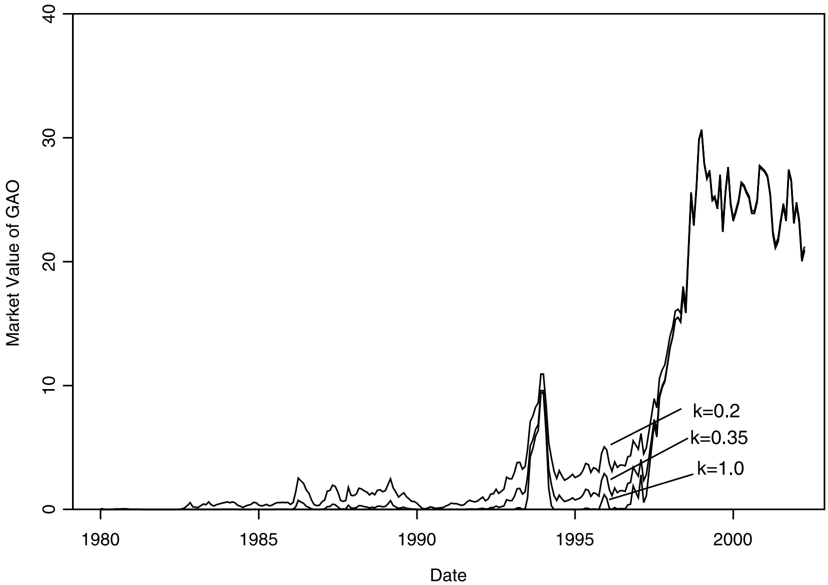


Figure 5: Market values of GAO with 10 years to maturity using $\kappa = 0.2, 0.35, 1.00$; cost per £100 current portfolio value. Stock prices are assumed to be independent of interest rates. $\sigma = 0.025$; PMA92(C20) mortality.

price is close to the strike price. If the option is very far out of the money or deeply in the money the price of the option is relatively insensitive to the volatility assumption. This same intuition is at work here.

The option value also depends on the assumed value of the autoregression factor κ. Figure 5 shows how the cost varies with three values; $\kappa = 0.2$, 0.35 and, 1.0. Holding everything else equal, smaller values for κ lead to an increase in the option price. This is because as κ decreases the overall volatility increases — periods of low interest are more likely to be prolonged with high autocorrelation — and higher volatility causes the option price to increase. Note that for $\kappa = 0.2$ the GAO price is much larger than the benchmark value (based on $\kappa = 0.35$) for the out-of-the-money case.

At first sight it may be surprising that a simple one-factor model can give at least an approximate estimate of the market value of the option (per £100 portfolio value), since actual interest rate dynamics are much too complicated to be captured by such a model. One reason is that we have calibrated the model to the input term structure so that it reproduces the market prices of all the zero coupon bonds. In addition the prices are fairly robust to the volatility assumption for the realized market conditions, though possibly less robust to the autoregression assumptions, particularly for the out-of-the-money periods. However we stress that such a simple model will not be adequate for hedging purposes. Hull (2002) notes

> "The reality is that relatively simple one-factor models, if used carefully, usually give reasonable prices for instruments, but good hedging schemes must explicitly or implicitly assume many factors."

5. Hedging

In this section we discuss some of the issues involved in hedging the guaranteed annuity risk using traded securities. Although the full fledged guaranteed annuity option depends on three stochastic variables — interest rates, stock prices and mortality — here we just discuss the interest rate risk.

The hedging of long-term interest rate options is a difficult task. In order to implement an effective hedging strategy we require a robust and reliable model of interest rate dynamics over the long-term. The search for such a model remains an area of active research and, despite some useful progress, there appears to be no consensus on such a model. For a survey of some of the recent work in modelling term structure dynamics see Dai and Singleton (2003). Note that for risk management and hedging purposes we require a model that provides a good description of the actual movements in yield curves over time. In other words we need a model that describes interest rate movements under the real world measure. For pricing we only needed the Q-measure, with parameters determined from the current market structure.

We begin by reviewing the relationship between pricing and hedging in an ideal setting. Consider the standard no arbitrage pricing model, where there is a perfect frictionless market with continuous trading. If the market is complete, then any payoff can be hedged with traded securities. Since there is no arbitrage the current price of the derivative must be equal to the current price of the replicating portfolio. If an institution sells this derivative then it can take the premium (price) and set up the replicating portfolio. As time passes it can dynamically adjust the position, so that at maturity the value of the replicating portfolio is exactly equal to the payoff on the derivative. In an ideal world where the model assumptions are fulfilled, it should be possible to conduct this replication program without needing any additional funds. The initial price should be exactly sufficient.

In the real world the assumptions of these models are never exactly fulfilled. For example

- The asset price dynamics will not be correctly specified.
- It will not be feasible to rebalance the replicating portfolio on a continuous basis. Instead it has to be rebalanced at discrete intervals.
- There are transaction costs on trading.

The impact of these deviations from the idealized assumptions has been explored in the Black Scholes Merton world. We discuss these three possible deviations in turn.

If the process that generates the market prices deviates from the model implicit in the pricing formula there will be additional hedging errors. This is because the portfolio weights that would be required to replicate the payoff under the true model will be different from the portfolio weights computed under the assumed model. This point has been explored in the case of equity derivatives by several authors including Bakshi, Cao and Chen (1997, 2000), Chernov et al (2001) and Jiang and Oomen (2002), and in the case of equity-linked life insurance by Hardy (2003).

With discrete rebalancing, Boyle and Emanuel (1980) showed that if the portfolio is rebalanced at discrete intervals, there will be a hedging error which tends to zero as the rebalancing becomes more frequent. In the presence of transaction costs the frequency of rebalancing involves a trade-off between the size of the hedging error and the trading costs.

However in practice we may emphasize pricing at the expense of hedging by calibrating an *incorrect* model to give the accurate market price of a derivative. For example, quoted swaption and cap prices are universally based on the simple Black model. The Black model volatility that makes the market price equal to the model price has become a standard measure for conveying the price. However the Black model does not provide realistic dynamics for interest rates and so it is unsuitable for hedging and risk management applications. In the same way stock option prices when the asset price dynamics follow a process with stochastic volatility can still be quoted in terms of the Black Scholes implied volatility. We can always find the value of the Black Scholes volatility that reproduces the market price of the option even when the true dynamics include stochastic volatility. However as shown by Melino and Turnbull (1995) the use of the simple Black Scholes model, in the presence of stochastic volatility, may lead to large and costly hedging errors, especially for long dated options.

In the case of stochastic interest rates, several studies have shown that it is possible to have a simple model that does a reasonable job of pricing interest rate derivatives even though the model is inadequate for hedging purposes. Canabarro (1995) uses a two factor simulated economy to show that, although one-factor models produce accurate prices for interest rate derivatives, these models lead to poor hedging performance. Gupta and Subrahmanyam (2001) use actual price data to show that, while a one-factor model is adequate for pricing caps and floors, a two factor model performs better in hedging these types of derivatives.

Litterman and Scheinkman (1991) demonstrated that most of the variation in interest rates could be explained by three stochastic factors. Dai and Singleton (2000) examine three factor models of the so called affine class. The classical Cox Ingersoll Ross (1985) model and the Vasicek model are the best known examples of the affine class. These models have the attractive property that bond prices become exponentials of affine functions and are easy to evaluate. Dai and Singleton find reasonable empirical support for some versions of the three factor affine model using swap market data for the period April 1987 to August 1996.

In the context of guaranteed annuity options we require an interest rate model that describes interest rate behavior over a longer time span. Ahn, Dittmar and Gallant (2002) provide support for quadratic term structure models. They are known as quadratic models because the short term rate of interest is a quadratic function of the underlying state variables. Their empirical tests use US bond data for the period 1946-1991 and they conclude that the quadratic three factor model

> "...provides a fairly good description of term structure dynamics and captures these dynamics better than the preferred affine term structure model of Dai and Singleton."

Bansal and Zhou (2002) show that the affine models are also dominated by their proposed regime switching model. Their empirical test is based on US interest rate data for the period 1964-1995. Even a casual inspection of the data suggests the existence of different regimes. They conclude that standard models, including the affine models with up to three factors, are sharply rejected by the data. Regime switching models have been extensively used by Hardy (2003) to model equity returns in the context of pricing and risk management of equity indexed annuities.

The interest rate exposure in a guaranteed annuity option is similar to that under a long dated swaption. Hence it is instructive to examine some recent results on hedging swaptions. This is a topic of current interest as evidenced by papers by Andersen and Andreasen (2002), Fan, Gupta and Ritchken (2001), Driessen, Klaasen and Melenberg (2002), and Longstaff, Santa-Clara and Schwartz (2001). The main conclusion of these papers is that multi-factor models are necessary for good hedging results. However it should be noted that the empirical tests in these papers tend to use relatively short observation periods — around three to five years being typical. Swaption data is unavailable for long periods since the instruments first were created in the late 1980's. Hence these models are being tested over the 1995-2000 period when interest rates were fairly stable. If the swaption data were available over longer periods, it seems likely that a regime switching interest rate model would be required to do an adequate hedging job.

In Section 6 we consider the GAO hedge using highly simplified assumptions for equities and interest rates.

6. The Equity Risk

We have already shown in equation (1) that the size of the payoff on the UK guaranteed annuity option is proportional to the amount of the maturity proceeds. This amount will depend on the stock market performance over the life of the contract. For unit-linked policies the maturity amount depends directly on the performance of the underlying fund. In the case of with profits contracts the policy proceeds depend on the bonuses declared by the insurance company and the size of these bonuses is directly related to stock market performance. In general the stock market returns are passed through to the policyholders in reversionary and terminal bonuses. In this section we discuss how the inclusion of equity risk impacts the pricing and risk management of the guaranteed annuity options.

We have seen in Section 3 that under some strong assumptions about the joint dynamics of interest rates we can obtain simple pricing formula. Specifically if we assume that equity returns are independent of interest rates, and that interest rates are governed by a one-factor Hull-White model, we can obtain a simple valuation formula for the price of the guaranteed annuity option. The formula can be modified to handle the case when there is correlation between stocks and bonds. Ballotta and Haberman have also derived a formula under these assumptions.

We can illustrate the key issues involved in pricing and hedging when the equity risk is included by considering a simpler contract than the guaranteed annuity option. This contract has the following payoff at time T

$$P(T, T+j) = S(T) \max\left(D(T, T+j) - K_j, 0\right)$$

It corresponds to an option on the zero coupon bond which matures at time $(T+j)$ and where the payoff is directly related to the value of the reference index. This contract includes no mortality risk here and there is just one zero coupon bond at maturity rather than a linear combination of zero coupon bonds. However this simpler contract captures the key dependencies of the guaranteed annuity option. The full guaranteed annuity option with correlation can then be determined, using $D(t,T)$ as numeraire, and the forward measure Q_T as the equivalent martingale measure:

$$V(t) = D(t, T) \sum_{j=1}^{J} {}_{T+j-t}p_x E_{Q_T}\left[P(T, T+j)\right]$$

We can derive a closed form expression for each term in $V(t)$ if, for example, we assume that under the forward measure Q_T the random variables $S(T)$ and $D(T, T+j)$ have a bivariate lognormal distribution.

For simplicity we treat the period from t to T as a single time step, with variance-covariance matrix for $\log S(T) | \mathcal{F}_t$ and $\log D(T, T+j) | \mathcal{F}_t$:

$$\Sigma_{\mathbf{j}} = \begin{bmatrix} \sigma_S^2 & \rho\sigma_D\sigma_S \\ \rho\sigma_D\sigma_S & \sigma_D^2 \end{bmatrix}.$$

Note that the values of ρ and σ_D would both vary with j.
Assume also that under the forward measure Q_T the means of $\log S(T)/S(t) | \mathcal{F}_t$ and $\log D(T, T+j) / D(t, T+j) | \mathcal{F}_t$ are μ_S and μ_D respectively.
From the martingale numeraire property, we know that

$$E_{Q_T}\left[\frac{D(T, T+j)}{D(T, T)} \middle| \mathcal{F}_t\right] = \frac{D(t, T+j)}{D(t, T)}$$

and from the lognormal distribution we know

$$E_{Q_T}\left[D(T, T+j) \middle| \mathcal{F}_t\right] = D(t, T+j)\, e^{\mu_D + \sigma_D^2/2}$$

so that

$$e^{\mu_D + \sigma_D^2/2} = \frac{1}{D(t, T)}$$

Similarly, $e^{\mu_S + \sigma_S^2/2} = \dfrac{1}{D(t, T)}$

This is analogous to the risk-free property in a standard Black-Scholes framework, that all assets have the same expected accrual under the Q-measure.

Using properties of the bivariate lognormal distribution (see, for example, chapter four of Boyle et al (1998)), we have

$$\begin{aligned} E_{Q_T}\left[P(T, T+j) \mid \mathcal{F}_t\right] = \\ & S(t)\, e^{\mu_S + \sigma_S^2/2} \left\{ D(t, T+j)\, e^{\mu_D + \sigma_D^2/2 + \rho\sigma_S\sigma_D} N(h_3) - K_j N(h_4) \right\} \qquad (11) \\ & = \frac{S(t)}{D(t,T)} \left\{ \frac{D(t,T+j)}{D(t,T)} e^{\rho\sigma_S\sigma_D} N(h_3) - K_j N(h_4) \right\} \end{aligned}$$

where $h_3 = \dfrac{\log \frac{D(t,T+j)}{D(t,T)K_j} + \rho\sigma_D\sigma_S}{\sigma_D} + \dfrac{\sigma_D}{2}$ and $h_4 = h_3 - \sigma_D$

So the option price at t for the annuity payment at $T+j$ is

$$ {}_{T+j-t}P_x\, S(t) \left\{ \frac{D(t,T+j)}{D(t,T)} e^{\rho\sigma_S\sigma_D} N(h_3) - K_j N(h_4) \right\} \qquad (12)$$

Formula (12) incorporates both equity risk and interest rate risk. The fact that the option is proportional to the equity $S(t)$ arises because it is a form of quanto option, where the payoff is essentially in units of stock rather than in cash. Wilkie et al (2003) discuss this quanto feature. The moneyness of the option is entirely related to the interest rate risk. Once the option is in the money, then every extra £ of stock accumulation increases the cost of the GAO proportionately. This leads to perverse incentives for the insurer acting as fund manager.

We see that the option price is an increasing function of the correlation coefficients ρ. Indeed the price is quite sensitive to the values of ρ and for plausible parameter values the option price with $\rho = 0$ is roughly double that using $\rho = -1$, and is around one-half the price using $\rho = 1$, for all j. Correlations are notoriously difficult to forecast and so we conclude that when the equity risk is assumed to be correlated with the interest rate risk, pricing the option becomes more difficult. Of course this modifies our earlier conclusions about the effectiveness of a one-factor model in pricing the guaranteed annuity option. Our earlier model assumed that the stock price movements were independent of interest rate movements.

It is now well established in the empirical literature that equity prices do not follow a simple lognormal process. There is mounting evidence that some type of stochastic volatility model does a better job of modelling equity returns. Hardy (2003) provides evidence that regime switching model does a good job of fitting the empirical distribution of monthly stock returns. Andersen Benzoni and Lund (2002) demonstrate that both stochastic volatility and jump components are present in the S&P daily index returns. Several authors[8] have

[8] See Bakshi, Cao and Chen (1997, 2000).

shown that these models may produce significant pricing deviations from the lognormal Black Scholes option prices. This deviation depends to some extent on the moneyness and the term of the option. It will be less critical for shorter, in-the-money options which are not very sensitive to the volatility.

When we turn to hedging matters become worse. There are two reasons. First we require a good model of the joint dynamics of bonds and equities that will be robust over long time periods. There appears to be no obvious model that would fulfill these requirements. The lognormal assumption for the zero coupon bond prices is a particular problem when, in practice, these prices are highly auto-correlated. Second even if we are willing to adopt the pricing model in (12) the resulting hedging implementation leads to some practical problems.

To hedge the option based on this simple model we would need to invest in three securities. The first is an investment in the underlying equity index equal to the current market value of the option. We denote the number of units invested in the index by $H_1(t)$ where

$$H_1(t) = \left(\frac{D(t, T+j)\, e^{\rho\sigma_D \sigma_S}}{D(t, T)} N(h_3) - K_j N(h_4) \right)$$

The second consists of an investment of $H_2(t)$ units of the zero coupon bond which matures at time $(T+j)$, where

$$H_2(t) = S(t) \frac{e^{\rho\sigma_D \sigma_S}}{D(t, T)} N(h_3)$$

The third consists of an investment of $H_3(t)$ units of the zero coupon bond which matures at time (T), where

$$H_3(t) = -S(t) \frac{D(t, T+j)\, e^{\rho\sigma_D \sigma_S}}{D(t, T)^2} N(h_3)$$

Note that the value of the initial hedge is

$$H_1(t)\, S(t) + H_2(t)\, D(t, T+j) + H_3(t)\, D(t, T) = S(t)\, H_1(t),$$

which is equal to the initial price of the option since the last two terms on the left hand side cancel one another.

Suppose the hedge is to be rebalanced at time $(t+h)$. Just before rebalancing the value of the hedge portfolio is

$$H_1(t)\, S(t+h) + H_2(t)\, D(t+h, T+j) + H_3(t)\, D(t+h, T)$$

where $S(t+h)$, $D(t+h, T+j)$, $D(t+h, T)$ denote the market prices at time $(t+h)$ of the three hedge assets. The new hedging weights $H_i(t+h)$, $i = 1, 2, 3$ are computed based on these new asset prices and the value of the revised hedge is

$$H_1(t+h)\,S(t+h) + H_2(t+h)\,D(t+h, T+j) + H_3(t+h)\,D(t+h, T).$$

If the value of the hedge portfolio after rebalancing increases funds need to be added. If the value of the hedge portfolio after rebalancing goes down funds can be withdrawn. In an idealized world the hedge would be self financing. However in practice hedging is done discretely, there are transactions costs and the market movements can deviate significantly from those implied by the model. These slippages can lead to considerable hedging errors. Some numerical examples using a similar approach are given in Hardy (2003).

7. The Mortality Risk

We noted earlier that there was a dramatic improvement in annuitants mortality over the relevant period. This improvement was not anticipated when the contracts were designed. The effect of this improvement was to increase the value of the interest rate guarantee by raising the threshold interest rate at which the guarantee became effective. The structure of the guaranteed annuity option means that the policyholder's option is with respect to two random variables: future interest rates and future mortality rates. To isolate the mortality option, suppose that all interest rates are deterministic but that future mortality rates are uncertain. In this case the option to convert the maturity proceeds into a life annuity is an option on future mortality rates. If, on maturity, the mortality rates have improved[9] above the level assumed in the contract the policyholder will obtain a higher annuity under the guarantee. On the other hand if life expectancies are lower than those assumed in the contract the guarantee is of no value since policyholder will obtain a higher annuity in the open market. When interest rates are stochastic the mortality option interacts with the interest rate option as we saw in Section Two.

Milevsky and Promsilow (2001) have recently analyzed the twin impacts of stochastic mortality and stochastic interest rates in their discussion of guaranteed purchase rates under variable annuities in the United States. They model the mortality option by modeling the traditional actuarial force of mortality as a random variable. The expectation of this random variable corresponds to the classical actuarial force of mortality. They show that, under some assumptions, the mortality option can — in principle — be hedged by the insurance company selling more life insurance. The intuition here is that if people live longer the losses on the option to annuitize will be offset by profits on the life policies sold.

There may be some practical difficulties in implementing this. First, it may not be possible to sell the insurance policies to the same type of policyholders who hold the GAO contracts. Secondly, to implement the mortality hedging strategy the insurer requires a good estimate of the distribution of future mortality. Harking back to the UK case it would have been most unlikely for any

[9] When we write about mortality rates improving we mean that people are living longer.

insurer in the 1970's to predict accurately the distribution of future mortality rates. If the insurer has a sufficiently accurate estimate of the distribution of the future mortality rates to conduct an effective hedging strategy, then it should be able to project future mortality improvements to minimize the mortality risk under the guarantee.

The mortality risk exposure facing insurers under the guaranteed annuity options could have been eliminated at inception by a different contract design. Instead of guaranteeing to pay a fixed annual amount the insurer could have guaranteed to use a certain pre specified interest rate in conjunction with the prevailing mortality assumption (in use at the time of retirement). Under this revised contract design there is no additional liability incurred if mortality improves and annuities become more expensive. This adjustment would have significantly reduced the liabilities under the guaranteed annuity options.

8. Hedging with Swaptions

Swaps have became enormously important financial instruments for managing interest rate risks. They are often more suitable than bonds for hedging interest rate risk because the swap market is more liquid. Furthermore, while it can be difficult to short a bond, the same exposure can easily be arranged in the swap market by entering a payer swap. Options to enter swap contracts are known as swaptions and there is now a very liquid market in long dated swaptions, where the option maturities can extend for ten years and the ensuing swap can last for periods up to thirty years. Pelsser (2003) shows that long dated receiver swaptions are natural vehicles for dealing with the interest rate risk under guaranteed annuity options. In this section we discuss the feasibility of using this approach.

Upon maturity, the owner of the swaption will only exercise it if the option is in the money. Suppose the swaption gives its owner the option to enter a receiver swap when the swaption matures. The counter party that enters (or is long) a receiver swap agrees to pay the floating interest rate (e.g. Libor or Euribor) and in return receive the fixed rate: known as the *swap rate*. If a firm owns a receiver swaption with a strike price of 7% it will compare the market swap rate with the strike rate when the swaption reaches maturity. For example if the market swap rate at maturity is 5% then the firm should optimally exercise the swaption because the guaranteed rate of 7% provides a better deal. It is preferable to receive fixed rate coupons of 7% than the market rate of 5%. By entering a receiver swaption an institution protects itself against the risk that interest rates will have fallen when the swaption matures. This is exactly the type of interest rate risk exposure in the guaranteed annuity option.

Pelsser shows how to incorporate mortality risk to replicate the expected payoff under the guaranteed annuity option. He assumes, as we do, that the mortality risk is independent of the financial risk and that the force of mortality (hazard rate) is deterministic. He derives an expression for the price of the guaranteed annuity option as a portfolio of long dated receiver swaptions. The advantage of his approach is that the swaptions incorporate the right type of interest rate options. Pelsser calls this approach the static hedge since there is no need for dynamic hedging. This is an advantage given the difficulty of hedging long-term interest rate options with more basic securities such as bonds and swaps. However the swaption approach still has problems in dealing with the stock price risk and the risk of increasing longevity.

The presence of equity risk means that the number of swaptions has to be adjusted in line with index movements. During a period of rising equity returns an insurer would have to keep purchasing these swaptions and this would become very expensive as the swaptions began to move into the money. In these circumstances the liability under the guarantee is open ended. The swaption solution does not deal with the equity risk.

9. Lessons

We have discussed the three major types of risks in the guaranteed annuity option and examined the pricing and the feasibility of hedging the risk under these contracts. In this section we will explore the extent to which the approaches discussed in his paper could or should have been applied. We also suggest that this episode has implications for the eduction and training of the actuarial profession particulary in connection with its exposure to ideas in modern financial economics.

It is worth emphasizing that when these guarantees were being written, the UK actuarial profession was still using deterministic methods to value liabilities. In particular valuation and premium calculations were based on a single deterministic interest rate. These methods were enshrined in the educational syllabus and rooted in current practice. Such methods are incapable of dealing adequately with options.

The relevant UK actuarial textbook used at the time, Fisher and Young (1965) in discussing guaranteed annuity options stated:

> "If, when the maturity date arrives, the guaranteed annuity rate is not as good as the office's own rates or a better purchase can be made elsewhere the option will not be exercised. The office cannot possibly gain from the transaction and should, therefore, at least in theory, guarantee only the lowest rate that seem likely in the foreseeable future.

However no guidance was provided as to what level this rate should be. Fisher and Young did suggest that conservative assumptions should be used and that allowance should be made for future improvements in mortality.

> "The option may not be exercised until a future date ranging perhaps from 5 to 50 years hence, and since it will be relatively easy to compare the yield under the option with the then current yields it is likely to be exercised against the office. The mortality and interest rate assumptions should be conservative."

The standard actuarial toolkit in use at the time was incapable of assessing the risks under this type of guarantee. However the guarantees were granted and they gave rise to a serious risk management problem that jeopardized the solvency of a number of UK companies. For many companies, the first time that the guaranteed annuity option for maturing contract became in-the-money was in October 1993. In December 1993, Equitable Life announced that it would cut the terminal bonuses in the case of policyholders who opted for the guarantee. This meant that the guaranteed annuity option policyholders who exercised their guarantee ended up paying for their own guarantee. The affected policyholders argued that Equitable's action made a mockery of their guarantee. The validity of this controversial approach became the subject of a protracted legal dispute. Eventually, in July 2000 the House of Lords settled the matter. It ruled against the Equitable and decreed that the practice of cutting the terminal bonuses to pay for the guarantee was disallowed. Equitable faced an immediate liability of 1.4 billion pounds to cover its current liability for the guaranteed annuity options and in December 2000 was forced to close its doors to new business. The oldest life insurance company in the world was felled by the guaranteed annuity option.

This entire episode should provide salutary lessons for the actuarial profession. It is now clear that the profession could have benefited from greater exposure to the paradigms of modern financial economics, to the difference between diversifiable and non-diversifiable risk, and to the application of stochastic simulation in asset-liability management would have enabled insurers to predict, monitor and manage the exposure under the guarantee.

10. Acknowledgments

The authors thank Myung Chon, Antoon Pelsser and two anonymous referees for many very helpful comments.

11. References

Ahn, D-H., Dittmar, R.F. and Gallant, A.R. (2002) Quadratic term structure models: theory and evidence. *Review of Financial Studies* **15**, 243-288.

Andersen, L. and Andreasen, J. (2001) "Factor dependence in bermudan swaptions: fact or fiction?" *Journal of Financial Economics* **62** 3-37.

Andersen, T.G., Benzoni, L. and Lund, J. (2002) An empirical investigation of continuous-time equity return models. J*ournal of Finance*, 1239-1284.

Bakshi, G., Cao, C. and Chen, Z. (1997) Empirical performance of alternative option pricing models. *Journal of Finance* **52**, 2003-2049.

Bakshi, G., Cao, C. and Chen, Z. (2000) Pricing and hedging long-term options. *Journal of Econometrics* **94**, 277-318.

Ballotta, L. and Haberman, S. (2002) Valuation of guaranteed annuity options. *Working Paper*, Department of Actuarial Science and Statistics, City University, London, UK.

Bansal, R. and Zhou, H. (2002) Term structure of interest rates with regime shifts. *Journal of Finance* **57**, 1997-2043.

Björk, T. (1998) *Arbitrage theory in continuous time*. Oxford University Press.

Bolton, M.J., Carr, D.H., Collis, P.A. et al (1997) Reserving for annuity guarantees. *Report of the Annuity Guarantees Working Party*, Institute of Actuaries, London, UK.

BOYLE, P.P. and EMANUEL, D. (1980) Discretely adjusted option hedges. *Journal of Financial Economics* **8**, 259-282.

BOYLE, P.P. and SCHWARTZ, E.S. (1977) Equilibrium prices of guarantees under equity-linked contracts, *Journal of Risk and Insurance* **44**, 4, 639-660.

BOYLE, P.P., COX, S., DUFRESNE, D., GERBER, H., MUELLER, H., PEDERSEN, H., PLISKA, S., SHERRIS, M., SHIU, E. and TAN, K.S. (1998) *Financial Economics*. The Actuarial Foundation, Chicago, USA.

CANNABERO, E. (1995) Where do one-factor interest rate models fail? *Journal of Fixed Income* **5**, 31-52.

CHERNOF, M., GALLANT, A.R., GHYSELS, E. and TAUCHEN, G. (2001) Alternative models for stock price dynamics. *Working Paper*, University of North Carolina.

COX, J.C., INGERSOLL, J.E. and ROSS, S.A. (1985) A theory of the term structure of interest rates. *Econometrica* **53**, 385-467.

DAI, Q. and SINGLETON, K. (2000) Specification analysis of affine term structure models. *Journal of Finance* **55**, 1943-1978.

DAI, Q. and SINGLETON, K. (2003) Term structure dynamics in theory and reality. *Review of Financial Studies*, forthcoming.

DRIESSEN, J., KLAASSEN, P. and MELENBERG, B. (2003) The performance of multi-factor term structure models for pricing and hedging caps and swaptions, *Journal of Financial and Quantitative Analysis*, forthcoming.

DUNBAR, N. (1999) Sterling Swaptions under New Scrutiny. *Risk*, December 33-35.

FAN, R., GUPTA, A. and RITCHKEN, P. (2001) On the performance of multi-factor term structure models for pricing caps and swaptions, *Working Paper Case Western University*, Weatherhead School of Management.

FISHER, H.F. and YOUNG, J. (1965) *Actuarial Practice of Life Assurance*, Cambridge University Press.

GUPTA, A. and SUBRAHMANYAM, M.G. (2001) An Examination of the Static and Dynamic Performance of Interest Rate Option Pricing Models in the Dollar Cap-Floor Markets. *Working Paper, Case Western Reserve University*, Weatherhead School of Management.

HARDY, M.R. (2003) *Investment Guarantees: Modeling and Risk Management for Equity-Linked Life Insurance*, Wiley.

HULL, J. (2002) *Options Futures and Other Derivatives*, Prentice Hall.

HULL, J. and WHITE A. (1990) "Pricing Interest Rate Derivative Securities", *Review of Financial Studies* **3(4)**, 573-592.

JAMSHIDIAN, F. (1989) "An Exact Bond Option Formula", *Journal of Finance* **44(1)**, 205-209.

JIANG, G.J. and OOMEN, R.C.A. (2002) Hedging Derivatives Risk, *Working Paper*, University of Arizona.

LITTERMAN, T. and SCHEINKMAN, J. (1991) Common factors affecting bond returns. *Journal of Fixed Income* **1**, 62-74.

LONGSTAFF, F., SANTA-CLARA, P. and SCHWARTZ, E. (2001) Throwing Away a Billion Dollars. *Journal of Financial Economics* **63**, 39-66.

Maturity Guarantees Working Party (MGWP) (1980) Report of the Maturity Guarantees Working Party. *Journal of the Institute of Actuaries* **107**, 102-212.

MELINO A. and S. Turnbull, M. (1995) Mis-specification and the Pricing and Hedging of long-term Foreign Currency Options. *Journal of International Money and Finance* **14.3**, 373-393.

MILEVSKY, M.A., and PROMISLOW, S.D. (2001) Mortality derivatives and the option to annuitize. *Insurance: Mathematics and Economics* **29(3)**, 299-316.

NOWMAN, K.B. (1997) Gaussian Estimation of Single-Factor Continuous Time Models of the Term Structure of Interest Rates. *Journal of Finance* **52**, 1695-1706.

PELSSER, A. (2003) Pricing and Hedging Guaranteed Annuity Options via Static Option Replication. *Insurance: Mathematics and Economics*, forthcoming.

VASICEK, O.A. (1977) "An Equilibrium Characterization of the Term Structure". *Journal of Financial Economics* **5**, 177-188.

WILKIE, A.D., WATERS, H.R. and YANG, S. (2003) Reserving, Pricing and Hedging for Policies with Guaranteed Annuity Options. Paper presented to the Faculty of Actuaries, Edinburgh, January 2003. *British Actuarial Journal*, forthcoming.

YANG, S. (2001) *Reserving, Pricing and Hedging for Guaranteed Annuity Options*. Phd Thesis, Department of Actuarial Mathematics and Statistics, Heriot Watt University, Edinburgh.

YU, J. and PHILLIPS, P. (2001) A Gaussian Approach for Continuous Time Models of the Short Term Interest Rates. *The Econometrics Journal* **4(2)**, 211-225.

PHELIM BOYLE
Centre for Advanced Studies in Finance
School of Accountancy
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1

MARY HARDY
Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1

A DISCRETE TIME BENCHMARK APPROACH FOR INSURANCE AND FINANCE

Hans Bühlmann[1] and Eckhard Platen[2]

Abstract

This paper proposes a consistent approach to discrete time valuation in insurance and finance. This approach uses the growth optimal portfolio as reference unit or benchmark. When used as benchmark, it is shown that all benchmarked price processes are supermartingales. Benchmarked fair price processes are characterized as martingales. No measure transformation is needed for the fair pricing of insurance policies and derivatives. The standard actuarial pricing rule is obtained as a particular case of fair pricing when the contingent claim is independent from the growth optimal portfolio.

1991 *Mathematics Subject Classification:* primary 90A12; secondary 60G30, 62P20.
JEL Classification: G10, G13

Keywords and phrases

Financial and insurance market model, benchmark approach, growth optimal portfolio, numeraire portfolio, fair pricing, actuarial pricing, unit linked insurance.

1. Introduction

There exists a stream of literature that exploits the concept of a *growth optimal portfolio* (GOP), originally developed by Kelly (1956) and later extended and discussed, for instance, in Long (1990), Artzner (1997), Bajeux-Besnainou & Portait (1997), Karatzas & Shreve (1998), Kramkov & Schachermayer (1999), Korn (2001) and Goll & Kallsen (2002). Under certain assumptions the GOP coincides with the *numeraire portfolio*, which makes prices, when expressed in units of this particular portfolio, into martingales under the given probability

[1] Department of Mathematics, Eidgenössische Technische Hochschule, 8092 Zürich, Switzerland.
[2] University of Technology Sydney, School of Finance & Economics and Department of Mathematical Sciences, PO Box 123, Broadway, NSW, 2007, Australia.

measure. In Kramkov & Schachermayer (1999) and Platen (2001, 2002) it was demonstrated that prices when benchmarked by the GOP can become supermartingales. The notion of a numeraire portfolio was recently extended by Becherer (2001), taking into account benchmarked prices that are supermartingales when an equivalent local martingale measure exists. In standard cases with an equivalent martingale measure the numeraire portfolio has been shown to coincide with the inverse of the deflator or state price density, see Constatinides (1992), Duffie (1996) or Rogers (1997). Furthermore, in Bühlmann (1992, 1995) and Bühlmann et al. (1998) the deflator has been suggested for the modeling of financial and insurance markets. Similarly, in Platen (2001, 2002, 2004) a financial market has been constructed by characterization of the GOP as benchmark portfolio.

Within this paper we follow a discrete time benchmark approach, where we characterize key features of a financial and insurance market via the GOP. We do not assume the existence of an equivalent martingale measure. The concept of fair pricing is introduced, where fair prices of insurance policies and derivatives are obtained via conditional expectations with respect to the real world probability measure. This provides a consistent basis for pricing that is widely applicable in insurance but also in derivative pricing. Examples of a discrete time market will be given that illustrate some key features of the benchmark approach.

2. Discrete Time Market

Let us consider a discrete time market that is modeled on a given probability space $(\Omega, \mathcal{A}, P)$. Asset prices are assumed to change their values only at the given discrete times

$$0 \leq t_0 < t_1 < \ldots < t_n < \infty$$

for fixed $n \in \{0, 1, \ldots\}$. The information available at time t in this market is described by $\mathcal{A}_{t_i}$. In this paper we consider $d+1$ primary securities, $d \in \{1, 2, \ldots\}$, which generate interest, dividend, coupon or other payments as income or loss, incurred from holding the respective asset. We denote by $S_i^{(j)}$ the nonnegative value at time t_i of a primary security account. This account holds only units of the jth security and all income is reinvested into this account. The 0th primary security account is the domestic savings account. According to the above description, the domestic savings account $S^{(0)}$ is then a roll-over short term bond account, where the interest payments are reinvested at each time step. If the jth primary security is a share, then $S_i^{(j)}$ is the value at time t_i of such shares including accrued dividends. Thus, the quantity $S_i^{(j)}$ represents the jth cum-dividend share price at time t_i. We assume that

$$S_i^{(j)} > 0 \tag{2.1}$$

almost surely for all $j \in \{0, 1, \ldots, d\}$.

Now, we introduce the *growth ratio* $h^{(j)}_{i+1}$ of the jth primary security account at time t_{i+1} in the form

$$h^{(j)}_{i+1} = \begin{cases} \dfrac{S^{(j)}_{i+1}}{S^{(j)}_i} & \text{for} \quad S^{(j)}_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

for $i \in \{0, 1, \ldots, n-1\}$ and $j \in \{0, 1, \ldots, d\}$. Note that the *return* of $S^{(j)}$ at time t_{i+1} equals $h^{(j)}_{i+1} - 1$. In our context the concept of a growth ratio will be more convenient than that of a return. We assume that $h^{(j)}_{i+1}$ is $\mathcal{A}_{t_{i+1}}$-measurable and almost surely finite. The growth rate of the domestic savings account $S^{(0)}$ shall be strictly positive, that is

$$h^{(0)}_{i+1} > 0 \tag{2.3}$$

almost surely for all $i \in \{0, 1, \ldots, n-1\}$ with $S^{(0)}_0 = 1$. We can express the price of the jth primary security account at time t_i, that is usually the jth cum-dividend share price $S^{(j)}_i$, in the form

$$S^{(j)}_i = S^{(j)}_0 \prod_{l=1}^{i} h^{(j)}_l \tag{2.4}$$

for $i \in \{0, 1, \ldots, n\}$ and $j \in \{0, 1, \ldots, d\}$. Note that due to assumptions (2.1) and (2.3) we have for the savings account

$$S^{(0)}_i > 0 \tag{2.5}$$

for all $i \in \{0, 1, \ldots, n\}$.

In the given discrete time market it is possible to form *self-financing portfolios* containing the above primary security accounts, where the changes in the value of the portfolio are only due to changes in primary security accounts. Since we will only consider self-financing portfolios we omit in the following the word "self-financing". For the characterization of a strictly positive portfolio at time t_i it is sufficient to describe the *proportion* $\pi^{(j)}_i \in (-\infty, \infty)$ of its value that at this time is invested in the jth primary security account, $j \in \{0, 1, \ldots, d\}$. Obviously, the proportions add to one, that is

$$\sum_{j=0}^{d} \pi^{(j)}_i = 1 \tag{2.6}$$

for all $i \in \{0, 1, \ldots, d\}$. The vector process $\pi = \left\{\pi_i = \left(\pi^{(0)}_i, \pi^{(1)}_i, \ldots, \pi^{(d)}_i\right),\ i \in \{0, 1, \ldots, n\}\right\}$ denotes the corresponding process of proportions. We assume that π_i is $\mathcal{A}_{t_i}$-measurable, which means that the proportions at a given time do not depend on any future events. The value of the corresponding portfolio at time t_i is denoted by $S_i^{(\pi)}$ and we write $S^{(\pi)} = \left\{S_i^{(\pi)}, i \in \{0, 1, \ldots, n\}\right\}$. Obviously, we obtain the *growth ratio* $h_l^{(\pi)}$ of this portfolio at time t_l in the form

$$h_l^{(\pi)} = \sum_{j=0}^{d} \pi_{l-1}^{(j)} h_l^{(j)} \tag{2.7}$$

for $l \in \{1, 2, \ldots, n\}$, where its value at time t_i is given by the expression

$$S_i^{(\pi)} = S_0^{(\pi)} \prod_{l=1}^{i} h_l^{(\pi)} \tag{2.8}$$

for $i \in \{0, 1, \ldots, n\}$.

3. Discrete Time Market of Finite Growth

Let us denote by $\mathcal{V}$ the set of all *strictly positive portfolio* processes $S^{(\pi)}$. This means, for a portfolio process $S^{(\pi)} \in \mathcal{V}$ it holds $h_{i+1}^{(\pi)} \in (0, \infty)$ almost surely for all $i \in \{0, 1, \ldots, n-1\}$. Due to (2.5) $\mathcal{V}$ is not empty. We define for a given portfolio process $S^{(\pi)} \in \mathcal{V}$ with corresponding process of proportions π its *growth rate* $g_i^{(\pi)}$ at time t_i by the conditional expectation

$$g_i^{(\pi)} = E\left(\log\left(h_{i+1}^{(\pi)}\right) \middle| \mathcal{A}_{t_i}\right) \tag{3.1}$$

for all $i \in \{0, 1, \ldots, n-1\}$. This allows us to introduce the *optimal growth rate* $\underline{g}_i$ at time t_i as the supremum

$$\underline{g}_i = \sup_{S^{(\pi)} \in \mathcal{V}} g_i^{(\pi)} \tag{3.2}$$

for all $i \in \{0, 1, \ldots, n-1\}$.

If the optimal growth rate could reach an infinite value, then the corresponding portfolio would have unlimited growth. We exclude such unrealistic behaviour by introducing the following natural condition.

Assumption 3.1 *We assume that the given discrete time market is of* finite growth*, that is*

$$\max_{i \in \{0, 1, \ldots, n-1\}} \underline{g}_i < \infty, \tag{3.3}$$

almost surely.

Furthermore, it is natural to assume that our discrete time market is such that a portfolio exists, which attains the optimal growth rate.

Assumption 3.2 *There exists a portfolio $S^{(\underline{\pi})} \in \mathcal{V}$ with corresponding process of proportions $\underline{\pi}$ and*

$$S_0^{(\underline{\pi})} = 1, \tag{3.4}$$

such that

$$g_i^{(\underline{\pi})} = \underline{g}_i \tag{3.5}$$

and

$$E\left(\frac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}}\middle| \mathcal{A}_{t_i}\right) < \infty \tag{3.6}$$

for all $i \in \{0, 1, \ldots, n-1\}$ and $S^{(\pi)} \in \mathcal{V}$. Such a portfolio is called a growth optimal portfolio *(GOP).*

Without conditions (3.3) and (3.5) there is no basis for considering GOPs. Also condition (3.6) is a very natural condition, which only assumes the integrability of ratios of growth rates and thus allows to form conditional expectations. There is an extremely wide range of models that satisfy the Assumptions 3.1 and 3.2. These cover most established discrete time models used in insurance and finance.

From the viewpoint of an investor, a *growth optimal portfolio* (GOP), can be interpreted as a best performing portfolio because there is no other strictly positive portfolio that in the long term can outperform its optimal growth rate. The GOP has also another remarkable property, which we derive in the following. Let us study the situation that an investor puts almost all of his wealth in a GOP $S^{(\underline{\pi})}$ and invests a vanishing small proportion $\theta \in \left(0, \frac{1}{2}\right)$ into an alternative portfolio $S^{(\pi)} \in \mathcal{V}$. We call the resulting portfolio the *interpolated portfolio* $V^{\theta,\pi,\underline{\pi}} \in \mathcal{V}$. It exhibits by (2.7) at time t_{i+1} the growth ratio

$$h_{i+1}^{\theta,\pi,\underline{\pi}} = \frac{V_{i+1}^{\theta,\pi,\underline{\pi}}}{V_i^{\theta,\pi,\underline{\pi}}} = \theta h_{i+1}^{(\pi)} + (1-\theta) h_{i+1}^{(\underline{\pi})} \tag{3.7}$$

with corresponding growth rate

$$g_i^{\theta,\pi,\underline{\pi}} = E\left(\log\left(h_{i+1}^{\theta,\pi,\underline{\pi}}\right)\middle| \mathcal{A}_{t_i}\right) \tag{3.8}$$

for $i \in \{0, 1, \ldots, n-1\}$. To study the rate of change in the growth rate of the interpolated portfolio let us define its *derivative* in the direction of the alternative portfolio $S^{(\pi)} \in \mathcal{V}$ at time t_i, that is the limit

$$\left.\frac{\partial g_i^{\theta,\pi,\underline{\pi}}}{\partial \theta}\right|_{\theta=0+} = \lim_{\theta \to 0+} \frac{1}{\theta}\left(g_i^{\theta,\pi,\underline{\pi}} - g_i^{(\underline{\pi})}\right) \tag{3.9}$$

for $i \in \{0, 1, \ldots, n-1\}$. We prove in Appendix A the following fundamental identity, which will give us access to the understanding of the central role of the GOP in pricing.

Theorem 3.3 *For a portfolio* $S^{(\pi)} \in \mathcal{V}$ *and* $i \in \{0, 1, \ldots n-1\}$ *the derivative of the growth rate of the interpolated portfolio at time* t_i *equals*

$$\left.\frac{\partial g_i^{\theta,\pi,\underline{\pi}}}{\partial \theta}\right|_{\theta=0+} = E\left(\frac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}}\,\middle|\,\mathcal{A}_{t_i}\right) - 1. \tag{3.10}$$

One observes that from (3.9), (3.5) and (3.2), we must have by the optimality property of the GOP that

$$\left.\partial g_i^{\theta,\pi,\underline{\pi}}\right|_{\theta=0+} \leq 0, \tag{3.11}$$

which leads by the identity (3.10) directly to the following important result.

Corollary 3.4 *A portfolio process* $S^{(\underline{\pi})} \in \mathcal{V}$ *is growth optimal if and only if all portfolios* $S^{(\pi)} \in \mathcal{V}$, *when expressed in units of* $S^{(\underline{\pi})}$, *are* $(\underline{\mathcal{A}}, P)$*-supermartingales, that is*

$$E\left(\frac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}}\,\middle|\,\mathcal{A}_{t_i}\right) \leq 1 \tag{3.12}$$

for all $i \in \{0, 1, \ldots, n-1\}$.

Corollary 3.4 reveals a fundamental property of the GOP. It says, all nonnegative securities, when expressed in units of the GOP are supermartingales. Note that we did not make any major assumptions on the given discrete time market. Under the additional assumption on the existence of an equivalent local martingale measure, a similar result has been obtained for semimartingale markets in Becherer (2001). Corollary 3.4 is proved without the explicit assumption on the existence of an equivalent risk neutral measure. The simple and direct proof of Theorem 3.3 in the Appendix avoids the technical machinery employed in Becherer (2001). In addition, our approach is constructive and the fundamental equation (3.10) can be used to establish further identities or inequalities in risk management.

Let us consider two nonnegative portfolios that are both growth optimal, see (3.5). According to Corollary 3.4 the first portfolio, when expressed in units of the second, must be a supermartingale. Additionally, by the same argument the second, expressed in units of the first, must be also a supermartingale. This can only be true if both processes are identical, which yields the following result.

Corollary 3.5 *The value process of the GOP is unique.*

Note that the stated uniqueness of the GOP does *not* imply that its proportions $\underline{\pi}$ have to be unique.

4. Fair Portfolios

In what follows we call prices, which are expressed in units of the GOP, *benchmarked* prices and their growth ratios *benchmarked* growth ratios. The condition (3.6) guarantees the integrability of benchmarked growth ratios and prices The benchmarked price $\hat{S}_i^{(\pi)}$ at time t_i of a portfolio $S^{(\pi)}$ is defined by the relation

$$\hat{S}_i^{(\pi)} = \frac{S_i^{(\pi)}}{S_i^{(\underline{\pi})}} \tag{4.1}$$

for all $i \in \{0, 1, \ldots, n\}$. By Corollary 3.4, the benchmarked price of a strictly positive portfolio $S^{(\pi)} \in \mathcal{V}$ is a supermartingale, which means by (3.12), (4.1), and (2.8) that

$$\hat{S}_i^{(\pi)} \geq E\left(\hat{S}_k^{(\pi)} \middle| \mathcal{A}_{t_i}\right), \tag{4.2}$$

for all $k \in \{0, 1, \ldots, n\}$ and $i \in \{0, 1, \ldots, k\}$.

In common actuarial and financial valuations in competitive, liquid markets a price is typically chosen such that seller and buyer have no systematic advantage or disadvantage. The problem of such a description is hidden in the fact that one must specify the reference unit or numeraire and the corresponding probability measure that both buyers and sellers use to calculate their expected payoff. If one chooses the real world measure as obvious probability measure, then one needs still to determine the reference unit. We know from Long (1990) that under certain conditions benchmarked prices are martingales. In markets with a corresponding equivalent risk neutral martingale measure this price corresponds to the risk neutral price. For this reason we choose in our more general setting the GOP as numeraire. By using the real world probability measure to form expectations and the GOP as numeraire it follows from Corollary 3.4, as shown in (4.2), that any strictly positive portfolio price, when expressed in units of the GOP, must be a supermartingale. This could give an advantage to the seller of the portfolio $S^{(\pi)}$ if the equality in (4.2) is a strict one. Its expected future benchmarked payoff is in such a case less than its present value. The only situation when buyers and sellers are equally treated is when the benchmarked price process $\hat{S}^{(\pi)}$ is an $(\underline{\mathcal{A}}, P)$-martingale, that means

$$\hat{S}_k^{(\pi)} = E\left(\hat{S}_k^{(\pi)} \middle| \mathcal{A}_{t_i}\right) \tag{4.3}$$

for all $k \in \{0, 1, \ldots, n\}$ and $i \in \{0, 1, \ldots, k\}$. Equation (4.3) means that the actual benchmarked price $\hat{S}_i^{(\pi)}$ is the best forecast of its future benchmarked values.

Equivalently to (4.3) we have by (2.8) for the corresponding portfolio process $S^{(\pi)}$ that

$$E\left(\frac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}}\middle|\mathcal{A}_{t_i}\right) = 1 \tag{4.4}$$

for all $i \in \{0, 1, \ldots, n-1\}$. This leads us naturally to the *concept of fair pricing*, see also Platen (2002):

Definition 4.1 *We call a value process* $V = \{V_k, k \in \{0, 1, \cdots, n\}\}$ fair *if its benchmarked value* $\hat{V}_k = \frac{V_k}{S_k^{(\underline{\pi})}}$ *forms an* $(\underline{\mathcal{A}}, P)$*-martingale.*

By Definition 4.1 and application of Theorem 3.3 we directly obtain the following interesting characterization of fair prices.

Corollary 4.2 *A given portfolio process* $S^{(\pi)}$ *is fair if and only if*

$$\left.\frac{\partial g_i^{\theta,\pi,\underline{\pi}}}{\partial \theta}\right|_{\theta=0+} = 0 \tag{4.5}$$

for all $i \in \{0, 1, \ldots, n-1\}$.

Intuitively, Corollary 4.2 expresses the fact that a portfolio is fair if the maximum that the growth rate of the corresponding interpolated portfolio attains, is a genuine maximum. This typically means that the GOP proportions must satisfy the usual first order conditions in the direction of the portfolio. This will happen if $S^{(\underline{\pi})}$ is in the interior of $\mathcal{V}$ as in this case the derivative at zero may be taken from both sides.

5. A Two Asset Example

To illustrate key features of the given discrete time benchmark approach, let us consider a simple example of a market with two primary security accounts. The two primary securities are the domestic currency, which is assumed to pay zero interest, and a stock that pays zero dividends. The savings account at time t_i is here simply the constant $S_i^{(0)} = 1$ for $i \in \{0, 1, \ldots, n\}$. The stock price $S_i^{(1)}$ at time t_i is given by the expression

$$S_i^{(1)} = S_0^{(1)} \prod_{l=1}^{i} h_l^{(1)}, \tag{5.1}$$

for $i \in \{0, 1, \ldots, n\}$. Here the growth ratio $h_l^{(1)} \in (0, \infty)$ at time t_l is assumed to be a random variable that can reach values, which are arbitrarily close

close to 0 and ∞. Since the GOP has always to be strictly positive we must have

$$\underline{\pi}_i^{(1)} \in [0,1] \tag{5.2}$$

for all $i \in \{0,1,\ldots,n\}$. By (2.6) the GOP proportions $\underline{\pi}_i^{(0)}$ and $\underline{\pi}_i^{(1)}$ are such that

$$\underline{\pi}_i^{(0)} = 1 - \underline{\pi}_i^{(1)} \tag{5.3}$$

for all $i \in \{0,1,\ldots,n\}$. Obviously, the set $\mathcal{V}$ of strictly positive, portfolios $S^{(\pi)}$ is then characterized by those portfolios with proportion $\pi_i^{(1)} \in [0,1]$ for all $i \in \{0,1,\ldots,n\}$. The growth rate $g_i^{(\pi)}$ at time t_i for a portfolio $S^{(\pi)} \in \mathcal{V}$ is according to (3.1) given by the expression

$$g_i^{(\pi)} = E\left(\log\left(1 + \pi_i^{(1)}\left(h_{i+1}^{(1)} - 1\right)\right)\middle| \mathcal{A}_{t_i}\right) \tag{5.4}$$

for all $i \in \{0,1,\ldots,n-1\}$. Let us now compute the optimal growth rate of this market, see (3.1). The first derivative of $g_i^{(\pi)}$ with respect to $\pi_i^{(1)}$ is

$$\frac{\partial g_i^{(\pi)}}{\partial \pi_i^{(1)}} = E\left(\frac{h_{i+1}^{(1)} - 1}{1 + \pi_i^{(1)}\left(h_{i+1}^{(1)} - 1\right)}\middle| \mathcal{A}_{t_i}\right) \tag{5.5}$$

and the second derivative has the form

$$\frac{\partial^2 g_i^{(\pi)}}{\partial \left(\pi_i^{(1)}\right)^2} = -E\left(\frac{\left(h_{i+1}^{(1)} - 1\right)^2}{\left(1 + \pi_i^{(1)}\left(h_{i+1}^{(1)} - 1\right)\right)^2}\middle| \mathcal{A}_{t_i}\right) \tag{5.6}$$

for $i \in \{0,1,\ldots,n-1\}$. We note that the second derivative is always negative, which indicates that the growth rate has at most one maximum. However, this maximum may refer to a proportion that does not belong to the interval $[0,1]$. To clarify such a situation we compute with (5.5) the values

$$\left.\frac{\partial g_i^{(\pi)}}{\partial \pi_i^{(1)}}\right|_{\pi_i^{(1)}=0} = E\left(h_{i+1}^{(1)}\middle| \mathcal{A}_{t_i}\right) - 1 \tag{5.7}$$

and

$$\left.\frac{\partial g_i^{(\pi)}}{\partial \pi_i^{(1)}}\right|_{\pi_i^{(1)}=1} = 1 - E\left(\frac{1}{h_{i+1}^{(1)}}\middle| \mathcal{A}_{t_i}\right) \tag{5.8}$$

for $i \in \{0,1,\ldots,n-1\}$. Due to (5.6) the first derivative $\frac{\partial g_i^{(\pi)}}{\partial \pi_i^{(1)}}$ is decreasing for $\pi_i^{(1)}$ increasing. If

$$E\left(\left(h_{i+1}^{(1)}\right)^{\lambda}\middle|\mathcal{A}_{t_i}\right) \geq 1 \tag{5.9}$$

for both $\lambda = 1$ and $\lambda = -1$, then (5.7) and (5.8) are of opposite sign and hence there exists some $\underline{\pi}_i^{(1)} \in [0, 1]$ such that

$$\left.\frac{\partial g_i^{(\pi)}}{\partial \pi_i^{(1)}}\right|_{\pi_i^{(1)} = \underline{\pi}_i^{(1)}} = 0 \tag{5.10}$$

for $i \in \{0, 1, \ldots, n-1\}$. Otherwise, if condition (5.9) is violated, then the optimal proportion is to be chosen at one of the boundary points. In this case the derivative (5.5) will not be zero at the optimal proportion and we obtain not a genuine maximum for the optimal proportion.

To check whether particular primary securities and portfolios are fair we now specify in our example the distribution of the growth ratios. Let us consider the case when the growth ratio $h_i^{(1)}$ is independent of the past and lognormally distributed such that

$$\log\left(h_i^{(1)}\right) \sim \mathcal{N}\left(\mu\Delta, \sigma^2\Delta\right) \tag{5.11}$$

with mean $\mu\Delta$, variance $\sigma^2\Delta > 0$ and time step size $\Delta = t_{i+1} - t_i > 0$.

1. At first, we clarify when the derivative $\frac{\partial g_i^{(\pi)}}{\partial \pi_i^{(1)}}$ can become zero for $\pi_i^{(1)} \in [0, 1]$. Because of

$$E\left(\left(h_{i+1}^{(1)}\right)^{\lambda}\middle|\mathcal{A}_{t_i}\right) = \exp\left\{\left(\lambda\mu + \frac{\sigma^2}{2}\right)\Delta\right\} \tag{5.12}$$

for $\lambda = 1$ and $\lambda = -1$, (5.10) can only hold for $|\mu| \leq \frac{\sigma^2}{2}$. In this case it is also possible to show for $\Delta \to 0$ that the optimal proportion $\underline{\pi}_i^{(1)}$ for the GOP reaches asymptotically the value

$$\lim_{\Delta \to 0} \underline{\pi}_i^{(1)} = \frac{1}{2} + \frac{\mu}{\sigma^2},$$

with limits

$$\lim_{\Delta \to 0} E\left(\frac{h_{i+1}^{(0)}}{h_{i+1}^{(\underline{\pi})}}\middle|\mathcal{A}_{t_i}\right) = 1$$

and

$$\lim_{\Delta \to 0} E\left(\frac{h_{i+1}^{(1)}}{h_{i+1}^{(\underline{\pi})}}\middle|\mathcal{A}_{t_i}\right) = 1$$

and it follows for all strictly positive portfolios $S^{(\pi)} \in \mathcal{V}$ that

$$\left. \frac{\partial g_i^{\theta,\pi,\underline{\pi}}}{\partial \theta} \right|_{\theta=0} = 0$$

for $i \in \{0, 1, \ldots, n-1\}$. Thus by Corollary 4.2 all portfolios $S^{(\pi)} \in \mathcal{V}$ are fair if the absolute mean to variance ratio is less than $\frac{1}{2}$, that is $\frac{|\mu|}{\sigma^2} \leq \frac{1}{2}$. This means, for all strictly positive benchmarked portfolios the expected log-return of $S^{(1)}$ is not allowed to be greater than half of its squared variance.

2. In the case $\frac{\mu}{\sigma^2} < -\frac{1}{2}$, when the stock significantly underperforms, then the situation is different. The optimal proportion is

$$\underline{\pi}_i^{(1)} = 0$$

for all $i \in \{0, 1, \ldots, n-1\}$. For the GOP this requires to hold all investments in the savings account. Here we get

$$E\left(\left. \frac{h_{i+1}^{(1)}}{h_{i+1}^{(\underline{\pi})}} \right| \mathcal{A}_{t_i} \right) = \exp\left(\left(\mu + \frac{\sigma^2}{2} \right) \Delta \right) < 1,$$

which shows that the benchmarked stock price process $\hat{S}^{(1)} = \frac{S^{(1)}}{S^{(\underline{\pi})}}$ is a strict supermartingale and not a martingale. Thus $S^{(1)}$ is not fair according to Definition 4.1. Alternatively, we can check by Corollary 4.2 whether $S^{(1)}$ is fair. For the portfolio π with all wealth invested in stock, that is $\pi_i = \left(\pi_i^{(0)}, \pi_i^{(1)}\right) = (0, 1)$, we obtain the derivative of the corresponding interpolated portfolio in the form

$$\left. \frac{\partial g_i^{\theta,\pi,\underline{\pi}}}{\partial \theta} \right|_{\theta=0+} = \exp\left(\left(\mu + \frac{\sigma^2}{2} \right) \Delta \right) - 1 < 0,$$

which shows by Corollary 4.1 that $S^{(1)}$ is not fair. On the other hand $\hat{S}^{(0)}$ is clearly a martingale and thus fair.

3. For $\frac{\mu}{\sigma^2} > \frac{1}{2}$ the stock is performing extremely well. The optimal propotion is

$$\underline{\pi}_i^{(1)} = 1$$

for $i \in \{0, 1, \ldots, n-1\}$. This means, for sufficiently large mean of the logarithm of the growth ratio of the stock one has to hold for the GOP all investments in the stock. In this case we get

$$E\left(\frac{h_{i+1}^{(0)}}{h_{i+1}^{(\underline{\pi})}}\bigg|\mathcal{A}_{t_i}\right) = \exp\left(\left(-\mu + \frac{\sigma^2}{2}\right)\Delta\right) < 1,$$

which says that the benchmarked domestic savings account $\hat{S}^{(0)} = \frac{S^{(0)}}{S^{(\underline{\pi})}}$ is a strict supermartingale. This means that $\hat{S}^{(0)}$ is not a martingale and thus by Definition 4.1 not fair. However, note that $\hat{S}^{(1)} = 1$ is a martingale. For $\pi_i = \left(\pi_i^{(0)}, \pi_i^{(1)}\right) = (1,0)$ we have then

$$\left.\frac{\partial g_i^{\theta,\pi,\underline{\pi}}}{\partial\theta}\right|_{\theta=0+} = \exp\left(\left(-\mu + \frac{\sigma^2}{2}\right)\Delta\right) - 1 < 0.$$

This confirms also by Corollary 4.2 that $S^{(0)}$ is not fair.

This example demonstrates that benchmarked prices are not always martingales. However, these benchmarked prices become martingales if the corresponding derivative of the growth rate of the interpolated portfolio in the direction of the security is zero, as follows from Corollary 4.2. Furthermore, the given log-normal example indicates that discrete time markets with securities, where the mean to variance ratio of the excess log-return over the risk free rate exceeds one half, may not be fair.

6. Fair Pricing of Contingent Claims

Now, let us consider a *contingent claim* H_i, which is an $\mathcal{A}_{t_i}$-measurable, possibly negative payoff, expressed in units of the domestic currency and has to be paid at a maturity date t_i, $i \in \{1, 2, \ldots, n\}$. Note that the claim H_i is not only contingent on the information provided by the observed primary security accounts $S_l^{(j)}$ up until time t_i, $j \in \{0, 1, \ldots, d\}$, $l \in \{0, 1, \ldots, i\}$, but as well on additional information contained in $\mathcal{A}_{t_i}$ as, for instance, the occurrence of defaults or insured events. Following our previous discussion and Definition 4.1 we obtain directly the following formula for the fair price of a contingent claim.

Corollary 6.1 *The* fair price $U_k^{(H_i)}$ *at time t_k for the contingent claim H_i satisfies the* fair pricing formula

$$U_k^{(H_i)} = S_k^{(\underline{\pi})} E\left(\frac{H_i}{S_i^{(\underline{\pi})}}\bigg|\mathcal{A}_{t_k}\right), \tag{6.1}$$

for $k \in \{0, 1, \ldots, i\}$.

Obviously, by (4.10) all fair contingent claim prices have a corresponding *benchmarked fair price* of the type

$$\hat{U}_k^{(H_i)} = \frac{U_k^{(H_i)}}{S_k^{(\underline{\pi})}} \tag{6.2}$$

for all $k \in \{0, 1, \ldots, i\}$, $i \in \{0, 1, \ldots, n\}$, where the process $\hat{U}^{(H_i)} = \{U_k^{(H_i)},\ k \in \{0, 1, \ldots, i\}\}$ forms an $(\underline{\mathcal{A}}, P)$-martingale according to Definition 4.1. The argument can be easily extended to sums of contingent claims with $\underline{\mathcal{A}}$-adapted maturity dates. Note that all fair portfolios and fair contingent claim prices form a price system, where benchmarked prices are $(\underline{\mathcal{A}}, P)$-martingales.

If there exists only one equivalent risk neutral martingale measure, then the pricing formula (6.1) is the standard *risk neutral pricing formula*, used in finance, see Platen (2001, 2002, 2004). However note, in this paper we do *not* assume the existence of such an equivalent risk neutral martingale measure and consider a more general framework.

Formally, one can extend (6.1) also for assessing the accumulated value for cashflows that occurred in the past, that is for $k \in \{i+1, i+2, \ldots\}$. Then we obtain

$$U_k^{(H_i)} = \frac{H_i}{S_i^{(\underline{\pi})}} S_k^{(\underline{\pi})} \tag{6.3}$$

for $i \in \{0, 1, \cdots\}$ and $k \in \{i+1, i+2, \ldots\}$. In (6.3) we express the with earnings accumulated t_k-value of the payment H_i made at time t_i. This interpretation is important for insurance accounting as will be discussed below.

An important case arises when a contingent claim H_i with maturity t_i is *independent* of the value $S_i^{(\underline{\pi})}$ of the GOP. Then by using (6.1) its fair price at time t_k is obtained by the formula

$$U_k^{(H_i)} = E\left(H_i \middle| \mathcal{A}_{t_k}\right) P_k^i, \tag{6.4}$$

where

$$P_k^i = E\left(\frac{S_k^{(\underline{\pi})}}{S_i^{(\underline{\pi})}} \middle| \mathcal{A}_{t_k} \right) \tag{6.5}$$

is the fair value at time t_k of the *zero coupon bond* with maturity t_i for $k \in \{0, 1, \ldots, i\}$, $i \in \{0, 1, \ldots, n\}$. The formula (6.4) reflects the classical *actuarial pricing formula* that has been applied by actuaries for centuries to project future cashflows into present values, though with an "artificial" not financial market oriented understanding of P_k^i. Thus it turns out that the actuarial pricing approach is in this particular case generalized by the fair pricing concept that we introduced above through Definition 4.1. Note, in this case the knowledge of the particular dynamics of the GOP is *not* necessary since the zero coupon bond P_k^i carries the relevant information needed from the GOP.

7. Fair Pricing of Sequences of Cashflows

For the pricing of an insurance policy the actuarial task is the valuation of a *sequence of cashflows* $X_0, X_1, \dots, X_n$, which are paid at the times $t_0, t_1, \dots, t_n$, respectively. After each payment, its value is invested by the insurance company in a strictly positive portfolio, characterized by a process of proportions π. Here we choose an arbitrary process of proportions π, representing the investment portfolio of the insurance company. The benchmarked fair price $\hat{Q}_0$ at time t_0 for the above sequence of cashflows is according to (6.2) given by the expression

$$\hat{Q}_0 = E\left(\sum_{k=0}^{n} \frac{X_k}{S_k^{(\pi)}} \,\middle|\, \mathcal{A}_{t_0} \right). \tag{7.1}$$

It follows that the benchmarked fair value $\hat{Q}_i$ at time t_i for $i \in \{0, 1, \dots, n-1\}$ of this sequence of cashflows equals the sum

$$\hat{Q}_i = \hat{C}_i + \hat{R}_i \tag{7.2}$$

for $i \in \{0, 1, \dots, n\}$. Here we obtain

$$\hat{C}_i = \frac{1}{S_i^{(\pi)}} \sum_{k=0}^{i} X_k \prod_{l=k}^{i-1} h_{l+1}^{(\pi)}, \tag{7.3}$$

which expresses the benchmarked value of the already *accumulated payments*. Furthermore,

$$\hat{R}_i = E\left(\sum_{k=1+1}^{n} \frac{X_k}{S_k^{(\pi)}} \,\middle|\, \mathcal{A}_{t_i} \right) \tag{7.4}$$

is the benchmarked fair price at time t_i for the remaining payments, which is called the *prospective reserve*, see Bühlmann 95. It is easy to check that the process $\hat{Q} = \{\hat{Q}_i, i \in \{0, 1, \dots, n\}\}$ forms an $(\underline{\mathcal{A}}, P)$-martingale for all choices of π by the insurance company. When expressed in units of the domestic currency, we have at time t_i for the above sequence of cashflows the fair value

$$Q_i = S_i^{(\pi)} \hat{Q}_i \tag{7.5}$$

for all $i \in \{0, 1, \dots, n\}$.

The above result is important, for instance, for the fair pricing of life insurance policies. Each insurance carrier can choose its own process of proportions π to invest the payments that arise. However, the GOP, which is needed to value the prospective reserve, must be the same for all insurance companies in the same market. Above we clarified the role of the GOP for pricing the prospective reserve. We point out that the above analysis says nothing about the performance and riskiness of different investment strategies that the insurance carrier can choose. The growth rate for the investment portfolio becomes optimal, if the proportions of the GOP are used. If the insurance company aims to maximize the growth rate of its investments, then the fair pricing of an insurance policy and the optimization of the investment portfolio both involve the GOP.

8. Unit Linked Insurance Contracts

In the insurance context we look again at the cashflows $X_0, X_1, \ldots, X_n$ but assume a specific form for these random variables. Intuitively, they stand now for unit linked claims and premiums. Hence they can be of either sign. The cashflow at time t_i is of the form

$$X_i = D_i S_i^{(\pi)} \tag{8.1}$$

for $i \in \{1, 2, \ldots, n\}$. The payments are linked to some strictly positive reference portfolio $S^{(\pi)} \in \mathcal{V}$ with given proportions π. The insurance contract specifies the reference portfolio $S^{(\pi)}$ and the random variables D_i, which are contingent on the occurrence of insured events during the period $(t_{i-1}, t_i]$, for instance, death, disablement or accidents.

The standard actuarial technique treats such contracts by using the reference portfolio process $S^{(\pi)}$ as numeraire and then deals with the unit linked random variables $D_0, D_1, \ldots, D_n$ at interest rate zero. It is reasonable to assume that these random variables are $\underline{\mathcal{A}}$-adapted and independent of the reference portfolio process $S^{(\pi)}$.

The standard actuarial value $W_i^{(\pi)}$ of the payment stream at time t_i is determined by the accumulated payments $C_i^{(\pi)}$ and the properly defined liability or prospective reserve r_i. The standard actuarial methodology assumes that the insurer invests all accumulated payments in the reference portfolio $S^{(\pi)}$. Then one obtains for $W_i^{(\pi)}$, when expressed in units of the domestic currency, the expression

$$W_i^{(\pi)} = C_i^{(\pi)} + r_i \tag{8.2}$$

with *accumulated payments*

$$C_i^{(\pi)} = S_i^{(\pi)} \sum_{k=1}^{i} D_k \tag{8.3}$$

and the liability or *actuarial prospective reserve*

$$r_i = S_i^{(\pi)} E\left(\sum_{k=i+1}^{n} D_k \,|\, \mathcal{A}_{t_i} \right) \tag{8.4}$$

for $i \in \{0, 1, \ldots, n\}$. Observe the difference between $W_i^{(\pi)}$ and Q_i as defined in (7.5). Hence the standard actuarial pricing and fair pricing will, in general, lead to different results. As we have seen previously in (6.4) this is to be expected when the cashflows are not independent of the GOP.

The benchmarked value $\hat{W}_i^{(\pi)} = \frac{W_i^{(\pi)}}{S_i^{(\pi)}}$ at time t_i for the cashflows of this unit linked insurance contract is then by (8.2) of the form

$$\hat{W}_i^{(\pi)} = \frac{C_i^{(\pi)} + r_i}{S_i^{(\pi)}} \tag{8.5}$$

for $i \in \{0, 1, \ldots, n\}$. On the other hand, the benchmarked fair value $\hat{Q}_i^{(\pi)}$ at time t_i of the cashflows of this contract is according to (7.1) - (7.5) given by the expression

$$\hat{Q}_i^{(\pi)} = \frac{C_i^{(\pi)} + R_i}{S_i^{(\pi)}} \tag{8.6}$$

with *fair prospective reserve*

$$R_i = S_i^{(\pi)} E\left(\sum_{k=i+1}^{n} \frac{D_k S_k^{(\pi)}}{S_k^{(\pi)}} \middle| \mathcal{A}_{t_i} \right) \tag{8.7}$$

for $i \in \{0, 1, \ldots, n\}$. Under the natural condition of nonnegative fair prospective reserves one can prove that the benchmarked fair prospective reserve is less or equal the actuarial prospective reserve. The proof of the following inequality relies on the supermartingale property of $\frac{S_k^{(\pi)}}{S_k^{(\pi)}}$, $k \in \{0, 1, \ldots, n\}$ and is shown in Appendix B.

Lemma 8.1 *If*

$$E\left(\sum_{k=m+1}^{n} D_k \middle| \mathcal{A}_{t_m} \right) \geq 0 \tag{8.8}$$

for all $m \in \{0, 1, \ldots, n-1\}$*, then*

$$R_i \leq r_i \tag{8.9}$$

for all $i \in \{0, 1, \ldots, m-1\}$*.*

As by (8.4) we have

$$r_m = S_m^{(\pi)} E\left(\sum_{k=m+1}^{n} D_k \middle| \mathcal{A}_{t_m} \right),$$

the condition (8.8) of the lemma means that the insurance contract defines a cashflow whose actuarial prospective reserve never becomes negative. This is usually observed as a practical constraint, since insurance products that allow for negative reserves have many defects. From (8.5) and (8.6) we immediately have under condition (8.8) the inequality

$$\hat{Q}_i^{(\pi)} \leq \hat{W}_i^{(\pi)}$$

for $i \in \{0, 1, \ldots, n\}$. Reverting to property (8.9) we observe that there is, in general, a nonnegative difference

$$r_i - R_i \geq 0 \tag{8.10}$$

between the actuarial and the fair prospective reserve. This difference is a consequence of the classical actuarial price calculation leading to the prospective reserve r_i in (8.4). Of course, the actuarial and the fair prospective reserve coincide if one uses the GOP as reference portfolio.

Conclusion

We have shown that the growth optimal portfolio plays a central role for pricing in finance and insurance markets. The concept of fair contingent claim pricing has been introduced. Fair price processes, when measured in units of the growth optimal portfolio, form martingales. For contingent claims that are independent of the growth optimal portfolio fair prices also coincide with the classical actuarial prices, however, in general, this is not the case.

Acknowledgement

The second author likes to express his gratitude to ETH Zurich for its kind hospitality during several visits. Both authors thank Freddy Delbaen, Walter Schachermayer, Martin Schweizer and Albert Shiryaev for valuable suggestions for improvements of this work. The paper has gone through different versions. The referees have been very helpful to identify its weak spots.

A. Appendix

Proof of Theorem 3.3

For $\theta \in \left(0, \frac{1}{2}\right)$ and $S^{(\pi)} \in \mathcal{V}$ we consider the interpolated portfolio $V^{\theta, \pi, \underline{\pi}} \in \mathcal{V}$, that is with growth ratio

$$h_{i+1}^{\theta, \pi, \underline{\pi}} > 0 \tag{A.1}$$

given in (3.7) for $i \in \{0, 1, \ldots, n-1\}$. One can then show, using $\log(x) \leq x - 1$ and (3.7), that

$$G_{i+1}^{\theta, \pi, \underline{\pi}} = \frac{1}{\theta} \log\left(\frac{h_{i+1}^{\theta, \pi, \underline{\pi}}}{h_{i+1}^{(\underline{\pi})}}\right) \leq \frac{1}{\theta}\left(\frac{h_{i+1}^{\theta, \pi, \underline{\pi}}}{h_{i+1}^{(\underline{\pi})}} - 1\right) = \frac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}} - 1 \tag{A.2}$$

and

$$G_{i+1}^{\theta, \pi, \underline{\pi}} \geq -\frac{1}{\theta}\left(\frac{h_{i+1}^{(\underline{\pi})}}{h_{i+1}^{\theta, \pi, \underline{\pi}}} - 1\right) = \frac{h_{i+1}^{(\pi)} - h_{i+1}^{(\underline{\pi})}}{h_{i+1}^{\theta, \pi, \underline{\pi}}}. \tag{A.3}$$

We obtain in (A.3) for $h_{i+1}^{(\pi)} - h_{i+1}^{(\underline{\pi})} \geq 0$ because of $h_{i+1}^{\theta, \pi, \underline{\pi}} > 0$ the inequality

$$G_{i+1}^{\theta, \pi, \underline{\pi}} \geq 0 \tag{A.4}$$

and for $h_{i+1}^{(\pi)} - h_{i+1}^{(\underline{\pi})} < 0$ from (A.3) because of $\theta \in \left(0, \frac{1}{2}\right)$, $h_{i+1}^{(\pi)} \geq 0$, and $h_{i+1}^{\theta, \pi, \underline{\pi}} > 0$ that

$$G_{i+1}^{\theta, \pi, \underline{\pi}} \geq -\frac{h_{i+1}^{(\underline{\pi})}}{h_{i+1}^{\theta, \pi, \underline{\pi}}} = -\frac{1}{1 - \theta + \theta \dfrac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}}} \geq -\frac{1}{1-\theta} \geq -2. \tag{A.5}$$

Summarizing (A.2)-(A.5) we have for $i \in \{0, 1, \ldots, n-1\}$ and $S^{(\pi)} \in \mathcal{V}$ the upper and lower bounds

$$-2 \leq G_{i+1}^{\theta, \pi, \underline{\pi}} \leq \frac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}} - 1, \tag{A.6}$$

where by (3.6)

$$E\left(\frac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}}\right) < \infty. \tag{A.7}$$

Then by using (A.6) and (A.7) it follows by the Dominated Convergence Theorem that

$$\begin{aligned}\left.\frac{\partial g_i^{\theta,\pi,\underline{\pi}}}{\partial\theta}\right|_{\theta=0} &= \lim_{\theta\to 0+} E\left(G_{i+1}^{\theta,\pi,\underline{\pi}} \,\middle|\, \mathcal{A}_{t_i}\right)\\ &= E\left(\lim_{\theta\to 0+} G_{i+1}^{\theta,\pi,\underline{\pi}} \,\middle|\, \mathcal{A}_{t_i}\right)\\ &= E\left(\left.\frac{\partial}{\partial\theta}\log\left(\frac{h_{i+1}^{\theta,\pi,\underline{\pi}}}{h_{i+1}^{(\underline{\pi})}}\right)\right|_{\theta=0} \,\middle|\, \mathcal{A}_{t_i}\right)\\ &= E\left(\frac{h_{i+1}^{(\pi)}}{h_{i+1}^{(\underline{\pi})}} \,\middle|\, \mathcal{A}_{t_i}\right)-1 \end{aligned}\tag{A.8}$$

for $i\in\{0,1,\ldots,n-1\}$ and $S^{(\pi)}\in\mathcal{V}$. This proves equation (3.10). □

B. Appendix

Proof of Lemma 8.1

We have

$$E\left(\sum_{k=i+1}^{n}\frac{D_k\,S_k^{(\pi)}}{S_k^{(\underline{\pi})}}\,\middle|\,\mathcal{A}_{t_{n-1}}\right)\le E\left(\sum_{k=i+1}^{n}\frac{D_k\,S_{k\wedge(n-1)}^{(\pi)}}{S_{k\wedge(n-1)}^{(\underline{\pi})}}\,\middle|\,\mathcal{A}_{t_{n-1}}\right)$$

if $E(D_n|\mathcal{A}_{t_{n-1}})\geq 0$,

$$E\left(\sum_{k=i+1}^{n}\frac{D_k\,S_{k\wedge(n-1)}^{(\pi)}}{S_{k\wedge(n-1)}^{(\underline{\pi})}}\,\middle|\,\mathcal{A}_{t_{n-2}}\right)\le E\left(\sum_{k=i+1}^{n}\frac{D_k\,S_{k\wedge(n-2)}^{(\pi)}}{S_{k\wedge(n-2)}^{(\underline{\pi})}}\,\middle|\,\mathcal{A}_{t_{n-2}}\right)$$

if $E(D_n+D_{n-1}|\mathcal{A}_{t_{n-2}})\geq 0$

$$\vdots$$

$$\begin{aligned}E\left(\sum_{k=i+1}^{n}\frac{D_k\,S_{k\wedge(i+1)}^{(\pi)}}{S_{k\wedge(i+1)}^{(\underline{\pi})}}\,\middle|\,\mathcal{A}_{t_i}\right)&\le E\left(\sum_{k=i+1}^{n}\frac{D_k\,S_{k\wedge i}^{(\pi)}}{S_{k\wedge i}^{(\underline{\pi})}}\,\middle|\,\mathcal{A}_{t_i}\right)\\ &=\frac{r_i}{S_i^{(\underline{\pi})}}\end{aligned}$$

if $E(D_n + D_{n-1} + \ldots + D_{i+1} \mid \mathcal{A}_{t_i}) \geq 0$, for $i \in \{0,1,\ldots,n-1\}$. Taking conditional expectation with respect to $\mathcal{A}_{t_i}$, the inequalities above become a chain, whose first member equals $\frac{R_i}{S_i^{(\pi)}}$, and the last member becomes $\frac{r_i}{S_i^{(\pi)}}$.

This proves (8.9). □

References

ARTZNER, P. (1997) On the numeraire portfolio. In *Mathematics of Derivative Securities*, pp. 53-58. Cambridge University Press.

BAJEUX-BESNAINOU, I. and PORTAIT, R. (1997) The numeraire portfolio: A new perspective on financial theory. *The European Journal of Finance* **3**, 291-309.

BECHERER, D. (2001) The numeraire portfolio for unbounded semimartingales. *Finance Stoch.* **5**, 327-341.

BÜHLMANN, H. (1992) Stochastic discounting. *Insurance: Mathematics and Economics* **11**, 113-127.

BÜHLMANN, H. (1995) Life insurance with stochastic interest rates. In G. Ottaviani (Ed.), *Financial Risk and Insurance*, pp. 1-24. Springer.

BÜHLMANN, H., DELBAEN, F., EMBRECHTS, P. and SHIRYAEV, A. (1998) On Esscher transforms in discrete finance models. *ASTIN Bulletin* **28(2)**, 171-186.

CONSTATINIDES, G.M. (1992) A theory of the nominal structure of interest rates. *Rev. Financial Studies* **5**, 531-552.

DUFFIE, D. (1996) *Dynamic Asset Pricing Theory* (2nd ed.). Princeton, University Press.

GOLL, T. and KALLSEN, J. (2003) A complete explicit solution to the log-optimal portfolio problem. *Adv. in Appl. Probab.* **13(2)**, 774-799.

HEATH, D. and PLATEN, E. (2002) Pricing and hedging of index derivatives under an alternative asset price model with endogenous stochastic volatility. In J. Yong (Ed.), *Recent Developments in Mathematical Finance*, pp. 117-126. World Scientific.

KARATZAS, I. and SHREVE, S.E. (1998) *Methods of Mathematical Finance*, Volume 39 of *Appl. Math.* Springer.

KELLY, J.R. (1956) A new interpretation of information rate. *Bell Syst. Techn. J.* **35**, 917-926.

KORN, R. (2001) Value preserving strategies and a general framework for local approaches to optimal portfolios. *Math. Finance* **10(2)**, 227-241.

KRAMKOV, D.O. and SCHACHERMAYER, W. (1999) The asymptotic elasticity of utility functions and optimal investment in incomplete markets. *Ann. Appl. Probab.* **9**, 904-950.

LONG, J.B. (1990) The numeraire portfolio. *J. Financial Economics* **26**, 29-69.

PLATEN, E. (2001) A minimal financial market model. In *Trends in Mathematics*, pp. 293-301. Birkhäuser.

PLATEN, E. (2002) Arbitrage in continuous complete markets. *Adv. in Appl. Probab.* **34(3)**, 540-558.

PLATEN, E. (2004). A class of complete benchmark models with intensity based jumps to appear in J. Appl. Probab. **4(1)**.

ROGERS, L.C.G. (1997) The potential approach to the term structure of interest rates and their exchange rates. *Math. Finance* **7**, 157-176.

A UNIFIED APPROACH TO GENERATE RISK MEASURES

BY

MARC J. GOOVAERTS[1,2], ROB KAAS[2], JAN DHAENE[1,2], QIHE TANG[2]

ABSTRACT

The paper derives many existing risk measures and premium principles by minimizing a Markov bound for the tail probability. Our approach involves two exogenous functions $v(S)$ and $\phi(S,\pi)$ and another exogenous parameter $\alpha \leq 1$. Minimizing a general Markov bound leads to the following unifying equation:

$$E[\phi(S,\pi)] = \alpha E[v(S)].$$

For any random variable, the risk measure π is the solution to the unifying equation. By varying the functions ϕ and v, the paper derives the mean value principle, the zero-utility premium principle, the Swiss premium principle, Tail VaR, Yaari's dual theory of risk, mixture of Esscher principles and more. The paper also discusses combining two risks with super-additive properties and sub-additive properties. In addition, we recall some of the important characterization theorems of these risk measures.

KEYWORDS

Insurance premium principle, Risk measure, Markov inequality

1. INTRODUCTION

In the economic and actuarial financial literature the concept of insurance premium principles (risk measures) has been studied from different angles. An insurance premium principle is a mapping from the set of risks to the reals, cf. e.g. Gerber (1979). The reason to study insurance premium principles is the well-known fact in the actuarial field that if the premium income equals the expectation of the claim size or less, ruin is certain. In order to keep the ruin probability restricted one considers a risk characteristic or a risk measure for calculating premiums that includes a safety loading. This concept is essential for the economics of actuarial evaluations. Several types of insurance premium principles have been studied and characterized by means of axioms as in

[1] CRIS, Catholic University of Leuven
[2] Department of Quantitative Economics, University of Amsterdam.

Goovaerts *et al.* (1984). On the other hand, desirable properties for premiums relevant from an economic point of view have been considered. An insurance premium principle is often considered as the "price" of a risk (or of a tail risk in reinsurance), as the value of a stochastic reserve, or as an indication of the maximal probable loss. This gives the relation to ordering of risks that is recently developed in the actuarial literature. In Artzner (1999), see also Artzner *et al.* (1999), a risk measure is also defined as a mapping from the set of r.v.'s to the reals. It could be argued that a risk measure is a broader concept than an insurance premium calculation principle. Indeed, for a risk X, the probability $\rho(X) = \Pr[X > 1.10\mathrm{E}[X]]$ is a risk measure, but this is not a premium calculation principle because tacitly it is assumed that premiums are expressed in monetary units. However, assuming homogeneity for a risk measure, hence $\rho(aX) = a\rho(X)$ for all real $a > 0$ and all risks X, implies that $\rho(X)$ allows changing the monetary units. On the other hand, because the parameters appearing in the insurance premium principles may depend on monetary units, the class of insurance premiums contains the risk measures that are homogeneous as a special case. In addition, let X be a risk variable with finite expectation and let u be an initial surplus. Defining a transformed random variable describing risk as

$$Y = X + \frac{\alpha}{u}(X - \mathrm{E}[X])^2$$

also allows risk measures to depend on other monetary quantities. Consequently it is difficult to give a distinction between insurance premium principles and homogeneous risk measures. Different sets of axioms lead to different risk measures. The choice of the relevant axioms of course depends on the economics of the situations for which it is used. Desirable properties might be different for actual calculation of premiums, for reinsurance premiums, or for allocation, and so on.

In this paper we present a unified approach to some important classes of premium principles as well as risk measures, based on the Markov inequality for tail probabilities. We prove that most well-known insurance premium principles can be derived in this way. In addition, we will refer to some of the important characterization theorems of these risk measures.

Basic material on utility theory and insurance goes back to Borch (1968, 1974), using the utility concept of von Neumann and Morgenstern (1944). The foundation of premium principles was laid by Bühlmann (1970) who introduced the zero-utility premium, Gerber (1979) and comprehensively by Goovaerts *et al.* (1984). The utility concept, the mean-value premium principle as well as the expected value principle can be deduced from certain axioms. An early source is Hardy *et al.* (1952). The Swiss premium calculation principle was introduced by Gerber (1974) and De Vijlder and Goovaerts (1979). A multiplicative equivalent of the utility framework has led to the Orlicz principle as introduced by Haezendonck and Goovaerts (1982). A characterization for additive premiums has been introduced by Gerber and Goovaerts (1981), and led to the so-called mixture of Esscher premium principles. More recently, Wang (1996) introduced in the actuarial literature the distortion functions into the framework of risk measures, using Yaari's (1987) dual theory of choice under risk. This approach can also be introduced in an axiomatic way. Artzner (1999) restricted the class of Orlicz premium principles by adding the requirement of translation invariance to its axioms, weakened by Jarrow (2002). This has mathematical consequences that are sometimes contrary to practical insurance applications. In the 1980's the practical significance of the basic axioms has been discussed; see Goovaerts *et al.* (1984). On the same grounds Artzner (1999) provided an argumentation for selecting a set of desirable axioms. In Goovaerts *et al.* (2003) it is argued that there are no sets of axioms generally valid for all types of risky situations. There is a difference in desirable properties when one considers a risk measure for allocation of capital, a risk measure for regulating purposes or a risk measure for premiums. There is a parallel with mathematical statistics, where characteristics of distributions may have quite different meanings and uses, like e.g. the mean to measure central tendency, the variance to measure spread, the skewness to reflect asymmetry and the peakedness to measure the thickness of the tails. In an actuarial context, risk measures might have different properties than in other economic contexts. For instance, if we cannot assume that there are two different reinsurers willing to cover both halves of a risk separately, the risk measure (premium) for the entire risk should be larger than twice the original risk measure.

This paper aims to introduce many different risk measures (premium principles) now available, each with their desirable properties, within a unified framework based on the Markov inequality. To give an idea how this is achieved, we give a simple illustration.

Example 1.1. *The exponential premium is derived as the solution to the utility equilibrium equation*

$$\mathrm{E}\left[-e^{-\beta(w-X)}\right] = \mathrm{E}\left[-e^{-\beta(w-\pi)}\right], \tag{1.1}$$

where w is the initial capital and $u(x) = -e^{-\beta x}$ is the utility attached to wealth level x. This is equivalent to

$$\mathrm{E}\left[e^{-\beta(\pi-X)}\right] = 1, \tag{1.2}$$

hence we get the explicit solution

$$\pi = \frac{1}{\beta}\log\mathrm{E}\left[e^{\beta X}\right]. \tag{1.3}$$

Taking $Y = e^{\beta X}$ and $y = e^{\beta\pi}$ and applying $\Pr[Y > y] \le \frac{1}{y}\mathrm{E}[Y]$ (Markov inequality), we get the following inequality for the survival probabilities with X:

$$\Pr[X > \pi] \le \frac{1}{e^{\beta\pi}}\mathrm{E}\left[e^{\beta X}\right]. \tag{1.4}$$

For this Markov bound to be non-trivial, the r.h.s. of (1.4) must be at most 1. *It equals* 1 *when π is equal to the exponential(β) premium with X. This procedure leads to an equation which gives the premium for X from a Markov bound.*

Two more things must be noted. First, for fixed π*, we write the bound in (1.4) as* $f(\beta) = \mathrm{E}[e^{\beta(X-\pi)}]$*. Since* $f''(\beta) = \mathrm{E}[(X-\pi)^2 e^{\beta(X-\pi)}] > 0$ *the function* $f(\beta)$ *is convex in some relevant* β*-region. The risk aversion* β_0 *for which this bound* $f(\beta)$ *is minimal has* $f'(\beta_0) = 0$*, hence* $\pi = \mathrm{E}[Xe^{\beta_0 X}] / \mathrm{E}[e^{\beta_0 X}]$*, which is the Esscher premium for X with parameter* β_0*. This way, also the Esscher premium has been linked to a Markov bound. The r.h.s. of (1.4) equals* 1 *for* $\beta = 0$ *as well, and is less than or equal to* 1 *for* β *in the interval* $[0,\beta_1]$*, where* β_1 *is the risk aversion for which the exponential premium equals* π*. Second, if* $\pi = \frac{1}{\beta}\log \mathrm{E}[e^{\beta X}]$ *holds, we have the following exponential upper bound for tail probabilities: for any* $k > 0$*,*

$$\Pr[X > \pi + k] \leq \frac{1}{e^{\beta(\pi+k)}} \mathrm{E}\left[e^{\beta X}\right] = e^{-\beta k}. \tag{1.5}$$

Using variations of the Markov bound above, the various equations that generate various premium principles (or risk measures) can be derived. Section 2 presents a method to do this, Section 3 applies this method to many such principles, and discusses their axiomatic foundations as well as some other properties; Section 4 concludes.

2. Generating Markovian risk measures

Throughout this paper, we denote the cumulative distribution function (cdf) of a random variable S by F_S. For any non-negative and non-decreasing function $v(s)$ satisfying

$$\mathrm{E}[v(S)] < +\infty, \tag{2.1}$$

we define an associated r.v. S^* having a cdf with differential

$$\mathrm{d}F_{S^*}(s) = \frac{v(s)\mathrm{d}F_S(s)}{\mathrm{E}[v(S)]}, \quad -\infty < s < +\infty. \tag{2.2}$$

It is easy to prove that

$$\Pr[S > \pi] \leq \Pr\left[S^* > \pi\right], \quad -\infty < \pi < +\infty. \tag{2.3}$$

For any Lebesgue measurable bivariate function $\phi(\cdot,\cdot)$ satisfying

$$\phi(s,\pi) \geq \mathrm{I}_{(s>\pi)}, \tag{2.4}$$

we have the following inequalities:

$$\Pr\left[S^* > \pi\right] = \mathrm{E}\left[\mathrm{I}_{(S^*>\pi)}\right] \leq \mathrm{E}\left[\phi\left(S^*,\pi\right)\right]. \tag{2.5}$$

Then it follows from (2.3) that

$$\Pr[S > \pi] \leq \frac{\mathrm{E}[\phi(S,\pi)v(S)]}{\mathrm{E}[v(S)]}. \tag{GMI}$$

This is a generalized version of the Markov inequality, which has $S \geq 0$ with probability 1 and $\pi \geq 0$, $\phi(s,\pi) = s/\pi$ and $v(s) \equiv 1$. Therefore, we denote it by the acronym [GMI]. Similar discussions can be found in Runnenburg and Goovaerts (1985), where the functions $v(\cdot)$ and $\phi(\cdot,\cdot)$ are specified as $v(\cdot) \equiv 1$ and $\phi(s,\pi) = f(s)/f(\pi)$, respectively, for some non-negative and non-decreasing function $f(\cdot)$.

For the inequality [GMI] to make sense, the bivariate function $\phi(s,\pi)$ given in (2.4) and the r.v. S should satisfy

$$\mathrm{E}\left[\phi(S,\pi)\,v(S)\right] < +\infty \tag{2.6}$$

for all relevant π. Note that if (2.6) holds for some $-\infty < \pi < +\infty$ then (2.1) does as well. By assuming (2.6), it is clear that the family of r.v.'s S considered in the inequality [GMI] is restricted, in the sense that the right tail of S can not be arbitrarily heavy. For the given functions $\phi(\cdot,\cdot)$ and $v(\cdot)$ as above, we introduce below a family of all admissible r.v.'s that satisfy (2.6):

$$\mathbb{S}_{\phi,v} = \left\{S : \mathrm{E}\left[\phi(S,\pi)v(S)\right] < +\infty \quad \text{for all large } \pi\right\}. \tag{2.7}$$

Sometimes we are interested in the case that there exists a minimal value $\pi_M^{(\alpha)}$ such that [GMI] gives a bound

$$\Pr\left[S > \pi_M\right] \leq \frac{\mathrm{E}\left[\phi(S,\pi_M)\,v(S)\right]}{\mathrm{E}\left[v(S)\right]} \leq \alpha \leq 1. \tag{2.8}$$

Note that (2.8) produces an upper bound for the α-quantile $q_\alpha(S)$ of S.
For each $0 \leq \alpha \leq 1$, the restriction (2.4) on $\phi(\cdot,\cdot)$ allows us further to introduce a subfamily of $\mathbb{S}_{\phi,v}$ as follows:

$$\mathbb{S}_{\phi,v,\alpha} = \left\{S : \frac{\mathrm{E}\left[\phi(S,\pi)\,v(S)\right]}{\mathrm{E}\left[v(S)\right]} \leq \alpha \quad \text{for all large } \pi\right\}. \tag{2.9}$$

If in (2.4) the function $\phi(s,\pi)$ is strictly smaller than 1 for at least one point (s,π), then it is not difficult to prove that there are some values of $0 \leq \alpha < 1$ such that the subfamilies $\mathbb{S}_{\phi,v,\alpha}$ are not empty. We also note that $\mathbb{S}_{\phi,v,\alpha}$ increases in $\alpha \geq 0$.

Hereafter, for a real function $f(\cdot)$ defined on an interval D and a constant b in the range of the function $f(\cdot)$, we write an equation $f(\pi) = b$ with the understanding that its root is the minimal value of π satisfying the inequalities $f(\pi) \leq b$ and $\max\{f(x)\,|\,x \in (\pi - \varepsilon,\, \pi + \varepsilon) \cap D\} \geq b$ for any $\varepsilon > 0$. With this convention, the minimal value $\pi_M^{(\alpha)}$ such that the second inequality in (2.8) holds is simply the solution of the equation

$$\frac{\mathrm{E}\left[\phi(S,\pi_M)\,v(S)\right]}{\mathrm{E}\left[v(S)\right]} = \alpha. \tag{UE$_\alpha$}$$

When $\alpha = 1$, we call

$$\frac{\mathrm{E}\left[\phi\left(S, \pi_M\right) v(S)\right]}{\mathrm{E}\left[v(S)\right]} = 1 \qquad \text{[UE]}$$

the unifying equation, or [UE] in acronym. This equation will act as the unifying form to generate many well-known risk measures. The equation [UE] gives the minimal percentile for which the upper bound for the tail probability of S still makes sense. It will turn out that these minimal percentiles correspond to several well-known premium principles (risk measures). It is clear that the solution of the equation [UE] is not smaller than the minimal value of the r.v. S.

Definition 2.1. *Let S be an admissible r.v. from the family* $\mathbb{S}_{\phi,v,\alpha}$ *for some* $0 \leq \alpha \leq 1$, *where* $\phi = \phi(\cdot,\cdot)$ *and* $v = v(\cdot)$ *are two given measurable functions with* ϕ *satisfying (2.4) and v non-negative and non-decreasing. The solution* $\pi_M^{(\alpha)}$ *of the equation [UE*$_\alpha$*] is called a* Markovian risk measure *of the r.v. S at level* α.

Remark 2.1. *About the actuarial meaning of the ingredients* $\phi(\cdot,\cdot)$, $v(\cdot)$ *and* α *in Definition 2.1 we remark that* α *represents a confidence bound, which in practical situations is determined by the regulator or the management of an insurance company. In principle the actuarial risk measures considered are intended to be approximations (on the safe side) for the VaR of order* α, *and to have some desirable actuarial properties such as additivity, subadditivity, or superadditivity, according to actuarial applications for calculating solvency margins, for RBC calculations, as well as for the top-down approach of premiums calculations. The functions* $\phi(\cdot,\cdot)$ *and* $v(\cdot)$ *are introduced to derive bounds for the VaR, so that these bounds have some desirable properties for applications. In addition, because a risk measure provides an upper bound for the VaR, it might be interesting to determine the minimal value of the risk measure attached by the different choices of* $\phi(\cdot,\cdot)$ *and* $v(\cdot)$.

Remark 2.2. *Clearly, given the ingredients* $\phi(\cdot,\cdot)$, $v(\cdot)$ *and* α, *the Markovian risk measure* $\pi_M^{(\alpha)}(S)$ *involves only the distribution of the admissible r.v. S. A Markovian risk measure provides an upper bound for the VaR at the same level. By selecting appropriate functions* ϕ *the Markovian risk measures can reflect desirable properties when adding r.v.'s in addition to their dependence structure.*

Remark 2.3. *Let* X_1 *and* X_2 *be two admissible r.v.'s, with Markovian risk measures* $\pi_M^{(\alpha)}(X_1)$ *and* $\pi_M^{(\alpha)}(X_2)$. *Then we have*

$$\Pr\left[X_1 > \pi_M^{(\alpha)}(X_1)\right] \leq \alpha, \qquad \Pr\left[X_2 > \pi_M^{(\alpha)}(X_2)\right] \leq \alpha. \tag{2.10}$$

We can obtain from the equation [UE$_\alpha$*] that*

$$\Pr\left[X_1 + X_2 > \pi_M^{(\alpha)}(X_1) + \pi_M^{(\alpha)}(X_2)\right] \leq \alpha \tag{2.11}$$

- *in case X_1 and X_2 are independent when the risk measure $\pi_M^{(\alpha)}$ involved is subadditive for sums of independent risks;*
- *in case X_1 and X_2 are comonotonic when the risk measure $\pi_M^{(\alpha)}$ involved is subadditive for sums of comonotonic risks;*
- *for any X_1 and X_2, regardless of their dependence structure, in case a subadditive risk measure $\pi_M^{(\alpha)}$ is applied.*

3. Some Markovian risk measures

In what follows we will provide a list of important insurance premium principles (or risk measures) and show how they can be derived from the equation [UE]. We will also list a set of basic underlying axioms. In practice, for different situations different sets of axioms are needed.

3.1. The mean value principle

The mean value principle has been characterized by Hardy *et al.* (1952); see also Goovaerts *et al.* (1984), Chapter 2.8, in the framework of insurance premiums.

Definition 3.1. *Let S be a risk variable. For a given non-decreasing and non-negative function $f(\cdot)$ such that* $\mathrm{E}[f(S)]$ *converges, the* mean value risk measure $\pi = \pi_f$ *is the root of the equation* $f(\pi) = \mathrm{E}[f(S)]$.

Clearly, we can obtain the mean value risk measure by choosing in the equation [UE] the functions $\phi(s,\pi) = f(s)/f(\pi)$ and $v(\cdot) \equiv 1$. As verified in Goovaerts *et al.* (1984), p. 57-61, this principle can be characterized by the following axioms (necessary and sufficient conditions):

A1.1. $\pi(c) = c$ for any degenerate risk c;

A1.2. $\Pr[X \leq Y] = 1 \Longrightarrow \pi(X) \leq \pi(Y)$;

A1.3. If $\pi(X) = \pi(X')$, Y is a r.v. and I is a Bernoulli variable independent of the vector $\{X, X', Y\}$, then $\pi(IX + (1-I)Y) = \pi(IX' + (1-I)Y)$.

Remark 3.1. *This last axiom can be expressed in terms of distribution functions by assuming that mixing F_Y with F_X or with $F_{X'}$ leads to the same risk measure, as long as the mixing weights are the same.*

Remark 3.2. *Under the condition that* $\mathrm{E}[f(S)]$ *converges one obtains as an upper bound for the survival probability*

$$\Pr[S > \pi + u] \leq \frac{\mathrm{E}[f(S)]}{f(\pi + u)} = \frac{f(\pi)}{f(\pi + u)}. \tag{3.1}$$

Specifically, when $\mathrm{E}[e^{\lambda S}] < \infty$ *for some* $\lambda > 0$ *one obtains* $e^{-\lambda u}$ *as an upper bound for the probability* $\Pr[S > \pi + u]$*, see the example in Section 1. In case π_α is the root of* $\mathrm{E}[f(S)] = \alpha f(\pi)$*, by the inequality [GMI] one gets* $\Pr[S \geq \pi_\alpha] \leq \alpha$.

3.2. The zero-utility premium principle

The zero-utility premium principle was introduced by Bühlmann (1970).

Definition 3.2. *Let* $u(\cdot)$ *be a non-decreasing utility function. The* zero-utility premium $\pi(S)$ *is the solution of* $u(0) = \mathrm{E}[u(\pi - S)]$.

We assume that either the risk or the utility function is bounded from above. Because $u(\cdot)$ and $u(\cdot) + c$ define the same ordering in expected utility, the utility is determined such that $u(x) \to 0$ as $x \to +\infty$. To obtain the zero-utility premium principle, one chooses in the equation [UE] the functions $\phi(s,\pi) = u(\pi - s)/u(0)$ and $v(\cdot) \equiv 1$.

In order to relate the utility to the VaR one should proceed as follows. By the inequality [GMI], we get

$$\Pr[S > \pi_\alpha] \le \mathrm{E}\left[\frac{-u(\pi_\alpha - S)}{-u(0)}\right] = \alpha, \tag{3.2}$$

where π_α is the solution of the equation $\mathrm{E}[u(\pi - S)] = \alpha u(0)$. The result obtained here requires that the utility function $u(\cdot)$ is bounded from below. However, this restriction can be weakened by considering limits for translated utility functions.

Let the symbol $\preceq_{eu}$ represent the weak order with respect to the zero-utility premium principle, that is, $X \preceq_{eu} Y$ means that X is preferable to Y. We write $X \sim_{eu} Y$ if both $X \preceq_{eu} Y$ and $Y \preceq_{eu} X$. It is well-known that the preferences of a decision maker between risks can be described by means of comparing expected utility as a measure of the risk if they fulfill the following five axioms which are due to von Neumann and Morgenstern (1944) (combining Denuit *et al.* (1999) and Wang and Young (1998)):

A2.1. If $F_X = F_Y$ then $X \sim_{eu} Y$;

A2.2. The order $\preceq_{eu}$ is reflexive, transitive and complete;

A2.3. If $X_n \preceq_{eu} Y$ and $F_{X_n} \to F_X$ then $X \preceq_{eu} Y$;

A2.4. If $F_X \ge F_Y$ then $X \preceq_{eu} Y$;

A2.5. If $X \preceq_{eu} Y$ and if the distribution functions of X'_p and Y'_p are given by $F_{X'_p}(x) = pF_X(x) + (1-p)F_Z(x)$ and $F_{Y'_p}(x) = pF_Y(x) + (1-p)F_Z(x)$ where F_Z is an arbitrary distribution function, then $X'_p \preceq_{eu} Y'_p$ for any $p \in [0,1]$.

From these axioms, the existence of a utility function $u(\cdot)$ can be proven, with the property that $X \preceq_{eu} Y$ if and only if $\mathrm{E}[u(-X)] \ge \mathrm{E}[u(-Y)]$.

3.3. The Swiss premium calculation principle

The Swiss premium principle was introduced by Gerber (1974) to put the mean value principle and the zero-utility principle in a unified framework.

Definition 3.3. *Let $w(\cdot)$ be a non-negative and non-decreasing function on $\mathbb{R}$ and $0 \leq z \leq 1$ be a parameter. Then the* Swiss premium principle $\pi = \pi(S)$ *is the root of the equation*

$$\mathrm{E}\left[w(S - z\pi)\right] = w\left((1 - z)\pi\right). \tag{3.3}$$

This equation is the special case of [UE] with $\phi(s,\pi) = w(s - z\pi) / w((1 - z)\pi)$ and $v(\cdot) \equiv 1$. It is clear that $z = 0$ provides us with the mean value premium, while $z = 1$ gives the zero-utility premium. Recall that by the inequality [GMI], the root π_α of the equation $\mathrm{E}[w(S - z\pi)] = \alpha w((1 - z)\pi)$ determines an upper bound for the VaR_α.

Remark 3.3. *Because one still may choose $w(\cdot)$, it can be arranged to have supplementary properties for the risk measures. Indeed if we assume that $w(\cdot)$ is convex, we have*

$$X \leq_{cx} Y \Longrightarrow \pi(X) \leq \pi(Y). \tag{3.4}$$

See for instance Dhaene et al. (2002a, b) for the definition of $\leq_{cx}$ (convex order). For two random pairs (S_1, S_2) and $(\widetilde{S}_1, \widetilde{S}_2)$ with the same marginal distributions, we call (S_1, S_2) more related than $(\widetilde{S}_1, \widetilde{S}_2)$ if the probability $\Pr[S_1 \leq x, S_2 \leq y]$ *that S_1 and S_2 are both small is larger than that for $\widetilde{S}_1$ and $\widetilde{S}_2$, for all x and y; see e.g. Kaas et al. (2001), Chapter 10.6. In this case one gets from* (3.4)

$$\pi(\widetilde{S}_1 + \widetilde{S}_2) \leq \pi(S_1 + S_2). \tag{3.5}$$

The risk measure of the sum of a pair of r.v.'s with the same marginal distributions depends on the dependence structure, and in this case increases with the degree of dependence between the terms of the sum.

Remark 3.4. *Gerber (1974) proves the following characterization: Let $w(\cdot)$ be strictly increasing and continuous, then the Swiss premium calculation principle generated by $w(\cdot)$ is additive for independent risks if and only if $w(\cdot)$ is exponential or linear.*

3.4. The Orlicz premium principle

The Orlicz principle was introduced by Haezendonck and Goovaerts (1982) as a multiplicative equivalent of the zero-utility principle. To introduce this premium principle, they used the concept of a Young function ψ, which is a mapping from $\mathbb{R}_0^+$ into $\mathbb{R}_0^+$ that can be written as an integral of the form

$$\psi(x) = \int_0^x f(t)\,\mathrm{d}t, \qquad x \geq 0, \tag{3.6}$$

where f is a left-continuous, non-decreasing on $\mathbb{R}_0^+$ satisfying $f(0) = 0$ and $\lim_{x \to +\infty} f(x) = +\infty$. It is seen that a Young function ψ is absolutely continuous, convex and strictly increasing, and has $\psi'(0) = 0$. We say that ψ is normalized if $\psi(1) = 1$.

Definition 3.4. *Let ψ be a normalized Young function. The root of the equation*

$$E[\psi(S/\pi)] = 1 \tag{3.7}$$

is called the Orlicz premium principle *of the risk S.*

The unified approach follows from the equation [UE] with $\phi(s,\pi)$ replaced by $\psi(s/\pi)$ and $v(s) \equiv 1$. The Orlicz premium satisfies the following properties:

A4.1. $\Pr[X \le Y] = 1 \Longrightarrow \pi(X) \le \pi(Y)$;

A4.2. $\pi(X) = 1$ when $X \equiv 1$;

A4.3. $\pi(aX) = a\pi(X)$ for any $a > 0$ and any risk X;

A4.4. $\pi(X + Y) \le \pi(X) + \pi(Y)$.

Remark 3.5. *A4.3 above says that the Orlicz premium principle is positively homogenous. In the literature, positive homogeneity is often confused with currency independence. As an example, we look at the standard deviation principle* $\pi_1(X) = E[X] + \alpha \cdot \sigma[X]$ *and the variance principle* $\pi_2(X) = E[X] + \beta \cdot \mathrm{Var}[X]$, *where α and β are two positive constants, α is dimension-free but the dimension of* $1/\beta$ *is money. Clearly $\pi_1(X)$ is positive homogenous but $\pi_2(X)$ is not. But it stands to reason that when applying a premium principle, if the currency is changed, so should all constants having dimension money. So going from BFr to Euro, where* 1 *Euro* $\approx$ 40 *BFr, the value of β in $\pi_2(X)$ should be adjusted by the same factor. In this way both $\pi_1(X)$ and $\pi_2(X)$ are independent of the monetary unit.*

Remark 3.6. *These properties remain exactly the same for risks that may also be negative, such as those used in the definition of coherent risk measures by Artzner (1999). Indeed if $\pi(-1) = -1$ and one extends these properties to r.v.'s supported on the whole line $\mathbb{R}$, then*

$$\pi(X + a - a) \le \pi(X + a) - a. \tag{3.8}$$

Hence $\pi(X + a) \ge \pi(X) + a$ and consequently $\pi(X + a) = \pi(X) + a$.

The interested reader is referred to Haezendonck and Goovaerts (1982). If in addition translation invariance is imposed for non-negative risks, it turns out that the only coherent risk measure for non-negative risks within the class of Orlicz principles is an expectation $\pi(X) = E[X]$.

Remark 3.7. *The Orlicz principle can also be generalized to cope with VaR_α. Actually, from the inequality [GMI], the solution π_α of the equation* $E[\psi(S/\pi)] = \alpha$ *gives* $\Pr[S > \pi_\alpha] \le \alpha$.

3.5. More general risk measures derived from Markov bounds

For this section, we confine to risks with the same mean. We consider more general risk measures derived from Markov bounds, applied to sums of pairs of r.v.'s, which may or may not be independent. The generalization consists in the fact that we consider the dependence structure to some extent in the risk premium, letting the premium for the sum $X+Y$ depend both on the distribution of the sum $X+Y$ and on the distribution of the sum X^c+Y^c of the comonotonic (maximally dependent) copies of the r.v.'s X and Y. Because of this, we denote the premium for the sum $X+Y$ by $\pi(X, Y)$ rather than by $\pi(X+Y)$. When the r.v.'s X and Y are comonotonic, however, there is no difference in understanding between the two symbols $\pi(X, Y)$ and $\pi(X+Y)$.

Taking $\pi(X)$ simply equal to $\pi(X,0)$, we consider the following properties:

A5.1. $\pi(aX)=a\pi(X)$ for any $a>0$;

A5.2. $\pi(X+b)=\pi(X)+b$ for any $b\in\mathbb{R}$;

A5.3$_a$. $\pi(X,Y)\le\pi(X)+\pi(Y)$;

A5.3$_b$. $\pi(X,Y)\ge\pi(X)+\pi(Y)$;

A5.3$_c$. $\pi(X,Y)=\pi(X)+\pi(Y)$.

Remark 3.8. *A5.3$_a$ describes the subadditivity property, which is realistic only in case diversification of risks is possible. However, this is rarely the case in insurance. Subadditivity gives rise to easy mathematics because distance functions can be used. The superadditivity property for a risk measure (that is not Artzner coherent) is redundant in the following practical situation of capital allocation or solvency assessment. Suppose that two companies with risks X_1 and X_2 merge and form a company with risk X_1+X_2. Let d_1, d_2 and d denote the allocated capitals or solvency margins. Then, with probability* 1,

$$(X_1+X_2-d_1-d_2)_+\le(X_1-d_1)_++(X_2-d_2)_+. \tag{3.9}$$

This inequality expresses the fact that, with probability 1, *the residual risk of the merged company is smaller than the risk of the split company. In case $d\ge d_1+d_2$, one gets, also with probability* 1,

$$(X_1+X_2-d)_+\le(X_1-d_1)_++(X_2-d_2)_+. \tag{3.10}$$

Hence in case one calculates the capitals d_1, d_2 and d by means of a risk measure it should be superadditive (or additive) to describe the economics in the right way. Subadditivity is only based on the idea that it is easier to convince the shareholders of a conglomerate in failure to provide additional capital than the shareholders of some of the subsidiaries. Recent cases indicate that for companies in a financial distress situation splitting is the only way out.

Remark 3.9. *It should also be noted that subadditivity cannot be used as an argument for a merger of companies to be efficient. The preservation* (3.9) *of the inequality of risks with probability one expresses this fact; indeed*

$$\pi\big((X_1+X_2-d_1-d_2)_+\big)\le\pi\big((X_1-d_1)_++(X_2-d_2)_+\big) \tag{3.11}$$

expresses the efficiency of a merger. It has nothing to do with the subadditivity. A capital $d<d_1+d_2$, for instance derived by a subadditive risk measure, can only

be considered if the dependence structure allows it. For instance if d is determined as the minimal root of the equation

$$\pi\big((X_1 + X_2 - d)_+\big) = \pi\big((X_1 - d_1)_+ + (X_2 - d_2)_+\big), \tag{3.12}$$

d obviously depends on the dependence between X_1 and X_2. Note that in $(X_1 - d_1)_+ + (X_2 - d_2)_+$, the two terms are dependent. Taking this dependence into account, the risk measure providing the capitals d, d_1 and d_2 will not always be subadditive, nor always superadditive, but may instead exhibit behavior similar to the VaR, see Embrechts et al. (2002).

Let $\psi(\cdot)$ be a non-decreasing, non-negative, and convex function on $\mathbb{R}$ satisfying $\lim_{x\to+\infty} \psi(x) = +\infty$. For fixed $0 < p < 1$, we get, by choosing $v(\cdot) = 1$, the equality

$$\phi(X, Y, \pi) = \frac{1}{\psi(1)} \cdot \psi\left(\frac{\big(X + Y - F^{-1}_{X^c+Y^c}(p)\big)_+}{\pi - F^{-1}_{X^c+Y^c}(p)}\right) \tag{3.13}$$

and by solving [UE] for π, the following risk measure for the sum of two r.v.'s:

$$\mathrm{E}\left[\psi\left(\frac{\big(X + Y - F^{-1}_{X^c+Y^c}(p)\big)_+}{\pi(X,Y) - F^{-1}_{X^c+Y^c}(p)}\right)\right] = \psi(1) \tag{3.14}$$

for some parameter $0 < p < 1$. Hereafter, the pth quantile of a r.v. X with d.f. F_X is, as usual, defined by

$$F_X^{-1}(p) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq p\}, \quad p \in [0,1]. \tag{3.15}$$

It is easily seen that there exists a unique constant $a(p) > 0$ such that

$$\mathrm{E}\left[\psi\left(\frac{\big(X + Y - F^{-1}_{X^c+Y^c}(p)\big)_+}{a(p)}\right)\right] = \psi(1). \tag{3.16}$$

Thus $\pi(X,Y) = F^{-1}_{X^c+Y^c}(p) + a(p)$. Especially, letting Y be degenerate at 0 we get $\pi(X) > F_X^{-1}(p)$.

Now we check that A5.1, A5.2 and A5.3$_a$ are satisfied by π (subadditive case). In fact, the proofs for the first two axioms are trivial. As for A5.3$_a$, we derive

$$\mathrm{E}\left[\psi\left(\frac{\big(X + Y - F^{-1}_{X^c+Y^c}(p)\big)_+}{\pi(X) + \pi(Y) - F^{-1}_{X^c+Y^c}(p)}\right)\right]$$

$$\begin{aligned}
&\le \mathrm{E}\left[\psi\left(\frac{\left(X-F_X^{-1}(p)\right)_+ + \left(Y-F_Y^{-1}(p)\right)_+}{\pi(X)+\pi(Y)-F_X^{-1}(p)-F_Y^{-1}(p)}\right)\right]\\
&= \mathrm{E}\left[\psi\left(\frac{\pi(X)-F_X^{-1}(p)}{\pi(X)+\pi(Y)-F_X^{-1}(p)-F_Y^{-1}(p)}\cdot\frac{X-F_X^{-1}(p)}{\pi(X)-F_X^{-1}(p)}\right.\right.\\
&\qquad\left.\left.+\frac{\pi(Y)-F_Y^{-1}(p)}{\pi(X)+\pi(Y)-F_X^{-1}(p)-F_Y^{-1}(p)}\cdot\frac{Y-F_Y^{-1}(p)}{\pi(Y)-F_Y^{-1}(p)}\right)\right]\\
&\le \mathrm{E}\left[\frac{\pi(X)-F_X^{-1}(p)}{\pi(X)+\pi(Y)-F_X^{-1}(p)-F_Y^{-1}(p)}\cdot\psi\left(\frac{X-F_X^{-1}(p)}{\pi(X)-F_X^{-1}(p)}\right)\right]\\
&\qquad+\mathrm{E}\left[\frac{\pi(Y)-F_Y^{-1}(p)}{\pi(X)+\pi(Y)-F_X^{-1}(p)-F_Y^{-1}(p)}\cdot\psi\left(\frac{Y-F_Y^{-1}(p)}{\pi(Y)-F_Y^{-1}(p)}\right)\right]\\
&= \psi(1)\\
&= \mathrm{E}\left[\psi\left(\frac{\left(X+Y-F_{X^c+Y^c}^{-1}(p)\right)}{\pi(X,Y)-F_{X^c+Y^c}^{-1}(p)}\right)\right].
\end{aligned} \tag{3.17}$$

This proves A5.3$_a$.

Remark 3.10. *If the function $\psi(\cdot)$ above is restricted to satisfy $\psi(1)=\psi'(1)$, then it can be proven that the risk measure*

$$\pi_l(X) = F_X^{-1}(p) + \mathrm{E}\left[\left(X - F_X^{-1}(p)\right)_+\right] \tag{3.18}$$

gives the lowest generalized Orlicz measure. In fact, since ψ is convex on $\mathbb{R}$ and satisfies $\psi(1)=\psi'(1)$, we have

$$\psi\left((x)_+\right) \ge \psi(1)\cdot(x)_+ \quad \text{for any } x\in\mathbb{R}. \tag{3.19}$$

Let $\pi(X)$ be a generalized Orlicz risk measure of the risk X, that is, $\pi(X)$ is the solution of the equation

$$\mathrm{E}\left[\psi\left(\frac{\left(X-F_X^{-1}(p)\right)_+}{\pi(X)-F_X^{-1}(p)}\right)\right] = \psi(1). \tag{3.20}$$

By (3.19) and recalling that $\pi(X) > F_X^{-1}(p)$, *we have*

$$\psi(1) = \mathrm{E}\left[\psi\left(\frac{\left(X - F_X^{-1}(p)\right)_+}{\pi(X) - F_X^{-1}(p)}\right)\right] \geq \psi(1) \cdot \mathrm{E}\left[\frac{\left(X - F_X^{-1}(p)\right)_+}{\pi(X) - F_X^{-1}(p)}\right], \tag{3.21}$$

which implies that

$$\pi(X) \geq F_X^{-1}(p) + \mathrm{E}\left[\left(X - F_X^{-1}(p)\right)_+\right] = \pi_l(X). \tag{3.22}$$

Remark 3.11. *Now we consider the risk measure*

$$\mathrm{E}\left[\psi\left(\frac{\left(X - F_X^{-1}(p)\right)_+}{\pi(X) - F_X^{-1}(p)}\right)\right] = 1 - p \tag{3.23}$$

for some parameter $0 < p < 1$. *Similarly as in Remark 3.12, if the function* $\psi(\cdot)$ *is restricted to satisfy* $\psi(1) = \psi'(1)$, *we obtain the lowest risk measure as*

$$\pi(X) = F_X^{-1}(p) + \frac{1}{1-p}\,\mathrm{E}\left[\left(X - F_X^{-1}(p)\right)_+\right] = \mathrm{E}\left[X \,\middle|\, X > F_X^{-1}(p)\right]. \tag{3.24}$$

Remark 3.12. *Another choice is to consider the root of the equation*

$$\mathrm{E}\left[\frac{1}{\psi(1)} \cdot \psi\left(\frac{|X - \mathrm{E}[X]|}{\pi(X) - \mathrm{E}[X]}\right)\right] = \alpha, \tag{3.25}$$

defining, in general terms, a risk measure for the deviation from the expectation. As a special case when $\psi(t) \equiv t^2 \mathrm{I}_{(t \geq 0)}$ *one gets*

$$\pi(X) = \mathrm{E}[X] + \frac{\sigma[X]}{\sqrt{\alpha}}. \tag{3.26}$$

Note that both (3.24) *and* (3.26) *produce an upper bound for the* α-*quantile* $q_\alpha(X)$ *of X.*

Remark 3.13. *For a risk variable X, one could consider a risk measure* $\pi_c(X)$ *which is additive, and define another risk measure* $\rho(X)$, *where the deviation* $a(X) = \rho(X) - \pi_c(X)$ *is determined by*

$$\mathrm{E}\left[\psi\left(\frac{|X - \pi_c(X)|}{a(X)}\right)\right] = \psi(1). \tag{3.27}$$

Here the role of $\pi_c(X)$ *is to measure central tendency while* $a(X)$ *measures the deviation of the risk variable X from* $\pi_c(X)$. *If* $\pi_c(X)$ *is positively homogenous,* *translation invariant and additive, then $\rho(X)$ is positively homogenous and translation invariant. The measure $\rho(X)$ may be subadditive or superadditive, depending on the convexity or concavity of the function $\psi(\cdot)$.*

3.6. Yaari's dual theory of choice under risk

Yaari (1987) introduced the dual theory of choice under risk. It was used by Wang (1996), who introduced distortion functions in the actuarial literature. A distortion function is defined as a non-decreasing function $g : [0,1] \longrightarrow [0,1]$ such that $g(0) = 0$ and $g(1) = 1$.

Definition 3.5. *Let S be a non-negative r.v. with d.f. F_S, and $g(\cdot)$ be a distortion function defined as above. The* distortion risk measure *associated with the distortion function g is defined by*

$$\pi = \int_0^{+\infty} g\left(1 - F_S(x)\right) \mathrm{d}x. \tag{3.28}$$

Choosing the function $\phi(\cdot,\cdot)$ in the equation [UE] such that $\phi(s,\pi) = s/\pi$ and using the left-hand derivative $g'_-(1-F_S(s))$ instead of $v(s)$, using integration by parts we get the desired unifying approach. The choice of $v(\cdot)$, which at first glance may look artificial, is very natural if one wants to have $\mathrm{E}[v(S)] = 1$.

This risk measure can be characterized by the following axioms:

A6.1. $\Pr[X \le Y] = 1 \Longrightarrow \pi(X) \le \pi(Y)$;

A6.2. If risks X and Y are comonotonic then $\pi(X+Y) = \pi(X) + \pi(Y)$;

A6.3. $\pi(1) = 1$.

Remark 3.14. *It is clear that this principle results in large upper bounds because*

$$\Pr[X \ge \pi + u] \le \mathrm{E}\left[\frac{X \cdot g'_-(1 - F_X(X))}{\pi + u}\right] = \frac{\pi}{\pi + u}. \tag{3.29}$$

It is also clear that the set of risks for which π is finite contains all risks with finite mean.

3.7. Mixtures of Esscher principles

The mixture of Esscher principles was introduced by Gerber and Goovaerts (1981). It is defined as follows:

Definition 3.6. *For a bounded r.v. S, we say a principle $\pi = \pi(S)$ is a* mixture of Esscher principles *if it is of the form*

$$\pi_F(S) = F(-\infty)\phi(-\infty) + \int_{-\infty}^{+\infty} \phi(t)\mathrm{d}F(t) + (1 - F(+\infty))\phi(+\infty), \tag{3.30}$$

where F is a non-decreasing function satisfying $0 \le F(t) \le 1$ *and* ϕ *is of the form*

$$\phi(t) = \phi_S(t) = \frac{\mathrm{d}}{\mathrm{d}t} \log \mathrm{E}\left[e^{tS}\right], \quad t \in \mathbb{R}. \tag{3.31}$$

Actually we can regard F as a possibly defective cdf with mass at both $-\infty$ and $+\infty$. Since the variable S is bounded, $\phi(-\infty) = \min[S]$ and $\phi(+\infty) = \max[S]$. In addition, $\phi_S(t)$ is the Esscher premium of S with parameter $t \in \mathbb{R}$.

In the special case where the function F is zero outside the interval $[0, \infty]$, the mixture of Esscher principles is a mixture of premiums with a non-negative safety loading coefficient. We show that in this case the mixture of Esscher premiums can also be derived from the Markov inequality. Actually,

$$\pi_F(S) = \int_0^{+\infty} \phi(t)\mathrm{d}F(t) + (1 - F(+\infty))\,\phi(+\infty) = \int_{[0,+\infty]} \phi(t)\mathrm{d}F(t). \tag{3.32}$$

It can be shown that the mixture of Esscher principles is translation invariant. Hence in what follows, we simply assume, without loss of generality, that $\min[S] \ge 0$ because otherwise a translation on S can be used. We notice that, for any $t \in [0, +\infty]$,

$$\phi(t) = \frac{\mathrm{E}\left[Se^{tS}\right]}{\mathrm{E}\left[e^{tS}\right]} \ge \mathrm{E}[S]. \tag{3.33}$$

The inequality (3.33) can, for instance, be deduced from the fact that the variables S and e^{tS} are comonotonic, hence positively correlated. Since we have assumed that $\min[S] \ge 0$, now we choose in [GMI] the functions $v(\cdot) \equiv 1$ and $\phi(s,\pi) = s/\pi$, then we obtain that

$$\Pr[S > \pi] \le \mathrm{E}[\phi(S,\pi)] \le \frac{1}{\pi}\mathrm{E}[S] \le \frac{1}{\pi}\int_{[0,+\infty]} \phi(t)\mathrm{d}F(t), \tag{3.34}$$

where the last step in (3.34) is due to the inequality (3.33) and the fact that $F([0,+\infty]) = 1$. Letting the r.h.s. of (3.34) be equal to 1, we immediately obtain (3.32).

We now verify another result: the tail probability $\Pr[S > \pi + u]$ decreases exponentially fast in $u \in [0, +\infty)$. The proof is not difficult. Actually, since the risk variable S is bounded, it holds for any $\alpha > 0$ that

$$\Pr[S > \pi + u] \le \exp\{-\alpha(\pi + u)\} \cdot \mathrm{E}[\exp\{\alpha S\}]. \tag{3.35}$$

Hence, in order to get the announced result, it suffices to prove that, for some $\alpha > 0$,

$$\mathrm{E}[\exp\{\alpha S\}] \le \exp\{\alpha\pi\} = \exp\left\{\alpha \int_{[0,+\infty]} \phi(t)\mathrm{d}F(t)\right\},$$

or equivalently to prove that, for some $\alpha > 0$,

$$\log \mathrm{E}[\exp\{\alpha S\}] \le \alpha \int_{[0,+\infty]} \phi(t)\mathrm{d}F(t). \tag{3.36}$$

In the trivial case where the risk S is degenerate, both sides of (3.36) are equal for any $\alpha > 0$. If F is not degenerate, the Esscher premium $\phi(t)$ is strictly increasing in $t \in [0,+\infty]$, and we can find some $\alpha_0 > 0$ such that

$$\phi(\alpha) \le \int_{[0,+\infty]} \phi(t)\mathrm{d}F(t) \tag{3.37}$$

holds for any $\alpha \in [0, \alpha_0]$. Thus in any case we obtain that (3.36) holds for any $\alpha \in [0, \alpha_0]$. We summarize:

Remark 3.15. *For the mixture of Esscher premiums π defined above, if F is concentrated on* $[0,+\infty]$*, then*

$$\Pr[S > \pi + u] \le \exp\{-\alpha_0 u\} \tag{3.38}$$

holds for any $u \ge 0$, where the constant $\alpha_0 > 0$ is the solution of the equation

$$\phi(\alpha) = \int_{[0,+\infty]} \phi(t)\mathrm{d}F(t). \tag{3.39}$$

The mixture of Esscher premiums is characterized by the following axioms; see Gerber and Goovaerts (1981):

A7.1. $\phi_{X_1}(t) \le \phi_{X_2}(t) \quad \forall t \in \mathbb{R} \Longrightarrow \pi_F(X_1) \le \pi_F(X_2)$;

A7.2. It holds for any two independent risks X_1 and X_2 that

$$\pi_F(X_1 + X_2) = \pi_F(X_1) + \pi_F(X_2). \tag{3.40}$$

Hence this risk measure is additive for independent risks. When the function F in (3.32) is non-zero only on the interval $[0,\infty]$, the premium contains a positive safety loading.

4. Conclusions

This paper shows how many of the usual premium calculation principles (or risk measures) can be deduced from a generalized Markov inequality. All risk measures provide information concerning the VaR, as well as the asymptotic behavior of $\Pr[S > \pi + u]$. Therefore, the effect of using a risk measure and requiring additional properties is equivalent to making a selection of admissible risks. Notice that when using a risk measure, additional requirements are usually needed about convergence of certain integrals. In this way, the set of admissible risks is restricted, e.g. the one having finite mean, finite variance, finite moment generating function and so on.

References

Artzner, Ph. (1999) Application of coherent risk measures to capital requirements in insurance. *North American Actuarial Journal* **3(2)**, 11-25.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999) Coherent measures of risk. *Mathematical Finance* **9**, 203-228.

Borch, K. (1968) *The economics of uncertainty*. Princeton University Press, Princeton.

Borch, K. (1974) *The mathematical theory of insurance*. Lexington Books, Toronto.

Bühlmann, H. (1970) *Mathematical Methods in Risk Theory*. Springer-Verlag, Berlin.

Denuit, M., Dhaene, J. and Van Wouwe, M. (1999) The economics of insurance: a review and some recent developments. *Mitt. Schweiz. Aktuarver.* **2**, 137-175.

De Vijlder, F. and Goovaerts, M.J. (1979) An invariance property of the Swiss premium calculation principle. *Mitt. Verein. Schweiz. Versicherungsmath.* **79(2)**, 105-120.

Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R. and Vyncke, D. (2002a) The concept of comonotonicity in actuarial science and finance: theory. *Insurance Math. Econom.* **31(1)**, 3-33.

Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R. and Vyncke, D. (2002b) The concept of comonotonicity in actuarial science and finance: application. *Insurance Math. Econom.* **31(2)**, 133-161.

Embrechts, P., McNeil, A.J. and Straumann, D. (2002) Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 176-223. Cambridge Univ. Press, Cambridge.

Gerber, H.U. (1974) On additive premium calculation principles. *ASTIN Bulletin* **7**, 215-222.

Gerber, H.U. (1979) *An Introduction to Mathematical Risk Theory*. Huebner Foundation Monograph **8**, distributed by Richard D. Irwin, Inc., Homewood, Illinois.

Gerber, H.U. and Goovaerts, M.J. (1981) On the representation of additive principles of premium calculation. *Scand. Actuar. J.* **4**, 221-227.

Goovaerts, M.J., De Vijlder, F. and Haezendonck, J. (1984) *Insurance Premiums.* North-Holland Publishing Co., Amsterdam.

Goovaerts, M.J., Kaas, R. and Dhaene, J. (2003) Economic capital allocation derived from risk measures, *North American Actuarial Journal* **7(2)**, 44-59.

Haezendonck, J. and Goovaerts, M.J. (1982) A new premium calculation principle based on Orlicz norms. *Insurance Math. Econom.* **1(1)**, 41-53.

Hardy, G.H., Littlewood, J.E. and Pólya, G. (1952) *Inequalities.* 2nd ed. Cambridge University Press.

Jarrow, R. (2002) Put option premiums and coherent risk measures. *Math. Finance* **12(2)**, 135-142.

Kaas, R., Goovaerts, M.J., Dhaene, J. and Denuit, M. (2001) *Modern Actuarial Risk Theory.* Dordrecht: Kluwer Acad. Publ.

Runnenburg, J.Th. and Goovaerts, M.J. (1985) Bounds on compound distributions and stop-loss premiums. *Insurance Math. Econom.* **4(4)**, 287-293.

von Neumann, J., and Morgenstern, O. (1944) *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, New Jersey.

Wang, S. (1996) Premium calculation by transforming the layer premium density. *ASTIN Bulletin* **26(1)**, 71-92.

Wang, S. and Young, V.R. (1998) Ordering risks: expected utility theory versus Yaari's dual theory of risk. *Insurance Math. Econom*. **22(2)**, 145-161.

Yaari, M.E. (1987) The dual theory of choice under risk. *Econometrica* **55(1)**, 95-115.

Marc Goovaerts
Center for Risk and Insurance Studies (CRIS)
Katholieke Universiteit Leuven
B-3000 Leuven, Belgium
E-mail: Marc.Goovaerts@econ.kuleuven.ac.be

Rob Kaas
Department of Quantitative Economics
University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
E-mail: r.kaas@uva.nl

Jan Dhaene
Center for Risk and Insurance Studies (CRIS)
Katholieke Universiteit Leuven
B-3000 Leuven, Belgium
E-mail: Jan.Dhaene@econ.kuleuven.ac.be

Qihe Tang
Department of Quantitative Economics
University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
E-mail: q.tang@uva.nl

OPTIMAL DYNAMIC XL REINSURANCE

BY

CHRISTIAN HIPP AND MICHAEL VOGT

University of Karlsruhe, Germany

ABSTRACT

We consider a risk process modelled as a compound Poisson process. We find the optimal dynamic unlimited excess of loss reinsurance strategy to minimize infinite time ruin probability, and prove the existence of a smooth solution of the corresponding Hamilton-Jacobi-Bellman equation as well as a verification theorem. Numerical examples with exponential, shifted exponential, and Pareto claims are given.

KEYWORDS

Stochastic control, Ruin probability, XL reinsurance

1. INTRODUCTION

Assume an insurance company has the possibility to choose and buy dynamically an unlimited excess of loss reinsurance. For this situation, stochastic control theory is used to derive the optimal reinsurance strategy which minimizes ruin probability when the reinsurer computes his premium according to the expected value principle. The corresponding problem has been solved by Schmidli (2000) for the case of dynamic proportional reinsurance.

We model the risk process R_t of an insurance company by a Lundberg process with claim arrival intensity λ and absolutely continuous claim size distribution Q. The number of claims A_t in a time interval $(0,t]$ is a Poisson process with intensity λ, and the claim sizes U_i, $i = 1, 2, \ldots$ are positive iid variables independent of A_t. Let T_i be the occurrence time of the i-th claim, $i = 1, 2, \ldots, c$ the premium intensity of the insurer which contains a positive safety loading

$$c > \lambda E[U_i],$$

and $s = R_0$ the initial reserve. Then – without reinsurance – the surplus of the insurance company at time t is

$$R_t = s + ct - \sum_{i=1}^{A_t} U_i .$$

The reinsurer uses the expected value principle with safety loading $\theta > 0$ for premium calculation. We assume $(1 + \theta)\lambda E[U_i] > c$, because otherwise the insurer could get rid of all his risk by reinsuring his total portfolio.

Excess of loss reinsurance is a non proportional risk sharing contract in which, for a given retention level $b \geq 0$ a claim of size U is divided into the cedent's payment $\min\{U, b\}$ and the reinsurer's payment $(U - b)^+ = U - \min\{U, b\}$. In this paper the retention level is assumed to be chosen dynamically, i.e. the insurer adjusts the retention level b_t at every time $t \geq 0$, based on the information available just before time t: If $\mathcal{F}_t$ is the sigma-field generated by $R_u, u \leq t$, then b_t is assumed to be predictable (a pointwise limit of left continuous $\mathcal{F}_t$ adapted processes), i.e. it is a measurable function of s and the times and sizes of claims occurring before t. It can be represented by a sequence of functions $\pi_n, n = 0, 1, 2, \ldots$ with $\pi_n : \mathbb{R}^{2n+1} \to \mathbb{R}$ measurable and

$$b_t = \pi_n(T_1, \ldots, T_n, U_1, \ldots, U_n, t - T_n) \text{ for } T_n < t \leq T_{n+1}.$$

We will show that the optimal reinsurance strategy exists and is given via a feedback equation of the following form:

$$b_t = b(R_{t-}^b),$$

where R_t^b is the surplus process with strategy b_t, and $b(s)$ is a measurable function. In particular, the optimal strategy is Markovian, i.e. it depends on the actual surplus only and not on the history of the process. Let b_t be an arbitrary dynamic reinsurance strategy. Then with $\rho = (1 + \theta)\lambda$,

$$R_t^b = s + ct - \rho \int_0^t E\left[(U - b_x)^+\right] dx - \sum_{i=1}^{A_t} \min\left\{U_i, b_{T_i}\right\} \tag{1}$$

is the surplus process under the strategy b_t.

Our aim is to minimize ruin probability which is the same as maximizing survival probability. The ruin time τ_b is the first time the surplus of the insurance company ever becomes negative using reinsurance strategy b_t. It is given by

$$\tau_b = \inf\left\{t \geq 0 : R_t^b < 0\right\}.$$

Then we can write the ruin probability as

$$\psi_b(s) = P(\tau_b < \infty).$$

With

$$\delta_b(s) = P\left(\tau_b = \infty \mid R_0 = s\right) = 1 - \psi_b(s)$$

we will compute the function

$$\delta(s) = \sup_b \{\delta_b(s)\},$$

and find an optimal strategy b_t^*, such that $\delta(s) = \delta_{b^*}(s)$.

A more realistic problem would have a loading of the reinsurer which varies with the retention level (e.g. if instead of the expected value principle one would use the variance principle). Furthermore, one should also consider limited XL-covers, and then both, the retention and the limit, will be considered as control variables, see [7].

2. Hamilton-Jacobi-Bellman equation

The computation of the optimal reinsurance strategy is based on the classical Hamilton-Jacobi-Bellman equation which can be derived heuristically considering (1) on a short time interval $[0,\Delta]$ in which a constant strategy b is used. One of the following two cases can occur:

1. There is no claim in $[0,\Delta]$, which happens with probability $1 - \lambda\Delta + o(\Delta)$. Then the reserve of the company at time Δ is given by

$$R_\Delta = s + (c - \rho E[(U-b)^+])\Delta.$$

2. There is exactly one claim with claim size $U \sim Q$ in $(0,\Delta]$ and this happens with probability $\lambda\Delta + o(\Delta)$. Then the reserve can be written as

$$R_\Delta = s + (c - \rho E[(U-b)^+])\Delta - \min\{U,b\}.$$

Taking expectations and averaging over all possible claim sizes, we arrive at the equation

$$\delta_b(s) = (\lambda\Delta + o(\Delta))\, E\left[\delta\left(s + \left(c - \rho E\left[(U-b)^+\right]\right) - \min\{U,b\}\right)\right]$$
$$+ (1 - \lambda\Delta + o(\Delta))\,\delta\left(s + \left(c - \rho E\left[(U-b)^+\right]\right)\Delta\right) + o(\Delta).$$

For $\Delta \to 0$ we obtain for a smooth function $\delta(s)$

$$0 = \lambda E[\delta(s - \min\{U,b\})] - \lambda\delta(s) + \left(c - \rho E[(U-b)^+]\right)\delta'(s)$$

and finally by maximizing over all possible values for b the Hamilton-Jacobi-Bellman equation for our optimization problem:

$$0 = \sup_{b>0}\left\{\lambda E[\delta(s - \min\{U,b\}) - \delta(s)] + (c - \rho E[(U-b)^+])\delta'(s)\right\} \qquad (2)$$

An optimal strategy is derived from a solution $(\delta(s), b^*(s))$ of the equation (2), where $b^*(s)$ is the point at which the supremum in (2) is attained.

The insurance company has a non negative net premium income if

$$c \geq \rho E[(U-b)^+].$$

Let $\underline{b}$ be the value where equality holds:

$$c = \rho E[(U-\underline{b})^+].$$

Since we are looking for a nondecreasing solution of equation (2) we can rewrite it as

$$\delta'(s) = \inf_{b > \underline{b}} \left\{ \lambda \frac{\delta(s) - E[\delta(s - \min\{U, b\})]}{c - \rho E\left[(U-b)^+\right]} \right\}. \tag{3}$$

3. Existence of a solution

In this section we shall prove the existence of a solution of equation (2). This will be done through a monotonicity argument, similar to the approach in [2].

Theorem 1 *Assume the claim size distribution Q is absolutely continuous. There exists a nondecreasing solution* $V(s)$ *of the Hamilton-Jacobi-Bellman equation (2) which is continuous on* $[0,\infty)$, *continuously differentiable on* $(0,\infty)$, *with* $V(s) = 0$ *for* $s < 0$, *and* $V(s) \to 1$ *for* $s \to \infty$.

Proof. Define a sequence $V_n(s)$ via $V_0(s) = \delta_0(s)$, the ruin probability without reinsurance (which means $b = \infty$ or $b = M$ if $P(U \geq M) = 0$) for $n = 0$, and through the recursion

$$V'_{n+1}(s) = \inf_{b > 0} \left\{ \lambda \frac{V_n(s) - E\left[V_n(s - \min\{U, b\})\right]}{c - \rho E[(U-b)^+]} \right\}, n = 0, 1, \ldots. \tag{4}$$

We show by induction that $V_n'(s)$, $n = 0, 1, 2, \ldots$ is a decreasing sequence. For $n = 0$ we have

$$V_0'(s) = \lambda \frac{V_0(s) - E[V_0(s-U)]}{c}$$

(see [1]. p. 4) and from (4) we get for $n = 0$:

$$V_1'(s) = \inf_{b>0} \lambda \left\{ \frac{V_0(s) - E[V_0(s - \min\{U, b\})]}{c - \rho E[(U-b)^+]} \right\}.$$

Thus we have $V_1'(s) \leq V_0'(s)$ for all $s \geq 0$. Now let $n \geq 1$ and s be fixed. For all b we have

$$
\begin{aligned}
V'_{n+1}(s)(c-\rho E[(U-b)^+]) &\leq \lambda V_n(s)-\lambda E[Vn(s-\min\{U,b\})] \\
&= \lambda E\left[\int_{s-\min\{U,b\}}^{s} V'_n(u)du\right] \\
&\leq \lambda E\left[\int_{s-\min\{U,b\}}^{s} V'_{n-1}(u)du\right] \\
&= \lambda V_{n-1}(s)-\lambda E\left[V_{n-1}(s-\min\{U,b\})\right].
\end{aligned}
$$

Here we used the induction hypothesis $V'_n(s) \leq V'_{n-1}(s)$ for all $s \geq 0$. Since b was arbitrary, we can switch to the infimum which gives us the required result

$$
V'_{n+1}(s) \leq V'_n(s).
$$

So $V'_n(s)$ is a decreasing sequence of continuous functions, and since $V'_n(s) > 0$ the sequence $V'_n(s)$ converges to a function $g(s)$, and with

$$
V(s) = 1 - \int_s^\infty g(u)du
$$

we have a nondecreasing continuous function $V(s)$ satisfying

$$
g(s) = \inf_{b \geq \underline{b}} \left\{ \lambda \frac{V(s) - E[V(s - \min\{U,b\})]}{c - \rho E\left[(U-b)^+\right]} \right\}.
$$

What is left is a proof for continuity of $g(s)$: then

$$
V'(s) = g(s)
$$

is continuous, and $V(s)$ satisfies equation (2). We first show that $g(s) > 0$ for all $s \geq 0$. The function $g(s)$ is the limit of the functions $V'_n(s)$. If the infimum in (4) is not attained in $[\underline{b}, s]$ then it is attained at $b = \infty$, and hence $V'_n(s)$ and $V'_0(s)$ are proportional for small s, i.e.

$$
V'_n(s) \propto V'_0(s), 0 \leq s < \underline{b}.
$$

Furthermore,

$$
g(0) = \frac{\lambda V(0)}{c} > 0,
$$

which implies $g(s) > 0$ for $0 \leq s < \underline{b}$. Assume that

$$
s_0 = \inf\{s: g(s) = 0\} < \infty.
$$

Then $s_0 \geq \underline{b}$, and there exists $s_0 \leq s < s_0 + \underline{b}$ for which $g(s) = 0$ or

$$\inf_{b \geq \underline{b}} \{V(s) - E[V(s - \min\{U, b\}]\} =$$
$$V(s) - E[V(s - \min\{U, \underline{b}\})] = 0,$$

i.e. $V(s) = V(s - \underline{b})$ (notice that $P(U > \underline{b}) > 0$). Then

$$0 = \int_{s-\underline{b}}^{s} g(u)du \geq \int_{s-\underline{b}}^{s_0} g(u)du$$

which contradicts the choice of s_0.

We next show that in the definition of the functions $V_n(s)$, $s \leq K$, the infimum can be restricted to the region $[b_1, \infty]$, where $b_1 > \underline{b}$. Assume the contrary, i.e. there exists a sequence $0 \leq s_n \leq K$ and $b_n \to \underline{b}$ such that

$$V'_{n+1}(s_n) \geq \lambda \frac{V_n(s_n) - E[V_n(s_n - \min\{U, b_n\})]}{c - \rho E[(U - b_n)^+]} - \frac{1}{n} \geq V'_{n+1}(s_n) - \frac{1}{n}.$$

Since $0 \leq V'_n(s) \leq V'_0(s)$ and $c - \rho E[(U - b_n)^+] \to 0$, we obtain

$$V_n(s_n) - E[V_n(s_n - \min\{U, b_n\})] \to 0,$$

and therefore for each accumulation point s_0 of the sequence s_n

$$V(s_0) - E\left[V\left(s_0 - \min\{U, \underline{b}\}\right)\right] = 0 = g(s_0),$$

a contradiction.

Finally, the relation

$$|g(x) - g(y)|$$
$$\leq \sup_{b \geq b_1} \lambda \frac{V(x) - E[V(x - \min\{U, b\})]}{c - \rho E[(U - b)^+]} - \lambda \frac{V(y) - E[V(y - \min\{U, b\})]}{c - \rho E[(U - b)^+]}$$

for $x, y \geq 0$ implies continuity of $g(s)$. ■

Remark 1 Let

$$V(s, b) = \lambda \frac{V(s) - E[V(s - \min\{U, b\})]}{c - \rho E[(U - b)^+]}$$

where $V(s)$ is a smooth solution of the Bellman equation (3) with the properties of Theorem 1. Then the infimum over $b \geq \underline{b}$ is either

$$\frac{\lambda}{c}(V(s) - E[V(s - U)])$$

(no reinsurance or $b = \infty$) or

$$\lambda \frac{V(s) - E[V(s-U)] - V(0)P\{U \geq s\}}{c - \rho E[(U-s)^+]}$$

$(b = s)$ or

$$\inf_{\underline{b} < b < s} \lambda \left\{ \frac{V(s) - E[V(s - \min\{U, b\})]}{c - \rho E[(U-b)^+]} \right\}.$$

Since $V(s, b)$ has a continuous derivative w.r.t. b in $(\underline{b}, s)$ this last infimum – if attained in this interval – is attained at the point b for which this derivative is zero or

$$\lambda V'(s-b) = \rho V'(s).$$

So in each case there is (possibly more than) one point $b \in (\underline{b}, \infty]$ at which the infimum is attained, and a measurable selection of these points yields a measurable function $b(s)$. The corresponding strategy b_t^* is admissible, it can be represented by

$$\pi_n(T_1, \ldots, T_n, U_1, \ldots, U_n, t) = b(R(T_n) + B(t)),$$

where $R(T_n)$ is a measurable function of $T_1, \ldots, T_n, U_1, \ldots, U_n$ and

$$B(t) = ct - \rho \int_0^t E[(U - b_x)^+] dx.$$

The retention $b = \infty$ (no reinsurance) will be optimal for small values of s: For $s \leq \underline{b} < b$ we have

$$V(s, b) = \lambda \frac{V(s) - E[V(s-U)]}{c - \rho E[(U-b)^+]}$$

which is maximal for $b = \infty$.

Remark 2. The function $V(s)$ will not be concave in general: notice that the survival probability $\delta_0(s)$ will not be concave in general, and $V'(s)$ will be proportional to $\delta_0'(s)$ for $0 \leq s \leq \underline{b}$. However, if $\delta_0(s)$ is concave, the function $V(s)$ constructed above will be concave, too. To see this we have to show that all the above functions $V_n(s)$ are concave which is done by induction. If $V_n(s)$ is concave, then for $s \geq 0$ and $h > 0$ we have for arbitrary b

$$\begin{aligned} & V_n(s+h) - E[V_n(s+h-\min\{U,b\})] \\ & = E\int_{s+h-\min\{U,b\}}^{s+h} V_n'(u)du \\ & \leq E\int_{s-\min\{U,b\}}^{s} V_n'(u)du \\ & = V_n(s) - E\left[V_n(s - \min\{U,b\})\right] \end{aligned}$$

and hence

$$V'_{n+1}(s+h) \le V'_{n+1}(s).$$

4. Verification theorem

In this section we will show that the strategy b_t^* derived from the maximizer $b^*(s)$ in (2) maximizes the survival probability. This is done through the following verification theorem. Notice that this theorem also implies uniqueness of the solution.

Theorem 2 *The strategy b_t^* maximizes survival probability: For any* s > 0 *and arbitrary predictable strategy b_t with survival probability $\delta(s)$ we have*

$$V(s) \ge \delta(s),$$

with equality for $b_t = b_t^*$.

Proof. Let $V(s)$ be the smooth solution of (3) constructed in chapter 3, for which

$$0 \le V(s) \le 1$$

and

$$\lim_{s \to \infty} V(s) = 1.$$

We write $R(t)$ and $R^*(t)$ for the risk process of the insurance company with reinsurance strategy b_t and b_t^*, respectively, and initial capital s. Let τ and τ^* be the corresponding ruin times, X_t^*, X_t the stopped processes and W_t^*, W_t the stopped processes, transformed by $V(s)$, i.e.

$$W_t^* = V(X_t^*) = V\left(R^*\left(\min\{t, \tau^*\}\right)\right),$$
$$W_t = V(X_t) = V\left(R\left(\min\{t, \tau\}\right)\right).$$

Then, as in [5], p. 80, (2.16), we obtain

$$E[W_t] = V(s) + E\left[\int_0^t V'(X_s)\left(c - \rho E[(U - b_s)^+]\right) ds \right.$$
$$\left. + \lambda \int_0^t E[V(X_s - \min\{U, b_s\}) - V(X_s)] ds, \right]$$

and a corresponding formula for W_t^*. From the Hamilton-Jacobi-Bellman equation (2) we see that for all $t > 0$

$$E[W_t^*] = V(s) \geq E[W_t]. \tag{5}$$

Assume first that the predictable strategy b_t satisfies

$$b_t \geq B > 0 \text{ for all } t \geq 0, \tag{6}$$

where B satisfies $P\{U > B\} > 0$. We show that in this case the process $R(t)$ is unbounded on $\{\tau = \infty\}$. For this we prove

$$P\{R(t) \leq M \text{ for all } t \geq 0 \text{ and } \tau = \infty\} = 0 \tag{7}$$

for all $M > 0$. With $n > (M + c)/B$ the probability of more than n claims of size larger than B in an interval of length 1 is positive. Since the claims process has stationary and independent increments, with probability 1 there are more than n such claims in an interval $[t, t+1]$. For $R(t) \leq M$ we have

$$R(t+1) \leq M + c - nB < 0,$$

i.e. $\tau < \infty$. This proves (7).

For arbitrary ε we now construct a strategy b_t^+ with risk process $R^+(t)$ and ruin time τ^+ such that $P\{\tau = \infty \text{ and } \tau^+ < \infty\} < \varepsilon$ and $R^+(t) \to \infty$ on $\{\tau = \infty$ and $\tau^+ = \infty\}$. Let $M > s$ be sufficiently large such that $1 - \delta_0(M) < \varepsilon$ let $T = \inf\{t : R(t) = M\}$ which is finite almost everywhere on $\{\tau = \infty\}$, and define

$$b_t^+ = \begin{cases} b_t & \text{if } t \leq T \\ \infty & \text{if } t > T. \end{cases}$$

The strategy b_t^+ is predictable, and

$$P\{\tau = \infty, \tau^+ < \infty\} \leq 1 - \delta_0(M) < \varepsilon.$$

Furthermore, $T < \infty$ implies $R^+(t) \to \infty$.

Now repeat the above reasoning leading to (5) for $R^+(t)$ instead of $R(t)$. We obtain

$$E\left[V\left(R^*\left(\min\{t, \tau^*\}\right)\right)\right] = V(s) \geq E\left[V\left(R^+\left(\min\{t, \tau^+\}\right)\right)\right],$$

and with $t \to \infty$ we arrive with $V(R^*(\tau^*)) = 0$ and $V(R^+(\tau^+)) = 0$ at

$$\begin{aligned} P\{\tau^* = \infty\} \geq V(s) &\geq P\{\tau = \infty \text{ and } \tau^+ = \infty\} \\ &\geq P\{\tau = \infty\} - \varepsilon. \end{aligned}$$

Since ε was arbitrary, this is our assertion for the special case of a strategy b_t with property (6). In particular, since any solution $V(s)$ of (3) in the sense of

Theorem 1 will produce a strategy satisfying (6), we have uniqueness of the solution and $V(s) = P\{\tau^* = \infty\}$.

Next we show that for premium intensities $c, \bar{c}$ with $\bar{c} > c$ we have

$$W'(s) \geq \overline{W}'(s) \text{ for all } s \geq 0,$$

where $W(s)$ and $\overline{W}(s)$ are solutions to (3) with c and $\bar{c}$, respectively and $W(0) = \overline{W}(0) = \alpha$. Notice, that $W(s)$ and $\overline{W}(s)$ do not solve (3) in the sense of Theorem 1, the conditions $W(s) \to 1$ for $s \to \infty$ and $\overline{W}(s) \to 1$ for $s \to \infty$ will not hold. Let $W_n(s), \overline{W}_n(s)$ be the sequences constructed in the proof of Theorem 1 converging to $W(s)$ and $\overline{W}(s)$ with $W(s)$, respectively $\overline{W}(s)$ defined by

$$W(s) = \alpha + \int_0^s g(u)du \text{ and } \overline{W}(s) = \alpha + \int_0^s \bar{g}(u)du,$$

where $g(s)$, $\bar{g}(s)$ are the limits of the sequences $W_n'(s)$, $\bar{W}_n'(s)$. We prove by induction that

$$W_n'(s) \geq \overline{W}_n'(s), \quad n = 0, 1, 2, \ldots. \tag{8}$$

For $n = 0$ we have

$$W_0'(s) = \frac{\lambda}{c}(W_0(s) - E[W_0(s-U)]),$$
$$\overline{W_0'}(s) = \frac{\lambda}{\bar{c}}(\overline{W}_0(s) - E[\overline{W}_0(s-U)]).$$

At $s = 0$ we have $W_0'(s) > \bar{W}_0'(s)$. Assume now that

$$s_0 = \inf\{s\colon W_0'(s) \leq \overline{W}_0'(s)\} < \infty.$$

By continuity, $s_0 > 0$. Then

$$\overline{W}'_0(s_0) = \frac{\lambda}{\bar{c}} E\left[\int_{s_0 - U}^{s_0} \overline{W}'_0(u)du\right] \leq \frac{\lambda}{\bar{c}} E\left[\int_{s_0 - U}^{s_0} W_0'(u)du\right] < W_0'(s_0),$$

a contradiction. Assume now that (8) holds for n. Then for all $b > 0$

$$W_n(s) - E[W_n(s - \min\{U, b\})] = E\left[\int_{s - \min\{U,b\}}^{s} W_n'(u)du\right] \geq$$
$$E\left[\int_{s - \min\{U,b\}}^{s} \overline{W_n'}(u)du\right] = \overline{W}_n(s) - E\left[\overline{W}_n(s - \min\{U, b\})\right]$$

which implies $W_{n+1}'(s) \geq \overline{W}_{n+1}'(s)$ and finally the desired result $W(s) \geq \overline{W}(s)$ for all s.

Now let c_n converge monotonically to c from above, and $W_n(s)$ the corresponding solutions of (3) with c_n instead of c. Then the sequence of functions $W_n'(s)$ is monotone and bounded by $W_0'(s)$, the function corresponding to c. Let $g(s)$ and $W(s)$ be the limits of $W_n'(s)$ and $W_n(s)$, respectively. As in the proof of Theorem 1 we obtain continuity of $g(s)$, and so $W'(s) = g(s)$ and $W(s)$ is a solution of (3) with c. Uniqueness of the solution for every α implies $W(s) = W_0(s)$.

To obtain a solution $V(s)$ of (3) satisfying $V(s) \to 1$ for $s \to \infty$ let $V_n(s)$ be a sequence of functions with

$$V_n(s) = \frac{V_n(0)}{\alpha} W_n(s),$$

then for $n \to \infty$ we have

$$V_n(s) \to \frac{\gamma}{\alpha} W_0(s) \text{ or}$$
$$V_n(s) \to \gamma' V(s).$$

For $s \to \infty$ we have

$$1 = \frac{\gamma}{\alpha} W_0(s) = \gamma' V(s)$$

and therewith $\gamma' = 1$. The same argumentation with $\overline{V}_n(s)$ instead of $V_n(s)$ leads us to $\overline{V}(s)$ with $\overline{V}(s) \geq V(s)$ for $\bar{c} > c$. For fixed s and arbitrary small $\varepsilon > 0$ we can find $\bar{c} > c$ for which

$$\overline{V}(s) < V(s) + \varepsilon.$$

Let $\overline{R}(t)$ be the risk process with strategy b_t, premium intensity $\bar{c}$ and $\bar{\tau}$ its ruin time. Then on $\{\tau = \infty\}$ we have $\overline{R}(t) \to \infty$ and hence, with $\bar{\tau} \geq \tau$

$$\begin{aligned} P\{\tau = \infty\} &= \lim_{t\to\infty} E\left[\overline{V}(\overline{R}(\min\{t,\bar{\tau}\}))1_{\{\tau = \infty\}}\right] \\ &\leq \lim_{t\to\infty} E\left[\overline{V}(\overline{R}(\min\{t,\bar{\tau}\}))\right] \leq \overline{V}(s) < V(s) + \varepsilon. \end{aligned}$$

So with $\varepsilon \to 0$

$$P\{\tau^* = \infty\} \geq V(s) \geq P\{\tau = \infty\}$$

which proves the verification theorem. ■

5. Numerical Examples

Here we present numerical computations for three different claim size distributions. Our first example has exponential claim sizes with mean $1/m$. Even in this simple case it seems to be impossible to find an analytical solution of (2).

The survival probability of an insurance company using no reinsurance, i.e. $b_t = \infty$ for all t, can be expressed explicitly by

$$\delta(s) = 1 - \frac{\lambda}{mc}\exp\left(-\left(m - \frac{\lambda}{c}\right)s\right) \tag{9}$$

(see [4]. p. 164). We will use the same parameters as in [6], i.e. $m = 1$, (which implies $\mu = 1$), $\lambda = 1$, premium rate $c = 1.5$ and $\rho = 1.7$. Since $V(0)$ is unknown we start with $V(0) = \delta(0)$ from (9) and norm the function $V(s)$ replacing $V(s)$ by $V(s)/V(s_1)$ where s_1 is sufficiently large. Figure 1 gives the survival probabilities for no reinsurance (lower graph) and for optimal excess of loss reinsurance (upper graph) for reserves $s \in [0, 15]$. We see that optimal excess of loss reinsurance gives a considerably higher survival probability. Figure 2 gives the optimal strategy $b^*(s)$ for values $s \in [0, 5]$; for $s \geq 5$ the optimal strategy is nearly constant. For small s the optimal strategy is, as expected, to keep the whole risk. At the point $s \approx 0.376$ the optimal strategy is $b(s) = s$, which means that independent of the following claim size the reserve remains nonnegative immediately after the claim. For $s \gtrsim 0.797$ we have to choose strategies $b(s) < s$ and the optimal strategy tends to be constant. Figure 3 is used to explain the optimal strategy presented in Figure 2. For each curve we fixed s ($s = 0.4$ at upper graph, then $s = 0.59, s = 0.6, s = 0.61, s = 0.8$ and finally $s = 0.9$ at lowest graph) and calculated $V(s, b)$ (defined in Remark 1) for varying $b \in [0.15, 1]$. For s small $V(s, b)$ is minimized for $b = \infty$. For $s \in [0.376, 0.797]$ the minimum is achieved at the jump, which means $b = s$. For larger values of s, here for

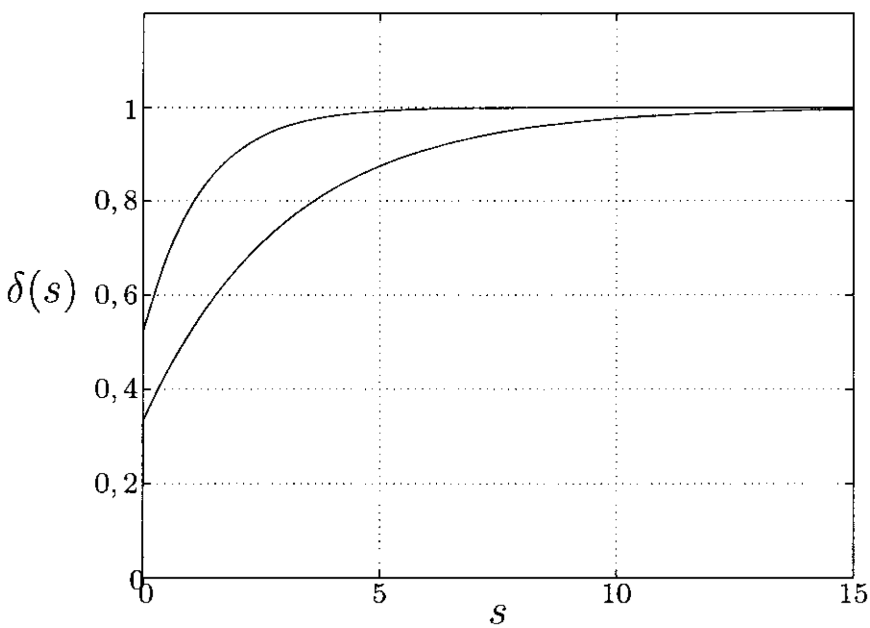


Figure 1: Survival probabilities

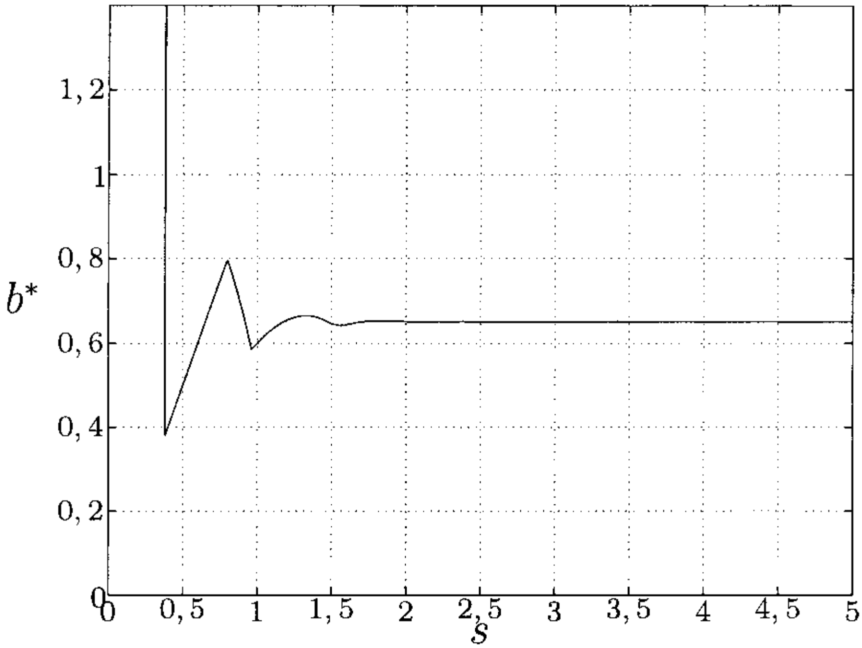


Figure 2: Optimal strategy for exponential distribution

example $s = 0.9$ the minimal $V(s, b)$ is achieved at a point before the jump. Looking at

$$E[\delta(s - \min\{U, b\})] = \int_0^b \delta(s - x)f(x)dx + \delta(s - b)(1 - F(b))$$

we can see why the jumps occur, if $b > s$ the term $\delta(s - b)(1 - F(b))$ equals zero. Figure 4 gives the optimal strategy $b^*(s), s \in [0, 15]$ in the case of a non concave solution $V(s)$. To achieve such a $V(s)$ we use a distribution with density

$$p(x) = m(\exp(-m(x - 1))), \quad x > 1$$

which is an exponential distribution shifted by 1, and solve the corresponding Hamilton-Jacobi-Bellman equation for parameters $m = 1$, $\lambda = 1$ and premium rates $c = 3$ and $\rho = 3.5$. Notice that in this case we have to choose $c > 2$ to keep the condition $c > \lambda E[U_i]$. In the last example we consider Pareto distributed claim sizes with parameter $a = 2$, i.e. claims with density

$$p(x) = 2(1 + x)^{-3}, x > 0.$$

Like in the first example we choose $\lambda = 1$ and the premium rates $c = 1.5$ and $\rho = 1.7$. Without reinsurance the survival probability at $s = 0$ is

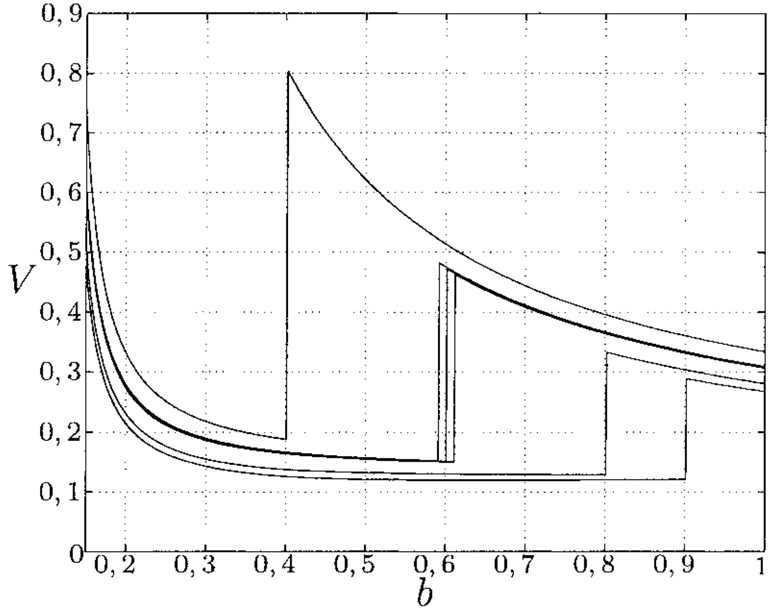


Figure 3: $V(s,b)$ for different values of s and varying b

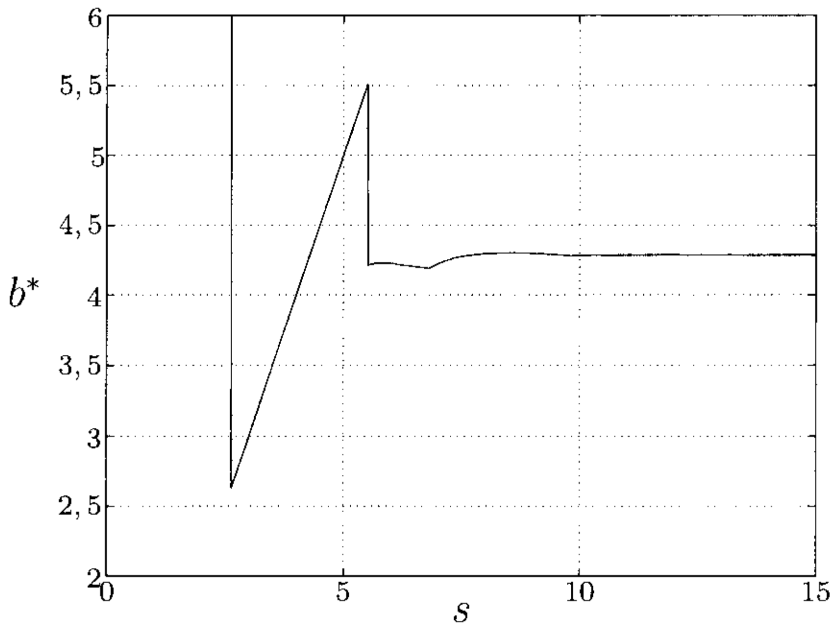


Figure 4: Optimal strategy for shifted exponential distribution

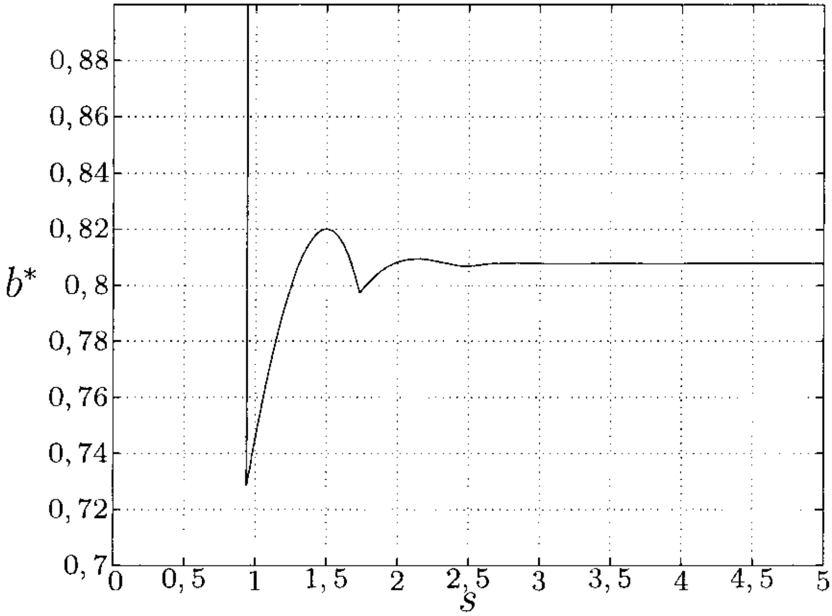


Figure 5: Optimal strategy for Pareto distribution

$$\delta(0) = 1 - \frac{\lambda}{c(a-1)}.$$

In Figure 5 we show the optimal reinsurance strategy for $s \in [0, 5]$ Contrary to the case of exponential distributed claim sizes there exists no interval in which we can choose $b^*(s) = s$. The optimal strategy for large values of s is constant, $b^*(s) \approx 0.8077$ for $s = 5$.

References

[1] Grandell, J. (1991) *Aspects of Risk Theory*. Springer.

[2] Hipp, C. and Taksar, M. (2000) Stochastic Control for Optimal New Business. *Insurance: Mathematics and Economics* **26**, 185-192.

[3] Hipp, C. and Plum, M. (2000) Optimal Investment for Insurers. I*nsurance, Mathematics and Economics* **27**, 215-228.

[4] Rolski, T., Schmidli, H., Schmidt, V. and Teugels, J. (1998) Stochastic Processes for Insurance and Finance. *Wiley Series in Probability and Statistics* **15.**

[5] Schäl, M. (1998) On piecewise deterministic Markov control processes: Control of jumps and of risk processes in insurance. *Insurance: Mathematics and Economics* **22**, 75-91.

[6] Schmidli, H. (2000) Optimal Proportional Reinsurance Policies in a Dynamic Setting. Research Report 403, Dept. Theor. Statis, Arhus University. 16.

[7] Vogt, M. (2003) Optimale dynamische Rückversicherung – ein Kontrolltheoretischer Ansatz. Dissertation, Universität Karlsruhe (TH).

COMMON POISSON SHOCK MODELS:
APPLICATIONS TO INSURANCE AND CREDIT RISK MODELLING

BY

FILIP LINDSKOG* AND ALEXANDER J. MCNEIL*

ABSTRACT

The idea of using common Poisson shock processes to model dependent event frequencies is well known in the reliability literature. In this paper we examine these models in the context of insurance loss modelling and credit risk modelling. To do this we set up a very general common shock framework for losses of a number of different types that allows for both dependence in loss frequencies across types and dependence in loss severities. Our aims are threefold: to demonstrate that the common shock model is a very natural way of approaching the modelling of dependent losses in an insurance or risk management context; to provide a summary of some analytical results concerning the nature of the dependence implied by the common shock specification; to examine the aggregate loss distribution that results from the model and its sensitivity to the specification of the model parameters.

1. INTRODUCTION

Suppose we are interested in losses of several different types and in the numbers of these losses that may occur over a given time horizon. More concretely, we might be interested in insurance losses occurring in several different lines of business or several different countries. In credit risk modelling we might be interested in losses related to the default of various types of counterparty. Further suppose that there are strong a priori reasons for believing that the frequencies of losses of different types are dependent. A natural approach to modelling this dependence is to assume that all losses can be related to a series of underlying and independent *shock* processes. In insurance these shocks might be natural catastrophes; in credit risk modelling they might be a variety of economic events such as local or global recessions; in operational risk modelling they might be the failure of various IT systems. When a shock occurs this may cause losses of several different types; the common shock causes the numbers of losses of each type to be dependent.

* Research of the first author was supported by Credit Suisse Group, Swiss Re and UBS AG through RiskLab, Switzerland. We thank in particular Nicole Bäuerle for commenting on an earlier version of this paper.

This kind of construction is very familiar in the *reliability* literature where the failure of different kinds of system components is modelled as being contingent on independent shocks that may affect one or more components. It is commonly assumed that the different varieties of shocks arrive as independent *Poisson processes*, in which case the counting processes for the different loss types are also Poisson and can be easily handled analytically. In reliability such models are known as *fatal shock models*, when the shock always destroys the component, and *non-fatal shock models*, or *not-necessarily-fatal shock models*, when components have a chance of surviving the shock. A good basic reference on such models is Barlow and Proschan (1975) and the ideas go back to Marshall and Olkin (1967).

In this paper we set up a very general Poisson shock model; the dimension is arbitrary and shocks may be fatal or not-necessarily-fatal. We review and generalise results for the multivariate Poisson process counting numbers of failures of different types. We also consider the modelling of dependent severities. When a loss occurs, whether in insurance or credit risk modelling, a loss size may be assigned to it. It is often natural to assume that losses of different types caused by the same underlying shock also have dependent severities. We set up general multivariate compound Poisson processes to model the losses of each type. Our interest focusses on three distributions in particular, and their sensitivity to the finer details of the parameterization of the model:

- The multivariate compound Poisson distribution of the cumulative losses of different types at some fixed point in time.
- The multivariate exponential distribution of the times to the first losses of each type.
- The univariate compound Poisson aggregate loss distribution at a fixed time point.

There have been a number of other related papers in this area in recent years, particularly concentrating on the second of these issues. In Savits (1988) nonhomogeneous Poisson shock processes are investigated and the effect of different mean functions for the shock processes on the distributional properties of the joint component lifetimes is studied. In Li and Xu (2001) the authors investigate stochastic bounds and dependence properties of the joint component lifetime distribution for rather general shock arrival processes. In particular the effect of dependent interarrival times of the shocks and the effect of simultaneous shock arrivals on the joint component lifetime distribution are investigated; the joint impact of these two types of dependency on the behaviour of the system is analysed.

The present paper is structured as follows. In Section 2 we describe the general not-necessarily-fatal-shock model with dependent loss frequencies and dependent loss severities. In Section 3 we ignore loss severities and examine the multivariate distribution of loss frequencies and the consequences for the aggregate loss frequency distribution of specifying the shock structure in different ways. An important key to analysing the model is to see that it may be written in terms of an *equivalent fatal shock model*. This facilitates the approximation of the aggregate loss frequency distribution using the *Panjer recursion* approach and also makes it very easy to analyse the multivariate exponential distribution of the *times to the first losses* of each type. In section 4 the analysis is generalised by including dependent loss severities. The dependence in severities is created using *copula techniques* and the object of interest is now the tail of the overall aggregate loss distribution. Sections 3 and 4 are illustrated with a stylized insurance example; Section 5 consists of an extended example of how the model might be applied to the modelling of portfolio credit risk.

2. The Model

2.1. Loss Frequencies

Suppose there are m different types of shock or event and, for $e = 1, \ldots, m$, let

$$\left\{N^{(e)}(t), t \geq 0\right\}$$

be a Poisson process with intensity $\lambda^{(e)}$ recording the number of events of type e occurring in $(0, t]$. Assume further that these shock counting processes are independent. Consider losses of n different types and, for $j = 1, \ldots, n$, let

$$\left\{N_j(t), t \geq 0\right\}$$

be a counting process that records the *frequency* of losses of the jth type occurring in $(0, t]$.

At the rth occurrence of an event of type e the Bernoulli variable $I_{j,r}^{(e)}$ indicates whether a loss of type j occurs. The vectors

$$\mathbf{I}_r^{(e)} = \left(I_{1,r}^{(e)}, \ldots, I_{n,r}^{(e)}\right)'$$

for $r = 1, \ldots, N^{(e)}(t)$ are considered to be independent and identically distributed with a multivariate Bernoulli distribution. In other words, each new event represents a new independent opportunity to incur a loss but, for a fixed event, the loss trigger variables for losses of different types may be dependent. The form of the dependence depends on the specification of the multivariate Bernoulli distribution and independence is a special case. We use the following notation for p-dimensional marginal probabilities of this distribution (where the subscript r is dropped for simplicity).

$$P\left(I_{j_1}^{(e)} = i_{j_1}, \ldots, I_{j_p}^{(e)} = i_{j_p}\right) = p_{j_1, \ldots, j_p}^{(e)}\left(i_{j_1}, \ldots, i_{j_p}\right), i_{j_1}, \ldots, i_{j_p} \in \{0, 1\}.$$

We also write $p_j^{(e)}(1) = p_j^{(e)}$ for one-dimensional marginal probabilities, so that in the special case of conditional independence we have

$$p_{j_1, \ldots, j_p}^{(e)}(1, \ldots, 1) = \prod_{k=1}^{p} p_{j_k}^{(e)}$$

The counting processes for events and losses are thus linked by

$$N_j(t) = \sum_{e=1}^{m} \sum_{r=1}^{N^{(e)}(t)} I_{j,r}^{(e)}. \tag{1}$$

Under the Poisson assumption for the event processes and the Bernoulli assumption for the loss indicators, the loss processes $\{N_j(t), t \geq 0\}$ are clearly Poisson themselves, since they are obtained by superpositioning m independent (possibly thinned) Poisson processes generated by the m underlying event processes. $(N_1(t), \ldots, N_n(t))'$ can be thought of as having a *multivariate Poisson* distribution.

However the total number of losses $N(t) = \sum_{j=1}^{n} N_j(t)$ is in general not Poisson but rather *compound Poisson*. It is the sum of m independent compound Poisson distributed random variables as can be seen by writing

$$N(t) = \sum_{e=1}^{m} \sum_{r=1}^{N^{(e)}(t)} \sum_{j=1}^{n} I_{j,r}^{(e)}. \tag{2}$$

The compounding distribution of the eth compound Poisson process is the distribution of $\sum_{j=1}^{n} I_j^{(e)}$, which in general is a sum of dependent Bernoulli variables. We return to the compound Poisson nature of the process $\{N(t), t \geq 0\}$ after generalising it in the next section.

2.2. Adding Dependent Severities

We can easily add *severities* to our multivariate Poisson model. Suppose that when the rth event of type e occurs a potential loss of type j with severity $X_{j,r}^{(e)}$ can occur. Whether the loss occurs or not is of course determined by the value of the indicator $I_{j,r}^{(e)}$, which we assume is independent of $X_{j,r}^{(e)}$. The potential losses $\{X_{j,r}^{(e)}, r = 1, \ldots, N^{(e)}(t), e = 1, \ldots, m\}$ are considered to be iid with distribution F_j. Potential losses of different types caused by the same event may however be dependent. We consider that they have a joint distribution function F. That is, for a vector $\mathbf{X}_r^{(e)}$ of potential losses generated by the same event we assume

$$\mathbf{X}_r^{(e)} = \left(X_{1,r}^{(e)}, \ldots, X_{n,r}^{(e)}\right)' \sim F.$$

In a more general model it would be possible to make the multivariate distribution of losses caused by the same event depend on the nature of the underlying event e. However, in practice it may make sense to assume that there is a single underlying multivariate severity distribution which generates the severities for all event types. This reflects the fact that it is often standard practice in insurance to model losses of the same type type as having an identical claim size distribution, without necessarily differentiating carefully between the events that caused them.

The aggregate loss process for losses of type j is a compound Poisson process given by

$$Z_j(t) = \sum_{e=1}^{m} \sum_{r=1}^{N^{(e)}(t)} I_{j,r}^{(e)} X_{j,r}^{(e)}. \tag{3}$$

The aggregate loss caused by losses of all types can be written as

$$Z(t) = \sum_{e=1}^{m} \sum_{r=1}^{N^{(e)}(t)} \sum_{j=1}^{n} I_{j,r}^{(e)} X_{j,r}^{(e)} = \sum_{e=1}^{m} \sum_{r=1}^{N^{(e)}(t)} \mathbf{I}_r^{(e)\prime} \mathbf{X}_r^{(e)}, \tag{4}$$

and is again seen to be a sum of m independent compound Poisson distributed random variables, and therefore itself compound Poisson distributed. Clearly (2) is a special case of (4) and (1) is a special case of (3). Thus we can understand all of these processes by focusing on (4). The compound Poisson nature of $Z(t)$ can be clarified by rewriting this process as

$$Z(t) \overset{\mathrm{d}}{=} \sum_{s=1}^{S(t)} Y_s,$$

where $\{S(t), t \geq 0\}$ is a Poisson process with intensity $\lambda = \sum_{e=1}^{m} \lambda^{(e)}$, counting all shocks s generated by all event types, and where the random variables $Y_1, \ldots, Y_{S(t)}$ are iid and independent of $\{S(t),\ t \geq 0\}$. Y_1 has the stochastic representation

$$Y_1 \overset{\mathrm{d}}{=} \left(\sum_{e=1}^{m} 1_{\left(\sum_{j=1}^{e-1} \lambda^{(j)}/\lambda,\ \sum_{j=1}^{e} \lambda^{(j)}/\lambda\right)}(U) \mathbf{I}^{(e)} \right)' \mathbf{X},$$

where U, $\mathbf{I}^{(e)}$, $\mathbf{X}$ are independent, U is uniformly distributed on (0,1), $\mathbf{I}^{(e)}$ is a generic random vector of indicators for shocks of event type e, and $\mathbf{X}$ is a generic random vector of severities caused by the same shock. In words: a shock s is of event type e with probability $\lambda^{(e)}/\lambda$.

We consider two examples that fit into the framework of the model we have set up. The first one, an insurance application of the model, we continue to develop throughout the paper. The second one, a credit risk application, is presented separately in Section 5.

2.3. Insurance example: natural catastrophe modelling

Fix $n = 2$, $m = 3$. Let $N_1(t)$ and $N_2(t)$ count windstorm losses in France and Germany respectively. Suppose these are generated by three different kinds of windstorm that occur independently. $N^{(1)}(t)$ counts west European windstorms; these are likely to cause French losses but no German losses. $N^{(2)}(t)$ counts central European windstorms; these are likely to cause German losses but no French losses. $N^{(3)}(t)$ counts pan-European windstorms, which are likely to cause both French and German losses.

3. The Effect of Dependent Loss Frequencies

To begin with we look at the distribution of the random vector $(N_1(t),\ldots, N_n(t))'$, particularly with regard to its univariate and bivariate margins as well as the correlation structure. Part 2 of the following proposition is from Barlow and Proschan (1975), p. 137.

Proposition 1.

1. $\{(N_1(t),\ldots,N_n(t))', t \geq 0\}$ *is a multivariate Poisson process with*

$$E(N_j(t)) = t \sum\nolimits_{e=1}^{m} \lambda^{(e)} p_j^{(e)}. \tag{6}$$

2. *The two-dimensional marginals are given by*

$$P(N_j(t) = n_j, N_k(t) = n_k) = e^{-\lambda t (p_{j,k}(1,1) + p_{j,k}(1,0) + p_{j,k}(0,1))} \times$$
$$\sum_{i=0}^{\min\{n_j, n_k\}} \frac{\left(\lambda t p_{j,k}(1,1)\right)^i \left(\lambda t p_{j,k}(1,0)\right)^{n_j - i} \left(\lambda t p_{j,k}(0,1)\right)^{n_k - i}}{i!\left(n_j - i\right)!\left(n_k - i\right)!}, \tag{7}$$

where $\lambda = \sum\nolimits_{e=1}^{m} \lambda^{(e)}$ *and*

$$p_{j,k}(i_j, i_k) = \lambda^{-1} \sum_{e=1}^{m} \lambda^{(e)} p_{j,k}^{(e)}(i_j, i_k),\ i_j, i_k \in \{0,1\}.$$

3. *The covariance and correlation structure is given by*

$$\operatorname{cov}\left(N_j(t), N_k(t)\right) = t \sum_{e=1}^{m} \lambda^{(e)} p_{j,k}^{(e)}(1,1) \tag{8}$$

and

$$\rho\left(N_j(t), N_k(t)\right) = \frac{\sum_{e=1}^{m} \lambda^{(e)} p_{j,k}^{(e)}(1,1)}{\sqrt{\left(\sum_{e=1}^{m} \lambda^{(e)} p_j^{(e)}\right)\left(\sum_{e=1}^{m} \lambda^{(e)} p_k^{(e)}\right)}}.$$

Proof

1. obvious using thinning and superposition arguments for independent Poisson processes.
2. is found in Barlow and Proschan (1975), p. 137.
3. is a special case of Proposition 7 (part 2).

□

Clearly, from Proposition 1 part 3, a necessary condition for $N_j(t)$ and $N_k(t)$ to be independent is that $p_{j,k}^{(e)}(1,1) = 0$ for all e; i.e. it must be impossible for losses of types j and k to be caused by the same event. If for at least one event it is possible that both loss types occur, then we have positive correlation between loss numbers. However Proposition 1, part 2 allows us to make a stronger statement.

Corollary 2. *$N_j(t)$ and $N_k(t)$ are independent if and only if $p^{(e)}_{j,k}(1,1) = 0$ for all e.*

Note that if $p^{(e)}_{j,k}(1,1) = 0$ for all j,k with $j \neq k$, then

$$\begin{aligned} P\left(\mathbf{I}^{(e)}_r = \mathbf{0}\right) &= 1 - P\left(\cup_{j=1}^{n}\left\{I^{(e)}_{j,r} = 1\right\}\right) \\ &= 1 - \left(\sum_{j=1}^{n} p^{(e)}_j - \sum_{j<k} p^{(e)}_{j,k}(1,1) + \ldots + (-1)^{n-1} p^{(e)}_{1,\ldots,n}(1,\ldots,1)\right) \\ &= 1 - \sum_{j=1}^{n} p^{(e)}_j . \end{aligned}$$

Hence if $\sum_{j=1}^{n} p^{(e)}_j > 1$ for some e, then $p^{(e)}_{j,k}(1,1) > 0$ for some $j \neq k$, or equivalently:

Corollary 3. *If $\sum_{j=1}^{n} p^{(e)}_j > 1$ for some e, then $N_1(t), \ldots, N_n(t)$ are not independent.*

Thus if we begin by specifying univariate conditional loss probabilities $p^{(e)}_j$ it is not always true that a shock model can be constructed which gives independent loss frequencies.

We have already noted that the process of total loss numbers $N(t) = \sum_{j=1}^{n} N_j(t)$ is in general not Poisson (but rather a sum of independent compound Poissons). If there is positive correlation between components $N_j(t)$ then $\{N(t), t \geq 0\}$ itself cannot be a Poisson process since it is *overdispersed* with respect to Poisson. It can easily be calculated (see Proposition 9 later) that

$$\operatorname{var}(N(t)) = \sum_{j=1}^{n}\sum_{k=1}^{n} \operatorname{cov}\left(N_j(t), N_k(t)\right) > E(N(t)) \tag{9}$$

Suppose we define a new vector of independent Poisson distributed loss counters $\hat{N}_j(t)$ such that $\hat{N}_j(t) \overset{d}{=} N_j(t)$. Clearly $\hat{N}(t) = \sum_{j=1}^{n} \hat{N}_j(t)$ is Poisson distributed and

$$\operatorname{var}\left(\hat{N}(t)\right) = E\left(\hat{N}(t)\right) = E(N(t)).$$

The case where the components $N_j(t)$ are dependent is clearly more dangerous (in the sense of higher variance) than the case with independent components. Although the expected number of total losses is the same in both cases the variance is higher in the dependent case and, using (9) and (8), we can calculate the inflation of the variance that results from dependence.

3.1. Insurance example (continued)

Consider a 5 year period and suppose French losses occur on average 5 times per year and German losses on average 6 times per year; in other words we assume $\lambda_1 = 5$ and $\lambda_2 = 6$. We consider three models for the dependence between these loss frequencies.

- **Case 1: No common shocks.** If there are no common shocks, then $N(5) = N_1(5) + N_2(5)$ has a Poisson distribution with intensity $\lambda = \lambda_1 + \lambda_2 = 5 + 6 = 11$.

In reality we believe that there are common shocks, in our case particularly the pan-European windstorms. Suppose west, central and pan-European windstorms occur on average 4, 3 and 3 times per year respectively. In terms of event intensities we have

$$\lambda^{(1)} = 4, \lambda^{(2)} = 3 \text{ and } \lambda^{(3)} = 3.$$

In terms of the indicator probabilities we assume that empirical evidence and expert judgement has been used to estimate

$$p_1^{(1)} = 1/2,\ p_2^{(1)} = 1/4,\ p_1^{(2)} = 1/6,\ p_2^{(2)} = 5/6,\ p_1^{(3)} = 5/6 \text{ and } p_2^{(3)} = 5/6$$

which means that, although unlikely, west European windstorms can cause German losses and central European windstorms can cause French losses. Note that these choices provide an example where the assumption of no common shocks is not only unrealistic but also impossible. To see this consider Corollary 3 and note that $p_1^{(3)} + p_2^{(3)} > 1$.

To make sure that our estimates of event frequencies and indicator probabilities tally with our assessment of loss frequencies we must have that

$$\lambda_j = \lambda^{(1)} p_j^{(1)} + \lambda^{(2)} p_j^{(2)} + \lambda^{(3)} p_j^{(3)},\ j = 1, 2.$$

However the specification of the univariate indicator probabilities is insufficient to completely specify the model. We need to fix the dependence structure of the bivariate indicators $\left(I_1^{(e)}, I_2^{(e)}\right)'$ for $e = 1, 2, 3$. For simplicity we will consider two possibilities.

- **Case 2: Independent indicators.**

$$p_{1,2}^{(e)}(1,1) = p_1^{(e)} p_2^{(e)} \text{ for } e = 1, 2, 3.$$

- **Case 3: Positively dependent indicators.**

$$p_{1,2}^{(e)}(1,1) \geq p_1^{(e)} p_2^{(e)} \text{ for } e = 1,2,3.$$

To be specific in Case 3 we will consider $p_{1,2}^{(e)}(1,1) = \min\left(p_1^{(e)}, p_2^{(e)}\right)$, which is the strongest possible dependence between the indicators, sometimes known as comonotonicity; the random variables $X_1, \ldots, X_n$ are said to be comonotonic if there exist increasing functions $v_1, \ldots, v_n : \mathbb{R} \to \mathbb{R}$, and a random variable Z such that $(X_1, \ldots, X_n)' \stackrel{d}{=} (v_1(Z), \ldots, v_n(Z))'$. For more on comonotonicity see Wang and Dhaene (1998) and the references therein. See also Joe (1997) for some discussion of dependence bounds in multivariate Bernoulli models. In terms of interpretation in our application this means:

- if a west European windstorm causes a German loss, then with certainty it also causes a French loss;
- if a central European windstorm causes a French loss, then with certainty it also causes a German loss;
- if a pan-European windstorm causes one kind of loss, then with certainty it causes the other kind of loss.

For cases 1, 2 and 3 we get $\text{var}(N(5)) = 55, 85$ and 95 respectively. Of more interest than the variance as a measure of the riskiness of $N(5)$ are the tail probabilities $P(N(5) > k)$. In this example these probabilities can be calculated analytically using formula (7) for the bivariate frequency function. The left plot in Figure 1 shows exceedence probabilities $P(N(5) > k)$, for

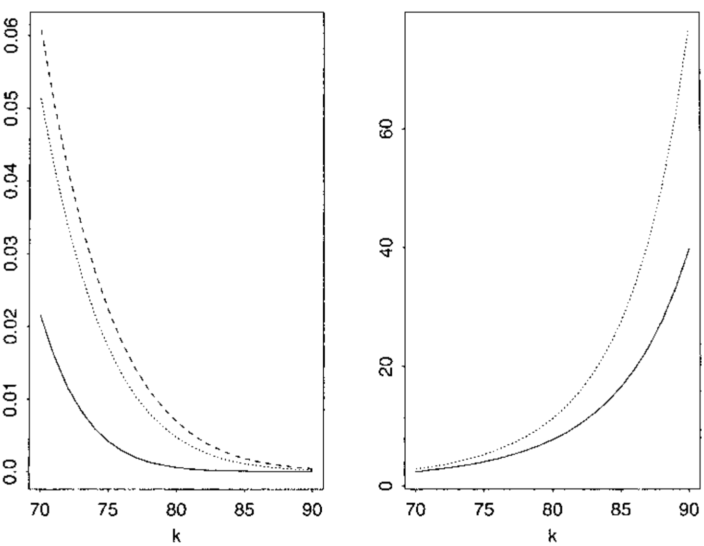


Figure 1: Left: Exceedence probabilities $P(N(5) > k)$ for $k = 70, 71, \ldots, 90$, for case 1 (lower), 2 (middle) and 3 (upper). Right: Ratios of such exceedence probabilities for cases 1-2 (lower) and 1-3 (upper).

$k = 70, 71, \ldots, 90$, for the three cases. The right plot shows by which factor such an exceedence probability is underestimated by case 1 if the correct model would be given by case 2 or 3. Clearly, both the presence of common shocks and then the subsequent addition of dependent indicators have a profound effect on the aggregate frequency distribution of $N(5)$.

3.2. The Equivalent Fatal Shock Model

The not-necessarily-fatal shock model set up in the previous section has the nice property of being easily interpreted. As we will now show this model has an equivalent representation as a fatal shock model. Basically, instead of counting all shocks, we only count loss-causing shocks. From this representation we can draw a number of non-trivial conclusions about our original model.

Let $\mathcal{S}$ be the set of non-empty subsets of $\{1, \ldots, n\}$. For $s \in \mathcal{S}$ we introduce a new counting process $\tilde{N}_s(t)$, which counts shocks in $(0, t]$ resulting in losses of all types in s *only*. Thus if $s = \{1, 2, 3\}$, then $\tilde{N}_s(t)$ counts shocks which cause simultaneous losses of types 1, 2 and 3, but not of types 4 to n. We have

$$\tilde{N}_s(t) = \sum_{e=1}^{m} \sum_{r=1}^{N^{(e)}(t)} \sum_{s':s' \supseteq s} (-1)^{|s'|-|s|} \prod\nolimits_{k \in s'} I_{k,r}^{(e)},$$

where $\sum_{s':s' \supseteq s} (-1)^{|s'|-|s|} \prod_{k \in s'} I_{k,r}^{(e)}$ is an indicator random variable which takes the value 1 if the rth shock of type e causes losses of all type in s only, and the value 0 otherwise. Furthermore let $\tilde{N}(t)$ count all shocks in $(0, t]$ which result in losses of any kind. Clearly we have

$$\tilde{N}(t) = \sum_{s \in \mathcal{S}} \tilde{N}_s(t).$$

The key to a fatal shock representation is the following result.

Proposition 4.

1. $\{\tilde{N}_s(t), t \geq 0\}$ *for* $s \in \mathcal{S}$ *are independent Poisson processes with intensities*

$$\lambda_s = \sum_{e=1}^{m} \lambda^{(e)} \sum_{s':s' \supseteq s} (-1)^{|s'|-|s|} p_{s'}^{(e)},$$

where $p_{s'}^{(e)} = P\left(\prod_{k \in s'} I_{k,r}^{(e)} = 1\right)$, *and*

2. $\{\tilde{N}(t), t \geq 0\}$ *is a Poisson process with intensity*

$$\tilde{\lambda} = \sum_{s \in \mathcal{S}} \lambda_s = \sum_{e=1}^{m} \lambda^{(e)} \left(1 - P\left(\mathbf{I}_r^{(e)} = \mathbf{0}\right)\right).$$

Proof. Let $J_{s,r}^{(e)} = \sum_{s':s'\supseteq s}(-1)^{|s'|-|s|}\prod_{k\in s'} I_{k,r}^{(e)}$. First note that the random variable $J_{s,r}^{(e)}$ takes values in $\{0,1\}$, and that $P\left(J_{s,r}^{(e)} = 1\right) = \sum_{s':s'\supseteq s}(-1)^{|s'|-|s|} p_{s'}^{(e)}$, where $p_{s'}^{(e)} = P\left(\prod_{k\in s'} I_{k,r}^{(e)} = 1\right)$, does not depend on r. Hence $\left\{\sum_{r=1}^{N^{(e)}(t)} J_{s,r}^{(e)}, t \geq 0\right\}$ is obtained by thinning the Poisson process $\left\{N^{(e)}(t), t \geq 0\right\}$, and is therefore a Poisson process with intensity $\lambda^{(e)}\sum_{s':s'\supseteq s}(-1)^{|s'|-|s|} p_{s'}^{(e)}$. $\left\{\tilde{N}_s(t), t\geq 0\right\}$ is obtained by superpositioning the independent Poisson processes $\left\{\sum_{r=1}^{N^{(e)}(t)} J_{s,r}^{(e)}, t\geq 0\right\}$ for $e = 1,\dots,m$, and is therefore a Poisson process with intensity $\lambda_s = \sum_{e=1}^{m}\lambda^{(e)}$ $\sum_{s':s'\supseteq s}(-1)^{|s'|-|s|}p_{s'}^{(e)}$. Since $P\left(\sum_{s\in\mathcal{S}}\sum_{s':s'\supseteq s}(-1)^{|s'|-|s|}\prod_{k\in s'} I_{k,r}^{(e)} = 1)\right)$ (the probability that the rth shock of type e causes at least one loss) does not depend on r, thinning and superpositioning arguments give that $\{\tilde{N}(t), t\geq 0\}$ is a Poisson process with intensity $\tilde{\lambda} = \sum_{s\in\mathcal{S}}\lambda_s = \sum_{e=1}^{m}\lambda^{(e)}\left(1 - P\left(\mathbf{I}_r^{(e)} = \mathbf{0}\right)\right)$. Each jump in the process $\left\{\tilde{N}(t), t\geq 0\right\}$ corresponds to a jump in exactly one of the processes $\left\{\tilde{N}_s(t), t\geq 0\right\}$ for $s\in\mathcal{S}$. Given a jump in $\left\{\tilde{N}(t), t\geq 0\right\}$, the probability of the jump being in $\left\{\tilde{N}_s(t), t\geq 0\right\}$ is given by $q_s = \lambda_s/\tilde{\lambda}$ for $s\in\mathcal{S}$. Order the $l = |\mathcal{S}| = 2^n - 1$ non-empty subsets of $\{1,\dots,n\}$ in some arbitrary way.

Then

$$P\left(\tilde{N}_{s_1}(t) = n_1,\dots,\tilde{N}_{s_l}(t) = n_l \mid \tilde{N}(t) = \tilde{n}\right) = \begin{cases} \tilde{n}!\prod_{j=1}^{l}\left(q_{s_j}^{n_j}\big/ n_j!\right), & \tilde{n} = \sum_{j=1}^{l} n_j, \\ 0 & ,\tilde{n} \neq \sum_{j=1}^{l} n_j. \end{cases}$$

and hence

$$\begin{aligned} P(\tilde{N}_{s_1}(t) = n_1,\dots,\tilde{N}_{s_l}(t) = n_l) &= P\left(\tilde{N}(t) = \sum_{j=1}^{l} n_j\right)\left(\sum_{j=1}^{l} n_j\right)!\prod_{j=1}^{l}\frac{q_{s_j}^{n_j}}{n_j!} \\ &= \prod_{j=1}^{l} e^{-\lambda_{s_j}t}\frac{\left(\lambda_{s_j}t\right)^{n_j}}{n_j!} = \prod_{j=1}^{l} P\left(\tilde{N}_{s_j}(t) = n_j\right). \end{aligned}$$

It follows that the processes $\left\{\tilde{N}_s(t), t\geq 0\right\}$ for $s\in\mathcal{S}$ are independent Poisson processes.

□

Since the Poisson processes $\left\{\tilde{N}_s(t), t\geq 0\right\}$ for $s\in\mathcal{S}$ are independent and since the loss counting processes may be written as

$$N_j(t) = \sum_{s:j\in s}\tilde{N}_s(t),$$

it also follows that we have obtained a fatal shock model representation for the original not-necessarily-fatal set-up.

Furthermore, since $\lambda_s = 0$ for all s with $|s| \geq 2$ if and only if $p_{j,k}^{(e)}(1,1) = 0$ for all e and all j,k with $j \neq k$, Corollary 2 concerning pairwise independence can be strengthened.

Corollary 5. $N_1(t), \ldots, N_n(t)$ are independent if and only if $p_{j,k}^{(e)}(1,1) = 0$ for all e and all j,k with $j \neq k$.

A direct consequence of the fatal shock model representation of the original not-necessarily-fatal shock model is that the multivariate distribution of the times to first losses can be easily analysed. Let $T_j = \inf\{t : N_j(t) > 0\}$ denote the time to the first loss of type j. We now consider briefly the distribution of $(T_1, \ldots, T_n)'$ whose dependence structure is well understood. For $s \in \mathcal{S}$ let $Z_s = \inf\{t : \tilde{N}_s(t) > 0\}$. $\{Z_s\}_{s \in \mathcal{S}}$ are independent exponential random variables with parameters $\{\lambda_s\}_{s \in \mathcal{S}}$. Hence

$$T_j = \inf\{t : N_j(t) > 0\} = \inf\left\{t : \sum_{s : j \in s} \tilde{N}_s(t) > 0\right\} = \min_{s : j \in s} Z_s$$

and $(T_1, \ldots, T_n) = \left(\min_{s : 1 \in s} Z_s, \ldots, \min_{s : n \in s} Z_s\right)$. Survival probabilities for $(T_1, \ldots, T_n)$ can be calculated as follows.

$$\begin{aligned}
&P(T_1 > t_1, \ldots, T_n > t_n) \\
&= P\left(\cap_i \tilde{N}_{\{i\}}(t_i) = 0, \cap_{i<j} \tilde{N}_{\{i,j\}}\left(\max(t_i, t_j)\right) = 0, \ldots, \tilde{N}_{\{1,\ldots,n\}}(\max(t_1, \ldots, t_n)) = 0\right) \\
&= \prod_i P\left(\tilde{N}_{\{i\}}(t_i) = 0\right) \prod_{i<j} P\left(\tilde{N}_{\{i,j\}}\left(\max(t_i, t_j)\right) = 0\right) \cdots P\left(\tilde{N}_{\{1,\ldots,n\}}(\max(t_1, \ldots, t_n)) = 0\right) \\
&= \exp\left(-\sum_i \lambda_{\{i\}} t_i - \sum_{i<j} \lambda_{\{i,j\}} \max(t_i, t_j) - \ldots - \lambda_{\{1,\ldots,n\}} \max(t_1, \ldots, t_n)\right) \qquad (10)
\end{aligned}$$

The multivariate exponential distribution with this joint survival probability is the multivariate exponential distribution of Marshall and Olkin (Marshall and Olkin (1967)). The distribution has been studied extensively, see Barlow and Proschan (1975), Joe (1997), Marshall and Olkin (1967) or Nelsen (1999). The multivariate exponential distribution of Marshall and Olkin has the property that

$$P(T_1 > t_1 + s_1, \ldots, T_n > t_n + s_n | T_1 > t_1, \ldots, T_n > t_n) = P(T_1 > s_1, \ldots, T_n > s_n),$$

for all $t_1, \ldots, t_n, s_1, \ldots, s_n > 0$. This is the multivariate version of the lack of memory property which is well known for the univariate exponential distribution. Not that this does not apply to general multivariate distributions with exponential marginals. The expression for the joint survival probability (10) might not be very convenient to work with if the model was set up as a not-necessarily-fatal shock model. However, it can easily be rewritten in a more convenient form.

$$P(T_1 > t_1, \ldots, T_n > t_n) = \exp\left(-\sum_{e=1}^{m} \lambda^{(e)} \left[\sum_i p_i^{(e)} t_i + \sum_{i<j} p_{i,j}^{(e)}(1,1)\min(t_i, t_j) + \ldots + p_{1,\ldots,n}^{(e)}(1,\ldots,1)\min(t_1,\ldots,t_n)\right]\right).$$

Recall that for a Poisson process with intensity μ, the time to the kth jump is $\Gamma(k, 1/\mu)$-distributed, where $\Gamma(\cdot,\cdot)$ denotes the Gamma distribution. Hence the time to the kth loss-causing shock is $\Gamma(k, 1/\tilde{\lambda})$-distributed, where $\tilde{\lambda} = \sum_{e=1}^{m} \lambda^{(e)}\left(1 - P\left(\mathbf{I}_r^{(e)} = \mathbf{0}\right)\right)$. The time to the kth loss is $\inf\{t : N(t) \geq k\}$, where

$$N(t) = \sum_{i=1}^{n} i \sum_{s:|s|=i} \tilde{N}_s(t).$$

$\{N(t), t \geq 0\}$ is in general not a Poisson process but rather a compound Poisson process, the time to the kth jump is still $\Gamma(k, 1/\tilde{\lambda})$-distributed but there are non unit jump sizes. By noting that the probability that the time to the kth loss is less than or equal to t can be expressed as $P(N(t) \geq k)$, it is clear that the distribution of the time to the kth loss can be fully understood from the distribution of $N(t)$ for $t \geq 0$, and this distribution can be evaluated using Panjer recursion or other methods.

3.3. Panjer Recursion

If there are common shocks, then $N(t) = \sum_{j=1}^{n} N_j(t)$ does not have a Poisson distribution. In our insurance example we have considered only two loss types and it is thus easy to calculate the distribution of $N(t)$ directly using convolution and the bivariate frequency function in (7). A more general method of calculating the probability distribution function of $N(t)$, which will also work in higher dimensional examples, is Panjer recursion (Panjer (1981)). We use the notation of the preceding section. In addition, let W_i denote the number of losses due to the ith loss-causing shock. The total number of losses, $N(t)$, has the stochastic representation

$$N(t) \stackrel{d}{=} \sum_{k=1}^{\tilde{N}(t)} W_k,$$

where $W_1, \ldots, W_{\tilde{N}(t)} (\stackrel{d}{=} W)$ are iid and independent of $\tilde{N}(t)$. The probability $P(N(t) = r)$ can now easily be calculated using Panjer recursion.

Proposition 6.

$$P(N(t)=r)=\begin{cases}\sum_{k=1}^{\min(r,n)} \frac{\tilde{\lambda} tk}{r} P(W=k)P(N(t)=r-k), & r\geq 1,\\ \exp(-\tilde{\lambda} t), & r=0,\end{cases} \tag{11}$$

where

$$P(W=k)=\begin{cases}\tilde{\lambda}^{-1}\sum_{e=1}^{m}\lambda^{(e)}\left(\sum_{s:|s|=k} p_s^{(e)}+\sum_{i=1}^{n-k}(-1)^i \frac{\binom{n}{k}\binom{n-k}{i}}{\binom{n}{k+i}}\sum_{s:|s|=k+i} p_s^{(e)}\right), & k<n,\\ \tilde{\lambda}^{-1}\sum_{e=1}^{m}\lambda^{(e)} p_{\{1,\dots,n\}}^{(e)}, & k=n,\end{cases}$$

Proof. The formula (11) follows from Theorem 4.4.2, p. 119 in Rolski, Schmidli, Schmidt and Teugels (1998), and that the maximum number of losses due to a loss-causing shock is n. The probability that a loss-causing shock causes exactly k losses is given by $P(W=k)=\tilde{\lambda}^{-1}\sum_{s:|s|=k}\lambda_s$, where

$$\sum_{s:|s|=k}\lambda_s=\sum_{e=1}^{m}\lambda^{(e)}\sum_{s:|s|=k}\sum_{s':s'\supseteq s}(-1)^{|s'|-|s|}p_{s'}^{(e)}.$$

The expression of the probability that a loss-causing shock causes n losses can be simplified to

$$P(W=n)=\tilde{\lambda}^{-1}\sum_{e=1}^{m}\lambda^{(e)}p_{\{1,\dots,n\}}^{(e)}=\tilde{\lambda}^{-1}\sum_{e=1}^{m}\lambda^{(e)}p_{1,\dots,n}^{(e)}(1,\dots,1).$$

For $k\leq n$ we note that there are $\binom{n}{k}$ sets s with $|s|=k$, and for each such s there are $\binom{n-k}{i}$ sets of size $k+i$ $(i\in\{1,\dots,n-k\})$ which contain s as a proper subset. Hence

$$\sum_{s:|s|=k}\sum_{s':s'\supseteq s}(-1)^{|s'|-|s|}p_{s'}^{(e)} \tag{12}$$

consists of $\binom{n}{k}\binom{n-k}{i}$ terms $(-1)^{|s'|-|s|}p_{s'}^{(e)}$ for which $|s'|=k+i$ and $s\subset s'$. Since there are $\binom{n}{k+i}$ sets s' with $|s'|=k+i$ it follows that (12) is equal to

$$\sum_{s:|s|=k}p_s^{(e)}+\sum_{i=1}^{n-k}(-1)^i\frac{\binom{n}{k}\binom{n-k}{i}}{\binom{n}{k+i}}\sum_{s:|s|=k+i}p_s^{(e)}.$$

□

For large n, say $n > 100$, the usefulness of the Panjer recursion scheme relies heavily on the calculation of $\sum_{s:|s|=k} p_s^{(e)}$ for $k \in \{1, \ldots, n\}$. We now look at two specific assumptions on the multivariate Bernoulli distribution of $\mathbf{I}^{(e)}$ conditional on a shock of type e. The assumption of conditional independence is attractive for computations since in this case

$$\sum_{s:|s|=k} p_s^{(e)} = \sum_{j_1=1}^{n} \sum_{j_2>j_1} \cdots \sum_{j_k>j_{k-1}} p_{j_1}^{(e)} p_{j_2}^{(e)} \ldots p_{j_k}^{(e)}.$$

Under the assumption of conditional comonotonicity

$$\sum_{s:|s|=k} p_s^{(e)} = \sum_{j_1=1}^{n} \sum_{j_2>j_1} \cdots \sum_{j_k>j_{k-1}} \min\left(p_{j_1}^{(e)}, p_{j_2}^{(e)}, \ldots, p_{j_k}^{(e)}\right).$$

The latter assumption leads to very efficient computations of $\sum_{s:|s|=k} p_s^{(e)}$. Let

$$\left(p_{\pi_1}^{(e)}, p_{\pi_2}^{(e)}, \ldots, p_{\pi_n}^{(e)}\right)'$$

denote the sorted vector of univariate conditional indicator probabilities, such that $p_{\pi_1}^{(e)} \leq p_{\pi_2}^{(e)} \leq \ldots \leq p_{\pi_n}^{(e)}$. Then

$$\sum_{s:|s|=k} p_s^{(e)} = \sum_{i=1}^{n} \binom{n-i}{k-1} p_{\pi_i}^{(e)},$$

where $\binom{n-i}{k-1}$ is the number of subsets of size k of $\{1, \ldots, n\}$ with i as smallest element.

4. The Effect of Dependent Severities

We now consider adding severities to our shock model and study the multivariate distribution of $(Z_1(t), \ldots, Z_n(t))'$. Again we can calculate first and second moments of the marginal distributions and correlations between the components.

Proposition 7.

1. $\{(Z_1(t), \ldots, Z_n(t))', t \geq 0\}$ *is a multivariate compound Poisson process. If* $E(|X_j|) < \infty$, *then*

$$E\left(Z_j(t)\right) = E\left(X_j\right) E\left(N_j(t)\right).$$

2. *If* $E\left(X_j^2\right), E\left(X_k^2\right) < \infty$, *then the covariance and correlation structure is given by*

$$\operatorname{cov}(Z_j(t), Z_k(t)) = E(X_j X_k) \operatorname{cov}(N_j(t), N_k(t))$$

and

$$\rho\left(Z_j(t), Z_k(t)\right) = \frac{E\left(X_j X_k\right)}{\sqrt{E\left(X_j^2\right) E\left(X_k^2\right)}} \rho\left(N_j(t), N_k(t)\right).$$

Proof

1. is easily established from formula (3).
2. We observe that $\forall j, k \in \{1, \dots, n\}$,

$$\begin{aligned}
\operatorname{cov}\left(Z_j(t), Z_k(t)\right) &= \sum_{e=1}^{m} \operatorname{cov}\left(\sum_{r=1}^{N^{(e)}(t)} I_{j,r}^{(e)} X_{j,r}^{(e)}, \sum_{r=1}^{N^{(e)}(t)} I_{k,r}^{(e)} X_{k,r}^{(e)}\right) \\
&\quad + \sum_{e=1}^{m} \sum_{f \neq e} \operatorname{cov}\left(\sum_{r=1}^{N^{(e)}(t)} I_{j,r}^{(e)} X_{j,r}^{(e)}, \sum_{r=1}^{N^{(f)}(t)} I_{k,r}^{(f)} X_{k,r}^{(f)}\right) \\
&= \sum_{e=1}^{m} \operatorname{cov}\left(\sum_{r=1}^{N^{(e)}(t)} I_{j,r}^{(e)} X_{j,r}^{(e)}, \sum_{r=1}^{N^{(e)}(t)} I_{k,r}^{(e)} X_{k,r}^{(e)}\right) \\
&= E\left(X_j X_k\right) \sum_{e=1}^{m} E\left(N^{(e)}(t)\right) E\left(I_j^{(e)} I_k^{(e)}\right) \\
&= E\left(X_j X_k\right) t \sum_{e=1}^{m} \lambda^{(e)} p_{j,k}^{(e)}(1,1) \\
&= E\left(X_j X_k\right) \operatorname{cov}\left(N_j(t), N_k(t)\right)
\end{aligned}$$

□

Now consider the distribution of the total loss $Z(t) = \sum_{j=1}^{n} Z_j(t)$. The expected total loss is easily calculated to be

$$E(Z(t)) = \sum_{j=1}^{n} E\left(X_j\right) E\left(N_j(t)\right),$$

and higher moments of $Z(t)$ can also be calculated, by exploiting the compound Poisson nature of this process as shown in (5). Since $\{Z(t), Z \geq 0\}$ is the most general aggregate loss process that we study in this paper we collect some useful moment results for this process.

Lemma 8. *The pth derivative of the characteristic function $\varphi_{Z(t)}$ of $Z(t)$ satisfies*

$$\varphi_{Z(t)}^{(p)}(x) = \lambda t \sum_{k=0}^{p-1} \binom{p-1}{k} \varphi_Y^{(k+1)}(x) \varphi_{Z(t)}^{(p-k-1)}(x), \qquad (13)$$

where φ_Y denotes the characteristic function of Y.

Proof. The formula (13) is proved by induction using the binomial relation $\binom{p-1}{k} + \binom{p-1}{k+1} = \binom{p}{k+1}$. The induction step reads as follows. Fix p and suppose that (13) holds. Then

$$\begin{aligned}
\varphi_{Z(t)}^{(p+1)}(x) &= \lambda t \sum_{k=0}^{p-1} \binom{p-1}{k} \left(\varphi_Y^{(k+2)}(x) \varphi_{Z(t)}^{(p-k-1)}(x) + \varphi_Y^{(k+1)}(x) \varphi_{Z(t)}^{(p-k)}(x) \right) \\
&= \lambda t \sum_{k=0}^{p-2} \binom{p-1}{k} \varphi_Y^{(k+2)}(x) \varphi_{Z(t)}^{(p-k-1)}(x) + \lambda t \varphi_Y^{(p+1)}(x) \varphi_{Z(t)}^{(0)}(x) \\
&\quad + \lambda t \varphi_Y^{(1)}(x) \varphi_{Z(t)}^{(p)}(x) + \lambda t \sum_{k=0}^{p-2} \binom{p-1}{k+1} \varphi_Y^{(k+2)}(x) \varphi_{Z(t)}^{(p-k-1)}(x) \\
&= \lambda t \varphi_Y^{(1)}(x) \varphi_{Z(t)}^{(p)}(x) + \lambda t \sum_{k=1}^{p-1} \binom{p}{k} \varphi_Y^{(k+1)}(x) \varphi_{Z(t)}^{(p-k)}(x) \\
&\quad + \lambda t \varphi_Y^{(p+1)}(x) \varphi_{Z(t)}^{(0)}(x) \\
&= \lambda t \sum_{k=0}^{p} \binom{p}{k} \varphi_Y^{(k+1)}(x) \varphi_{Z(t)}^{(p-k)}(x).
\end{aligned}$$

□

Using Lemma 8, the moments of $Z(t)$ can be calculated as follows.

Proposition 9.

1. *If they exist, the 2nd and 3rd order central moment of $Z(t)$ are given by*

$$E\left((Z(t) - E(Z(t)))^p \right) = \lambda t E(Y^p), \quad p = 2, 3, \qquad (14)$$

where $\lambda = \sum_{e=1}^{m} \lambda^{(e)}$ and

$$E(Y^p) = \frac{1}{\lambda} \sum_{j_1=1}^{n} \cdots \sum_{j_p=1}^{n} E(X_{j_1} \cdots X_{j_p}) \sum_{e=1}^{m} \lambda^{(e)} p_{j_1,\ldots,j_p}^{(e)}(1,\ldots,1). \qquad (15)$$

2. *Whenever they exist, the non-central moments of $Z(t)$ are given recursively by*

$$E\left(Z(t)^p \right) = \lambda t \sum_{k=0}^{p-1} \binom{p-1}{k} E\left(Y^{k+1} \right) E\left(Z(t)^{p-k-1} \right),$$

with $E(Y^{k+1})$ given by (15).

Proof

1. For a compound Poisson process of the form (5) the formula (14) is well known. We can calculate that for all p

$$\begin{aligned} E(Y^p) &= \sum_{e=1}^{m} \frac{\lambda^{(e)}}{\lambda} E\left((\mathbf{I}^{(e)\prime}\mathbf{X})^p\right) \\ &= \sum_{e=1}^{m} \frac{\lambda^{(e)}}{\lambda} E\left(\sum_{j_1=1}^{n} \cdots \sum_{j_p=1}^{n} I_{j_1}^{(e)} \ldots I_{j_p}^{(e)} X_{j_1} \ldots X_{j_p} \right) \\ &= \lambda^{-1} \sum_{e=1}^{m} \lambda^{(e)} \sum_{j_1=1}^{n} \cdots \sum_{j_p=1}^{n} E\left(X_{j_1} \ldots X_{j_p}\right) p_{j_1,\ldots,j_p}^{(e)}(1,\ldots,1) \\ &= \lambda^{-1} \sum_{j_1=1}^{n} \ldots \sum_{j_p=1}^{n} E\left(X_{j_1} \ldots X_{j_p}\right) \sum_{e=1}^{m} \lambda^{(e)} p_{j_1,\ldots,j_p}^{(e)}(1,\ldots,1). \end{aligned}$$

2. If $E(|Z(t)|^p) < \infty$, then $\varphi_{Z(t)}^{(k)}(0) = i^k E(Z(t)^k)$ for $k = 1, \ldots, p$. The conclusion follows by applying this to (13).

□

We are particularly interested in the effect of different levels of dependence between both loss frequencies and loss severities on the tail of the distribution of $Z(t)$, and on higher quantiles of this distribution. The distribution of $Z(t)$ is generally not available analytically but, given the ease of simulating from our Poisson common shock model, it is possible to estimate quantiles empirically to a high enough degree of accuracy that differences between different dependence specifications become apparent.

It is also possible, given the ease of calculating moments of $Z(t)$, to use a moment fitting approach to approximate the distribution of $Z(t)$ with various parametric distributions, and we implement this approach in the following example.

4.1. Insurance example (continued)

Assume that French and German severities are Pareto(4,3) distributed, i.e.

$$F_i(x) = P(X_i \le x) = 1 - \left(\frac{3}{3+x}\right)^4, \; E(X_i) = 1, \; E\left(X_i^2\right) = 3, \; E\left(X_i^3\right) = 27, \quad i = 1, 2.$$

We have to fix the dependence structure of potential losses $(X_1, X_2)'$ at the same shock. We do this using the copula approach. The copula C of $(X_1, X_2)'$ is the distribution function of $\left(F_1(X_1), F_2(X_2)\right)'$. The distribution function of $(X_1, X_2)'$ can be expressed in terms of C as

$$F(x_1, x_2) = C\left(F_1(x_1), F_2(x_2)\right).$$

For more on copulas see Embrechts, McNeil and Straumann (2001), Nelsen (1999) or Joe (1997). We consider three cases.

- Independent severities:

$$F(x_1, x_2) = F_1(x_1)\,F_2(x_2).$$

- Positively dependent severities with Gaussian dependence:

$$F(x_1, x_2) = C_\rho^{\mathrm{Ga}}(F_1(x_1), F_2(x_2)),$$

  where

$$C_\rho^{\mathrm{Ga}}(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{\frac{-(s^2 - 2\rho st + t^2)}{2(1-\rho^2)}\right\} dsdt.$$

  and $\rho \in (0, 1)$.

- Positively dependent severities with Gumbel dependence:

$$F(x_1, x_2) = C_\theta^{\mathrm{Gu}}(F_1(x_1), F_2(x_2)),$$

  where

$$C_\theta^{\mathrm{Gu}}(u, v) = \exp\left(-\left\{(-\log u)^\theta + (-\log v)^\theta\right\}^{1/\theta}\right),$$

  and $\theta > 1$.

For both of the positive dependence models we will parameterize the copulas such that Kendall's rank correlation (τ) (see e.g. Embrechts, McNeil and Straumann (2001) for details) between X_1 and X_2 is 0.5. This is achieved by setting

$$\rho = \sin\left(\frac{\pi}{2}\tau\right) \text{ and } \theta = \frac{1}{1-\tau}.$$

As we have discussed there are several possibilities for modelling the tail of $Z(5)$. One approach is to fit a heavy-tailed generalised F-distribution (referred to as a generalised Pareto distribution in Hogg and Klugman (1994)) to $Z(5)$ using moment fitting with the first three moments. The distribution function is given by

$$H_{\alpha,\lambda,k}(x) = G\left(2k, 2\alpha, \frac{\alpha}{k\lambda}x\right) \text{ for } \alpha > 0, \lambda > 0, \mathrm{k} > 0,$$

where $G(\nu_1, \nu_2, \cdot)$ is the distribution function for the F-distribution with ν_1 and ν_2 degrees of freedom. The nth moment exists if $\alpha > n$ and is then given by

$$\lambda^n \left(\prod_{i-0}^{n-1}(k+i)\right) \Big/ \left(\prod_{i=1}^{n}(\alpha - i)\right).$$

By calculating the first three moments of $Z(5)$ for different frequency and severity dependencies we fit generalised F-distributions and study the difference in tail behaviour. Figure 2 shows quantiles of generalised F-distributions determined by fitting the first three moments to $Z(5)$ for case 1, 2 and 3 and for different dependence structures between the severities. It clearly shows the effect of common shocks on the tail of $Z(5)$ and perhaps even more the drastic change in tail behaviour when adding moderate dependence between the severities.

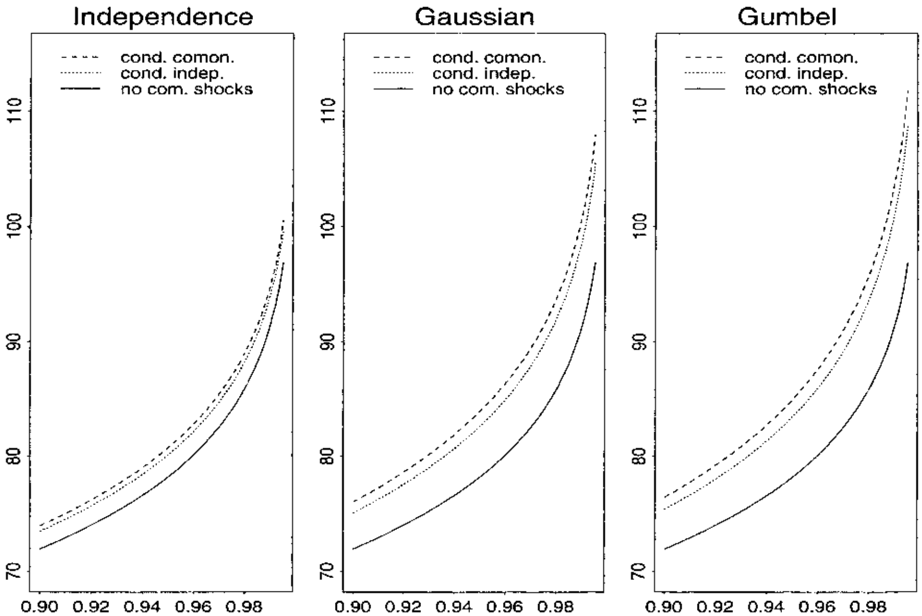


Figure 2: The curves from lower to upper show the quantiles $H^{-1}_{\alpha,\lambda,k}(q)$ of moment fitted generalised F-distributions for case 1, 2 and 3 and $q \in [0.900, 0.995]$. The first three moments coincide with those of $Z(5)$.

It should be noted that the quantile estimates of $Z(5)$ given by moment fitted generalised F-distributions are slight overestimates of the true quantiles for $\alpha \in [0.900, 0.995]$. However the accuracy is sufficient to show the major differences between the quantile curves of $Z(5)$ for our different copula choices.

5. Applying the Methodology to Portfolio Credit Risk

We consider here the problem of quantifying risk in large portfolios of defaultable assets, the simplest example being loan portfolios. Models that are used for this purpose address the phenomenon of *dependent defaults* and a number of approaches have been suggested in the literature in recent years and implemented in widely-used industry solutions such as the model proposed by the KMV corporation (KMV-Corporation 1997), the model of the RiskMetrics group (RiskMetrics-Group 1997), or CreditRisk$^+$, developed by Credit Suisse Financial Products (Credit-Suisse-Financial-Products 1997). In this section we describe how the common Poisson shock model provides a simple alternative framework for modelling portfolio credit risk.

We emphasise that our focus here, and that of the industry models, is on *overall portfolio credit risk*. We are less concerned with providing detailed analysis of the individual default potential of a single credit risk, such as is required in the pricing of defaultable bonds or standard credit derivatives. For this purpose the class of *default intensity models* has emerged as the most important; see for example Jarrow and Turnbull (1995).

5.1. Poisson shock models for defaults

We consider a portfolio of loans and develop a shock model by considering that every counterparty in the portfolio defines a loss type and that a variety of different kinds of *economic shock* may lead to the default of these counterparties: *global* shock events may potentially affect all counterparties; *sector* shock events may affect only certain kinds of company, such as companies in a particular geographical area or companies concentrated on a particular industry; *idiosyncratic* shock events (such as an episode of bad management) may affect only individual counterparties; one might also think of *endogenous* shock events where the default of important primary counterparties might affect other counterparties, so that default was *contagious*.

In all cases the common shock construction means that defaults of individual counterparties are modelled by the first events in a series of dependent Poisson process. Suppose the random vector $\mathbf{T} = T_1, \ldots, T_n)'$ describes the times to default for the n counterparties in the portfolio. In Section 3.2 we observed that this vector of default times has a multivariate exponential distribution with the Marshall-Olkin survival copula. Suppose that time is measured in years and that we are interested in the portfolio credit loss distribution for a time horizon of 1 year; suppose further that the exposures (loan sizes) are known and given by $e_1, \ldots, e_n$. If we neglect interest rates and assume that nothing is recovered from defaulted firms then the overall portfolio loss is given by

$$L = \sum_{i=1}^{n} e_i 1_{\{T_i \leq 1\}}.$$

Frey and McNeil (2001) have shown that the distribution of L is fully determined by the set of individual default probabilities $\{p_i = P(T_i \leq 1), i = 1, \ldots, n\}$ and the copula C of the vector $\mathbf{T}$; the exponential distributional form for the margins of $\mathbf{T}$ is not a critical feature of the model. They also show that the tail of the distribution of L and related risk measures are often much more sensitive to the assumptions about the dependence between the defaults as summarised by the copula, than they are to the accurate specification of individual default probabilities.

5.2. Relation of the shock model to standard models

All other models that have been suggested for portfolio credit risk also imply multivariate distributions for the vector of default times of the portfolio members. Regardless of whether these models are set up as one-period or multi-period models, or whether they assume constant intensities of default (as in the Poisson shock model), non-homogeneous intensities or stochastic intensities, they all imply a distribution for the counterparty survival times. From model to model, these distributions will vary with respect to both their marginal distributions and their copulas, but as far as determining the loss distribution is concerned, if they have been calibrated to give broadly similar individual default probabilities it is the copula that will be decisive in determining the tail of the portfolio loss distribution.

Both KMV and CreditMetrics may be considered to descend from the firm-value model of Merton (1974), where default is modelled as occurring when the asset value of a company falls below its liabilities, and asset value changes are considered to have a multivariate normal distribution. Although this appears very different to the Poisson shock model, as far as the loss distribution for a fixed time horizon is concerned, both KMV and CreditMetrics are in fact structurally equivalent to a model in which default times have a multivariate exponential distribution with the Gaussian copula (i.e. the copula that describes the dependence inherent in a multivariate normal distribution); see Li (1999) and Frey and McNeil (2001) for more detail. Thus the crucial difference in the one-period framework lies in the fact that these industry models imply a Gaussian copula whereas common shocks imply a Marshall-Olkin copula to describe the dependence of the survival times.

The CreditRisk+ model employs a mixture-modelling philosophy which assumes that *conditional* on a vector of independent gamma-distributed macroeconomic factors, the default of a counterparty occurs independently of other counterparties and is the first event in a Poisson process with an intensity that depends on these factors. Survival times are not exponential (but rather conditionally exponential), but for portfolio risk modelling in a one-period setting, this is again not a decisive factor. Assuming that the model has been calibrated to give plausible values for individual default probabilities, it is the copula of the default times that is most important in determining the overall loss distribution, although this copula is difficult to isolate in closed form in the general version of CreditRisk+.

There have been a number of papers on the subject of extending the intensity-based approach to modelling the default of single counterparties to obtain models for dependent defaults of several counterparties, principally with the problem of pricing so-called basket credit derivatives in mind; see Schoenbucher and Schubert (2001) for a useful summary of these approaches.

We mention in particular a model of Duffie and Singleton (1995) where individual defaults follow Cox processes with stochastic intensities. These intensities may jump (by a random amount) when certain common shocks occur which have the potential to affect all counterparties, or when idiosyncratic shocks occur affecting an individual company; as in our model shocks occur as Poisson processes. Our model can be thought of a cruder version of the Duffie & Singleton model with constant and deterministic intensities for individual defaults.

5.3. Setting up the shock model

Consider a loan portfolio consisting of n obligors. Suppose the counterparties can be divided into K geographical or industry sectors. We consider a model where obligors are subject to idiosyncratic, sector and global shocks, so that there are a total of $m = n + K + 1$ shock event processes.

Suppose that the jth obligor belongs to sector $k = k_j$ where $k \in \{1,\ldots,K\}$. From formula (6) we know that $N_j(t)$, the number of defaults of obligor j in $(0,t]$ is Poisson with intensity given by

$$\lambda_j = \lambda^{(j)} + p_j^{(n+k)} \lambda^{(n+k)} + p_j^{(m)} \lambda^{(m)}, \quad k = k(j),$$

where the three terms represent the contributions to the default intensity of idiosyncratic, sector and global events respectively. Note that in general this intensity will be set so low that the probability of a firm defaulting more than once in the period of interest can be considered negligible.

This is a very general model and to obtain a model that we would have a hope of calibrating in a real application we need to drastically reduce the number of parameters in the model. We assume first that companies can be grouped together into *rating classes* within which default rates can be considered constant and known. It is very common in portfolio default risk modelling to base the assessment of default intensities for individual companies on information about historical default rates for similarly rated companies. Suppose that the jth obligor belongs to rating category $l = l(j)$ where $l \in \{1,\ldots,L\}$. We assume for the overall default intensity λ_j that

$$\lambda_j = \lambda_{\text{total},\,l}, \quad l = l(j), j = 1,\ldots,n. \tag{16}$$

To achieve (16) we assume that the rate of occurrence of idiosyncratic shocks also depends only on the rating category and we adopt the notation

$$\begin{aligned} \lambda^{(j)} &= \lambda_{\text{idio},\,l}, \quad l = l(j), j = 1,\ldots,n, \\ \lambda^{(n+k)} &= \lambda_{\text{sector},\,k}, \quad k = 1,\ldots,K, \\ \lambda^{(m)} &= \lambda_{\text{global}}. \end{aligned}$$

Clearly we now have a total of $L + K + 1$ shock intensities to set.

We assume also that the conditional default probabilities given the occurrence of sector shocks only depend on the rating class of the company and write for an obligor j

$$p_j^{(n+k)} = s_{k,l}, \quad l = l(j).$$

We assume moreover that the default indicators for several companies in the same sector are conditionally independent given the occurrence of an event in that sector. Analogously, we assume that the conditional default probabilities given the occurrence of global shocks depend on both rating class and sector of the company and write

$$p_j^{(m)} = g_{k,l}, \quad l = l(j),\ k = k(j).$$

We assume that the default indicators for any group of companies are conditionally independent given the occurrence of a global event.

In total we have $2KL$ conditional default probabilities to set and we have the system of equations

$$\lambda_{\text{total},\,l} = \lambda_{\text{idio},\,l} + s_{k,l}\lambda_{\text{sector},\,k} + g_{k,l}\lambda_{\text{global}},\ k = 1,\ldots,K, l = 1,\ldots,L,$$

subject to the constraint, imposed by (16), that

$$s_{k,l}\lambda_{\text{sector},\,k} + g_{k,l}\lambda_{\text{global}} = s_{k',l}\lambda_{\text{sector},\,k'} + g_{k',l}\lambda_{\text{global}},\ \forall k \neq k'. \tag{16}$$

5.4. Understanding the factors determining the risk

We are interested in the behaviour of $N(t)$, the total number of defaults, for fixed t. If we suppose that the individual default rates have been fixed then $E(N(t))$ has been fixed. However, depending how we set the various shock intensities and individual default probabilities the risk inherent in $N(t)$ may vary considerably. If we measure risk by variance we can get analytically an idea of which factors affect the risk by considering

$$\operatorname{var}(N(t)) - E(N(t)) = \sum_{j_1, j_2 : j_1 \neq j_2} \operatorname{cov}\left(N_{j_1}(t), N_{j_2}(t)\right).$$

For simplicity we set $t = 1$ and consider a model with one rating class ($L = 1$) and assume that the conditional default probabilities do not depend on the sector for all global shocks ($g_{k,1} = g, \forall k$). Let there be n_k obligors in sector k. We have for $j_1 \neq j_2$ that

$$\operatorname{cov}\left(N_{j_1}(1), N_{j_2}(1)\right) = \begin{cases} g^2\lambda_{\text{global}} & k(j_1) \neq k(j_2), \\ g^2\lambda_{\text{global}} + s_k^2\lambda_{\text{sector},\,k} & k(j_1) = k(j_2) = k, \end{cases}$$

which allows us to calculate that

$$\mathrm{var}(N(1)) - E(N(1)) = (n^2 - n)g^2 \lambda_{\text{global}} + \sum_{k=1}^{K} \left(n_k^2 - n_k\right) s_k^2 \lambda_{\text{sector},k}.$$

In view of (17) we have that $s_k \lambda_{\text{sector},k}$, the intensity of default due to sector shocks in sector k, must be equal for all k in this special case. If we write $\delta_{\text{global}} = g\lambda_{\text{global}}$ and $\delta_{\text{sector}} = s_k \lambda_{\text{sector},k}$, $\forall k$, for the default intensities due to global or sector causes we obtain finally

$$\mathrm{var}(N(1)) - E(N(1)) = \delta_{\text{global}} \left(n^2 - n\right) g + \delta_{\text{sector}} \sum_{k=1}^{K} (n_k^2 - n_k) s_k.$$

This expression allows us to draw two broad conclusions about the riskiness of the model as measured by the variance of the number of defaults.

- The higher the portion of the default intensity that we attribute to sector and global shocks (common shocks) the riskier the model. As before, the only way we can have a Poisson distribution for $N(1)$ is if the default intensity can be attributed entirely to idiosyncratic shocks.
- Suppose we assume common shocks and fix the portions of the default intensity that we attribute to common shocks (δ_{global} and δ_{sector}). The overall risk also depends on how we set the conditional default probabilities g and s_k. Low shock intensities and high conditional default probabilities are riskier than the other way around.

These conclusions are confirmed in the following simulation example where we allow two rating categories and more heterogeneous conditional default probabilities.

5.5. A simulation study

In our examples we take $t = 1$ year and consider $K = 4$ sectors and $L = 2$ rating categories; we assume that overall default rates for these categories are $\lambda_{\text{total},1} = 0.005$ and $\lambda_{\text{total},2} = 0.02$. Let $n_{k,l}$ denote the number of companies in rating class l and sector k. We set

$$n_{1,1} = 10000, n_{2,1} = 20000, n_{3,1} = 15000, n_{4,1} = 5000,$$
$$n_{1,2} = 10000, n_{2,2} = 25000, n_{3,2} = 10000, n_{4,2} = 5000.$$

In the following two cases we investigate the sensitivity of the tail of $N(1)$ to the specification of model parameters. Results are based on 10000 simulated realizations of $N(1)$ and models are compared with respect to estimates of the 95% and 99% quantiles.

- **Case 1**
  We study the effects of increasing the intensity of the common shocks and decreasing the intensity of the idiosyncratic shocks when the univariate

conditional default probabilities are held constant. We set the values of these parameters to be

$$(s_{1,1}, s_{2,1}, s_{3,1}, s_{4,1}, s_{1,2}, s_{2,2}, s_{3,1}, s_{4,2})$$
$$= (0.25, 0.08, 0.05, 0.1, 1, 0.3, 0.25, 0.25) 10^{-2}$$
$$(g_{1,1}, g_{2,1}, g_{3,1}, g_{4,1}, g_{1,2}, g_{2,2}, g_{3,1}, g_{4,2})$$
$$= (0.25, 0.1, 0.4, 0.1, 1, 0.5, 1.5, 1) 10^{-2}.$$

We have some flexibility in choosing the intensities

$$(0.005, 0.02, 0.0, 0.0, 0.0, 0.0, 0.0) \to (0.004, 0.016, 0.2, 1.0, 0.4, 0.8, 0.2) \to$$
$$(0.002, 0.008, 0.6, 3.0, 1.2, 2.4, 0.6) \to (0.0, 0.0, 1.0, 5.0, 2.0, 4.0, 1.0)).$$

Hence we start with the special case of no common shocks and a situation where every individual default process $N_j(t)$ is independent Poisson and the total number of defaults $N(t)$ is Poisson. In the second model we still attribute 80% of the default intensities λ_j to idiosyncratic shocks, but we now have 20% in common shocks. In the third model we have 60% in common shocks and in the final model we have only common shocks. The effect of the increasing portion of defaults due to common shocks on the distribution of $N(1)$ is seen in Figure 4 and empirical quantiles of $N(1)$ are given in Table 1.

- **Case 2**
  Suppose we attribute 40% of defaults for companies in both ratings classes to idiosyncratic shocks and 60% to common shocks. That is we assume

$$(\lambda_{\text{idio},1}, \lambda_{\text{idio},2}) = (0.002, 0.008).$$

Suppose, for both rating classes, we attribute to sector specific causes, 20% of defaults of sector 1 companies, 50% of defaults of sector 2 companies, 10% of defaults of sector 3 companies and 40% of defaults of sector 4 companies. Moreover we believe that the frequencies of sector and global shocks are in the ratio

$$\lambda_{\text{sector},1} : \lambda_{\text{sector},2} : \lambda_{\text{sector},3} : \lambda_{\text{sector},4} : \lambda_{\text{global}} = 1:5:2:4:1$$

We have now specified the model up to a single factor f. For any $f \geq 0.05$ the following choices of model parameters would satisfy our requirements

$$(\lambda_{\text{sector},1}, \lambda_{\text{sector},2}, \lambda_{\text{sector},3}, \lambda_{\text{sector},4}, \lambda_{\text{global}})$$
$$= f(0.2, 1.0, 0.4, 0.8, 0.2)$$
$$(s_{1,1}, s_{2,1}, s_{3,1}, s_{4,1}, s_{1,2}, s_{2,2}, s_{3,1}, s_{4,2})$$
$$= \frac{1}{f}(0.5, 0.25, 0.125, 0.25, 2, 1, 0.5, 1) 10^{-2}$$

$$(g_{1,1}, g_{2,1}, g_{3,1}, g_{4,1}, g_{1,2}, g_{2,2}, g_{3,1}, g_{4,2}) \\ = \frac{1}{f}(1, 0.25, 1.25, 0.5, 4, 1, 5, 2)10^{-2}.$$

The condition $f \geq 0.05$ is to ensure that $s_1^1, \ldots, s_2^4$, $g_1^1, \ldots, g_2^4 \leq 1$. When f is increased by a factor Δf the intensities of the common shocks are increased by a factor Δf and the univariate conditional default probabilities are decreased by a factor $1/\Delta f$. The effect of increasing f on the distribution of $N(1)$ is seen in figure 3, where histograms are plotted by row for $f = 1,2,4,8$. The key message is as anticipated that low shock intensities and high conditional default probabilities are more riskier than the other way around. Values for the empirical 95th and 99th percentiles of the distribution of $N(1)$ are given in Table 1.

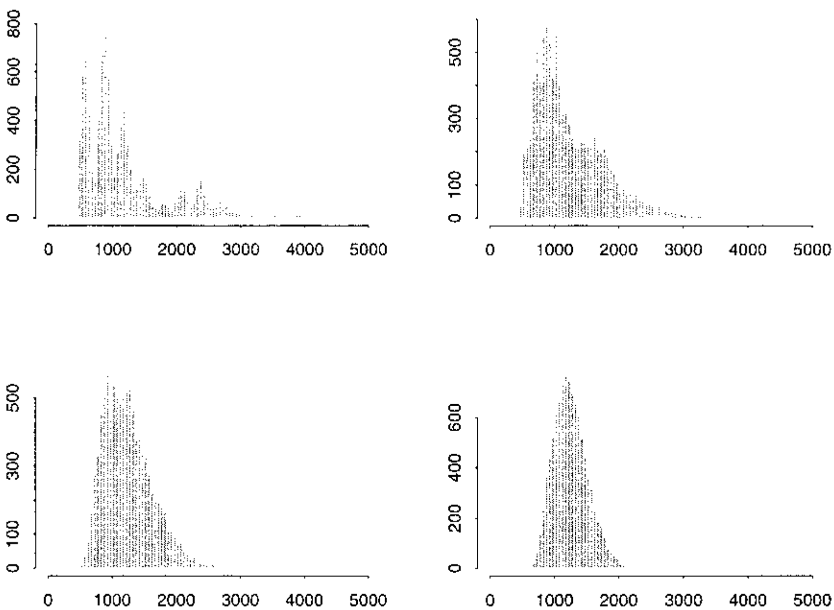


Figure 3: Histograms of 10000 independent simulations of $N(1)$, the number of defaults in a one year period, for $f = 1$ (upper left), $f = 2$ (upper right), $f = 4$ (lower left) and $f = 8$ (lower right).

TABLE 1

EMPIRICAL QUANTILES OF $N(1)$ CORRESPONDING TO THE SAMPLES OF SIZE 10000 SHOWN IN FIGURES 3 AND 4.

	Case 1				***Case 2***			
	f=1	***f=2***	***f=4***	***f=8***	***f=1***	***f=2***	***f=4***	***f=8***
$a = 0.95$	2742	2307	1957	1734	1308	1769	2106	2346
$a = 0.99$	3898	2889	2381	1972	1331	2180	2622	2948

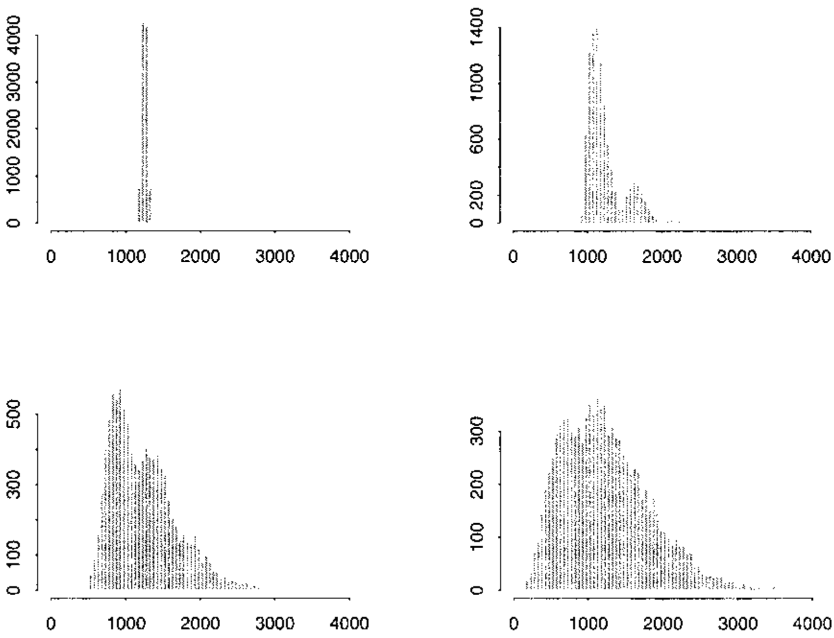


Figure 4: Histograms of 10000 independent simulations of $N(1)$ when increasing the intensities of the common shocks and decreasing the intensities of the idiosyncratic shocks while holding the univariate conditional default probabilities fixed. (1) upper left, (2) upper right, (3) lower left, (4) lower right.

5.6. Conclusion

Clearly the calibration of such models is a difficult and highly judgemental enterprise and the method would seem most useful as a broad brush approach to assessing the risk of a portfolio about which relatively little is known; it might be useful for instance in generating possible future loss scenarios under a variety of assumptions about the frequency and severity of economic downturns. Obviously the higher the number of rating classes and sectors that are introduced the more difficult the calibration will prove to be. Our analysis in Section 5.4 and our simulations in Section 5.5 suggest that calibration might proceed along the following lines.

1. For each combination of rating class and sector, historical data on defaults should be used to estimate what proportions can be attributed to idiosyncratic, sector or global causes. In determining these proportions the constraints imposed by (16) must be respected.
2. Having carved up the individual intensities into these three portions we should then attempt to determine the relative intensity of sector and global events.
3. Bearing in mind that the conditional default probabilities can have a profound impact on the loss distribution we now fix the absolute intensity of sector and global shocks as well as these conditional default probabilities;

for conservatism we err on the side of underestimating shock intensities and overestimating conditional default probabilities.

References

Barlow, R., and Proschan, F. (1975) *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart & Winston, New York.

Credit-Suisse-Financial-Products (1997) CreditRisk+ a Credit Risk Management Framework. Technical Document, available from htpp://www.csfb.com/creditrisk.

Duffie, D., and Singleton, K. (1998) Simulating Correlated Defaults. Working paper, Graduate School of Business, Stanford University.

Embrechts, P., McNeil, A., and Straumann, D. (2001) Correlation and dependency in risk management: properties and pitfalls. In *Risk Management: Value at Risk and Beyond*, ed. by M. Dempster. Cambridge University Press, Cambridge.

Frey, R., and McNeil, A. (2001): Modelling dependent defaults. Preprint, ETH Zürich, available from http://www.math.ethz.ch/~frey.

Hogg, R., and Klugman, S. (1984) *Loss Distributions*. Wiley, New York.

Jarrow, R., and Turnbull, S. (1995) Pricing Derivatives on Financial Securities Subject to Credit Risk. *Journal of Finance*, L(1), 83-85.

Joe, R. (1997) *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

KMV-Corporation (1997) Modelling Default Risk. Technical Document, available from http://www.kmv.com.

Li, D. (1999) On Default Correlation: A Copula Function Approach. Working paper, RiskMetrics Group, New York.

Li, H., and Xu, S. (2001) Stochastic bounds and dependence properties of survival times in a multicomponent shock model. *Journal of Multivariate Analysis* **76**, 63-89.

Marshall, A., and Olkin, I. (1967) A multivariate exponential distribution. *Journal of American Statistical Association* **62**, 30-44.

Merton, R. (1974) On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *Journal of Finance* **29**, 449-470.

Nelsen, R.B. (1999) *An Introduction to Copulas*. Springer, New York.

Panjer, H. (1981) Recursive evaluation of a family of compound distributions. *ASTIN Bulletin* **12**, 22-26.

RiskMetrics-Group (1997) CreditMetrics – Technical Document, available from http://www.riskmetrics.com/research.

Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. (1998) *Stochastic Processes for Insurance and Finance*. Wiley, Chichester.

Savits, T. (1988) Some multivariate distributions derived from a non-fatal shock model. *Journal of Applied Probability* **25**, 383-390.

Schönbucher, P., and Schubert, D. (2001) Copula-dependent default risk in intensity models. Working paper.

Wang, S., and Dhaene, J. (1998) Comonotonicity, correlation order and premium principles. *Insurance: Mathematics and Economics* **22**, 235-242.

Filip Lindskog
Risklab
Federal Institute of Technology
ETH Zentrum
CH-8092 Zurich
Tel.: +41 1 632 67 41
Tel.: +41 1 632 10 85
lindskog@math.ethz.ch

Alexander J. McNeil
Department of Mathematics
Federal Institute of Technology
ETH Zentrum
CH-8092 Zurich
Tel.: +41 1 632 61 62
Tel.: +41 1 632 15 23
mcneil@math.ethz.ch

ASYMPTOTIC DEPENDENCE OF REINSURANCE AGGREGATE CLAIM AMOUNTS

ANA J. MATA*

Heriot-Watt University, United Kingdom

ABSTRACT

In this paper we study the asymptotic behaviour of the joint distribution of reinsurance aggregate claim amounts for large values of the retention level under various dependence assumptions. We prove that, under certain dependence assumptions, for large values of the retention level the ratio between the joint distribution of the aggregate losses and the product of the marginal distributions converges to a constant value that only depends on the frequency parameters.

KEYWORDS

Dependent risks, excess of loss, reinsurance layers, multivariate Panjer recursion, asymptotic independence.

1. INTRODUCTION

Recently the importance of modelling dependent insurance and reinsurance risks has attracted the attention of actuarial practitioners and scientists. Even though classical theories have been developed under the assumption of independence between risks, there are practical cases where this assumption is not valid.

In a recent paper Embrechts, McNeil and Straumann (2001) wrote:

"Although insurance has traditionally been built on the assumption of independence and the law of large numbers has governed the determination of premiums, the increasing complexity of insurance and reinsurance products has led recently to increased actuarial interest in the modelling of dependent risks..."

Although the literature on dependence between risks in insurance portfolios is increasing rapidly, very few authors have applied these development to practical problems, for example reinsurance modelling.

In this paper we study the problem of dependence between risks from the reinsurer's point of view when he provides excess of loss cover for two dependent risks under different dependence assumptions. When the reinsurer undertakes excess of loss reinsurance for a portfolio, in particular for catastrophe excess of

* e-mail: ana.mata@cnare.com

loss, the probability that a claim will impact the reinsurance layer is very small. Therefore, in many cases the correlation coefficient between aggregate claim amounts for the reinsurer becomes very small. We could then be tempted to think that the dependence between portfolios disappears as we look at the tail of the distribution and that, therefore, we could assume independence. It has been largely discussed in the literature that the linear correlation coefficient is not a satisfactory measure of dependence in the non-normal case, see, for example, Embrechts, McNeil and Straumann (2001).

In this paper we look at the effect of different dependence assumptions and their effect on the joint distribution of the aggregate claim amounts compared to the product of the marginal distributions when the retention or attachment is large (hence, the probability of a claim to the layer tends to zero). In Section 2, we describe a model used for insurance and reinsurance aggregate claim amounts that are subject to the same events. In Section 2.1, we discuss how to calculate the distribution of the sum of aggregate claim amounts under different dependence assumptions. In Section 3, we define a measure of asymptotic dependence. We use this concept to study the effect of large values of the retention level on the joint distribution of reinsurance aggregate losses under the dependence assumptions described in Section 2. Numerical illustrations and discussion of the results are presented in Sections 3.2 and 3.3.

2. A dependence model for reinsurance aggregate losses

In this section we describe a model that has been typically used in the actuarial literature to model insurance aggregate claim amounts that are exposed to the same events or claims. This model has been proposed, for example, in Sundt (1999) and Ambagaspitiya (1999) where they develop multivariate recursions to calculate the joint distribution of the aggregate claim amounts. We will assume that there are only two portfolios, however the results can be generalised for any number of risks or portfolios.

The Model: Two risks or portfolios are affected by the same events, therefore they are subject to the same frequency distribution. This model is the general model described in Sundt (1999). One of the most common applications of this model is, for example, catastrophe reinsurance where several portfolios are exposed to the same events. Also in fire insurance, the same fire can cause damage to neighbouring buildings or properties insured under different policies by the same insurer.

Assumptions:

1. Let N be the total number of claims in a fixed period of time. It is assumed that N belongs to the Panjer class of counting distribution, i.e. there exists constants a and b such that N satisfies

$$P(N=n) = \left(a + \frac{b}{n}\right) P(N=n-1) \qquad \text{for } n = 1, 2, \ldots$$

2. Let $\{X_i\}_{i\geq 1}$ and $\{Y_i\}_{i\geq 1}$ be sequences of i.i.d. random variables representing the claim amounts for each risk. We assume that (X_i, Y_i) are i.i.d. pairs from a bivariate distribution.
3. N is independent of $(X_i, Y_i)\ \forall i$.

Then the aggregate claim amounts for these risks are

$$S_1 = \sum_{i=1}^{N} X_i \qquad \text{and} \qquad S_2 = \sum_{i=1}^{N} Y_i \tag{1}$$

The claim amounts for each risk could have various interpretations:

1. From the primary insurer's point of view, the losses due to the ith event (X_i, Y_i) may be dependent or independent.
2. From the reinsurer's point of view (X_i, Y_i) may represent excess of loss claims due to the same event from different underlying risks whose individual losses may be dependent or independent. In reinsurance, this model could also be used for reinsurance losses for two excess of loss layers from the same underlying risk. In this case the aggregate losses are dependent not only through the number of events, but also through the claim distribution for the primary risk.

2.1. Joint distribution of dependent aggregate claim amounts

Sundt (1999) and Ambagaspitiya (1999) developed multivariate recursions that allow us to calculate the joint distribution of the aggregate claim amounts under the assumptions of the model described in Section 2. As we discussed above, the individual claim amounts for each portfolio are not necessarily independent and we assume that they are integer-valued random variables. The joint probability function is given by $p(x,y)$ for $x = 0,1,2,\ldots,\ y = 0,1,2,\ldots$ in appropriate units.

The aggregate claim amounts are as given in formula (1), and the recursion for the joint distribution of (S_1, S_2) is as follows:

$$g(s_1,s_2) = \sum_{u=0}^{s_1}\left(a + \frac{bu}{s_1}\right)\sum_{v=0}^{s_2} p(u,v)\, g(s_1-u, s_2-v), \tag{2}$$

for $s_1 = 1,2,\ldots,\ s_2 = 0,1,2,\ldots$

$$g(s_1,s_2) = \sum_{v=0}^{s_2}\left(a + \frac{bv}{s_2}\right)\sum_{u=0}^{s_1} p(u,v)\, g(s_1-u, s_2-v), \tag{3}$$

for $s_1 = 0,1,2,\ldots, s_2 = 1,2,\ldots$ See Sundt (1999).

In many cases the insurer/reinsurer would only be interested in calculating the distribution of the sum of the total losses for both risks. For example, if we are interested in calculating how much capital we must allocate (under some criteria) to each portfolio separately or to the combined portfolio, then we would be interested in the distribution of the sum of the corresponding aggregate claim amounts.

To calculate the distribution of the sum of dependent aggregate claim amounts under the assumptions of the model described above it is not necessary to calculate the joint distribution. In the next section we discuss in more detail how this is possible.

2.2. Distribution of the sum of dependent aggregate claim amounts

Under the assumptions of the dependence model described in Section 2 the sum of the aggregate claim amounts is given by:

$$S = S_1 + S_1 = \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} (X_i + Y_i). \tag{4}$$

Therefore if we can calculate the distribution of the sum $X_i + Y_i$ for each $i \geq 1$, then the distribution of S can be calculated using Panjer recursion for univariate compound random variables. We denote $U_i = X_i + Y_i$.

Given the joint distribution of the individual claims for the ith event, the distribution of U_i is given by:

$$P(U_i = u) = P(X_i + Y_i = u) = \sum_{m=0}^{u} P(X_i = u - m, Y_i = m), \qquad \text{for } u = 0, 1, \ldots.$$

If X_i and Y_i are independent then $P(U_i = u)$ is given by the convolution of the marginal distributions.

When we consider two excess of loss layers from the same risk, e.g. (m_1, m_2) and (m_2, m_3), if Z_i represents the claim amount due to the ith event for the primary insurer, then the losses for the reinsurer are

$$X_i = \min\left(\max\left(Z_i - m_1, 0\right), m_2 - m_1\right) \qquad \text{and}$$

$$Y_i = \min\left(\max\left(Z_i - m_2, 0\right), m_3 - m_2\right).$$

Hence, $X_i + Y_i$ represents the losses for the combined layer (m_1, m_3), whose distribution can be easily calculated from the distribution of Z_i. However, Mata (2000) showed that for layers of the same risk that are subject to different aggregate conditions such as reinstatements and aggregate deductibles, the distribution of the sum of aggregate losses for two or more layers is not equivalent to the distribution of total aggregate losses for the combined layer. Therefore, the bivariate recursion given in formulae (2) and (3) must be used in these cases.

Example 1. Assume a reinsurer is considering to provide excess of loss cover for the following two layers: $10\, xs\, 20$ and $10\, xs\, 30$ from any two risks (in appropriate units and currency). The reinsurer is given the following information about the underlying risks:

1. Both primary risks are exposed to the same possible events of claims. N, the number of claims during the period of coverage, follows a Poisson distribution with parameter $\lambda = 1$.
2. The individual claims for the ith event for each primary risk, X_i and Y_i, have the same marginal distribution. We assume that the claim size distribution follows a Pareto distribution with parameters $\alpha = 3$ and $\beta = 10$ and probability density function:

$$f(x) = \frac{\alpha\beta^{\alpha}}{(\beta + x)^{\alpha+1}} \qquad \text{for } x > 0.$$

Therefore, for each event the reinsurer's claim amounts are:

$$X_i^R = \min\left(\max\left(0, X_i - 20\right), 10\right) \qquad \text{and} \qquad Y_i^R = \min\left(\max\left(0, Y_i - 30\right), 10\right),$$

hence, the reinsurer's aggregate claim amounts are:

$$S_1^R = \sum_{i=1}^{N} X_i^R \qquad \text{and} \qquad S_2^R = \sum_{i=1}^{N} Y_i^R$$

Since we do not have any extra information about the individual claim amounts for each risk, there are many dependence structures that can be used in order to calculate the joint distribution of the individual claim amounts. Even if we were given the marginal distributions and the correlation coefficient there are several possibilities for the joint distribution of the individual claim amounts, see, for example, Embrechts, McNeil and Straumann (2001). Let us study the following three set ups:

(a) The individual claim amounts X_i and Y_i are independent.
(b) The individual claim amounts are dependent and their joint distribution follows a bivariate Pareto distribution with parameters $(\alpha, \beta_1, \beta_2)$ and joint probability density function

$$f(x,y) = \frac{\alpha(\alpha+1)}{\beta_1\beta_2}\left(-1 + \frac{x+\beta_1}{\beta_1} + \frac{y+\beta_2}{\beta_2}\right)^{-(\alpha+2)}$$

In this example $\alpha = 3$, $\beta_1 = \beta_2 = 10$. For more details about the multivariate Pareto distribution see, for example, Mardia *et al* (1979).
(c) The layers belong to the same underlying risk in which case notice that they are consecutive layers.

Each of these set ups satisfies the assumptions of the model described in Section 2. It can be seen that the covariance between the aggregate claims amounts under the assumptions of the model presented in Section 2 is given by:

$$Cov(S_1, S_2) = E(N)Cov(X_i, Y_i) + Var(N)E[X_i]E[Y_j] \qquad \text{for } i \neq j \tag{5}$$

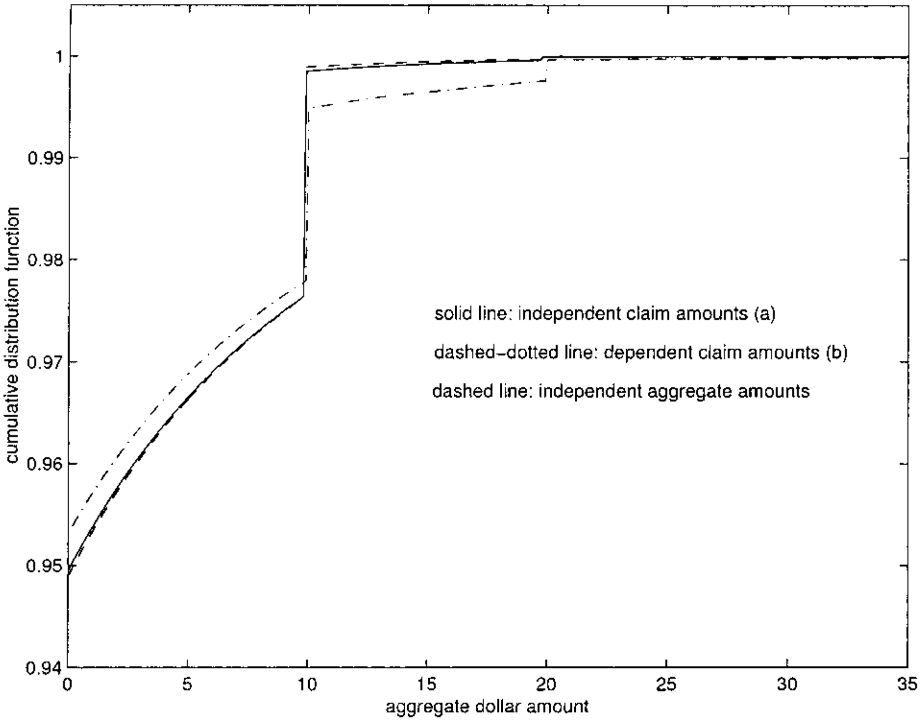


FIGURE 1: Distribution of $S_1^R + S_2^R$, cases (a) and (b).

It is interesting to note that the correlations calculated according to (5) are quite different from each other in the three cases considered in this example. The correlations are:

(a) $\rho\left(S_1^R, S_2^R\right) = 0.019$

(b) $\rho\left(S_1^R, S_2^R\right) = 0.206$

(c) $\rho\left(S_1^R, S_2^R\right) = 0.761$

Figure 1 shows the c.d.f. of $S_1^R + S_2^R$ when the layers belong to different risks, i.e. dependence assumptions as in (a) and (b). The dashed line in Figure 1 represent the c.d.f. of $S_1^R + S_2^R$ when we assume these risks are completely independent, i.e. ignoring that both risks are exposed to the same claims. We observe that under the simplest dependence model (a), where the dependence arises only through the common number of events, the distribution of the total aggregate losses is very close to the distribution of total losses under the assumption of independence. However, when more complex dependence assumptions are built in, such as the bivariate Pareto claim distribution, the distribution of total aggregate losses is very different to the distribution under the independence assumption. In particular we notice that under the dependence assumption in (b) the tail of the distribution is significantly heavier than when independence is assumed.

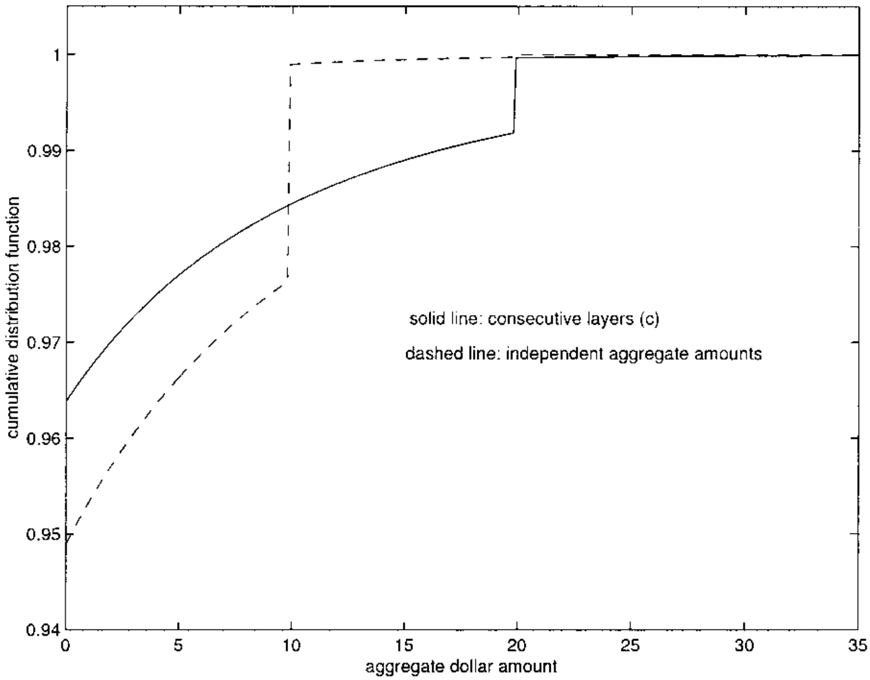


FIGURE 2: Distribution of $S_1^R + S_2^R$, cases (c).

Figure 2 shows the c.d.f. of $S_1^R + S_2^R$ when the layers belong to the same risk and the distribution of $S_1^R + S_2^R$ if independence is assumed. We notice that the distribution of total losses for consecutive layers would be totally mis-estimated if the dependence structure is ignored. This is of course due to the fact that when layers belong to the same risk there would be a positive claim in the second layer only when the claim for the first layers is a full loss. Hence, for layers of the same risk the claim amount dependence has more effect than the frequency dependence.

With this numerical example we have shown how different dependence assumptions may impact the distribution of total losses for the reinsurer. Therefore, by not taking into account how dependence arises one could mis-estimate the overall risk. This is of particular importance when pricing multi-layer excess of loss treaties, where dependence arises not only through he number of claims, but also through the claim size distribution.

3. ASYMPTOTIC BEHAVIOUR OF DEPENDENT REINSURANCE AGGREGATE CLAIM AMOUNTS

In the previous section we looked at the distribution of the sum of dependent aggregate claim amounts from the reinsurer's point of view. In Figures 1 and 2 we observed the effect of different dependence structures that might be used to model reinsurance aggregate claim amounts.

Clearly, given the distribution of the individual claim amounts for the primary insurer and the distribution of the number of claims, the choice of retention level M completely determines the distribution of the aggregate claim amounts for the reinsurer. For large values of the retention level, the probability of a claim to the reinsurance layer is small. Hence the probability of zero losses is very high, and therefore the correlation coefficient is small.

Based on the results showed in Figures 1 and 2, where we observed that under some dependence assumptions the distribution of total aggregate losses for the reinsurer is very close to the distribution of total aggregate losses if independence is assumed, it is our objective to give some insight to the following question:

Are the reinsurer's aggregate claim amounts from different but dependent risks approximately independent for large values of the retention levels?

In the next section we give some theoretical insight into the asymptotic behaviour of the distribution of the aggregate claim amounts for large values of the retention levels under different dependence assumptions.

3.1. On measures of asymptotic independence for reinsurance aggregate claim amounts

In order to provide some answers to the question outlined above we start by giving the definition of asymptotic independence which will be referred to in the remainder of the paper. For large values of the retention levels the probability of a non-zero loss for the reinsurer tends to zero. Hence we use the following definition of asymptotic independence.

Definition 1 *Suppose two sequences of random variables* $\{V_n\}$ and $\{W_n\}$ *are dependent for each n. If these random variables satisfy*

$$\lim_{n \to \infty} \frac{P(V_n \in A, W_n \in B)}{P(V_n \in A)\, P(W_n \in B)} = 1, \tag{6}$$

for all sets A and B that have positive probability, then it is said that V_n *and* W_n *are asymptotically independent. We will refer to the ratio in (6) as the dependence ratio.*

We prove below that under certain dependence assumptions the reinsurance aggregate losses satisfy the condition given in (6) for some sets A and B, but not *for all* sets. We set out below the assumptions we require to prove this result.

Assumptions and notation:

1. The primary insurance risks satisfy the dependence assumptions of the model described in Section 2, but we assume that the claim amounts for

the ith event are independent. Hence, the dependence structure arises only through the common number of events N.

2. Individual claim amounts for the primary insurer X_i and Y_i are integer-valued independent random variables that take values $x = 0, 1, 2, \ldots$ and $y = 0, 1, 2, \ldots$ in appropriate units. We assume that the probability functions of the individual claim amounts do not have fin§(ite upper limit. We denote $p_1(x)$ and $p_2(y)$ as the probability functions of the individual claim amounts for each portfolio.
3. The common number of claims, N, belongs to Panjer's class of counting distributions. Thus, a and b will represent the constants of Panjer's class.
4. We denote by $P_N(t)$ the probability generating function of N which is defined as

$$P_N(t) = E\left[t^N\right]$$

5. Let $\{M_{1,n}\}_{n\geq 1}$ and $\{M_{2,n}\}_{n\geq 1}$ be sequences of integer numbers representing ting the retention levels of the excess of loss reinsurance for each risk. These sequences satisfy $M_{i,n} > M_{i,n-1}$ for $i = 1, 2$. For a given value of the retention, the reinsurer's claim amounts for the ith event are:

$$X_i^R\left(M_{1,n}\right) = \max\left(X_i - M_{1,n}, 0\right) \qquad Y_i^R\left(M_{2,n}\right) = \max\left(Y_i - M_{2,n}, 0\right).$$

Therefore, the aggregate claim amounts for the reinsurer are:

$$S_1^R(M_{1,n}) = \sum_{i=1}^{N} X_i^R\left(M_{1,n}\right) \quad \text{and} \quad S_2^R(M_{2,n}) = \sum_{i=1}^{N} Y_i^R\left(M_{2,n}\right). \tag{7}$$

The distribution functions of the aggregate losses are functions of the retention levels $\{M_{2,n}\}_{n\geq 1}$ and $\{M_{2,n}\}_{n\geq 1}$.

6. The retention levels are such that the reinsurer's aggregate claim amounts satisfy

$$\lim_{n\to\infty} P\left(S_i^R\left(M_{i,n}\right) = 0\right) = 1, \qquad \text{for } i = 1, 2. \tag{8}$$

7. The probability functions for the individual claim amounts for the reinsurer are $p_{1,n}(x) = P\left(X_i^R(M_{1,n}) = x\right)$ and $p_{2,n}(y) = P\left(Y_i^R\left(M_{2,n}\right) = y\right)$ for $x, y = 0, 1, 2, \ldots,$
8. We assume that the probability functions for the individual claim amounts satisfy

$$\lim_{n\to\infty} \frac{p_i(x + M_{i,n})}{p_i(y + M_{i,n})} = C(x, y) \qquad \text{for } x = 1, 2, \ldots, y$$

for $y = 1, 2, \ldots$ where $C(x,y)$ is a constant that only depends on x and y and $0 \leq C(x, y) < \infty$.

9. The probability functions for the aggregate claim amounts are $g_{1,n}(s_1) = P\left(S_1^R\left(M_{1,n}\right) = s_1\right)$ and $g_{2,n}(s_2) = P\left(S_2^R\left(M_{2,n}\right) = s_2\right)$, for $s_1, s_2 = 0, 1, 2, \ldots$. We assume that $g_{i,n}(s_i) > 0$ for $i = 1, 2$ and for $s_1, s_2 = 0, 1, 2, \ldots$
10. The joint probability function for the aggregate claim amounts is defined as $g_n(s_1, s_2) = P\left(S_1^R\left(M_{1,n}\right) = s_1,\ S_2^R\left(M_{2,n}\right) = s_2\right)$, for $s_1, s_2 = 0, 1, 2, \ldots$

Proposition 1. *Under the assumptions outlined above the aggregate claim amounts for the reinsurer defined in (7) satisfy:*

a) $$\lim_{n \to \infty} \frac{g_n(0,0)}{g_{1,n}(0)g_{2,n}(0)} = 1$$

b) $$\lim_{n \to \infty} \frac{P\left(S_1^R\left(M_{1,n}\right) \le s_1, S_2^R\left(M_{2,n}\right) \le s_2\right)}{P\left(S_1^R\left(M_{1,n}\right) \le s_1\right) P\left(S_2^R\left(M_{2,n}\right) \le s_2\right)} = 1 \quad \textit{for all } s_1, s_2 = 0, 1, 2, \ldots$$

c) $$\lim_{n \to \infty} \frac{g_n(s_1, 0)}{g_{1,n}(s_1)g_{2,n}(0)} = 1 \qquad \textit{for } s_1 = 1, 2, \ldots$$

d) $$\lim_{n \to \infty} \frac{g_n(0, s_2)}{g_{1,n}(0)g_{2,n}(s_2)} = 1 \qquad \textit{for } s_2 = 1, 2, \ldots$$

e) $$\lim_{n \to \infty} \frac{g_n(s_1, s_2)}{g_{1,n}(s_1)g_{2,n}(s_2)} = 1 + \frac{1}{a+b} \qquad \textit{for } s_1, s_2 = 1, 2, \ldots$$

Proof. The proof of this proposition is essentially an induction based proof. In order to avoid confusion with the details of the algebraic proof we leave the analytical proof for the Appendix and we concentrate in the interpretation of the results and the assumptions.

From the results shown in Proposition 1 we make the following remarks:

1. Note that for the Binomial, Poisson and Negative Binomial distribution it always holds that $a + b > 0$. Hence, $1 + \frac{1}{a+b} > 1$. In other words, the dependence ratio is always greater than or equal to 1.
2. The statement in b) implies that when we consider cumulative distributions we are including the value of zero which has a high probability for large values of the retention level. This result explains the behaviour observed in Figures 1 and 2 where we considered the cumulative distribution function of the sum of the reinsurer's aggregate claim amounts. In other words, if zero is included in the probability being evaluated the probability would tend to 1 due to assumption 6 above.
3. In Proposition 1 we assumed one of the simplest cases of dependence in insurance/reinsurance risks. Hence, under more complicated assumptions of dependence between risks the dependence ratio might converge to a different value. In the next section we compare numerically the asymptotic behaviour of the dependence ratio under various dependence assumptions.

4. Statement e) shows cases where the ratio between the joint distribution and the product of the marginal distribution does not tend to 1, which proves the fact that even under the simplest dependence assumption independence cannot be assumed. However, if $a + b$ takes large values, the limit would be close to 1. For example, when N follows a Poisson distribution with parameter λ, $a + b = \lambda$ which is the expected number of common events per unit of time. If we increase λ we are increasing the dependence parameter as the expected number of common events becomes larger. Nevertheless, by increasing λ the limit in e) would be closer to 1 which implies that the joint distribution is closer to the independent case. This shows how counter-intuitive results can be when the independence assumption is relaxed.

The following proposition shows another case when the result in e) also holds.

Proposition 2 *Assume that two risks follow assumptions 1 and 3 of Proposition 1. Let $\{M_{1,n}\}$ and $\{M_{2,n}\}$ be sequences of real numbers representing the reinsurance retention levels for each portfolio. These sequences are such that the reinsurer's aggregate claim amounts satisfy the condition in (8) and that for each n*

$$P\left(S_1^R\left(M_{1,n}\right)=0\right) = P\left(S_2^R\left(M_{2,n}\right)=0\right) = \alpha_n.$$

Then, the reinsurance aggregate claim amounts defined in (7) satisfy

$$\lim_{n\to\infty} \frac{P\left(S_1^R\left(M_{1,n}\right)>0, S_2^R\left(M_{2,n}\right)>0\right)}{P\left(S_1^R M_{1,n}>0\right) P\left(S_2^R\left(M_{2,n}\right)>0\right)} = 1+\frac{1}{a+b}.$$

Proof. We give detailed analytical proof of this result in the Appendix.

Note that the assumptions for Proposition 2 are more general than the assumptions we made for Proposition 1. Since the result in Proposition 2 refers to the joint survival function evaluated at zero, the claim size distribution could have a continuous density function. Also the retention levels are not required to be sequences of integer values, it is enough that $\{M_{1,n}\}_{n\geq 1}$ and $\{M_{2,n}\}_{n\geq 1}$ are sequences of real numbers such that:

$$\lim_{n\to\infty} P\left(S_i^R\left(M_{i,n}\right)=0\right)=1 \qquad \text{for } i=1,2$$

For Proposition 2, we assumed that both aggregate claim amounts for the reinsurer have the same probability of being zero. Nevertheless, in Example 2 we show that this is not a necessary assumption. It seems to be sufficient that when n tends to infinity the probabilities of being zero tend to one. We discuss this in more detail in Example 2.

3.2. Numerical illustrations

It is our objective in this section to illustrate numerically the results shown in previous section. We compare numerically the behaviour of the dependence ratio as defined in (6) under all dependence assumptions described in Example 1 above.

Example 2. Assume that two insurance portfolios follow the distributional assumptions as in Example 1. For each risk the reinsurer takes a layer of size 10 with deductibles $M_{1,n} = n$ and $M_{2,n} = 10 + n$ for $n = 0, 1, 2, \ldots$. Therefore, for each event the reinsurer receives claims for the following amounts

$$X_i^R(M_{1,n}) = \min(\max(0, X_i - n), 10) \quad \text{and}$$
$$Y_i^R(M_{2,n}) = \min(\max(0, Y_i - 10 - n), 10).$$

We consider the three set ups as described in Example 1. In this example N follows a Poisson distribution with parameter $\lambda = 1$. Hence, it follows that $1 + \frac{1}{a+b} = 2$.

Figure 3 shows the dependence ratio $\dfrac{P(S_1^R(M_{1,n}) > 0, S_2^R(M_{2,n}) > 0)}{P(S_1^R(M_{1,n}) > 0)\,P(S_2^R(M_{2,n}) > 0)}$ as $n \to \infty$. We observe that the asymptotic behaviour of the joint survival function at zero is very different under the three dependence assumptions. In the case of independent claim amounts, the dependence ratio converge to $1 + \frac{1}{a+b}$ as shown in Proposition 2. However, when the claim amounts are dependent, as in cases (b) and (c), the dependence ratio tends to infinity. We also notice that for layers of the same risk the dependence ratio goes to infinity faster since

$$P(S_1^R(M_{1,n}) > 0, S_2^R(M_{2,n}) > 0) = P(S_2^R(M_{2,n}) > 0)$$

and therefore,

$$\lim_{n \to \infty} \frac{P(S_1^R(M_{1,n}) > 0, S_2^R(M_{2,n}) > 0)}{P(S_1^R(M_{1,n}) > 0)\,P(S_2^R(M_{2,n}) > 0)} = \lim_{n \to \infty} \frac{1}{(S_1^R(M_{1,n}) > 0)} = \infty.$$

Figure 4 shows the behaviour of $\dfrac{g_n(0,1)}{g_{1,n}(0)\,g_{2,n}(1)}$ for large values of the retention n. We note that for layers of the same risk the dependence ratio has a constant value of zero since it is not possible that the second layer takes a positive value if the first layer is zero. However, for layers of separate risks the dependence ratio tends to one. We also observe that in the case of layers from different underlying risks with dependent claim amounts the dependence ratio converges to one, but slower than in the case of independent claims amounts.

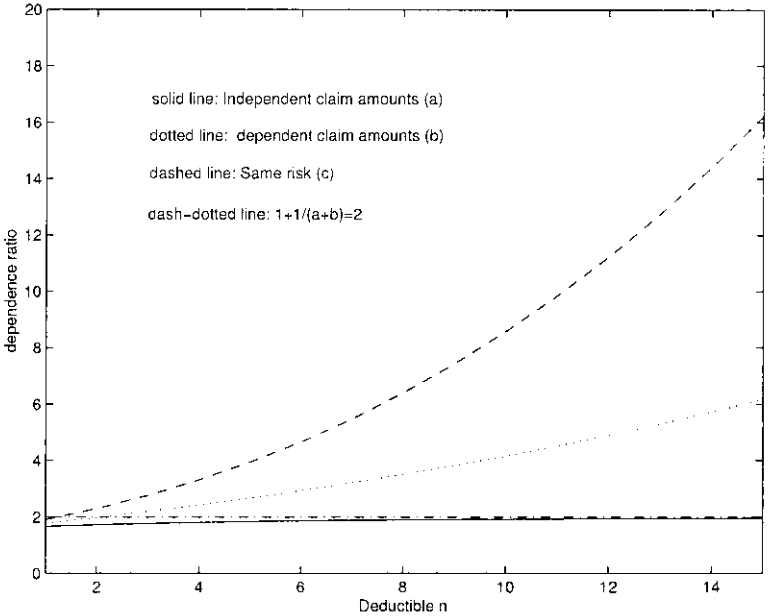


FIGURE 3: Behaviour of $\dfrac{P\left(S_1^R(M_{1,n}) > 0,\, S_2^R(M_{2,n}) > 0\right)}{P\left(S_1^R(M_{1,n}) > 0\right) P\left(S_2^R(M_{2,n}) > 0\right)}$, for Poisson ($\lambda = 1$)

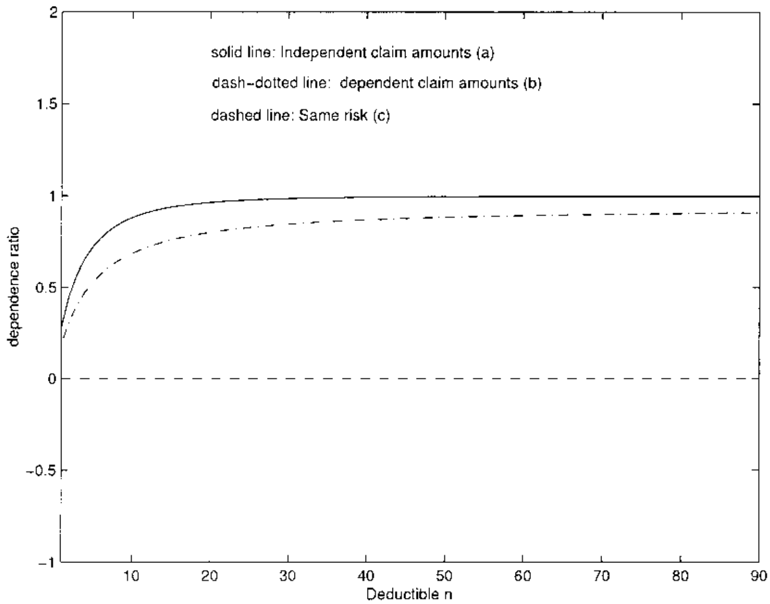


FIGURE 4: Behaviour of $\dfrac{g_n(0,1)}{g_{1,n}(0)\,g_{2,n}(1)}$, for Poisson ($\lambda = 1$)

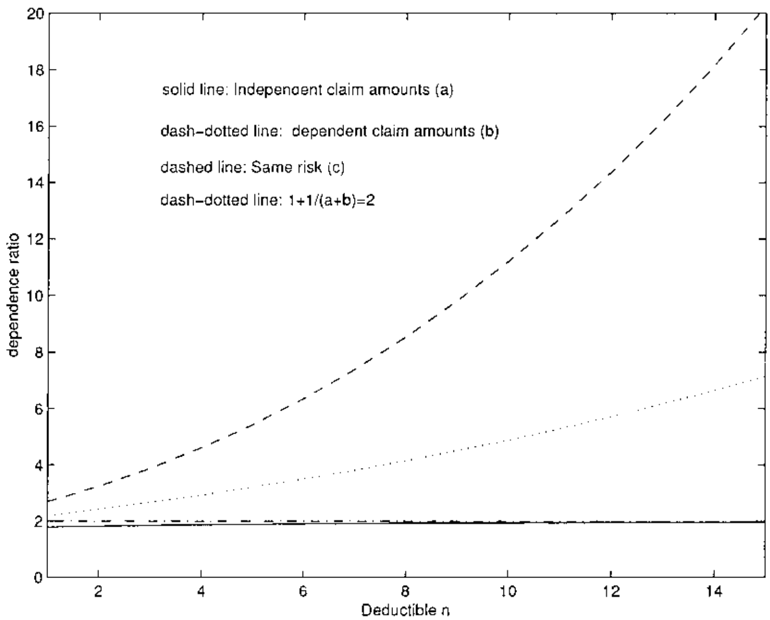


FIGURE 5: Behaviour of $\dfrac{P\left(S_1^R(M_{1,n}) > 2,\ S_2^R(M_{2,n}) > 2\right)}{P\left(S_1^R(M_{1,n}) > 2\right) P\left(S_2^R(M_{2,n}) > 2\right)}$, for Poisson ($\lambda = 1$)

Figure 5 shows the behaviour of

$$\frac{P\left(S_1^R(M_{1,n}) > 2, S_2^R(M_{2,n}) > 2\right)}{P\left(S_1^R(M_{1,n}) > 2) P(S_2^R(M_{2,n}) > 2\right)} \qquad \text{as } n \to \infty$$

We notice that the asymptotic behaviour of joint survival functions for $s_1, s_2 > 0$ (in this case $s_1 = s_2 = 2$) is very similar to the asymptotic behaviour of the joint survival functions for $s_1 = s_2 = 0$. In the case of independent claim amounts the dependence ratio for the joint survival function also converges to $1 + \dfrac{1}{a+b}$ as shown in Proposition 2 for $s_1 = s_2 = 0$.

3.3. Comments on the assumptions in Section 3.1 in practical applications

The results shown in Section 3.1 are a step towards a better understanding of the effect of dependence between reinsurance risks that have small probabilities of large losses. In order to get a better understanding of the results of Propositions 1 and 2 one has to look closely to each of the assumptions made. We discuss below the relevance of each assumption.

1. **Assumption 5:** Note that in Example 2 we have assumed that the excess of loss layers for the reinsurer have a finite upper limit equal to 10. Therefore, assumption 5 does not seem to be a restriction in practical terms. However, the assumption that the reinsurer takes excess of loss layers with infinite limit facilitated the analytical proof.
2. **Assumption 6:** It is not unreasonable to assume that for non proportional reinsurance the retention or deductible is such that the probability of claims affecting the reinsurer is very small. This is a typical assumption in practice.
3. **Assumption 8:** We assumed that the probability function is such that $\lim_{M\to\infty}\dfrac{f(x+M)}{f(y+M)} = C(x,y)$ for $0 < x \le y$, where $0 \le C(x,y) < \infty$ and analytically this is the key assumption for the proof of Proposition 1. Although this property seems to be related to the theory of slowly or regularly varying functions (see, for example, Embrechts, Mikosh and Klüppelberg (1997)) it is in fact a weaker condition as some density functions satisfy assumption 8 but are not regularly varying functions. For example, the Exponential distribution satisfies the condition in assumption 8, however its probability density function is not a regularly varying function. On the other hand, the density function of a Pareto distribution is a regularly-varying function and it also satisfies the condition in assumption 8. The condition in assumption 8 is satisfied by most of the continuous loss distributions used to model insurance/reinsurance losses, such as: Exponential, Gamma, Log-normal, Pareto and Generalised Pareto. The Normal distribution does not satisfy this property as it can be seen that $\lim_{m\to\infty}\dfrac{f(x+m)}{f(m+y)} = \infty$. In practical cases the loss distributions for insurance claims are usually skewed and heavy-tailed, and therefore, the Normal distribution is not a reasonable loss distribution for practical use.
4. **Proposition 2.** In this proposition we assumed that both aggregate claim amounts have the same probability of being zero. Example 2 shows that this seems not to be a restriction. In fact in Example 2 the layers are such that for each n we have

$$P\left(S_1^R\left(M_{1,n}\right)=0\right) < P\left(S_2^R\left(M_{2,n}\right)=0\right).$$

However, for any $\epsilon > 0$ there is M such that for $n \ge M$

$$P\left(S_2^R\left(M_{2,n}\right)=0\right) - P\left(S_1^R\left(M_{1,n}\right)=0\right) < \epsilon,$$

and as n tends to infinite both probabilities tend to one.

4. Conclusions

Modelling dependencies between risks has become an area of increased research interest in actuarial science. Although many authors have emphasized the importance of differentiating between correlation and dependence, in practice, when one thinks of dependencies, inevitably the correlation coefficient is the first thing that comes to mind.

The numerical examples in this paper showed that even when the correlation coefficient becomes small dependence cannot be ignored. Failure to identify dependence between risks may lead to underestimation of the overall risk. This is particularly relevant when pricing risks or managing aggregation of risk exposure.

Throughout this paper we have looked at dependencies from the reinsurer's point of view where there is a very small probability of very large losses. Loosely speaking, the main result states that for large values of the retention levels the dependence ratio converges to a constant defined by the frequency distribution. This constant is always greater than or equal to one. Intuitively, if the aggregate losses are dependent only through the number of events, one would be inclined to think that if the expected number of events increases then the dependence becomes stronger. However, we showed that when the number of events follows a Poisson distribution the larger the expected number of events the joint distribution of aggregate losses gets closer to the product of the marginal distributions which is the distribution of independent aggregate claim amounts.

Modelling dependencies is an area with a vast possibility for research. It is of particular importance to extend the ideas presented in this paper of comparing the joint distribution with the product of the marginal distributions for more general right tail dependence models, for example by looking at multivariate extreme value distributions or extreme value copulas. This comparison is always helpful in practice when often the practical actuary is interested in the impact of making simplified assumptions to facilitate the implementation of new models and techniques.

Acknowledgements

This research was carried out while the author was a research student in the Department of Actuarial Mathematics and Statistics at Heriot-Watt University, Edinburgh. The author would like to thank Professor Howard Waters for his invaluable advise, helpful comments and patience throughout the numerous versions of this paper. Many thanks also to the helpful comments of the anonymous referees and my colleague Mr. Parr Schoolman that significantly helped to improve the presentation of the results. The author is also thankful for the financial support of The British Council and FUNDAYACUCHO from Venezuela.

References

Ambagaspitiya, R.S. (1999) On the distribution of two classes of correlated aggregate claims. *Insurance: Mathematics and Economics* 24:301-308

EMBRECHTS, P., MCNEIL, A.J. and STRAUMANN, D. (2001) Correlations and dependency in risk management: properties and pitfalls *In Risk Management: Value at Risk and Beyond, ed. by M. Dempters and H.K. Moffat*. Cambridge University Press.

EMBRECHTS, P., MIKOSH, T. and KLÜPPELBERG, C. (1997) *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

MARDIA, K.V., KENT, J.T. and BIBBY, J.M. (1979) *Multivariate Analysis.* Academic Press, London.

MATA, A.J. (2000) Pricing excess of loss reinsurance with reinstatemets. *ASTIN Bulletin* 30(2): 349-368.

PANJER, H.H. (1981) Recursive evaluation of a family of compound distributions. *ASTIN Bulletin* 12(1): 21-26.

SUNDT, B. (1999) On multivariate Panjer recursions. *ASTIN Bulletin* 29(1): 29-46.

APPENDIX: ANALYTICAL PROOF OF PROPOSITIONS 1 AND 2

This Appendix gives the analytical proof of Propositions 1 and 2 in Section 3.1.

Proposition 1: *Under the assumptions outlined in Section 3.1. the aggregate claim amounts for the reinsurer defined in* (7) *satisfy:*

a) $$\lim_{n \to \infty} \frac{g_n(0,0)}{g_{1,n}(0) g_{2,n}(0)} = 1$$

b) $$\lim_{n \to \infty} \frac{P\left(S_1^R\left(M_{1,n}\right) \le s_1, S_2^R\left(M_{2,n}\right) \le s_2\right)}{P\left(S_1^R\left(M_{1,n}\right) \le s_1\right) P\left(S_2^R\left(M_{2,n}\right) \le s_2\right)} = 1 \qquad \text{for all } s_1, s_2 = 0, 1, 2, \ldots$$

c) $$\lim_{n \to \infty} \frac{g_n(s_1, 0)}{g_{1,n}(s_1)\, g_{2,n}(0)} = 1 \qquad \text{for } s_1 = 1, 2, \ldots$$

d) $$\lim_{n \to \infty} \frac{g_n(0, s_2)}{g_{1,n}(0)\, g_{2,n}(s_2)} = 1 \qquad \text{for } s_2 = 1, 2, \ldots$$

e) $$\lim_{n \to \infty} \frac{g_n(s_1, s_2)}{g_{1,n}(s_1) g_{2,n}(s_2)} = 1 + \frac{1}{a+b} \qquad \text{for } s_1, s_2 = 1, 2, \ldots$$

Proof.

Reminder: in what follows n represents the indexation of the retention level that are increasing sequences (tending to infinity) as defined in the assumptions in Section 3.1.

a) Since the aggregate claim amounts satisfy the condition (8) in assumption 6, we have that $\lim_{n\to\infty} g_{i,n}(0) = 1$ for $i = 1, 2$. This also implies that $\lim_{n\to\infty} p_{i,n}(0) = 1$ for $i = 1, 2$. The joint probability of being zero is given by $g_n(0,0) = P_N\left(p_{1,n}(0)\, p_{2,n}(0)\right)$, where $P_N(t)$ is the probability generating function of N. Therefore from assumption 6 it also holds that $\lim_{n\to\infty} g_n(0,0) = 1$. Then we directly obtain the result in a).
b) From a) we have that the joint probability of being zero tends to one as well as the probability of each aggregate claim amount being zero. Hence, the result in b) follows directly since we are considering cumulative probabilities which include the value of zero.
c) Since we have assumed that the random variables are integer-valued we can evaluate the joint distribution of the aggregate claim amounts using the bivariate recursion proposed by Sundt (1999) defined in formulae (2) and (3). We will prove the statement in c) by induction and we do the basic step for $s_1 = 1$.

$$\lim_{n\to\infty}\frac{g_n(1,0)}{g_{1,n}(1)g_{2,n}(0)} = \lim_{n\to\infty}\frac{\frac{1}{1-ap_{1,n}(0)p_{2,n}(0)}(a+b)p_{1,n}(1)p_{2,n}(0)\,g_n(0,0)}{\left(\frac{1}{1-ap_{1,n}(0)}(a+b)p_{1,n}(1)g_{1,n}(0)\right)g_{2,n}(0)}$$

$$= \frac{1-a}{1-a},$$

where the last inequality is due to the result in a). The limit above is equal to 1, only when $a \neq 1$. For the Poisson, Negative Binomial and Binomial distributions a takes the values: $a=0, a=(1-p)$ and $a=-p/(1-p)$, respectively. Hence $a \neq 1$ for the three distributions that belong to Panjer's class. Therefore the limit above is equal to 1. We now state the inductive hypothesis: Let us assume that the result in c) holds for all the pairs $(s_1, 0)$ such that $s_1 = 1, 2, \ldots, X$, then we have to prove that it is also true for $(X+1, 0)$ (note that X in this context does not represent a random variable, it is an index in the induction proof). From the recursion in (2) we have

$$g_n(X+1,0) = \frac{p_{2,n}(0)}{1-ap_{1,n}(0)p_{2,n}(0)}\sum_{u=1}^{X+1}\left(a+\frac{bu}{X+1}\right)p_{1,n}(u)g_n(X+1-u,0)$$

Using Panjer's univariate recursion for $g_{1,n}(X+1)$, we have

$$\frac{g_n(X+1,0)}{g_{1,n}(X+1)g_{2,n}(0)} =$$

$$\frac{\frac{p_{2,n}(0)g_{2,n}(0)[1-ap_{1,n}(0)]}{1-ap_{1,n}(0)p_{2,n}(0)}\sum_{u=1}^{X+1}\left(a+\frac{bu}{X+1}\right)p_{1,n}(u)g_{1,n}(X+1-u)\frac{g_n(X+1-u,0)}{g_{1,n}(X+1-u)\,g_{2,n}(0)}}{g_{2,n}(0)\sum_{u=1}^{X+1}\left(a+\frac{bu}{X+1}\right)p_{1,n}(u)g_{1,n}(X+1-u)}$$

Since $\lim_{n\to\infty} p_{1,n}(0) = \lim_{n\to\infty} p_{2,n}(0) = 1$, and using the inductive hypothesis, for any $\epsilon > 0$ there exists K such that for $n \geq K$

$$1-\epsilon < \frac{p_{2,n}(0)[1-ap_{1,n}(0)]}{1-ap_{1,n}(0)p_{2,n}(0)}\,\frac{g_n(X+1-u,0)}{g_{1,n}(X+1-u)\,g_{2,n}(0)} < 1+\epsilon,$$

for $u = 1, 2, \ldots, X+1$. Hence for $n \geq K$

$$1-\epsilon < \frac{g_n(X+1,0)}{g_{1,n}(X+1)\,g_{2,n}(0)} < 1+\epsilon,$$

which proves c).

d) The prove of the statement in d) follows the same argument as c) but using the recursion in formula (3) instead of (2).

e) To prove the statement in e) we will use the results in a), c) and d). We start by proving the result for $s_1 = 1$ and $s_2 = 1$. Using the recursion in formula (2) we have

$$\lim_{n\to\infty}\frac{g_n(1,1)}{g_{1,n}(1)g_{2,n}(1)}=$$

$$\lim_{n\to\infty}\frac{1}{1-ap_{1,n}(0)p_{2,n}(0)}$$

$$\left[\frac{ap_{1,n}(0)p_{2,n}(1)g_n(1,0)}{g_{1,n}(1)g_{2,n}(1)}+\frac{(a+b)p_{1,n}(1)\sum_{v=0}^{1}p_{2,n}(v)\,g_n(0,1-v)}{g_{1,n}(1)\,g_{2,n}(1)}\right]$$

In the limit above we have three terms, we will analyse each term separately. Using Panjer's univariate recursion for $g_{2,n}(1)$ the limit for the first term can be calculated as follows:

$$\lim_{n\to\infty}\frac{ap_{1,n}(0)\,p_{2,n}(1)\,g_{2,n}(0)}{\frac{1}{1-ap_{2,n}(0)}(a+b)p_{2,n}(1)g_{2,n}(0)}\frac{g_n(1,0)}{g_{1,n}(1)g_{2,n}(0)}=\frac{a(1-a)}{a+b},$$

since from part c) we know that $\lim_{n\to\infty}\frac{g_n(1,0)}{g_{1,n}(1)\,g_{2,n}(0)}=1$. For the second term we use the result in d) and Panjer's univariate recursion for $g_{1,n}(1)$, therefore the limit for the second term is

$$\lim_{n\to\infty}\frac{(a+b)p_{1,n}(1)p_{2,n}(0)g_{1,n}(0)}{\frac{1}{1-ap_{1,n}(0)}(a+b)p_{1,n}(1)g_{1,n}(0)}\frac{g_n(0,1)}{g_{1,n}(0)g_{2,n}(1)}=1-a.$$

And finally for the third term we use the result in a) together with Panjer's univariate algorithms for $g_{1,n}(1)$ and $g_{2,n}(1)$, hence the limit is

$$\lim_{n\to\infty}\frac{(a+b)p_{1,n}(1)\,g_{1,n}(0)\,p_{2,n}(1)\,g_{2,n}(0)}{g_{1,n}(1)\,g_{2,n}(1)}\frac{g_n(0,0)}{g_{1,n}(0)\,g_{2,n}(0)}=(1-a)\left(\frac{1-a}{a+b}\right)$$

Putting these three results together we obtain

$$\lim_{n\to\infty}\frac{g_n(1,1)}{g_{1,n}(1)g_{2,n}(1)}=\frac{1}{(1-a)}\left[\frac{a(1-a)}{a+b}+(1-a)+\frac{(1-a)^2}{a+b}\right]=1+\frac{1}{a+b}.$$

We have to use a bivariate induction to prove the result in e). We state the inductive hypothesis as follows:
Assume that

$$\lim_{n\to\infty}\frac{g_n(s_1,s_2)}{g_{1,n}(s_1)g_{2,n}(s_2)}=1+\frac{1}{a+b}$$

for all (s_1,s_2) such that $s_1=1,2,\ldots,X$ and $s_2=1,2,\ldots,Y$, together with the results in a), c) and d). Therefore using this hypothesis we need to prove that the result in statement e) holds in the following three cases:

(i) $(X+1, y)$ such that $y = 1, 2, \dots, Y$
(ii) For all $(x, Y+1)$ such that $x = 1, 2, \dots, X$
(iii) For $(X+1,\ Y+1)$.

Note that X and Y should not be confused with random variables. In each case above the argument is similar except that in (i) we use the recursion in (2) whereas in (ii) we use the recursion in formula (3). We will prove the result only in case (i), the other cases follow. Let us fix y such that $y = 2, 3, \dots, Y$. Together with the inductive hypothesis and a), c) and d) we also assume that the statement in e) holds for all the pairs $(X+1,\ s_2)$ such that $s_2 = 1, \dots, y-1$, then we want to prove the result for $(X+1,\ y)$. Using the recursion in (2) to evaluate $g_n(X+1, y)$ we have

$$\lim_{n\to\infty} \frac{g_n(X+1,y)}{g_{1,n}(X+1)\, g_{2,n}(y)} =$$

$$\lim_{n\to\infty} \frac{1}{1-ap_{1,n}(0)p_{2,n}(0)} \left[\frac{ap_{1,n}(0)\sum_{v=1}^{y} p_{2,n}(v) g_n(X+1,\ y-v)}{g_{1,n}(X+1) g_{2,n}(y)} \right.$$
$$\left. + \frac{\sum_{u=1}^{X+1}\left(a+b\frac{u}{X+1}\right) p_{1,n}(u) \sum_{v=0}^{y} p_{2,n}(v) g_n(X+1-u,\ y-v)}{g_{1,n}(X+1)\, g_{2,n}(y)} \right]$$

To be able to use the results in a), c) and d) we must separate those terms for which one of the entries is zero in the evaluation of g_n from the terms where both entries are greater than zero. Doing so we obtain the following result

$$\lim_{n\to\infty} \frac{1}{1-ap_{1,n}(0)\, p_{2,n}(0)} \times$$
$$\left[\frac{ap_{1,n}(0)\left(\sum_{v=1}^{y-1} p_{2,n}(v)\, g_n(X+1,\ y-v) + p_{2,n}(y)\, g_n(X+1, 0)\right)}{g_{1,n}(X+1)\, g_{2,n}(y)} + \right.$$
$$\frac{\sum_{u=1}^{X}\left(a+b\frac{u}{X+1}\right) p_{1,n}(u)\left(\sum_{v=0}^{y-1} p_{2,n}(v)\, g_n(X+1-u,\ y-v)\right)}{g_{1,n}(X+1)\, g_{2,n}(y)} +$$
$$\frac{\sum_{u=1}^{X}\left(a+b\frac{u}{X+1}\right) p_{1,n}(u)\, p_{2,n}(Y)\, g_n(X+1-u, 0)}{g_{1,n}(X+1)\, g_{2,n}(y)} +$$
$$\left. \frac{(a+b)\, p_{1,n}(X+1)\left(\sum_{v=0}^{y-1} p_{2,n}(v)\, g_n(0,\ y-v) + p_{2,n}(y)\, g_n(0,0)\right)}{g_{1,n}(X+1)\, g_{2,n}(y)} \right]$$

Now we can use the same method of multiplying and dividing each term that contains g_n by the corresponding product of the marginal distributions.

Then for those terms where one or both entries is zero the ratio tends to one, and for the ratios where both entries are greater than zero the ratio tends to $1+\frac{1}{(a+b)}$ due to the inductive hypothesis. Therefore, the limit above can be written as follows

$$\lim_{n\to\infty}\frac{g_n(X+1,y)}{g_{1,n}(X+1)g_{2,n}(y)}=\lim_{n\to\infty}\frac{1}{1-ap_{1,n}(0)p_{2,n}(0)}\times$$

$$\left[\left(\frac{ap_{1,n}(0)\,g_{1,n}(X+1)}{g_{1,n}(X+1)}\right)\left(\frac{\left(1+\frac{1}{a+b}\right)\sum_{v=1}^{y-1}p_{2,n}(v)\,g_{2,n}(y-v)+p_{2,n}(y)\,g_{2,n}(0)}{g_{2,n}(y)}\right)+\right.$$

$$+\left(\frac{\sum_{u=1}^{X}\left(a+b\,\frac{u}{X+1}\right)p_{1,n}(u)\,g_{1,n}(X+1-u)}{g_{1,n}(X+1)}\right)\times$$

$$\left(\frac{\left(1+\frac{1}{a+b}\right)\sum_{v=0}^{y-1}\;p_{2,n}(v)\,g_{2,n}(y-v)+p_{2,n}(y)\,g_{2,n}(0)}{g_{2,n}(y)}\right)+$$

$$\left.+\left(\frac{(a+b)p_{1,n}(X+1)\,g_{1,n}(0)}{g_{1,n}(X+1)}\right)\left(\frac{p_{2,n}(0)\,g_{2,n}(y)+\sum_{v=1}^{y}p_{2,n}(v)\,g_{2,n}(y-v)}{g_{2,n}(y)}\right)\right]. \qquad \text{(A.1)}$$

From condition (8) in assumption 6 it follows that $\lim_{n\to\infty}g_{i,n}(x)=0$ for all $x=1,2,3,\ldots$ and for $i=1,2$. The same result holds for $p_{i,n}(x)$. Using these results we have

$$\lim_{n\to\infty}\sum_{v=1}^{y}p_{2,n}(v)\,g_{2,n}(y-v)=0,$$

since each term tends to zero. However, when we divide the above sum by $g_{2,n}(y)$ we obtain the following result

$$\lim_{n\to\infty}\frac{\sum_{v=1}^{y}\;p_{2,n}(v)\,g_{2,n}(y-v)}{\frac{1}{1-ap_{2,n}(0)}\sum_{v=1}^{y}\left(a+b\frac{v}{y}\right)p_{2,n}(v)\,g_{2,n}(y-v)}=$$

$$\lim_{n\to\infty}\frac{p_{2,n}(1)\,g_{2,n}(y-1)+\cdots+p_{2,n}(y-1)\,g_{2,n}(1)+p_{2,n}(y)\,g_{2,n}(0)}{\frac{1}{1-ap_{2,n}(0)}\left(\left(a+\frac{b}{y}\right)p_{2,n}(1)\,g_{2,n}(y-1)+\cdots+(a+b)\,p_{2,n}(y)\,g_{2,n}(0)\right)}$$

We observe that in the limit above each term tends to zero, however the last term contains $g_{2,n}(0)$ which tends to one as n tends to infinity. Therefore, we divide each term by $p_{2,n}(y)$ and we obtain

$$\lim_{n\to\infty} \frac{\dfrac{p_{2,n}(1)\, g_{2,n}(y-1)}{p_{2,n}(y)} + \cdots + g_{2,n}(0)}{\dfrac{1}{1-ap_{2,n}(0)}\left(\left(a+\dfrac{b}{y}\right)\dfrac{p_{2,n}(1)\, g_{2,n}(y-1)}{p_{2,n}(y)} + \cdots + (a+b)\, g_{2,n}(0)\right)} = \frac{1-a}{a+b}.$$

The last equality is due to the result in assumption 8 where we assumed that $\lim_{n\to\infty} \frac{p_{i,n}(x)}{p_{i,n}(y)} = C(x,y)$ which is a constant for $i = 1,2$, and $\lim_{n\to\infty} g_{i,n}(x) = 0$ for all $x = 1,2,\ldots$. From the discussion above we obtain directly the following results

$$\lim_{n\to\infty} \frac{\sum_{v=1}^{y-1} p_{2,n}(v)\, g_{2,n}(y-v)}{g_{2,n}(y)} = 0$$

$$\lim_{n\to\infty} \frac{\sum_{u=1}^{X}\left(a+b\dfrac{u}{X+1}\right) p_{1,n}(u)\, g_{1,n}(X+1-u)}{g_{1,n}(X+1)} = 0$$

$$\lim_{n\to\infty} \frac{(a+b)\, p_{1,n}(X+1)\, g_{1,n}(0)}{g_{1,n}(X+1)} = 1-a$$

Hence from all the above, we can evaluate the limit in formula (A.1) and we obtain the following result

$$\lim_{n\to\infty} \frac{g_n(X+1,Y)}{g_{1,n}(X+1)\, g_{2,n}(Y)} = \frac{1}{1-a}\left[a\left(\frac{1-a}{a+b}\right) + (1-a)\left(1+\frac{1-a}{a+b}\right)\right] = 1 + \frac{1}{a+b},$$

which is the result shown in e).

Proposition 2: *Suppose that two risks follow assumptions 1 and 3 of Proposition 1. Let $\{M_{1,n}\}$ and $\{M_{2,n}\}$ be sequences of real numbers representing the reinsurance retention levels for each portfolio. These sequences are such that the reinsurer's aggregate claim amounts satisfy the condition in (8) and that for each n*

$$P\left(S_1^R\left(M_{1,n}\right)=0\right) = P\left(S_2^R\left(M_{2,n}\right)=0\right) = \alpha_n.$$

Then, the reinsurance aggregate claim amounts defined in (7) satisfy

$$\lim_{n\to\infty} \frac{P\left(S_1^R\left(M_{1,n}\right)>0, S_2^R\left(M_{2,n}\right)>0\right)}{P\left(S_1^R\left(M_{1,n}\right)>0\right) P\left(S_2^R\left(M_{2,n}\right)>0\right)} = 1 + \frac{1}{a+b}.$$

Proof. Note that in this case we do not require that the individual claim amounts are integer-valued random variables. We also do not need that the retention levels are integer numbers. We denote by $p_{i,n}(0)$ the probability that an individual claim amount for the reinsurer is zero and by $g_{i,n}(0)$ the probability that the aggregate claim amount for the reinsurer is zero for retention level $M_{i,n}$ for $i = 1, 2$.

If $n \to \infty$ then $\alpha_n \to 1^-$. We can write the probabilities of being zero in terms of the probability generating function as

$$g_{i,n}(0) = P\left(S_i^R\left(M_{i,n}\right) = 0\right) = P_N\left(p_{i,n}(0)\right) = \alpha_n \qquad \text{for } i = 1, 2,$$

where $P_N(t)$ is the probability generating function of the frequency distribution.

Hence, $p_{i,n}(0) = P_N^{-1}(\alpha_n)$ for $i = 1, 2$, provided that the inverse of the probability generating function exists. For the Poisson, the Negative Binomial and the Binomial distributions the inverse of the probability generating function can be written explicitly.

As in part a) of Proposition 1 we can write the joint probability of the aggregate claim amounts being zero as follows

$$g_n(0,0) = P_N\left(p_{1,n}(0)\, p_{2,n}(0)\right) = P_N\left(\left(P_N^{-1}(\alpha_n)\right)^2\right).$$

Therefore,

$$\lim_{n \to \infty} \frac{P\left(S_1^R\left(M_{1,n}\right) > 0, S_2^R\left(M_{2,n}\right) > 0\right)}{P\left(S_1^R\left(M_{1,n}\right) > 0\right) P\left(S_2^R\left(M_{2,n}\right) > 0\right)} = \lim_{n \to \infty} \frac{1 - g_{1,n}(0) - g_{2,n}(0) + g_n(0,0)}{\left(1 - g_{1,n}(0)\right)\left(1 - g_{2,n}(0)\right)}$$

$$= \lim_{\alpha_n \to 1^-} \frac{1 - 2\alpha_n + P_N\left(\left(P_N^{-1}(\alpha_n)\right)^2\right)}{(1 - \alpha_n)^2} = \frac{0}{0}.$$

Applying L'Hospital rule twice the limit above can be calculated as follows

$$\lim_{\alpha_n \to 1^-} \left[2 \frac{d^2}{d\alpha_n^2} P_N\left(\left(P_N^{-1}(\alpha_n)\right)^2\right) \frac{(P_N^{-1}(\alpha_n))^2}{\left(\frac{d}{d\alpha_n} P_N\left(P_N^{-1}(\alpha_n)\right)\right)^2} + \right.$$

$$\left. \frac{1}{\frac{d}{d\alpha_n} P_N\left(P_N^{-1}(\alpha_n)\right)} - \frac{d^2}{d\alpha_n^2} P_N\left(P_N^{-1}(\alpha_n)\right) \frac{\left(P_N^{-1}(\alpha_n)\right)}{\left(\frac{d}{d\alpha_n} P_N\left(P_N^{-1}(\alpha_n)\right)\right)^2} \right]$$

From the properties of the probability generating function we have $P_N(1) = 1$ and therefore $P_N^{-1}(1) = 1$. Moreover, the probability generating function satisfies $\frac{d}{dt} P_N(1) = E[N]$. Therefore, the limit above is given by

$$\lim_{n \to \infty} \frac{P\left(S_1^R\left(M_{1,n}\right) > 0, S_2^R\left(M_{2,n}\right) > 0\right)}{P\left(S_1^R\left(M_{1,n}\right) > 0\right) P\left(S_2^R\left(M_{2,n}\right) > 0\right)} = \frac{\frac{d^2}{dt^2} P_N(1)}{(E[N])^2} + \frac{1}{E[N]}.$$

Therefore, by evaluating the limit above with the corresponding values for each of the distributions that belong to Panjer's class it follows that

$$\lim_{n\to\infty} \frac{P\left(S_1^R\left(M_{1,n}\right)>0, S_2^R\left(M_{2,n}\right)>0\right)}{P\left(S_1^R\left(M_{1,n}\right)>0\right)P\left(S_2^R\left(M_{2,n}\right)>0\right)} = 1+\frac{1}{a+b}.$$

Ana J. Mata
CNA Re[1]
333. S. Wabash Av., 39th floor
Chicago IL 60604
USA

[1] The statements, analysis and opinions set forth in this article are solely those of the author and do not reflect the statements, opinions or analysis of any third party, including Continental Casualty Company or any of its subsidiary or affiliated companies.

THE MARKOV CHAIN MARKET

BY

RAGNAR NORBERG

ABSTRACT

We consider a financial market driven by a continuous time homogeneous Markov chain. Conditions for absence of arbitrage and for completeness are spelled out, non-arbitrage pricing of derivatives is discussed, and details are worked out for some cases. Closed form expressions are obtained for interest rate derivatives. Computations typically amount to solving a set of first order partial differential equations. An excursion into risk minimization in the incomplete case illustrates the matrix techniques that are instrumental in the model.

KEYWORDS

Continuous time Markov chains, Martingale analysis, Arbitrage pricing theory, Risk minimization, Unit linked insurance.

INTRODUCTION

A. Prospectus

The theory of diffusion processes, with its wealth of powerful theorems and model variations, is an indispensable toolbox in modern financial mathematics. The seminal papers of Black and Scholes and Merton were crafted with Brownian motion, and so was the major part of the plethora of papers on arbitrage pricing theory and its ramifications that followed over the past good quarter of a century.

A main course of current research, initiated by the martingale approach to arbitrage pricing Harrison and Kreps (1979) and Harrison and Pliska (1981), aims at generalization and unification. Today the core of the matter is well understood in a general semimartingale setting, see e.g. Delbaen and Schachermayer (1994). Another course of research investigates special models, in particular Levy motion alternatives to the Brownian driving process, see e.g. Eberlein and Raible (1999). Pure jump processes have found their way into finance, ranging from plain Poisson processes introduced in Merton (1976) to fairly general marked point processes, see e.g. Björk et al. (1997). As a pedagogical exercise, the market driven by a binomial process has been intensively studied since it was proposed in Cox et al. (1979).

The present paper undertakes to study a financial market driven by a continuous time homogeneous Markov chain. The idea was launched in Norberg (1995) and reappeared in Elliott and Kopp (1998), the context being modeling of the spot rate of interest. These rudiments will here be developed into a model that delineates a financial market with a locally risk-free money account, risky assets, and all conceivable derivatives. The purpose of this exercise is two-fold: In the first place, there is an educative point in seeing how well established theory turns out in the framework of a general Markov chain market and, in particular, how and why it differs from the familiar Brownian motion driven market. In the second place, it is worthwhile investigating the potential of the model from a theoretical as well as from a practical point of view. Further motivation and discussion of the model is given in Section 5.

B. Contents of the paper

We hit the road in Section 2 by recapitulating basic definitions and results for the continuous time Markov chain. We proceed by presenting a market featuring this process as the driving mechanism and by spelling out conditions for absence of arbitrage and for completeness. In Section 3 we carry through the program for arbitrage pricing of derivatives in the Markov chain market and work out the details for some special cases. Special attention is paid to interest rate derivatives, for which closed form expressions are obtained. Section 4 addresses the Föllmer-Sondermann-Schweizer theory of risk minimization in the incomplete case. Its particulars for the Markov chain market are worked out in two examples, first for a unit linked life endowment, and second for hedging strategies involving a finite number of zero coupon bonds. The final Section 5 discusses the versatility and potential uses of the model. It also raises the somewhat intricate issue of existence and continuity of the derivatives involved in the differential equations for state prices, which finds its resolution in a forthcoming paper. Some useful pieces of matrix calculus are placed in the Appendix.

C. Notation

Vectors and matrices are denoted by boldface letters, lower and upper case, respectively. They may be equipped with top-scripts indicating dimensions, e.g. $\mathbf{A}^{n\times m}$ has n rows and m columns. We may write $\mathbf{A} = (a^{ef})_{e\in\mathcal{E}}^{f\in\mathcal{F}}$ to emphasize the ranges of the row index e and the column index f. The transpose of $\mathbf{A}$ is denoted by $\mathbf{A}'$. Vectors are taken to be of column type, hence row vectors appear as transposed (column) vectors. The identity matrix is denoted by $\mathbf{I}$, the vector with all entries equal to 1 is denoted by $\mathbf{1}$, and the vector with all entries equal to 0 is denoted by $\mathbf{0}$. By $\mathbf{D}_{e=1,\dots,n}(a^e)$, or just $\mathbf{D}(\mathbf{a})$, is meant the diagonal matrix with the entries of $\mathbf{a} = (a^1,\dots,a^n)'$ down the principal diagonal. The n-dimensional Euclidean space is denoted by $\mathbb{R}^n$, and the linear subspace spanned by the columns of $\mathbf{A}^{n\times m}$ is denoted by $\mathbb{R}(\mathbf{A})$.

The cardinality of a set $\mathcal{Y}$ is denoted by $|\mathcal{Y}|$. For a finite set it is just its number of elements.

2. The Markov chain market

A. The continuous time Markov chain

At the base of everything is some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\{Y_t\}_{t\geq 0}$ be a continuous time Markov chain with finite state space $\mathcal{Y} = \{1,\ldots,n\}$. We take the paths of Y to be right-continuous and Y_0 deterministic. Assume that Y is time homogeneous so that the transition probabilities

$$p_t^{ef} = \mathbb{P}\left[Y_{\tau+t} = f \,\middle|\, Y_\tau = e\right]$$

depend only on the length of the time period. This implies that the transition intensities

$$\lambda^{ef} = \lim_{t \searrow 0} \frac{p_t^{ef}}{t}, \tag{2.1}$$

$e \neq f$, exist and are constant. To avoid repetitious reminders of the type "$e, f \in \mathcal{Y}$", we reserve the indices e and f for states in $\mathcal{Y}$ throughout. We will frequently refer to

$$\mathcal{Y}^e = \left\{f; \lambda^{ef} > 0\right\},$$

the set of states that are directly accessible from state e, and denote the number of such states by

$$n^e = \left|\mathcal{Y}^e\right|.$$

Put

$$\lambda^{ee} = -\lambda^{e\cdot} = -\sum_{f; f \in \mathcal{Y}^e} \lambda^{ef}$$

(minus the total intensity of transition out of state e). We assume that all states intercommunicate so that $p_t^{ef} > 0$ for all e, f (and $t > 0$). This implies that $n^e > 0$ for all e (no absorbing states). The matrix of transition probabilities,

$$\mathbf{P}_t = \left(p_t^{ef}\right),$$

and the infinitesimal matrix,

$$\mathbf{\Lambda} = \left(\lambda^{ef}\right),$$

are related by (2.1), which in matrix form reads $\mathbf{\Lambda} = \lim_{t \searrow 0} \frac{1}{t}(\mathbf{P}_t - \mathbf{I})$, and by the forward and backward Kolmogorov differential equations,

$$\frac{d}{dt}\mathbf{P}_t = \mathbf{P}_t \mathbf{\Lambda} = \mathbf{\Lambda} \mathbf{P}_t. \tag{2.2}$$

Under the side condition $\mathbf{P}_0 = \mathbf{I}$, (2.2) integrates to

$$\mathbf{P}_t = \exp(\mathbf{\Lambda} t). \tag{2.3}$$

The matrix exponential is defined in the Appendix, from where we also fetch the representation (A.3):

$$\mathbf{P}_t = \mathbf{\Phi} \mathbf{D}_{e=1,\dots,n}\left(e^{\rho^e t}\right) \mathbf{\Phi}^{-1} = \sum_{e=1}^{n} e^{\rho^e t} \boldsymbol{\varphi}^e \boldsymbol{\psi}^{e\prime}. \tag{2.4}$$

Here the first eigenvalue is $\rho^1 = 0$, and the corresponding eigenvectors are $\boldsymbol{\varphi}^1 = \mathbf{1}$ and $\boldsymbol{\psi}^{1\prime} = (p^1, \dots, p^n) = \lim_{t \nearrow \infty} (p_t^{e1}, \dots, p_t^{en})$, the stationary distribution of Y. The remaining eigenvalues, $\rho^2, \dots, \rho^n$, have strictly negative real parts so that, by (2.4), the transition probabilities converge exponentially to the stationary distribution as t increases.

Introduce

$$I_t^e = 1[Y_t = e], \tag{2.5}$$

the indicator of the event that Y is in state at time t, and

$$N_t^{ef} = \left|\{\tau; 0 < \tau \le t, Y_{\tau -} = e, Y_\tau = f\}\right|, \tag{2.6}$$

the number of direct transitions of Y from state e to state $f \in \mathcal{Y}^e$ in the time interval $(0, t]$. For $f \notin \mathcal{Y}^e$ we define $N_t^{ef} \equiv 0$. The assumed right-continuity of Y is inherited by the indicator processes I^e and the counting processes N^{ef}. As is seen from (2.5), (2.6), and the obvious relationships

$$Y_t = \sum_e e I_t^e, \quad I_t^e = I_0^e + \sum_{f;\, f \neq e} \left(N_t^{fe} - N_t^{ef}\right),$$

the state process, the indicator processes, and the counting processes carry the same information, which at any time t is represented by the sigma-algebra $\mathcal{F}_t^Y = \sigma\{Y_\tau; 0 \le \tau \le t\}$. The corresponding filtration, denoted by $\mathbf{F}^Y = \{\mathcal{F}_t^Y\}_{t \ge 0}$, is taken to satisfy the usual conditions of right-continuity and completeness, and $\mathcal{F}_0$ is assumed to be trivial.

The compensated counting processes M^{ef}, $e \neq f$, defined by

$$dM_t^{ef} = dN_t^{ef} - I_t^e \lambda^{ef}\, dt \tag{2.7}$$

and $M_0^{ef} = 0$, are zero mean, square integrable, mutually orthogonal martingales with respect to $(\mathbf{F}^Y, \mathbb{P})$. We feel free to use standard definitions and results from counting process theory and refer to Andersen et al. (1993) for a background.

We now turn to the subject matter of our study and, referring to introductory texts like Björk (1998) and Pliska (1997), take basic concepts and results from arbitrage pricing theory as prerequisites.

B. The continuous time Markov chain market

We consider a financial market driven by the Markov chain described above. Thus, Y_t represents the state of the economy at time t, $\mathcal{F}_t^Y$ represents the information available about the economic history by time t, and $\mathbf{F}^Y$ represents the flow of such information over time.

In the market there are $m+1$ basic assets, which can be traded freely and frictionlessly (short sales are allowed, and there are no transaction costs). A special role is played by asset No. 0, which is a "locally risk-free" *bank account* with state-dependent interest rate

$$r_t = r^{Y_t} = \sum_e I_t^e r^e,$$

where the state-wise interest rates r^e, $e=1,\ldots,n$, are constants. Thus, its price process is

$$S_t^0 = \exp\left(\int_0^t r_u \, du\right) = \exp\left(\sum_e r^e \int_0^t I_u^e \, du\right),$$

where $\int_0^t I_u^e \, du$ is the total time spent in economy state e during the period $[0,t]$. The dynamics of this price process is

$$dS_t^0 = S_t^0 r_t \, dt = S_t^0 \sum_e r^e I_t^e \, dt.$$

The remaining m assets, henceforth referred to as *stocks*, are risky, with price processes of the form

$$S_t^i = \exp\left(\sum_e \alpha^{ie} \int_0^t I_u^e \, du + \sum_e \sum_{f \in \mathcal{Y}^e} \beta^{ief} N_t^{ef}\right), \tag{2.8}$$

$i=1,\ldots,m$, where the α^{ie} and β^{ief} are constants and, for each i, at least one of the β^{ief} is non-null. Thus, in addition to yielding state-dependent returns of the same form as the bank account, stock No. i makes a price jump of relative size

$$\gamma^{ief} = \exp\left(\beta^{ief}\right) - 1$$

upon any transition of the economy from state e to state f. By the general Itô's formula, its dynamics is given by

$$dS_t^i = S_{t-}^i \left(\sum_e \alpha^{ie} I_t^e \, dt + \sum_e \sum_{f \in \mathcal{Y}^e} \gamma^{ief} \, dN_t^{ef}\right). \tag{2.9}$$

(Setting $S_0^i = 1$ for all i is just a matter of convention; it is the relative price changes that matter.)

Taking the bank account as numeraire, we introduce the discounted asset prices $\tilde{S}_t^i = S_t^i / S_t^0$, $i = 0, \ldots, m$. The discounted price of the bank account is $\tilde{S}_t^0 \equiv 1$, which is certainly a martingale under any measure. The discounted stock prices are

$$\tilde{S}_t^i = \exp\left(\sum_e \left(\alpha^{ie} - r^e \right) \int_0^t I_u^e \, du + \sum_e \sum_{f \in y^e} \beta^{ief} N_t^{ef} \right), \tag{2.10}$$

with dynamics

$$d\tilde{S}_t^i = \tilde{S}_{t-}^i \left(\sum_e \left(\alpha^{ie} - r^e \right) I_t^e \, dt + \sum_e \sum_{f \in y^e} \gamma^{ief} \, dN_t^{ef} \right), \tag{2.11}$$

$i = 1, \ldots, m$.

C. Portfolios

A dynamic *portfolio* or *investment strategy* is an $m + 1$-dimensional stochastic process

$$\boldsymbol{\theta}_t = \left(\theta_t^0, \ldots, \theta_t^m \right),$$

where θ_t^i represents the number of units of asset No i held at time t. The portfolio θ must adapted to $\mathbf{F}^Y$ and the shares of stocks, $(\theta_t^1, \ldots, \theta_t^m)$, must also be $\mathbf{F}^Y$-predictable. For a sufficiently rigorous treatment of the concept of predictability, see Andersen et al. (1993). For our purposes it suffices to know that any left-continuous or deterministic process is predictable, the intuition being that the value of a predictable process at any time is determined by the strictly past history of the driving process Y. We will comment on these assumptions at a later point when the prerequisites are in place.

The *value* of the portfolio $\boldsymbol{\theta}$ at time t is

$$V_t^\theta = \boldsymbol{\theta}_t' \mathbf{S}_t = \sum_{i=0}^m \theta_t^i S_t^i.$$

Henceforth we will mainly work with discounted prices and values and, in accordance with (2.10), equip their symbols with a tilde. The discounted value of the portfolio at time t is

$$\tilde{V}_t^\theta = \boldsymbol{\theta}_t' \tilde{\mathbf{S}}_t. \tag{2.12}$$

The strategy $\boldsymbol{\theta}$ is *self-financing* (SF) if $dV_t^\theta = \boldsymbol{\theta}_t' d\mathbf{S}_t$ or (recall $d\tilde{S}_t^0 = 0$)

$$d\tilde{V}_t^\theta = \boldsymbol{\theta}_t' d\tilde{\mathbf{S}}_t = \sum_{i=1}^m \theta_t^i \, d\tilde{S}_t^i. \tag{2.13}$$

D. Absence of arbitrage

Let

$$\tilde{\mathbf{\Lambda}} = (\tilde{\lambda}^{ef})$$

be an infinitesimal matrix that is equivalent to $\mathbf{\Lambda}$ in the sense that $\tilde{\lambda}^{ef} = 0$ if and only if $\lambda^{ef} = 0$. By Girsanov's theorem for counting processes (see e.g. Andersen et al. (1993)) there exists a measure $\tilde{\mathbb{P}}$, equivalent to $\mathbb{P}$, under which Y is a Markov chain with infinitesimal matrix $\tilde{\mathbf{\Lambda}}$. Consequently, the processes M^{ef}, $e \in \mathcal{Y}$, $f \in \mathcal{Y}^e$, defined by

$$d\tilde{M}_t^{ef} = dN_t^{ef} - I_t^e \tilde{\lambda}^{ef}\, dt, \tag{2.14}$$

and $\tilde{M}_0^{ef} = 0$, are zero mean, mutually orthogonal martingales with respect to $(\mathbf{F}^Y, \tilde{\mathbb{P}})$. Rewrite (2.11) as

$$d\tilde{S}_t^i = \tilde{S}_{t-}^i \left[\sum_e \left(\alpha^{ie} - r^e + \sum_{f \in \mathcal{Y}^e} \gamma^{ief} \tilde{\lambda}^{ef} \right) I_t^e\, dt + \sum_e \sum_{f \in \mathcal{Y}^e} \gamma^{ief}\, d\tilde{M}_t^{ef} \right], \tag{2.15}$$

$i = 1, \ldots, m$. The discounted stock prices are martingales with respect to $(\mathbf{F}^Y, \tilde{\mathbb{P}})$ if and only if the drift terms on the right vanish, that is,

$$\alpha^{ie} - r^e + \sum_{f \in \mathcal{Y}^e} \gamma^{ief} \tilde{\lambda}^{ef} = 0, \tag{2.16}$$

$e = 1, \ldots, n$, $i = 1, \ldots, m$. From general theory it is known that the existence of such an equivalent martingale measure $\tilde{\mathbb{P}}$ implies absence of arbitrage. The relation (2.16) can be cast in matrix form as

$$r^e \mathbf{1} - \boldsymbol{\alpha}^e = \mathbf{\Gamma}^e \tilde{\boldsymbol{\lambda}}^e, \tag{2.17}$$

$e = 1, \ldots, n$, where $\mathbf{1}$ is $m \times 1$ and

$$\alpha^e = \left(\alpha^{ie}\right)_{i = 1, \ldots, m}, \quad \mathbf{\Gamma}^e = \left(\gamma^{ief}\right)_{i = 1, \ldots, m}^{f \in \mathcal{Y}^e}, \quad \tilde{\boldsymbol{\lambda}}^e = \left(\tilde{\lambda}^{ef}\right)_{f \in \mathcal{Y}^e}.$$

The existence of an equivalent martingale measure is equivalent to the existence of a solution $\tilde{\boldsymbol{\lambda}}^e$ to (2.17) with all entries strictly positive. Thus, the market is arbitrage-free if (and we can show only if) for each e, $r^e\mathbf{1} - \boldsymbol{\alpha}^e$ is in the interior of the convex cone of the columns of $\mathbf{\Gamma}^e$.

Assume henceforth that the market is arbitrage-free so that (2.15) reduces to

$$d\tilde{S}_t^i = \tilde{S}_{t-}^i \sum_e \sum_{f \in \mathcal{Y}^e} \gamma^{ief}\, d\tilde{M}_t^{ef}. \tag{2.18}$$

Inserting (2.18) into (2.13), we find

$$d\tilde{V}_t^{\theta} = \sum_e \sum_{f \in \mathcal{Y}^e} \sum_{i=1}^{m} \theta_t^i \tilde{S}_{t-}^i \gamma^{ief} \, d\tilde{M}_t^{ef}, \tag{2.19}$$

which means that the value of an SF portfolio is a martingale with respect to $(\mathbf{F}^Y, \tilde{\mathbb{P}})$ and, in particular,

$$\tilde{V}_t^{\theta} = \tilde{\mathbb{E}}[\tilde{V}_T^{\theta} \mid \mathcal{F}_t] \tag{2.20}$$

for $0 \leq t \leq T$. Here $\tilde{\mathbb{E}}$ denotes expectation under $\tilde{\mathbb{P}}$. (The tilde, which in the first place was introduced to distinguish discounted values from the nominal ones, is also attached to the equivalent martingale measure because it arises from the discounted basic price processes.)

We remind of the standard proof of the result that the existence of an equivalent martingale measure implies absence of arbitrage: Under (2.20) one can not have $\tilde{V}_0^{\theta} = 0$ and at the same time have $\tilde{V}_T^{\theta} \geq 0$ almost surely and $\tilde{V}_T^{\theta} > 0$ with positive probability.

We can now explain the assumptions made about the components of the portfolio $\boldsymbol{\theta}_t$. The adaptedness requirement is commonplace and says just that the investment strategy must be based solely on the currently available information. Without this assumption it is easy to construct examples of arbitrages in the present and in any other model, and the theory would become void just as would practical finance if investors could look into the future. The requirement that $(\theta^1, \ldots, \theta^m)$ be $\mathbf{F}^Y$-predictable means that investment in stocks must be based solely on information from the strict past. Also this assumption is omnipresent in arbitrage pricing theory, but its motivation is less obvious. For instance, in the Brownian world ‘predictable’ is the same as ‘adapted’ due to the (assumed) continuity of Brownian paths. In the present model the two concepts are perfectly distinct, and it is easy to explain why a trade in stocks cannot be based on news reported at the very instant where the trade is made. The intuition is that e.g. a crash in the stock market cannot be escaped by rushing money over from stocks to bonds. Sudden jumps in stock prices, which are allowed in the present model, must take the investor by surprise, else there would be arbitrage. This is seen from (2.19). If the θ_t^i, $i = 1, \ldots, m$, could be any adapted processes, then we could choose them in such a manner that $d\tilde{V}_t^{\theta} \geq 0$ almost surely and strictly positive with positive probability. For instance, we could take them such that

$$\tilde{V}_t^{\theta} = \sum_e \sum_{f \in \mathcal{Y}^e} \int_0^t \tilde{M}_\tau^{ef} d\tilde{M}_\tau^{ef} = \frac{1}{2} \sum_e \sum_{f \in \mathcal{Y}^e} \left(\left(\tilde{M}_t^{ef}\right)^2 + \sum_{\tau \in (0,t]} \left(\Delta \tilde{M}_\tau^{ef}\right)^2 \right).$$

Clearly, $\tilde{V}_T^{\theta}$ is non-negative and attains positive values with positive probability while $\tilde{V}_0^{\theta} = 0$, hence $\boldsymbol{\theta}$ would be an arbitrage.

E. Attainability

A T-claim is a contractual payment due at time T. More precisely, it is an $\mathcal{F}_T^Y$-measurable random variable H with finite expected value. The claim is *attainable* if it can be perfectly duplicated by some SF portfolio $\boldsymbol{\theta}$, that is,

$$\tilde{V}_T^{\theta} = \tilde{H}. \tag{2.21}$$

If an attainable claim should be traded in the market, then its price must at any time be equal to the value of the duplicating portfolio in order to avoid arbitrage. Thus, denoting the price process by π_t and, recalling (2.20) and (2.21), we have

$$\tilde{\pi}_t = \tilde{V}_t^{\theta} = \tilde{\mathbb{E}}\left[\tilde{H} \,\middle|\, \mathcal{F}_t\right], \tag{2.22}$$

or

$$\pi_t = \tilde{\mathbb{E}}\left[e^{-\int_t^T r} H \,\middle|\, \mathcal{F}_t\right]. \tag{2.23}$$

(We use the short-hand $e^{-\int_t^T r}$ for $e^{-\int_t^T r_u\,du}$.)

By (2.22) and (2.19), the dynamics of the discounted price process of an attainable claim is

$$d\tilde{\pi}_t = \sum_e \sum_{f \in \mathcal{Y}^e} \sum_{i=1}^m \theta_t^i \tilde{S}_{t-}^i \gamma^{ief} \, d\tilde{M}_t^{ef}. \tag{2.24}$$

F. Completeness

Any T-claim H as defined above can be represented as

$$\tilde{H} = \tilde{\mathbb{E}}[\tilde{H}] + \int_0^T \sum_e \sum_{f \in \mathcal{Y}^e} \eta_t^{ef} d\tilde{M}_t^{ef}, \tag{2.25}$$

where the η_t^{ef} are $\mathbf{F}^Y$-predictable processes (see Andersen et al. (1993)). Conversely, any random variable of the form (2.25) is, of course, a T-claim. By virtue of (2.21) and (2.19), attainability of H means that

$$\begin{aligned} \tilde{H} &= \tilde{V}_0^{\theta} + \int_0^T d\tilde{V}_t^{\theta} \\ &= \tilde{V}_0^{\theta} + \int_0^T \sum_e \sum_{f \in \mathcal{Y}^e} \sum_i \theta_t^i \tilde{S}_{t-}^i \gamma^{ief} \, d\tilde{M}_t^{ef}. \end{aligned} \tag{2.26}$$

Comparing (2.25) and (2.26), we see that H is attainable iff there exist predictable processes $\theta_t^1, \ldots, \theta_t^m$ such that

$$\sum_{i=1}^{m} \theta_t^i \tilde{S}_{t-}^i \gamma^{ief} = \eta_t^{ef},$$

for all e and $f \in \mathcal{Y}^e$. This means that the n^e-vector

$$\boldsymbol{\eta}_t^e = \left(\eta_t^{ef}\right)_{f \in \mathcal{Y}^e}$$

is in $\mathbb{R}(\boldsymbol{\Gamma}^{e\prime})$.

The market is *complete* if every T-claim is attainable, that is, if every n^e-vector is in $\mathbb{R}(\boldsymbol{\Gamma}^{e\prime})$. This is the case if and only if rank $(\boldsymbol{\Gamma}^e) = n^e$, which can be fulfilled for each e only if $m \geq \max_e n^e$, i.e. the number of risky assets is no less than the number of sources of randomness.

3. Arbitrage-pricing of derivatives in a complete market

A. Differential equations for the arbitrage-free price

Assume that the market is arbitrage-free and complete so that the price of any T-claim is uniquely given by (2.22) or (2.23).

Let us for the time being consider a T-claim that depends only on the state of the economy and the price of a given stock at time T. To simplify notation, we drop the top-script indicating this stock throughout and write just

$$S_t = \exp\left(\sum_e \alpha^e \int_0^t I_u^e \, du + \sum_e \sum_{f \in \mathcal{Y}^e} \beta^{ef} N_t^{ef}\right).$$

Thus, the claim is of the form

$$H = h^{Y_T}(S_T) = \sum_e I_T^e h^e(S_T). \tag{3.1}$$

Examples are a European call option defined by $H = (S_T - K)^+$, a caplet defined by $H = (r_T - g)^+ = (r^{Y_T} - g)^+$, and a zero coupon T-bond defined by $H = 1$.

For any claim of the form (3.1) the relevant state variables involved in the conditional expectation (2.23) are (S_t, t, Y_t). This is due to the form of the stock price, by which

$$S_T = S_t \exp\left(\sum_e \alpha^e \int_t^T I_u^e \, du + \sum_e \sum_{f \in \mathcal{Y}^e} \beta^{ef} \left(N_T^{ef} - N_t^{ef}\right)\right), \tag{3.2}$$

and the Markov property, by which the past and the future are conditionally independent, given the present state Y_t. It follows that the price π_t is of the form

$$\pi_t = \sum_{e=1}^{n} I_t^e v^e(S_t, t), \tag{3.3}$$

where the functions

$$v^e(s,t) = \tilde{\mathbb{E}}\left[e^{-\int_t^T r} H \,\middle|\, Y_t = e,\ S_t = s\right] \tag{3.4}$$

are the state-wise prices. Moreover, by (3.2) and the homogeneity of Y, we obtain the representation

$$v^e(s,t) = \mathbb{E}\left[h^{Y_{T-t}}\left(s\,S_{T-t}\right)\middle|\, Y_0 = e\right]. \tag{3.5}$$

The discounted price (2.22) is a martingale with respect to $(\mathbf{F}^Y, \tilde{\mathbb{P}})$. Assume that the functions $v^e(s,t)$ are continuously differentiable. Applying Itô to

$$\tilde{\pi}_t = e^{-\int_0^t r} \sum_{e=1}^n I_t^e v^e\left(S_t, t\right), \tag{3.6}$$

we find

$$\begin{aligned} d\tilde{\pi}_t &= e^{-\int_0^t r} \sum_e I_t^e \Big(- r^e v^e\left(S_t,t\right) + \frac{\partial}{\partial t} v^e\left(S_t,t\right) + \frac{\partial}{\partial s} v^e\left(S_t,t\right) S_t \alpha^e \Big) dt \\ &\quad + e^{-\int_0^t r} \sum_e \sum_{f \in \mathcal{Y}^e} \Big(v^f\left(S_{t-}\left(1+\gamma^{ef}\right), t\right) - v^e\left(S_{t-},t\right)\Big) dN_t^{ef} \\ &= e^{-\int_0^t r} \sum_e I_t^e \Big(- r^e v^e\left(S_t,t\right) + \frac{\partial}{\partial t} v^e\left(S_t,t\right) + \frac{\partial}{\partial s} v^e\left(S_t,t\right) S_t \alpha^e \\ &\quad + \sum_{f \in \mathcal{Y}^e} \Big(v^f\left(S_{t-}\left(1+\gamma^{ef}\right), t\right) - v^e\left(S_{t-},t\right)\Big) \tilde{\lambda}^{ef} \Big) dt \\ &\quad + e^{-\int_0^t r} \sum_e \sum_{f \in \mathcal{Y}^e} \Big(v^f\left(S_{t-}\left(1+\gamma^{ef}\right), t\right) - v^e\left(S_{t-},t\right)\Big) d\tilde{M}_t^{ef}. \end{aligned} \tag{3.7}$$

By the martingale property, the drift term must vanish, and we arrive at the non-stochastic partial differential equations

$$\begin{aligned} &- r^e v^e(s,t) + \frac{\partial}{\partial t} v^e(s,t) + \frac{\partial}{\partial s} v^e(s,t)\, s\alpha^e \\ &\quad + \sum_{f \in \mathcal{Y}^e} \Big(v^f\left(s\left(1+\gamma^{ef}\right), t\right) - v^e(s,t)\Big) \tilde{\lambda}^{ef} = 0 \end{aligned} \tag{3.8}$$

with side conditions

$$v^e(s,T) = h^e(s), \tag{3.9}$$

$e = 1, \ldots, n$.

In matrix form, with

$$\mathbf{R} = \mathbf{D}_{e=1,\ldots,n}(r^e), \qquad \mathbf{A} = \mathbf{D}_{e=1,\ldots,n}(\alpha^e)$$

and other symbols (hopefully) self-explaining, the differential equations and the side conditions are

$$\begin{gathered} -\mathbf{R}\mathbf{v}(s,t) + \frac{\partial}{\partial t}\mathbf{v}(s,t) + s\mathbf{A}\frac{\partial}{\partial s}\mathbf{v}(s,t) + \tilde{\mathbf{\Lambda}}\mathbf{v}\left(s(1+\gamma),t\right) = \mathbf{0}, \\ \mathbf{v}(s,T) = \mathbf{h}(s). \end{gathered} \tag{3.11}$$

There are other ways of obtaining the differential equations. One is to derive them from the integral equations obtained by conditioning on whether or not the process Y leaves its current state in the time interval $(t,T]$ and, in case it does, on the time and the direction of the transition. This approach is taken in Norberg (2002) and is a clue in the investigation of the assumed continuous differentiability of the functions v^e.

Before proceeding we render a comment on the fact that the price of a derivative depends on the drift parameters α^e of the stock prices as is seen from (3.8). This is all different from the Black-Scholes-Merton model in which a striking fact is that the drift parameter does not appear in the derivative prices. There is no contradiction here, however, as both facts reflect the paramount principle that the equivalent martingale measure arises from the path properties of the price processes of the basic assets and depends on the original measure only through its support. The drift term is a path property in the jump process world but not in the Brownian world. In the Markov chain market the pattern of direct transitions as given by the $\mathcal{Y}^e$ is a path property, but apart from that the intensities $\mathbf{F}$ do not affect the derivative prices.

B. Identifying the strategy

Once we have determined the solution $v^e(s,t)$, $e = 1,\ldots,n$, the price process is known and given by (3.3).

The duplicating SF strategy can be obtained as follows. Setting the drift term to 0 in (3.7), we find the dynamics of the discounted price;

$$d\tilde{\pi}_t = e^{-\int_0^t r} \sum_e \sum_{f \in \mathcal{Y}^e} \left(v^f\left(S_{t-}\left(1+\gamma^{ef}\right), t\right) - v^e\left(S_{t-}, t\right)\right) d\tilde{M}_t^{ef}. \tag{3.12}$$

Identifying the coefficients in (3.12) with those in (2.24), we obtain, for each time t and state e, the equations

$$\sum_{i=1}^m \theta_t^i S_{t-}^i \gamma^{ief} = v^f\left(S_{t-}\left(1+\gamma^{ef}\right), t\right) - v^e\left(S_{t-}, t\right), \tag{3.13}$$

$f \in \mathcal{Y}^e$. The solution $(\theta_t^{i,e})_{i=1,\ldots,m}$ certainly exists since rank$(\mathbf{\Gamma}^e) \le m$, and it is unique if rank$(\mathbf{\Gamma}^e) = m$. It is a function of t and $\mathbf{S}_{t-}$ and is thus predictable.

Finally, θ^0 is determined upon combining (2.12), (2.22), and (3.6):

$$\theta_t^0 = e^{-\int_0^t r} \left(\sum_{e=1}^n I_t^e v^e(S_t, t) - \sum_{i=1}^m \theta_t^i S_t^i \right).$$

This function is not predictable.

C. The Asian option

As an example of a path-dependent claim, let us consider an Asian option, which is a T-claim of the form $H = \left(\frac{1}{T}\int_0^T S_\tau \, d\tau - K\right)^+$, where $K \ge 0$. The price process is

$$\begin{aligned}\pi_t &= \tilde{\mathbb{E}}\left[e^{-\int_t^T r} \left(\frac{1}{T} \int_0^T S_\tau \, d\tau - K \right)^+ \middle| \mathcal{F}_t^Y \right] \\ &= \sum_{e=1}^n I_t^e v^e\left(S_t, t, \int_0^t S_\tau \, d\tau\right),\end{aligned}$$

where

$$v^e(s,t,u) = \tilde{\mathbb{E}}\left[e^{-\int_t^T r} \left(\frac{1}{T} \int_t^T S_\tau \, d\tau + \frac{u}{T} - K \right)^+ \middle| Y_t = e, \; S_t = s \right].$$

The discounted price process is

$$\tilde{\pi}_t = e^{-\int_0^t r} \sum_{e=1}^n I_t^e v^e\left(t, S_t, \int_0^t S_\tau \, d\tau\right).$$

We are lead to partial differential equations in three variables.

D. Interest rate derivatives

A particularly simple, but important, class of claims are those of the form $H = h^{Y_T}$. Interest rate derivatives of the form $H = h(r_T)$ are included since $r_t = r^{Y_t}$. For such claims the only relevant state variables are t and Y_t, so that the function in (3.4) depends only on t and e. The differential equations (3.8) and the side condition (3.9) reduce to

$$\frac{d}{dt} v_t^e = r^e v_t^e - \sum_{f \in \mathcal{Y}^e} \left(v_t^f - v_t^e\right) \lambda^{ef}, \tag{3.14}$$

$$v_T^e = h^e. \tag{3.15}$$

In matrix form:

$$\frac{d}{dt}\mathbf{v}_t = \left(\tilde{\mathbf{R}} - \tilde{\boldsymbol{\Lambda}}\right)\mathbf{v}_t,$$
$$\mathbf{v}_T = \mathbf{h}.$$

Similar to (2.3) we arrive at the explicit solution

$$\mathbf{v}_t = \exp\{(\tilde{\boldsymbol{\Lambda}} - \mathbf{R})(T-t)\}\,\mathbf{h}. \tag{3.16}$$

It depends on t and T only through $T-t$.

In particular, the zero coupon bond with maturity T corresponds to $\mathbf{h} = \mathbf{1}$. We will henceforth refer to it as the T-bond in short and denote its price process by $p(t,T)$ and its state-wise price functions by $\mathbf{p}(t,T) = \left(p^e(t,T)\right)_{e=1,\ldots,n}$;

$$\mathbf{p}(t,T) = \exp\{(\tilde{\boldsymbol{\Lambda}} - \mathbf{R})(T-t)\}\,\mathbf{1}. \tag{3.17}$$

For a call option on a U-bond, exercised at time $T(<U)$ with price K, $\mathbf{h}$ has entries $h^e = (p^e(T,U) - K)^+$.

In (3.16)-(3.17) it may be useful to employ the representation (A.3),

$$\exp\{(\tilde{\boldsymbol{\Lambda}} - \mathbf{R})(T-t)\} = \tilde{\boldsymbol{\Phi}}\mathbf{D}_{e=1,\ldots,n}\left(e^{\tilde{\rho}^e(T-t)}\right)\tilde{\boldsymbol{\Phi}}^{-1}. \tag{3.18}$$

4. Risk minimization in incomplete markets

A. Incompleteness

The notion of incompleteness pertains to situations where there exist contingent claims that cannot be duplicated by an SF portfolio and, consequently, do not receive unique prices from the no arbitrage postulate alone. In Paragraph 2F we alluded to incompleteness arising from a scarcity of traded assets, that is, the discounted basic price processes are incapable of spanning the space of all martingales with respect to $(\mathbf{F}^Y, \tilde{\mathbb{P}})$ and, in particular, reproducing the value (2.25) of every financial derivative.

B. Risk minimization

Throughout this section we will mainly be working with discounted prices and values without any other mention than the tilde notation. The reason is that the theory of risk minimization rests on certain martingale representation results that apply to discounted prices under a martingale measure. We will be content to give just a sketchy review of some main concepts and results from the seminal paper of Föllmer and Sondermann (1986) on risk minimization.

Let $\tilde{H}$ be a T-claim that is not attainable. This means that an *admissible* portfolio $\boldsymbol{\theta}$ satisfying

$$\tilde{V}_T^{\theta} = \tilde{H}$$

cannot be SF. The *cost* by time t of an admissible portfolio $\boldsymbol{\theta}$ is denoted by $\tilde{C}_t^{\theta}$ and is defined as that part of the portfolio value that has not been gained from trading:

$$\tilde{C}_t^{\theta} = \tilde{V}_t^{\theta} - \int_0^t \boldsymbol{\theta}_\tau{}' d\tilde{\mathbf{S}}_\tau .$$

The *risk* at time t is defined as the mean squared outstanding cost,

$$\tilde{R}_t^{\theta} = \tilde{\mathbb{E}}\left[\left(\tilde{C}_T^{\theta} - \tilde{C}_t^{\theta}\right)^2 \middle| \mathcal{F}_t\right]. \tag{4.1}$$

By definition, the risk of an admissible portfolio $\boldsymbol{\theta}$ is

$$\tilde{R}_t^{\theta} = \tilde{\mathbb{E}}\left[\left(\tilde{H} - \tilde{V}_t^{\theta} - \int_t^T \boldsymbol{\theta}_\tau{}' d\tilde{\mathbf{S}}_\tau\right)^2 \middle| \mathcal{F}_t\right],$$

which is a measure of how well the current value of the portfolio plus future trading gains approximates the claim. The theory of risk minimization takes this entity as its objective function and proves the existence of an optimal admissible portfolio that minimizes the risk (4.1) for all $t \in [0, T]$.

The proof is constructive and provides a recipe for determining the optimal portfolio. One commences from the *intrinsic value* of $\tilde{H}$ at time t defined as

$$\tilde{V}_t^H = \tilde{\mathbb{E}}[\tilde{H} \mid \mathcal{F}_t]. \tag{4.2}$$

This is the martingale that at any time gives the optimal forecast of the claim with respect to mean squared prediction error under the chosen martingale measure. By the Galchouk-Kunita-Watanabe representation, it decomposes uniquely as

$$\tilde{V}_t^H = \tilde{\mathbb{E}}[\tilde{H}] + \int_0^t \boldsymbol{\theta}_t^{H\prime} d\tilde{\mathbf{S}}_t + L_t^H, \tag{4.3}$$

where L^H is a martingale with respect to $(\mathbf{F}, \tilde{\mathbb{P}})$ which is orthogonal to the martingale $\tilde{\mathbf{S}}$. The portfolio $\boldsymbol{\theta}^H$ defined by this decomposition minimizes the risk process among all admissible strategies. The minimum risk is

$$\tilde{R}_t^H = \tilde{\mathbb{E}}\left[\int_t^T d\left\langle L^H \right\rangle_\tau \middle| \mathcal{F}_t\right]. \tag{4.4}$$

C. Unit-linked insurance

As the name suggests, a life insurance product is said to be *unit-linked* if the benefit is a certain share of an asset (or portfolio of assets). If the contract stipulates a prefixed minimum value of the benefit, then one speaks of *unit-linked insurance with guarantee*. A risk minimization approach to pricing and hedging of unit-linked insurance claims was first taken by Møller (1998), who worked with the Black-Scholes-Merton financial market. We will here sketch how the analysis goes in our Markov chain market, which is a particularly suitable partner for the life history process since both are intensity-driven.

Let T_x be the remaining life time of an x years old who purchases an insurance at time 0, say. The conditional probability of survival to age $x + u$, given survival to age $x + t$ $(0 \le t < u)$, is

$$\mathbb{P}\left[T_x > u \,\middle|\, T_x > t\right] = e^{-\int_t^u \mu_{x+s} ds}, \tag{4.5}$$

where μ_y is the mortality intensity at age y. Introduce the indicator of survival to age $x + t$, $I_t = 1[T_x > t]$, and the indicator of death before time t, $N_t = 1[T_x \le t] = 1 - I_t$. The latter is a (very simple) counting process with intensity $I_t \mu_{x+t}$, and the associated $(\mathbf{F}, \mathbb{P})$ martingale M is given by

$$dM_t = dN_t - I_t \mu_{x+t} dt. \tag{4.6}$$

Assume that the life time T_x is independent of the economy Y. We will be working with the martingale measure $\tilde{\mathbb{P}}$ obtained by replacing the intensity matrix $\mathbf{\Lambda}$ of Y with the martingalizing $\tilde{\mathbf{\Lambda}}$ and leaving the rest of the model unaltered.

Consider a unit-linked pure endowment benefit payable at a fixed time T, contingent on survival of the insured, with sum insured equal to the price S_T of the (generic) stock, but guaranteed no less than a fixed amount g. This benefit is a contingent T-claim,

$$H = (S_T \vee g) I_T.$$

The single premium payable as a lump sum at time 0 is to be determined. Let us assume that the financial market is complete so that every purely financial derivative has a unique price process. Then the intrinsic value of H at time t is

$$\tilde{V}_t^H = \tilde{\pi}_t I_t \, e^{-\int_t^T \mu},$$

where $\tilde{\pi}_t$ is the discounted price process of the derivative $S_T \vee g$, and we have used the somewhat sloppy abbreviation $\int_t^T \mu_{x+u} du = \int_t^T \mu$.

Using Itô together with (4.5) and (4.6) and the fact that M_t and $\tilde{\pi}_t$ almost surely have no common jumps, we find

$$
\begin{aligned}
d\tilde{V}_t^H &= d\tilde{\pi}_t I_{t-} e^{-\int_t^T \mu} + \tilde{\pi}_{t-} I_{t-} e^{-\int_t^T \mu} \mu_{x+t} dt + (0 - \tilde{\pi}_{t-} e^{-\int_t^T \mu}) dN_t \\
&= d\tilde{\pi}_t I_t e^{-\int_t^T \mu} - \tilde{\pi}_t e^{-\int_t^T \mu} dM_t.
\end{aligned}
$$

It is seen that the optimal trading strategy is that of the price process of the sum insured multiplied with the conditional probability that the sum will be paid out, and that

$$
dL_t^H = -e^{-\int_t^T \mu} \tilde{\pi}_t \, dM_t.
$$

Using $d\langle M \rangle_t = I_t \mu_{x+t} \, dt$ (see Andersen et al. (1993)), the minimum risk (4.4) now assumes the form

$$
\tilde{R}_t^H = \tilde{\mathbb{E}}\left[\int_t^T e^{-2\int_\tau^T \mu} \tilde{\pi}_\tau^2 I_\tau \mu_{x+\tau} d\tau \,\middle|\, \mathcal{F}_t\right] = I_t e^{-2\int_0^t r} \sum_e I_t^e R^e(S_t, t), \tag{4.7}
$$

where

$$
R^e(s,t) = \tilde{\mathbb{E}}\left[\int_t^T e^{-2\int_\tau^T \mu} e^{-2\int_t^\tau r} \pi_\tau^2 I_\tau \mu_{x+\tau} d\tau \,\middle|\, S_t = s,\ Y_t = e,\ I_t = 1\right].
$$

Working along the lines of the proof of (3.8), this time starting from the martingale

$$
\begin{aligned}
M_t^R &= \tilde{\mathbb{E}}\left[\int_0^T e^{-2\int_\tau^T \mu} \tilde{\pi}_\tau^2 I_\tau \mu_{x+\tau} d\tau \,\middle|\, \mathcal{F}_t\right] \\
&= \int_0^t e^{-2\int_\tau^T \mu} e^{-2\int_0^\tau r} \pi_\tau^2 I_\tau \mu_{x+\tau} d\tau + I_t e^{-2\int_0^t r} \sum_e I_t^e R^e(S_t, t),
\end{aligned}
$$

we obtain the differential equations

$$
\begin{aligned}
&\left(\pi_t^2 - R^e(s,t)\right)\mu_{x+t} - 2r^e R^e(s,t) + \frac{\partial}{\partial t} R^e(s,t) + \frac{\partial}{\partial s} R^e(s,t) s\alpha^e \\
&+ \sum_{f \in \mathcal{Y}^e} \left(R^f\left(s\left(1+\gamma^{ef}\right), t\right) - R^e(s,t)\right)\tilde{\lambda}^{ef}.
\end{aligned} \tag{4.8}
$$

These are to be solved in parallel with the differential equations (3.8) and are subject to the conditions

$$
R^e(s,t) = 0. \tag{4.9}
$$

D. Trading with bonds: How much can be hedged?

It is well known that in a model with only one source of randomness, like the Black-Scholes-Merton model, the price process of one zero coupon bond will determine the value process of any other zero coupon bond that matures at an earlier date. In the present model this is not the case, and the degree of incompleteness of a given bond market is therefore an issue.

Suppose an agent faces a contingent T-claim and is allowed to invest only in the bank account and a finite number m of zero coupon bonds with maturities T_i, $i = 1,\ldots,m$, all post time T. The scenario could be that regulatory constraints are imposed on the investment strategy of an insurance company. The question is, to what extent can the claim be hedged by self-financed trading in these available assets?

An allowed SF portfolio $\boldsymbol{\theta}$ has a discounted value process $\tilde{V}_t^{\theta}$ of the form

$$d\tilde{V}_t^{\theta} = \sum_{i=1}^{m} \theta_t^i \sum_e \sum_{f \in \mathcal{Y}^e} \left(\tilde{p}^f(t,T_i) - \tilde{p}^e(t,T_i)\right) d\tilde{M}_t^{ef} = \sum_e d\tilde{\mathbf{M}}_t^{e\,\prime} \mathbf{Q}_t^e \boldsymbol{\theta}_t,$$

where $\boldsymbol{\theta}$ is predictable, $\tilde{\mathbf{M}}_t^e = (\tilde{M}_t^{ef})_{f \in \mathcal{Y}^e}$ is the n^e-dimensional vector comprising the non-null entries in the e-th row of $\tilde{\mathbf{M}}_t = (\tilde{M}_t^{ef})$, and

$$\mathbf{Q}_t^e = \mathbf{Y}^e \mathbf{Q}_t,$$

where

$$\mathbf{Q}_t = \left(\tilde{p}^e(t,T_i)\right)_{e=1,\ldots,n}^{i=1,\ldots,m} = \left(\tilde{\mathbf{p}}(t,T_i),\ \ldots,\ \tilde{\mathbf{p}}(t,T_m)\right), \tag{4.10}$$

and $\mathbf{Y}^e$ is the $n^e \times n$ matrix which maps $\mathbf{Q}_t$ to $(\tilde{p}^f(t,T_i) - \tilde{p}^e(t,T_i))_{f \in \mathcal{Y}^e}^{i=1,\ldots,m}$. If e.g. $\mathcal{Y}^n = \{1,\ldots,p\}$, then $\mathbf{Y}^n = (\mathbf{I}^{p\times p}, \mathbf{0}^{p\times(n-p-1)}, -\mathbf{1}^{p\times 1})$.

The sub-market consisting of the bank account and the m zero coupon bonds is complete in respect of T-claims iff the discounted bond prices span the space of all martingales with respect to $(\mathbf{F}^Y, \tilde{\mathbb{P}})$ over the time interval $[0,T]$. This is the case iff, for each e, $\text{rank}(\mathbf{Q}_t^e) = n^e$. Now, since $\mathbf{Y}^e$ obviously has full rank n^e, the rank of $\mathbf{Q}_t^e$ is determined by that of $\mathbf{Q}_t$ in (4.10). We will argue that, typically, $\mathbf{Q}_t$ has full rank. Thus, suppose $\mathbf{c} = (c_1,\ldots,c_m)'$ is such that

$$\mathbf{Q}_t \mathbf{c} = \mathbf{0}^{n\times 1}.$$

Recalling (3.17), this is the same as

$$\sum_{i=1}^{m} c_i \exp\{(\tilde{\boldsymbol{\Lambda}} - \mathbf{R})T_i\}\mathbf{1} = \mathbf{0},$$

or, by (3.18) and since $\tilde{\boldsymbol{\Phi}}$ has full rank,

$$\mathbf{D}_{e=1,\ldots,n}\left(\sum_{i=1}^{m} c_i e^{\tilde{\rho}^e T_i}\right) \tilde{\boldsymbol{\Phi}}^{-1}\mathbf{1} = \mathbf{0}. \tag{4.11}$$

Since $\tilde{\boldsymbol{\Phi}}^{-1}$ has full rank, the entries of the vector $\tilde{\boldsymbol{\Phi}}^{-1}\mathbf{1}$ cannot be all null. Typically all entries are non-null, and we assume this is the case. Then (4.11) is equivalent to

$$\sum_{i=1}^{m} c_i e^{\tilde{\rho}^e T_i} = 0, \quad e = 1,\ldots,n. \tag{4.12}$$

Using the fact that the generalized Vandermonde matrix has full rank (see Gantmacher (1959)), we know that (4.12) has a non-null solution $\mathbf{c}$ if and only if the number of distinct eigenvalues $\tilde{\rho}^e$ is less than m. The role of the Vandermonde matrix in finance is the topic of a parallel paper by the author, Norberg (1999).

In the case where $\mathrm{rank}(\mathbf{Q}_t^e) < n^e$ for some e we would like to determine the Galchouk-Kunita-Watanabe decomposition for a given $\mathcal{F}_T^Y$-claim. The intrinsic value process (4.2) has dynamics of the form

$$d\tilde{V}_t^H = \sum_e \sum_{f \in \mathcal{Y}^e} \eta_t^{ef}\, d\tilde{M}_t^{ef} = \sum_e d\tilde{\mathbf{M}}_t^{e\prime} \boldsymbol{\eta}_t^e, \tag{4.13}$$

where $\boldsymbol{\eta}_t^e = (\eta_t^{ef})_{f \in \mathcal{Y}^e}$ is predictable. We seek its appropriate decomposition (4.3) into

$$\begin{aligned} d\tilde{V}_t^H &= \sum_i \theta_t^i\, d\tilde{p}(t,T_i) + \sum_e \sum_{f \in \mathcal{Y}^e} \zeta_t^{ef}\, d\tilde{M}_t^{ef} \\ &= \sum_e \sum_{f \in \mathcal{Y}^e} \sum_i \theta_t^i \left(\tilde{p}^f(t,T_i) - \tilde{p}^e(t,T_i)\right) d\tilde{M}_t^{ef} + \sum_e \sum_{f \in \mathcal{Y}^e} \zeta_t^{ef}\, d\tilde{M}_t^{ef} \\ &= \sum_e d\tilde{\mathbf{M}}_t^{e\prime} \mathbf{Q}_t^e \boldsymbol{\theta}_t^e + \sum_e d\tilde{\mathbf{M}}_t^{e\prime} \boldsymbol{\zeta}_t^e, \end{aligned}$$

such that the two martingales on the right hand side are orthogonal, that is,

$$\sum_e I_{t-}^e \sum_{f \in \mathcal{Y}^e} \left(\mathbf{Q}_t^e \boldsymbol{\theta}_t^e\right)' \tilde{\boldsymbol{\Lambda}}^e \boldsymbol{\zeta}_t^e = 0,$$

where $\tilde{\boldsymbol{\Lambda}}^e = \mathbf{D}(\tilde{\boldsymbol{\lambda}}^e)$. This means that, for each e, the vector $\boldsymbol{\eta}_t^e$ in (4.13) is to be decomposed into its $\langle\, ,\rangle_{\tilde{\boldsymbol{\Lambda}}^e}$ projections onto $\mathbb{R}(\mathbf{Q}_t^e)$ and its orthocomplement. From (A.4) and (A.5) we obtain

$$\mathbf{Q}_t^e \boldsymbol{\theta}_t^e = \mathbf{P}_t^e \boldsymbol{\eta}_t^e,$$

where

$$\mathbf{P}_t^e = \mathbf{Q}_t^e \left(\mathbf{Q}_t^{e\prime} \tilde{\boldsymbol{\Lambda}}^e \mathbf{Q}_t^e\right)^{-1} \mathbf{Q}_t^{e\prime} \tilde{\boldsymbol{\Lambda}}^e,$$

hence

$$\boldsymbol{\theta}_t^e = \left(\mathbf{Q}_t^{e\prime} \tilde{\boldsymbol{\Lambda}}^e \mathbf{Q}_t^e\right)^{-1} \mathbf{Q}_t^{e\prime} \tilde{\boldsymbol{\Lambda}}^e \boldsymbol{\eta}_t^e. \tag{4.14}$$

Furthermore,

$$\boldsymbol{\zeta}_t^e = \left(\mathbf{I} - \mathbf{P}_t^e\right)\boldsymbol{\eta}_t^e, \tag{4.15}$$

and the minimum risk (4.4) is

$$\tilde{R}_t^H = \tilde{\mathbb{E}}\left[\int_t^T \sum_e \sum_{f \in \mathcal{Y}^e} I_\tau^e \tilde{\lambda}^{ef} \left(\zeta_\tau^{ef}\right)^2 d\tau \,\middle|\, \mathcal{F}_t\right]. \tag{4.16}$$

The computation goes as follows: The coefficients η^{ef} involved in the intrinsic value process (4.13) and the state-wise prices $p^e(t,T_i)$ of the T_i-bonds are obtained by simultaneously solving (3.8) and (3.14), starting from (3.11) and (3.15), respectively, and at each step computing the optimal trading strategy $\boldsymbol{\theta}$ by (4.14) and the ζ from (4.15). The risk may be computed in parallel by solving differential equations for suitably defined state-wise risk functions. The relevant state variables depend on the nature of the T-claim as illustrated in the previous paragraph.

5. Discussion of the model

A. Versatility of the Markov chain

By suitable specification of $\mathcal{Y}$, $\boldsymbol{\Lambda}$, and the asset parameters r^e, α^{ie}, and β^{ief}, we can make the Markov chain market reflect virtually any conceivable feature a real world market may have. We can construct a non-negative mean reverting interest rate. We can design stock markets with recessions and booms, bullish and bearish trends, and crashes and frenzies and other extreme events (not in the mathematical sense of the word, though, since the intensities and the jump sizes are deterministic). We can create forgetful markets and markets with long memory, markets with all sorts of dependencies between assets — hierarchical and others. In the huge class of Markov chains we can also find an approximation to virtually any other theoretical model since the Markov chain models are dense in the model space, roughly speaking. In particular, one can construct a sequence of Markov chain models such that the compensated multivariate counting process converges weakly to a given multivariate Brownian motion. An obvious route from Markov chains to Brownian motion goes via Poisson processes, which we will now elaborate a bit upon.

B. Poisson markets

A Poisson process is totally memoryless whereas a Markov chain recalls which state it is in at any time. Therefore, a Poisson process can be constructed by suitable specification of the Markov chain Y. There are many ways of doing it, but a minimalistic one is to let Y have two states $\mathcal{Y} = \{1,2\}$ and intensities $\lambda^{12} = \lambda^{21} = \lambda$. Then the process N defined by $N_t = N_t^{12} + N_t^{21}$ (the total number of transitions in $(0,t]$) is Poisson with intensity λ since the transitions counted by N occur with constant intensity λ.

Merton (1976) introduced a simple Poisson market with

$$S_t^0 = e^{rt},$$
$$S_t^1 = e^{\alpha t + \beta N_t},$$

where r, α, and β are constants, and N is a Poisson process with constant intensity λ. This model is accommodated in the Markov chain market by letting Y be a two-state Markov chain as prescribed above and taking $r^1 = r^2 = r$, $\alpha^1 = \alpha^2 = \alpha$, and $\beta^{12} = \beta^{21} = \beta$. The no arbitrage condition (2.17) reduces to $\tilde{\lambda} > 0$, where $\tilde{\lambda} = (r - \alpha)/\gamma$ and $\gamma = e^{\beta} - 1$. When this condition is fulfilled, $\tilde{\lambda}$ is the intensity of N under the equivalent martingale measure.

The price function (3.5) now reduces to an expected value in the Poisson distribution with parameter $\tilde{\lambda}(T-t)$:

$$\begin{aligned} v(s,t) &= \tilde{\mathbb{E}}\left[e^{-r(T-t)} h\left(s\, e^{\alpha(T-t) + \beta N_{T-t}}\right)\right] \\ &= e^{-(r+\tilde{\lambda})(T-t)} \sum_{n=0}^{\infty} \frac{(\tilde{\lambda}(T-t))^n}{n!} h\left(s\, e^{\alpha(T-t)+\beta n}\right). \end{aligned} \tag{5.1}$$

A more general Poisson market would have stock prices of the form

$$S_t^i = \exp\left(\alpha^i t + \sum_{j=1}^{n} \beta^{ij} N_t^j\right),$$

$i = 1, \ldots, m$, where the N^j are independent Poisson processes. The Poisson processes can be constructed by the recipe above from independent Markov chains Y^j, $j = 1, \ldots, n$, which constitute a Markov chain, $Y = (Y^1, \ldots, Y^n)$.

C. On differentiability and numerical methods

The assumption that the functions $v^e(s,t)$ are continuously differentiable is not an innocent one and, in fact, it typically does not hold true. An example is provided by the Poisson market in the previous paragraph. From the explicit formula (5.1) it is seen that the price function inherits the smoothness properties of the function h, which typically is not differentiable everywhere and may even have discontinuities. For instance, for $h(s) = (s-K)^+$ (European call) the function v is continuous in both arguments, but continuous differentiability fails to hold on the curves $\{(s,t);\, s\, e^{\alpha(T-t)+n\beta} = K\}$, $n = 0, 1, 2, \ldots$ This warning prompts a careful exploration and mapping of the Markov chain terrain. That task is a rather formidable one and is not undertaken here. Referring to Norberg (2002), let it suffice to report the following: From a recursive system of backward integral equations it is possible to locate the positions of all points (s, t) where the functions v^e are non-smooth. Equipped with this knowledge one can arrange a numerical procedure with controlled global error, which amounts to solving the differential equations where they are valid and gluing the piece-wise solutions together at the exceptional points where they are not. For interest rate derivatives, which involve only ordinary first order differential equations, these problems are less severe and standard methods for numerical computation will do.

Acknowledgments

This work was partly supported by the Mathematical Finance Network under the Danish Social Science Research Council, Grant No. 9800335. A first draft, worked out during visits at the Centre for Actuarial Studies, University of Melbourne, and at the Department of Mathematics, Stanford University, was presented as invited lecture to the 1999 QMF Conference.

References

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. Springer-Verlag.

Björk, T., Kabanov, Y. and Runggaldier, W. (1997) Bond market structures in the presence of marked point processes. *Mathematical Finance* **7**, 211-239.

Björk, T. (1998) *Arbitrage Theory in Continuous Time*, Oxford University Press.

Cox, J., Ross, S. and Rubinstein, M. (1979) Option pricing: A simplified approach. *J. of Financial Economics* **7**, 229-263.

Delbaen, F. and Schachermayer, W. (1994) A general version of the fundamental theorem on asset pricing. *Mathematische Annalen* **300**, 463-520.

Eberlein, E. and Raible, S. (1999) Term structure models driven by general Lévy processes. *Mathematical Finance* **9**, 31-53.

Elliott, R.J. and Kopp, P.E. (1998) *Mathematics of financial markets*, Springer-Verlag.

Föllmer, H. and Sondermann, D. (1986) Hedging of non-redundant claims. In *Contributions to Mathematical Economics in Honor of Gerard Debreu*, 205-223, eds. Hildebrand, W., Mas-Collel, A., North-Holland.

Gantmacher, F.R. (1959) *Matrizenrechnung* II, VEB Deutscher Verlag der Wissenschaften, Berlin.

Harrison, J.M. and Kreps, D.M. (1979) Martingales and arbitrage in multi-period securities markets. *J. Economic Theory* **20**, 381-408.

Harrison, J.M. and Pliska, S. (1981) Martingales and stochastic integrals in the theory of continuous trading. *J. Stoch. Proc. and Appl.* **11**, 215-260.

Karlin, S. and Taylor, H. (1975) *A first Course in Stochastic Processes*, 2nd. ed., Academic Press.

Merton, R.C. (1976) Option pricing when underlying stock returns are discontinuous. *J. Financial Economics* **3**, 125-144.

Møller, T. (1998) Risk minimizing hedging strategies for unit-linked life insurance. *ASTIN Bulletin* **28**, 17-47.

Norberg, R. (1995) A time-continuous Markov chain interest model with applications to insurance. *J. Appl. Stoch. Models and Data Anal.*, 245-256.

Norberg, R. (1999) On the Vandermonde matrix and its role in mathematical finance. *Working paper No. 162*, Laboratory of Actuarial Math., Univ. Copenhagen.

Norberg, R. (2002) Anomalous PDEs in Markov chains: domains of validity and numerical solutions. Research Report, Department of Statistics, London School of Economics: http://stats.lse.ac.uk/norberg. (*Submitted to Finance and Stochastics*).

Pliska, S.R. (1997) *Introduction to Mathematical Finance*, Blackwell Publishers.

A. Appendix: Some useful matrix results

A diagonalizable square matrix $\mathbf{A}^{n\times n}$ can be represented as

$$\mathbf{A} = \mathbf{\Phi}\mathbf{D}_{e=1,\ldots,n}\left(\rho^e\right)\mathbf{\Phi}^{-1} = \sum_{e=1}^{n}\rho^e\phi^e\psi^{e'}, \tag{A.2}$$

where the ϕ^e are the columns of $\mathbf{\Phi}^{n\times n}$ and the $\psi^{e'}$ are the rows of $\mathbf{\Phi}^{-1}$. The ρ^e are the eigenvalues of $\mathbf{A}$, and ϕ^e and $\psi^{e'}$ are the corresponding eigenvectors, right and left, respectively. Eigenvectors (right or left) corresponding to eigenvalues that are distinguishable and non-null are mutually orthogonal. These results can be looked up in e.g. Karlin and Taylor (1975).

The exponential function of $\mathbf{A}^{n\times n}$ is the $n\times n$ matrix defined by

$$\exp(\mathbf{A}) = \sum_{p=0}^{\infty}\frac{1}{p!}\mathbf{A}^p = \mathbf{\Phi}\mathbf{D}_{e=1,\ldots,n}\left(e^{\rho^e}\right)\mathbf{\Phi}^{-1} = \sum_{e=1}^{n}e^{\rho^e}\phi^e\psi^{e'}, \tag{A.3}$$

where the last two expressions follow from (A.2). This matrix has full rank.

If $\mathbf{\Lambda}^{n\times n}$ is positive definite symmetric, then $\langle\boldsymbol{\eta}_1,\boldsymbol{\eta}_2\rangle_{\mathbf{\Lambda}} = \boldsymbol{\eta}_1'\mathbf{\Lambda}\boldsymbol{\eta}_2$ defines an inner product on $\mathbb{R}^n$. The corresponding norm is given by $\|\boldsymbol{\eta}\|_{\mathbf{\Lambda}} = \langle\boldsymbol{\eta},\boldsymbol{\eta}\rangle_{\mathbf{\Lambda}}^{1/2}$. If $\mathbf{Q}^{n\times m}$ has full rank m ($\leq n$), then the $\langle\cdot,\cdot\rangle_{\mathbf{\Lambda}}$-projection of $\boldsymbol{\eta}$ onto $\mathbb{R}(\mathbf{Q})$ is

$$\boldsymbol{\eta}_{\mathbf{Q}} = \mathbf{P}_{\mathbf{Q}}\boldsymbol{\eta}, \tag{A.4}$$

where the projection matrix (or projector) $\mathbf{P}_{\mathbf{Q}}$ is

$$\mathbf{P}_{\mathbf{Q}} = \mathbf{Q}(\mathbf{Q}'\mathbf{\Lambda}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{\Lambda}. \tag{A.5}$$

The projection of $\boldsymbol{\eta}$ onto the orthogonal complement $\mathbb{R}(\mathbf{Q})^{\perp}$ is

$$\boldsymbol{\eta}_{\mathbf{Q}^{\perp}} = \boldsymbol{\eta} - \boldsymbol{\eta}_{\mathbf{Q}} = (\mathbf{I} - \mathbf{P}_{\mathbf{Q}})\boldsymbol{\eta}.$$

Ragnar Norberg
London School of Economics and Political Science
Department of Statistics
Houghton Street, London WC2A 2AE
United Kingdom
E-mail: R.Norberg@lse.ac.uk

PENSION FUNDING AND THE ACTUARIAL ASSUMPTION CONCERNING INVESTMENT RETURNS

BY

M. Iqbal Owadally

Abstract

An assumption concerning the long-term rate of return on assets is made by actuaries when they value defined-benefit pension plans. There is a distinction between this assumption and the discount rate used to value pension liabilities, as the value placed on liabilities does not depend on asset allocation in the pension fund. The more conservative the investment return assumption is, the larger planned initial contributions are, and the faster benefits are funded. A conservative investment return assumption, however, also leads to long-term surpluses in the plan, as is shown for two practical actuarial funding methods. Long-term deficits result from an optimistic assumption. Neither outcome is desirable as, in the long term, pension plan assets should be accumulated to meet the pension liabilities valued at a suitable discount rate. A third method is devised that avoids such persistent surpluses and deficits regardless of conservatism or optimism in the assumed investment return.

Keywords

Actuarial valuation, funding method, intervaluation gains and losses.

1. Introduction

Actuaries periodically value defined benefit pension plans to recommend suitable contribution rates. A number of valuation assumptions are made for this purpose concerning various uncertain factors affecting the value of pension obligations and the funding for these obligations. This set of valuation assumptions is usually called the valuation basis. Different bases may be required for different purposes. For example, in certain jurisdictions, technical solvency bases may be specified by regulation. There may also be a different set of projection assumptions, usually scenario-based or stochastic, to investigate pension benefit amendments, asset-liability management or other issues.

Actuarial valuations for funding purposes, that is, with the objective of recommending a contribution rate are considered in this paper. A deterministic valuation basis is typically employed. Factors of a demographic nature about

ASTIN BULLETIN, Vol. 33, No. 2, 2003, pp. 289-312

which assumptions are made include the mortality of plan participants at various ages, as well as their disability and withdrawal rates from the plan. Assumptions about economic factors such as price and wage inflation are also required when pensions are a function of final or career-average salary and when they are indexed with price inflation. An assumption about investment returns on the pension plan assets is also made.

If the pension liability exceeds the plan assets, then an unfunded liability (or deficit) exists. The unfunded liability varies over time as actual experience generally does not unfold exactly according to actuarial valuation assumptions. Suitable methods of pension funding generate a schedule of contributions that satisfies two objectives. First, unfunded liabilities must be paid off and there must be enough funds to pay benefits as and when they are due. Second, the contributions that are required from the sponsor and members of the plan must be stable over time.

In this paper, we investigate the effect on pension funding of deviation of actual experience from the actuarial investment return assumption. The relevance of this assumption is discussed in section 2. A simple model is described in section 3. It is used to investigate pension funding under two common funding methods, in sections 4 and 5, and under a variation described in section 6 which has the useful property of yielding full funding independently of the investment return assumption. Finally, a numerical example is given in section 7.

A list of important symbols is given here for ease of reference:

AL	actuarial liability
B	benefit paid every year
C_t	pension contribution paid at start of year $(t, t+1)$
F_t	market value of pension plan assets at time t
i	actual rate of return on plan assets
i_A	actuarial assumption for rate of return on plan assets
i_L	actuarial assumption for rate to discount pension liabilities
K	parameter in spreading of gains and losses (equation (20))
K_1, K_2	parameters in modified spreading of gains and losses (equation (36))
L_t	actuarial intervaluation loss in year $(t-1, t)$
m	amortization period for gains and losses in section 4 (equation (13))
n	amortization period for initial unfunded liability (equation (10))
NC	normal cost or normal contribution rate
P_t	payment for initial unfunded liability at time t (equation (10))
S_t	supplementary contribution paid at the start of year $(t, t+1)$
u, u_A, u_L	$1+i$, $1+i_A$, $1+i_L$ respectively
U_t	unamortized part of initial unfunded liability at time t (equation (11))
UL_t	unfunded liability $= AL - F_t$
v, v_A, v_L	$(1+i)^{-1}$, $(1+i_A)^{-1}$, $(1+i_L)^{-1}$ respectively

2. Investment Return Assumption

The actuarial investment return assumption, henceforth denoted by i_A, is an assumption concerning the long-term rate of return on pension plan assets. Funding for pension benefits involves the substitution of contribution income (from plan participants and sponsor) by investment income from accumulated assets. It is well-known that the choice of i_A (and indeed of other valuation assumptions) affects the incidence of contribution payments and pace of funding: see for example Berin (1989, p. 93) and Trowbridge and Farr (1976, p. 27). The more optimistic the investment return assumption is, the larger the investment return is assumed to be in any given year, and the smaller the contribution that is initially required. If insufficient assets are eventually accumulated compared to the pension liability (that is, if a deficit emerges), then higher contributions than otherwise necessary will eventually be required. Conversely, the more conservative i_A is, the larger the contribution that is initially required and, if surpluses emerge, smaller contributions than otherwise necessary, will eventually be required. Thus, the schedule of contribution payments is accelerated the more i_A is conservative, and it is slowed down the more i_A is optimistic. The actuarial choice of i_A is therefore a means of controlling the pace of funding in the pension plan (Daykin, 1976; Trowbridge and Farr, 1976, p. 27).

Another key actuarial valuation assumption is the interest rate assumption (i_L) used to discount pension liabilities. As pension liabilities are not generally traded, they must be priced by comparison with similar asset cash flows. In theory, pension liabilities should be valued using market discount rates, suitably risk-adjusted, or at the rates implied in asset portfolios that are dedicated or matched by cash flow to these liabilities. In practice, more approximate methods are used. Pension liabilities have a long duration and are usually discounted at a single term-independent discount rate which is typically based on corporate bond yields to reflect the risk of default from the sponsor.

In classical actuarial valuation methodology (for example, Trowbridge and Farr, 1976), i_A and i_L are identical. More recent actuarial practice distinguishes between the two assumptions: see for example Actuarial Standard of Practice No. 27 of the Actuarial Standards Board (1996) in the United States. The U.S. pension accounting standard FAS87 also distinguishes between the liability discount rate and the assumption for the "expected long-term rate of return on plan assets". Thornton and Wilson (1992) refer to a "dual-interest" valuation method, used in the United Kingdom, whereby i_A is a "best-estimate assumption" of investment return on the actual asset portfolio and i_L is a "prudent estimate" of investment return based on a hypothetical asset portfolio that matches pension liabilities.

The distinction between the pension liability discount rate assumption and the investment return assumption is often blurred in practice because it is assumed that they are numerically equal. Actuarial Standard of Practice No. 27 of the U.S. Actuarial Standards Board (1996) states that "generally, the appropriate discount rate is the same as the investment return assumption". This presumes that the pension fund is invested in assets that closely match or hedge or immunize the pension liability so that approximately equal discount rates apply to both asset and liability cash flows. In practice, asset allocation may involve a mismatch between assets and liabilities. For example, asset managers may have a rate-of-return objective involving a benchmark portfolio or index set without reference to the liabilities (McGill *et al.*, 1996, p. 659). It is also generally difficult to hedge pension liabilities perfectly with normal market instruments, because of the risk of default from the plan sponsor and because final-salary pensions are related to economic wage inflation.

In this paper, the assumed rates on assets and liabilities (i_A and i_L respectively) are taken to be conceptually distinct (although they could be numerically equal). The aim of this paper is to investigate the effect on pension funding of actual investment returns being different from the assumed investment return on assets.

3. Model

A simplified model of a defined benefit pension plan is used here. For details of the model, refer to Dufresne (1988, 1989) and Owadally and Haberman (1999). A stationary pension plan population is assumed, with fixed mortality and withdrawal rates at different ages. The only benefit that is provided in the model plan is a final-salary pension paid at normal retirement age. There is no inflation on salaries and it is also postulated that actuarial valuation assumptions remain unchanged over time. This leads to a significant simplification in that the payroll, the pension benefit B paid out every year, as well as the combination of actuarial liability AL and normal cost NC generated by a given actuarial cost method, are constant. Trowbridge (1952) shows that an equation of equilibrium holds:

$$AL = (1 + i_L)(AL + NC - B), \tag{1}$$

where i_L is the interest rate used to discount pension liability cash flows. (Alternatively, one may assume that benefits in payment are indexed with wage inflation so that, when measured net of wage inflation, the payroll as well as B, AL and NC are all constant. All quantities must then be considered net of wage inflation).

Assuming that contributions C_t and benefits B are paid at the start of year $(t, t+1)$, the value of the pension fund F_t at time t follows a simple recurrence relation:

$$F_{t+1} = (1+i)(F_t + C_t - B), \tag{2}$$

where i is the actual rate of return earned on the pension plan assets. The unfunded liability is defined as the excess of actuarial liability over assets:

$$UL_t = AL - F_t. \tag{3}$$

It is assumed that all actuarial valuation assumptions, other than i_A, are borne out by experience. In other words, demographic and economic experience unfold in accordance with actuarial valuation assumptions, except that the actual investment rate of return i may differ from the assumed investment rate of return i_A.

An intervaluation loss L_t during year $(t, t+1)$ is the change in unfunded liability as a result of actual experience deviating from actuarial valuation assumptions (Dufresne, 1989). A gain is defined as a negative loss. More specifically, an *asset* loss is the unexpected increase in unfunded liability that is attributable to the actual investment return being less than the investment return assumption. The contribution that is paid at the start of year $(t, t+1)$ is equal to the normal cost NC plus a supplementary contribution S_t which is paid to amortize past intervaluation losses and any initial unfunded liability:

$$C_t = NC + S_t. \tag{4}$$

Letting $v_L = (1+i_L)^{-1}$, it follows from equations (1)-(4) that

$$UL_{t+1} = AL + (1+i)\left(UL_t - S_t - v_L AL\right). \tag{5}$$

Actual experience does not deviate from actuarial assumptions except possibly in investment returns. Therefore, only asset gains or losses occur. An expression for the asset loss is obtained by Dufresne (1989) as follows. Had a rate of return of i_A been earned on the plan assets (instead of the actual rate of return i), the unfunded liability at the end of year $(t, t+1)$ would have been $UL^A_{t+1} = AL + (1+i_A)(UL_t - S_t - v_L AL)$, by comparison with equation (5). Therefore the intervaluation loss in year $(t, t+1)$ is

$$L_{t+1} = UL_{t+1} - UL^A_{t+1} \tag{6}$$

$$= UL_{t+1} - AL - (1+i_A)\left(UL_t - S_t - v_L AL\right) \tag{7}$$

$$= (i - i_A)\left(UL_t - S_t - v_L AL\right). \tag{8}$$

Equation (8) shows that the asset intervaluation loss L_{t+1} in year $(t, t+1)$ arises because the actual return on assets in that year (i) is different from the assumed return (i_A). Equation (7) may be rewritten as

$$UL_{t+1} - u_A UL_t = L_{t+1} - u_A\left(S_t - (v_A - v_L)AL\right), \tag{9}$$

where $u_A = 1 + i_A$ and $v_A = (1+i_A)^{-1}$.

The supplementary contribution S_t in equation (4) pays off over time past intervaluation losses as well as any initial unfunded liability at time 0. The initial unfunded liability may arise because of past service liabilities, or because of a change in the valuation basis or an amendment to benefit rules.

Assume henceforth that $L_t = 0$ for $t \le 0$, $UL_t = 0$ for $t < 0$, and that the initial unfunded liability UL_0 is amortized over a finite period of n years at rate i_A by means of payments

$$P_t = \begin{cases} UL_0 / \ddot{a}_{\overline{n}|}, & 0 \le t \le n-1, \\ 0, & t \ge n. \end{cases} \tag{10}$$

In equation (10), $\ddot{a}_{\overline{n}|} = (1 - v_A^n) / (1 - v_A)$ denotes the present value of an annuity-certain of term n payable in advance and calculated at rate i_A. The unamortized part of the initial unfunded liability at time t is

$$U_t = \begin{cases} UL_0 \, \ddot{a}_{\overline{n-t}|} / \ddot{a}_{\overline{n}|}, & 0 \le t \le n-1, \\ 0, & t \ge n. \end{cases} \tag{11}$$

Observe that

$$u_A U_t - U_{t+1} = u_A P_t. \tag{12}$$

4. Amortizing Gains and Losses

Dufresne (1989) describes a funding method whereby the supplementary contribution S_t, in equation (4), is calculated to amortize past intervaluation gains and losses. His analysis may be extended by allowing for a distinction between the liability valuation rate (i_L) and the investment return assumption (i_A), as well as by explicitly amortizing the initial unfunded liability:

$$S_t = \sum_{j=0}^{m-1} \frac{L_{t-j}}{\ddot{a}_{\overline{m}|}} + (v_A - v_L) AL + P_t. \tag{13}$$

In equation (13), $\ddot{a}_{\overline{m}|} = (1 - v_A^m) / (1 - v_A)$ is the present value of an annuity-certain over m years payable in advance and calculated at assumed rate i_A. The supplementary contribution consists of level amortization payments for intervaluation losses over the past m years, an adjustment for the difference between assumed rates on assets and liabilities, as well as an amortization payment for the initial unfunded liability.

Replacing S_t from equation (13) into equation (9) and using equation (12) yields

$$(UL_{t+1} - U_{t+1}) - u_A (UL_t - U_t) = L_{t+1} - u_A \sum_{j=0}^{m-1} \frac{L_{t-j}}{\ddot{a}_{\overline{m}|}}. \tag{14}$$

The unfunded liability at the end of the year is therefore the accumulation of the unfunded liability at the start of the year plus the loss that emerges during the year less the accumulated value of payments made in respect of past losses.

It is easily verified that the solution of equation (14) is

$$UL_t - U_t = \sum_{j=0}^{m-1} \frac{\ddot{a}_{\overline{m-j}|}}{\ddot{a}_{\overline{m}|}} L_{t-j}. \tag{15}$$

For details of this solution, see Dufresne (1989). Note also equation (12) for the initial unfunded liability and recall that the annuities are valued at rate i_A.

When the funding method in equation (13) is used, a unit loss that emerged j years ago is completely paid off if $j \geq m$, but further payments of $1/\ddot{a}_{\overline{m}|}$ for the next $m-j$ years are outstanding if $0 \leq j \leq m-1$. The present value of these payments is $\ddot{a}_{\overline{m-j}|}/\ddot{a}_{\overline{m}|}$. Equation (15) shows that the unfunded liability is the present value of the payments that remain to be made in respect of losses that are not yet paid off, together with the unamortized part of the initial unfunded liability.

As in Dufresne (1989), replace S_t from equation (13) and UL_t from equation (15) into equation (8), and use equation (12), to obtain:

$$L_{t+1} = (i - i_A)\left[\sum_{j=0}^{m-1} L_{t-j}\left(\ddot{a}_{\overline{m-j}|} - 1\right)/\ddot{a}_{\overline{m}|} - v_A\left(AL - U_{t+1}\right)\right]. \tag{16}$$

If the actual rate of return on plan assets in a given year is the same as the assumed rate of return (that is, if $i = i_A$), no intervaluation loss emerges in that year ($L_t = 0 \;\forall\; t$ from equation (16)) and the unfunded liability consists only of the unamortized part of the initial unfunded liability ($UL_t = U_t$ for $t \geq 0$ from equation (15)).

Dufresne (1989) obtains a sufficient condition for the convergence of $\{L_t\}$, $\{UL_t\}$ and $\{S_t\}$ as $t \to \infty$. The following result is due to Dufresne (1989).

RESULT 1. *Provided that* $\left|i - i_A\right| \sum_{j=0}^{m-1}\left(\ddot{a}_{\overline{m-j}|} - 1\right)/\ddot{a}_{\overline{m}|} < 1$,

$$\lim_{t\to\infty} L_t = \frac{-(i - i_A)v_A AL}{1 - (i - i_A)\sum_{j=0}^{m-1}\left(\ddot{a}_{\overline{m-j}|} - 1\right)/\ddot{a}_{\overline{m}|}}, \tag{17}$$

$$\lim_{t\to\infty} UL_t = \sum_{j=0}^{m-1} \frac{\ddot{a}_{\overline{m-j}|}}{\ddot{a}_{\overline{m}|}} \lim_{t\to\infty} L_t, \tag{18}$$

$$\lim_{t\to\infty} S_t = \frac{m}{\ddot{a}_{\overline{m}|}} \lim_{t\to\infty} L_t + (v_A - v_L)AL. \tag{19}$$

The only differences between (17)-(19) and the results of Dufresne (1989) are that the annuities are valued at rate i_A here and there is an explicit term for the difference between i_A and i_L in equation (19). Equations (17)-(19) follow from equations (16), (15) and (13). (Recall that $U_t = 0$ for $t \geq n$ from equation (11) since the initial unfunded liability is amortized over a finite period n.)

COROLLARY 1. *Assume that* $\left|i - i_A\right| \sum_{j=0}^{m-1}\left(\ddot{a}_{\overline{m-j}|} - 1\right)/\ddot{a}_{\overline{m}|} < 1$.
If $i_A = i$, *then* $\lim UL_t = 0$. *If* $i_A > i$, *then* $\lim UL_t > 0$. *If* $i_A < i$, *then* $\lim UL_t < 0$.

Corollary 1 confirms the observations made in section 2: if the actuarial investment return assumption is optimistic (that is, $i_A > i$), then a persistent deficit occurs ($\lim UL_t > 0$); on the other hand, if the investment return assumption is conservative (that is, $i_A < i$), then a persistent surplus occurs ($\lim UL_t < 0$). Note also that, if $i_A \neq i$, $\lim UL_t$ depends on the period m over which gains and losses are amortized.

5. Spreading Gains and Losses

Dufresne (1988) discusses another funding method that is used to determine contributions. This method is widely used in the United Kingdom and is also implicit in actuarial cost methods such as the Aggregate and Frozen Initial Liability methods (Trowbridge and Farr, 1976, p. 85). The equations in Dufresne (1988) may also be extended to allow for the distinction between the rate at which liabilities are discounted and the investment return assumption, as well as for the separate treatment of the initial unfunded liability.

The supplementary contribution paid in year $(t, t+1)$ is

$$S_t = \sum_{j=0}^{\infty}(1-K)\,K^j u_A^j L_{t-j} + (v_A - v_L)\,AL + P_t, \tag{20}$$

where $0 \leq K < v_A$. In this alternative method, a unit loss is paid off by means of a sequence of exponentially declining payments, $\{(1-K)K^j u_A^j,\ j = 0, 1, \ldots\}$, the unit loss being paid off in perpetuity since $\sum_{j=0}^{\infty}(1-K)\,K^j u_A^j \cdot v_A^j = 1$. The larger the parameter K, the slower the loss is paid off. The loss is never completely defrayed, except in the limit as $t \to \infty$, but Trowbridge and Farr (1976) point out that this is not a weakness as intervaluation losses occur randomly in practice and are never completely removed. This funding method is commonly referred to as "spreading" gains and losses, by contrast with the method in section 4 which involves amortizing gains and losses (McGill *et al.*, 1996, p. 525; Berin, 1989, p. 18; Dufresne, 1988).

Replacing S_t from equation (20) into equation (9) and using equation (12) yields

$$(UL_{t+1} - U_{t+1}) - u_A(UL_t - U_t) = L_{t+1} - u_A \sum_{j=0}^{\infty}(1-K)\,K^j u_A^j L_{t-j}. \tag{21}$$

Recall that $L_t = 0$ for $t \leq 0$, $UL_t = 0$ for $t < 0$, and $UL_0 = U_0$. It is easily verified, from equation (21), that

$$UL_t - U_t = \sum_{j=0}^{\infty} K^j u_A^j L_{t-j}. \tag{22}$$

Compare equation (15) when losses are amortized to equation (22) when losses are spread.

Equation (22) is sensible since, for a unit loss that emerged j years ago, the following sequence of payments is outstanding: $\{(1-K)K^l u_A^l,\ l = j, j+1, \ldots\}$.

The present value of these payments is $\sum_{l=j}^{\infty}(1-K)K^l u_A^l \cdot v_A^{l-j} = K^j u_A^j$. Equation (22) thus shows that, at any time t, the unfunded liability is the present value of payments yet to be made in respect of all past and present losses, together with the unamortized part of the initial unfunded liability.

The supplementary contribution S_t in this method may be calculated directly as a proportion $1 - K$ of the unfunded liability, together with an adjustment for the difference between assumed rates on assets and liabilities and for the separate amortization of the initial unfunded liability. Comparing equations (20) and (22),

$$S_t = (1-K)\left(UL_t - U_t\right) + \left(v_A - v_L\right)AL + P_t. \tag{23}$$

For simplicity, Dufresne (1988) disregards the separate treatment of initial unfunded liability and the distinction between i_A and i_L and considers only $S_t = (1 - K)UL_t$. Dufresne (1988) also states that the parameter K is usually calculated as $K = 1 - 1/\ddot{a}_{\overline{M}|}$. M is typically between 1 and 10 years in the United Kingdom. Thus, if $M = m$, the first payment made in respect of a unit loss is $1/\ddot{a}_{\overline{m}|}$ under both the amortization and spreading funding methods (equations (13) and (20) respectively).

Replace S_t from equation (20) and UL_t from equation (22) into equation (8), and use equation (12), to obtain:

$$L_t = \left(i - i_A\right)\left[\sum_{j=0}^{\infty} K^{j+1}u_A^j L_{t-j} - v_A\left(AL - U_{t+1}\right)\right]. \tag{24}$$

Compare equation (16) when losses are amortized to equation (24) when losses are spread. If the actuarial assumption as to the rate of investment return on plan assets equals the actual rate of return (that is, if $i = i_A$), then no loss emerges ($L_t = 0\ \forall\ t$ from equation (24)) and the unfunded liability consists only of the unamortized part of the initial unfunded liability ($UL_t = U_t$ for $t \geq 0$ from equation (22)).

From equation (24),

$$L_{t+1} - u_A K L_t = \left(i - i_A\right)\left[KL_t - v_A\left(AL - U_{t+1}\right) + K\left(AL - U_t\right)\right], \tag{25}$$

which is a first-order linear difference equation that simplifies to

$$L_{t+1} - uK\, L_t = -\,v_A\left(i - i_A\right)\left[\left(AL - U_{t+1}\right) - u_A K\left(AL - U_t\right)\right]. \tag{26}$$

Recall from equation (11) that $U_t = 0$ for $t \geq n$. Provided $|uK| < 1$, it follows from equation (26) that

$$\lim_{t\to\infty} L_t = -\,AL\left(i - i_A\right)v_A \frac{1 - u_A K}{1 - uK}. \tag{27}$$

In equation (20), K was defined to be such that $0 \leq K < v_A$. Provided $|uK| < 1$, the right hand side of equation (22) is also absolutely convergent and

$$\lim_{t\to\infty} UL_t = (1-u_A K)^{-1} \lim_{t\to\infty} L_t. \tag{28}$$

$\lim S_t$ may be found from equations (23) and (28). This is summarised in the following result.

RESULT 2 *Provided* $|uK| < 1$,

$$\lim_{t\to\infty} L_t = -AL(i-i_A)v\frac{v_A-K}{v-K}, \tag{29}$$

$$\lim_{t\to\infty} UL_t = AL\frac{v-v_A}{v-K}, \tag{30}$$

$$\lim_{t\to\infty} S_t = AL(1-K)\frac{v-v_A}{v-K} + AL(v_A-v_L). \tag{31}$$

In contrast with Dufresne (1988), we have allowed for separate amortization of the initial unfunded liability and also for the possibility that the actuarial assumptions i_A and i_L are different, and we have also derived equations pertaining to the intervaluation loss L_t. Result 2 may alternatively be obtained, as in Dufresne (1988), by substituting S_t from equation (23) into equation (5) giving a first-order difference equation

$$(UL_{t+1}-U_{t+1}) - uK(UL_t-U_t) = (1-uv_A)(AL-U_{t+1}), \tag{32}$$

which solves to

$$UL_t - U_t = (1-uv_A)\sum_{j=0}^{t-1}(uK)^j\left(AL-U_{t-j}\right). \tag{33}$$

Corollary 2 hereunder follows directly from equation (30):

COROLLARY 2 *Assume that* $|uK| < 1$. *If* $i_A = i$, *then* $\lim UL_t = 0$. *If* $i_A > i$, *then* $\lim UL_t > 0$. *If* $i_A < i$, *then* $\lim UL_t < 0$.

Compare Corollary 1 with Corollary 2. Under both amortization and spreading, the choice of the actuarial investment return assumption i_A affects the long-term funding status of the pension plan. Note also from equation (30) that, when $i_A \neq i$, $\lim UL_t$ depends on the parameter K that is used to spread gains and losses.

6. MODIFIED SPREADING OF GAINS AND LOSSES

If the actual investment return deviates from the actuarial investment return assumption, then persistent underfunding or overfunding will occur in the long term, as shown in Corollaries 1 and 2 in both of the preceding methods. Persistent deficits jeopardize the security of pension benefits for plan members since, in the event of sponsor insolvency, there will not be enough funds to meet benefit obligations. On the other hand, excessive surpluses are also undesirable as funds are being diverted from productive activity in the company. Plan participants may also demand that surpluses be distributed to them in the form of improved benefits (McGill *et al.*, 1996, p. 592-4).

In practice, the emergence of persistent surpluses or deficits causes actuaries to revise their actuarial valuation assumptions. Nevertheless, it is of interest to devise a funding method that avoids systematic surpluses and deficits.

Suppose that a constant stream of intervaluation losses of size $\ell > 0$ occurs in the pension plan. If losses are being amortized as in the method of section 4, then a positive unfunded liability (that is, a deficit) occurs since, from equation (15) and for $t \geq n$,

$$UL_t = \ell \sum_{j=0}^{m-1} \frac{\ddot{a}_{\overline{m-j}|}}{\ddot{a}_{\overline{m}|}} > 0. \tag{34}$$

Likewise, a deficit occurs if losses are being spread, as in section 5, since, from equation (22) and for $t \geq n$,

$$UL_t = \ell / (1 - u_A K) > 0. \tag{35}$$

This suggests a variation on the spreading of losses. Consider a new funding method, which is referred to henceforth as "modified spreading of gains and losses", where supplementary contributions are calculated to pay off intervaluation losses and the initial unfunded liability as follows:

$$S_t = \sum_{j=0}^{\infty} \left(\alpha_1 K_1^j - \alpha_2 K_2^j\right) u_A^j L_{t-j} + (v_A - v_L) AL + P_t \tag{36}$$

where

$$\alpha_1 = (1 - u_A K_1)(1 - K_1) / u_A (K_2 - K_1), \tag{37}$$

$$\alpha_2 = (1 - u_A K_2)(1 - K_2) / u_A (K_2 - K_1), \tag{38}$$

and where $0 \leq K_1 < v_A$ and $0 \leq K_2 < v_A$ and $K_1 \neq K_2$.

In this method, a unit loss is liquidated by means of an infinite sequence of payments $\{(\alpha_1 K_1^j - \alpha_2 K_2^j) u_A^j,\ j = 0, 1, \ldots\}$ and is paid off in perpetuity since

$$\sum_{j=0}^{\infty} \left(\alpha_1 K_1^j - \alpha_2 K_2^j\right) u_A^j \cdot v_A^j = \frac{\alpha_1}{1 - K_1} - \frac{\alpha_2}{1 - K_2} = 1. \tag{39}$$

Replacing S_t from equation (36) into equation (9) and using equation (12) yields

$$(UL_{t+1} - U_{t+1}) - u_A (UL_t - U_t) = L_{t+1} - u_A \sum_{j=0}^{\infty} \left(\alpha_1 K_1^j - \alpha_2 K_2^j\right) u_A^j L_{t-j.} \tag{40}$$

Now define

$$\beta_1 = (1 - u_A K_1) / u_A (K_2 - K_1), \tag{41}$$

$$\beta_2 = (1 - u_A K_2) / u_A (K_2 - K_1). \tag{42}$$

Noting that $\alpha_1 = \beta_1(1 - K_1)$ and $\alpha_2 = \beta_2(1 - K_2)$ and $\beta_1 - \beta_2 = 1$, the right hand side of equation (40) may be rewritten as

$$\begin{aligned} L_{t+1} &+ u_A \sum_{j=0}^{\infty} \left(\beta_1 K_1^{j+1} - \beta_2 K_2^{j+1}\right) u_A^j L_{t-j} - u_A \sum_{j=0}^{\infty} \left(\beta_1 K_1^j - \beta_2 K_2^j\right) u_A^j L_{t-j} \\ &= u_A \sum_{j=-1}^{\infty} \left(\beta_1 K_1^{j+1} - \beta_2 K_2^{j+1}\right) u_A^j L_{t-j} - u_A \sum_{j=0}^{\infty} \left(\beta_1 K_1^j - \beta_2 K_2^j\right) u_A^j L_{t-j} \\ &= \sum_{j=0}^{\infty} \left(\beta_1 K_1^j - \beta_2 K_2^j\right) u_A^j L_{t+1-j} - u_A \sum_{j=0}^{\infty} \left(\beta_1 K_1^j - \beta_2 K_2^j\right) u_A^j L_{t-j}, \end{aligned} \tag{43}$$

which, upon comparison with the left hand side of equation (40), yields

$$UL_t - U_t = \sum_{j=0}^{\infty} \left(\beta_1 K_1^j - \beta_2 K_2^j\right) u_A^j L_{t-j}. \tag{44}$$

Compare equations (15), (22) and (44).

Under the method of equation (36), for a unit loss that emerged j years ago, the following sequence of payments is yet to be made: $\{(\alpha_1 K_1^l - \alpha_2 K_2^l) u_A^l, l = j, j+1, \ldots\}$. The present value of these outstanding payments is therefore

$$\sum_{l=j}^{\infty} \left(\alpha_1 K_1^l - \alpha_2 K_2^l\right) u_A^l \cdot v_A^{l-j} = \left[\frac{\alpha_1 K_1^j}{1 - K_1} - \frac{\alpha_2 K_2^j}{1 - K_2}\right] u_A^j = \left(\beta_1 K_1^j - \beta_2 K_2^j\right) u_A^j. \tag{45}$$

Equation (44) thus shows that, at any time t, the unfunded liability is the present value of payments yet to be made in respect of all past and present losses, together with the unamortized part of the initial unfunded liability.

The following proposition is proven in the Appendix.

PROPOSITION 1 *Provided that*

$$\min(i, i_A) > -100\%, \tag{46}$$

$$i - i_A < 100\% + i_A, \tag{47}$$

$$0 \le \min(K_1, K_2) < \max(K_1, K_2) < \min(v, v_A), \tag{48}$$

then

$$\lim_{t \to \infty} L_t = -AL(i - i_A)v, \tag{49}$$

$$\lim_{t \to \infty} UL_t = 0, \tag{50}$$

$$\lim_{t \to \infty} S_t = AL(v - v_L). \tag{51}$$

The sufficient conditions (46)-(48) in Proposition 1 are not very restrictive. (Necessary and sufficient conditions are discussed in the Appendix.) Condition (46) is easily satisfied under normal economic conditions. Condition (47) also holds in practice. Long-run economic growth means that the actuarial assumption i_A as to the long-term rate of return on plan assets is positive ($i_A > 0$). Condition (47) then requires that the actuarial investment return assumption i_A does not underestimate the actual return on assets i by 100% or more. Condition (48) is also easily met in practice. For example, if $\max(i, i_A) = 15\%$, then $0 \leq K_1 < 0.87$ and $0 \leq K_2 < 0.87$ with $K_1 \neq K_2$ means that condition (48) holds.

COROLLARY 3 *Assume that conditions* (46)-(48) *hold. Then,* $\lim UL_t = 0$, *irrespective of whether* $i_A = i$ *or* $i_A > i$ *or* $i_A < i$.

Compare Corollaries 1, 2 and 3. Corollary 3 states that, under the modified spreading funding method described by equation (36), the pension plan is fully funded in the long term, irrespective of the deviation of the investment return assumption from the actual return on the pension plan assets (provided that the mild conditions (46)-(48) hold). Furthermore, $\lim UL_t$ is independent of the funding method parameters K_1 and K_2.

The choice of i_A affects the progression of funding in the short term, but i_A does not affect the funding position asymptotically. In fact, one could arbitrarily set $i_A = i_L$ as under the classical actuarial valuation methodology described in section 2 and effectively dispense with an investment return assumption i_A that is distinct from the rate i_L at which the pension liability is valued.

Corollary 3 may be explained as follows. Suppose that a constant stream of intervaluation losses of size $\ell = 0$ occurs in the pension plan. Recall that this results in a persisting deficit when losses are being either amortized or spread: see equations (34) and (35) respectively. By contrast, under the method of equation (36), a constant stream of losses of size $\ell \neq 0$ results in zero unfunded liability because, from equation (44) and for $t \geq n$,

$$UL_t = \ell \sum_{j=0}^{\infty} \left(\beta_1 K_1^j - \beta_2 K_2^j\right) u_A^j = \ell \left[\frac{\beta_1}{1 - u_A K_1} - \frac{\beta_2}{1 - u_A K_2}\right] = 0 \tag{52}$$

where we use equations (41) and (42).

It was shown that the spreading method of equation (20) could be calculated more directly in terms of the unfunded liability, in equation (23). This may also be achieved here. The following proposition is proven in the Appendix.

PROPOSITION 2 *The funding method described in equation* (36) *is equivalently achieved by calculating supplementary contributions as follows:*

$$S_t = \lambda_1 \left(UL_t - U_t\right) + \lambda_2 \sum_{j=0}^{\infty} \left(UL_{t-j} - U_{t-j}\right) + \left(v_A - v_L\right) AL + P_t, \qquad (53)$$

where $\lambda_1 = 1 - u_A K_1 K_2$ *and* $\lambda_2 = v_A (1 - u_A K_1)(1 - u_A K_2)$.

Trowbridge and Farr (1976, p. 62) state that "easy computations" are a desirable characteristic of a funding method. Equation (53) provides a straightforward way of computing contributions from year to year as only the historic sum of unfunded liabilities need be stored and updated.

Compare equations (23) and (53). The second term on the right hand side of equation (53) represents a historic sum (without interest) of past unfunded liabilities. Contributions are therefore paid until surpluses and deficits cancel each other out and the unfunded liability is zero. Modified spreading of gains and losses, in the representation of equation (53), is similar to a method described by Balzer (1982) in the context of a general insurance system (see also Taylor, 1987). Balzer (1982) refers to a summation term similar to the second term on the right hand side of equation (53) as supplying an "integral action" which adjusts for a "persisting stream of unpredicted claims".

7. NUMERICAL EXAMPLE

An illustration of the previous results is given here and is based on the following:

Demographic projections: Mortality: English Life Table No. 12 (males). Plan population: stationary with single entry age of 20 and single retirement age of 65.
Salary: Constant throughout working lifetime.
Benefit: A level pension at age 65 paying 2/3 of annual salary.
Economic projections: No inflation. Assets earn a constant rate of return of 4.5%.
Initial unfunded liability: Zero. (Alternatively, assume that UL_0 is being separately amortized as in equation (10) and that $UL_t - U_t$, rather than UL_t is evaluated below.)
Actuarial valuations: Frequency: yearly. Actuarial cost method: unit credit.
Actuarial assumptions: Fixed with valuation assumptions $i_L = 4\%$, $i_A = 1\%$, 4.5% and 6%. Other valuation assumptions are identical to projection assumptions.
Valuation data: Number of entrants and payroll are calculated such that the yearly benefit outgo B is normalized to 1. Actuarial liability $AL = 16.94$, normal cost $NC = 0.3486$, both expressed as a proportion of B.
Funding method parameters: Amortization: $m = 5$. Spreading: $K = 1 - 1/\ddot{a}_{\overline{5|}}$. Modified spreading: $K_1 = K$, $K_2 = 0.8$.

When $i = i_A = 4.5\%$, numerical work (not shown here) shows that neither gain nor loss arises and the funded ratio (that is, ratio of fund value to actuarial

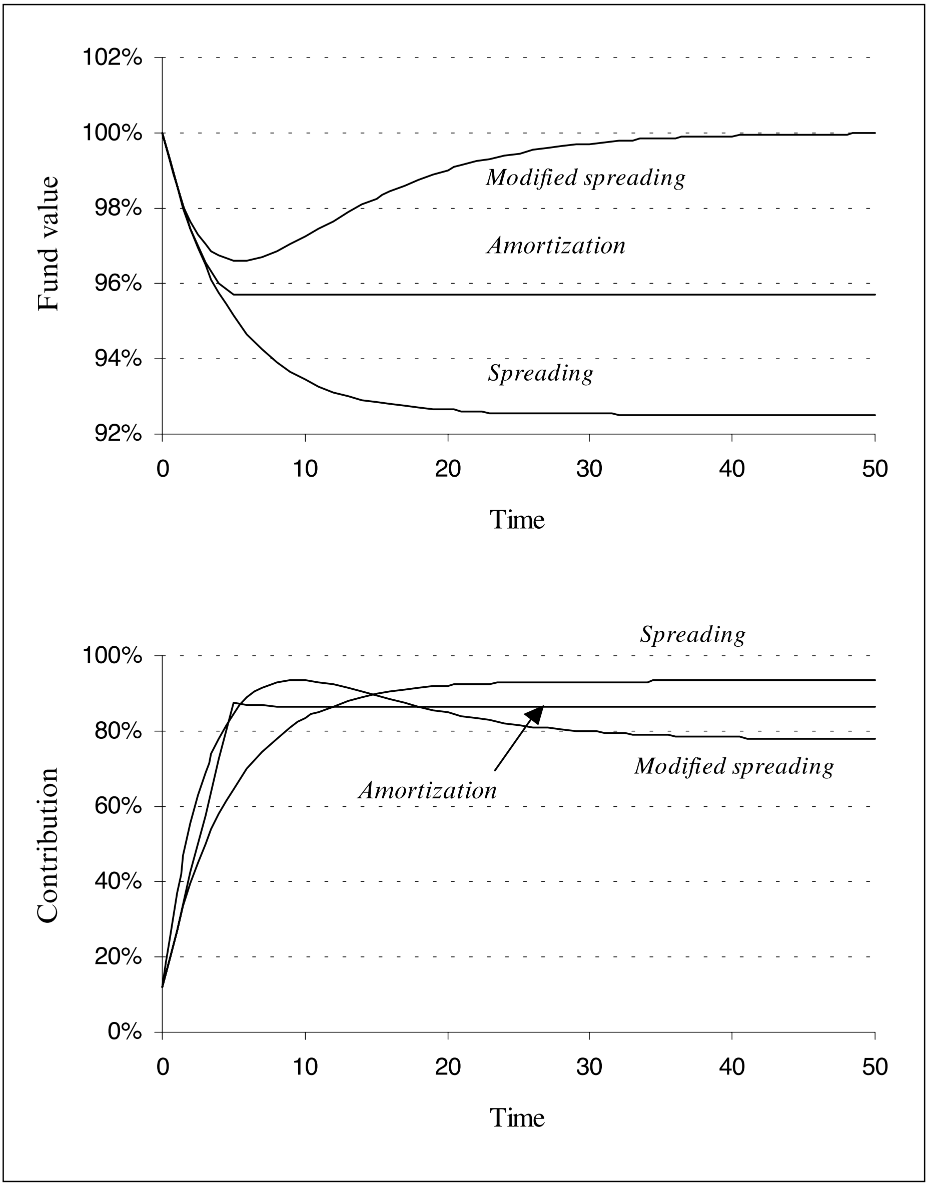


Figure 1: Fund value (per cent of actuarial liability) and contribution (per cent of normal cost) against time (years) when $i_A = 6\%$ and $i = 4.5\%$ for amortization, spreading and modified spreading.

liability) remains at 100%, while the contribution paid is equal to the normal cost, for all three methods. This accords with Corollaries 1, 2 and 3 when $i_A = i$.

When $i = 4.5\%$ and $i_A = 6\%$, the investment return assumption is optimistic. Fund values (as a percentage of actuarial liability) and contributions (as a percentage of normal cost) over time are exhibited in Table 1. See also Figure 1.

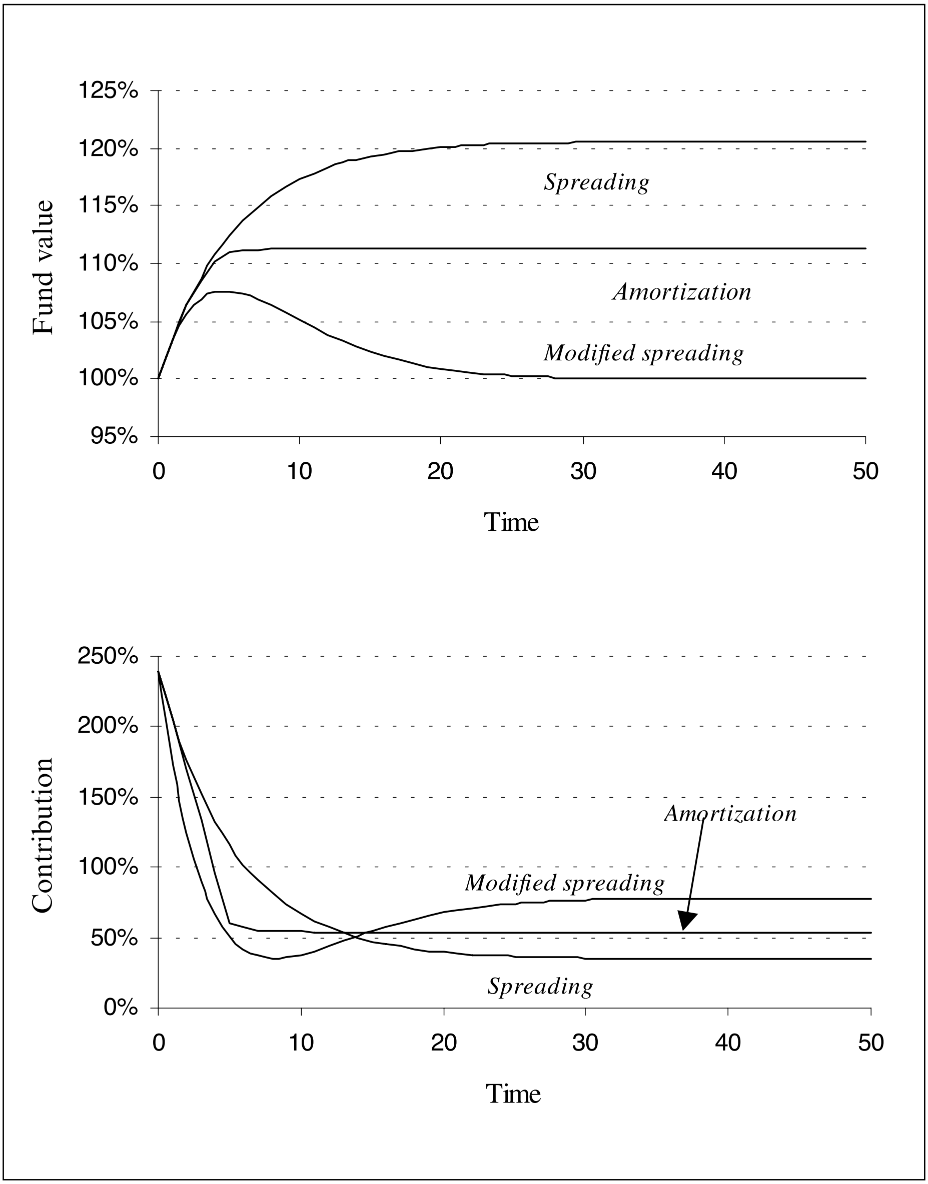


Figure 2: Fund value (per cent of actuarial liability) and contribution (per cent of normal cost) against time (years) when $i_A = 1\%$ and $i = 4.5\%$ for amortization, spreading and modified spreading.

A contribution that is equal to 11.8% of normal cost is required initially under all three methods. Under amortization, the required contribution levels off at 86.6% of normal cost and an unfunded liability of 4.3% of actuarial liability remains. Under spreading, the contribution rises steadily to 93.3% of normal cost and an unfunded liability of 7.5% of actuarial liability is left eventually.

TABLE 1

FUND VALUE (PER CENT OF ACTUARIAL LIABILITY) AND CONTRIBUTION (PER CENT OF NORMAL COST) WHEN $i_A = 6\%$ AND $i = 4.5\%$ FOR AMORTIZATION (A), SPREADING (S) AND MODIFIED SPREADING (MS).

Time	Fund value (%)			Contribution (%)		
	A	S	MS	A	S	MS
0	100.0	100.0	100.0	11.8	11.8	11.8
2	97.4	97.4	97.6	42.5	39.7	55.6
4	96.0	95.8	96.7	72.5	58.1	78.2
6	95.7	94.6	96.6	87.0	70.1	88.8
8	95.7	93.9	96.8	86.6	78.1	92.9
10	95.7	93.4	97.2	86.6	83.3	93.4
12	95.7	93.1	97.7	86.6	86.7	92.3
14	95.7	92.9	98.1	86.6	89.0	90.5
16	95.7	92.8	98.4	86.6	90.5	88.4
18	95.7	92.7	98.8	86.6	91.4	86.5
20	95.7	92.6	99.0	86.6	92.1	84.8
25	95.7	92.6	99.5	86.6	92.9	81.7
30	95.7	92.5	99.7	86.6	93.2	79.8
35	95.7	92.5	99.8	86.6	93.3	78.8
40	95.7	92.5	99.9	86.6	93.3	78.3
45	95.7	92.5	100.0	86.6	93.3	78.0
50	95.7	92.5	100.0	86.6	93.3	77.8

TABLE 2

FUND VALUE (PER CENT OF ACTUARIAL LIABILITY) AND CONTRIBUTION (PER CENT OF NORMAL COST) WHEN $i_A = 1\%$ AND $i = 4.5\%$ FOR AMORTIZATION (A), SPREADING (S) AND MODIFIED SPREADING (MS).

Time	Fund value (%)			Contribution (%)		
	A	S	MS	A	S	MS
0	100.0	100.0	100.0	238.8	238.8	238.8
2	106.3	106.3	105.7	169.1	175.9	124.1
4	110.2	110.7	107.5	96.5	132.3	66.3
6	111.2	113.8	107.3	57.1	102.2	41.7
8	111.2	115.9	106.3	54.6	81.3	35.2
10	111.3	117.3	105.1	54.1	66.9	37.7
12	111.3	118.3	103.8	54.1	56.9	44.0
14	111.3	119.0	102.8	54.1	50.0	51.2
16	111.3	119.5	102.0	54.1	45.3	57.8
18	111.3	119.9	101.3	54.1	42.0	63.4
20	111.3	120.1	100.9	54.1	39.7	67.7
25	111.3	120.4	100.3	54.1	36.6	74.2
30	111.3	120.5	100.1	54.1	35.3	76.7
35	111.3	120.6	100.0	54.1	34.9	77.5
40	111.3	120.6	100.0	54.1	34.7	77.7
45	111.3	120.6	100.0	54.1	34.6	77.7
50	111.3	120.6	100.0	54.1	34.5	77.7

Under modified spreading, the required contribution stabilizes at about 78% of normal cost with the plan being fully funded eventually. This therefore agrees with Corollaries 1, 2 and 3 when $i_A > i$: long-run deficits occur under amortization and spreading, but not under modified spreading. Furthermore, numerical experiments suggest that the long-run unfunded liabilities that occur under amortization and spreading are larger, the larger the deviation between actual and assumed returns.

Note that the pension fund is ultimately in balance under all three methods. For example, under amortization, using units of yearly benefit outgo, a fund of 95.7% × 16.94 = 16.21 yields investment income of 16.21 × 4.5% = 0.7295 at the end of the year. At the start of the year, the present value of this income is 0.7295 / 1.045 = 0.698. Contribution income is 86.6% × 0.3486 = 0.301. Total income is 0.698 + 0.301 = 1 which balances the benefit of 1 that is paid out. The balance occurs at different levels under the three methods. Under modified spreading, the fund is eventually in equilibrium in such a way that the pension plan is fully funded.

When $i = 4.5\%$ and $i_A = 1\%$, a conservative investment return assumption is being made. See Table 2 and Figure 2. A large contribution (more than double the normal cost) is required initially under all three funding methods. Intervaluation gains lead initially to falling contributions under all three methods (at about the same rate). Ultimately, the lowest contribution (at only 35% of normal cost) is generated when spreading is used, but this is at the expense of a large surplus in the pension fund of 20% of actuarial liability. On the other hand, the surplus is only 5% of actuarial liability within 10 years, and under 1% within 20 years, when modified spreading is used. This also agrees with Corollaries 1, 2 and 3 when $i_A < i$.

8. Conclusion

The investment return assumption made by actuaries when valuing defined benefit pension plans and its relevance to the pace of funding for pension benefits was discussed in section 2. It was argued that this assumption is theoretically distinct from the discount rate that is used to value pension liabilities, although they may be equal in practice. A simplified model pension plan was posited in Section 3, where actuarial liability, normal cost and benefit outgo were constant. The only intervaluation gains and losses allowed in the model resulted from actual investment return deviating from the actuarial investment return assumption.

Two practical funding methods were described in sections 4 and 5 and it was shown, in both cases, that a conservative investment return assumption leads to a long-term surplus whereas an optimistic investment return assumption leads to a long-term deficit. Both long-term surpluses and deficits were deemed to be undesirable. Surpluses may entail expensive demands for benefit enhancements from plan members during wage negotiations and also involves the diversion of capital away from projects within the sponsoring corporation. Deficits may endanger the security of pension benefits should the plan sponsor become insolvent. A funding method was devised and described in section 6 that avoids such persistent surpluses and deficits, under mild stability conditions, independently of the conservatism or optimism in the actuarial investment return assumption. A simple way of implementing this funding method was derived in terms of the historic sum of past unfunded liabilities. A numerical illustration of these results was provided in section 7.

The analysis in this paper yielded closed-form mathematical solutions but this required simplistic modelling assumptions. Future research should relax these restrictive assumptions. First, only asset gains and losses were considered. Mortality, withdrawal, inflation and other factors are also variable and should be incorporated in the model. Second, these factors are uncertain and inter-valuation gains and losses are random. A stochastic approach following Dufresne (1988, 1989) and Owadally and Haberman (1999), who investigate pension funding with random investment returns, should be illuminating. It will enable a more realistic comparison of the various funding methods to be made in terms of the variance of fund values and contributions. The efficient choice of parameters K_1 and K_2 under modified spreading of gains and losses can also then be investigated.

Acknowledgement

The author would like to thank Steve Haberman for his comments on this paper.

Appendix

Proof of Proposition 1

It is easy to show, from equations (37), (38), (41) and (42) that

$$\alpha_1 - \alpha_2 = 1 + v_A - K_1 - K_2, \tag{54}$$

$$\alpha_1 K_2 - \alpha_2 K_1 = v_A - K_1 K_2, \tag{55}$$

$$\beta_1 - \beta_2 = 1, \tag{56}$$

$$\beta_1 K_2 - \beta_2 K_1 = v_A. \tag{57}$$

Replace S_t from equation (36) and UL_t from equation (44) into equation (8), and use equation (12), to obtain:

$$L_{t+1} = (i - i_A)\left[\sum_{j=0}^{\infty}\left[(\beta_1 - \alpha_1) K_1^j - (\beta_2 - \alpha_2) K_2^j\right] u_A^j L_{t-j} - v_A (AL - U_{t+1})\right]. \tag{58}$$

This may be rewritten using the lag or backward shift operator B as follows:

$$B^{-1} L_t = (i - i_A)\left[\frac{\beta_1 - \alpha_1}{1 - u_A K_1 B} L_t - \frac{\beta_2 - \alpha_2}{1 - u_A K_2 B} L_t - B^{-1} v_A (AL - U_t)\right]. \tag{59}$$

Note from equations (54)-(57) that

$$(\beta_1 - \alpha_1) - (\beta_2 - \alpha_2) = K_1 + K_2 - v_A, \tag{60}$$

$$(\beta_1 - \alpha_1) K_2 - (\beta_2 - \alpha_2) K_1 = K_1 K_2. \tag{61}$$

Multiply both sides of equation (59) by $(1 - u_A K_1 B)(1 - u_A K_2 B)B$ and use the two equations above:

$$\begin{aligned}(1 - u_A K_1 B)(1 - u_A K_2 B) L_t = (i - i_A)\big[&(K_1 + K_2 - v_A) BL_t - u_A K_1 K_2 B^2 L_t \\ &- (1 - u_A K_1 B)(1 - u_A K_2 B) v_A (AL - U_t)\big].\end{aligned} \tag{62}$$

Collect terms in L_t on the left hand side to obtain a second order linear difference equation for L_t:

$$\begin{aligned}&\left\{1 - B\left[uK_1 + uK_2 - uv_A + 1\right] + B^2\left[uu_A K_1 K_2\right]\right\} L_t \\ &\qquad = -(1 - u_A K_1 B)(1 - u_A K_2 B) v_A (i - i_A)(AL - U_t).\end{aligned} \tag{63}$$

Difference equation (63) has a quadratic characteristic equation,

$$P(z) = z^2 - z\left[uK_1 + uK_2 - uv_A + 1\right] + uu_A K_1 K_2 = 0, \tag{64}$$

whose roots must be less than one in magnitude for $\{L_t\}$ to converge as $t \to \infty$. Necessary and sufficient conditions for this for a general quadratic equation are given by Marden (1966):

$$|P(0)| < 1 \Rightarrow |uu_A K_1 K_2| < 1, \tag{65}$$

$$P(1) > 0 \Rightarrow uv_A(1 - u_A K_1)(1 - u_A K_2) > 0, \tag{66}$$

$$P(-1) > 0 \Rightarrow uv_A[2u_A(v + u_A K_1 K_2) - (1 - u_A K_1)(1 - u_A K_2)] > 0. \tag{67}$$

It is now shown that inequalities (65)-(67) follow from the sufficient conditions in Proposition 1. Note first that condition (46) may be rewritten as $0 < \min(u, u_A) \leq \max(u, u_A)$. Conditions (46) and (48) thus imply that

$$0 \leq \min(u, u_A)\min(K_1, K_2) < \max(u, u_A)\max(K_1, K_2) < 1. \tag{68}$$

Hence, inequality (65) follows from sufficient conditions (46) and (48).

Next, note from the inequalities (68) that

$$0 < 1 - \max(u, u_A)\max(K_1, K_2) < 1 - \min(u, u_A)\min(K_1, K_2) \leq 1, \tag{69}$$

and therefore that

$$0 < (1 - u_A K_1) \leq 1 \quad \text{and} \quad 0 < (1 - u_A K_2) \leq 1. \tag{70}$$

Hence, inequality (66) follows from sufficient conditions (46) and (48).

Finally, condition (47) may be written as $u < 2u_A$ or $2u_A v > 1$, by virtue of condition (46). It follows from inequalities (70) that

$$(1 - u_A K_1)(1 - u_A K_2) \leq 1 < 2u_A v \leq 2u_A v + 2u_A^2 K_1 K_2 \tag{71}$$

$$\Rightarrow 2u_A(v + u_A K_1 K_2) - (1 - u_A K_1)(1 - u_A K_2) > 0. \tag{72}$$

Hence, inequality (67) follows from sufficient conditions (46), (47) and (48).

Let the roots of the characteristic equation (64) be ν_1 and ν_2. If $\nu_1 \neq \nu_2$, L_t in equation (63) has a solution of the form $L_t = A\nu_1^t + B\nu_2^t + L$, where A, B, $L \in \mathbb{R}$. If sufficient conditions (46)-(48) hold, then $|\nu_1| < 1$ and $|\nu_2| < 1$, the sequence $\{L_t\}$ converges to L and, furthermore, the series $\sum_{j=0}^{\infty}(L_j - L)$ is absolutely convergent.

Assuming convergence, it is clear from equation (63) that

$$\begin{aligned} L = \lim_{t\to\infty} L_t &= \frac{-(1 - u_A K_1)(1 - u_A K_2) v_A (i - i_A) AL}{1 - [uK_1 + uK_2 - uv_A + 1] + [uu_A K_1 K_2]} \\ &= -AL(i - i_A) v \frac{v_A (1 - u_A K_1)(1 - u_A K_2)}{v_A (1 - u_A K_1)(1 - u_A K_2)} = -AL(i - i_A) v, \end{aligned} \tag{73}$$

which proves equation (49).

The limit in equation (51) is obtained by resorting to equation (36):

$$S_t = \sum_{j=0}^{\infty}\left(\alpha_1 K_1^j - \alpha_2 K_2^j\right)u_A^j\left(L_{t-j} - L\right) + \sum_{j=0}^{\infty}\left(\alpha_1 K_1^j - \alpha_2 K_2^j\right)u_A^j L + \left(v_A - v_L\right)AL + P_t. \tag{74}$$

As $t \to \infty$, the first sum on the right hand side of equation (74) vanishes since both $\sum_{j=0}^{\infty}(\alpha_1 K_1^j - \alpha_2 K_2^j)u_A^j$ and $\sum_{j=0}^{\infty}(L_j - L)$ are absolutely convergent and their Cauchy product is also absolutely convergent. As $t \to \infty$, the second sum on the right hand side of equation (74) converges to

$$\left(\frac{\alpha_1}{1-u_A K_1} - \frac{\alpha_2}{1-u_A K_2}\right) \times L = v_A \times -AL\left(i - i_A\right)v = -AL\left(v_A - v\right), \tag{75}$$

where use is made of equations (37) and (38). P_t also vanishes as $t \to \infty$ from equation (10). Hence, $\lim_{t\to\infty} S_t = -AL(v_A - v) + AL(v_A - v_L) = AL(v - v_L)$.

Finally, the limit in equation (50) is obtained by taking limits on each term on the right hand side of equation (44) which may be rewritten as follows:

$$UL_t - U_t = \sum_{j=0}^{\infty}\left(\beta_1 K_1^j - \beta_2 K_2^j\right)u_A^j\left(L_{t-j} - L\right) + \sum_{j=0}^{\infty}\left(\beta_1 K_1^j - \beta_2 K_2^j\right)u_A^j L. \tag{76}$$

As $t \to \infty$, the first sum on the right hand side of equation (76) vanishes since both $\sum_{j=0}^{\infty}(\beta_1 K_1^j - \beta_2 K_2^j)u_A^j$ and $\sum_{j=0}^{\infty}(L_j - L)$ are absolutely convergent and their Cauchy product is also absolutely convergent. As $t \to \infty$, the second sum on the right hand side of equation (76) converges to zero since

$$\sum_{j=0}^{\infty}\left(\beta_1 K_1^j - \beta_2 K_2^j\right)u_A^j = \frac{\beta_1}{1-u_A K_1} - \frac{\beta_2}{1-u_A K_2} = 0, \tag{77}$$

where use is made of equations (41) and (42). Hence, $\lim_{t\to\infty} UL_t = 0$. □

Proof of Proposition 2

Rewrite equation (44) in terms of the lag or backward shift operator B:

$$\begin{aligned} UL_t - U_t &= \left[\frac{\beta_1}{1-u_A K_1 B} - \frac{\beta_2}{1-u_A K_2 B}\right] L_t \\ &= \frac{(\beta_1 - \beta_2) - (\beta_1 K_2 - \beta_2 K_1)u_A B}{(1-u_A K_1 B)(1-u_A K_2 B)} L_t. \end{aligned} \tag{78}$$

Using equations (56) and (57), the numerator on the right hand side of the above equation simplifies and

$$UL_t - U_t = \frac{1-B}{(1-u_A K_1 B)(1-u_A K_2 B)} L_t. \tag{79}$$

Likewise, rewrite equation (36) in terms of the backward shift operator B and use equations (54) and (55) to simplify:

$$\begin{aligned} S_t &= \left[\frac{\alpha_1}{1-u_A K_1 B} - \frac{\alpha_2}{1-u_A K_2 B}\right] L_t + (v_A - v_L) AL + P_t \\ &= \frac{(1+v_A-K_1-K_2)-(v_A-K_1K_2)u_A B}{(1-u_A K_1 B)(1-u_A K_2 B)} L_t + (v_A - v_L) AL + P_t. \end{aligned} \tag{80}$$

Cancel L_t from equations (79) and (80) and simplify:

$$\begin{aligned} &S_t - (v_A - v_L) AL - P_t \\ &= \frac{(1+v_A-K_1-K_2)-(v_A-K_1K_2)u_A B}{1-B} (UL_t - U_t) \\ &= \left[1 - u_A K_1 K_2 + \frac{v_A (1-u_A K_1)(1-u_A K_2)}{1-B}\right] (UL_t - U_t) \\ &= (1-u_A K_1 K_2)(UL_t - U_t) + v_A (1-u_A K_1)(1-u_A K_2) \sum_{j=0}^{\infty} \left(UL_{t-j} - U_{t-j}\right), \end{aligned} \tag{81}$$

which is equation (53). □

References

Actuarial Standards Board (1996) *Actuarial Standard of Practice No. 27: Selection of Economic Assumptions for Measuring Pension Obligations*. Pensions Committee of the Actuarial Standards Board, American Academy of Actuaries, Washington, DC.

Balzer, L.A. (1982) Control of insurance systems with delayed profit/loss-sharing feedback and persisting unpredicted claims. *Journal of the Institute of Actuaries* **109**, 285-316.

Berin, B.N. (1989) *The Fundamentals of Pension Mathematics*. Society of Actuaries, Schaumburg, Illinois.

Daykin, C.D. (1976) Long-term rates of interest in the valuation of a pension fund. *Journal of the Institute of Actuaries* **21**, 286-340.

Dufresne, D. (1988) Moments of pension contributions and fund levels when rates of return are random. *Journal of the Institute of Actuaries* **115**, 535-544.

Dufresne, D. (1989) Stability of pension systems when rates of return are random. *Insurance: Mathematics and Economics* **8**, 71-76.

Marden, M. (1966) *The Geometry of Polynomials*, 2nd ed. American Mathematical Society, Providence, Rhode Island.

McGill, D.M., Brown, K.N., Haley, J.J. and Schieber, S.J. (1996) *Fundamentals of Private Pensions*, 7th ed. University of Pennsylvania Press, Philadelphia, Pennsylvania.

Owadally, M.I. and Haberman, S. (1999) Pension fund dynamics and gains/losses due to random rates of investment return. *North American Actuarial Journal* **3(3)**, 105-117.

Taylor G.C. (1987) Control of unfunded and partially funded systems of payments. *Journal of the Institute of Actuaries* **114**, 371-392.

Thornton, P.N. and Wilson, A.F. (1992) A realistic approach to pension funding. *Journal of the Institute of Actuaries* **119**, 229-312.

TROWBRIDGE, C.L. (1952) Fundamentals of pension funding. *Transactions of the Society of Actuaries* **4**, 17-43.

TROWBRIDGE, C.L. and FARR, C.E. (1976) *The Theory and Practice of Pension Funding*. Richard D. Irwin, Homewood, Illinois.

M. IQBAL OWADALLY
Faculty of Actuarial Science and Statistics
Cass Business School
City University
106 Bunhill Row, London EC1Y 8TZ, England
Email: iqbal@city.ac.uk

CHAIN LADDER BIAS

BY

GREG TAYLOR

ABSTRACT

The chain ladder forecast of outstanding losses is known to be unbiased under suitable assumptions. According to these assumptions, claim payments in any cell of a payment triangle are dependent on those from preceding development years of the same accident year. If all cells are assumed stochastically independent, the forecast is no longer unbiased. Section 5 shows that, under rather general assumptions, it is biased upward. This result is linked to earlier work on some stochastic versions of the chain ladder.

KEYWORDS

Chain ladder, IBNR.

1. INTRODUCTION

The chain ladder (CL) approach to estimation of a loss reserve is well known. It is described, for example, by Taylor (2000).

Its origins are not altogether clear, but it seems likely that it originated as a heuristic device. As such, it may be viewed as a non-parametric estimator. The precise definition is given in Sections 2 and 3.

Kremer (1982) recognised that the CL involved a log-linear cross-classification structure. A number of parametric stochastic versions of the CL developed from this, eg Hertig (1985), Renshaw (1989), Verrall (1989, 1990, 1991).

Mack (1994) pointed out that these stochastic models gave mean estimates of liability that differed from the "classical" CL estimate. While the form of stochastic model underlying the classical CL was speculative, due to the latter's heuristic nature, Mack suggested one. It is distribution free. Details are given in Section 2. Mack also identified the differences between this and the other stochastic models.

Whereas the cross-classified models typically assume stochastic independence of all cells in the data set, the CL (in Mack's formulation) does not. It was shown by Mack (1993) that the algorithm of the classical CL produced unbiased forecasts of liability under its own assumptions.

However, it does not necessarily do so under the alternative assumption of independence between all cells. Some papers have studied the bias in estimates of liability in the parametric cross-classified models mentioned above, but little is known of the bias in the classical CL forecast when all cells are independent.

The purpose of the present paper is to investigate the direction of bias in this case.

2. Framework and notation

Consider a square array X of stochastic quantities $X(i,j) \geq 0$, $i = 0,1,\ldots,I$; $j = 0,1,\ldots,I$.

Denote row sums and column sums as follows:

$$R(i,j) = \sum_{h=0}^{j} X(i,h) \tag{2.1}$$

$$C(i,j) = \sum_{g=0}^{i} X(g,j). \tag{2.2}$$

In addition introduce the following notation for the total sum over a rectangular subset of X:

$$\begin{aligned} T(i,j) &= \sum_{g=0}^{i} \sum_{h=0}^{j} X(g,h) \\ &= \sum_{g=0}^{i} R(g,j) \\ &= \sum_{h=0}^{j} C(i,h) \end{aligned} \tag{2.3}$$

Generally, in the following, any summation of the form $\sum_{a}^{b}$ with $b < a$ will be taken to be zero.

In a typical loss reserving framework, i denotes accident period, j development period, and available data will consist of observations on the triangular subset Δ of X:

$$\Delta = \{X(i,j), i = 0,1,\ldots,I;\ j = 0,1,\ldots,I-i\} \tag{2.4}$$

Figure 2.1 illustrates the situation.

Still in a loss reserving context, Δ would represent some form of claims experience, eg claim counts or claim amounts. The loss reserving problem consists of forecasting the lower triangle in Figure 2.1, conditional on Δ. There is particular interest in forecasting $R(i,I) \mid \Delta$, $i = 1,\ldots,I$. In standard loss reserving parlance, the $X(i,j)$ are usually referred to as **incremental** quantities, or just **increments**, and the $R(i,j)$ as their **cumulative** equivalents.

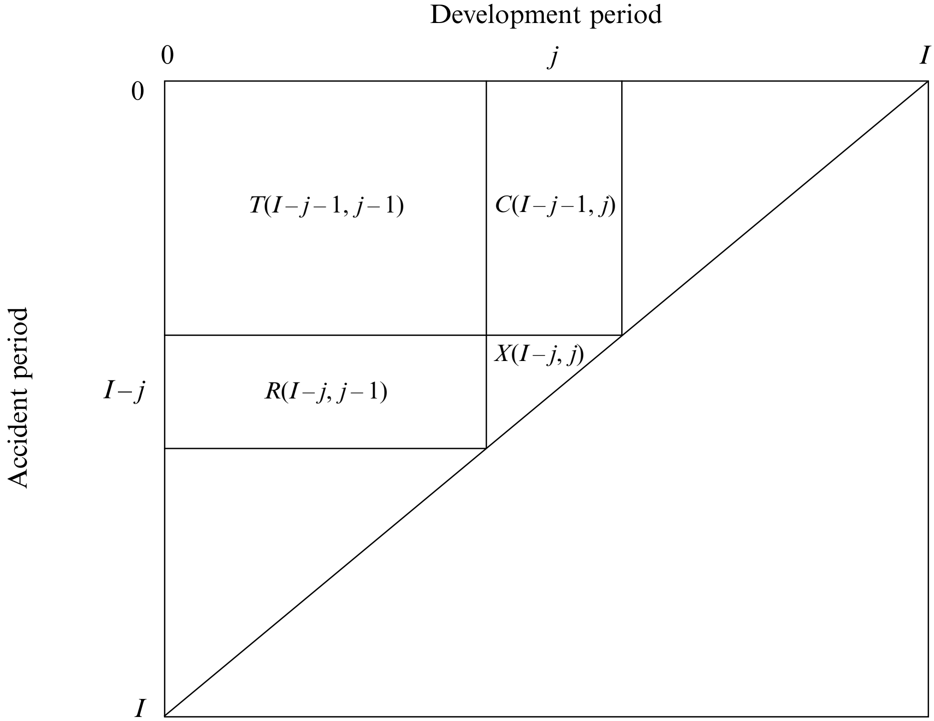


Figure 2.1: Data array

Matrix notation

Analysis of CL bias will proceed by examination of certain derivatives of the CLF defined in Section 3. These depend ultimately on derivatives of the $\hat{v}(k)$. Quantities such as $\partial^2\hat{v}(k)/\partial^2 X(i,j)$ require evaluation, and a matrix notation will be useful for keeping organization among the indexes.

In the following, all vectors will be of dimension $(I+1)$, and all matrices of dimension $(I+1)\times(I+1)$.

Let X denote the matrix $[X(i,j)], i,j=0,1,\ldots,I$. Also let e_k denote the natural basis vector:

$$e_k^T = \Big(0,\ldots,0,\underset{(k)}{1},0,\ldots,0\Big),\ k=0,1,\ldots,I \tag{2.5}$$

and define

$$f_k^T = \sum_{r=0}^{k} e_k^T = \Big(1,\ldots,\underset{(k)}{1},0,\ldots,0\Big) \tag{2.6}$$

where the upper T denotes transposition.

Under this notation, $X(i,j)$ is selected from X as follows:

$$X(i,j) = e_i^T X e_j. \tag{2.7}$$

Similarly,

$$R(i,j) = e_i^T X f_j \tag{2.8}$$

$$C(i,j) = f_i^T X e_j \tag{2.9}$$

$$T(i,j) = f_i^T X f_j. \tag{2.10}$$

3. Chain ladder forecast

Define the **age-to-age factor**

$$\begin{aligned}\hat{\nu}(j) &= T(I-j-1, j+1)/T(I-j-1, j)\\ &= 1 + C(I-j-1, j+1)/T(I-j-1, j)\end{aligned} \tag{3.1}$$

and

$$\hat{R}(i,I) = R(i,I-i)\prod_{k=I-i}^{I-1}\hat{\nu}(k). \tag{3.2}$$

The value of $\hat{R}(i,I)$ calculated in this way will be referred to as the **chain ladder forecast** (CLF) of $R(i,I)$.

4. Chain ladder models

4.1. Independent accident periods

The CLF has been formulated in Section 3 just as an algorithm. No model for the data X has yet been stated.

It is evident that the properties of the CLF will depend on the model. This and the next sub-section consider two alternative models. The first is represented by the following two assumptions.

Assumption 1. The increments of different accident periods are stochastically independent in the sense that

$$Prob[X(i,j) \mid \Delta] = Prob[X(i,j) \mid X(i,0),\ldots,X(i,I-i)].$$

Assumption 2. $E[R(i,j+1) \mid X(i,0),X(i,1),\ldots,X(i,j)] = \nu(j)R(i,j).$ (4.1)

Remark 1. It follows from Assumptions 1 and 2 that

$$E[R(i,j+1) \mid \Delta] = \nu(j) R(i,j) \tag{4.2}$$

for any $j \geq I - i$ (ie future j).

Since $R(i,j+1) = R(i,j) + X(i,j+1)$, one may re-write (4.2) in the form:

$$E[X(i,j+1) \mid \Delta] = [\nu(j) - 1] R(i,j) \tag{4.3}$$

Remark 2. It is clear from (4.3) that $R(i,j)$ and $X(i,j+1)$ may fail to be independent. Generally, under Assumptions 1 and 2, the $X(i,j)$ for fixed i may fail to be independent.

Theorem 1 (Mack). Under Assumptions 1 and 2,

(1) $\hat{\nu}(j)$ is an unbiased estimator of $\nu(j)$ for $j = 0,1,\ldots,I-1$; and
(2) the CLF $\hat{R}(i,I)$ is an unbiased estimator of $E[R(i,I) \mid \Delta]$ for $i = 1,2,\ldots,I$;

provided that both estimators exist.

Proof. See Mack (1993). □

It is also convenient to re-write (4.2) in the form:

$$E\left[R(i,j+1)/R(i,j) \mid \Delta\right] = \nu(j). \tag{4.4}$$

4.2. Independent increments

Replace Assumptions 1 and 2 by 1a, 2a and 3 as follows.

Assumption 1a. All increments $X(i,j)$ are stochastically independent.

Assumption 2a. $E[R(i,j+1)]/E[R(i,j)] = \eta(j)$. (4.5)

The assumption is written in this form in order to relate it to Assumption 2. It is useful to note that an equivalent, and more natural, assumption is that

$$E[X(i,j)] = \alpha(i)\beta(j) \tag{4.5a}$$

for parameters $\alpha(i), \beta(j)$, $i,j = 0,1,\ldots,I$.

Define the set

$$D_i = \{(g,h) : g \leq I-k-1, h \leq k+1,\ k = I-i,\ldots,I-1\} \tag{4.6}$$

and

$$E_i = \{(g,h) : g \leq I-k-1, h \leq k,\ k = I-i,\ldots,I-1\}.$$

Assumption 3. $T(g,h) > 0$ for each $(g,h) \in E_i$.

Remark 3. It is implicit in Assumption 2a that $E[R(i,j)] \neq 0$. By the assumed non-negativity of the $X(i,j)$, $E[R(i,j)] > 0$ for each i, j.

Remark 4. A comparison of (4.4) and (4.5) indicates that $\nu(j)$ and $\eta(j)$ are different quantities (for fixed j) since

$$E\left[R(i,j+1)\,/\,R(i,j)\right] \neq E\left[R(i,j+1)\right]/\,E\left[R(i,j)\right]. \tag{4.7}$$

This fact was pointed out by Mack (1994).

By Assumption 3, applied to (3.1), all $\hat{\nu}(k)$ appearing in (3.2) are defined and strictly positive. Given the assumed non-negativity of the $X(i,j)$, Assumption 3 is both necessary and sufficient for the chain ladder forecast to make sense. Comment on the non-negativity requirement will be made in Section 6.

The conditions of Theorem 1 no longer hold, and so the CLF is not necessarily unbiased.

5. Chain ladder bias

The following is a somewhat technical result, but has been included here rather than in the appendix because the symmetric appearance of rows and columns of X in the second derivative of $Y(i)$ is interesting.

Theorem 2. Define

$$Y(i) = \prod_{k=I-i}^{I-1} \hat{\nu}(k). \tag{5.1}$$

Then

$$\frac{\partial^2 Y(i)}{\partial X^2(g,h)} = 0 \text{ for } (g,h) \notin D_i; \tag{5.2}$$

for $(g,h) \in D_i$ and $h \leq I-i$,

$$\frac{1}{Y(i)}\frac{\partial^2 Y(i)}{\partial X^2(g,h)} = 2\sum_{k=I-i}^{I-g-1} \frac{C(I-k-1,k+1)}{T(I-k-1,k+1)\,T(I-k-1,k)} \text{ x } \left[\frac{1}{T(i-1,I-i)} + \sum_{l=I-i+1}^{k} \frac{R(I-l,l)}{T(I-l-1,l)\,T(I-l,l)}\right] \tag{5.3}$$

for $(g,h) \in D_i$ and $h > I-i$,

$$\frac{1}{Y(i)}\frac{\partial^2 Y(i)}{\partial X^2(g,h)} = 2\sum_{k=h}^{I-g-1}\frac{C(I-k-1,k+1)}{T(I-k-1,k+1)\,T(I-k-1,k)} \times \sum_{l=h}^{k}\frac{R(I-l,l)}{T(I-l-1,l)\,T(I-l,l)}. \tag{5.4}$$

These results depend on Assumption 3 for the existence of (5.3) and (5.4), but do **not** depend on the Assumptions 1a and 2a.

Proof. See appendix. □

Theorem 3. Under Assumptions 1a, 2a and 3, and if $X(g,h)$ is not degenerate for at least one $(g,h) \in D_i$, the CLF $\hat{R}(i,I)$ is biased upward as an estimate of $E[R(i,I)]$ in the sense that

$$E\left[\hat{R}(i,I) \middle| R(i,I-i)\right] > R(i,I-i)\frac{E[R(i,I)]}{E[R(i,I-i)]}. \tag{5.5}$$

Also,

$$E\left[\hat{R}(i,I)\right] > E[R(i,I)]. \tag{5.6}$$

Proof. See appendix. □

Remark 5. An alternative form of (5.6) is:

$$E_{R(i,I-i)}E\left[\hat{R}(i,I) \middle| R(i,I-i)\right] > E[R(i,I)]. \tag{5.7}$$

Note that, in general, (5.5) does **not** imply that

$$E\left[\hat{R}(i,I) \middle| R(i,I-i)\right] > E[R(i,I)].$$

A few words in interpretation of Theorem 3. First define

$$\begin{aligned} L(i,j) &= R(i,I) - R(i,j) \\ &= \sum_{h=j+1}^{I} X(i,h) \end{aligned} \tag{5.8}$$

which is the required loss reserve in respect of accident period i at the end of development period j.

The forecast of $L(i,j)$ associated with the CLF of $R(i,I)$ is

$$\hat{L}(i,j) = \hat{R}(i,I) - R(i,j). \tag{5.9}$$

Now consider the CLF $\hat{L}(i, I-i)$, taken on the last diagonal of Δ, and how it is conditioned by Δ. By (3.2),

$$\hat{L}(i,I-i) = R(i,I-i)\left[\prod_{k=I-i}^{I-1} \hat{\nu}(k) - 1\right] \tag{5.10}$$

ie is proportional to $R(i,I-i)$.

By Theorem 1,

$$E\left[\hat{L}(I-i)\,|\,\Delta\right] = E\left[L(I-i)\,|\,\Delta\right]. \tag{5.11}$$

in the **dependent case**, when Assumptions 1 and 2 hold. The CLF of loss reserve is **conditionally unbiased**.

On the other hand, in the **independent case**, when Assumption 1a holds, $L(i,j)$ and $R(i,j)$ are stochastically independent since they involve disjoint sets of the $X(i,h)$ (see (2.1) and (5.8)). Hence

$$E\left[L(i,I-i)\,|\,\Delta\right] = E\left[L(i,I-i)\right]. \tag{5.12}$$

If Assumptions 2a and 3 also hold, then Theorem 3 may be applied to (5.12) to yield

$$E\left[\hat{L}(i,I-i)\right] > E\left[L(i,I-i)\,|\,\Delta\right] = E\left[L(i,I-i)\right] \tag{5.13}$$

In this case, the CLF of loss reserve, taken **unconditionally**, is **biased upward**. Note that it makes no sense to discuss whether the CLF is conditionally biased in this case. For (5.10) shows that

$$E\left[\hat{L}(i,I-i)\,|\,\Delta\right] \text{ is proportional to } R(i,I-i) \tag{5.14}$$

whereas (5.12) shows that

$$E\left[L(i,I-i)\,|\,\Delta\right] \text{ is independent of } R(i,I-i). \tag{5.15}$$

Then, whether $\hat{L}(i,I-i)$ is conditionally biased upward or downward depends on $R(i,I-i)$.

6. Negative increments

It has been assumed throughout that all $X(i,j) \geq 0$. It is evident that some positivity assumption is required for Theorem 3 to hold. If, for example, one were to apply the theorem to data that were subject to Assumption 3, and then reverse the signs of all $X(i,j)$, one would obtain a **downward** bias for the CLF applied to the modified data.

It is evident that Theorem 3 would continue to hold under weaker assumptions. For example, if the requirement on $X(i,j)$ were weakened to require only that $\text{Prob}[X(i,j) > \delta > 0] > 1 - \varepsilon$ for some $\varepsilon \geq 0$. Then $\text{Prob}[\partial^2 f / \partial X^2(g,h) > 0]$ could be made arbitrarily close to 1 for $(g,h) \in D_i$, so that (A.25) held with strict inequality, and Theorem 3 followed, also with strict inequality.

However, finding necessary conditions on the $X(i,j)$ to ensure that Theorem 3 holds does not appear easy. In considering such conditions, one may write (A.40) in the form:

$$\frac{1}{Y}\frac{\partial^2 Y}{\partial X^2(g,h)} = 2\sum_{k=h}^{I-g-1}\left[\frac{1}{T(I-k-1,k)} - \frac{1}{T(I-k-1,k+1)}\right] \text{x}$$
$$\sum_{l=h}^{k}\left[\frac{1}{T(I-l-1,l)} - \frac{1}{T(I-l,l)}\right]$$

so that, when the $X(i,j)$ need not be non-negative, a sufficient condition for the left side to be positive is that

$$1/T(g+1,\, h) < 1/T(g,h) \tag{6.1}$$

and

$$1/T(g,\, h+1) < 1/T(g,h) \tag{6.2}$$

for all (g,h).

It is tempting then to contemplate stochastic versions of the inequalities, such as

$$E[1/T(g+1,\, h)] < E[1/T(g,h)] \tag{6.3}$$

$$E[1/T(g,\, h+1)] < E[1/T(g,h)] \tag{6.4}$$

or such.

However, the choice of such conditions to lead from Theorem 2 to the proof of (A.25), the key to Theorem 3, is not clear.

7. Conclusion

Theorem 3 shows that, under certain distribution free conditions, **the CLF is biased upward**. A simulation test of prediction bias in the chain ladder and other models was carried out by Stanard (1985). One of his experiments dealt with the case in which the total number of claims in an accident year is a Poisson variate and is multinomially distributed over development years. It may be shown that distinct cells in a row of the claim count triangle are then stochastically independent, and so Theorem 3 applies.

Stanard's simulations did in fact find upward bias in the CLF.

8. Acknowledgements

Professor Klaus Schmidt and Dr Klaus Thomas Hess of Technische Universität Dresden gave generously of their time to improve earlier drafts of this paper. Their thorough review identified one major blunder and produced a number of constructive suggestions, leading to other enhancements.

Dr Alois Gisler read even earlier drafts of this paper and provided a number of valuable comments and insights. The paper possibly would not have occurred without his assistance.

The author also acknowledges helpful comment by Dr Thomas Mack.

References

Hertig, J. (1985) A statistical approach to the IBNR-reserves in marine reinsurance. *ASTIN Bulletin*, **15**, 171-183.

Kallenberg, O. (1997) Foundations of modern probability (2nd ed). Springer.

Kremer, E. (1982) IBNR claims and the two-way model of ANOVA. *Scandinavian Actuarial Journal*, 47-55.

Mack, T. (1993) Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, **23**, 213-221.

Mack, T. (1994) Which stochastic model is underlying the chain ladder method? *Insurance: mathematics and economics*, **15**, 133-138.

Renshaw, A.E. (1989) Chain ladder and interactive modelling (claims reserving and GLIM). *Journal of the Institute of Actuaries*, **116**, 559-587.

Stanard, J.N. (1985) A simulation test of prediction errors of loss reserve estimation techniques. *Proceedings of the Casualty Actuarial Society*, **72**, 124-148.

Taylor, G.C. (1986) Loss reserving in non-life insurance. North-Holland, Amsterdam, Netherlands.

Taylor, G. (2000) Loss reserving: an actuarial perspective. Kluwer Academic Publishers, Dordrecht, Netherlands.

Verrall, R.J. (1989) A state space representation of the chain ladder linear model. *Journal of the Institute of Actuaries*, **116**, 589-609.

Verrall, R.J. (1990) Bayes and empirical bayes estimation for the chain ladder model. *Astin Bulletin*, **20**, 217-243.

Verrall, R.J. (1991) On the estimation of reserves from loglinear models. *Insurance: mathematics and economics*, **10**, 75-80.

Appendix

Proof of theorems

Lemma 1. Suppose that

$$y = \prod_{i=1}^{n} f_i(x) \tag{A.1}$$

with $x = (x_1, \ldots, x_m)^T$ and $f_i(x) > 0$ for each i. Then

$$\frac{1}{y}\frac{\partial^2 y}{\partial x_k^2} = \sum_{i=1}^{n} \frac{1}{f_i}\frac{\partial^2 f_i}{\partial x_k^2} + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \frac{1}{f_i f_j}\frac{\partial f_i}{\partial x_k}\frac{\partial f_j}{\partial x_k}.$$

Lemma 2 (derivatives of age-to-age factors).

$$\frac{1}{\hat{\nu}(k)}\frac{\partial \hat{\nu}(k)}{\partial X(g,h)} = \frac{\delta(h=k+1)}{T(I-k-1,k+1)} - \frac{\delta(g<I-k)\,\delta(h \le k)\,C(I-k-1,k+1)}{T(I-k-1,k)\,T(I-k-1,k+1)} \tag{A.2}$$

where $\delta(.)$ is defined as follows:

$$\begin{aligned} \delta(C) &= 1 \text{ for condition } C \text{ true;} \\ &= 0 \text{ for condition } C \text{ false;} \end{aligned} \tag{A.3}$$

and it is understood that (A.2) applies to **past observations** ($g + h \le I$).

Further,

$$\frac{1}{\hat{\nu}(k)}\frac{\partial^2 \hat{\nu}(k)}{\partial X^2(g,h)} = \frac{2\delta(g<I-k)\,\delta(h \le k)\,C(I-k-1,k+1)}{T^2(I-k-1,k)\,T(I-k-1,k+1)} \tag{A.4}$$

where this result is again understood to apply only to past observations.

Proof. By (2.10),

$$\partial T(i,j) / \partial X(g,h) = f_i^T e_g e_h^T f_j. \tag{A.5}$$

Write the quantity of interest $\hat{\nu}(k)$ in the abbreviated form $\nu = N/D$ where N and D denote numerator and denominator respectively. If N and D depend on variable X, then

$$\begin{aligned} \frac{1}{\nu}\frac{\partial \nu}{\partial X} &= \frac{\partial \log \nu}{\partial X} = \frac{\partial(\log N - \log D)}{\partial X} \\ &= \frac{1}{N}\frac{\partial N}{\partial X} - \frac{1}{D}\frac{\partial D}{\partial X} \\ &= \frac{1}{ND}\left(D\frac{\partial N}{\partial X} - N\frac{\partial D}{\partial X}\right). \end{aligned} \tag{A.6}$$

Now for $\nu = \hat{\nu}(k)$,

$$N = T(I-k-1, k+1) = f_{I-k-1}^T X f_{k+1} \tag{A.7}$$

$$D = T(I-k-1, k) = f_{I-k-1}^T X f_k \tag{A.8}$$

where the matrix notation of Section 2 has been used.

First derivatives of N and D are given by (A.5). Substitution of these into (A.6) yields

$$\begin{aligned} \frac{1}{\hat{\nu}(k)} \frac{\partial \hat{\nu}(k)}{\partial X(g,h)} &= [T(I-k-1,k)\,T(I-k-1,k+1)]^{-1} \times \\ & f_{I-k-1}^T X \left(f_k f_{k+1}^T - f_{k+1} f_k^T \right) e_h e_g^T f_{I-k-1} \end{aligned} \tag{A.9}$$

Note that

$$\begin{aligned} f_k f_{k+1}^T - f_{k+1} f_k^T &= f_k \left(f_k + e_{k+1} \right)^T - \left(f_k + e_{k+1} \right) f_k^T \\ &= f_k e_{k+1}^T - e_{k+1} f_k^T. \end{aligned} \tag{A.10}$$

Substitute (A.10) into (A.9) and recall (2.9) and (2.10) to obtain

$$\begin{aligned} \frac{1}{\hat{\nu}(k)} \frac{\partial \hat{\nu}(k)}{\partial X(g,h)} &= [T(I-k-1,k)\,T(I-k-1,k+1)]^{-1} \times \\ & \left[T(I-k-1,k)\, e_{k+1}^T - C(I-k-1,k+1) f_k^T \right] e_h e_g^T f_{I-k-1} \end{aligned} \tag{A.11}$$

Note that

$$f_k^T e_h = \delta(h \le k). \tag{A.12}$$

By means of this and similar relations, (A.11) reduces to:

$$\begin{aligned} \frac{1}{\hat{\nu}(k)} \frac{\partial \hat{\nu}(k)}{\partial X(g,h)} &= [T(I-k-1,k)\,T(I-k-1,k+1)]^{-1} \delta(g \le I-k-1) \\ & \times [\delta(h = k+1)\, T(I-k-1,k) - \delta(h \le k)\, C(I-k-1,k+1)] \end{aligned} \tag{A.13}$$

Note that

$$\begin{aligned} & \delta(g \le I-k-1)\,\delta(h = k+1) \\ &= \delta(g \le I-h)\,\delta(h=k+1) = \delta(g+h \le I)\,\delta(h=k+1) = \delta(h=k+1) \end{aligned} \tag{A.14}$$

under the conditions of the lemma.

Substituting (A.14) into (A.13) yields (A.2) as required.

To prove (A.4), return to (A.6). In this abbreviated notation there, take a second derivative:

$$\frac{\partial}{\partial X}\left(\frac{1}{\nu}\frac{\partial \nu}{\partial X}\right) = \frac{1}{\nu}\frac{\partial^2 \nu}{\partial X^2} - \left(\frac{1}{\nu}\frac{\partial \nu}{\partial X}\right)^2 \tag{A.15}$$

Now (A.6) yields

$$\frac{\partial}{\partial X}\left(\frac{1}{\nu}\frac{\partial \nu}{\partial X}\right) = \frac{1}{N}\frac{\partial^2 N}{\partial X^2} - \left(\frac{1}{N}\frac{\partial N}{\partial X}\right)^2 - \frac{1}{D}\frac{\partial^2 D}{\partial X^2} + \left(\frac{1}{D}\frac{\partial D}{\partial X}\right)^2 \tag{A.16}$$

By (A.5), (A.7) and (A.8), $\partial^2 N/\partial X^2 = \partial^2 D/\partial X^2 = 0$, so that (A.16) reduces to

$$\begin{aligned}\frac{\partial}{\partial X}\left(\frac{1}{\nu}\frac{\partial \nu}{\partial X}\right) &= \left(\frac{1}{D}\frac{\partial D}{\partial X} - \frac{1}{N}\frac{\partial N}{\partial X}\right)\left(\frac{1}{D}\frac{\partial D}{\partial X} + \frac{1}{N}\frac{\partial N}{\partial X}\right) \\ &= -\left(\frac{1}{\nu}\frac{\partial \nu}{\partial X}\right)\left(\frac{1}{D}\frac{\partial D}{\partial X} + \frac{1}{N}\frac{\partial N}{\partial X}\right)\end{aligned} \tag{A.17}$$

by (A.6).

Substitution of (A.17) in (A.15), and use of (A.6) again yields

$$\frac{1}{\nu}\frac{\partial^2 \nu}{\partial X^2} = -2\left(\frac{1}{\nu}\frac{\partial \nu}{\partial X}\right)\left(\frac{1}{D}\frac{\partial D}{\partial X}\right) \tag{A.18}$$

By (A.8),

$$\partial D/\partial X\,(g,h) = f^T_{I-k-1} e_g e_h^T f_k = \delta(g < I - k)\,\delta(h \le k) \tag{A.19}$$

Substitute (A.2) and (A.19) into (A.18) and note that the term involving $\delta(h = k+1)$ vanishes because $\delta(h = k+1)\,\delta(h \le k) = 0$.

The result is (A.4), as required. □

Lemma 3 (multivariate Jensen inequality). Let $X = (X_1, \ldots, X_m)^T$ where the X_k are stochastically independent random variables. Let $f: R^m \to R$ be twice differentiable in all its arguments, and suppose that

$$\partial^2 f(X)/\partial X_k^2 \ge 0 \text{ for all } X \text{ and for } k = 1, 2, \ldots, m. \tag{A.20}$$

Then

$$E[f(X)] \ge f(E[X]). \tag{A.21}$$

If strict inequality holds in (A.20) for at least one k, and X_k is not degenerate, then strict inequality holds in (A.21).

Proof. An elegant proof of (A.21) appears in Kallenberg (1997, p. 49). It is reproved below in order to obtain the strict inequality.

Let $\mu = (\mu_1, \ldots, \mu_m)^T = E[X]$. Expand $f(X)$ as the Taylor series:

$$f(X) = f(X_1,\ldots,X_{m-1},\mu_m) + (X_m - \mu_m)\,\partial f(X_1,\ldots,X_{m-1},\mu_m)/\partial X_m + \tfrac{1}{2}(X_m-\mu_m)^2\,\partial^2 f(X_1,\ldots,X_{m-1},\xi_m)/\partial X_m^2 \tag{A.22}$$

where $\xi_m = \mu_m + \theta_m(X_m - \mu_m)$ for some $0 < \theta_m < 1$.

Now expand $f(X_1, \ldots, X_{m-1}, \mu_m)$ similarly, then $f(X_1, \ldots, X_{m-2}, \mu_{m-1}, \mu_m)$, and so on to obtain

$$f(X) = f(\mu) + \sum_{k=1}^{m}(X_k-\mu_k)\,\partial f(X_1,\ldots,X_{k-1},\mu_k,\ldots,\mu_m)/\partial X_k + \sum_{k=1}^{m}\tfrac{1}{2}(X_k-\mu_k)^2\,\partial^2 f(X_1,\ldots,X_{k-1},\xi_k,\mu_{k+1},\ldots,\mu_m)/\partial X_k^2. \tag{A.23}$$

Take expectations on both sides of (A.23). Note that

$$\begin{aligned} &E\left[(X_k-\mu_k)\,\partial f(X_1,\ldots,X_{k-1},\mu_k,\ldots,\mu_m)/\partial X_k\right] \\ &= E(X_k-\mu_k)\,E\left[\partial f(X_1,\ldots,X_{k-1},\mu_k,\ldots,\mu_m)/\partial X_k\right] \\ &= 0 \end{aligned} \tag{A.24}$$

where the middle step follows from the stochastic independence of X_k from $(X_1, \ldots, X_{k-1})$.

By (A.20),

$$E\left[(X_k-\mu_k)^2\,\partial^2 f(X_1,\ldots,X_{k-1},\xi_k,\mu_{k+1},\ldots,\mu_m)/\partial X_k^2\right] \ge 0 \tag{A.25}$$

with strict inequality if strict inequality holds in (A.20) and X_k is not degenerate.

The lemma then follows. □

Proof of Theorem 2. Consider $Y(i)$ defined by (5.1), with $\hat{v}(k)$ defined by (3.1) and (2.3). The observations $X(g,h)$ involved in the $\hat{v}(k)$ constituting $Y(i)$ are just those in D_i. This justifies (5.2).

Now consider $(g,h) \in D_i$. Note that Lemma 1 is applicable to $Y(i)$ because of (5.1). Hence

$$\frac{1}{Y(i)}\frac{\partial^2 Y(i)}{\partial X^2(g,h)} = \sum_{k=I-i}^{I-1}\frac{1}{\hat{v}(k)}\frac{\partial^2\hat{v}(k)}{\partial X^2(g,h)} + \sum_{\substack{k,l=I-i\\k\neq l}}^{I-1}\frac{1}{\hat{v}(k)\hat{v}(l)}\frac{\partial\hat{v}(k)}{\partial X(g,h)}\frac{\partial\hat{v}(l)}{\partial X(g,h)}. \quad \text{(A.26)}$$

In this equation $I-i \le k \le I-1$. Since $(g,h)\in D_i$ it is also the case that $h-1 \le k \le I-1-g$. This yields $\max\{I-i, h-1\} \le k \le I-1-g$.

Therefore, there are two cases to be considered:

- In the case $h \le I-i$ one has $I-i \le k \le I-1-g$
- In the case $h \ge I-i+1$ one has $h-1 \le k \le I-1-g$.

Case I: $h \le I-i$.

In this case

$$g < I-k, \quad h \le k. \quad \text{(A.27)}$$

Under these conditions, combination of (3.1) with (A.2) gives

$$\partial\hat{v}(k)/\partial X(g,h) = -C(I-k-1, k+1)/T^2(I-k-1, k). \quad \text{(A.28)}$$

Similarly, combination of (3.1) with (A.4) gives

$$\partial^2\hat{v}(k)/\partial X^2(g,h) = 2C(I-k-1, k+1)/T^3(I-k-1, k). \quad \text{(A.29)}$$

Substitution of (3.1), (A.28) and (A.29) into (A.26) yields

$$\begin{aligned}\frac{1}{Y(i)}\frac{\partial^2 Y(i)}{\partial X^2(g,h)} &= 2\sum_{k=I-i}^{I-g-1}\frac{C(I-k-1,k+1)}{T(I-k-1,k+1)\,T^2(I-k-1,k)}\\ &\quad + 2\sum_{k=I-i}^{I-g-1}\sum_{l=I-i}^{k-1}\frac{C(I-k-1,k+1)\,C(I-l-1,l+1)}{T(I-k-1,k+1)\,T(I-k-1,k)\,T(I-l-1,l+1)\,T(I-l-1,l)}\\ &= 2\sum_{k=I-i}^{I-g-1}\frac{C(I-k-1,k+1)}{T(I-k-1,k+1)\,T(I-k-1,k)}\\ &\quad\times\left[\frac{1}{T(I-k-1,k)} + \sum_{l=I-i}^{k-1}\frac{C(I-l-1,l+1)}{T(I-l-1,l+1)\,T(I-l-1,l)}\right]\end{aligned} \quad \text{(A.30)}$$

Note that Assumption 3 implies that $T(g,h) > 0$ for $(g,h)\in D_i$, and so guarantees the existence of all ratios in (A.30).

Note also that, by (2.2), the second member within the square bracket may be expanded as follows:

$$\frac{C(I-l-1,l+1)}{T(I-l-1,l+1)\,T(I-l-1,l)} = \frac{1}{T(I-l-1,l)} - \frac{1}{T(I-l-1,l+1)} \quad \text{(A.31)}$$

Substitute (A.31) into the square bracket in (A.30) to obtain

$$\begin{aligned}&\frac{1}{T(I-k-1,k)}+\sum_{l=I-i}^{k-1}\frac{1}{T(I-l-1,l)}-\sum_{l=I-i+i}^{k}\frac{1}{T(I-l,l)}\\&=\frac{1}{T(i-1,I-i)}+\sum_{l=I-i+1}^{k}\left[\frac{1}{T(I-l-1,l)}-\frac{1}{T(I-l,l)}\right]\\&=\frac{1}{T(i-1,I-i)}+\sum_{l=I-i+1}^{k}\frac{R(I-l,l)}{T(I-l-1,l)T(I-l,l)}\end{aligned}\tag{A.32}$$

by (2.1).

Substitute (A.32) for the square bracket in (A.30) to obtain

$$\begin{aligned}\frac{1}{Y(i)}\frac{\partial^2 Y(i)}{\partial X^2(g,h)}&=2\sum_{k=I-i}^{I-g-1}\frac{C(I-k-1,k+1)}{T(I-k-1,k+1)(I-k-1,k)}\times\\&\left[\frac{1}{T(i-1,I-i)}+\sum_{l=I-i+1}^{k}\frac{R(I-l,l)}{T(I-l-1,l)T(I-l,l)}\right].\end{aligned}\tag{A.33}$$

This proves (5.3).

Case II: $h > I-i.$

It follows from the argument immediately preceding the proof of Case I that

$$g < I-k,\ \ k \geq h-1 \tag{A.34}$$

(A.34) may be written as the two sub-cases:

$k = h-1$

and

$g < I-k,\ h \leq k.$

Then Lemma 2 gives

$$\frac{\partial\hat{\nu}(h-1)}{\partial X(g,h)}=\frac{1}{T(I-h,h-1)} \tag{A.35}$$

$$\frac{\partial^2\hat{\nu}(h-1)}{\partial X^2(g,h)}=0 \tag{A.36}$$

$$\frac{\partial\hat{\nu}(k)}{\partial X(g,h)}=-\frac{C(I-k-1,k+1)}{T^2(I-k-1,k)},\ \ k\geq h,\ g\leq I-k-1 \tag{A.37}$$

$$\frac{\partial^2 \hat{\nu}(k)}{\partial X^2(g,h)} = \frac{2C(I-k-1,k+1)}{T^3(I-k-1,k)}, \ \ k \geq h, \ g \leq I-k-1. \tag{A.38}$$

Substitute (A.35) – (A.38) into (A.26) to obtain

$$\begin{aligned}
\frac{1}{Y(i)}\frac{\partial^2 Y(i)}{\partial^2 X(g,h)} &= \frac{1}{\hat{\nu}(h-1)}\frac{\partial^2 \hat{\nu}(h-1)}{\partial X^2(g,h)} + \sum_{k=h}^{I-g-1}\frac{1}{\hat{\nu}(k)}\frac{\partial^2 \hat{\nu}(k)}{\partial X^2(g,h)} \\
&+ \sum_{\substack{k,l=h \\ k\neq l}}^{I-g-1}\frac{1}{\hat{\nu}(k)\hat{\nu}(l)}\frac{\partial \hat{\nu}(k)}{\partial X(g,h)}\frac{\partial \hat{\nu}(l)}{\partial X(g,h)} \\
&+ \frac{2}{\hat{\nu}(h-1)}\frac{\partial \hat{\nu}(h-1)}{\partial X(g,h)}\sum_{k=h}^{I-g-1}\frac{1}{\hat{\nu}(k)}\frac{\partial \hat{\nu}(k)}{\partial X(g,h)} \\
&= 2\sum_{k=h}^{I-g-1}\frac{C(I-k-1,k+1)}{T(I-k-1,k+1)\,T^2(I-k-1,k)} \\
&+ 2\sum_{k=h}^{I-g-1}\frac{C(I-k-1,k+1)}{T(I-k-1,k+1)\,T(I-k-1,k)}\sum_{l=h}^{k-1}\frac{C(I-l-1,l+1)}{T(I-l-1,l+1)\,T(I-l-1,l)} \\
&- \frac{2}{T(I-h,h)}\sum_{k=h}^{I-g-1}\frac{C(I-k-1,k+1)}{T(I-k-1,k+1)\,T(I-k-1,k)}.
\end{aligned} \tag{A.39}$$

As in Case I, Assumption 3 guarantees the existence of all the ratios in (A.39).

Now apply (A.31) and use the same mode of calculation as led from (A.30) to (A.33). Then (A.39) becomes:

$$\begin{aligned}
\frac{1}{Y(i)}\frac{\partial^2 Y(i)}{\partial X^2(g,h)} &= 2\sum_{k=h}^{I-g-1}\frac{C(I-k-1,k+1)}{T(I-k-1,k+1)\,T(I-k-1,k)} \times \\
&\sum_{l=h}^{k}\frac{R(I-l,l)}{T(I-l-1,l)\,T(I-l,l)}.
\end{aligned} \tag{A.40}$$

This proves (5.4). □

Proof of Theorem 3. Consider $Y(i)$ defined by (5.1). By Theorem 2,

$\partial^2 Y(i)/\partial X^2(g,h) \geq 0$ for all (g,h)

with strict inequality for some (g,h), namely those in D_i. It follows from a multivariate form of Jensen's inequality (see Lemma 3) that

$$E[Y(i)] > \bar{Y}(i) = \prod_{k=I-i}^{I-1}\bar{\nu}(k) \tag{A.41}$$

where $\bar{Y}(i)$ is the value obtained by replacing each $X(g,h)$ in $Y(i)$ by its expectation, and $\bar{v}(k)$ is similarly defined.

By (3.1),

$$\bar{v}(k) = E[T(I-k-1,k+1)] / E[T(I-k-1,k)]$$
$$= \sum_{g=0}^{I-k-1} E[R(g,k+1)] / \sum_{g=0}^{I-k-1} E[R(g,k)] \quad \text{[by (2.3)]} \qquad \text{(A.42)}$$
$$= \eta(k), \quad \text{by (4.5)}.$$

Substitute (A.42) in (A.41):

$$E[Y(i)] > \prod_{k=I-i}^{I-1} \eta(k) = E[R(i,I)] / E[R(i,I-i)] \qquad \text{(A.43)}$$

by (4.5). This proves (5.5).

Now take expectations on both sides of (3.2):

$$E\left[\hat{R}(i,I)\right] = E[R(i,I-i)]\, E[Y] > E[R(i,I)], \quad \text{by (A.43)}. \qquad \text{(A.44)}$$

The first step leading to (A.44) is justified as follows. $Y(i)$ is defined by (5.1) and $\hat{v}(k)$ by (3.1), which shows that the rows of X involved in $Y(i)$ are $0, 1, \ldots, i-1$. Thus, $R(i, I-i)$ and $Y(i)$ are stochastically independent.

□

Greg Taylor
Taylor Fry Consulting Actuaries
Level 8, 30 Clarence Street
Sydney NSW 2000
Australia

Professorial Associate,
Centre for Actuarial Studies
Faculty of Economics and Commerce
University of Melbourne
Parkville VIC 3052
Australia

Phone: 61-2-9249-2901
Fax: 61-2-9249-2999
greg@taylorfry.com.au

CLAIMS RESERVING USING TWEEDIE'S COMPOUND POISSON MODEL

BY

MARIO V. WÜTHRICH

ABSTRACT

We consider the problem of claims reserving and estimating run-off triangles. We generalize the gamma cell distributions model which leads to Tweedie's compound Poisson model. Choosing a suitable parametrization, we estimate the parameters of our model within the framework of generalized linear models (see Jørgensen-de Souza [2] and Smyth-Jørgensen [8]). We show that these methods lead to reasonable estimates of the outstanding loss liabilities.

KEYWORDS

Claims Reserving, Run-off Triangles, IBNR, Compound Poisson Model, Exponential Family, GLM, MSEP.

INTRODUCTION

Claims reserving and IBNR estimates are classical problems in insurance mathematics. Recently Jørgensen-de Souza [2] and Smyth-Jørgensen [8] have fitted Tweedie's compound Poisson model to insurance claims data for tarification. Using the connection between tarification and claims reserving analysis (see Mack [3]), we translate the fitting procedure to our run-off problem. Our model should be viewed within the context of stochastic methods for claims reserving. For excellent overviews on this topic we refer to England-Verrall [1] and Taylor [9].

The starting point of this work was the gamma cell distributions model presented in Section 7.5 of Taylor [9]. The gamma cell distributions model assumes that every cell of the run-off triangle consists of r_{ij} independent payments which are gamma distributed with mean τ_{ij} and shape parameter γ. These assumptions enable the calculation of convoluted distributions of incremental payments. Unfortunately, this model does not allow one to estimate e.g. the mean square error of prediction (MSEP), since one has not enough information. We assume that the number of payments r_{ij} are realisations of random variables R_{ij}, i.e. the number of payments R_{ij} and the size of the individual payments $X_{ij}^{(k)}$ are both modelled stochastically. This can be done assuming that

R_{ij} is Poisson distributed. These assumptions lead to *Tweedie's compound Poisson model* (see e.g. Jørgensen-de Souza [2]). Choosing a clever parametrization for Tweedie's compound Poisson model, we see that the model belongs to the exponential dispersion familiy with variance function $V(\mu) = \mu^p$, $p \in (1,2)$, and dispersion ϕ. It is then straightforward to use generalized linear model (GLM) methods for parameter estimations. A significant first step into that direction has been done by Wright [11].

In this work we study a version of Tweedie's compound Poisson model with constant dispersion ϕ (see Subsection 4.1). This model should be viewed within the context of the over-dispersed Poisson model (see Renshaw-Verrall [6] or England-Verrall [1], Section 2.3) and the Gamma model (see Mack [3] and England-Verrall [1], Section 3.3): The over-dispersed Poisson model and the Gamma model correspond to the two extreme cases $p = 1$ and $p = 2$, resp. Our extension closes continuously the gap between these two models, since $p \in (1,2)$. To estimate p we additionally use the information r_{ij} which is not used in the parameter estimations for $p = 1$ and $p = 2$. Though we have one additional parameter, we obtain in general better estimates since we also use more information and have more degrees of freedom.

Moreover, our parametrization is such that the variance parameters p and ϕ are orthogonal to the mean parameter. This leads to a) efficient parameter estimations (fast convergence), b) good estimates of MSEP.

At the end of this article we demonstrate the method using motor insurance datas. Our results are compared to several different classical methods. Of course, in practice it would not be wise to trust in just these methods. It should be pointed out that the methods presented here are all payment based. Usually it is also interesting to compare payment based results to results which rely on total claims incurred datas (for an overview we refer to Taylor [9] and the references therein).

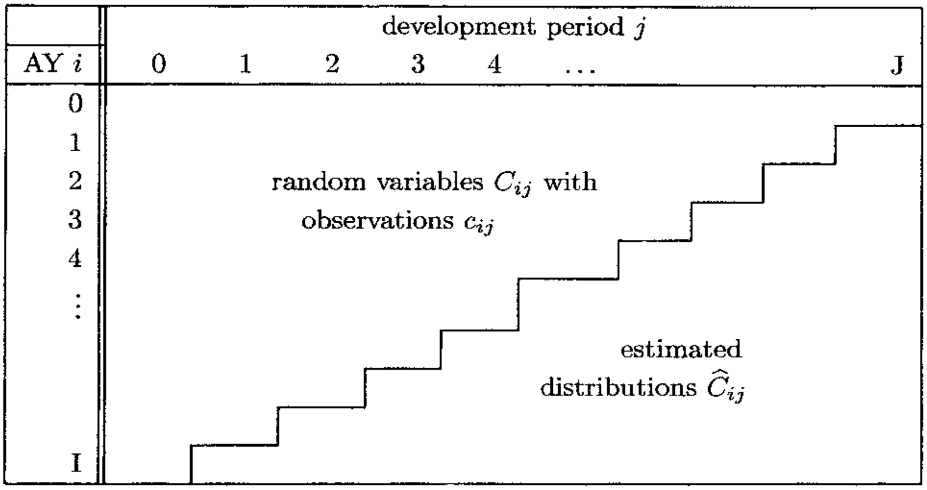
In the next section we define the model. In Section 3 we recall the definition of Tweedie's compound Poisson model. In Section 4 we apply Tweedie's compound Poisson model to our run-off problem. In Section 5 we give an estimation procedure for the mean square error of prediction (MSEP). Finally, in Section 6 we give the examples.

2. Definition of the model

We use the following (well-known) structure for the run-off patterns: the accident years are denoted by $i \le I$ and the development periods are denoted by $j \le J$. We are interested in the random variables C_{ij}. C_{ij} denote the incremental payments for claims with origin in accident year i during development period j. Usually one has observations c_{ij} of C_{ij} for $i + j \le I$ and one tries to complete (estimate) the triangle for $i + j > I$. The following illustration may be helpful.

Definition of the model:

1. The number of payments R_{ij} are independent and Poisson distributed with parameter $\lambda_{ij} w_i > 0$. The weight $w_i > 0$ is an appropriate measure for the volume.



2. The individual payments $X_{ij}^{(k)}$ are independent and gamma distributed with mean $\tau_{ij} > 0$ and shape parameter $\gamma > 0$.
3. R_{ij} and $X_{mn}^{(k)}$ are independent for all indices. We define the incremental payments paid in cell (i, j) as follows

$$C_{ij} = 1_{\{R_{ij} > 0\}} \cdot \sum_{k=1}^{R_{ij}} X_{ij}^{(k)} \quad \text{and} \quad Y_{ij} = C_{ij} / W_i. \tag{2.1}$$

Remarks:

- There are several different possibilities to choose appropriate weights w_i, e.g. the number of policies or the total number of claims incurred, etc. If one chooses the total number of claims incurred one needs first to estimate the number of IBNyR cases (cases incurred but not yet reported).
- Sometimes it is also convenient to define R_{ij} as the number of claims with origin in i which have at least one payment in period j.
- Y_{ij} denotes the normalized incremental payments in cell (i,j).
- One easily sees that conditionally, given R_{ij}, the random variable C_{ij} is gamma distributed with mean $R_{ij}\tau_{ij}$ and shape parameter $R_{ij}\gamma$ (for $R_{ij} > 0$).

3. Tweedie's compound Poisson model

In this section we formulate our model in a reparametrized version, this has already been done in the tarification problems of [2] and [8]. Therefore we try to keep this section as short as possible and give the main calculations in Appendix A.

For the moment we skip the indices i and j. The distribution Y (for given weight w) is parametrized by the three parameters λ, τ and γ. We now choose new parameters μ, ϕ and p such that the density of Y can be written as, $y \geq 0$, (see (A.2) below and formula (12) in [2])

$$f_Y(y;\mu,\phi/w,p) = c(y;\phi/w,p)\exp\left\{\frac{w}{\phi}\left(y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right)\right\}, \tag{3.1}$$

where $c(y;\phi/w,p)$ is given in Appendix A and

$$p = (\gamma+2)/(\gamma+1) \in (1,2), \tag{3.2}$$

$$\mu = \lambda \cdot \tau, \tag{3.3}$$

$$\phi = \lambda^{1-p}\tau^{2-p}/(2-p). \tag{3.4}$$

If we set $\theta = \mu^{1-p}/(1-p)$ we see that the density of Y can be written as (see also [2], formula (12))

$$f_Y(y;\mu,\phi/w,p) = c(y;\phi/w,p)\exp\left\{\frac{w}{\phi}\left(y\theta - \kappa_p(\theta)\right)\right\}, \quad \text{with } \kappa_p(\theta) = \frac{1}{2-p}\left((1-p)\,\theta\right)^{\frac{2-p}{1-p}}. \tag{3.5}$$

Hence, the distribution of Y belongs to the exponential dispersion family with parameters μ, ϕ and $p \in (1,2)$ (see e.g. McCullagh-Nelder [5], Section 2.2.2). We write for $p \in (1,2)$

$$Y \backsim \mathrm{ED}^{(p)}(\mu,\phi/w). \tag{3.6}$$

For $Y \backsim \mathrm{ED}^{(p)}(\mu,\phi/w)$ we have (see [2] Section 2.2)

$$E[Y] = \kappa_p'(\theta) = \mu, \tag{3.7}$$

$$\mathrm{Var}(Y) = \frac{\phi}{w}\cdot V(\mu) = \frac{\phi}{w}\mu^p. \tag{3.8}$$

ϕ is the so-called dispersion parameter and $V(\cdot)$ the variance function with $p \in (1,2)$. For our claims reserving problem we consider the following situation:

Constant dispersion ϕ (see Subsection 4.1): $p \in (1,2)$ and Y_{ij} are independent with

$$Y_{ij} \backsim \mathrm{ED}^{(p)}\left(\mu_{ij},\phi/w_i\right) \Rightarrow E\left[Y_{ij}\right] = \mu_{ij} \text{ and } \mathrm{Var}\left(Y_{ij}\right) = \frac{\phi}{w_i}\mu_{ij}^p. \tag{3.9}$$

Interpretation and Remarks:

- Tweedie [10] seems to be the first one to study the compound Poisson model with gamma severeties from the point of view of exponential dispersion models. For this reason this model is known as Tweedie's compound Poisson model in the literature, see e.g. [8].
- $p = (\gamma + 2) / (\gamma + 1)$ is a function of γ (shape parameter of the single payments distributions $X_{ij}^{(k)}$). Hence the shape parameter γ determines the behaviour of the variance function $V(\mu) = \mu^p$. Furthermore we have chosen a parametrization (μ, ϕ, p) such that μ is *orthogonal* to (ϕ, p) in the sense that the Fisher information matrix is zero in the off-diagonal (see e.g. [2], page 76, or [8]). I.e. our parametrization focuses attention to variance parameters (ϕ, p) and a mean parameter μ which are orthogonal to each other. This orthogonality has many advantages to alternative parametrizations. E.g. we have efficient algorithms for parameter estimations which typically rapidly converge (see Smyth [7]). Moreover the estimated standard errors of μ, which are of most interest, do not require adjustments by the standard errors of the variance parameters, since these are orthogonal.
- Our model closes continuously the gap between the over-dispersed Poisson Model (see Renshaw-Verrall [6] or England-Verrall [1], Section 2.3) where we have a linear variance function ($p = 1$):

$$\mathrm{Var}\left(Y_{ij}\right) = \phi / w_i \cdot \mu_{ij}, \tag{3.10}$$

  and the Gamma model (see Mack [3] and England-Verrall [1], Section 3.3) where

$$\mathrm{Var}\left(Y_{ij}\right) = \phi / w_i \cdot \mu_{ij}^2. \tag{3.11}$$

  In our case p is estimated from the data using additionally the information r_{ij} (see (4.6)). The information r_{ij} is not used in the boundary cases $p = 1$ and $p = 2$.
- Naturally in our model we have $p \in (1,2)$, since $\gamma > 0$. We estimate p from the data, so theoretically the estimated p could lie outside the interval $[1,2]$ which would mean that none of our models fits to the problem (e.g. $p = 0$ implies normality, $p = 3$ implies the inverse Gaussian model). In all our claims reserving examples we have observed that the estimated p was lying strictly within (1,2).

4. Application of Tweedie's model to claims reserving

4.1. Constant dispersion parameter model

We assume that all the Y_{ij} are independent with $Y_{ij} \sim \mathrm{ED}^{(p)}(\mu_{ij}, \phi/w_i)$, i.e. Y_{ij} belongs to the exponential dispersion family with $p \in (1,2)$, and

$$E\left[Y_{ij}\right] = \mu_{ij} \quad \text{and} \quad \text{Var}\left(Y_{ij}\right) = \frac{\phi}{w_i}\mu_{ij}^p. \tag{4.1}$$

We use the notation $\boldsymbol{\mu} = (\mu_{00}, \ldots, \mu_{IJ})'$. Given the observations $\{(r_{ij}, y_{ij}),\ i+j \le I,\ \sum_{ij} r_{ij} > 0\}$, the log-likelihood function for the parameters $(\boldsymbol{\mu}, \phi, p)$ is given by (see Appendix A and [2], Section 3)

$$L(\boldsymbol{\mu}, p, \phi) = \sum_{i,j} \left[r_{ij} \log\left\{ \frac{(w_i/\phi)^{\gamma+1} y_{ij}^{\gamma}}{(p-1)^{\gamma}(2-p)} \right\} - \log\left(r_{ij}!\,\Gamma\left(r_{ij}\gamma\right) y_{ij}\right) \right.$$
$$\left. + \frac{w_i}{\phi} \left\{ y_{ij} \frac{\mu_{ij}^{1-p}}{1-p} - \frac{\mu_{ij}^{2-p}}{2-p} \right\} \right]. \tag{4.2}$$

Formula (4.2) immediately shows that given p the observations $y_{ij} = c_{ij}/w_i$ are sufficient for MLE estimation of μ_{ij} (one does not need r_{ij}). Moreover, for constant ϕ, the dispersion parameter has no influence on the estimation of $\boldsymbol{\mu}$.

Next we assume a multiplicative model (often called chain-ladder type structure): i.e. there exist parameters $\alpha(i)$ and $f(j)$ such that for all $i \le I$ and $j \le J$

$$\mu_{ij} = \alpha(i) \cdot f(j). \tag{4.3}$$

After suitable normalization, α can be interpreted as the expected ultimate claim in accident year i and f is the proportion paid in period j. It is now straightforward to choose the logarithmic link function

$$\eta_{ij} = \log(\mu_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta}, \tag{4.4}$$

where $\boldsymbol{\beta} = (\log\alpha(0), \ldots, \log\alpha(I), \log f(0), \ldots, \log f(j))'$ and $X = (\mathbf{x}_{00}, \ldots, \mathbf{x}_{IJ})$ is the appropriate design matrix.

Parameter estimation:

a) **For *p* known.** We deal with a generalized linear model (GLM) of the form (4.1)-(4.4). Hence we can use standard software packages for the estimation of $\boldsymbol{\mu}$.

b) **For *p* unknown.** Usually p and γ, resp., are unknown. Henceforth we study the profile likelihood for γ (here we closely follow [2] Section 3.2): For $\boldsymbol{\mu}$ and p given, the MLE of ϕ is given by (see (4.2))

$$\hat{\phi}_p = \frac{-\sum_{i,j} w_i \left(y_{ij} \frac{\mu_{ij}^{1-p}}{1-p} - \frac{\mu_{ij}^{2-p}}{2-p} \right)}{(1+\gamma)\sum_{i,j} r_{ij}}. \tag{4.5}$$

From this we obtain the profile likelihood for p and γ, resp., $\left(\sum_{ij} r_{ij} > 0\right)$ as

$$L_\mu(p) = L(\boldsymbol{\mu}, p, \hat{\phi}_p) = (1+\gamma)\sum_{i,j} r_{ij}\left(\log\frac{w_i}{\hat{\phi}_p} - 1\right) + \sum_{i,j} r_{ij}\log\left(\frac{1}{2-p}\left(\frac{y_{ij}}{p-1}\right)^{\gamma}\right) - \sum_{i,j}\log\Gamma\left(r_{ij}\gamma\right). \tag{4.6}$$

Given $\boldsymbol{\mu}$, the parameter p is estimated maximizing (4.6).

c) **Finally we combine a) and b).** The main advantage of our parametrization is (as already mentioned above) the orthogonality of $\boldsymbol{\mu}$ and (ϕ, p). $\boldsymbol{\mu}$ can be estimated as if (ϕ, p) were known and vice versa. Alternating the updating procedures for $\boldsymbol{\mu}$ and (ϕ, p) leads to an efficent algorithm: Set initial value $p^{(0)}$ and estimate $\boldsymbol{\mu}^{(1)}$ via a). Then estimate $p^{(1)}$ from $\boldsymbol{\mu}^{(1)}$ via (4.6), and iterate this procedure. We have seen that typically one obtains very fast convergence of $(\boldsymbol{\mu}^{(k)}, p^{(k)})$ to some limit (for our examples below we needed only 4 iterations).

4.2. Dispersion modelling

So far we have always assumed that ϕ is constant over all cells (i, j). If we consider the definitions (3.3) and (3.4) we see that every factor which increases λ increases the mean μ and decreases the dispersion ϕ because $p \in (1,2)$. Increasing the average payment size τ increases both the mean and the dispersion. Changing λ and τ such that $\lambda^{1-p}\,\tau^{2-p}$ remains constant has only an effect on the mean μ. Hence it is necessary to model both the mean and the dispersion in order to get a fine structure, i.e. model μ_{ij} and ϕ_{ij} for each cell (i,j) individually and estimate p. Such a model has been studied in the context of tarification by Smyth-Jørgensen [8].

We do not further follow these ideas here since we have seen that in our situation such models are over-parametrized. Modelling the dispersion parameters while also trying to optimize the power of the variance function allows too many degrees of freedom: e.g. if we apply the dispersion modelling model to the data given in Example 6.1 one sees that p is blown up when allowing the dispersion parameters to be modelled too. It is even possible that there is no unique solution when modelling ϕ_{ij} and p at the same time (in all our examples we have observed rather slow convergence even when choosing "meaningful" initial values which indicates this problematic).

5. Mean square error of prediction

To estimate the mean square error of prediction (MSEP) we proceed as in England-Verrall [1]. Assume that the incremental payments C_{ij} are independent, and $\widehat{C_{ij}}$ are unbiased estimators depending only on the past (and hence are independent from C_{ij}). Assume $\hat{\eta}_{ij}$ is the GLM estimate for $\eta_{ij} = \log \mu_{ij}$, then (see e.g. [1], (7.6)-(7.7))

$$\begin{aligned} \mathrm{MSEP}_{C_{ij}}\left(\widehat{C_{ij}}\right) &= E\left[\left(C_{ij} - \widehat{C_{ij}}\right)^2\right] = \mathrm{Var}\left(C_{ij}\right) + \mathrm{Var}\left(\widehat{C_{ij}}\right) \\ &\approx \phi w_i \cdot \mu_{ij}^p + \left(w_i \mu_{ij}\right)^2 \mathrm{Var}\left(\hat{\eta}_{ij}\right). \end{aligned} \tag{5.1}$$

The last term is usually available from standard statistical software packages, all the other parameters have been estimated before. The first term in (5.1) denotes the process error, the last term the estimation error.

The estimation of the MSEP for several cells (i, j) is more complicated since we obtain correlations from the estimates. We define Δ to be the unknown triangle in our run-off pattern. Define the total outstanding payments

$$C = \sum_{(i,j) \in \Delta} C_{ij} \quad \text{and} \quad \hat{C} = \sum_{(i,j) \in \Delta} \widehat{C_{ij}}. \tag{5.2}$$

Then

$$\begin{aligned} \mathrm{MSEP}_C(\hat{C}) = E\left[\left(C - \hat{C}\right)^2\right] &\approx \sum_{(i,j) \in \Delta} \phi_{ij} w_i \cdot \mu_{ij}^p + \sum_{(i,j) \in \Delta} \left(w_i \mu_{ij}\right)^2 \mathrm{Var}\left(\hat{\eta}_{ij}\right) \\ &+ \sum_{\substack{(i_1,j_1),(i_2,j_2) \in \Delta, \\ (i_1,j_1) \neq (i_2,j_2)}} w_{i_1} \mu_{i_1 j_1} w_{i_2} \mu_{i_2 j_2} \mathrm{Cov}\left(\hat{\eta}_{i_1 j_1}, \hat{\eta}_{i_2 j_2}\right). \end{aligned}$$

The evaluation of the last term needs some care: Usually one obtains a covariance matrix for the estimated GLM parameters $\log \alpha(i)$ and $\log f(j)$. This covariance matrix needs to be transformed into a covariance matrix for η with the help of the design matrices.

6. Example

Example 6.1.

We consider Swiss Motor Insurance datas. We consider 9 accident years over a time horizon of 11 years. Since we want to analyze the different methods rather mechanically, this small part of the truth is already sufficient for drawing conclusions.

TABLE 6.2

OBSERVATIONS FOR THE NORMALIZED INCREMENTAL PAYMENTS $Y_{ij} = C_{ij}/w_i$.

y_{ij}	Development period j										
AY i	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	157.95	65.89	7.93	3.61	1.83	0.55	0.14	0.22	0.01	0.14	0.00
1	176.86	60.31	8.53	1.41	0.63	0.34	0.49	1.01	0.38	0.23	
2	189.67	60.03	10.44	2.65	1.54	0.66	0.54	0.09	0.19		
3	189.15	57.71	7.77	3.03	1.43	0.95	0.27	0.61			
4	184.53	58.44	6.96	2.91	3.46	1.12	1.17				
5	185.62	56.59	5.73	2.45	1.05	0.93					
6	181.03	62.35	5.54	2.43	3.66						
7	179.96	55.36	5.99	2.74							
8	188.01	55.86	5.46								

TABLE 6.3

NUMBER OF PAYMENTS R_{ij} AND VOLUME w_i.

r_{ij}	Development period j											
AY i	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	w_i
0	6'229	3'500	425	134	51	24	13	12	6	4	1	112'953
1	6'395	3'342	402	108	31	14	12	5	6	5		110'364
2	6'406	2'940	401	98	42	18	5	3	3			105'400
3	6'148	2'898	301	92	41	23	12	10				102'067
4	5'952	2'699	304	94	49	22	7					99'124
5	5'924	2'692	300	91	32	23						101'460
6	5'545	2'754	292	77	35							94'753
7	5'520	2'459	267	81								92'326
8	5'390	2'224	223									89'545

Remark: As weights w_i we take the number of reported claims (the number of IBNyR claims with reporting delay of more than two years is almost zero for this kind of business).

a) Tweedie's compound Poisson model with constant dispersion.

We assume that Y_{ij} are independent with $Y_{ij} \sim \mathrm{ED}^{(p)}(\mu_{ij}, \phi/w_i)$ (see (4.1)). Define the total outstanding payments C as in (5.2). If we start with initial value $p^{(0)} = 1.5 \in (1,2)$ and then proceed the estimation iteration as in Subsection 4.1, we observe that already after 4 iterations we have sufficiently converged to equilibrium (for the choice of p one should also have a look at Figure 1):

TABLE 6.4

ESTIMATION OF p.

Iteration k	**0**	**1**	**2**	**3**	**4**
$p^{(k)}$	1.5000	1.1743	1.1741	1.1741	1.1741
Outstanding payments $\hat{C}^{(k)}$		1'431'266	1'451'288	1'451'300	1'451'299

For $p = 1.1741$ the GLM output is as follows: Dispersion $\hat{\phi} = 29'281$ and parameter estimates:

TABLE 6.5

PARAMETERS α AND f FOR $p = 1.1741$.

j	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$\widehat{\log\alpha}(j)$	–5.862	–5.825	–5.762	–5.782	–5.777	–5.819	–5.792	–5.837	–5.809		
$\widehat{\log f}(j)$	11.01	9.90	7.79	6.78	6.45	5.51	5.13	5.08	4.17	4.16	0.00

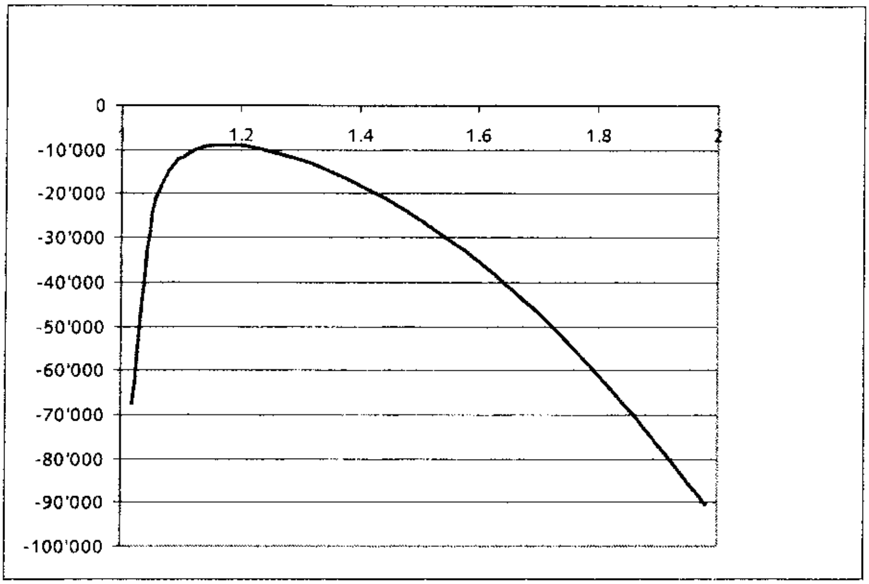


Figure 1: Profile likelihood function $L_\mu(p)$ (see (4.6)).

Altogether this leads to the following result:

TABLE 6.6

ESTIMATED OUTSTANDING PAYMENTS FROM TWEEDIE'S COMPOUND POISSON MODEL.

	Tweedie constant ϕ = 29'281 and p = 1.1741				
AY i	**Outst. payments**	**MSEP$^{1/2}$**	**in %**	**Estimation error**	**Process error**
1	326	2'636	808.9%	1'867	1'860
2	21'565	26'773	124.2%	15'584	21'770
3	40'716	35'515	87.2%	19'122	29'927
4	89'278	53'227	59.6%	25'940	46'479
5	138'338	65'977	47.7%	30'529	58'489
6	204'269	80'815	39.6%	35'191	72'751
7	360'117	111'797	31.0%	45'584	102'082
8	596'690	149'775	25.1%	61'212	136'695
Total	1'451'299	271'503	18.7%	179'890	203'355

The results in Table 6.6 show that there is considerable uncertainty in the reserve estimates, especially in the old years where the outstanding payments are small. This comes from the fact that we have only little information to estimate $f(j)$ for large j and it turns out that the parameter estimation error lives on the same scale as the process error. For young accident years we have on the one hand a lot of information to estimate $f(j)$ for small j and on the other hand $f(j)$ for j large has a rather small influence on the overall outstanding payments estimate for young accident years in our example. Therefore the relative prediction error is smaller for young accident years

b) Over-dispersed Poisson and Gamma Model.

We first compare our result to the two boundary cases $p = 1$ and $p = 2$. These models are described in Renshaw-Verrall [6] or England-Verrall [1], Section 2.3 (over-dispersed Poisson model) and Mack [3] or England-Verrall [1], Section 3.3 (Gamma model). We also refer to (3.10)-(3.11). We obtain the following results:

TABLE 6.7

ESTIMATED OUTSTANDING PAYMENTS FROM THE OVER-DISPERSED POISSON MODEL.

	Over-dispersed Poisson model with ϕ = 36'642 and p = 1				
AY i	**Outst. payments**	**MSEP$^{1/2}$**	**in %**	**Estimation error**	**Process error**
1	330	4'947	1500.7%	3'520	3'475
2	21'663	34'776	160.5%	20'386	28'174
3	41'007	46'070	112.3%	24'896	38'763
4	88'546	65'229	73.7%	31'786	56'961
5	140'150	80'795	57.6%	37'316	71'662
6	204'157	95'755	46.9%	41'089	86'491
7	362'724	125'433	34.6%	49'421	115'286
8	602'784	161'023	26.7%	61'978	148'618
Total	1'461'360	371'208	21.7%	216'965	231'403

TABLE 6.8

ESTIMATED OUTSTANDING PAYMENTS FROM THE GAMMA MODEL.

	Gamma model with ϕ = 29'956 and p = 2				
AY i	**Outst. payments**	**MSEP$^{1/2}$**	**in %**	**Estimation error**	**Process error**
1	447	346	77.3%	255	233
2	20'248	13'602	67.2%	8'527	10'597
3	40'073	20'127	50.2%	13'178	15'213
4	122'899	56'984	46.4%	37'465	42'936
5	121'740	50'091	41.1%	35'106	35'730
6	221'524	91'174	41.2%	66'731	62'126
7	331'115	147'730	44.6%	107'386	101'451
8	527'988	250'816	47.5%	194'155	158'784
Total	1'386'034	336'842	24.3%	265'771	206'950

Conclusions: It is not very surprising that the over-dispersed Poisson model gives a better fit than the Gamma model (especially for young accident years we have a huge estimation error term in the Gamma model, see Table 6.8). Tweedie's compound Poisson model converges to the over-dispersed Poisson model for $p \to 1$ and to the Gamma model for $p \to 2$. For our data set $p = 1.1741$ is close to 1, hence we expect that Tweedie's compound Poisson results are close to the over-dispersed Poisson results. Indeed, this is the case (see Tables 6.6 and 6.7). Moreover we observe that the estimation error term is essentially smaller in Tweedie's model than in the over-dispersed Poisson model. Two main reasons for this fact are 1) For the parameter estimations in Table 6.6 we additionally use the information coming from the number of payments r_{ij} (which is used for the estimation of p). 2) In our model, the variance parameters (ϕ, p) are orthogonal to μ, hence their uncertainties have no influence on the parameter error term coming from $\mathrm{Var}(\hat{\mu})$.

c) Mack's model and log-normal model.

A classical non-parametric model is the so-called chain-ladder method where we apply Mack's formulas (see Mack [4]) for the MSEP estimation. We apply the chain-ladder method to the cumulative payments

$$D_{ij} = \sum_{k=0}^{j} C_{ik} = w_i \sum_{k=0}^{j} Y_{ik}. \tag{6.1}$$

We choose the chain-ladder factors and the estimated standard errors as follows (for the definition of $f(j)$ and $\sigma_j^2 = \alpha_j^2$ we refer to Mack [4], formulas (3) and (5)). Of course there is unsufficient information for the estimation of σ_{10}. Since it is not our intention to give good strategies for estimating ultimates (this would go beyond the scope of this paper) we have just chosen a value which looks meaningful.

TABLE 6.9

CHAIN-LADDER PARAMETERS IN MACK'S MODEL.

j	1	2	3	4	5	6	7	8	9	10
$f(j)$	1.3277	1.0301	1.0107	1.0076	1.0030	1.0020	1.0019	1.0008	1.0008	1.0000
σ_j	157.28	34.16	14.17	23.31	5.70	7.78	8.67	3.89	3.00	0.50

This leads to the following result:

TABLE 6.10

ESTIMATED OUTSTANDING PAYMENTS FROM MACK'S MODEL.

	Chain-ladder estimates				
AY i	**Outst. payments**	**MSEP$^{1/2}$**	**in %**	**Estimation error**	**Process error**
1	330	3'740	1134.6%	2'661	2'627
2	21'663	19'903	91.9%	11'704	16'099
3	41'007	30'090	73.4%	15'954	25'512
4	88'546	57'012	64.4%	26'295	50'585
5	140'150	71'511	51.0%	31'476	64'212
6	204'157	75'522	37.0%	31'746	68'526
7	362'724	138'915	38.3%	49'300	129'872
8	602'784	156'413	25.9%	54'293	146'688
Total	1'461'360	286'752	19.6%	177'616	225'120

A look at the results shows that Tweedie's compound Poisson model is close to the chain-ladder estimates. For the outstanding payments this is not surprising since for $p = 1.1741$, we expect that Tweedie's estimate for the outstanding payments is close to the Poisson estimate (which is identical with the chain-ladder estimate). For the error terms it is more surprising that they are so similar. The reason for this similarity is not so clear because we have estimated a different number of parameters with a different number of observations. Furthermore, MSEP is obtained in completely different ways (see also discussion in [1], Section 7.6).

An other well-known model is the so-called parametric chain-ladder method, which is based on the log-normal distribution (see Taylor [9], Section 7.3). We assume that

$$\log\left(D_{i,j+1}/D_{i,j}\right) \sim \mathcal{N}\left(\xi_j, \sigma_j^2\right) \text{ and are independent.} \tag{6.2}$$

This model is different from the one usually used in claims reserving, which would apply to incremental data (see e.g. [1], Section 3.2). We have chosen the model from Taylor [9] because it is very easy to handle.

Living in a "normal" world we estimate the parameters as in Taylor [9], formulas (7.11)-(7.13): i.e. since we assume that the parameters only depend on the development period, we take the unweighted averages to estimate ξ_j and the canonical variance estimate for σ_j^2. This implies:

TABLE 6.11

PARAMETER ESTIMATES IN THE LOG-NORMAL MODEL.

j	1	2	3	4	5	6	7	8	9	10
$\xi(j)$	0.2832	0.0293	0.0106	0.0077	0.0030	0.0020	0.0019	0.0008	0.0008	0.0000
σ_j	0.0274	0.0067	0.0027	0.0046	0.0011	0.0015	0.0016	0.0007	0.0004	0.0001

The prediction errors are estimated according to Taylor [9], formulas (7.29)-(7-35). This leads to the following result:

TABLE 6.12

ESTIMATED OUTSTANDING PAYMENTS FROM THE LOG-NORMAL MODEL.

	Log-normal model				
AY i	**Outst. payments**	**MSEP$^{1/2}$**	**in %**	**Estimation error**	**Process error**
1	330	3'905	1183.7%	2'761	2'761
2	21'603	14'297	66.2%	8'412	11'561
3	40'814	26'680	65.4%	13'991	22'717
4	88'535	53'940	60.9%	25'130	47'728
5	140'739	69'027	49.0%	30'676	61'836
6	205'396	71'506	34.8%	31'043	64'416
7	367'545	131'216	35.7%	49'386	121'568
8	608'277	147'156	24.2%	54'163	136'826
Total	1'473'238	271'252	18.4%	170'789	210'733

The log-normal model gives estimates for the outstanding payments which are close to the chain-ladder estimates, and hence are close to Tweedie's estimates. We have very often observed this similarity. One remarkable difference between Tweedie's MSEP estimates and log-normal MSEP estimates is, that the Tweedie model gives more weight to the uncertainties for high development periods where one has only a few observations. This may come from the fact that for the chain-ladder model we consider cumulative data. This cumulation has already some smoothing effect.

CONCLUSIONS

Of course, we start the actuarial analysis of our claims reserving problem by the chain-ladder method. The chain-ladder reserve can very easily be calculated.

But we believe that it is also worth to perform Tweedie's compound Poisson method. Using the additional information r_{ij} one obtains an estimate for the variance function $V(\mu) = \mu^p$. If p is close to 1, Tweedie's compound Poisson method supports that the chain-ladder estimate. Whereas for p different from 1 it is questionable to believe in the chain-ladder reserve, since Tweedie's model tells us that we should rather consider a different model (e.g. the Gamma model for p close to 2).

A. Reparametrization

We closely follow [2]. We skip the indices i, j. The joint density of (Y, R) is

$$\begin{aligned} f_{R,Y}(r,y;\lambda,\tau,\gamma)\,dy &= P[y<Y<y+dy\,|\,R=r]\cdot P[R=r] \\ &= \frac{(\gamma/\tau)^{r\gamma}(yw)^{r\gamma-1}}{\Gamma(r\gamma)}\exp\left\{-\frac{\gamma}{\tau}yw\right\}\cdot\frac{(w\lambda)^r}{r!}\exp\{-w\lambda\}\,w\,dy \\ &= \left\{\lambda(\gamma/\tau)^{\gamma}y^{\gamma}w^{\gamma+1}\right\}^r\frac{1}{r!\Gamma(r\gamma)y}\exp\left\{\frac{w}{\phi}\left(-y\frac{\gamma\phi}{\tau}-\lambda\phi\right)\right\}dy \\ &= \left\{\frac{(w/\phi)^{\gamma+1}y^{\gamma}}{(p-1)^{\gamma}(2-p)}\right\}^r\frac{1}{r!\Gamma(r\gamma)y}\exp\left\{\frac{w}{\phi}\left(y\frac{\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}\right)\right\}dy. \end{aligned} \tag{A.1}$$

Hence the density of Y can be obtained summing over all possible values of R:

$$\begin{aligned} f_Y(y;\mu,\phi/w,p) &= \sum_r f_{R,Y}(r,y;\lambda,\tau,\gamma) \\ &= \left(\sum_r\left\{\frac{(w/\phi)^{\gamma+1}y^{\gamma}}{(p-1)^{\gamma}(2-p)}\right\}^r\frac{1}{r!\Gamma(r\gamma)y}\right)\exp\left\{\frac{w}{\phi}\left(y\frac{\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}\right)\right\} \\ &= c(y;\phi/w,p)\cdot\exp\left\{\frac{w}{\phi}\left(y\cdot\theta-\kappa_p(\theta)\right)\right\}. \end{aligned} \tag{A.2}$$

This proves that Y belongs to the exponential dispersion family $\mathrm{ED}^{(p)}(\mu, \phi/w)$.

Acknowledgements

The author thanks the anonymous referees for their remarks which have substantially improved this manuscript, especially concerning Subsection 4.2 and the examples section.

References

[1] England, P.D., and Verrall, R.J. (2002) Stochastic claims reserving in general insurance, Institute of Actuaries and Faculty of Actuaries, http://www.actuaries.org.uk/files/pdf/sessional/sm0201.pdf

[2] Jørgensen, B., and de Souza, M.C.P. (1994) Fitting Tweedie's compound Poisson model to insurance claims data, *Scand. Actuarial J.*, 69-93.

[3] Mack, T. (1991) A simple parametric model for rating automobile insurance or estimating IBNR claims reserves, *Astin Bulletin* **21**, 93-109.

[4] Mack, T. (1997) Measuring the variability of chain ladder reserve estimates, *Claims Reserving Manual* **2**, D6.1-D6.65.

[5] McCullagh, P., and Nelder, J.A. (1989) Generalized linear models, 2nd edition, Chapman and Hall.

[6] Renshaw, A.E., and Verrall, R.J. (1998) A stochastic model underlying the chain-ladder technique, *B.A.J.* **4**, 903-923.

[7] Smyth, G.K. (1996) Partitioned algorithms for maximum likelihood and other nonlinear estimation, *Statistics and Computing* **6**, 201-216.

[8] Smyth, G.K., and Jørgensen, B. (2002), Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling, *Astin Bulletin* **32**, 143-157.

[9] Taylor, G. (2000), Loss reserving: an actuarial perspective, Kluwer Academic Publishers.

[10] Tweedie, M.C.K. (1984), An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions. Proceeding of the Indian Statistical Golden Jubilee International Conference*, Eds. J.K. Ghosh and J. Roy, 579-604, Indian Statistical Institute, Calcutta.

[11] Wright, T.S. (1990), A stochastic method for claims reserving in general insurance, *J.I.A.* **117,** 677-731.

Mario V. Wüthrich
Winterthur Insurance
Römerstrasse 17, P.O. Box 357,
CH-8401 Winterthur
Switzerland
E-mail: mario.wuethrich@winterthur.ch

INTEREST-RATE CHANGES AND THE VALUE OF A NON-LIFE INSURANCE COMPANY

BY

THOMAS ALBRECHT, ALLIANZ AG

ABSTRACT

How does a change in the risk-free interest-rate affect the value of a non-life insurance company or portfolio? Risk managers typically argue that there should be little impact as long as assets and liabilities are properly matched. However, the risk-management perspective focuses on existing assets and liabilities, while neglecting the value of future business potential. This paper argues that interest-rate changes can have a significant impact on the appraisal value of a non-life insurance company, even if assets and liabilities are matched. This impact can be positive as well as negative, depending on the underlying parameters. Relevant parameters include reserving intensity, combined ratio, business growth-rate, asset allocation, risk-capital relative to premium income and the correlation between interest-rate and technical insurance results.

KEYWORDS

Valuation, Interest-Rates, Asset-Liability-Management

INTEREST-RATE CHANGES AND THE VALUE OF A NON-LIFE INSURANCE COMPANY

What impact does a change in the risk-free interest-rate have on the present value of a non-life insurance company or portfolio?

The topic of interest-rate changes has received substantial attention in the case of life-insurance companies (see, among others, BABBEL (1995), HOLSBOER (2000), DICKINSON (2000) and SIGLIENTI (2000)), as in several countries the industry has had to cope with very low market interest-rates, while long-term contracts with guaranteed minimum returns had to be honored.

For non-life companies, the issue has been discussed less extensively. Several authors have analyzed proper asset-liability-management in the face of interest-rate risks. However, their work tends to focus on existing assets and liabilities (e.g. D'ARCY / GORVETT (2000) and CAMPBELL (1997)), while neglecting the impact of interest-rate changes on the value of future business potential. Such an approach is appropriate for short-term risk-management purposes (i.e. safeguarding the required solvency-capital at any given point in time) as well as for evaluating a pure run-off portfolio. To assess the interest sensitivity of a going-concern appraisal value (or market value), an extended framework is needed. So far, attempts in this direction have been made in the context of dynamic financial analysis[1], and in a paper by PANNING (1995). However, PANNING – among other simplifications – only analyses the case of a break-even combined ratio and uses discount-rates without risk-adjustment[2]. This paper tries to provide a more extended treatment of the issues involved.

1. BASIC DCF-VALUATION OF A NON-LIFE INSURANCE COMPANY

The available literature on non-life insurance DCF-valuation is comparatively small. A discussion of methodology and relevant problems can be found in ALBRECHT (2001), COPELAND / KOLLER / MURRIN (2000) and HARTUNG (2000).

The profit and loss account of a non-life insurer can – for example - be summarized in the following way:

Earned net premiums	
–	losses incurred
–	administration and acquisition expenses
=	technical result
+	investment income on insurance reserves
+	investment income on equity
+/–	other íncome / expense
=	profit before tax
+/–	taxes
=	profit after tax
–	retained profit
=	dividend

1 Dynamic financial analysis subjects financial models to scenario testing or stochastic simulation to assess the impact of future events on the financial position of a company. For an overview, see D'ARCY / GORVETT / HERBERS / HETTINGER (1997) or CASUALTY ACTUARIAL SOCIETY (1999).

2 Another paper dealing with market values, STAKING / BABBEL (1995), empirically examines the effect of asset duration (a proxy for interest-risk exposure) on the market-value of non-life insurers (more specifically, on Tobin's q). However, no attempt is made to discuss the effect of interest-rate *changes* on company value. Instead, the paper only deals with the connection between interest *exposure* and valuation premiums.

To calculate the company value using a flow-to-equity approach, the future (potential) dividends are discounted at an adequate discount-rate.

If we use the simplifying assumption that dividends grow at a constant annual rate g (this will typically not hold in real-life situations, but does not restrict the general analysis conducted in this article), yearly dividends can be written as follows:

year	***1***	***2***	***3***	***4***
dividend	a	$a(1+g)$	$a(1+g)^2$	$a(1+g)^3$
present value	$a/(1+d)$	$a(1+g)/(1+d)^2$	$a(1+g)^2/(1+d))^3$	$a(1+g)^3/(1+d)^4$

As a perpetuity, the present value of this dividend stream can be written as: $a/(d-g)$

notation: a expected amount of next dividend payment
g growth-rate
d discount-rate

Assuming that profit retention is determined by the company's growth (the required risk-capital will typically increase roughly proportional to premium income), company value V can be written as (in % of mark-to-market equity):

$$V = ((r + ix + tr\,np)(1-t) - g)/(i + pr - g)$$

with: $a = (r + i\,x + tr\,np)(1-t) - g$
= profit after tax and after growth-related profit-retention

d = $i + pr$
= risk-free interest-rate + adequate risk-premium

notation: i risk-free interest-rate
r expected investment return on shareholders' funds[3], with $r = i + z >= i$
z risk-premium earned on investments (i.e. expected return in excess of risk-free rate)
x insurance reserves in % of equity
tr technical result in % of net premiums (tr = 1 – combined ratio; combined ratio = cr)
np net premiums in % of equity

[3] It is assumed that reserves are invested at the risk-free rate, whereas any risky investments are financed through the company's equity. All returns should be 'normalized', i.e. reflect expected average future earnings including expected (unrealized) capital gains.

t tax-rate
pr risk-premium as part of discount-rate

Some comments on plausible real-life parameters:

- *i* is defined as constant for all maturities. This is obviously not compatible with reality and is only assumed to simplify the model. DRUKARCZYK (1998), p. 330ff., describes how to model interest-rates in more detail.
- *pr* should be equal to the market risk-premium, if the company's investment and insurance risks imply a beta of roughly one. However, depending on the specific company – and especially on the investment policies followed –, beta may be considerably higher or lower than one[4]. Annual equity market risk-premiums are typically assumed in the range of 4 – 6 percentage points, though views on adequate risk-premiums differ widely[5].
- The assumed long-term growth-rate should be lower than the long-term interest-rate, if the empirical patterns of the past do not fundamentally change in the future (the nominal growth-rate of developed economies is typically lower than the nominal interest-rate). This obviously does not rule out higher growth-rates for the short- and mid-term in specific cases (or even the insurance industry as a whole).
- The risk-premium *z* earned on investments financed by shareholders' funds should not exceed the risk-premium pr used in the cost-of-capital: The risk-level of the company assumed by choosing pr would already be fully made up of investment risk, if value = equity and $z = pr$ (Implicitly, it is assumed that the company earns risk-equivalent returns on investments and does not have the ability to earn excess-return through superior investment skills. If this were not the case, z might not be purely a risk-premium, but could also include excess returns earned). If value >> equity, then $z > pr$ is in principle possible. However, since the technical insurance business should also be risky, and therefore required to earn a risk-premium, the assumption $z < pr$ seems sensible. For more details on the decomposition of systematic risk in investment and insurance-risk, see ALBRECHT (2001).
- A long-tail line of business with high insurance reserves *x* will likely imply comparatively high combined ratios, as there will be more financial investments and therefore a better financial result to subsidise the technical insurance losses (e.g. SWISS RE (2001)).
- A higher proportion of equity relative to net premiums will typically be necessary for long-tail businesses with high amounts of reserving. Alternatively, for a given line-of-business, a high proportion of equity should imply

[4] The risk of the company to be valued may depend on the choice of parameters: This paper assumes that an interest-rate change does not affect the risk-premium. However, this may be incorrect, as an interest-rate increase leads to higher risk-free payments on investments, thus potentially affecting the riskiness of the total cash-flow stream.

[5] ALBRECHT (1999b), chapter 2, discusses the existing evidence on equity market risk-premiums.

a low risk-premium *pr*, as the risk per unit of capital will be lowered, if the amount of capital is increased.

All other things being equal, an increase in the technical result (a decrease in the combined ratio) obviously increases company value *V*. Higher claims reserves (a longer duration of the run-off) also imply an increase in *V*, as they positively impact investment earnings. A higher risk-premium, on the other hand, negatively impacts *V*.

Not immediately clear is the effect of a higher growth-rate (lower discounting, but also higher profit retention), the net-effect hinges on the profitability of growth (i.e. if profit retention leads to sufficient future profits to earn cost-of-capital for the retained capital)[6].

Also not immediately clear is the effect of a higher interest-rate. This effect is the main topic of this paper.

2. Effects of Interest-Rate Changes on Company Value

The asset-liability-management literature – e.g. D'Arcy / Gorvett (2000) and Campbell (1997) – typically compares the interest-sensitivity of existing assets and liabilities: If both sides of the balance-sheet react to interest-rate changes in exactly the same way, the residual value of balance-sheet equity is not affected.

This approach is different from looking at the present value of future cashflows: Conceptually, the present value can be disaggregated into three components:

(1) The present value of future investment income and maturity refunds on *existing* assets.
(2) The present value of future claims payments on *existing* business.
(3) The present value of net payments from *future* insurance business.

Asset-liability management typically looks at (1) and (2), while neglecting (3). In contrast, the DCF-valuation formula for the appraisal value incorporates (1), (2) and (3).

In the context of this paper, the effect of interest-rate changes on the full appraisal value (fundamental market-value) of an insurance company is examined. Conceptually, this is similar to the work of Panning (1995) and the field of dynamic financial analysis.

We can calculate the partial derivative of DCF-company value with respect to i to derive the value impact of a change in i. In doing so, we can either assume that only i alone changes, or that there are correlations with other variables (e.g. inflation, insurance-rates, share prices).

---

[6] $dV/dg = (-(i + pr - g) + (r + i * x + tr * np)(1 - t) - g) / (i + pr - g)^2$
> 0 for $(r + i * x + tr * np)(1 - t) > i + pr$
i.e. for return on equity > discount-rate

At first it is assumed that all investments are short-term (or at variable interest-rates). Subsequently, adjustments are made for the case of asset-liability-matching or other longer-term investment strategies.

2.1. Pure Changes in the Real Interest-Rate with All Other Parameters Unchanged

Let us assume that an interest-rate change does not affect any other parameter. Then, the partial derivative of company value V with respect to i is:

$$dV/di = (((1+x)(i+pr-g)-((i+z)+i*x+tr*np))(1-t)+g)/(i+pr-g)^2$$
$$> 0 \quad \text{for} \quad ((1+x)(pr-g)-tr\,np-z)(1-t)+g) > 0$$

(Note that $r = i + z$. The derivation assumes that the technical result is unaffected by interest-rate changes, $d\,tr\,/\,d\,i = 0$. Empirical evidence on this assumption is given in section 2.3. Also, it is assumed that accounting is based on undiscounted reserves. In the case of reserve-discounting, an increase in i decreases today's reserves, improving today's technical result. However, the discount unwinds in the future, leaving no overall net effect, apart from potential taxation issues.)

All other things being equal, it therefore follows that:

- The higher the *cost-of-capital risk-premium pr* and/or the *combined ratio cr* and/or the *tax-rate t*, the more likely will a higher interest-rate *increase* company value *V*.
- The higher the *insurance reserves x*, the more likely will a higher interest-rate *increase V* for $pr > g$, and *decrease V* for $pr < g$.
- The higher the *growth-rate g* and/or the *investment risk-premium z*, the more likely will a higher interest-rate *decrease V*[7].
- The higher *premium income relative to equity*, the more likely will a higher interest-rate *increase V* for $cr > 100\%$, and *decrease V* for $cr < 100\%$.

How can these effects be explained economically?

- The *combined ratio (cr)* is (by assumption) unrelated to interest-rates. According to the derivation above, an increase in the interest-rate implies a change in company value of $-\,tr\,np\,(1-t) = (cr-1)\,np\,(1-t)$. For combined ratios above 100%, the change in V is positive (the technical result taken on its own contributes a negative present value, which is reduced by higher interest-rates), for lower combined ratios it is negative (the present value of positive technical results is also reduced by higher interest-rates). In general, the higher the combined ratio, the higher the increase in *V* (or the lower the decline in *V*) in case of an interest-rate increase.

[7] In the case of *g*, the effect may be reversed if *x* is very small and/or *t* very high.

- A higher *cost-of-capital risk-premium (pr)* makes value increases in case of higher interest-rates more likely, as it softens the relative increase of the discount rate (with increasing pr, a given absolute change in i implies a smaller percentage change in $d = i + pr - g$).
- A higher *investment risk-premium (z)* has the opposite effect: It lowers the relative increase in investment yield on shareholders' funds when interest-rates go up, as a given change in *i* implies a smaller percentage change in $r = i + z$. However, as long as $d > r$, the present value of the future return on shareholders' funds still goes up with increases in *i* (as the relative increase in the discount rate is still smaller than the relative increase in the investment yield).
- Future investment income is leveraged by the amount of *insurance reserves (x)*. If the growth-adjusted discount-rate ($d = i + pr - g$) is higher than the interest earned on investments (i.e. $pr > g$), higher interest-rates increase the present value of future investment income. This happens because an increase in *i* has a smaller absolute effect on the discount-rate than on investment yield.
- The effects of the *growth-rate (g)* are hardest to discuss. On one hand, growth-related profit retention reduces distributable profits. An increase in *i* lowers the (negative) present value impact of not being able to distribute those profits. On the other hand, higher growth increases future profits. Insofar as those future profits are not interest-sensitive, an increase in *i* lowers their present value. If the company is profitable (i.e. return on equity > cost of capital), the value of future profits will be higher than the value of required profit retention, thus implying a negative overall effect of higher growth in case of an interest increase. If a company is unprofitable (i.e. technical result and investment returns are insufficient to finance the profit-retention required for long-term growth), the effect can go into reverse. However, as future investment returns are interest-sensitive, their present value does not decrease with an interest-rate increase. Thus, higher growth-rates negatively impact value sensitivity when interest rates go up, even if a company is unprofitable, except if insurance reserves are very small.

Table 1 shows some parameter constellations that can typically be observed in real-life situations (the table assumes a tax-rate of 35%).

As can be seen, for combined ratios above 100% an interest-rate increase only implies a drop in value in the case of high growth rates. For combined ratios considerably below 100%, however, an interest-rate increase is value-reducing under many realistic sets of parameters: If a company e.g. operates with a combined ratio of 95%, an interest-rate increase would always be value-reducing, except if the company exhibits high insurance reserves as well as low growth-rates.

The extent of the change in value is especially high, if interest-rates and growth-rates are low: If growth is around zero and the interest-rate is 3%, an interest-rate change of only one percentage point changes the value by 10-20%. At g = 0 and i = 2%, di = 0.01 even implies value changes of up to 30%.

TABLE 1

VALUE-IMPACT OF INTEREST-RATE CHANGES USING DIFFERENT SETS OF PARAMETERS

assumed parameters					***increase in interest-rate***	***value-change at***
pr	***g***	**1/*np***	***x***	***z***	***implies value-increase if***	***di = 1%, cr=100%***[8]
4% / 6%	**0%**	**25%**	1	0%	*cr* > 98.0% / 97.5%	16.7% / 20.0%
4% / 6%	0%	25%	1	3%	*cr* > 98.8% / 98.3%	6.9% / 10.0%
4% / 6%	0%	25%	2	3%	*cr* > 97.8% / 96.3%	9.4% / 12.5%
4% / 6%	0%	25%	4	0%	*cr* > 95.0% / 92.5%	16.7% / 20.0%
4% / 6%	0%	25%	4	3%	*cr* > 95.8% / 93.3%	11.8% / 15.0%
4% / 6%	**5%**	25%	**1**	3%	*cr* > 99.4% / 98.4%	3.0% / 5.9%
4% / 6%	5%	25%	**2**	3%	*cr* > 99.6% / 98.1%	1.1% / 4.0%
4% / 6%	5%	25%	**4**	3%	*cr* > 100.1%/ 97.5%	– 0.1% / 2.8%
4% / 6%	**10%**	25%	**1**	0%	*cr* > 99.2% / 98.2%	4.0% / 7.0%
4% / 6%	10%	25%	1	3%	*cr* > 100.0%/ 99.0%	0.4% / 3.2%
4% / 6%	10%	25%	**2**	3%	*cr* > 101.5%/100.0%	– 2.6% / 0.1%
4% / 6%	10%	25%	**4**	0%	*cr* > 103.7%/101.2%	– 3.7% / 0.9%
4% / 6%	10%	25%	4	3%	*cr* > 104.5%/102.0%	– 4.2% / – 1.5%
4% / 6%	**0%**	**50%**	**1**	0%	*cr* > 96.0% / 94.0%	16.7% / 20.0%
4% / 6%	0%	50%	1	3%	*cr* > 97.5% / 95.5%	6.9% / 10.0%
4% / 6%	0%	50%	**4**	0%	*cr* > 90.0% / 85.0%	16.7% / 20.0%
4% / 6%	0%	50%	4	3%	*cr* > 91.5% / 86.5%	11.8% / 15.0%
4% / 6%	**5%**	50%	**1**	3%	*cr* > 98.7% / 95.2%	3.0% / 5.9%
4% / 6%	5%	50%	**4**	3%	*cr* > 100.2%/ 95.2%	– 0.1% / 2.8%
4% / 6%	**10%**	50%	**1**	0%	*cr* > 98.3% / 96.3%	4.0% / 7.0%
4% / 6%	10%	50%	1	3%	*cr* > 99.8% / 97.8%	0.4% / 3.2%
4% / 6%	10%	50%	**4**	0%	*cr* > 107.3%/102.3%	– 3.7% / 0.9%
4% / 6%	10%	50%	4	3%	*cr* > 108.8%/103.8%	– 4.2% / – 1.5%

2.2. Inflation-Induced Interest-Rate Changes

D'ARCY / GORVETT (2000) examine the impact of inflation on asset-liability management strategies for a run-off portfolio. They conclude that the "effective duration" of liabilities is lower, if interest-rate changes are correlated with changes in inflation. Put differently: If you have liabilities with a duration of *x* years, locking-in an equivalent asset-duration of *x* years only leads to proper asset-liability-matching, if the liability payments are not subject to changing

[8] Interest-rates assumed: 3% (g = 0), 8% (g = 5%), 13% (g = 10%). The higher the assumed interest-rate, the lower the effect of an interest-rate change on company value.

inflation-rates. If the liabilities are subject to changing inflation, the asset duration has to be shortened, so that the investment income received will also be subject to changes in inflation through the link between inflation and interest-rates.

The same logic will now be applied to the DCF-valuation formula.

In case of purely inflation-induced interest-rate changes, $dg/di = 1$ follows, if premiums and profit also grow at the rate of inflation. In this case we can write:

$$dV/d = ((1 + x)(1 - t) - 1)\,(i + pr - g)/(i + pr - g)^2 = (x(1 - \mathrm{t}) - \mathrm{t})\,/\,(i + pr - \mathrm{g})$$
$$> 0 \quad \text{for} \quad x\,(1 - t) > t \text{ and } i + pr > g$$
$$\text{for} \quad \text{(e.g.) } x >= 1 \text{ and } t < 0{,}5 \text{ and } i + pr > g$$

An increase in inflation with other parameters unchanged seems to have a positive effect on company value, if the – not very restrictive – assumptions $x >= 1$, $t < 0.5$ hold.

However, this derivation is incomplete and therefore incorrect: The deduced value increase results from higher nominal interest income on the investments that are funded by insurance reserves. But a higher rate of inflation also implies a nominal increase in future claims payments, leading to a worsening technical run-off result. In other words, the implicit assumption of $d\,(tr\ np)/d\,i = 0$ is no longer adequate. If claims inflation equals the general inflation rate, an inflation-induced interest-rate increase will not affect company value, as the technical result of subsequent calendar years will worsen when inflated claims are settled, exactly offsetting the higher nominal interest income. This becomes immediately apparent in a balance-sheet with discounted claims reserves: If the real interest-rate goes up, insurance reserves are discounted at a higher rate, implying a lower present value. If nominal interest-rate and inflation go up by the same amount, discount-rate and future nominal claims payments increase at the same rate, leaving the discounted reserves unchanged. In subsequent years, even though the higher nominal rate implies higher investment yields, reserves are also being inflated more quickly, with both effects offsetting each other.

For valuation purposes this implies that an increase in inflation has to be reflected not only in nominal interest-rate and growth, but also in the run-off result. In countries with high rates of inflation it may therefore be easier to use real instead of nominal figures.

However, a real approach neglects the potential negative tax effect of an inflation increase: For $t = 0$, the higher interest earned on shareholders' funds is exactly sufficient to finance the necessary growth in nominal equity, as $dg = di$. For $\mathrm{t} > 0$, interest on shareholders' funds goes up only by $(1 - \mathrm{t})\,di$, insufficient to fund growth of $dg = di$. In this case, company value goes down if inflation goes up. However, the effect depends on the exact tax regime in place (this paper assumes a corporate tax rate and an additional income tax-rate which is applied to dividends as well as fixed income investments. While this is a reasonable assumption in many countries, it clearly does not apply in all cases. Under new German tax-rules, for example, income tax on dividends is lower than on fixed income, with potential implications for the after-tax cost-of-capital calculation).

2.3. Correlation Between Interest-Rate Changes and Insurance Profit

So far it was assumed that interest-rate changes do not affect other parameters: Neither did we assume a correlation with equity returns on the investment side, nor a correlation with the technical insurance result.

It is frequently argued that an increase in interest-rates has a negative effect on share prices, as future cash-flows are discounted more steeply[9]. Strictly speaking this only holds for changes in the real interest-rate (not for purely inflation-induced interest-rate changes), and only if an increase in the real interest-rate is not correlated with an increase in the future real profit potential. As argued in ALBRECHT (1999a, p. 127ff.), there is (ambiguous) evidence for such a correlation, so the relationship is less clear than it may seem at first glance.

However, a rough look at empirical data seems to validate the negative correlation between (nominal) interest-rates and share-prices: For the years 1981-99, the correlation between yearly returns of the German DAX and the German money market yield was – 0.31. The change in the money market yield was also negatively correlated with the DAX-return, at – 0.21[10].

The technical result may also be correlated with the interest-rate: If industry participants think that higher interest-rates imply an improvement in the financial result, they might be tempted to increase the competitive pressure on rates (PANNING (1995), DICKINSON (2000)). In this case, an interest-rate increase should be correlated with a worsening technical result (a higher loss ratio).

The yearly loss ratios of German non-life insurers for the years 1966-99[11] are positively correlated with the money-market yield (+0.30). Using first differences (i.e. changes in the interest-rate correlated with changes in the loss ratios), the positive correlation persists (not lagged: +0.18; lagged one year: +0.24).

The correlation differs considerably between lines of business, however: While (e.g.) fire and transport had high positive correlations (+0.54 and +0.38), liability and accident showed negative correlations (– 0.49 and – .45).

Within the scope of this article, a detailed quantification of correlation effects shall not be attempted. For example, some of the correlations listed above might simply be spurious. However, it should be clear from the rough

---

[9] See e.g. D'ARCY / GORVETT (2000), p. 396.

[10] DAX-return taken from GDV, table 100. Money-market yield from ALBRECHT (1999a), app. A 2.1. Similarly, ALBRECHT (1998), p. 263, reports a positive correlation between DAX and REX in the period 1988-96. As REX is a fixed-income performance-index, higher interest-rates imply lower index-values. Consequently, a positive correlation between DAX and REX indicates a negative correlation between DAX and the general level of interest-rates.

[11] Source: GDV (several tables). Money-market yield from ALBRECHT (1999a), app. A 2.1.

calculations performed above that correlations may have a major impact and merit further study.

2.4. Effect of Asset Liability Matching and Other Investment Strategies

If fixed-income investments are chosen to mature according to the expected timing of claims payments, the average term to maturity of the portfolio will depend on the ratio of reserves to premium income. If the payment structure is constant over time and the portfolio does not grow, claims reserves equaling yearly claims imply a medium payment period of 6 months. In the case of unearned premiums, the average payment period is also 6 months, if contracts are spread evenly over time with yearly renewal[12].

The value-effect of such an investment strategy in the case of an interest-rate change can be approximated via the change in present value of a fixed interest payment over the average payment period. The present value of fixed-income investments (excluding reinvestment risk, i.e. assuming zero-bond investments) can be approximated as[13]:

$$\text{Present Value (in \% of company equity)} = x\,(1 + i_f\,(1 - t))^m / (1 + i + pr)^m$$

with: i_f fixed interest-rate
m average duration of insurance reserves

Table 2 shows some examples for value changes (in % of equity)[14]:

TABLE 2

CHANGE IN INTEREST-SENSITIVITY WHEN INVESTMENTS ARE MATCHED WITH LIABILITIES

claims reserves % of claims	*claims ratio*	*unearned pr. % of net pr.*	*equity % of net pr.*	*–> equals x*	*m*	*value effect of di = 0,01*
40%	50%	50%	25%	2.8	0.41	– 1.0%
100%	50%	50%	25%	4.0	0.50	– 1.8%
100%	50%	50%	50%	2.0	0.50	– 0.9%
200%	50%	50%	50%	3.0	0.83	– 2.2%
300%	50%	50%	50%	4.0	1.25	– 4.2%

[12] Assuming a stagnant portfolio, 1/12 of unearned premiums will be earned in 12 months' time, 1/12 in 11 months, etc. On average, the unearned premiums on the balance-sheet at any point in time will therefore remain for 6 months. If the portfolio is growing, a larger part of unearned premiums will be earned in later months, implying a longer average duration than 6 months.

[13] Calculation excludes investments backed by shareholders' funds. It may be argued that the discount-rate should not include the full risk-premium *pr*, as the fixed-income portfolio in itself is not subject to the full company risk. However, omitting *pr* hardly affects the interest-sensitivities.

[14] For $i_f = i = 0.05$, $pr = 0.04$ und $t = 0.35$. Changes in *pr* and *t* have no major effect on the result.

As can be seen, asset-liability-matching results in rather short terms to maturities for broadly diversified p/c-insurers: Even if parts of a large insurance portfolio have a very long run-off period, an average period until maturity of more than a year ($m > 1$) should be the exception rather than the rule, as it would imply a ratio of claims reserves to claims considerably in excess of 200%. As a consequence, an interest-rate increase has rather small effects on the present value of liability-matched asset portfolios for all but insurers with long tails (high values for x and m).

To give an example, the non-life operations of German insurer Allianz AG show total (net) claims provisions of 45.3 *b* Euro in 2001 and net earned premiums of 34.4 *b* Euro. The ratio of claims reserves to premiums is 132% (ALLIANZ (2002), p. 4 and p. 48), the average duration equals roughly 0.6 years (using the simplified assumptions of stable business volume and constant payment pattern).

For specialized long-tail insurers, the ratio can be much higher, though. One example for longer-tail business is MAT (Marine Aviation Transport). In its 2000 financial statements, AGF MAT (a subsidiary of Allianz AG) shows (net) claims provisions of 384 m Euro and net earned premiums of 132 *m* Euro. This equals a ratio of 291%. As there are no unearned premiums at year-end (MAT-contracts typically start at the beginning of the calendar-year), the ratio implies an average duration of roughly 1.5 years (17.5 months).

Other business lines, e.g. workers' compensation or bonding, may exhibit much longer average durations, though they are rarely run as stand-alone companies.

Insurers frequently hold portfolios with durations far in excess of asset-liability-matching. Table 3 shows the present value effects of an average of several years until maturity:

TABLE 3

CHANGE IN INTEREST-SENSITIVITY WHEN INVESTMENT MATURITIES ARE LENGHTENED

	value effect of di = 0,01 for time until maturity of			
x	***2***	***3***	***4***	***5 years***
1	– 1.6%	– 2.3%	– 2.9%	– 3.4%
2	– 3.2%	– 4.6%	– 5.8%	– 6.8%
4	– 6.5%	– 9.2%	– 11.6%	– 13.6%

If reserves are high compared to equity (high values of x), a portfolio duration of several years quickly leads to a double-digit impact on the company's mark-to-market equity, even when interest-rates increase by only one percentage point.

3. Case-studies on the Total Effect of Interest-Rate Changes

As discussed, different sets of parameters will imply very different reactions of company value to interest-rate changes. This chapter intends to give examples of companies that are quite susceptible or less susceptible to interest-rate risk[15].

Non-company-specific parameters used throughout are $i = 0.05$, $pr = 0.04$ and $t = 0.35$.

> *Company A is a retail insurer with a high franchise value (low combined ratio, high growth rate). Parameters are as follows: cr = 0.95, g = 0.05 x = 4, np = 4, z = 0.*

According to the DCF-valuation formula, A's appraisal value is 6.06 times equity. Assuming short-term investments, an interest-rate increase of $di = 0.01$ implies a new value of 5.50 times equity, i.e. a drop in value of 9%.

If A had opted for a long-term investment strategy ($m = 5$ years), value would have dropped by a further 0.14 times equity, i.e. the total interest-related drop would have been 12%. Compared to total company value, the effect of the investment strategy is small, as A's franchise value is far in excess of equity. Nevertheless, long-term investments *increase* A's sensitivity to interest-rate changes.

The effect occurs irrespective of tail: If A is assumed to have a much longer tail, interest-rate increases still imply a drop in value with short-term investment strategies, and a bigger drop with longer-term investments.

> *Company B is considerably less profitable, but also boasts a high growth-rate. Parameters are: cr = 1.02, g = 0.05, x = 4, np = 4, z = 0.*

B's appraisal value is 1.51 times equity. Assuming short-term investments, $di = 0.01$ implies a new value of 1.86 times equity, i.e. an increase in value of 23%.

If *B* had opted for a long-term investment strategy ($m = 5$ years), the negative value-impact would again have been 0.14 times equity. Total value after the interest-change would then be 1.72 times equity, an increase of 14% compared to the initial value.

For *B*, a long-term investment portfolio actually *decreases* the interest-sensitivity of the company's appraisal value, even though *B* is not a long-tail insurer. This happens because the future profits of the company are earned through investment income, while the technical result is negative. The higher investment income overcompensates the higher discount-rate for the parameters chosen.

B has a second option to lower its interest-sensitivity: It can increase the portion of equity that is invested in risky assets (e.g. stocks): Assuming e.g. $z = 0.03$, $di = 0.01$ only increases company value by 13% (short term

[15] Unless otherwise mentioned, only changes in the real rate are considered.

investments) / 6% (long-term investments). This happens because the additional expected return (z) is assumed to accrue irrespective of interest-rate increases. A higher interest-rate lowers the present value of the risk-premium earned, another counter-cyclical effect on *B*'s value.

> *Company C is a long-tail industrial insurer with low profitability (high combined ratio) and low growth perspectives. Parameters are: $cr = 1.1$, $g = 0$, $x = 8$, $np = 3$, $z = 0$.*

C's appraisal value is 1.08 times equity. Assuming short-term investments, $di = 0.01$ implies a new value of 1.56 times equity, i.e. an increase in value of 44%.

For *C*, the higher investment earnings on the long tail more than overcompensate the lower present value of future profits (which is assumed to be low anyway). However, for such an insurer, the higher profitability caused by higher interest-earnings is likely to spark more aggressive competition on rates, thus increasing long-term combined ratios – an effect not covered by applying the formula on a ceteris paribus basis.

If *C* had opted for a long-term investment strategy ($m = 5$ years), the negative value-impact would have been 0.27 times equity. Total value after the interest-change would then be 1.29 times equity, still a 19% increase compared to the initial value. Again, as with B, long-term investments *decrease* the interest-sensitivity of *C*'s value.

The examples discussed above illustrate the limitations of PANNING's (1995) analysis: PANNING argues that the duration of investments should generally be lengthened to counter-balance the interest-sensitivity of future business value[16]. In his analysis, future business value has a negative duration (i.e. the value increases when interest-rates go up), because the duration of future losses is higher than the duration of future premiums. However, this only holds under his non-general assumption of break-even premiums. For sufficiently profitable companies (like company *A*), the effect goes into reverse.

What would have happened to *A*, *B* and *C* in the case of inflation-induced interest-rate changes? Assuming that a change in inflation does not affect the underlying profitability of the insurance business[17], and also neglecting potential tax-effects, the only effect of such an interest-rate change would be on the present value of longer-term fixed-income investments. The higher the inflation-risk, the higher the value sensitivity of longer-term investments. This serves to make longer-term investments less attractive, irrespective of company characteristics. Companies A, B and C should all lower their investment durations if inflation-uncertainty increases.

---

[16] Assuming – as is done here – that future premiums are fixed, i.e. not interest-sensitive.

[17] The extensive literature on the effect of inflation on business profitability cannot be discussed here due to space constraints. However, it can be argued that the results are quite inconclusive.

4. Empirical Results for Germany

It can be empirically verified if the market capitalisation of companies is correlated with interest-rate changes. However, changes in market capitalisation can only correctly reflect the fundamental effects of interest-rate changes, if market participants under-stand and correctly price those underlying fundamental effects. As this cannot be taken for granted – given the complexities of those effects – any empirical results should be interpreted with caution.

To capture the effect of interest-rate changes, unexpected changes should be used, as expected changes will already be anticipated in the share price. Consequently, the variable used to explain share-price effects cannot be the change in money-market rates, but only a performance-index for fixed-income securities.

To examine, if and to what extent the monthly stock-returns of major German companies between January 1990 and May 2001 can be explained by the REX-return (REX is a performance-index based on German fixed-income government securities of different maturities; regression is performed on the basis of Bloomberg-data), the following regression was performed (results shown in table 4):

$$r_{\text{company, i}} = a + b * r_{\text{rex}, i} + \varepsilon_i$$

with: $r_{\text{company}, i}$ return of the company's stock in month i
$r_{\text{rex}, i}$ return of the REX-index in month i

TABLE 4

Empirical interest-sensitivity of German companies' market capitalisation

company	*coefficient*	*standard error*	R^2
insurance			
Allianz	2,47	0,64**	0,10
Munich Re	2,17	0,78**	0,05
banking			
Commerzbank	1,16	0,64	0,02
Deutsche Bank	0,69	0,74	0,01
Dresdner Bank	0,99	0,76	0,01
non-financial sector			
BASF	0,82	0,63	0,01
Bayer	1,29	0,61*	0,03
SAP	– 2,10	1,12	0,03
Siemens	0,67	0,78	0,01
Volkswagen	1,12	0,82	0,01

* = The two-sided hypothesis "coefficient = 0" can be rejected at 95% level.
** = The two-sided hypothesis "coefficient = 0"can be rejected at 99% level.

With the exception of SAP, the coefficient of the REX-return is positive in all cases, i.e. an increase of REX (a decrease in the return of fixed-income securities) tends to imply an increase in share prices. However, for banks as well as non-financial companies, the standard error is nearly as high as the coefficient, and R^2 is practically zero. The only exception are the two insurers Allianz and Munich Re: The standard error is low relative to the coefficient, and R^2 implies an explanatory content of 5% and 10% of total share-price variance. This may sound low at first glance, but given the multitude of effects that influence share-prices, a higher explanatory content of interest-rate changes alone cannot plausibly be expected.

The result of this – rather rough and superficial – empirical analysis is that interest-rate changes do not in general have a strong effect on the market capitalisation of German companies. However, interest-rate increases do have a discernible (negative) impact on the market capitalisation of German insurance companies.

So is this in line with the theoretical analysis?

The financial statements of Allianz AG for the financial year 2000 show lendings and fixed-income securities of 189 *b* Euro, of which 91 *b* Euro with a remaining term to maturity of more than 5 years. Roughly 2/3 of investments are linked to life/health-business, where the bulk of investment income is passed on to policyholders. Assuming proportional allocation, roughly 60 *b* Euro of total investments – of which 30 *b* Euro with a remaining term to maturity of more than 5 years – should directly affect company value. It therefore sounds plausible to assume an average term to maturity of 5 years for the 60 *b* Euro in total investments. If we compare this to the company's equity of 35 *b* Euro, the result is roughly $x = 2$, implying a negative value effect of –3.8% if interest-rates go up by one percentage point. At the same time, the base-effect of an interest-rate increase (as discussed in 2.1) should be close to zero in the case of Allianz AG, if we assume $pr = 0.04$, $x = 2$, cr around 1.00, and $g = 0.05$ or less. Finally, the market may expect a negative correlation between interest-rates and technical insurance-results. In total, the theoretical effect of an interest-rate increase on the value of Allianz AG may therefore truly be negative, just like the empirical results indicate.

However, the empirical results can only be interpreted with great caution, as multi-national companies like Allianz and Munich Re will not only be affected by changes in German interest-rates, but by other countries' interest-rates as well. Furthermore, as those companies also write life and health-business, they might be subject to other effects which are not covered by this analysis.

5. Conclusions

Without a doubt the analytical model developed in this paper is simplified in several respects. To give just one example, it could be explicitly modelled that interest-rate changes are not identical over the whole yield-curve, but instead tend to be less pronounced for long maturities[18].

Nevertheless, the analysis leads to some general conclusions. Specifically, it was shown that a change in the interest-rate can affect the value of a non-life insurance company positively as well as negatively, depending on the underlying parameters:

If the investments are predominantly long-term (more than would be the case for mere asset-liability-matching) and average combined ratios comparatively low, a value reduction in case of increasing interest-rates is likely. This effect seems to dominate for the German insurers Allianz and Munich Re.

However, there is no reason to generalize this result: Companies / portfolios with low expected growth-rates, combined ratios around or above 100% and/or a comparatively short-term investment portfolio will on the contrary experience an increase in fundamental DCF-value if the market interest-rate goes up.

The effects of an interest-rate change on company value can be summarized as follows:

- An increase in the *real* interest-rate can affect company value positively or negatively, depending on the underlying parameters. The different effects on company value are:
  1. A lower present value of run-off claims (→ increases value).
  2. A lower present value of existing fixed-income securities (→ lowers value).
  3. A lower present value of technical profits earned from future insurance business (combined ratio < 100%) or of deficits incurred (combined ratio > 100%) (→ increases or lowers value, depending on the combined ratio). In addition, if the combined ratio worsens because rising interest-rates increase the competitive pressure, this will negatively affect the present value.
  4. A lower or higher present value of investment income in connection with future business (the lower the growth-rate of future business, and the higher the risk- premium in the discount-rate, the more likely will a higher interest-rate imply a higher present value of future investment income from future reserves).
- On the other hand, an increasing rate of inflation (with a corresponding rise in the *nominal* interest-rate) does not influence company value in a world without taxes – provided investments are short-term/variable rate – because higher interest-income will be compensated by higher nominal claims payments and a higher discount-rate for future profits (although it is possible that an increase in inflation may have consequences on real parameters and therefore company value). Longer-term fixed-income investments, however, imply a decrease in company value in case of an interest-rate increase, as claims inflation will not be compensated by higher investment returns until investments are reinvested after maturity. A further decrease in value results

[18] This follows from the market participants' expectations of future interest-rate changes. See Levin (1996), Anker (1993) or Albrecht (1999a). The simple assumption of a parallel shift in the yield-curve, as used in this paper, implies an over-estimation of long-term effects in section 2.1 relative to section 2.4.

from taxation effects under the tax regime assumed in this paper (though other tax regimes may have different effects).

References

ALBRECHT, T. (2001) Discounted-Cash-Flow Bewertung bei Sachversicherungsunternehmen, *Finanz Betrieb 5/2001*, p. 302-307.

ALBRECHT, T. (1999a) *Die Wahl der Zinsbindungsdauer*, Hamburg.

ALBRECHT, T. (1999b) *Asset Allocation und Zeithorizont*, Bad Soden/Ts.

ALBRECHT, T. (1998) Die Vereinbarkeit der Value-at-Risk-Methode in Banken mit anteilseignerorientierter Unternehmensführung, *Zeitschrift für Betriebswirtschaft* **68**, 259-273

AGF MAT (2001) *Annual Report 2000.*

ALLIANZ GROUP (2001, 2002) *Group Financial Statements 2000, 2001*

ANKER, P. (1992) *Zinsstruktur und Zinsprognose*, Giessen.

BABBEL, D.F. (1995) Asset-Liability Matching in the Life-Insurance Industry, *The Financial Dynmics of the Insurance Industry*, Altman and Vanderhoof (eds.), Chapter 11.

CAMPBELL, F. (1997) Asset/Liability Management for Property/Casualty Insurers, *The Handbook of Fixed Income Securities*, Fabozzi (ed.), Chapter 51.

CASUALTY ACTUARIAL SOCIETY (1999) Overview of Dynamic Financial Analysis, website of the Casualty Actuarial Society, *www.casact.org/research/dfa.*

COPELAND, T., KOLLER, T. and MURRIN, J. (2000) *Valuation*, 3rd Edition, New York (et al.)

D'ARCY, S. and GORVETT, R. (2000) Measuring the Interest Rate Sensitivity of Loss Reserves, *Proceedings of the Casualty Actuarial Society*, **87**, 365-400.

D'ARCY, S., GORVETT, R., HERBERS, J. and HETTINGER, T. (1997) Building a DFA Analysis Model that Flies, *Contingencies Magazine*, Nov./Dec. 1997, reprinted on the website of the Casualty Actuarial Society, *www.casact.org/coneduc/specsem/98dfa/dfamodel.htm.*

DRUKARCZYK, J. (1998) *Unternehmensbewertung*, 2nd Edition, München.

GDV (2001) *Statistisches Taschenbuch der Versicherungswirtschaft 2000*, download from *www.gdv.de.*

HARTUNG, T. (2000) *Unternehmensbewertung von Versicherungsgesellschaften*, Wiesbaden

HOLSBOER, J. H. (2000) The Impact of Low Interest Rates on Insurers, *Geneva Papers on Risk and Insurance*, **25**, 38-58.

LEVIN, F. (1996) *Die Erwartungstheorie der Zinsstruktur*, Frankfurt/M.

PANNING, W.H. (1995) Asset-Liability Management for a Going Concern, *The Financial Dynamics of the Insurance Industry*, Altman and Vanderhoof (eds.), Chapter 12.

SIGLIENTI, S. (2000) Consequences of the Reduction of Interest Rates on Insurance, *Geneva Papers on Risk and Insurance*, **25**, 63-77.

STAKING, K. and BABBEL, D. (1995) The Relation Between Capital Structure, Interest Rate Sensitivity, and Market Value in the Property-Liability Insurance Industry, *Journal of Risk and Insurance*, **62**, 690-718.

SWISS RE (2001) Rentabilität der Nichtleben-Versicherungswirtschaft: Zurück zum Underwriting, *Sigma 5/2001*, Zürich.

DR. THOMAS ALBRECHT
Treffauerstr. 38
D-81373 München
thomas.albrecht@allianz.com
Tel. +49 89 3800 7481
Fax +49 89 3800 19422

FAVORABLE ESTIMATORS FOR FITTING PARETO MODELS: A STUDY USING GOODNESS-OF-FIT MEASURES WITH ACTUAL DATA

BY

VYTARAS BRAZAUSKAS[1], ROBERT SERFLING[2]

ABSTRACT

Several recent papers treated robust and efficient estimation of tail index parameters for (equivalent) Pareto and truncated exponential models, for large and small samples. New robust estimators of "generalized median" (GM) and "trimmed mean" (T) type were introduced and shown to provide more favorable trade-offs between efficiency and robustness than several well-established estimators, including those corresponding to methods of maximum likelihood, quantiles, and percentile matching. Here we investigate performance of the above mentioned estimators on real data and establish — via the use of goodness-of-fit measures — that favorable theoretical properties of the GM and T type estimators translate into an excellent practical performance. Further, we arrive at guidelines for Pareto model diagnostics, testing, and selection of particular robust estimators in practice. Model fits provided by the estimators are ranked and compared on the basis of Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling statistics.

1. INTRODUCTION AND PRELIMINARIES

A single-parameter Pareto distribution plays a very significant role in actuarial modeling because of its conceptual simplicity and ease of applicability in practice. The cdf of the Pareto $\mathrm{P}(\sigma, \alpha)$ model is given by

$$F(x) = 1 - (\sigma/x)^{\alpha}, \quad x > \sigma, \tag{1}$$

where $\alpha > 0$ is the shape parameter that characterizes the tail of the distribution and $\sigma > 0$ is the scale parameter. When σ is assumed known, the $\mathrm{P}(\sigma, \alpha)$ model is called a single-parameter Pareto model. The assumption of σ *known* is quite typical in the actuarial literature because, as for example Philbrick (1985) states, "although there may be situations where this value must be estimated, in virtually all insurance applications this value will be selected in advance." (See also discussion by Rytgaard (1990).)

[1] Supported by a grant from the Actuarial Education and Research Fund.
[2] Supported by grants from the Casualty Actuarial Society and Society of Actuaries, with administrative support from the Actuarial Education and Research Fund, and by NSF Grant DMS-0103698.

In several recent papers Brazauskas and Serfling (2000a,b; 2001) treated robust and efficient estimation of the tail index α for various setups: for *large-* and *small-samples* and for *one-* and *two-parameter* models (corresponding to σ *known* or *unknown*). Developments presented there utilized a well-known equivalence relation between model (1) and the truncated exponential distribution $E(\mu, \theta)$ having cdf

$$G(z) = 1 - e^{-(z-\mu)/\theta}, \quad z > \mu, \tag{2}$$

for $\theta > 0$ and $-\infty < \mu < \infty$. Specifically, if random variable X has cdf F given by (1) then variable $Z = \log X$ has cdf G given by (2), with $\mu = \log \sigma$ and $\theta = \alpha^{-1}$.

In large-sample studies, for example, new robust estimators of "generalized median" (GM) type were introduced and "trimmed mean" (T) type estimators were adapted from the $E(\mu, \theta)$ model literature. These estimators were then compared with the maximum likelihood, quantile type, percentile matching, and other estimators. Using as *efficiency* criterion the *asymptotic relative efficiency* (ARE) with respect to the *maximum likelihood estimator* (MLE) and as *robustness criterion* the *breakdown point* (BP) (this is defined in Section 1.2), the GM type was seen to dominate all competitors, with the T type second best. From a practical point of view, the ARE is equivalent to the *accuracy* of the estimator and can be interpreted in terms of the length of the confidence interval (see Section 1.3 for precise discussion).

In the present paper we investigate performance of the above mentioned estimators on real data and establish — via the use of goodness-of-fit measures — that favorable theoretical properties of the GM and T type estimators translate into an excellent practical performance. The goodness-of-fit measures, defined in Section 1.1, are used here for two purposes: (i) to (formally) test the appropriateness of the estimated Pareto model for a particular data set when α is estimated by the MLE (this is defined in Section 1.3), and (ii) to evaluate and compare Pareto fits when various estimators (not only the MLE) of α are employed.

In the actuarial literature the issue of goodness-of-fit is addressed through a combination of informal methods and formal statistical tests. Most informal techniques are based on the difference (absolute or relative) between the fitted and empirical values of relevant quantities, such as the *number of claims* or *expected value of claims* for different claim layers. Additionally, for the Pareto model in particular, comparisons of $\hat{\alpha}$ with a typical value of α for the same insurance line of the entire industry are also used in the literature. (See Patrick (1980), Philbrick (1985), and Reichle and Yonkunas (1985).) Regarding formal approaches, tests based on *Kolmogorov-Smirnov (KS)* and χ^2 statistics seem to have a leading role (see, e.g., Philbrick and Jurschak (1981)). More extensive discussion on model validation principles is available in Klugman, Panjer, and Willmot (1998), Section 2.9.

As is well-known in the statistical literature (e.g., D'Agostino and Stephens (1986), p. 110), the χ^2 test is less powerful than tests based on the empirical cdf. Therefore, here we use three widely popular goodness-of-fit measures which are based on the empirical cdf — the above-mentioned KS statistic, the *Cramér-von Mises (CvM)* statistic, and the *Anderson-Darling (AD)* statistic. All these statistics (though emphasizing different aspects of discrepancy) measure the distance in some sense between the fitted model cdf $\hat{F}$ and the empirical cdf F_n. Thus estimators that lead to smaller values of these statistics are preferable.

The paper is organized as follows. First, in Section 2, we define precisely several estimators for the parameter α in (1). Next, in Section 3, the data sets are introduced, a method for data de-grouping is described, and preliminary data visualization and diagnostic tools are applied to the sets. Finally, comparisons and conclusions are presented in Section 4. Also, in Section 4.2, we arrive at guidelines for Pareto model diagnostics, testing, and selection of particular robust estimators in practice.

In the remainder of this introduction, we formulate precisely our performance criteria.

1.1. Goodness-of-Fit Measures

Let us consider a sample $X_1, \ldots, X_n$ and denote the ordered sample values by $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ and the empirical cdf by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}, \quad -\infty < x < \infty.$$

Also, for an estimator $\hat{\alpha}$, let $\hat{F}(X_{(j)})$ denote the probability assigned to $X_{(j)}$ by the model $P(\sigma, \hat{\alpha})$, for $j = 1, \ldots, n$. Note that $F_n(X_{(j)}) = j/n$, for $j = 1, \ldots, n$. The goodness of fit statistics are then defined as follows.

The KS statistic D_n:

$$D_n^+ = \max_{1 \le j \le n}\left(\frac{j}{n} - \hat{F}(X_{(j)})\right), \ D_n^- = \max_{1 \le j \le n}\left(\hat{F}(X_{(j)}) - \frac{j-1}{n}\right),$$
$$D_n = \max\left\{D_n^+, D_n^-\right\}.$$

The CvM statistic W_n^2:

$$W_n^2 = \sum_{j=1}^{n}\left(\hat{F}(X_{(j)}) - \frac{2j-1}{2n}\right)^2 + \frac{1}{12n}.$$

The AD statistic A_n^2:

$$A_n^2 = -n - \frac{1}{n}\sum_{j=1}^{n}\left\{(2j-1)\log\left(\hat{F}(X_{(j)})\right) + (2n+1-2j)\log\left(1 - \hat{F}(X_{(j)})\right)\right\}.$$

When the parameter α is estimated by $\hat{\alpha}_{\mathrm{ML}}$, critical values and formulas for significance levels for the statistics D_n, W_n^2, and A_n^2 are available in D'Agostino and Stephens (1986), Tables 4.11 and 4.12, pp. 135-136. Actually, the results developed there are for the model $E(\mu, \theta)$ but, due to the equivalence relation between (1) and (2), they can also be directly applied to the model $P(\sigma, \alpha)$.

1.2. Robustness Criterion: Breakdown Point

A popular and effective criterion for robustness of an estimator is its *breakdown point* (BP), loosely characterized as the largest proportion of corrupted sample observations that the estimator can cope with. In other words, the BP of an estimator measures the degree of resistance of the estimator to the influence of outlying observations which possibly (but not necessarily) represent contamination of a data set rather than merely unusually extreme observations generated by the target parametric model.

Brazauskas and Serfling (2000a,b) considered two types of contamination — upper and lower contamination — and, consequently, defined separate versions of BP:

> *Lower (Upper) Breakdown Point* (LBP/UBP): the largest proportion of *lower (upper)* sample observations which may be taken to a lower (an upper) limit without taking the estimator to a limit not depending on the parameter being estimated.

For modeling insurance loss data, however, contamination of the lower type is of lesser concern because the lower limit of losses is usually pre-defined by a contract. (For example, the lower limit can be represented as a deductible.) Thus, in the present treatment we favor estimators which have *nonzero* UBP.

1.3. Efficiency Criterion: Variance

If sample observations follow the postulated parametric model, then it is well-known that, for large data sets, the MLE attains (in its approximating normal distribution) the minimum possible variance among a large class of competing estimators. Therefore, it can be regarded as a quantitative benchmark for efficiency considerations. In particular, for the model $P(\sigma, \alpha)$ with σ known, the MLE of α is readily derived in Arnold (1983), and given by

$$\hat{\alpha}_{\mathrm{ML}} = \frac{n}{\sum_{i=1}^{n} \log\left(X_i / \sigma\right)}.$$

Its exact distribution theory is described by the statement that

$$\frac{2n\alpha}{\hat{\alpha}_{\mathrm{ML}}} \text{ has cdf } \chi^2_{2n},$$

where χ^2_ν denotes the chi-square distribution with ν degrees of freedom. This implies that $\hat{\alpha}_{\mathrm{ML}}$ is a biased estimator of α, but multiplication by the factor $(n-1)/n$ yields an unbiased version,

$$\hat{\alpha}_{\mathrm{MLU}} = \frac{n-1}{n}\,\hat{\alpha}_{\mathrm{ML}} = \frac{n-1}{\sum_{i=1}^{n}\log(X_i/\sigma)}.$$

For further details on exact distribution theory of $\hat{\alpha}_{\mathrm{ML}}$ see Rytgaard (1990).

Following techniques in Brazauskas and Serfling (2000a,b), it can be shown that for large sample size n, $\hat{\alpha}_{\mathrm{MLU}}$ is approximately normal with mean α and variance α^2/n. Moreover, other competing estimators $\hat{\alpha}$ for α considered here are approximately normal with mean α and variance $c\alpha^2/n$ for some constant $c > 1$ and large n. This means that confidence intervals for the parameter α based on the competing estimators will be $\sqrt{c}$ times wider than those based on the MLU. Such optimal precision of the MLU, however, is achieved at the price of robustness, which becomes crucial when the actual data departs from the assumed parametric model. Hence, the MLU is most *efficient* but is *nonrobust* with UBP = 0.

2. The Estimators

The MLE and MLU were given in Section 1.3. Here we introduce the other methods considered in this study for estimation of the parameter α. In particular, we present quantile, trimmed mean, and generalized median type estimators. For further details and discussion see Brazauskas and Serfling (2000a,b).

2.1. Quantile Type Estimators

Quantile type estimators of α are completely unaffected by additional information about σ. For this reason and for compatibility with the existing literature, we describe this approach here for the case when σ is treated as an unknown nuisance parameter.

Quantile estimators based on $k \geq 2$ (selected) quantile levels $0 < p_1 < \ldots < p_k < 1$ are defined as follows:

$$\hat{\alpha}_{\mathrm{Q}} = \left(\sum_{i=1}^{k} b_i \log X_{(\lceil np_i \rceil)}\right)^{-1},$$

$$\hat{\sigma}_{\mathrm{Q}} = \exp\left\{\log X_{(\lceil np_1 \rceil)} - u_1/\hat{\alpha}_{\mathrm{Q}}\right\},$$

with

$$b_1 = -\frac{1}{L}\frac{u_2 - u_1}{e^{u_2} - e^{u_1}},$$

$$b_i = \frac{1}{L}\left[\frac{u_i - u_{i-1}}{e^{u_i} - e^{u_{i-1}}} - \frac{u_{i+1} - u_i}{e^{u_{i+1}} - e^{u_i}}\right], \quad 2 \leq i \leq k-1,$$

$$b_k = \frac{1}{L}\frac{u_k - u_{k-1}}{e^{u_k} - e^{u_{k-1}}},$$

and

$$L = \sum_{i=2}^{k} \frac{(u_i - u_{i-1})^2}{e^{u_i} - e^{u_{i-1}}},$$

where $u_i = -\log(1-p_i)$, $1 \le i \le k$, and $\lceil x \rceil$ denotes the least integer $\ge x$. Such estimators were introduced and studied for the Pareto problem by Quandt (1966) for $k = 2$ and by Koutrouvelis (1981) for general $k \ge 2$.

Choosing the minimum of the determinant of the asymptotic covariance matrix of the estimators of σ and α as an optimality criterion, Koutrouvelis (1981) found that the optimal choice of p_1 is always

$$p_1^{\circ} = \frac{1}{n + 0.5},$$

and the remaining optimal quantile levels $p_2^{\circ}, \ldots, p_k^{\circ}$ are:

- For $k = 2$, take $p_2^{\circ} = 1 - (1 - p_1^{\circ})e^{-1.5936} \approx .80$.
- For $k = 5$, take $p_2^{\circ} = 1 - (1 - p_1^{\circ})e^{-0.6003} \approx .45$, $p_3^{\circ} = 1 - (1 - p_1^{\circ})e^{-1.3544} \approx .74$, $p_4^{\circ} = 1 - (1 - p_1^{\circ})e^{-2.3721} \approx .91$, and $p_5^{\circ} = 1 - (1 - p_1^{\circ})e^{-3.9657} \approx .98$.

We denote the optimal estimators of α by $\hat{\alpha}_{\mathrm{Q}}^{opt,k}$. We also consider a nonoptimal case (denoted by $\hat{\alpha}_{\mathrm{Q}}^{*}$):

- For $k = 5$, take $p_1 = .13$, $p_2 = .315$, $p_3 = .50$, $p_4 = .685$, and $p_5 = .87$.

Remark. When the number k of quantiles is chosen to equal the number of unknown parameters of the model, this method corresponds to what is called *percentile matching* by Klugman, Panjer, and Willmot (1998).

2.2. Trimmed Mean Estimators

For specified β_1 and β_2 satisfying $0 \le \beta_1, \beta_2 < 1/2$, a trimmed mean is formed by discarding the proportion β_1 lowermost observations and the proportion β_2 uppermost observations and averaging the remaining ones in some sense. In particular, for α we introduce the trimmed mean estimator

$$\hat{\alpha}_{\mathrm{T}} = \left(\sum_{i=1}^{n} c_{ni} \log\left(X_{(i)} / \sigma \right) \right)^{-1},$$

with $c_{ni} = 0$ for $1 \le i \le [n\beta_1]$, $c_{ni} = 0$ for $n - [n\beta_2] + 1 \le i \le n$, and $c_{ni} = 1/d(\beta_1, \beta_2, n)$ for $[n\beta_1] + 1 \le i \le n - [n\beta_2]$, where $[\cdot]$ denotes "greatest integer part", and

$$d(\beta_1, \beta_2, n) = \sum_{j=[n\beta_1]+1}^{n-[n\beta_2]} \sum_{i=0}^{j-1} (n-i)^{-1}.$$

These estimators correspond to the trimmed mean estimators introduced and studied by Kimber (1983a,b) for the equivalent problem of estimation of $\theta = \alpha^{-1}$ in the model $E(\mu, \theta)$ with μ known. The above c_{ni}'s are a choice making $\hat{\theta}_T = \hat{\alpha}_T^{-1}$ mean-unbiased for $\theta = \alpha^{-1}$.

2.3. Generalized Median Estimators

Generalized median (GM) statistics are defined by taking the median of the $\binom{n}{k}$ evaluations of a given kernel $h(x_1, \ldots, x_k)$ over all k-sets of the data. See Serfling (1984, 2000) for general discussion. In Brazauskas and Serfling (2000a), such estimators were considered for the parameter α in the case of σ known:

$$\hat{\alpha}_{\mathrm{GM}} = \mathrm{Median}\left\{h(X_{i_1}, \ldots, X_{i_k})\right\},$$

with a particular kernel $h(x_1, \ldots, x_k)$:

$$h(x_1, \ldots, x_k; \sigma) = \frac{1}{C_k} \frac{k}{\sum_{j=1}^{k} \log(x_j / \sigma)},$$

where C_k is a multiplicative median-unbiasing factor, i.e., chosen so that the distribution of $h(X_{i_1}, \ldots, X_{i_k}; \sigma)$ has *median* α. Values of C_k, for $k = 2{:}10$, are provided in the following table. (For $k > 10$, C_k is given by a very accurate approximation, $C_k \approx k/(k - 1/3)$.

k	2	3	4	5	6	7	8	9	10
C_k	1.1916	1.1219	1.0893	1.0705	1.0582	1.0495	1.0431	1.0382	1.0343

3. Data Sets and Preliminary Diagnostics

We choose three data sets for analysis in this study: Wind Catastrophes (1977), OLT Bodily Injury Liability Claims (1976), and Norwegian Fire Claims (1975). These data sets are of interest because they have been analyzed extensively in the actuarial literature. In this section we first present the data and briefly mention methods of analysis proposed in the literature. Then we describe a data de-grouping technique which we apply for the wind data, the liability data, and the Norwegian data. Finally, for an initial assessment of the validity of distributional assumptions, we provide histograms and *quantile-quantile* plots (QQ-plots) for each data set.

3.1. Wind Catastrophes (1977)

The Wind Catastrophes (1977) data set is taken from Hogg and Klugman (1984), p. 64. It represents 40 losses that occurred in 1977 due to wind-related catastrophes. The data were recorded to the nearest \$1,000,000 and include only those losses of $2,000,000 or more. The following display provides the losses (in millions of dollars):

2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	4	4	4	5
5	5	5	6	6	6	6	8	8	9	15	17	22	23	24	24	25	27	32	43

In Hogg and Klugman (1984) two parametric models were used to fit the wind data: truncated exponential (with the truncation point 1.5) and two-parameter Pareto. Derrig, Ostaszewski, and Rempala (2000) also studied this data set and, in addition to the above parametric models, used the empirical nonparametric and the bootstrap approaches to estimate the probability that a wind loss will exceed 29.5 (that is, $29,500,000). Further, Philbrick (1985), among several applications of the single-parameter Pareto distribution to real data, investigated a $P(\sigma, \alpha)$ fit to the wind data with the truncation point $\sigma = 2$. He advocated the use of the MLE for estimation of α but apparently was unaware that this estimator is biased.

3.2. OLT Bodily Injury Liability Claims (1976)

The complete OLT Bodily Injury Liability (1976) data set is available in Patrik (1980), p. 99. It is prepared by the Insurance Services Office and represents Owners, Landlords and Tenants (OLT) bodily injury liability losses for the policy limit $500,000 for policy year 1976 evaluated as of March 31, 1978. Patrik (1980) described general principles of selection, estimation, and testing of loss models for casualty insurance claims. For illustrative purposes he used the two-parameter Pareto distribution to fit the entire range of claims, including the OLT Bodily Injury Liability (1976) claims.

Here we follow Philbrick's (1985) approach and fit the single-parameter Pareto distribution only to the claims that are greater than $25,000. The grouped losses (recorded in thousands of dollars) in exceedance of this threshold are presented in Appendix, Table A.1.

3.3. Norwegian Fire Claims (1975)

This data set is one among 20 sets of Norwegian Fire Claims, for years 1972-1984 and 1986-1992, presented in Appendix I of Beirlant, Teugels, and Vynckier (1996). It represents the total damage done by 142 fires in Norway for the year 1975. (For this year a single-parameter Pareto distribution seems to provide a reasonably good fit to these data. Fits of similar quality are observed for several other years as well.) A priority of 500,000 Norwegian krones was in force, thus no claims below this limit were recorded. Actual losses (in thousands of Norwegian krones) are provided in Appendix, Table A.2.

The Norwegian fire claims for various years have been extensively analyzed by Beirlant, Teugels, and Vynckier (1996). Their approach is based on extreme value theory, which concentrates exclusively on the upper tail of the data. We will not pursue that approach here. However, we will follow techniques developed there for data diagnostics, namely, methods of Section 1.5 which describe how to construct specific QQ-plots.

3.4. A Method for Data De-grouping

We start with a motivational example based on the wind data. Losses recorded there are rounded to the nearest million which suggests that *actual* losses corresponding to 2, for example, were not exactly 2 but rather somewhere between 1.5 and 2.5. (This seems to be one of the reasons why Hogg and Klugman (1984) considered the left-hand endpoint 1.5 for the truncated exponential model.) To avoid ties and inappropriate clustering of claims due to such rounding, we apply a simple data de-grouping method as follows.

Let us continue with the wind data and, in particular, losses of size 2. It is reasonable to assume that actual observations that correspond to 2 are *equally spaced* (or, equivalently, *uniformly distributed*) on the interval (1.5, 2.5). Thus, for the wind data, instead of the 12 observations "2" we shall use 1.58, 1.65, 1.73, 1.81, 1.88, 1.96, 2.04, 2.12, 2.19, 2.27, 2.35, 2.42, as the actual data. More formally,

> if (A, B) is an interval of losses and m is the number of losses within (A, B), then m uniformly distributed losses $x_1, \ldots, x_m$ in that interval are found according to the formula:
>
> $$x_k = \left(1 - \frac{k}{m+1}\right) A + \frac{k}{m+1} B, \quad k = 1, \ldots, m.$$

We emphasize that such an approach neither distorts the original grouping nor changes the total loss amount within a group. It is easy to implement in practice and, most importantly, it makes the data continuous, thus allowing methods of estimation and goodness-of-fit to be applied directly. We apply this technique to all three data sets. Finally, one may also consider more sophisticated data de-grouping schemes by employing, for example, the beta family of distributions instead of uniform. In that case, however, additional information, such as mean and variance of losses within the interval, is required.

3.5. Preliminary Diagnostics

In Figure 1, we illustrate the results of preliminary diagnostics for the three data sets at hand. Three plots in the first column correspond to the Wind Catastrophes (1977), in the second column — OLT Bodily Injury Liability Claims (1976), and in the third column — Norwegian Fire Claims (1975) data. The following conclusions are quite evident:

- Histograms for all the data sets exhibit a similar shape for the underlying distribution. Two *one-parameter* models seem to be appropriate candidates: the truncated exponential, and the Pareto.

- The exponential QQ-plots clearly reveal that a truncated exponential model does not fit any of the three data sets (a good fit corresponds to a 45° line).
- In comparison with the exponential case, the Pareto QQ-plot shows *mild improvement* for the wind data, *significant improvement* for the liability data, and *nearly perfect fit* for the Norwegian fire data.

Remark. Plotting of lognormal or Weibull QQ-plots, for example, reveals significant improvements over the one-parameter Pareto model for the wind and the liability data. These distributions, however, are *two-parameter* models (thus, less parsimonious) and are not considered as competitors to the one-parameter Pareto model. □

4. Comparisons and Conclusions

In Section 4.1 we present summarizing tables for each data set, showing values of the estimates of α, values of the KS, CvM, and AD statistics, and ranks of the estimators based on these goodness-of-fit measures. In Section 4.2, conclusions are drawn and recommendations are provided. Performances of all estimators under consideration are compared by simultaneously examining their ranks (for all three data sets), their UBP's, and their variances.

Remarks. (i) Ranks to estimators are assigned as follows. The estimator with the lowest value for a selected goodness-of-fit measure receives rank 1, the estimator with the second lowest value (for the same measure) — receives rank 2, etc. The idea of ranking estimators or models based on a certain criterion is not new. It has been suggested and quite extensively dicussed by Klugman, Panjer, and Willmot (1998), Section 2.9.2.

(ii) For situations when the number $\binom{n}{k}$ of kernel evaluations needed for computation of $\hat{\alpha}_{\mathrm{GM}}$ becomes extremely large, we reduce the computational burden by randomly choosing 10^7 kernel evaluations if $\binom{n}{k}$ exceeds 10^7. Such an approach maintains a high degree of numerical accuracy (up to 3 decimal places) and renders the computational burden negligible. For instance, for the Norwegian data ($n = 142$), it requires only 150 seconds to compute (simultaneously) all estimators used in this study on a Pentium II 400MHz laptop computer. Further discussion on computational aspects of the GM estimators is available in Brazauskas and Serfling (2000a). □

4.1. Summary of Pareto Fits

Discussion of Table 4.1

The fitted model is $\mathrm{P}(\sigma = 1.5, \hat{\alpha})$ with values of $\hat{\alpha}$ ranging from 0.605 (for $\hat{\alpha}_{\mathrm{Q}}^{\mathrm{opt},2}$) to 0.791 (for $\hat{\alpha}_{\mathrm{Q}}^{\mathrm{opt},5}$). This range differs somewhat from the findings of Philbrick (1985), where the MLE value of 0.976 (for grouped data) is reported and compared to a typical parameter value of 1.0 for the property insurance line.

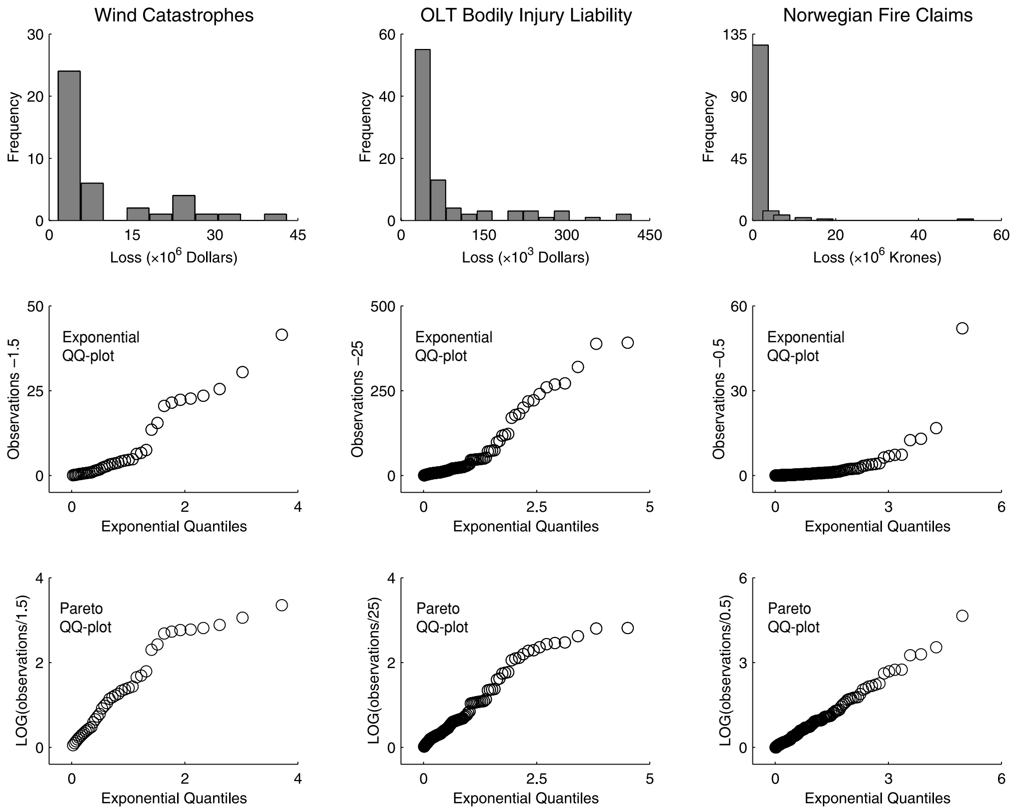


Figure 1: Preliminary diagnostics for all three data sets.

The main reason for such a difference is the choice of the truncation point. In Philbrick (1985), $\sigma = 2$ is used. However, in view of our discussion in Section 3.4 and Example 1 in Hogg and Klugman (1984), p. 64, we believe that the choice of 1.5 is more natural. Moreover, all three goodness-of-fit tests very strongly support the appropriateness of the $P(\sigma = 1.5,\ \hat{\alpha}_{ML} = 0.764)$ model with the goodness-of-fit values .1071 (KS), .1106 (CvM), .7329 (AD), and corresponding *p-values*: .51 (KS), .27 (CvM), .24 (AD). While the corresponding *p-values* for the $P(\sigma = 2.0,\ \hat{\alpha}_{ML} = 0.945)$ model (for the de-grouped data) are comparable for the CvM and AD statistics, the *p-value* for the KS statistic is substantially lower: .33 (KS), .30 (CvM), .23 (AD). Thus, based on this discussion, we choose the model $P(\sigma = 1.5,\ \hat{\alpha})$.

Table 4.1 suggests that, although the $P(\sigma = 1.5,\ \hat{\alpha}_{ML})$ model is accepted by all three tests, additional improvements of the fit are possible if we use the unbiased version MLU, which in turn can be even further improved by robust estimators. For example, the estimators $\hat{\alpha}_T$ (with $\beta_1 = 0$, $\beta_2 = .05$), $\hat{\alpha}_{GM}$ (with $k = 4$, $k = 5$, and $k = 10$), and $\hat{\alpha}^*_Q$, all have *uniformly* smaller ranks than $\hat{\alpha}_{MLU}$. □

TABLE 4.1

VALUES OF $\hat{\alpha}$, GOODNESS-OF-FIT STATISTICS, AND RANKS FOR THE WIND DATA.

Estimator	$\hat{\alpha}$	**KS**	**rank**	**CvM**	**rank**	**AD**	**rank**
MLU	.745	.0980	6	.0911	12	.6484	12
$Q^{opt,2}$	.605	.1320	14	.0956	13	.7939	13
Q^*, $k = 5$	.731	.0911	$2\frac{1}{2}$	.0792	10	.5999	10
$Q^{opt,5}$	.791	.1198	13	.1445	14	.8881	14
T, $\beta_1 = 0$, $\beta_2 = .05$	.707	.0932	4	.0642	7	.5457	5
T, $\beta_1 = 0$, $\beta_2 = .10$	.677	.1031	8	.0562	2	.5335	2
T, $\beta_1 = 0$, $\beta_2 = .15$	.664	.1077	11	.0568	4	.5487	6
T, $\beta_1 = 0$, $\beta_2 = .20$	.667	.1066	10	.0564	3	.5441	4
T, $\beta_1 = 0$, $\beta_2 = .25$	.673	.1045	9	.0561	1	.5368	3
GM, $k = 2$	.653	.1118	12	.0594	6	.5720	8
GM, $k = 3$	.692	.0981	7	.0587	5	.5316	1
GM, $k = 4$	.714	.0911	$2\frac{1}{2}$	.0679	8	.5576	7
GM, $k = 5$	.723	.0884	1	.0734	9	.5777	9
GM, $k = 10$	.744	.0975	5	.0901	11	.6445	11

DISCUSSION OF TABLE 4.2

The fitted model is P($\sigma = 25{,}000$, $\hat{\alpha}$) with values of $\hat{\alpha}$ ranging from 1.082 (for $\hat{\alpha}_{GM}$ with $k = 3$) to 1.172 (for $\hat{\alpha}_Q^{opt,2}$). Philbrick (1985) reports the MLE value of 1.108. This is slightly below the industry values of 1.245 (all classes liability; truncation point 25,000) and 1.159 (high severity liability; truncation point 30,000), which are available in Reichle and Yonkunas (1985), Appendix E.

TABLE 4.2

VALUES OF $\hat{\alpha}$, GOODNESS-OF-FIT STATISTICS, AND RANKS FOR THE LIABILITY DATA.

Estimator	$\hat{\alpha}$	**KS**	**rank**	**CvM**	**rank**	**AD**	**rank**
MLU	1.140	.0735	12	.0794	11	.6795	12
$Q^{opt,2}$	1.172	.0784	14	.0944	14	.7843	14
Q^*, $k = 5$	1.111	.0690	6	.0748	2	.6343	5
$Q^{opt,5}$	1.161	.0767	13	.0881	13	.7420	13
T, $\beta_1 = 0$, $\beta_2 = .05$	1.098	.0670	4	.0757	4	.6302	1
T, $\beta_1 = 0$, $\beta_2 = .10$	1.093	.0662	2	.0766	8	.6314	3
T, $\beta_1 = 0$, $\beta_2 = .15$	1.110	.0689	5	.0748	2	.6336	4
T, $\beta_1 = 0$, $\beta_2 = .20$	1.125	.0712	8	.0759	5	.6500	8
T, $\beta_1 = 0$, $\beta_2 = .25$	1.127	.0715	9	.0762	6	.6532	9
GM, $k = 2$	1.133	.0724	$10\frac{1}{2}$	.0775	$9\frac{1}{2}$	.6641	$10\frac{1}{2}$
GM, $k = 3$	1.082	.0656	1	.0795	12	.6395	7
GM, $k = 4$	1.094	.0664	3	.0764	7	.6310	2
GM, $k = 5$	1.113	.0693	7	.0748	2	.6359	6
GM, $k = 10$	1.133	.0724	$10\frac{1}{2}$	.0775	$9\frac{1}{2}$	.6641	$10\frac{1}{2}$

TABLE 4.3

VALUES OF $\hat{\alpha}$, GOODNESS-OF-FIT STATISTICS, AND RANKS FOR THE NORWEGIAN DATA.

Estimator	$\hat{\alpha}$	KS	rank	CvM	rank	AD	rank
MLU	1.209	.0517	13	.0353	$11\frac{1}{2}$	.3693	8
$Q^{opt,2}$	1.234	.0470	$3\frac{1}{2}$	.0351	$9\frac{1}{2}$	.3717	$10\frac{1}{2}$
Q^*, $k = 5$	1.232	.0473	5	.0348	8	.3698	9
$Q^{opt,5}$	1.203	.0529	14	.0367	13	.3759	13
T, $\beta_1 = 0$, $\beta_2 = .05$	1.221	.0494	8	.0341	1	.3645	$1\frac{1}{2}$
T, $\beta_1 = 0$, $\beta_2 = .10$	1.229	.0479	6	.0345	$5\frac{1}{2}$	.3674	7
T, $\beta_1 = 0$, $\beta_2 = .15$	1.234	.0470	$3\frac{1}{2}$	.0351	$9\frac{1}{2}$	.3717	$10\frac{1}{2}$
T, $\beta_1 = 0$, $\beta_2 = .20$	1.235	.0468	2	.0353	$11\frac{1}{2}$	.3728	12
T, $\beta_1 = 0$, $\beta_2 = .25$	1.226	.0485	7	.0343	$3\frac{1}{2}$	.3658	5
GM, $k = 2$	1.242	.0454	1	.0369	14	.3825	14
GM, $k = 3$	1.220	.0496	9	.0342	2	.3645	$1\frac{1}{2}$
GM, $k = 4$	1.217	.0502	10	.0343	$3\frac{1}{2}$	.3649	3
GM, $k = 5$	1.215	.0506	11	.0345	$5\frac{1}{2}$	.3655	4
GM, $k = 10$	1.214	.0508	12	.0346	7	.3659	6

In our case, the range of $\hat{\alpha}$ is in closer agreement with the industry values. A minor discrepancy between Philbrick and our MLE's is due to different data de-grouping approaches.

The goodness-of-fit tests show strong evidence that the P($\sigma = 25{,}000$, $\hat{\alpha}_{ML} = 1.153$) model is appropriate with the goodness-of-fit values (*p-values*): .0755 (.35) for KS, .0843 (.42) for CvM, .7153 (.26) for AD. Nonetheless, the fit based on the MLE is again uniformly improved by MLU, which is further improved by all T and GM estimators and by $\hat{\alpha}_Q^*$. □

DISCUSSION OF TABLE 4.3

The fitted model is P($\sigma = 500{,}000$, $\hat{\alpha}$) with values of $\hat{\alpha}$ ranging from 1.203 (for $\hat{\alpha}_Q^{opt,5}$) to 1.242 (for $\hat{\alpha}_{GM}$ with $k = 2$). The narrowness of the range points to a very good fit between the data and Pareto model, which was initially suggested above by the QQ-plot. Further, the three tests show extremely strong evidence in support of the P($\sigma = 500{,}000$, $\hat{\alpha}_{ML} = 1.218$) model with the goodness-of-fit values (*p-values*): .0500 (.70) for KS, .0343 (.89) for CvM, .3647 (.71) for AD. Therefore, it is not surprising that in this case the MLE fit is among the best, improved upon by only the $\hat{\alpha}_{GM}$ (with $k = 3$) and $\hat{\alpha}_T$ (with $\beta_1 = 0$, $\beta_2 = .05$) fits. □

4.2. Comparisons, Conclusions, and Recommendations

DISCUSSION OF TABLE 4.4

"Robustness versus efficiency" comparisons show that GM-type estimators dominate the competition. In particular, for a fixed variance (or UBP), any Q or T-type estimator can be improved upon by a GM estimator with as good

TABLE 4.4

COMPARISONS BASED ON RANKS, UBP, AND VARIANCE OF ESTIMATORS.

Estimator	Wind Catastrophes KS,	CvM,	AD	OLT Liability KS,	CvM,	AD	Norwegian Claims KS,	CvM,	AD	UBP	Variance ($\times\alpha^2/n$)
MLU	**6**	**12**	**12**	**12**	**11**	**12**	**13**	**$11\frac{1}{2}$**	**8**	**0**	**1**
$Q^{opt,2}$	14	13	13	14	14	14	$3\frac{1}{2}$	$9\frac{1}{2}$	$10\frac{1}{2}$	.203	1.541
Q^*, $k=5$	$2\frac{1}{2}$	10	10	6	2	5	5	8	9	.130	1.383
$Q^{opt,5}$	13	14	14	13	13	13	14	13	13	.019	1.079
T, $\beta_1=0$, $\beta_2=.05$	4	7	5	4	4	1	8	1	$1\frac{1}{2}$	.050	1.089
T, $\beta_1=0$, $\beta_2=.10$	8	2	2	2	8	3	6	$5\frac{1}{2}$	7	.100	1.180
T, $\beta_1=0$, $\beta_2=.15$	11	4	6	5	2	4	$3\frac{1}{2}$	$9\frac{1}{2}$	$10\frac{1}{2}$	.150	1.277
T, $\beta_1=0$, $\beta_2=.20$	10	3	4	8	5	8	2	$11\frac{1}{2}$	12	.200	1.383
T, $\beta_1=0$, $\beta_2=.25$	9	1	3	9	6	9	7	$3\frac{1}{2}$	5	.250	1.501
GM, $k=2$	12	6	8	$10\frac{1}{2}$	$9\frac{1}{2}$	$10\frac{1}{2}$	1	14	14	.293	1.280
GM, $k=3$	7	5	1	1	12	7	9	2	$1\frac{1}{2}$	.206	1.141
GM, $k=4$	$2\frac{1}{2}$	8	7	3	7	2	10	$3\frac{1}{2}$	3	.159	1.088
GM, $k=5$	1	9	9	7	2	6	11	$5\frac{1}{2}$	4	.129	1.061
GM, $k=10$	5	11	11	$10\frac{1}{2}$	$9\frac{1}{2}$	$10\frac{1}{2}$	12	7	6	.067	1.019

variance (UBP) and larger UBP (smaller variance). For example, $\hat{\alpha}_Q^{opt,5}$ with UBP = .019 and variance = 1.079 is improved upon by $\hat{\alpha}_{GM}(k=5)$, with UBP = .129 and variance = 1.061 and $\hat{\alpha}_{GM}(k=10)$, with UBP = .067 and variance = 1.019. Similarly $\hat{\alpha}_T$ ($\beta_1=0$, $\beta_2=.15$) with UBP = .150 and variance = 1.277 is improved upon by $\hat{\alpha}_{GM}$ ($k=3$), with UBP = .206 and variance = 1.141, and $\hat{\alpha}_{GM}$ ($k=4$), with UBP = .159 and variance = 1.088.

For the goodness-of-fit comparisons, if an estimator has at least 2 out of 3 ranks of corresponding statistics smaller than another estimator, then its performance is considered better. For the wind data, for example, $\hat{\alpha}_T$ ($\beta_1=0$, $\beta_2=.10$) with ranks (8, 2, 2) is better than $\hat{\alpha}_T$ ($\beta_1=0$, $\beta_2=.05$) with ranks (4, 7, 5) but worse than $\hat{\alpha}_{GM}$ ($k=3$) with ranks (7, 5, 1). This approach suggests that, for all data sets, $\hat{\alpha}_Q^*$ demonstrates the strongest performance among Q-type estimators but is outperformed by the best T-type ($\hat{\alpha}_T$ with $\beta_1=0$, $\beta_2=.05$ and $\beta_1=0$, $\beta_2=.10$) and by the best GM-type ($\hat{\alpha}_{GM}$ with $k=3$ and $k=4$) estimators. □

Conclusions

Based on the comparisons in Table 4.4, the following conclusions emerge:

- The GM-type estimators offer the best trade-offs between robustness and efficiency, which translates into an excellent performance in terms of goodness-of-fit. The best fits are provided by the $\hat{\alpha}_{GM}$ ($k=3$ and $k=4$) estimators, which offer moderate to high protection against contamination (UBP) and low to moderate sacrifice in accuracy (variance).

- The T-type estimators are slightly less competitive in terms of "robustness versus efficiency" comparisons. However, their goodness-of-fit performance is as good as that of the GM-type estimators. The best fits are provided by the $\hat{\alpha}_{\mathrm{T}}$ ($\beta_1 = 0$, $\beta_2 = .05$ and $\beta_1 = 0$, $\beta_2 = .10$) estimators, which offer low to moderate protection against contamination (UBP) and low to moderate sacrifice in accuracy (variance).
- The Q-type estimators are outperformed with respect to both criteria, "robustness versus efficiency" and goodness-of-fit, by the T and GM-type estimators and, thus, are less competitive.
- The nonrobust but most efficient MLU neither can improve nor be improved by any other estimator with respect to the "robustness versus efficiency" criterion, because it has the best variance and the worst UBP. However, its performance with respect to goodness-of-fit is *consistently* among the worst, implying that for "robustness versus efficiency" comparisons the robustness should be given a higher priority.

Practical Recommendations

When fitting Pareto models to loss data, the following steps are necessary:

1. If data are grouped or ties are present, de-group it using methods of Section 3.4; otherwise, go to the next step.
2. Use diagnostic tools — histogram and QQ-plots — to visually determine whether a Pareto model is appropriate.
3. Compute $\hat{\alpha}_{\mathrm{ML}}$ and apply the KS, CvM, and AD statistics to formally test if the Pareto model provides an adequate fit to the data. Note that tables with the critical values of these statistics are only available when α is estimated by MLE. (Tables are presented in D'Agostino and Stephens (1986), pp. 135-136, and Durbin (1975), Table 3.)
4. Compute $\hat{\alpha}$ using the MLU and Q, T, and GM-type estimators. If the range of $\hat{\alpha}$'s is narrow (as in the case of the Norwegian fire data), then the fit is very good and even the MLE can be relied on. However, if the range is relatively wide (as in the case of the wind data), then ranking of robust estimators has to be used to refine the fit.
5. In situations when all three goodness-of-fit tests support the Pareto model, the T-type estimators with 5%-10% trimming and the GM-type estimators with $k = 3$ and $k = 4$ perform the best. Estimators with high UBP should be applied if one of the tests rejects the Pareto model or if the QQ-plots are not satisfactory, or if the range of $\hat{\alpha}$'s is very wide.

5. Acknowledgments

The authors are very appreciative of constructive remarks provided by an anonymous referee. These have led to many improvements in the paper. Also, support of the first author by a grant from the Actuarial Education and Research Fund and support of the second author by grants from the Casualty Actuarial Society and Society of Actuaries, with administrative support from the Actuarial Education and Research Fund, and by NSF Grant DMS-0103698, are gratefully acknowledged.

References

ARNOLD, B.C. (1983). *Pareto Distributions*. International Cooperative Publishing House. Fairland, Maryland.

BEIRLANT, J., TEUGELS, J.L. and VYNCKIER, P. (1996) *Practical Analysis of Extreme Values*. Leuven University Press, Leuven, Belgium.

BRAZAUSKAS, V. and SERFLING, R. (2000a) Robust and efficient estimation of the tail index of a single-parameter Pareto distribution. *North American Actuarial Journal* **4(4)**, 12-27.

BRAZAUSKAS, V. and SERFLING, R. (2000b) Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics. *Extremes* **3(3)**, 231-249.

BRAZAUSKAS, V. and SERFLING, R. (2001) Small sample performance of robust estimators of tail parameters for Pareto and exponential models. *Journal of Statistical Computation and Simulation* **70(1)**, 1-19.

D'AGOSTINO, R.B. and STEPHENS, M.A. (1986) *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

DERRIG, R.A., OSTASZEWSKI, K.M. and REMPALA, G.A. (2000) Applications of resampling methods in actuarial practice. *Proceedings of the Casualty Actuarial Society* **LXXXVII**, 322-364.

DURBIN, J. (1975) Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika* **62**, 5-22.

HOGG, R.V. and KLUGMAN, S.A. (1984). *Loss Distributions*. Wiley, New York.

KIMBER, A.C. (1983a) Trimming in gamma samples. *Applied Statistics* **32**, 7-14.

KIMBER, A.C. (1983b) Comparison of some robust estimators of scale in gamma samples with known shape. *Journal of Statistical Computation and Simulation* **18**, 273-286.

KLUGMAN, S.A., PANJER, H.H. and WILLMOT, G.E. (1998) *Loss Models: From Data to Decisions.* Wiley, New York.

KOUTROUVELIS, I.A. (1981) Large-sample quantile estimation in Pareto laws. *Communications in Statistics, Part A – Theory and Methods* **10**, 189-201.

PATRIK, G. (1980) Estimating casualty insurance loss amount distributions. *Proceedings of the Casualty Actuarial Society* **LXVII**, 57-109.

PHILBRICK, S.W. (1985) A practical guide to the single parameter Pareto distribution. *Proceedings of the Casualty Actuarial Society* **LXXII**, 44-84.

PHILBRICK, S.W. and JURSCHAK, J. (1981) Discussion of "Estimating casualty insurance loss amount distributions." *Proceedings of the Casualty Actuarial Society* **LXVIII**, 101-106.

QUANDT, R.E. (1966) Old and new methods of estimation and the Pareto distribution. *Metrika* **10**, 55-82.

REICHLE, K.A. and YONKUNAS, J.P. (1985) Discussion of "A practical guide to the single parameter Pareto distribution." *Proceedings of the Casualty Actuarial Society* **LXXII**, 85-123.

RYTGAARD, M. (1990) Estimation in the Pareto distribution. *ASTIN Bulletin* **20(2)**, 201-216.

SERFLING, R. (1984) Generalized *L*-, *M*- and *R*-statistics. *Annals of Statistics* **12**, 76-86.

SERFLING, R. (2000) "Robust and nonparametric estimation via generalized L-statistics: theory, applications, and perspectives," In: *Advances in Methodological and Applied Aspects of Probability and Statistics*, Balakrishnan, N. (Ed.), pp. 197-217. Gordon & Breach.

Appendix

TABLE A.1

OLT Bodily Injury Liability Claims (1976) Data (× 1000 dollars).

Loss Amount	Number of Losses	Loss Amount	Number of Losses	Loss Amount	Number of Losses
25-30	11	70-75	9	220-230	1
30-35	18	75-80	1	240-250	2
35-40	9	95-100	4	260-270	1
40-45	4	120-130	2	280-290	1
45-50	11	140-150	3	290-300	2
50-55	3	190-200	1	340-350	1
55-60	2	200-210	2	410-420	2

Source: Patrik (1980), Appendix F, Part 1.

TABLE A.2.

Norwegian Fire Claims (1975) Data (× 1000 Norwegian krones).

500	552	600	650	798	948	1180	1479	2497	7371
500	557	605	672	800	957	1243	1485	2690	7772
500	558	610	674	800	1000	1248	1491	2760	7834
502	570	610	680	800	1002	1252	1515	2794	13000
515	572	613	700	826	1009	1280	1519	2886	13484
515	574	615	725	835	1013	1285	1587	2924	17237
528	579	620	728	862	1020	1291	1700	2953	52600
530	583	622	736	885	1024	1293	1708	3289	
530	584	632	737	900	1033	1298	1820	3860	
530	586	635	740	900	1038	1300	1822	4016	
540	593	635	748	910	1041	1305	1848	4300	
544	596	640	752	912	1104	1327	1906	4397	
550	596	650	756	927	1108	1387	2110	4585	
550	600	650	756	940	1137	1455	2251	4810	
551	600	650	777	940	1143	1475	2362	6855	

Source: Beirlant, Teugels, and Vynckier (1996), Appendix I.

Vytaras Brazauskas
Department of Mathematical Sciences
University of Wisconsin-Milwaukee
P.O. Box 413, Milwaukee, Wisconsin 53201, USA.
E-mail: vytaras@uwm.edu

Robert Serfling
Department of Mathematical Sciences,
University of Texas at Dallas, Richardson,
Texas 75083-0688, USA.
E-mail: serfling@utdallas.edu

THE VALUATION AND HEDGING OF VARIABLE RATE SAVINGS ACCOUNTS

BY

FRANK DE JONG[1] AND JACCO WIELHOUWER[2]

ABSTRACT

Variable rate savings accounts have two main features. The interest rate paid on the account is variable and deposits can be invested and withdrawn at any time. However, customer behaviour is not fully rational and withdrawals of balances are often performed with a delay. This paper focuses on measuring the interest rate risk of variable rate savings accounts on a value basis (duration) and analyzes the problem how to hedge these accounts. In order to model the embedded options and the customer behaviour we implement a partial adjustment specification. The interest rate policy of the bank is described in an error-correction model.

KEYWORDS

Term structure, duration, uncertain cash flow, variable rates of return JEL codes: C33, E43

1. INTRODUCTION

A major part of private savings is deposited in variable rate saving accounts, in the US also known as demand deposits. Typically, deposits can be invested and withdrawn at any time at no cost, which makes a savings account look similar to a money market account. However, the interest rate paid on savings accounts is often different from the money market rate. In Europe, the interest rate paid on the savings account can actually be higher or lower than the money market rate. Even when these interest rates differ, depositors do not immediately withdraw their money from savings accounts when rates on

* We thank Dennis Bams, Joost Driessen, D. Wilkie, participants at the AFIR 2000 colloquium, and two anonymous referees for comments on previous versions of the paper. The usual disclaimer applies.

[1] University of Amsterdam

[2] ING Group and CentER, Tilburg University

alternative investments are higher. Whatever the causes of this behaviour (market imperfections, transaction costs or other), these characteristics imply that the value of the savings accounts from the point of view of the issuing bank may be different from the nominal value of the deposits.

In the literature, the valuation of savings accounts is well studied. For example, Hutchison and Pennacchi (1996), Jarrow and Van Deventer (1998) and Selvaggio (1996) provide models for the valuation of such products. The first two papers build on the (extended) Vasicek (1977) model, whereas the latter paper uses a more traditional Net Present Value approach. In all these papers there is little explicit modeling of the dynamic evolution of the interest rate paid on the account and the balance, and how this evolution depends on changes in the term structure of market interest rates. For example, Jarrow and van Deventer's (1998) model is completely static in the sense that the interest rate paid on the account and the balance are linear functions of the current spot rate. In practice, it is well known that interest rates and balances are rather sluggish and often do not respond immediately to changes in the return on alternative investments, such as the money market rate. Typically, the interest rate paid on the account is set by the bank and the balance is determined by client behaviour. The balance depends, among other things, on the interest rate but also on the return on alternative investments. Because the paths of future interest rates and the adjustment of the balance determine the value of the savings accounts, an analysis of dynamic adjustment patterns is important.

In this paper, we analyze the valuation and hedging of savings deposits with an explicit model for the adjustment of interest rates and balances to changes in the money market rate. A recent paper by Janosi, Jarrow and Zullo (JJZ, 1999) presents an empirical analysis of the Jarrow and van Deventer (1998) model. They extend the static theoretical model to a dynamic empirical model, that takes the gradual adjustment of interest rates and balance into account. Our approach differs from the JJZ paper in several respects.

Firstly, we treat the term structure of discount rates as exogenous and calculate the value of the savings account by a simple Net Present Value equation. This approach, suggested by Selvaggio (1996) leads to simple valuation and duration formulas, and is applicable without assuming a particular term structure model. The drawback of the NPV approach is that we have to assume that the risk premium implicit in the discount factor is constant, but this may be a good first approximation because we want to concentrate on the effects of the dynamic adjustment of the interest rate paid on the account and balance and not on term structure effects.

Secondly, a difference between the JJZ model and ours is the modeling of the long run effects of discount rate shocks. In our model, there is a long run equilibrium, in which the difference between the interest rate paid on the account and the money market rate is constant, and the balance of the savings account is also constant (possibly around a trend). Short term deviations from these long run relations are corrected at a constant rate. This model structure is known in the empirical time series literature as an error correction model[3]. This model has some attractive properties, such as convergence of the effects of shocks to a long-run mean.

The interest rate sensitivity is quantified in a duration measure. We demonstrate that the duration depends on the adjustment patterns of interest rate paid on the account and balance. We pay particular attention to the implications of the model for the hedging of interest rate risk on savings deposits. We illustrate how to fund the savings deposits by a mix of long and short instruments that matches the duration of the savings account's liabilities.

The paper is organized as follows. First the valuation of the savings accounts is dealt with in 2. In 3 the models on the pricing policy and the customer behaviour are presented, and a discrete time version of the model is estimated for the Dutch savings accounts market. 4 deals with the duration of this product and 5 with hedging decisions. The paper is concluded in 6.

2. Valuation of variable rate savings accounts

The valuation problem of savings accounts and similar products was analyzed by Selvaggio (1996) and Jarrow and Van Deventer (1998). Their approach is to acknowledge that the liability of the bank equals the present value of future cash outflows (interest payments and changes in the balance). The present value of these flows does not necessarily equal the market value of the money deposited, and therefore the deposits may have some net asset value. Jarrow and Van Deventer (1998) treat the valuation of savings accounts in a no-arbitrage framework and derive the net asset value under a risk-neutral probability measure. However, in our paper we want to implement an empirical model for the savings rate and the balance, and therefore we need a valuation formula based on the *empirical* probability measure. We therefore adopt the approach proposed by Selvaggio (1996), who calculates the value of the liabilities as the expected present value of future cash flows, discounted at a discount rate which is equal to the risk free rate plus a risk premium[4]. Hence, the discount rate $R(t)$ can be written as

$$R(t) = r(t) + \gamma, \tag{1}$$

where $r(t)$ is the money market rate and γ is the risk premium. We can interpret this discount rate as the hurdle rate of the investment, that incorporates the riskiness of the liabilities, as in a traditional Net Present Value calculation.

The main assumption in this paper is that this risk premium is constant over time and does not depend on the level of the money market rate. This assumption is obviously a simplification. Any underlying formal term structure model, such as the Ho and Lee (1984) model, implies that risk premia depend on the

[3] We refer to Davidson et al. (1978) for an introduction to error correction models.
[4] Selvaggio (1996) calls the risk premium the Option Adjusted Spread

money market rate. However, the risk premia are typically small and since the focus of the paper is on modeling the dynamic adjustment of interest rates and balances, we ignore the variation in the risk premium and focus on the effect of shocks to the money market rate.

With this structure, the market value of liabilities is the expected discounted value of future cash outflows, i.e. interest payments on the account $i(t)$ and changes in the balance $D(t)$ [5]

$$L_D(0) = \mathrm{E}\left[\int_0^{\infty} e^{-Rs}[i(s)D(s) - D'(s)]\,ds\right]. \tag{2}$$

Notice that in this setup reinvestments of interest payments are counted as a part of deposit inflow $D'(t)$. Working out the integral over $D'(s)$ by partial integration we find that the value of the liabilities equals

$$L_D(0) = \mathrm{E}\left[\int_0^{\infty} e^{-Rs}[i(s) - R(s)]\,D(s)\,ds\right] + D(0). \tag{3}$$

Since the market value of the assets is equal to the initial balance, $D(0)$, the net asset value (i.e., the market value of the savings product from the point of view of the bank) is

$$V_D(0) = D(0) - L_D(0) = \mathrm{E}\left[\int_0^{\infty} e^{-Rs}[R(s) - i(s)]\,D(s)\,ds\right]. \tag{4}$$

For an interpretation of this equation, notice that $R(t) - i(t)$ is the difference between the bank's discount rate and the interest paid on the account. Additional savings generate value with return $R(t)$. The costs of these additional savings are $i(t)$, however. The difference $R(t) - i(t)$ therefore can be interpreted as a profit margin.

The net asset value is simply the present value of future profits (balance times profit margin). Therefore, the net asset value is positive if the interest rate paid on the account is on average below the discount rate. Obviously, the net asset value is zero if the interest rate paid on the account always equals the discount rate.

As an example, consider the situation where the interest rate paid on the account is always equal to the discount rate minus a fixed margin, $i(t) = R(t) - \mu$, and the discount rate is constant over time.[6] Moreover, assume that the balance is constant at the level D^*. In that case, the net asset value of the savings accounts is

$$V_D^* = \frac{\mu}{R} D^*. \tag{5}$$

---

[5] For notational clarity, the time variation in the discount rate R is suppressed. If the discount rate is time varying, the exact expression for the discount factor is $e^{-\int_0^{\infty} R(u)\,du}$.

[6] This is a special case of the Jarrow and Van Deventer (1998) model.

Intuitively, this is the value of a perpetuity with coupon rate μ and face value D^*. Figure 1 graphs the net asset value for different values of R and μ. For large profit margins and low discount rates, the net asset value can be a substantial fraction of the market value of the savings deposits.

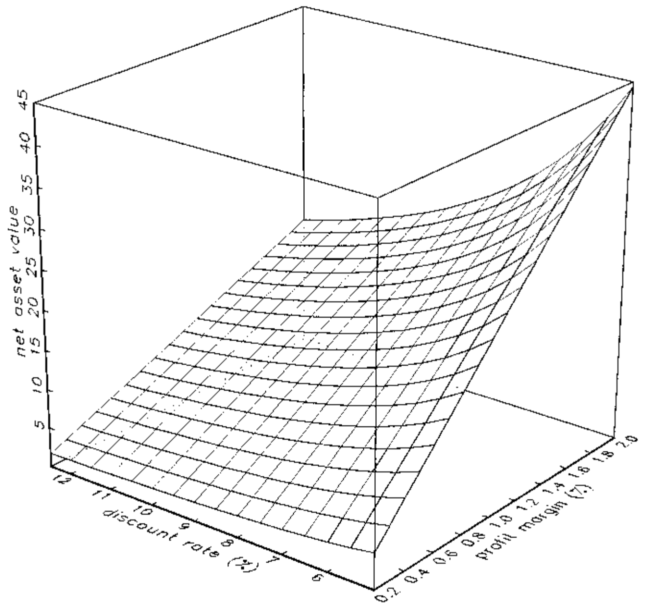


FIGURE 1: Net asset value
This figure shows the net asset value of a deposit of 100, as a function of the discount rate R and the profit margin μ

Obviously, this example describes the value in a static setting. For the interest rate sensitivity of the net asset value, we have to take into account that after a shock in interest rates, the interest rate paid on the account and the balance only gradually adjust to their new equilibrium values. In the next section we therefore present a model for the adjustment patterns of interest rate and balance after shocks to the discount rate. In the subsequent section we present discount rate sensitivity measures based on these adjustment patterns.

3. CLIENT AND BANK BEHAVIOUR

The analysis in the previous section shows that the net asset value of savings accounts depends on the specific pattern of the expected future interest rates and balances. The main difference between money market accounts and savings accounts is the sluggish adjustment of interest rates and balance to changes in the discount rate. In this section we model these adjustment processes. The models highlight the partial adjustment toward the long run equilibrium values of interest rates and balances. In the analysis, we take as given the path of the money market rate $r(t)$ and hence the path of the discount rate $R(t) = r(t) + \gamma$. We describe the stochastic evolution of the interest rate paid on savings deposits, $i(t)$, and the balance, $D(t)$, conditional on the path of the discount rate.

For the interest rate paid on savings accounts, we propose the following stochastic error correction specification

$$di(t) = \kappa[R(t) - \mu - i(t)]dt + \sigma_1 dW_1(t) \tag{6}$$

where $W_1(t)$ is a standard Brownian motion. This equation states that the interest rate adjusts to deviations between the long run value $R(t) - \mu$ and the current rate. We see this as the target policy rule of the bank that sets the interest rate. Deviations are corrected at speed $\kappa > 0$, and in the long run, expected interest rates are a margin μ below the discount rate $R(t)$. The stochastic term $W_1(t)$ models the deviations from the target policy rule. Such deviations could be due to sudden demand shocks, competition from other banks and the like.

For the balance we propose a partial adjustment specification

$$dD(t) = -\lambda\left[D(t) - D^*\right]dt - \eta\left[R(t) - \mu - i(t)\right]dt + \sigma_2 dW_2(t) \tag{7}$$

This specification has three components. Firstly, there is an autonomous convergence to a long run mean D^*, which is determined by a tradeoff by the clients between savings deposits and money market accounts. Secondly, there is an outflow of funds proportional to the excess of the discount rate over the savings rate. Thirdly, there is an unpredictable stochastic component.

This description with an autonomous convergence is especially suitable for a detrended time series. An autonomous convergence to a long run mean is expected in a detrended series for the balance. We detrend by defining the variable $D(t)$ as the fraction of total short term savings that is invested in variable rate savings accounts. In this case D^* is the long run fraction of total short term savings that is invested in variable rate savings accounts. In this way, the trend growth of the total savings market doesn't affect the empirical estimation and the duration analysis.

Working out the stochastic differential equations (6) and (7) gives:

$$i(t) = e^{-\kappa t} i(0) + \kappa \int_0^t e^{\kappa(s-t)}[R(s) - \mu]ds + \sigma_2 \int_0^t e^{\kappa(s-t)} dW_1(s), \tag{8a}$$

$$D(t) = D^* + e^{-\lambda t}\left(D(0) - D^*\right) - \eta \int_0^t e^{\lambda(s-t)}[R(s) - \mu - i(s)]ds + \sigma \int_0^t e^{\lambda(s-t)} dW_2(s). \tag{8b}$$

To interpret these equations, let's consider the situation where the discount rate R is constant over time. It is fairly easy to show that the effect of a change in the discount rate in this situation is given by the following partial derivatives

$$\frac{\partial i(t)}{\partial R} = \kappa \int_0^t e^{\kappa(s-t)} ds = 1 - e^{-\kappa t}, \tag{9a}$$

$$\frac{\partial D(t)}{\partial R} = -\eta \int_0^t e^{\lambda(s-t)} \left[1 - \frac{\partial i(s)}{\partial R}\right] ds$$

$$-\eta \int_0^t e^{\lambda(s-t)} e^{\kappa s} ds = -\eta \left(\frac{e^{-\lambda t} - e^{-\kappa t}}{\kappa - \lambda} \right). \tag{9b}$$

The long run derivative of the interest rate paid on the account is one, but in the short run the effect is less than one. If $\eta > 0$ and $\kappa > \lambda$ (which we show later is clearly the case empirically), the partial derivative of the balance is negative, and converges to zero in the long run.

These partial derivatives can be used to study the effects of a once-and-for-all shock to the discount rate, a kind of impulse response analysis. Starting from the equilibrium situation $D(0) = D^*$ and $i(0) = R - \mu$, the expected adjustment patterns are illustrated in Figure 2 for an increase in the discount rate by 1%. The parameter values are picked from the empirical estimates to be discussed shortly, and are equal to $\kappa = 0.79$, $\lambda = 0.048$ and $\eta = 0.43$ for the base case. We see that the interest rate doesn't follow the jump in the discount rate immediately but gradually adjusts to its new equilibrium value. The adjustment of the balance is more complex. Initially, the balance decreases because of withdrawals caused by the relatively low interest rate paid on the account. But as the interest rate increases, this effect becomes smaller and eventually the autonomous convergence of the balance to its long run level dominates. One interpretation of this is that clients who initially preferred the variable rate savings account to the money market account will return to variable rate savings accounts when the difference between the interest rate paid on the savings account and the money market rate reverts to the initial level.

Equations (9a) and (9b) also highlight the effects of the model parameters on the adjustment of interest rates and balance to a shock in the discount rate. The effect of η is obvious, it increases the impact of an interest rate shock. This effect may be important in the current market, as the increase in the use of internet for banking services and the resulting lower transaction and search costs will probably increase the interest rate sensitivity of the customers. The effect of the mean-reversion parameters κ and λ is more complicated. A higher value of λ speeds up the adjustment of the balance itself, but doesn't affect the interest rates. With a lower value of κ, both the adjustment of the interest rate and the balance are slower. The effect of the balance is a result of the dependence of the balance on the interest rate. These effects are illustrated in Figure 2, where the dashed line gives the adjustment pattern for a lower value of κ, and the dotted line the pattern with a higher value of λ.

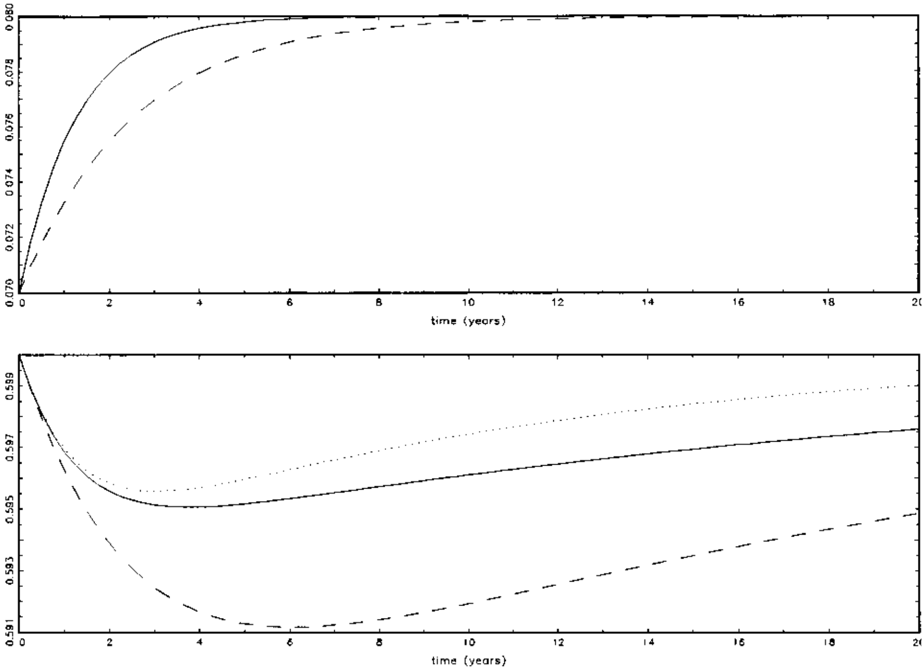


FIGURE 2: Adjustment of interest rate and balance of savings accounts
This figure shows the adjustment of interest rate (top panel) end balance (bottom panes) to a 1% shock in the discount rate. The solid line is the base case. The dashed line is for a smaller value of κ, the dotted line for a larger value of λ. The scale of the horizontal axis is years.

We now present some indicative estimates of the model parameters. This exercise is not meant to be a thorough empirical investigation of the adjustment pattern but merely serves as an illustration of the model. In order to translate the continuous time parameters to a discrete time setting, we use the following approximate[7] discretization of the continuous time model

$$\Delta i_t = \kappa\left[R_{t-1} - \mu - i_{t-1}\right]\Delta t + \epsilon_{1t}, \tag{10a}$$

$$\Delta D_t = -\lambda\left(D_{t-1} - D^*\right)\Delta t - \eta\left[R_t - \mu - i_t\right]\Delta t + \epsilon_{2t}. \tag{10b}$$

The discount rate is not directly observed in the data. Since a savings account shares characteristics of both a money market account and a long term deposit, its required rate of return (or discount rate) is proxied by a weighted average of the money market rate (r_t) and the long term bond yield (y_t).[8]

[7] This approximation is quite accurate. For example, the exact mean reversion parameter for the interest rate equation is $1 - \exp(-\kappa\Delta t)$, which for small values of κ or Δt is close to $\kappa\Delta t$.

[8] An alternative but equivalent way to justify this proxy is to assume that the risk premium of the savings deposit is a fraction of the risk premium on long term bonds.

We treat the weight δ as an unknown parameter which is estimated from the data. This leads to the following empirical model

$$\Delta i_t = \alpha_0 + \alpha_1 \Delta r_t + \alpha_2 \left[i_{t-1} - \{\delta r_{t-1} + (1-\delta) y_{t-1}\} \right] + e_{1,t}, \tag{11a}$$

$$D_t = \beta_0 + \beta_1 D_{t-1} + \beta_2 \left[i_{t-1} - \{\delta r_{t-1} - (1-\delta) y_{t-1}\} \right] + e_{2,t}. \tag{11b}$$

This model is slightly more general than the theoretical model because it contains an immediate, discrete adjustment of the interest rate to the money market rate. After this initial jump, the adjustment to the new equilibrium is gradual. This effect turned out to be so important empirically that we included it in the empirical model.[9] The parameters of the continuous time model can be solved from the following equations (with $\Delta t = 1/12$ for monthly data)

$$\kappa = -\alpha_2 / \Delta t,$$

$$\lambda = (1 - \beta_1)/\Delta t,$$

$$\eta = \beta_2 / \Delta t.$$

In fact, the long run deposit level and the average spread of the interest rates over the estimated discount rate could be unraveled from the constant terms of the model. These are not very accurately estimated however and we refrain from drawing inferences about these parameters from the estimates.

We use monthly data on interest rates and deposits from the Dutch savings account market. The interest rate paid on the account is taken from one of the price setters in the Dutch market. The sample period is 1982:12 to 1999:12, spanning 17 years which is slightly longer than the samples of Hutchison and Pennacchi (1996) or JJZ. To remove trends in the total savings volume, we define the balance D_t as the fraction of variable rate savings accounts to total savings. The following empirical estimates are obtained using least squares:

$$\Delta i_t = -0.084 + 0.072\,\Delta r_t - 0.066 \left[i_{t-1} - \{\delta r_{t-1} + (1-\delta) y_{t-1}\} \right] + e_{1,t}, \tag{12a}$$

$$D_t = 0.21 + 0.996 D_{t-1} + 0.039 \left[i_{t-1} - \{\delta r_{t-1} + (1-\delta) y_{t-1}\} \right] + e_{2,t}. \tag{12b}$$

The estimate of δ is around 0.2. These estimates imply the following annualized values for the continuous time parameters: $\kappa = 0.79$, $\lambda = 0.048$, and $\eta = 0.43$. Using these parameters we can solve the second equation for the steady state value of the fraction of variable rate savings deposits to total savings, $D^* = 0.58$.[10]

[9] Notice that including this term does not invalidate the duration analysis of the model, which is based on the gradual adjustment patterns only.

[10] The empirical average of D_t is 0.51

4. Duration

The previous section showed that the interest rate paid on the account and the balance of savings accounts are related to the discount rate. Therefore, the discount rate sensitivity of savings deposits will be different from the discount rate sensitivity of a money market account (which has a duration of zero). In this section, we study the sensitivity of the net asset value of a savings account to a parallel shift in the path of the discount rates. We study a shift from the original path $R(t)$ to $R(t) + \Delta R$, and evaluate the derivative in $\Delta R = 0$. With some abuse of notation, we will write the resulting expressions as $\partial V / \partial R$ but it should be kept in mind that this refers to a parallel shift in the path of discount rates. This approach is close to a traditional duration analysis, see e.g. Bierwag (1987), but we take into account the dependence of future cash flows on discount rates.

In the initial situation, the deposits are at their equilibrium value D^*. Differentiation of the net asset value with respect the discount rate gives

$$\frac{\partial V_D(0)}{\partial R} = \mathrm{E}\Big[-\int_0^\infty s e^{-Rs}[R(s) - i(s)]\, D(s) ds + \int_0^\infty e^{-Rs} \frac{\partial [R(s) - i(s)]}{\partial R} D(s) ds$$
$$+ \int_0^\infty e^{-Rs}[R(s) - i(s)] \frac{\partial D(s)}{\partial R} ds\Big] \tag{13}$$

The three components of this expression can be interpreted as follows:

1. the interest rate sensitivity of the expected discounted profits;
2. the change in the margin on the expected future balances;
3. the expected margin times increases or decreases in the balance of the deposit.

Notice that if the future balances do not change as a result of the interest rate change, and if the margin is constant, only the first term (the sensitivity of the present value of the profits) remains. The second and third term are specific for savings accounts with their slow adjustment of the interest rate and balance, and are therefore the most interesting for our analysis. We shall now discuss the duration of the accounts given the specific model for the evolution of interest rates and balances.

Assume again that $R(s) = R$ is constant, and that the initial situation is in equilibrium, $D(0) = D^*$ and $i(0) = R - \mu$. Under these initial conditions, the development of the interest rates and the balance can be derived from equations (8a) and (8b):

$$R - i(t) = R - e^{-\kappa t} i(0) - \kappa \int_0^t e^{\kappa(s-t)}[R - \mu] ds - \sigma_1 \int_0^t e^{\kappa(s-t)} dW_1(s)$$
$$= \mu - \sigma_1 \int_0^t e^{\kappa(s-t)} dW_1(s), \tag{14a}$$

$$D(t) = D^* + \eta \int_0^t e^{\lambda(s-t)}[R(s) - \mu - i(s)] ds + \sigma_2 \int_0^t e^{\lambda(s-t)} dW_2(s)$$
$$= D^* + \eta\sigma_1 \int_0^t e^{\lambda(s-t)} \int_0^s e^{\kappa(u-s)} dW_1(u) ds + \sigma_2 \int_0^t e^{\lambda(s-t)} dW_2(s). \tag{14b}$$

Assuming that the stochastic parts of the interest rate and the balance are uncorrelated, i.e. $\mathrm{Cov}(dW_1(t), dW_2(t)) = 0$, and noticing that the partial derivatives (9a) and (9b) are non-stochastic, we can work out the partial derivative of the value:

$$\begin{aligned}\frac{\partial V_D(0)}{\partial R} &= -\int_0^\infty s e^{-Rs}\,\mathrm{E}\{[R - i(s)]D(s)\}ds \\ &\quad + \int_0^\infty e^{-Rs}\frac{\partial [R - i(s)]}{\partial R}\mathrm{E}[D(s)]ds + \int_0^\infty e^{-Rs}\,\mathrm{E}[R - i(s)]\frac{\partial D(s)}{\partial R}ds \\ &= -\int_0^\infty s e^{-Rs}\mu D^* ds + \int_0^\infty e^{-Rs}e^{-\kappa s}D^* ds - \int_0^\infty e^{-Rs}\mu\eta\left(\frac{e^{-\lambda t} - e^{-\kappa t}}{\kappa - \lambda}\right)ds \\ &= -\frac{\mu}{R^2}D(0) + \frac{1}{R+\kappa}D(0) - \frac{\eta\mu}{\kappa - \lambda}\left(\frac{1}{R+\lambda} - \frac{1}{R+\kappa}\right). \end{aligned} \tag{15}$$

With an increase in the discount rate, the first term reflects the loss of value of the (perpetual) profit margin, the second term the discounted value of the interest payments not made on the original balance during the time the interest rate paid on the account ($i(t)$) is below the discount rate minus the profit margin ($R - \mu$), and the third term the discounted value of the profit foregone on the balance outflows.

We can transform this change of value to a duration measure if we assume that initially, the net asset value equals $V_D(0) = D(0)\mu/R$

$$Dur = -\frac{\partial V_D(0)}{\partial R}\frac{1}{V_D(0)} = \frac{1}{R} - \frac{1}{\mu}\frac{R}{R+\kappa} + \frac{\eta}{D(0)}\frac{1}{\kappa - \lambda}\left(\frac{R}{R+\lambda} - \frac{R}{R+\kappa}\right). \tag{16}$$

The first term reflects the duration of a perpetuity, and is determined by the present value of the profits in the steady state. The second term reflects the value of the lower interest rates paid on the existing balance, and is always negative. The third term is the duration of the profits on the additional balance outflows, and is positive under the assumption $\kappa > \lambda$. Especially when the margin μ is thin and the net asset value is low, the second term may dominate the other terms, leading to a negative duration for the net asset value of a savings account. In that case, an increase in the discount rate will increase the net asset value because for some time the interest rate paid on the savings account is lower than return on the assets deposited.

As an illustration Figure 3 shows the durations as a function of the discount rate R and the margin μ (the other parameters are put equal to the estimates of the previous section). We see that the duration is typically positive, except for low values of μ, and declines with the discount rate. Most of this effect is due to the duration of the discounted profit margin, $1/R$. Leaving out this term, we find the 'extra' duration of the net asset value induced by the sluggish adjustment pattern. Figure 4 shows these measures. Interestingly, the 'extra' duration is always negative, but converges to zero for relatively big profit margins μ.

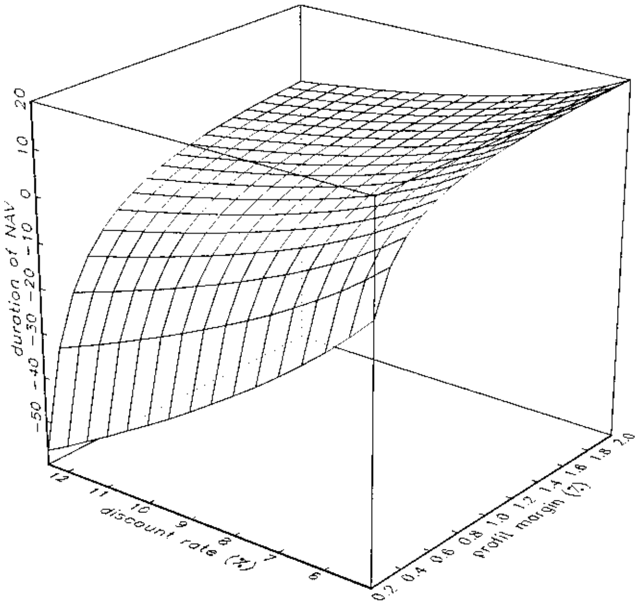


FIGURE 3: Duration of savings deposits
This figure shows the duration (in years) of savings deposits as a function of the discount rate (R) and the profit margin μ.

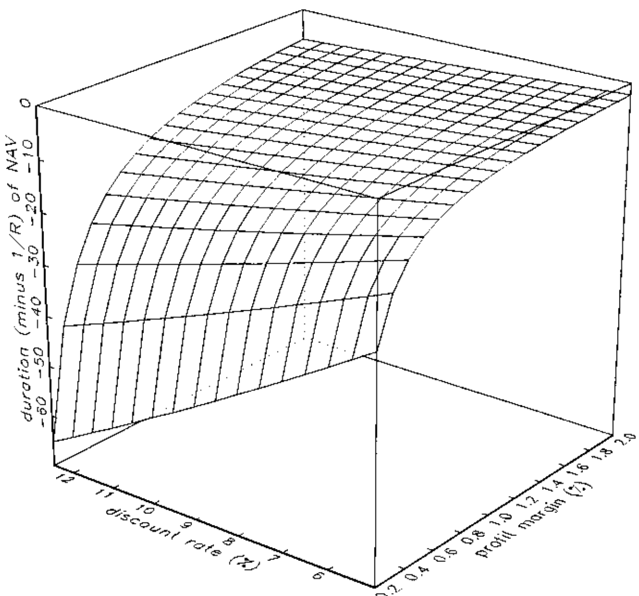


Figure 4: Duration of savings deposits (excluding profit margin)
This figure shows the extra duration (in years) of savings deposits, in excess of the duration of a perpetuity ($1/R$), as a function of the discount rate (R) and the profit margin μ.

5. Hedging

In this section we consider the problem of hedging the net asset value. Given the liability value L_D of the variable rate savings accounts, one can hedge the net asset value by immunization. For simplicity we assume the money deposited can be invested in two instruments, Long Investments (*LI*) and Short Investments (*SI*). The balance sheet of the bank then becomes

V_{LI}	L_D
V_{SI}	V_D

where V_D denotes the Net Asset Value. We now consider the construction of an investment portfolio where the interest rate risk on the net asset value is fully hedged, i.e. the net asset value V_D is not sensitive to the parallel shifts in the discount rate. From the balance sheet we see that this requires

$$\frac{\partial V_{SI}}{\partial R} + \frac{\partial V_{LI}}{\partial R} = \frac{\partial L_D}{\partial R}. \tag{17}$$

Of course, the solution to this equation, and hence the composition of the investment portfolio, depends on the durations of the short and long investments. As a simple example, consider the case where the short instrument has zero duration. In that case the investment in the long instrument is determined by

$$\frac{\partial V_{LI}}{\partial R} = \frac{\partial L_D}{\partial R}. \tag{18}$$

We can find $\frac{\partial L_D}{\partial R}$ from equations (4) and (13).

As an illustration, Figure 5 graphs the required position in long (10 year maturity) bonds in the hedge portfolio for different value of R and μ. As seen before, the duration of variable rate savings accounts may be negative, in particular when the profit margin μ is fairly small. In that case the bank can hedge the accounts by taking a long position in long investments. But if *Dur* is positive, which happens for example when the profit margin μ is fairly high, one should take a short position in the long asset. Alternatively, if one does not like to take short positions in bonds, one could use derivative instruments such as caps, which typically have a negative duration, or forward contracts.

6. Conclusion

This paper focuses on the valuation and interest rate sensitivity of variable rate savings accounts. The duration can be split in three different effects:

- the duration of the expected discounted profits;
- the change in margin on expected future balances due to a change in interest rate;
- the expected margin times increases or decreases in the balance of the account.

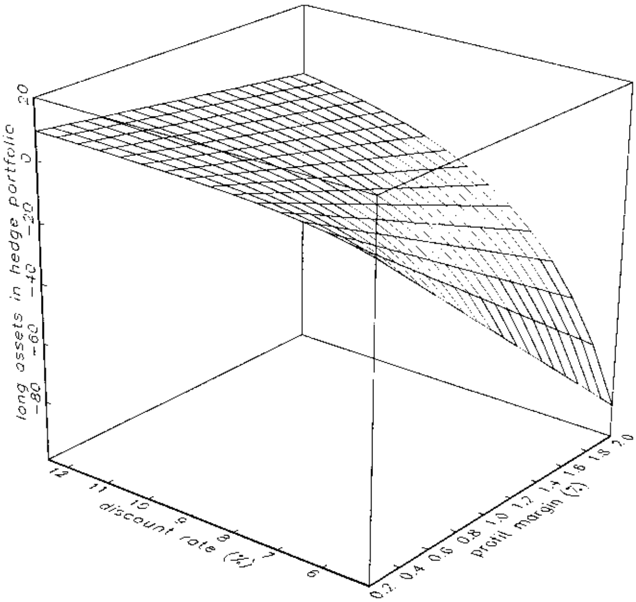


Figure 5: Hedge portfolio
This figure shows the position in long bonds (duration 10 years) in the hedge portfolio of a 100 deposit, as a function of the discount rate R and the profit margin μ.

The first element is the standard duration for products without embedded options. The second and third term are non-standard (for example, they are zero for a money market account) and arise due to the variable interest rate paid on the account and the option of the clients to withdraw and invest in the account at any time. The duration therefore crucially depends on the rapidness of the adjustment of the interest rate paid on the account to discount rate changes and on the reactions of the clients. These reactions will principally be determined by the clients interest rate sensitivity and by the market efficiency. The models are estimated for the Dutch savings account market. Duration curves are given for different margins.

When hedging the savings deposits, one can construct a portfolio with the same duration as the variable rate savings accounts. However, when one does not want to go short into a certain asset class, one might need to include derivatives (for example caps) to hedge these products, since it is possible to have negative durations. The intuition is that an interest rate increase might lead to a flight of clients to money market accounts. So buy ‘insurance’ when money market accounts are less attractive, which result in profits when interest rates spike up (the insurance pays out). The gain due to the caps in an increasing interest rate environment then offsets the loss in the savings accounts. Hedging in this way certainly smoothens the results on these products. Of course this can be achieved by going short in long assets as well.

For future research it might be interesting to analyze the second order effects. Then multiple immunization can be achieved with a portfolio with three asset classes. Finally, it is possible to make the discount rate a function of a number of interest rates with different maturities. This will of course increase the complexity of the model but allows for the calculation of key-rate durations.

References

Bierwag, G.O. (1987) *Duration Analysis*, Ballinger, Cambridge MA.

Davidson, J., Hendry, D.F., Srba, F. and Yeo, S. (1978) Econometric Modelling of the Aggregate Time Series Relationship between Consumer' Expenditure and Income in the United Kingdom, *Economic Journal*, **88**, 661-692.

Heath, D., Jarrow, R. and Morton, A. (1992) Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation, *Econometrica* **60**, 77-106.

Ho, T.S.Y. and Lee, S.-B. (1986) Term structure movements and the pricing of interest rate contingent claims, *Journal of Finance* **41**, 1011-1029.

Hull, J. (1993) *Options, Futures and other Derivative Securities*, second edition, Prentice-Hall.

Hutchison, D.E. and Pennacchi, G.G. (1996) Measuring Rents and Interest Rate Risk in Imperfect Financial Markets: The Case of Ratail Bank Deposits, *Journal of Financial and Quantitative Analysis* **31**, 401-417.

Janosi, T., Jarrow, R. and Zullo, F. (1999) An Empirical Analysis of the Jarrow-van Deventer Model for Valuing Non-Maturity Demand Deposits, *Journal of Derivatives*, Fall 1999, 8-31.

Jarrow, R.A., and van Deventer, D.R. (1998) The arbitarge-free valuation and hedging of savings accounts and credit card loans, *Journal of Banking and Finance* **22**, 249-272.

Selvaggio, R.D. (1996) Using the OAS Methodology to Value and Hedge Commercial Bank Retail Demand Deposit Premiums, *Chapter 12* 363-373.

Vasicek, O. (1977) An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**, 177-188.

Frank de Jong
Finance Group
Universiteit van Amsterdam
Roetersstraat 11
1018 WB, Amsterdam
the Netherlands
Phone: +31-20-5255815
Fax: +31-20-5255285
Email: f.c.j.m.dejong@uva.nl

PREDICTION OF STOCK RETURNS: A NEW WAY TO LOOK AT IT

BY

JENS PERCH NIELSEN[1] AND STEFAN SPERLICH[2]

ABSTRACT

While the traditional R^2 value is useful to evaluate the quality of a fit, it does not work when it comes to evaluating the predictive power of estimated financial models in finite samples. In this paper we introduce a validated R_V^2 value useful for prediction. Based on data from the Danish stock market, using this measure we find that the dividend-price ratio has predictive power. The best horizon for prediction seems to be four years. On a one year horizon, we find that while inflation and interest rate do not add to the predictive power of the dividend-price ratio then last years excess stock return does.

KEYWORDS

Prediction, Stock returns, Dividend price ratio, Cross Validation.

1. INTRODUCTION

Long term investors have the contradicting aims of minimizing risk and maximizing return over the long run. Much financial literature investigates trading patterns and strategy among long term investors, for example, Barber and Terrance (2000) argue for a buy-and-hold type of strategy that does not eat up returns by trading costs and many professional advisers argue that stocks are better over the long run, see Siegel (1998) and Jagannathan and Kocherlakota (1996) for particular easily read accounts on this. Other professional financial advisers say that expected returns in financial markets vary over time and contain a significant predictable component. Consequently time periods exist where the long term investor might choose to sell stocks and buy

* The research was supported by the Danish Social Science Research fund and the Spanish DGES BEC2001-1270. Comments and suggestions from Thomas Engsted, Carsten Tanggård and two anonymous referees are gratefully acknowledged.

[1] Codanhus, 60 Gammel Kongevej, DK-1790 Copenhagen V, Denmark. Email: npj@codan.dk.

[2] Universidad Carlos III de Madrid, Email: stefan@est-econ.uc3m.es.

bonds, because the return on stocks in these time periods do not match the risk involved. The dividend-price ratio and the earning-price ratio, in particular, has proven to have some predictive power for future stock returns, see Campbell, Lo, and MacKinlay (1997, Chapter 7) for an up-to-date account regarding of the dividend yield based predictability of stock returns, and see Shiller (2000, p.8) for a warning of an overvalued American stock exchange based on the earning-price ratio. Campbell, Lo, and MacKinlay (1997) argued that the predictable component of stock returns is increasing with the time horizon, since the measure of fit, the R^2, increases rapidly with the time horizon. In the actuarial literature Wilkie (1993) gets to a similar conclusion replicating the linear modeling approach of Fama and French (1988). The following quote is from Wilkie (1993, p.341), who found that predictive power seems to be strongest for a six and a half year time horizon for British data: "A 1% difference in the dividend yield at the time of purchase of the stock makes a difference… equivalent to about 4.1% a year compound for about six and a half years". So, there is a considerable financial and actuarial literature on the predictive power of the dividend yield, see also Richardson and Stock (1989) and Wilkie (1995). However, most literature uses traditional in-sample methodology like goodness of fit, the traditional R^2 value or parametric estimation procedures combined perhaps with some testing. In this paper we consider an adjusted measure of predictive power, the R_V^2 value, that is an out-of-sample measure in the sense that it measures how the model actually predicts into the future, see Eun and Resnick (1988,1994) for financial papers using a similar type of out-of-sample approach to evaluate their times series of stocks, bonds and exchange rates as we use in this paper to analyze our financial time series.

The paper follows its historical development. First we go through an analysis of Danish data from 1922 to 1996 and then we add the analysis of the updated data set from 1922 to 2001.

Regarding the data set 1922-1996

Based on our out-of-sample measure, it seems that the models have strongest predictive power for a time horizon of four years for Danish stock returns, at least with respect to our criterion. Dividend-price does seem to have predictive power whereas knowledge of inflation and short interest rates do not seem to add to this predictive power. However, our study shows that the one year lagged returns do. The best predictive filter on a one year basis turns out to be a two-dimensional fully nonparametric estimator based on the dividend-price ratio and last years lagged excess return. Last years excess return enters with a tendency towards reversal, such that good years tend to follow bad years and vice versa. The dividend-price ratio is, however, still the most indicative parameter while estimating the excess returns of the coming years.

Moreover, based on the current level on the dividend-price ratio in Denmark, around 1%, we concluded (in december 2000) that expected excess returns on stocks are indeed below zero, for all the considered time horizons with good prediction power, namely one, two, three, four and five years time horizons. Based on this finding we then concluded further that the current market and political situation in Denmark was out of balance, since all institutional investors heavily increased their percentages of stocks in their portfolios right then. On average, an increase from around 20% invested in stocks to around 40% invested in stocks have been seen for long term institutional investors in Denmark over the last seven years. The model of this paper argue that this strategy increases the risk without increasing the average return.[3] We believe that the considerations of this paper can be helpful while developing a modern information system for the long term investor.

Regarding the updated data set 1922-2001

It turns out that while there still is predictive power in the updated data set, it is much lower than in the original data set. This remarkable finding can have two explanations. Either the Danish main index, the KFX index, has had an exceptional behavior in some of the last five years or the entire world market has followed exceptional rules during this period. In either case, our findings show that some care has to be taken regarding predictive power of dividend yields and that further studies based on international data sets should be added to the current to get a fuller understanding of the problem. It could for example be relevant to consider a regime shift model such as the one of Harris (1999) to understand this question further. However, while a regime shift model does add to the understanding of historical facts, it does not help much when it comes to predictive power. For a deep insight into the nature of uncertainty in prediction, see Cairns (2000, p 314). In this paper we only consider the first of the three steps considered by Cairns, namely Method 1 that finds the best fit to a model according to a certain criterion. However, since our criterion is a validated measure of error, we implicitly take care of the errors dealt with in Cairns Method 2 and Method 3, that consider uncertainty due to parameter estimation and model estimation. Another relevant extension would be to combine the world wide data of Dimson, March and Staunton (2002) with the predictive methodology of this paper.

We first motivate our choice of regression variables by noting the basic relationship between stock returns and economic factors in Section 2. In Section 3 we describe our data. Our framework for prediction is given in Section 4 followed by the prediction results based on the data set 1922-1996 in Section 5 when dividend yield alone is used for prediction. In Section 6 we consider the use of more regression variables for the data set 1922-1996 and in Section 7 we shortly comment on the updated results for the data set 1922-2001.

---

[3] This remark is from a version of this paper dated february 2001. A talk based on the conclusions of this paper was given to the Danish Actuarial Society in december 2000 under the title "Be careful: the Danish stocks are too expensive".

2. The basic relationship between stock returns and economic factors

One traditional equation for the value of a stock is

$$P_t = \sum_{j=1}^{\infty}(1+\gamma)^{-j}(1+g)^{j-1}D_t,$$

where most of the entering quantities on the right hand side are unknown, γ, discount rate, g, constant growth of dividend yields, i inflation and D_t is the dividend paid out during the period t. This model was introduced to the financial theory by Williams (1938) and Gordon and Shapiro (1956). Campbell and Shiller (1988) referred to the model as the "dividend-ratio" in absence of uncertainty, see also Goetzman and Jorion (1993), Hodrick (1992), and Fama and French (1988). For simplicity the discount rate and the growth rate do not depend on time in this model although this is well known to be incorrect. The point of the above identity is however, that it shows that the price of stocks depend on quantities such that dividend yield, interest rate and inflation. The two latter being highly correlated with almost any relevant discount rate. It is also clear from the above identity that a decrease in discount rate, which is highly correlated with an increase in bond yield, are related to an increase in the stock return and vice versa.

3. The data and our definition of prediction

In this paper we use the annual Danish stock market data from Lund and Engsted (1996), respectively the extended sample period 1922-1996 from Engsted and Tanggård (2000). We have ourselves extended the period to 1922-2001. We consider the time series

$$W_t = (S_t, d_t, I_t, r_t),$$

where S_t is stock return, I_t is inflation and r_t is the short-term interest rate. The stock index is based on a value weighted portfolio of individual stocks chosen to obtain maximum coverage of the marked index of the Copenhagen Stock Exchange (CBS). Notice that CBS was open during the second world war. In constructing the data corrections were made for stock splits and new equity issues below market prices. Further, $d_t = D_t / P_t$ denotes (nominal) dividends D_t paid during year t divided by the (nominal) stock price P_t at the end of year t. The appendix in Lund and Engsted (1996) contains a detailed description of the data from where we have taken the following quote: "A nominal stock index and accompanying dividend series was constructed from the original price quotation sheets from the Copenhagen Stock Exchange. In order to avoid a possible tax-induced distortion due to the well known January effect, the stock index at the end of the year t is defined as the value in (mid) February of year $t+1$. Similarly, dividends for year t are defined as dividends paid out between February of year t and February of year $t+1$. However, no Danish companies pay dividends in

January, so the dividend series is effectively the dividends paid during the year t. Corrections are made for stock splits and new equity issues below the market price using techniques similar to those described in Shiller (1981). The stock index is a value-weighted portfolio consisting of approximately 16 individual stocks (companies), which are generally chosen in order to obtain the maximum coverage of the 'market' index of the Copenhagen Stock Exchange. The composition of the stock index is changed about every 10 years, and the weights for the individual stocks are only adjusted in connection with changes in the composition of the stock index."

We have updated the data set of Lund and Engsted (1996) following their original approach. The leading Danish stock index, the KFX, index has been used for this purpose. As a measure of the short-term interest rate, R_t, the discount rate of the Danish Central Bank's is used up to 1975, spliced together with a short-term zero-coupon yield for the period thereafter. In computing real values, we deflate nominal values by the consumption deflator[4]. The real excess stock return is defined as

$$S_t = \log\left\{\left(P_t + D_t\right)/P_{t-1}\right\} - r_{t-1},$$

where

$$r_t = \log\left(1 + R_t/100\right).$$

The resulting average of these excess stock returns are 2.5% for the period 1922-2001 (2.1% for 1922-1996) and 3.4% for the after war period 1948-2001 (3.2% for 1948-1996).

4. Our framework for prediction

The problem of prediction is considered as follows: Let $Y_t = \sum_{i=0}^{T-1} S_{t+i}$ be the excess stock return at time t over the next T years. We base our prediction on the assumption that Y_t can be approximated by a model of the form:

$$Y_t = g\left(W_{t-1}\right) + \epsilon_t, \qquad t \in \left\{K_1, \ldots, K_2\right\}, \tag{1}$$

where the error variable ϵ_t are mean zero stochastic variables given the past, $W_1, \ldots, W_{t-1}$ and $S_1, \ldots, S_{t-1}$[5]. Ideally we would like to be able to predict Y_t. We do, however, only have information of W_{t-1} and no information with respect to the error term ϵ_t. Therefore, estimating $g(\bullet)$ and using it for our

[4] The consumption series has been from Hansen (1974) and various publications from the Danish Central Statistical Bureau. The consumption series covers private consumption of durable and non-durable goods. Unfortunately, there is no price deflator for private consumption, so nominal prices, dividends and consumption are deflated using the consumption deflator for total consumption.

[5] Note that in our implementation W_{t-1} does not contain time lagged information as is the case e.g in the Wilkie model (1995). An investigation of time lagged variables would be an obvious extension of our approach to prediction.

prediction is the best we can do. Due to the definition of Y_t the time period (K_1, K_2) depends on T and is $\left(T_{first}, T_{last} - T + 1\right)$ with $T_{first} = 1923$ or 1949 and $T_{last} = 1996$ or 2001.

Let

$$X = \left\{\left(W_{K_1 - 1}, S_{K_1}\right), \ldots, \left(W_{K_2 - 1}, S_{K_2}\right)\right\}.$$

For t's where $K_1 \le (t - T) \le K_2$ we wish to be able to consider data points which exclude direct information about S_t. We therefore introduce

$$X^{(t)} = \left\{\left(W_{K_1 - 1}, S_{K_1}\right), \ldots, (W_{t-T-1}, S_{t-T}), (W_{t+T-1}, S_{t+T}), \ldots, \left(W_{K_{2-1}}, S_{K_2}\right)\right\}.$$

Now let the set H represent different estimation principles and let for $\hat{g}_h$, $h \in H$, be some estimator based on X and let $\hat{g}_h^{(t)}$ be the (equivalent) estimator based on $X^{(t)}$.

For a given time horizon T, we define the loss of the estimator $\hat{g}_h$ as

$$Q(\hat{g}_h) = \sum_{t = K_1}^{K_2} \left\{g(W_{t-1}) - \hat{g}_h(W_{t-1})\right\}^2 \tag{2}$$

which can be estimated by

$$\hat{Q}(\hat{g}_h) = \sum_{t = K_1}^{K_2} \left\{Y_t - \hat{g}_h^{(t)}(W_{t-1})\right\}^2, \tag{3}$$

i.e. we predict $g(W_{t-1})$ without the information contained in Y_t, compare also with Appendix 3. Notice that $Q(\hat{g}_h)$ is not estimated well by the goodness of fit measure

$$\overline{Q}(\hat{g}_h) = \sum_{t = K_1}^{K_2} \left\{Y_t - \hat{g}_h(W_{t-1})\right\}^2,$$

since this measure always will be in favor of the most complex model. Such complicated models are often in contradiction to the aim of predicting well.

While predicting, we find our optimal prediction scheme by minimizing $\hat{Q}(\hat{g}_h)$ over all principles h. This gives us the best predictor within H.

Now let h_0 correspond to the trivial prediction strategy based on the parametric model

$$Y_t = \mu + \epsilon_t, \tag{4}$$

where μ is estimated by $\hat{\mu} = (K_2 - K_1 + 1)^{-1} \sum_{t = K_1}^{K_2} Y_t$. For a given modeling and estimation principle h, we define our new R^2 value, that we call $R^2_{V,h}$, where V stands for validated, as

$$R^2_{V,h} = 1 - \frac{\hat{Q}(\hat{g}_h)}{\hat{Q}(\hat{g}_{h_0})}. \tag{5}$$

Notice that $R^2_{V,h}$ measures how well a given model and estimation principle h predicts compared to the simple estimation principle h_0. If $R^2_{V,h}$ is positive then we say that the modeling and estimation principle h predicts otherwise we say that the principle h does not predict. In the following we suppress h in the notation and rewrite $R^2_{V,h}$ as R^2_V. No confusion can occur since it will always be clear what h is under consideration. Note that $R^2_V \in (-\infty, 1]$ and $R^2_V > 0$. The interpretation of R^2_V is similar to the one of the classical R^2 that can be defined in a similar way as

$$R^2 = 1 - \frac{\overline{Q}(\hat{g}_h)}{\overline{Q}(\hat{g}_{\overline{h}_0})},$$

for a strategy h.

We illustrate the difference between our prediction procedure and traditional goodness of fit by considering two different estimators of stock returns. In particular, we consider nonparametric estimators based on the full data

$$W_t = \left(S_t, d_t, I_t, r_t\right) \tag{6}$$

and on the simple subset

$$\overline{W}_t = d_t = D_t \big/ P_t. \tag{7}$$

See Appendix 1 for the mathematical definition of the local linear kernel estimators used. The quality of fit of these two models are given in Figure 1, where the estimators of the regression function $g(\bullet)$ are used to fit next years stock return. We talk about fitting rather than predicting, because the graphs are based on an in-sample approach, where $g(\bullet)$ is estimated from the same stock returns as we fit. From the graphs it is quite clear that one can fit our data set pretty well from the full four dimensional time series, whereas the one dimensional time series consisting of the dividend yield alone fits the data much less. Based on a traditional goodness of fit measures as the R^2 value, see for example Kvålseth (1985), one would clearly prefer the four dimensional covariate to predict stock returns to the simpler one dimensional time series based on dividend yield alone. As a matter of fact, a goodness of fit type of procedure will always have a tendency to chose the most complicated model. Kvålseth (1985) is aware that goodness of fit has this problem and suggests a correction using degrees of freedom. In nonparametric regression, it is, however, unclear what degrees of freedom is. Hastie and Tibshirani (1990) give ad hoc suggestions that seemed to work in their simulations for testing using splines but did not work well in other contexts, see e.g. Sperlich, Linton and Härdle (1999) or Müller (1998). There are certainly other selection criteria for (usually particular) nonparametric models as e.g. the improved Akaike criterion of Hurvich, Simonoff and Tsai (1998). Inside their formulae appear also expressions we might interpret as approximations of the degrees of freedom but it is neither clear whether this criterion can be applied

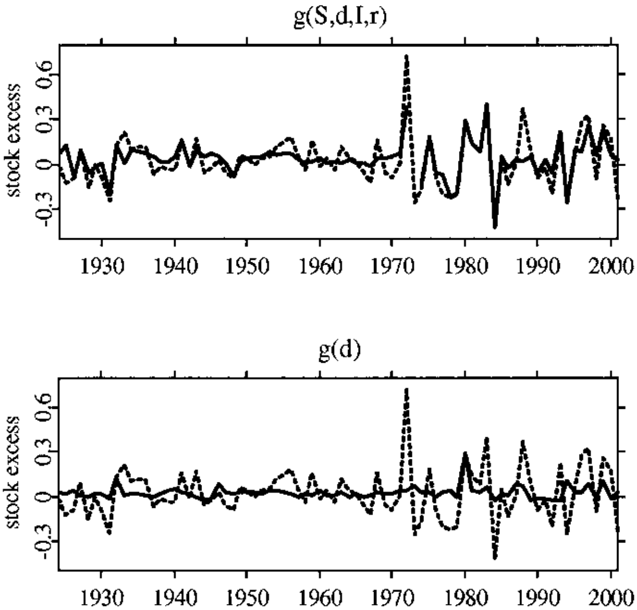


FIGURE 1: Fit of S_t using nonparametric regression (local linear) estimators. Upper curve: based on lagged stock excess, dividend, inflation and short term interest rate, i.e. W_t from (6). Lower curve: based on dividend by price ration, i.e. $\overline{W}_t$ from (7).

to all the smoothers we use here nor how to interpret the value the criterion takes. We therefore prefer our more straightforward prediction criterion. We will see later that, if we base our conclusions on R_V^2 values, then sometimes we end up with the opposite conclusion of that one we arrived at using the traditional goodness of fit R^2. Namely, we will see that the dividend yield can indeed help a bit to predict stock returns, while the nonparametric estimator of the full data set is too noisy to be useful for prediction and giving a strongly negative R_V^2 value.

We conclude by pointing out, that even when allowing for any kind of flexible model, it takes a selective choice of the most important explanatory variables to beat even the simple cross validated mean $\hat{Q}(g_{h_0})$ in practical prediction. Indeed, complexity is one of the worst enemies of a good prediction.

5. ESTIMATING AND EVALUATING THE POWER OF PREDICTION

In this section we enter the methodological question of finding a good estimator of prediction power, first we follow Campbell, Lo and Mackinlay (1997, p. 269) and calculate R^2 for different prediction horizons. As mentioned above we will first concentrate on the period up to 1996. In the first table, Table 1, we consider two versions of the regression

$$Y_t = S_{t+1} + \ldots + S_{t+T} = \alpha + \beta\delta_t + \epsilon_{t+T}, \tag{8}$$

where $\delta_t = d_t$ (left-hand columns) or $\delta_t = \ln(d_t)$ (right-hand columns).

TABLE 1

CLASSICAL R^2 VALUES FOR T-YEAR EXCESS STOCK RETURNS ON δ_t, MODEL (8)

horizon	$\delta_t = d_t$		$\delta_t = \ln(d_t)$	
T	***1923-1996***	***1949-1996***	***1923-1996***	***1949-1996***
1	3.8%	7.3%	3.2%	5.9%
2	8.8%	14.9%	6.6%	11.5%
3	13.0%	21.1%	10.5%	17.1%
4	17.5%	25.8%	14.2%	21.0%
5	18.7%	24.2%	15.7%	20.6%

We see that for the linear model the R^2 values are increasing with the time horizon, the same conclusion as Campbell, Lo and Mackinlay (1997, p. 269) arrived at for their American data set. This might imply that prediction over longer horizons is more easy than prediction over short horizons. In the next table, Table 2, we investigate this using our validated criterion for the linear model based on the dividend yield.

TABLE 2

R_V^2 VALUES FOR T-YEAR EXCESS STOCK RETURNS ON δ_t, MODEL (8)

horizon	$\delta_t = d_t$		$\delta_t = \ln(d_t)$	
T	***1923-1996***	***1949-1996***	***1923-1996***	***1949-1996***
1	–0.2%	1.4%	–1.1%	–0.3%
2	4.9%	8.2%	2.2%	3.0%
3	7.8%	14.2%	4.6%	7.7%
4	10.3%	16.0%	7.4%	9.4%
5	10.3%	9.5%	6.5%	0.5%
6	6.9%	–4.6%	5.2%	–19.5%

We see that, while the quality of prediction is smaller than the R^2 values considered above might suggest, the validated R_V^2 does indeed indicate predictive power of the dividend yield. The period with strongest predictive power seems to be a four year time period with an improved quality of prediction of around 10%. This corresponds to a 10% decrease of the variance of the error term involved in the prediction.

In Table 1 and 2 we see, that a linear regression based on the dividend yield itself instead of the logarithm to the dividend yield gives a better power of prediction. Notice that negative values do not occur in Table 1, but they are present in Table 2. Negative values can not occur with the classical R^2 measure. The classical R^2 measure always favors a more complicated model than the trivial one with a constant mean. The R_V^2 gives negative values in those cases where the prediction model is estimated to perform worse than the trivial model. Since it is indeed very difficult to predict stock markets, the R_V^2 measure will be negative for most attempted prediction models. The surprise here is perhaps that it actually does seem that the dividend yield has predictive power for most of the considered horizons.

Before we get to the nonparametric estimation, we first consider the period 1948-1996 once again, but this time estimation is performed using all the data from 1922 to 1996. However, only the time interval 1948-1996 is used while evaluating the predictive power of the filter. The results are presented in Table 3.

TABLE 3

PREDICTABILITY FOR 1948-1996 OF T-YEAR EXCESS STOCK RETURNS ON d_t, RESPECTIVELY ON $\ln(d_t)$, SIMPLE MODEL (8), EVALUATED WITH THE R_V^2 WHEN USING ALL DATA FROM 1922-1996 FOR PREDICTION

T	d_t	$\ln(d_t)$
1	3.3%	1.8%
2	10.8%	7.3%
3	16.7%	12.4%
4	19.8%	15.6%
5	18.0%	13.2%
6	17.8%	14.9%

The conclusion is that using the entire data set is better while predicting the post war period than just using the post war data itself. It seems that the increased estimation accuracy obtained by using more data outweigh the advantage of only using post war data while estimating the post war period. The argument for the latter methodology is off course that the post war period might be different in nature from the pre war period.

Finally we consider the power of prediction by choosing the functional relationship between the dividend-price ratio and the return by a nonparametric kernel estimator. Specifically, we use local linear kernel estimation what means that in the limit (with bandwidth $h \to \infty$) the function is linear, and thus, in the limit, coincides with the linear regression, see Appendix 1 and Appendix 2 for details. The bandwidth or smoothing parameter has been chosen such that it maximizes the R_V^2. Since this functional relationship can be arbitrary, the above discussion on using the raw dividend price ratio or taking the logarithm is irrelevant. We get the results drawn in Table 4.

TABLE 4

PREDICTABILITY OF T-YEAR EXCESS STOCK RETURNS ON USING NONPARAMETRIC MODELS AND MEASURED IN R_V^2. EXPLANATORY VARIABLE WAS DIVIDEND YIELD, d_t

T	***1923-1996***	***1949-1996***
1	–0.2%	3.3%
2	6.1%	11.5%
3	9.0%	20.7%
4	12.9%	24.5%
5	11.6%	21.7%
6	6.9%	17.8%

Again, when considering the post war period, then data from the entire period is used to fit the nonparametric functional relationship, and the evaluation of the quality of the fit is, however, based on the data in the post war period. While the nonparametric power of prediction for the period 1922-1996 is already slightly better than the strictly linear power of prediction, we see a clear improvement of prediction power for the nonparametric method when considering the period 1948-1996.

For a graphical visualization of the impact of the dividend-price ratio at excess stock returns, see Figure 2 and Figure 3 for respectively the one-year horizon and the four year horizon versions of the prediction of excess stock returns based on the dividend-price ratio. Both the parametric and nonparametric versions are shown. The graphs clearly indicate the impact of the dividend yield on future returns and we also see, that the Danish level of the dividend-price ratio around 1.5% (in 2000) was so low, that according our predictive filter it was a dangerous time to invest in stocks and we did not expect the average excess return on stocks to match this danger. As a matter of fact our model predicted excess returns in the year 2001 to have an average value

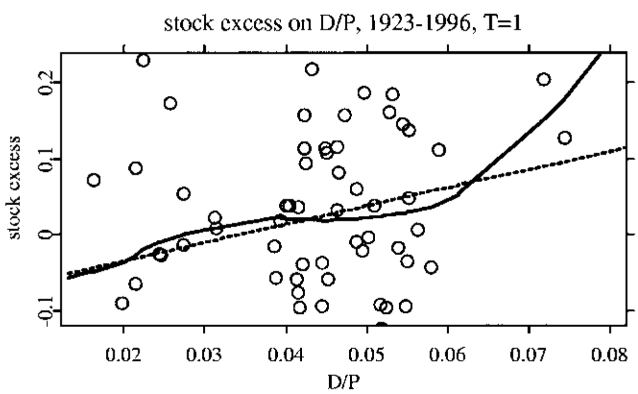


FIGURE 2: Optimal parametric (dashed) and nonparametric (solid) regression fit of stock excess on D/P with real data points.

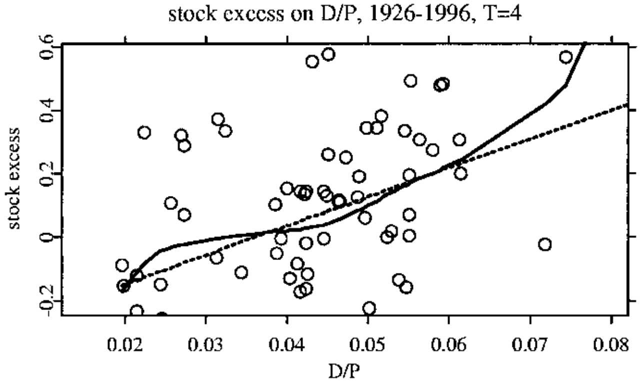


FIGURE 3: Optimal parametric (dashed) and nonparametric (solid) regression fit of stock excess on D/P and real data points.

below zero. So, at that time the extra risk inherent in investments in stocks was not followed by a corresponding extra return on stocks. As a consequence our advice to Danish long term investors was (and is) not to increase their percentage of stocks in their portfolio right now.[6]

6. LOOKING FOR THE BEST PREDICTION MODEL

In this section we investigate the potential advantages that we can obtain by including other variables than just dividend divided by price in our prediction. Due to the complexity of the study of the section, we have chosen to restrict our investigation to a time horizon of one year. Based on the considerations given in Section 2, we have chosen to consider a time series regression problem of the following form:

$$S_t = g(S_{t-1}, d_{t-1}, I_{t-1}, r_{t-1}) + \epsilon_t \tag{9}$$

using the data described in Section 3. The full four-dimensional model corresponds to estimate the function $g(\bullet)$ without any parametric assumptions nor assumptions of structure such as additivity or multiplicativity. This model is most often too complex for both to visualize and/or to predict well. The lack of prediction is due to the error of estimation rather than that the model is insufficient. Therefore we suggest some structure on $g(\bullet)$ to predict well. We have chosen to consider additive models such as

$$g(S, d, I, r) = c + g_1(S) + g_2(d) + g_3(I) + g_4(r), \tag{10}$$

[6] This remark is from a version of this paper dated february 2001.

compare also Appendix 2, especially for estimation.

Furthermore we consider both the situation where the entering g_i's are nonparametric and the situation where all the entering g_i's are parametric and follow a linear model. In our study we consider three types of models with all combinations of subsets of $(S_{t-1}, d_{t-1}, I_{t-1}, r_{t-1})$, namely

- Linear models
- Nonparametric additive models
- Fully nonparametric models

Note that we always applied local linear kernel smoothers applying the bandwidth h that maximizes the R_V^2, see Appendix 3. The more complex the model is, the bigger the estimation error will be but the smaller the modeling error will be. To be able to choose among the entering models, we use the validated R_V^2 defined in Section 5. All in all, we have 26 models to consider, namely 15 full models (that include automatically the 15 linear models) and 11 nonparametric additive models (leaving out the one-dimensional models that we counted among the full ones). As mentioned and explained in the appendices we always looked for the optimal bandwidths in the nonparametric procedures using Cross Validation, i.e. maximizing our R_V^2.

Some findings of the estimation respective model structure are the following.

Though the multidimensional nonparametric additive model reaches a positive R_V^2 for some of the considered models, the corresponding full model always did better. This is a clear indicator for having here a more complex structure than additivity. This is not surprising when we consider the complicated relationship between these variables as described in Section 2. From our calculated R_V^2 values we also concluded that the only linear model that does better than the simple constant is the linear model based on the dividend divided by price for the period 1948-2001 as described in the sections before. However, best among all estimators is the fully nonparametric two-dimensional model based on dividend divided by price and lagged excess stock return. This two-dimensional model has a R_V^2 value of 5.5% for the period 1923-1996 and 9.1% for the period 1948-1996. This is much better than the (negative) values of the R_V^2 obtained in Section 5.

Once again have a look on the relation excess returns to dividend by price. In Figure 5 we see the 3-D plot of the two-dimensional predictive filter based on the dividend-price ratio and the lagged excess return of stocks. Further, in Figure 4 we see three slices from this filter plotting the dependency on the dividend yield for three fixed values of excess returns: –25%, 0.7% and 29.5% corresponding to the lower 5% quantile, the median and the upper 95% quantile. We see a clear tendency of the excess stock return to be increasing with the dividend-price ratio. For small dividend yields (below the historical mean of about 4%) the stock return is decreasing with last years excess return. For higher dividend yields (above the historical mean of about 4%) the stock return is increasing with last years excess return. While the intuition of the dependency on the dividend yield is straightforward, it is less straightforward to understand the relationship between last years stock return and current

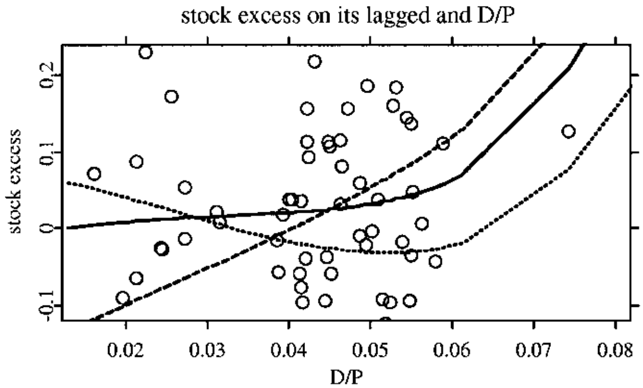


FIGURE 4: Nonparametric regression fits of stock excess on *D*/*P* and stock excess lagged fixed at –25% (dotted, starting above zero), at 1% (solid), and at 30% (dashed) for the period 1923-1996.

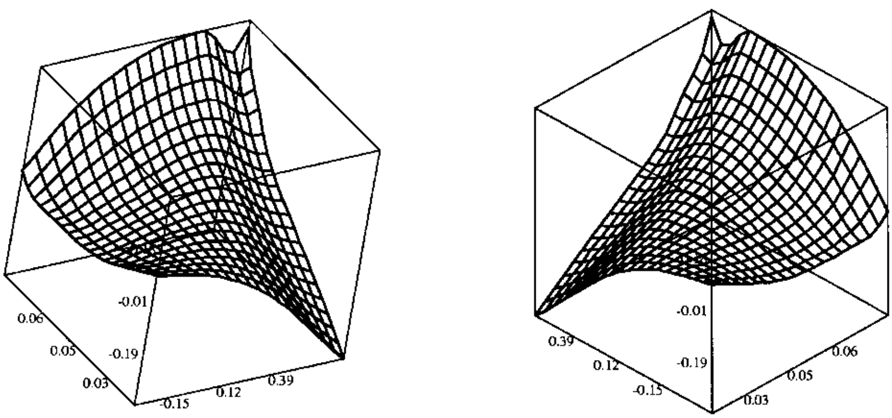


FIGURE 5: Two different views on the nonparametric regression fit of stock excess on *D*/*P* and stock excess lagged for the period 1923-1996.

stock return. However, this graph does show that Danish investors should have kept away for new investments in stocks in 2001, since they were just about to finish a magnificent year with a general Danish excess return on stocks above 30% resulting in a historical low dividend-price ratio of around 1.5%. A more detailed picture leading to the same type of conclusions can be found in the three dimensional plot in Figure 5.

7. EMPIRICAL RESULTS FOR 1922-2001

The statistical evidence based on the updated data set does not change the estimated curves and variables very much. However, the estimated predictive power of the updated data set leaves a much less optimistic impression of the possibility of predicting stock returns than the corresponding results based on the original data set. It is perhaps not surprising for followers of the stock market that the last five years, 1997-2001, have been unusual. Based on the updated data all considered linear models break down – they simply do not predict. This is in contradiction to classical studies like Fama and French (1988), Wilkie (1993) and others and need serious consideration in further work.

However, our main statements still hold when we extend the data set up to 2001. The relationship between the classical R^2 and the validated version, R_V^2, still play the same role and the full, not additive, nonparametric models still have predictive power.

The optimal R_V^2 is reached for $T = 4$ (time horizon) when only including $d = D/P$. Looking at $T = 1$, the best model for the period 1922-2001 now only uses last year stock return. The model based on both last years stock return and the dividend yield does, however, predicts almost as well. Their predictive powers are respectively, 1.0% and 0.9%. For the time period 1948-2001, we get a relatively impressive R_V^2 of 3.1% while including dividend yield and last years stock return. Optimal bandwidths are between $4\sigma_W$ and $4.8\sigma_W$, so the models are by far not linear. Here, σ_W is the vector of the standard deviations of the different regressors.

Let us first consider Table 5, the corresponding one to Table 1 in Section 5 where we looked at the classical R^2 values for T-year excess stock returns on δ_t, model (8). As one can see clearly, the model fits terribly bad compared to the results obtained for the time period 1922-1996. All R_V^2 values are negative for the linear and log-linear models. We therefore skip here the analogs for Tables 2 and 3 from Section 5.

TABLE 5

CLASSICAL R^2 VALUES FOR T-YEAR EXCESS STOCK RETURNS ON δ_t, MODEL (8)

horizon	$\delta_t = d_t$		$\delta_t = \ln(d_t)$	
T	***1923-2001***	***1949-2001***	***1923-2001***	***1949-2001***
1	1.6%	3.3%	0.8%	1.8%
2	1.8%	3.3%	0.3%	0.9%
3	1.8%	2.9%	0.2%	0.6%
4	2.6%	3.0%	0.2%	0.2%
5	1.8%	1.2%	0.1%	0.0%
5	1.2%	1.1%	0.0%	0.0%

Table 6 is the corresponding one to Table 4 in Section 5, i.e. we have drawn for the different time horizons T the R_V^2 obtained for the nonparametric model with d_t being the only regressor. The highest R_V^2 is given for $T = 4$ years, but now only with 6.7% for the whole, respectively 12.5% for the post war period.

TABLE 6

PREDICTABILITY OF T-YEAR EXCESS STOCK RETURNS USING NONPARAMETRIC MODELS AND MEASURED IN R_V^2. EXPLANATORY VARIABLE WAS DIVIDEND YIELD, d_t

T	***1923-2001***	***1949-2001***
1	–1.9%	–0.3%
2	0.0%	2.1%
3	–1.6%	3.5%
4	6.7%	12.5%
5	–10.7%	–5.9%
6	–20.4%	–11.9%

Finally, the Figures 2 to 4 showing the impact of dividend by price ratio $d_t = D_t / P_t$ on the real excess stock returns, stay quite the same when we include the years 1997-2001 into the estimation.

8. CONCLUSIONS

There are mainly three points we make. We first look for a reasonable measure for prediction power (the R_V^2). Second, we use this measure to evaluate the power of prediction of classical as well as more flexible methods. It turns out that the use of nonparametrics methodology and the inclusion of last years stock return significantly improve the level of prediction. Third, fixing the time horizon ($T = 1$) and using flexible methods, we ask for the best prediction model. Finally we illustrate how this can help us for a better understanding of the considered process (discussion of the Figures).

9. APPENDIX

Appendix 1. Local linear kernel regression

In this appendix we give a brief insight into the algorithms of nonparametric flexible function regression. In particular we explain the local linear smoothing. The basic idea is to construct an estimator that lays a smooth surface (or hyperplane), e.g. in the one dimensional case a smooth line, into the point cloud that presents its functional form. The smoothness of that surface can be (pre-) determined by choosing a respectively large *smoothing parameter h*, called bandwidth. Actually, often this parameter can also be data driven, see Appendix 3.

First, it is important to understand that this estimator works locally, e.g. we estimate the wanted function, the hyperplane, at each point we are interested in separately. This is, using the notation $E[Y|W] = g(w)$, $Y \in \mathbb{R}, W, w \in \mathbb{R}^d$ with $g(\cdot) : \mathbb{R}^d \to \mathbb{R}$, an unknown smooth function we estimate

$g(w_0)$ for some point $w_0 \in \mathbb{R}^d$. Having $(W_i, Y_i)_{i=1}^n$ observed, this is done by local least squares:

$$\begin{pmatrix} \hat{g}(x_0) \\ \widehat{\nabla g}(x_0) \end{pmatrix} = \underset{a_0, a_1}{argmin} \sum_{i=1}^{n} \left\{Y_i - a_0 - a_1^T (W_i - w_0)\right\}^2 K_h(W_i - w_0), \tag{11}$$

$a_0 \in \mathbb{R}, a_1 \in \mathbb{R}^d$ and $\nabla g(\cdot)$ being the gradient of $g(\cdot)$. Further, $K_h(v) = \prod_{j=1}^{d} \frac{1}{h} K\left(\frac{v_j}{h}\right)$ is a $\mathbb{R}^d \to \mathbb{R}$ weight function. In our calculations we chose $K(v) = \frac{15}{16}(1-u^2)^2 \mathbb{1}\{|u| \leq 1\}$. So we used a weighted least squares estimator for linear regression that becomes a local estimator due to the weights K_h giving a lot of weight to points (W_i, Y_i) where W_i is close to w_0 but zero weights to points far from w_0.

Here, in the weighting function comes the smoothing parameter h in: the larger h and consequently the environment with positive weighting, the smoother gets the resulting hyperplane, i.e. $h \to \infty$ gives a linear function whereas $h = 0$ yields interpolation of the Y_i's. Consistency, asymptotic theory and properties are well known and studied for the multivariate case in Ruppert and Wand (1994), for a general introduction see Fan and Gijbels (1996).

An often discussed question is how to choose bandwidth h in practice. As we are concerned about prediction, we take that bandwidth that is minimizing the "out of sample" prediction error using the Cross Validation measure, see Appendix 3.

Appendix 2. Local linear additive regression

We speak of a nonparametric additive model if $g(w)$, $w \in \mathbb{R}^d$ is of the form

$$g(w) = c + g_1(w_1) + g_2(w_2) + \cdots + g_d(w_d), \quad \text{with } c = E[Y], \tag{12}$$

and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ unknown smooth functions with $E[g_j(W_j)] = 0$ for identification. This is the natural extension of the classical linear regression model relaxing the restriction of linear impacts to arbitrary (but smooth) ones. Several procedures are known in the literature, see Sperlich (1998). In this article we focus only on the backfitting by Hastie, Tibshirani (1990). If $g(w)$ is really of additive form, then, under some regularity conditions, this gives us consistent estimators; if not, it tries to estimate the additive model that fits our data best. This is done by iteration: start with some initials $\hat{g}_j^{|0|}(\cdot)$ and $\hat{c} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then regress $Y - \hat{c} - \sum_{j \neq k}^{d} \hat{g}^{[r-1]}(W_j)$ on W_k to get $\hat{g}_k^{[r]}$ until convergence.

For the regression we applied the local linear kernel estimator. Again, bandwidths can be chosen using Cross Validation, presented in Appendix 3.

Appendix 3. Cross Validation

A typical question of interest, not only in prediction problems, is how to evaluate the different models. This concerns the model or variable selection as well as the bandwidth choice. In general, a natural way to evaluate an estimator is to look on the mean squared error or the expected squared difference between estimate and observation Y: $E[\{Y-\hat{g}(W)\}^2]$. This certainly has to be estimated. Additionally, as we speak about prediction, we would like to know how well the estimator works outside the considered sample. Both aspects are taken into account in the so called Cross Validation (CV) values, defined as

$$\text{CV} - \text{value} = \frac{1}{n}\sum_{l=1}^{n} \left\{y_l - \hat{g}^{(l)}(w_l)\right\}^2, \tag{13}$$

where $\hat{g}^{(l)}(w_l)$ is the considered estimator evaluated at point w_l but determined without observation (w_l, y_l). This CV-value is an approximation for the mean squared error (also for prediction) and a quite common used validation measure in nonparametric regression. For time series context and more references see e.g. Gyöfri, Härdle, Sarda, and Vieu (1990).

Remark: It is important to eliminate always all information that is aimed to predict from the estimation of $g(\bullet)$. So, if we predict the increase of assets over a period of 4 years, the estimator $\hat{g}^{(l)}$ is calculated not only without the l^{th} observation but also without the three years before and after year l.

How can it be used for bandwidth or model selection?

We give an example for bandwidth selection: write $\hat{g}$ as a function of the bandwidth ($\hat{g}_h$) and look for that h that minimizes

$$\text{CV}(h) = \frac{1}{n}\sum_{l=1}^{n} \left\{y_l - \hat{g}_h^{(l)}(w_l)\right\}^2.$$

This has been shown to give the optimal bandwidth in nonparametric regression, we again refer to Gyöfri et al. (1990).

Note finally, that minimizing the CV-value is equivalent to maximizing the R_V^2. So CV is directly used to find both: the optimal h for prediction and the optimal model for prediction.

References

Barber, B.M. and Terrance, O. (2000) Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance* **55**, 773-806.

Cairns, A.J.G. (2000) A discussion of parameter and model uncertainty in insurance. Insurance: *Mathematics and Economics* **27**, 313-330.

Campbell, J.Y. and Shiller, R.J. (1988) The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* **1**, 195-228.

Campbell, J.Y., Lo, A.W. and MacKinlay, A.C. (1997) *The Econometrics of financial markets*. Princeton University Press, Princeton, New Jersey.

Dimson, E., Marsh, P. and Staunton, M. (2002) *Triumph of the Optimists.* Princeton University Press.

ENGSTED, T. and TANGGÅRD, C. (2000) The Danish stock and bond markets: Comovement, return predictability and variance decomposition. *Journal of Empirical Finance* **8**, 243-271.

EUN, C.S. and RESNICK, B.G. (1988) Exchange rate uncertainty, forward contracts, and international portfolio selection. *Journal of Financ*e **43**, 197-215.

EUN, C.S. and RESNICK, B.G. (1994) Exchange rate uncertainty, forward contracts, and international portfolio selection. *Management Science* **40**, 140-161.

FAMA, E. and FRENCH, K. (1988) Dividend yields and expected stock returns. *Journal of Financial Economics* **22**, 3-25.

FAN, J. and GIJBELS, I. (1996) *Local polynomial regression*. Chapman and Hall, London.

GOETZMANN, W.N. and JORION, P. (1993) Testing the predictive power of dividend yields. *Journal of Finance* **48**, 663-679.

GORDON, M. and SHAPIRO, P. (1956) Capital Equilibrium analysis: The required rate of profit. *Management Science* **3**, 102-110.

GYÖFRI, L., HÄRDLE, W., SARDA, P. and VIEU, Ph. (1990) *Nonparametric Curve Estimation from Time Series*. Lecture Notes in Statistics. Heidelberg: Springer-Verlag.

HANSEN, S. (1974) *Økonomisk vækst i Danmark [Economic growth in Denmark]*, Akademisk Forlag, Copenhagen, Denmark.

HARRIS, G.R. (1999) Marcov chain Monte Carlo estimation of regime switching vector autoregressions. *ASTIN Bulletin* **29**, 47-79.

HASTIE, T.J. and TIBSHIRANI, R.J. (1990) *Generalized Additive Models*. Chapman and Hall.

HODRICK, R.J. (1992) Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies* **5**, 357-386.

HURVICH, C.M., SIMONOFF, J.S. and TSAI, C.-L. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, B* **60**, 271-293.

JAGANNATHAN, R. and KOCHERLAKOTA, N.R. (1996) Why should older people invest less in stocks than younger people. *Federal Reserve Bank of Minneapolis. Quaterly Review, Summer 1996*. 11-20.

KVÅLSETH, T. (1985) Cautionary note about R^2. *The American Statistician* **39**, 279-285.

LUND, J. and ENGSTED, T. (1996) GMM and present value tests of the C-CAPM: Evidence from the Danish, German, Swedish, and UK stock markets. *Journal of International Money and Finance* **15**, 497-521.

MÜLLER M. (1998) Computer-assisted Statistics Teaching in Network Environments. *COMPSTAT '98 Proceedings*, Bristol, UK.

RICHARDSON, M. and STOCK, J.H. (1989) Drawing inferences from statistics based on multi-year asset returns. *Journal of Financial Economics* **25**, 323-348.

RUPPERT, D. and WAND, M.P. (1994) Multivariate locally weighted least squares regression. *Annals of Statistics* **22**, 1346-1370.

SIEGEL, J.J. (1998) *Stocks for the long run*, 2nd ed. New York: McGraw-Hill.

SHILLER, R.J. (1981) Do stock prices move too much to be justified by subsequent changes in dividends. *American Economic Review* **71**, 421-436.

SHILLER, R.J. (2000) *Irrational exuberrance*. Princeton University Press, Princeton, New Jersey.

SPERLICH, S. (1998) *Additive Modelling and testing Model Specification*. Shaker Verlag, Aachen.

SPERLICH, S., LINTON, O., and HÄRDLE, W. (1999) Integration and backfitting methods in additive models-Finite sample properties and comparison. *Test* **8**, 419-458.

WILKIE, A.D. (1993) Can dividend yields predict share price changes? *Transactions of the 3rd International AFIR Colloquium, Rome*, **1**, 335-347.

WILKIE, A.D. (1995) More on a stochastic asset model for actuarial use. *British Actuarial Journal* **5**, 777-964.

WILLIAMS, J.B. (1938) *The Theory of Investment Value*. Harvard University Press, Cambridge.

SETTING A BONUS-MALUS SCALE IN THE PRESENCE OF OTHER RATING FACTORS: TAYLOR'S WORK REVISITED

BY

SANDRA PITREBOIS, MICHEL DENUIT AND JEAN-FRANÇOIS WALHIN

ABSTRACT

In this paper, we propose an analytic analogue to the simulation procedure described in Taylor (1997). We apply the formulas to a Belgian data set and discuss the interaction between a priori and a posteriori ratemakings.

KEYWORDS AND PHRASES

Bonus-Malus system, Markov chains, *a priori* ratemaking, experience rating

1. INTRODUCTION AND MOTIVATION

One of the main tasks of the actuary is to design a tariff structure that will fairly distribute the burden of claims among policyholders. To this end, he often has to partition all policies into homogeneous classes with all policyholders belonging to the same class paying the same premium. The classification variables introduced to partition risks into cells are called *a priori* variables (as their values can be determined before the policyholder starts to drive). In motor third-party liability (MTPL, in short) insurance, they include age, gender and occupation of the policyholders, type and use of their car, place where they live and sometimes even number of cars in the household or marital status. It is convenient to achieve *a priori* classification by resorting to generalized linear models (e.g. Poisson regression).

However, many important factors cannot be taken into account at this stage; think for instance of swiftness of reflexes, aggressiveness behind the wheel or knowledge of the highway code. Consequently, risk classes are still quite heterogeneous despite the use of many *a priori* variables. But it is reasonable to believe that these hidden factors are revealed by the number of claims reported by the policyholders over the successive insurance periods. Hence the amount of premium is adjusted each year on the basis of the individual claims experience in order to restore fairness among policyholders.

Rating systems penalizing insureds responsible for one or more accidents by premium surcharges (or *maluses*), and rewarding claim-free policyholders by awarding them discounts (or *bonuses*) are now in force in many developed countries. This *a posteriori* ratemaking is a very efficient way of classifying policyholders according to their risk. Besides encouraging policyholders to drive carefully (i.e. counteracting moral hazard), they aim to better assess individual risks. Such systems are called no-claim discounts, experience rating, merit rating, or Bonus-Malus systems (BMS, in short). We will adopt here the latter terminology. For a thorough presentation of the techniques relating to BMS, see Lemaire (1995).

When a BMS is in force, the amount of premium paid by the policyholder depends on the rating factors of the current period but also on claim history. In practice, a BMS consists of a finite number of levels, each with its own relative premium. New policyholders have access to a specified class. After each year, the policy moves up or down according to transition rules and to the number of claims at fault. The premium charged to a policyholder is obtained by applying the relative premium associated to his current level in the BMS to a base premium depending on his observable characteristics incorporated into the price list.

The problem addressed in this paper is the determination of the relative premiums attached to each of the levels of the BM scale when *a priori* classification is used by the company. The severity of the *a posteriori* corrections must depend on the extent to which amounts of premiums vary according to observable characteristics of policyholders. The key idea is that both *a priori* classification and *a posteriori* corrections aim to create tariff cells as homogeneous as possible. The residual heterogeneity inside each of these cells being smaller for insurers incorporating more variables in their *a priori* ratemaking, the *a posteriori* corrections must be softer for those insurers.

This paper is not conceptually innovating. All the ideas are contained in the seminal work by Taylor (1997). Our only contribution is to show how it is possible to avoid simulations by providing analytical formulas for the relative premiums attached to each level of the BM scale.

Our work is organized as follows. In Section 2, we briefly present the modelling used to compute pure premiums. Section 3 describes BM scales and their representation as Markov chains. Section 4 explains how to determine the relative premiums when *a priori* classification is in force or not. Section 5 describes several numerical illustrations. In Section 6, we show that it is possible to apply different *a posteriori* corrections according to *a priori* characteristics. The final Section 7 discusses some possible improvements and concludes.

2. Credibility updating formulas

Let $N_{it}, t = 1, 2, \ldots$, represent the number of claims incurred by policyholder i in period t. The annual expected claim frequency for policy i in year t is $\lambda_{it} = \mathbb{E}[N_{it}]$. It is expressed as the exponential transform of some predictor involving the characteristics of policyholder i in period t. Of course, all the risk factors cannot be taken into account at this stage.

Risk classes remain heterogeneous despite the use of many *a priori* risk characteristics. This residual heterogeneity can be represented by a random effect Θ_i superposed to the annual expected claim frequency. Specifically, given $\Theta_i = \theta$ the annual numbers of claims N_{it} are assumed to be independent and to conform to a Poisson distribution with mean $\lambda_{it}\theta$, i.e.

$$\Pr\left[N_{it}=k \mid \Theta_i=\theta\right]=\exp\left(-\theta\lambda_{it}\right)\frac{\left(\lambda_{it}\theta\right)^k}{k!},\quad k\in\mathbb{N}.$$

Moreover, all the Θ_i's are assumed to be independent and to follow a standard Gamma distribution with probability density function

$$u(\theta)=\frac{1}{\Gamma(a)}a^a\theta^{a-1}\exp(-a\theta),\quad \theta\in\mathbb{R}^+. \tag{2.1}$$

The latter is often referred to as the structure function of the portfolio. Since $\mathbb{E}[\Theta_i]=1$ we have that $\mathbb{E}[N_{it}]=\lambda_{it}$; λ_{it} is the expected claim number for a policyholder for which no information about past claims is available.

The premium is then adjusted over time with the help of credibility techniques. We assume that each policyholder has an unknown expected claim frequency ϑ_i, constant over time. Following the seminal work of Dionne and Vanasse (1989), the company approaches this unknown value with annual predictions of the form $\widehat{\vartheta}_{i1}=\lambda_{i1}$ and for $t\geq 2$,

$$\widehat{\vartheta}_{it}=\mathbb{E}\left[N_{it}\mid N_{i1},\ldots,N_{it-1}\right]=\lambda_{it}\frac{a+\sum_{\tau=1}^{t-1}N_{i\tau}}{a+\sum_{\tau=1}^{t-1}\lambda_{i\tau}}. \tag{2.2}$$

The latter Bayesian credibility estimator cannot be enforced in practice for MTPL, essentially due to commercial reasons and legal constraints. Instead, companies resort to BM scales, that may be considered as simplified versions of credibility theory formulas. Those are presented in the next section.

3. Markov Models for Practical BMS

3.1. BMS as Markov chains

In practice, insurance companies often resort to BM scales similar to those in Tables 5.4-5.6-5.8 and not on credibility coefficients like those of (2.2). Such scales possess a number of levels, $s+1$ say, numbered from 0 to s. A specified level is assigned to a new driver (often according to the use of the vehicle). Each claim free year is rewarded by a bonus point (i.e. the driver goes one level down). Claims are penalized by malus points (i.e. the driver goes up a certain number of levels each time he files a claim). We assume that the penalty is a given number of classes per claim. This facilitates the mathematical treatment of the problem but more general systems could also be considered. After sufficiently many claim-free years, the driver enters level 0 where he enjoys the maximal bonus.

In commercial BMS, the knowledge of the present level and of the number of claims of the present year suffice to determine the next level. This ensures that the BMS may be represented by a Markov chain: the future (the class for year $t+1$) depends on the present (the class for year t and the number of accidents reported during year t) and not on the past (the complete claim history and the levels occupied during years $1, 2, \ldots, t-1$). Sometimes, fictitious classes have to be introduced in order to meet this memoryless property. Indeed, in some BMS, policyholders occupying high levels are sent to the starting class after a few claimless years.

The relativity associated to level ℓ is denoted as r_ℓ; the meaning is that an insured occupying that level pays an amount of premium equals to r_ℓ% of the *a priori* premium determined on the basis of his observable characteristics.

3.2. Transient distributions

Let $p_{\ell_1\ell_2}(\vartheta)$ be the probability of moving from level ℓ_1 to level ℓ_2 for a policyholder with mean frequency ϑ. Further, $M(\vartheta)$ is the one-step transition matrix, i.e. $M(\vartheta) = \{p_{\ell_1\ell_2}(\vartheta)\}$, $\ell_1, \ell_2 = 0, 1, \ldots, s$. Taking the νth power of $M(\vartheta)$ yields the ν-step transition matrix whose element $(\ell_1\ell_2)$, denoted as $p^{(\nu)}_{\ell_1\ell_2}(\vartheta)$, is the probability of moving from level ℓ_1 to level ℓ_2 in ν transitions.

3.3. Stationary distribution

All BMS in practical use have a "best" level, with the property that a policy in that level remains in the same level after a claim-free period. In the following, we restrict attention to such non-periodic bonus rules. The transition matrix $\boldsymbol{M}(\vartheta)$ associated to such a BMS is regular, i.e. there exists some integer $\xi_0 \geq 1$ such that all entries of $\{\boldsymbol{M}(\vartheta)\}^{\xi_0}$ are strictly positive. Consequently, the Markov chain describing the trajectory of a policyholder with expected claim frequency ϑ accross the levels is ergodic and thus possesses a stationary distribution $\boldsymbol{\pi}(\vartheta) = (\pi_0(\vartheta), \pi_1(\vartheta), \ldots, \pi_s(\vartheta))^t$; $\pi_\ell(\vartheta)$ is the stationary probability for a policyholder with mean frequency ϑ to be in level ℓ i.e.

$$\pi_{\ell_2}(\vartheta) = \lim_{\nu \to +\infty} p^{(\nu)}_{\ell_1\ell_2}(\vartheta).$$

Note that $\boldsymbol{\pi}(\vartheta)$ does not depend on the starting class.

Let us now recall how to compute the $\pi_\ell(\vartheta)$'s. The vector $\boldsymbol{\pi}(\vartheta)$ is the solution of the system

$$\begin{cases} \boldsymbol{\pi}^t(\vartheta) = \boldsymbol{\pi}^t(\vartheta)\boldsymbol{M}(\vartheta), \\ \boldsymbol{\pi}^t(\vartheta)\boldsymbol{e} = 1 \end{cases}$$

where $\boldsymbol{e}$ is a column vector of 1's. Let $\boldsymbol{E}$ be the $(s+1) \times (s+1)$ matrix all of whose entries are 1, i.e. consisting of $s+1$ column vectors $\boldsymbol{e}$. Then, it can be shown that

$$\boldsymbol{\pi}^t(\vartheta) = \boldsymbol{e}^t(\mathbf{I} - \boldsymbol{M}(\vartheta) + \boldsymbol{E})^{-1},$$

which provides a direct method to get $\boldsymbol{\pi}(\vartheta)$. For a derivation of the latter result, see e.g. Rolski et al. (1999).

4. Determination of the relativities

4.1. Interaction between the BM scale and *a priori* ratemaking

Since the relativities attached to the different levels are the same whatever the risk class to which the policyholders belong, those scales overpenalize *a priori* bad risks. Let us explain this phenomenon, put in evidence by Taylor (1997). Over time, policyholders will be distributed over the levels of the bonus-malus scale. Since their trajectory is a function of past claims history, policyholders with low *a priori* expected claim frequencies will tend to gravitate in the lowest levels of the scale. Conversely for individuals with high *a priori* expected claim frequencies. Consider for instance a policyholder with a high *a priori* expected claim frequency, a young male driver living in a urban area, say. This driver is expected to report many claims (this is precisely why he has been penalized *a priori*) and so to be transferred to the highest levels of the BM scale. On the contrary, a policyholder with a low *a priori* expected claim frequency, a middle-aged lady living in a rural area, say, is expected to report few claims and so to gravitate in the lowest levels of the scale. The level occupied by the policyholders in the BM scale can thus be partly explained by their observable characteristics included in the price list. It is thus fair to isolate that part of the information contained in the level occupied by the policyholder that does not reflect observables characteristics. *A posteriori* corrections should be only driven by this part of the BM information.

Let us try to quantify these findings. To this end, we introduce the random variable L_ϑ valued in $\{0, 1, \ldots, s\}$ such that L_ϑ conforms to the distribution $\boldsymbol{\pi}(\vartheta)$ i.e.

$$\Pr[L_\vartheta = \ell] = \pi_\ell(\vartheta), \quad \ell = 0, 1, \ldots, s.$$

The variable L_ϑ thus represents the level occupied by a policyholder with annual expected claim frequency ϑ once the steady state has been reached.

Let us now pick at random a policyholder from the portfolio. Let us denote as Λ his (unknown) *a priori* expected claim frequency and as Θ the residual effect of the risk factors not included in the ratemaking. The actual (unknown) annual expected claim frequency of this policyholder is then $\Lambda\Theta$. Since the random effect Θ represents residual effects of hidden covariates, the random variables Λ and Θ may reasonably be assumed to be mutually independent. Let w_k be the weight of the kth risk class whose annual expected claim frequency is λ_k. Clearly, $\Pr[\Lambda = \lambda_k] = w_k$.

Now, let L be the BM level occupied by this randomly selected policyholder once the steady state has been reached. The distribution of L can be written as

$$\Pr[L = \ell] = \sum_k w_k \int_{\theta > 0} \pi_\ell(\lambda_k \theta) u(\theta) d\theta; \tag{4.1}$$

$\Pr[L = \ell]$ represents the proportion of the policyholders in level ℓ.

4.2. Norberg's predictive accuracy in segmented tariffs

Predictive accuracy is a useful measure of the efficiency of a BMS. The idea behind this notion is as follows. A BMS is good at discriminating among the good and the bad risks if the premium they pay is close to their "true" premium. According to Norberg (1976), once the number of classes and the transition rules have been fixed, the optimal relativity r_ℓ associated to level ℓ is determined by maximizing the asymptotic predictive accuracy.

As above, let $\Lambda\Theta$ be the true (unknown) expected claim frequency of a policyholder picked at random from the portfolio, where Θ admits the pdf (2.1) and $\Pr[\Lambda = \lambda_k] = w_k$, with $\mathbb{E}[\Lambda] = \overline{\lambda}$. Our aim is to minimize the expected squared difference between the "true" relative premium Θ and the relative premium r_L applicable to this policyholder (after the stationary state has been reached), i.e. the goal is to minimize

$$\begin{aligned}
\mathbb{E}[(\Theta - r_L)^2] &= \sum_{\ell=0}^{s} \mathbb{E}\left[(\Theta - r_\ell)^2 \mid L=\ell\right] \Pr[L = \ell] \\
&= \sum_{\ell=0}^{s} \int_{\theta>0} (\theta - r_\ell)^2 \Pr[L = \ell \mid \Theta = \theta]\, u(\theta)\, d\theta \\
&= \sum_{k} w_k \int_{\theta>0} \sum_{\ell=0}^{s} (\theta - r_\ell)^2 \pi_\ell(\lambda_k \theta)\, u(\theta)\, d\theta.
\end{aligned}$$

The solution is given by

$$\begin{aligned}
r_\ell &= \mathbb{E}[\Theta \mid L = \ell] \\
&= \mathbb{E}\left[\mathbb{E}[\Theta \mid L = \ell, \Lambda \mid L=\ell]\right] \\
&= \sum_{k} \mathbb{E}\left[\Theta \mid L = \ell, \Lambda = \lambda_k\right] \Pr\left[\Lambda = \lambda_k \mid L = \ell\right] \\
&= \sum_{k} \int_{\theta>0} \theta \frac{\Pr[L = \ell \mid \Theta = \theta, \Lambda = \lambda_k]\, w_k}{\Pr[L = \ell, \Lambda = \lambda_k]} u(\theta)\, d\theta \frac{\Pr[\Lambda = \lambda_k, L = \ell]}{\Pr[L = \ell]} \\
&= \frac{\sum_k w_k \int_{\theta>0} \theta \pi_\ell(\lambda_k \theta)\, u(\theta)\, d\theta}{\sum_k w_k \int_{\theta>0} \pi_\ell(\lambda_k \theta)\, u(\theta)\, d\theta}.
\end{aligned} \tag{4.2}$$

It is easily seen that $\mathbb{E}[r_L] = 1$, resulting in financial equilibrium once steady state is reached.

To end with, let us mention that if the insurance company does not enforce any *a priori* ratemaking system, all the λ_k's are equal to λ and reduces to the formula

$$r_\ell = \frac{\int_{\theta>0} \theta \pi_\ell(\overline{\lambda}\theta)\, u(\theta)\, d\theta}{\int_{\theta>0} \pi_\ell(\overline{\lambda}\theta)\, u(\theta)\, d\theta}$$

that has been derived in Norberg (1976).

5. Numerical illustrations

5.1. A priori ratemaking

The data used to illustrate this paper relate to a Belgian MTPL portfolio observed during the year 1997. The data set comprises 158,061 policies. The claim number distribution in the portfolio is described in Table 5.1. The overall mean claim frequency is 11.25%.

TABLE 5.1

Observed claims distribution in the Belgian MTPL portfolio.

Number k of claims reported	*Observed number of policies having reported k claims*
0	140 276
1	16 085
2	1 522
3	159
4	17
5	2
≥ 6	0

The following information is available on an individual basis: in addition to the number of claims filed by each policyholder and the exposure-to-risk from which these claims originate (i.e. the number of days the policy has been in force during 1997), we know the age of the policyholder in 1997 (18-21 years, 22-30, 31-55 or above 56), his/her gender (male-female), the kind of district where he/she lives (rural or urban), the fuel oils of the vehicle (gasoline or diesel), the power of the vehicle in kilowatts (less than 40 Kw, between 40 and 70 Kw or more than 70Kw), the use of the vehicle (leisure and commuting only, or also professionnal use), whether the vehicle has been classified as a sportscar by the company, whether the policyholder splits the payment of the premium (premium paid once a year versus premium splitted up), whether the policyholder subscribed other guarantees than MTPL (for instance material damage, theft, or comprehensive coverage in addition to MTPL).

A segmented tariff has been built on the basis of a Poisson regression model. Afterwards, geographical ratemaking has been performed following the method proposed by Boskov and Verrall (1994); see also Brouhns, Denuit, Masuy and Verrall (2002). This resulted in the definition of four zones. The final model was fitted by Poisson regression with the four zones that can be seen in Figure 5.1. A backward-type selection procedure eliminated some risk factors: use and sport were considered as non significant and were excluded from the Poisson model. This resulted in 1536 risk classes, each with its own *a priori* annual expected claim frequency. Table 5.2 displays the point estimates

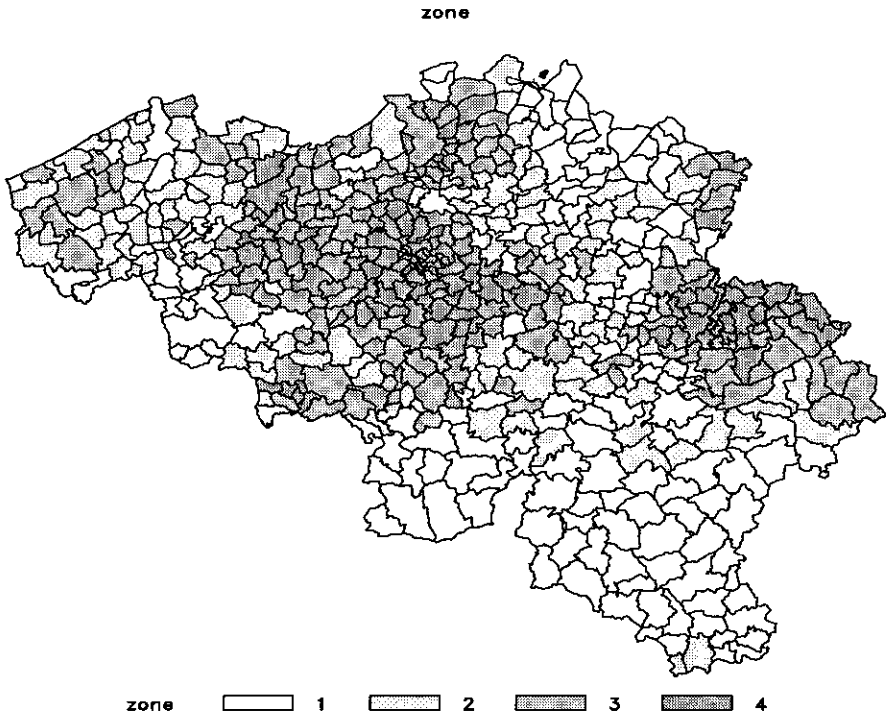


Figure 5.1: The four zones obtained with the Boskov-Verrall method.

of the regression coefficients β_0, β_1, … together with confidence intervals and p-values of test for the null hypothesis $\beta_j = 0$. Table 5.2 has been obtained with the SAS/STAT procedure GENMOD. Table 5.3 gives a part of the resulting price list. A "1" indicates the presence of the characteristic corresponding to the column. For a thorough description of the tariff construction, we refer the interested reader to Brouhns and Denuit (2003).

TABLE 5.2

SUMMARY OF THE POISSON FIT TO THE BELGIAN MTPL PORTFOLIO

Parameter	*DF*	*Estimate*	*Standard Error*	*Wald 95% Confidence Limits*		*Chi-Square*	*Pr > ChiSq*
Intercept β_0	1	–1.7326	0.0197	–1.7713	–01.6939	7701.76	<.0001
AGE 18-21	1	0.8219	0.0578	0.7086	0.9352	202.26	<.0001
AGE 22-30	1	0.3996	0.0184	0.3636	0.4357	472.45	<.0001
AGE >56	1	–0.2254	0.0185	– 0.2618	–0.1891	147.92	<.0001
AGE 31-55	0	0	0	0	0	.	.
GENDER woman	1	0.066	0.0165	0.0338	0.0983	16.1	<.0001
GENDER man	0	0	0	0	0	.	.

Parameter	*DF*	*Estimate*	*Standard Error*	*Wald 95% Confidence Limits*		*Chi-Square*	*Pr > ChiSq*
DISTRICT urban	1	0.2439	0.0153	0.214	0.2738	255.06	<.0001
DISTRICT rural	0	0	0	0	0	.	.
FUEL diesel	1	0.2074	0.0158	0.1764	0.2383	172.21	<.0001
FUEL gasoline	0	0	0	0	0	.	.
PAYMENT yearly	1	−0.2487	0.0147	−0.2776	−0.2198	284.53	<.0001
PAYMENT splitted	0	0	0	0	0	.	.
GARACCESS MTPL+	1	−0.1701	0.015	−0.1994	−0.1407	128.97	<.0001
GARACCESS MTPL only	0	0	0	0	0	.	.
POWER $>$70	1	0.1243	0.0198	0.0855	0.1631	39.38	<.0001
POWER $<$40	1	−0.0925	0.0185	−0.1288	−0.0562	24.95	<.0001
POWER 40-70	0	0	0	0	0	.	.
ZONE 1	1	−0.5492	0.0225	−0.5933	−0.5051	594.8	<.0001
ZONE 2	1	−0.3525	0.0199	−0.3916	−0.3135	313.2	<.0001
ZONE 3	1	−0.2301	0.0178	−0.2649	−0.1952	167.63	<.0001
ZONE 4	0	0	0	0	0	.	.

5.2. Scale –1/top

In this BM scale, the policyholders are classified according to the number of claim-free years since their last claim (0, 1, 2, 3, 4 or at least 5). After a claim all premiums reductions are lost. The transition rules are described in Table 5.4. Specifically, the starting class is the highest level 5. Each claim-free year is rewarded by one bonus class. In case an accident is reported, all the discounts are lost and the policyholder is transferred to level 5.

TABLE 5.4

TRANSITION RULES FOR THE BMS –1/TOP.

Starting level	*Level occupied if* **0** *claim is reported*	**≥ 1**
0	0	5
1	0	5
2	1	5
3	2	5
4	3	5
5	4	5

TABLE 5.3

A PRIORI RISK CLASSIFICATION

Fuel oil		*Premium splitting*		*Coverage*		*District*		*Gender*		*Power*		
Diesel	*Gas.*	*Annual*	*Splitted*	*MTPL*	*MTPL +*	*Rural*	*Urban*	*Woman*	*Man*	*40-70Kw*	*$>$70Kw*	*<40Kw*
1	0	1	0	1	0	1	0	0	1	0	0	1
0	1	1	0	1	0	0	1	1	0	0	0	1
1	0	1	0	1	0	1	0	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1	0	1	0
1	0	0	1	1	0	1	0	1	0	0	0	1
0	1	1	0	0	1	1	0	1	0	1	0	0
0	1	1	0	1	0	1	0	0	1	0	1	0
1	0	1	0	1	0	0	1	0	1	1	0	0
1	0	1	0	1	0	1	0	0	1	0	0	1
0	1	0	1	0	1	0	1	0	1	1	0	0
1	0	1	0	0	1	1	0	0	1	1	0	0
0	1	0	1	1	0	1	0	1	0	0	1	0
0	1	0	1	0	1	0	1	0	1	0	0	1
1	0	0	1	0	1	1	0	1	0	0	1	0
1	0	1	0	1	0	0	1	1	0	0	0	1
0	1	0	1	1	0	1	0	0	1	1	0	0
0	1	1	0	1	0	1	0	0	1	0	0	1
1	0	0	1	0	1	0	1	0	1	0	1	0
0	1	1	0	1	0	1	0	0	1	0	1	0
1	0	0	1	0	1	1	0	1	0	1	0	0
0	1	1	0	1	0	0	1	1	0	0	1	0
0	1	0	1	1	0	0	1	1	0	0	0	1
1	0	1	0	0	1	0	1	1	0	0	1	0
1	0	0	1	0	1	1	0	1	0	0	0	1
0	1	1	0	1	0	0	1	0	1	1	0	0

Age				*Zone*				*Annual*
18-21	***31-55***	***22-30***	***>55***	***Zone 1***	***Zone 2***	***Zone 3***	***Zone 4***	***Claim Freq.***
0	0	0	1	0	0	0	1	0.09858
0	1	0	0	0	1	0	0	0.09864
0	0	0	1	0	0	1	0	0.09873
0	0	0	1	0	1	0	0	0.09904
0	1	0	0	1	0	0	0	0.09909
0	1	0	0	0	0	0	1	0.09914
0	0	0	1	0	0	0	1	0.09940
0	0	0	1	0	1	0	0	0.09948
0	1	0	0	0	0	1	0	0.09979
0	0	0	1	0	0	1	0	0.09980
0	0	1	0	1	0	0	0	0.09981
0	1	0	0	1	0	0	0	0.09992
0	1	0	0	0	1	0	0	0.10011
0	0	0	1	0	1	0	0	0.10013
0	1	0	0	1	0	0	0	0.10013
0	1	0	0	0	1	0	0	0.10034
0	1	0	0	0	0	0	1	0.10047
0	0	0	1	1	0	0	0	0.10054
0	1	0	0	0	0	1	0	0.10062
0	0	0	1	0	0	1	0	0.10089
0	1	0	0	1	0	0	0	0.10096
0	0	0	1	0	1	0	0	0.10106
0	0	0	1	0	1	0	0	0.10118
0	1	0	0	0	1	0	0	0.10121
0	1	0	0	0	1	0	0	0.10139

Note that the philosophy behind such a BMS is different from credibility theory. Indeed, this BMS only aims to counteract moral hazard: it is in fact more or less equivalent to a deductible which is not paid at once but smoothed over the time needed to go back to the lowest class. Note however that this smoothed deductible only applies to the first claim.

The transition matrix $\boldsymbol{M}(\vartheta)$ associated to this BMS is given by

$$\boldsymbol{M}(\vartheta)=\begin{pmatrix} \exp(-\vartheta) & 0 & 0 & 0 & 0 & 1-\exp(-\vartheta) \\ \exp(-\vartheta) & 0 & 0 & 0 & 0 & 1-\exp(-\vartheta) \\ 0 & \exp(-\vartheta) & 0 & 0 & 0 & 1-\exp(-\vartheta) \\ 0 & 0 & \exp(-\vartheta) & 0 & 0 & 1-\exp(-\vartheta) \\ 0 & 0 & 0 & \exp(-\vartheta) & 0 & 1-\exp(-\vartheta) \\ 0 & 0 & 0 & 0 & \exp(-\vartheta) & 1-\exp(-\vartheta) \end{pmatrix}$$

It is easily checked that $p_{5\ell}^{(5)}(\vartheta)=\pi_{\ell}(\vartheta)$ for $\ell = 0,1,\ldots,5$, so that the system needs 5 years to reach stationarity (i.e. the time needed by the best policyholders starting from level 5 to arrive in level 0).

TABLE 5.5

NUMERICAL CHARACTERISTICS FOR THE SYSTEM –1/TOP

Level ℓ	**Pr** $[L=\ell]$	***Relativity*** $r_\ell = \mathbb{E}[\Theta \mid L=\ell]$ ***without*** **a priori** ***ratemaking***	***Relativity*** $r_\ell = \mathbb{E}[\Theta \mid L=\ell]$ ***with*** **a priori** ***ratemaking***	***Average a priori expected claim frequency in level*** ℓ $\mathbb{E}[\Lambda \mid L=\ell]$ ***with*** **a priori** ***ratemaking***
5	10.2%	166.6%	142.7%	12.8%
4	8.6%	154.4%	135.3%	12.5%
3	7.2%	143.8%	128.9%	12.2%
2	6.2%	134.6%	123.3%	12.0%
1	5.3%	126.5%	118.3%	11.8%
0	62.4%	70.8%	80.5%	10.6%

The results for the BM scale –1/top are displayed in Table 5.5. Specifically, the values in the third column are computed with the help of (4.3) with $\hat{a}=1.3671$ and $\hat{\overline{\lambda}}=0.1125$ Those values were obtained by fitting a Negative Binomial distribution to the portfolio observed claim frequencies given in Table 5.1. Integrations have been performed numerically with the QUAD procedure of SAS/IML. The fourth column is based on (4.2) with $\hat{a}=2.1368$ and the $\hat{\lambda}_k$'s obtained from *a priori* risk classification (i.e. from the $\hat{\beta}_j$'s displayed in Table 5.2). Once the steady state has been reached, the majority of the policies (62.4%) occupy level 0 and enjoy the maximum discount. The remaining 47.6% of the portfolio are distributed over levels 1-5, with about 10% in level 5 (those policyholders who just claimed). Concerning the relativities, the minimum percentage of 70.8% when the *a priori* ratemaking is not recognized becomes 80.5% where the relativities are adapted to the *a priori* risk classification. Similarly, the relativity attached to the highest level of 166.6% gets reduced to 142.7%. The severity of the *a posteriori* corrections is thus weaker once the *a priori* ratemaking is taken into account in the determination of the r_ℓ's. The last column of Table 5.5 indicates the extent to which *a priori* and *a posteriori* ratemakings interact. The numbers in this column are computed as

$$\begin{aligned}\mathbb{E}[\Lambda \mid L=\ell] &= \sum_k \lambda_k \Pr[\Lambda=\lambda_k \mid L=\ell] \\ &= \sum_k \lambda_k \frac{\Pr[L=\ell \mid \Lambda=\lambda_k]\, w_k}{\Pr[L=\ell]} \\ &= \frac{\sum_k \lambda_k w_k \int_{\theta>0} \pi_\ell(\lambda_k \theta) u(\theta) d\theta}{\sum_k w_k \int_{\theta>0} \pi_\ell(\lambda_k \theta) u(\theta) d\theta}. \end{aligned} \tag{5.1}$$

If $\mathbb{E}[\Lambda \mid L=\ell]$ is indeed increasing in the level ℓ, those policyholders who have been granted premium discounts at policy issuance (on the basis of their observable characteristics) will be also rewarded *a posteriori* (because they occupy the lowest levels of the BM scale). Conversely, the policyholders who have been penalized at policy issuance (because of their observable characteristics) will cluster in the highest BM levels and will consequently be penalized again. The average *a priori* expected claim frequency clearly increases with the level ℓ occupied by the policyholder.

5.3. Soft Taylor's scale (–1/+2)

Let us now consider the soft experience rating system defined in Taylor (1997). There are 9 BM levels. Level 6 is the starting level. A higher level number indicates a higher premium. If no claims have been reported by the policyholder then he moves one level down. If a number of claims, $n_t > 0$, has been reported during year t then the policyholder moves $2n_t$ levels up. The transition rules are described in Table 5.6.

Results are displayed in Table 5.7 which is the analogue of Table 5.5 for the BMS –1/+2. The BMS is perhaps too soft since the vast majority of the portfolio (about 75%) clusters in the super bonus level 0. The higher levels are occupied by a very small minority of drivers. Such a system does not really discriminate between good and bad drivers. Consequently, only those policyholders in level 0 get some discount whereas occupancy of any level 1-8 implies some penalty. Again, the a posteriori corrections are softened when a priori risk classification is taken into account in the determination of the r_ℓ's. The comments made for the scale –1/top still apply to this BMS.

TABLE 5.6

TRANSITION RULES FOR THE BMS –1/+2

Starting level	***Level occupied if*** **0**	**1**	**2**	**3**	**≥ 4**
claim(s) is/are reported					
8	7	8	8	8	8
7	6	8	8	8	8
6	5	8	8	8	8
5	4	7	8	8	8
4	3	6	8	8	8
3	2	5	7	8	8
2	1	4	6	8	8
1	0	3	5	7	8
0	0	2	4	6	8

TABLE 5.7

NUMERICAL CHARACTERISTICS FOR THE SYSTEM –1/+2

Level ℓ	**Pr** $[L = \ell]$	***Relativity*** $r_\ell = \mathbb{E}[\Theta \mid L = \ell]$ ***without*** **a priori** ***ratemaking***	***Relativity*** $r_\ell = \mathbb{E}[\Theta \mid L = \ell]$ ***with*** **a priori** ***ratemaking***	***Average a priori expected claim frequency in level*** ℓ $\mathbb{E}[\Lambda \mid L = \ell]$ ***with*** **a priori** ***ratemaking***
8	1.1%	325.3%	238.1%	17.2%
7	1.1%	294.0%	220.9%	16.2%
6	1.4%	258.0%	200.6%	15.2%
5	1.6%	234.0%	187.0%	14.5%
4	2.6%	194.5%	163.0%	13.5%
3	2.9%	179.2%	153.9%	13.1%
2	7.9%	133.9%	124.1%	12.0%
1	6.8%	127.2%	119.9%	11.8%
0	74.7%	75.6%	84.4%	10.7%

5.4. Severe Taylor's scale (-1/+4)

Let us finally consider the severe experience rating system defined in Taylor (1997). Again, there are 9 BM levels. Level 6 is the starting level. A higher level number indicates a higher premium. If no claims have been reported by the policyholder then he moves down one level. Each claim is now penalized by 4 levels (instead of 2 in the soft Taylor's scale). The transition rules are described in Table 5.8.

TABLE 5.8

TRANSITION RULES FOR THE BMS –1/+4.

Starting level	***Level occupied if*** **0** ***claim is reported***	***Level occupied if*** **1** ***claim is reported***	***Level occupied if*** **≥ 2** ***claim is reported***
8	7	8	8
7	6	8	8
6	5	8	8
5	4	8	8
4	3	8	8
3	2	7	8
2	1	6	8
1	0	5	8
0	0	4	8

Results are displayed in Table 5.9, the analogue of Tables 5.5 and 5.7. The interesting point is to compare results for the scale –1/+2 to those obtained for the scale –1/+4. The higher severity of the –1/+4 system results in more important premium discounts in the lowest part of the scale, and in reduced penalties for those occupying the highest levels. Similarly, the average *a priori* expected claim frequency for each level diminishes when the claims are more heavily penalized.

TABLE 5.9

NUMERICAL CHARACTERISTICS FOR THE SYSTEM –1/+4

Level ℓ	**Pr** $[L = \ell]$	***Relativity*** $r_\ell = \mathbb{E}[\Theta \mid L = \ell]$ ***without* a priori *ratemaking***	***Relativity*** $r_\ell = \mathbb{E}[\Theta \mid L = \ell]$ ***with* a priori *ratemaking***	***Average a priori expected claim frequency in level*** ℓ $\mathbb{E}[\Lambda \mid L = \ell]$ ***with* a priori *ratemaking***
8	4.6%	225.1%	180.7%	14.3%
7	4.3%	203.0%	167.3%	13.7%
6	4.0%	185.7%	156.9%	13.2%
5	3.8%	171.7%	148.6%	12.9%
4	7.0%	130.0%	121.1%	11.9%
3	6.1%	123.0%	116.8%	11.7%
2	5.3%	116.7%	112.8%	11.6%
1	4.7%	111.1%	109.2%	11.5%
0	60.3%	64.9%	76.5%	10.5%

6. *A POSTERIORI* CORRECTIONS DEPENDING ON *A PRIORI* CHARACTERISTICS

We know from credibility theory that the *a posteriori* corrections are functions of the *a priori* characteristics; see (2.2). On the contrary, when a BMS is in force, the same *a posteriori* corrections apply to all policyholders, whatever their *a priori* expected claim frequency. This of course induces unfairness in the portfolio.

In order to reduce the unfairness of the tariff, we could propose several BM scales, according to the *a priori* characteristics. Table 6.1 describes such a system where the company differentiates policyholders according to the type of district where they live (urban or rural). People living in urban areas have higher *a priori* expected claim frequencies. Thus, they should be more rewarded in case they do not file any claim and less penalized when they report accidents compared to people living in rural zones. This is indeed what we observe when we compare the relative premiums obtained for the system –1/+4: the maximal discount is 73.1% for urban policyholders, compared to 77.7% for rural ones. Similarly, the highest penalty is 176.6% for urbans against 183.0% for rurals.

TABLE 6.1

NUMERICAL CHARACTERISTICS FOR THE SYSTEM –1/+4 WITH THE DICHOTOMY URBAN/RURAL.

	Urban		*Rural*	
Level ℓ	*Relativity* $r_\ell = \mathbb{E}[\Theta \mid L = \ell]$ *with* **a priori** *ratemaking*	*Average a priori expected claim frequency level* ℓ $\mathbb{E}[\Lambda \mid L = \ell]$ *with* **a priori** *ratemaking*	*Relativity* $r_\ell = \mathbb{E}[\Theta \mid L = \ell]$ *with* **a priori** *ratemaking*	*Average a priori expected claim frequency in in level* ℓ $\mathbb{E}[\Lambda \mid L = \ell]$ *with* **a priori** *ratemaking*
8	176.6%	16.5%	183.0%	13.0%
7	162.5%	15.8%	169.8%	12.5%
6	151.6%	15.3%	159.6%	12.2%
5	142.9%	14.9%	151.4%	11.9%
4	116.8%	13.8%	122.9%	11.1%
3	112.2%	13.6%	118.7%	10.9%
2	108.1%	13.4%	114.8%	10.8%
1	104.3%	13.3%	111.2%	10.7%
0	73.1%	12.2%	77.7%	9.8%

7. DISCUSSION

All the techniques used in this paper resort to the stationary distribution of the scale. Therefore they can only be recommended if the steady state is reached after a relatively short period, as it is the case for the BM scale –1/top. It is worth mentioning that for the scale –1/top, the use of the stationary distribution for the computation yields higher premiums than those obtained using transient distributions, with the method of Børgan, Hoem and Norberg (1981).

The method described in the present paper can be extended to transient distributions, in the spirit of Børgan, Hoem and Norberg (1981). This may be interesting when a new scale is introduced or for BMS needing many years to reach their stationay regime.

If on a given market companies start to compete on the basis of BMS many policyholders could leave the portfolio after the occurrence of an accident, in order to avoid the resulting penalties. Those attritions can be incorporated in the model by adding an additional level to the Markov chain (in the spirit of Centeno and Silva (2001)). Transitions from a level of the BMS to this state represents a policyholder leaving the portfolio whereas transitions from this state to any level of the BMS means that a new policy enters the portfolio.

It has been assumed throughout this paper that the unknown expected claim frequencies were constant and that the random effects representing hidden characteristics were time-invariant. Dropping these assumptions makes the determination of the relativities much harder. We refer the interested reader to Brouhns, Guillén, Denuit and Pinquet (2003) for a thorough study of this general situation.

Acknowledgements

The authors wish to express their gratitude to the anonymous referees who have corrected a mistake in the original manuscript and whose numerous remarks and suggestions have considerably improved the text.

References

Borgan, Ø, Hoem, J.M. and Norberg, R. (1981) A nonasymptotic criterion for the evaluation of automobile bonus systems. *Scandinavian Actuarial Journal*, 265-178.

Boskov, M. and Verrall, R.J. (1994) Premium rating by geographical area using spatial models. *ASTIN Bulletin* **24**, 131-143.

Brouhns, N. and Denuit, M. (2003) Applications des modèles généralisés additifs à l'assurance automobile. Manuscript.

Brouhns, N., Denuit, M., Masuy, B. and Verrall, R. (2002) Ratemaking by geographical area in the Boskov and Verrall model: a case study using Belgian car insurance data. *actu-L* **2**, 3-28.

Brouhns, N., Guillén, M., Denuit, M. and Pinquet, J. (2003) Optimal bonus-malus scales in segmented tariffs. *Journal of Risk and Insurance*, in press.

Centeno, M. and Silva, J.M.A. (2001) Bonus systems in an open portfolio. *Insurance: Mathematics & Economics* **28**, 341-350.

Dionne, G., and Vanasse, C. (1989) A generalization of actuarial automobile insurance rating models: the Negative Binomial distribution with a regression component. *ASTIN Bulletin* **19**, 199-212.

Lemaire, J. (1995) *Bonus-Malus Systems in Automobile Insurance*. Kluwer Academic Publisher, Boston.

NORBERG, R. (1976) A credibility theory for automobile bonus system. *Scandinavian Actuarial Journal*, 92-107.

ROLSKI, T., SCHMIDLI, H., SCHMIDT, V. and TEUGELS, J. (1999) *Stochastic Processes for Insurance and Finance*. John Wiley & Sons, New York.

TAYLOR, G. (1997) Setting a Bonus-Malus scale in the presence of other rating factors. *ASTIN Bulletin* **27**, 319-327.

SANDRA PITREBOIS & JEAN-FRANÇOIS WALHIN
Secura Belgian Re
Avenue des Nerviens, 9-13 boîte 6
B-1040 Bruxelles, Belgium

MICHEL DENUIT
Institut de Statistique
Université Catholique de Louvain
Voie du Roman Pays, 20
B-1348 Louvain-la-Neuve, Belgium

JEAN-FRANÇOIS WALHIN & MICHEL DENUIT
Institut des Sciences Actuarielles
Université Catholique de Louvain
Grand-Rue, 54
B-1348 Louvain-la-Neuve, Belgium

BOOK REVIEWS

Modern Actuarial Risk Theory by Rob Kaas, Marc Goovaerts, Jan Dhaene and Michel Denuit [Kluwer Academic Publishers, Boston, 2001]

The publication of a book on Risk Theory is a sufficiently rare occurrence for it to be greeted enthusiastically. My enthusiasm for this book increased when I read in the authors' Preface that its intended readership is students in the final year of a bachelors program in quantitative economics or mathematical statistics or a masters program in actuarial science or in quantitative financial economics. Too few of the texts on risk theory are suitable for university students at this level. Gerber's (1979) book is a classic, but more suited to researchers, and, at the other end of the spectrum, Daykin *et al.*'s (1994) book covers many of the practical aspects of its subject well, but at the expense of a clear technical development. An exception to this is the excellent, albeit almost encyclopaedic, book by Klugman *et al.* (1998).

Modern Actuarial Risk Theory is a translation into English of a book which has been used in universities in The Netherlands and Belgium for more than ten years. The chapters in this book are:

1. Utility theory and insurance
2. The individual risk model
3. Collective risk models
4. Ruin theory
5. Premium principles
6. Bonus-malus systems
7. Credibility theory
8. Generalized linear models
9. IBNR techniques
10. Ordering of risks

There are a large number of, mostly short, exercises at the end of each chapter and a section at the end of the book containing answers or hints on how to complete the exercises.

The chapter titles are broadly in line with what I would expect to see in an undergraduate text on Risk Theory. However, the four authors have collectively made an enormous contribution to the development of actuarial science in recent years and some chapters of the book, notably Chapters 5 and 10, clearly reflect their interests.

A novel, and welcome, feature of such a book is the inclusion of generalized linear models (GLMs, Chapter 8). Such models are extremely useful in many branches of actuarial science and the authors demonstrate this in Chapter 9 where they model claims run-off data using a GLM and then show that several standard numerical methods, notably the chain ladder method, can be derived as special cases of this GLM. Surely this is a more satisfactory way to treat this subject than is usually found in textbooks.

A less welcome feature of the book is that it treats most topics in a mathematical way and provides little insurance context to motivate these topics. A good example of this is reinsurance. This is mentioned frequently throughout the book but there is no real discussion of how and why it operates. Presumably, lecturers teaching a course based on this book would be expected to provide this background material from other sources. Chapter 6 on Bonus-malus systems is an exception – it does have a good motivational introduction based on the Dutch system.

Another less welcome feature is that nowhere in the book, apart from one table in Chapter 7, do the authors use real insurance data to illustrate their technical development. This is a pity. The use of such data would inevitably require a brief description of the origin of the data, thereby giving the reader greater understanding of why the mathematical development is useful and an appreciation that Risk Theory is useful in practice. This is in marked contrast to the book by Klugman *et al.* (1998), where real data are used extensively.

Amazon's website gives the price of Modern Actuarial Risk Theory as US$144. This is on the high side, even by today's standards, for a "must buy" text for an undergraduate course. The same website gives the price of Loss Models as (no more than) US$110.

References

Daykin, C., Pentikäinen, T. and Pesonen, E. (1994) *Practical Risk Theory for Actuaries*. Chapman and Hall, London.

Gerber, H. (1979) *An Introduction to Mathematical Risk Theory*. S.S. Huebner Foundation, Philadelphia.

Klugman, S., Panjer, H. and Willmot, G. (1998) *Loss Models: From Data to Decisions*. Wiley, New York.

Howard Waters
May 2003

Mary Hardy: *Investment Guarantees: Modelling and risk management for equity-linked life insurance*. John Wiley & Sons. ISBN 0-471-39290-1, 2003.

It is always a pleasure to read something written by Professor Hardy. It is doubly so when it is a book on a subject that I have been long concerned with, and triply so when I and two colleagues, Dr Sheauwen Yang and Professor Howard Waters, have recently presented a long paper (2003) to the actuarial bodies in the United Kingdom on a rather similar subject, that of contracts with guaranteed annuity options (GAOs). (I shall refer to this paper and its authors as "WWY"). So I am very pleased to write this review, and to be able to recommend Professor Hardy's book most warmly.

Although I might have wished that I had written a book on this subject myself, this is not exactly the book that I would or could have written. Professor Hardy's approach is very similar to that of myself and my colleagues in relation to GAOs, but it also differs in a number of places, and she describes some things she has done that I have never attempted. I shall draw attention to our differences as we go along.

The book, according to her Introduction "is designed for all practitioners working in equity-linked insurance … It is written with actuaries in mind, but it should also be interesting to other investment professionals. [It] forms the basis for a one-semester graduate course for students of actuarial science, insurance and finance." In my view it succeeds well in these objectives. The actuarial material relating to mortality tables is tidied away into short Appendices. I am not sure that a practitioner who had no statistical or mathematical training at all could easily follow it, but it should present no difficulties to any actuary.

Equity-linked life assurance (as it is called in the U.S.A.) goes under several names: unit-linked in the United Kingdom, segregated funds in Canada. Many of the policies provide, or used to provide, guarantees of a minimum sum assured on maturity, and perhaps also on earlier death. The existence of these policies in the U.K. in the 1970s led to the seminal work done by the Maturity Guarantees Working Party (MGWP), whose report published in 1980, suggested setting up contingency reserves calculated as what are now called quantile reserves or value-at-risk reserves.

Equity-linked assurances with investment guarantees are the archetypal examples of a life insurance policy that contains benefits defined as the better of A and B, where A and B are amounts that are both defined in the policy. One can relate these to standard financial options by defining a new type of option, a *Maxi*, whose payoff at the expiry date is $\mathrm{Max}(A,B)$. A Maxi is easily related to the more usual *Call* and *Put* options. An equity-linked policy can be treated as an investment in ordinary shares ("equities" or "common stock") plus a put option, or as an investment in cash plus a call option, or as a maxi.

However, Professor Hardy's initial approach, like that of the MGWP and of WWY, is to ignore the financial option concepts, and to estimate quantile reserves (or better "conditional tail expectation" or CTE reserves) by the use of simulation. She describes this as "the actuarial approach", as opposed to the "dynamic hedging approach" of financial economists. However, many actuaries understand financial options, and many financial economists understand the necessity for contingency reserves, so the names are no more than convenient labels.

Chapter 1 of the book describes the types of policy considered, and the history and background. To do simulations one needs a stochastic simulation model, to replicate the "real world" and this is considered in Chapter 2. In this chapter the author describes several possible models and modifications thereof. Each of the models is fitted to two data sets, monthly values from about 1956 to 2000 of the TSE 300 index and the of the S&P 500 index. The models described include the independent lognormal, autoregressive AR(1) lognormal, and regime-switching lognormal (RSLN) models for the structure, ARCH and GARCH models for the residuals, also the empirical distribution, the Wilkie model, and vector autoregressive (VAR) models. It is clear that the author prefers the RSLN model.

One must emphasise that at this point we are seeking a model to represent the real world movement of economic variables, in this case the total returns on shares. We are not concerned with option pricing models. We would like a model of the real world that is as realistic as we can make it, and we can justify from the data, without its becoming intractable for simulation. If we wish to restrict ourselves to total returns on shares then the RSLN seems to have many advantages. But in general it seems to me to be a pity to look only at total returns. Share dividends, and share earnings are additional information, to which participants in the market do pay attention. The rate of inflation and interest rates, and for a country like Canada exchange rates and what is happening in the United States, may also be relevant, as Hardy later observes (on page 87). Therefore I would prefer to use an integrated model, on the lines of the Wilkie model, rather than model restricted to one series. However, I see no reason why we should put ourselves in the straightjacket of a VAR model. The relationships between variables may not be all strictly linear.

Harris (1999) has applied RSLN models to multivariate data. Whitten & Thomas (1999) apply a threshold model to multivariate data. In both cases there are multiple regimes (but restricted in their examples to two). All variables are in the same regime at once (but one could imagine a model where this did not apply). In the RSLN model the regime switches at random between models with specified probabilities. In the threshold model the regime is in one model or another depending on the value of an indicator variable in the previous period (whether or not inflation was higher or lower than 10%). An elaboration of Hardy's RSLN model would be to include the US index and the Canadian index in one model, and define four states where neither, one or other, or both are in the higher variance regime. This could take account of the connection between the states that is observed on page 87.

To use a model we must estimate parameters for it, and the next three chapters discuss this. In Chapter 3 Professor Hardy discusses the classical maximum likelihood estimation (MLE) method, how one derives the MLE parameter values, uses the information matrix to derive confidence intervals for and correlations between the parameter estimates, and then uses criteria such as the likelihood ratio test and Akaike criterion to choose between models. This is standard material, clearly presented.

In Chapter 4 the "left tail calibration method" is described. For the particular application low share returns are critical, so it is desirable that the left (negative) tail of the distribution is adequately modelled. It is clear that the monthly returns are "fat-tailed", which is why an ARCH or GARCH or RSLN model is much better than a simple lognormal model for representing the whole distribution. But it is possible to adjust the parameters of any of the models (usually just the standard deviation) so that the left tail is adequately "fattened". This usually means that the right tail is not fitted so well. The motivation for this process is also to meet the requirements of the Canadian Institute of Actuaries' report on segregated funds, which would allow a life office to use any model it wished provided that certain statistics in relation to the left tail are adhered to.

One method that Professor Hardy does not discuss is to use a fat-tailed distribution (other than a stable distribution) for the residuals. If Z represents the standardised (0,1) residuals, one can generate Z as $X_1 - X_2$ where X_1 and X_2 are distributed with any distributions defined on $(0, \infty)$, such as lognormal, gamma, Weibull, Pareto or many others. Since X_1 dominates the right tail and X_2 the left tail, one can fit the tails separately if one wishes. The MLE method would be difficult, but one can match higher moments or quantiles. The method is mentioned by WWY and seems worth considering.

In Chapter 5 we move on to Bayesian Markov Chain Monte Carlo (MCMC) methods. I have previously found these difficult to follow, I suppose because I have not in fact implemented them myself (I do not feel that I really understand a numerical mathematical method unless I have written a computer programme to implement it), but I find Hardy's explanation as clear as any I have seen so far. The advantage of the MCMC method, which indeed looks complicated as compared with the MLE method, is that it gives empirical, simulated, distributions for the parameters. The MLE method gives the covariance matrix of the parameter estimates, but one then assumes normality, and the results are only asymptotically normal. The MCMC method shows that the distribution of the parameter estimates is not as normal as one might have hoped.

This is important when we come later to discuss the effect of parameter uncertainty on the simulation results for the investment guarantees. One can allow for this by using a different set of parameters for each simulation, picked from a multivariate distribution of the parameters, using what WWY call a "hypermodel". MLE gives a multivariate normal distribution from which one can pick. MCMC gives an empirical multivariate distribution, with as many values to pick from as one has chosen to simulate in the MCMC procedure. There are both theoretical and practical considerations that might influence which method one chooses to use. Normal distributions can be awkward if the parameters are essentially positive (such as a variance) or restricted to a range such as $(0, 1)$ or $(-1, 1)$ (such as an autoregressive parameter), but one can transform the parameter (assume that log variance is normal), or just restrict it to the desired range (set any value greater than 1 to 1).

Using an empirical distribution requires large computer storage, which may or may not be a problem, and restricts the drawn parameter values to the range in the empirical distribution. If one simulates enough values by MCMC that may not be a problem; but it may put up the storage requirements. It seems that there would be more work to be done before one could say that MCMC methods should always be used, but I am sure that they should be tried out.

In Chapter 6 Professor Hardy shows how to model the guarantee liability using "the actuarial method", that is by setting up a contingency reserve at the start of the contract, which is invested in a specific, but unchanging, way, and which has a given chance (e.g. 99%) of being sufficient to meet the emerging liability.

The contracts that she describes in Canada have some features that may not be customary elsewhere, and this complicates things. Thus the policies usually have a guaranteed minimum benefit on death and also on maturity, though these may be defined differently; but also there may be multiple maturity dates, at each of which the policy may be "rolled over" for a further period; at that time if the guarantee is in the money, the insurer may pay out the difference; if it is out of the money, the guarantee may be reset at the higher current fund value; the policyholder may also have the option to reset at any time or at specified times for some minimum future period.

The methodology described allows for both deaths and withdrawals, and also for management charges and special charges for the guarantee. I sometimes feel that these practical complications, which of course must be allowed for by a real life office, serve to confuse the issue in a more theoretical exposition where one wishes to get over the fundamental principles. Fewer complications could have been included, but I do not feel that what is there is excessive.

However, although Professor Hardy shows how to obtain distributions of the costs, both on an emerging cash flow and on a present value basis, one thing that is missing here is how to calculated the charges, which is covered later in Chapter11.

At this stage it is also assumed, without discussion, that the guarantee reserve is invested in risk-free instruments. This is probably the best strategy for this type of contract. But an alternative would have been to invest the reserves in the same fund as the policy. For other types of contract this might prove to be the better. It should be investigated too, as is done to some extent in WWY.

Chapter 7 is entitled "A review of option pricing theory" and it performs that function quite satisfactorily. As the author remarks, those who are familiar with the Black-Scholes principles can pass it by.

In Chapter 8 Professor Hardy explains how the option pricing methodology, with dynamic hedging, can be applied to the specific problem of investment guarantees. Although what is presented is quite clear, I would have taken it more slowly and more fully. Thus I would have started by demonstrating (for the benefit of the sceptics) that, if the real world behaves in accordance with the P-measure probabilities in the option pricing model, then dynamic hedging according to the Q-measure calculations does indeed provide investment proceeds that are close to what is required, and the more frequent the hedging the smaller the variance of the hedging error. Then one can go on, as the author does, to show that, even if the real world behaves according to some other model, in this case the RSLN model, then the proceeds may not be too far out, provided some of the parameters, in particular the variances, are comparable.

The very important point is made, which can hardly be over-emphasised, that one needs two models for these calculations, one an option pricing model which is used for calculating the option price and hedging quantities at each time step, and the other a model that simulates the real world in whatever way one wishes. I believe that lack of clarity about this may cause much confusion.

Professor Hardy sensibly shows a numerical example of dynamic hedging for a 2-year contract with no mortality and withdrawals, before going on to the complications of dealing with these decrements. I am a great believer in showing the simple case first. If it is confused with too may irrelevant features, the important points may be lost.

One aspect where I was not entirely happy with the explanation is in relation to calculating the present value of the "margin offset" charge. In Canada the guarantees are explicitly charged for by making a charge on the fund units every month of α times the amount of the fund at that time. This is in addition to, or part of, a management charge per month, and the total of them is m times the amount of the fund each month. Thus the invested fund increases at a rate less than the total return on shares (even if were to be invested in the share index). This in fact makes the guarantee more likely to be in the money at maturity. But to calculate the value of the margin offset, Professor Hardy sums the monthly charges, discounted at the risk free rate, and then takes the expected value under the Q measure. If we ignore the charges and other details one can get the present value, A, discounted at the risk free monthly rate r as:

$$\text{A} = \text{E}_{\text{Q}}[\Sigma_{t=0,n-1}\ \alpha.S_t\ e^{-rt}]$$

Where n is the number of months and S_t is the share index value at time t. A is then equated to the initial value of the option, B, to get a value for α.

It does not seem immediately clear why the Q measure is used, but I think it can be explained: we (the life office) wish to set up the hedging portfolio for the whole option initially. We require therefore to borrow an amount B. We can repay the loan from the future margin offset charges that we shall receive. The amounts of these will depend on the fund performance. But if we borrow shares of value B (or denominate the loan as if it were in shares), then, using shares as the numeraire, we do know what we shall receive, and we can repay the loan exactly as we receive the charges. This would justify discounting at the rate of return on the shares, and the result is certain, so we do not need to take expectations. We therefore put:

$$\text{A} = \Sigma_{t=0,n-1}\ \alpha.S_t S_0/S_t = \alpha.S_0\ [\Sigma_{t=0,n-1}\ 1]$$

And the answer, after allowing for the complications we have missed out, is the same as Professor Hardy gets. However, the process of financing the initial option value by borrowing shares is not explained. Effectively, the future margin offsets are hedged, which justifies using the Q measure, but the hedging is static, not dynamic, except that some of the loan is repaid every month.

An aspect where Professor Hardy treats things differently from the way WWY do is in the dynamic hedging process. Just before each rebalancing date (at time t^-), the hedge portfolio has value $H(t^-)$; the desired value is $H(t)$, and the hedging error is the difference between these. Professor Hardy assumes that the difference is made up at once (or taken away if it is a surplus), so that the investments at time t^+ are always what is required by the hedging process. She then discounts the hedging errors at the risk free rate to get a present value for them. This implicitly assumes that the hedging errors are financed by, or invested in, the risk free asset. WWY treat the affair differently. They assume that all that is available is $H(t^-)$, and they make alternative assumptions about how it is invested: (i) the right amount could be put into shares, with the balance invested in the risk-free asset; or (ii) the right amount could be put into the risk-free asset with the balance in shares; or (iii) the amount available could be invested in the right proportions. Option (i) is equivalent to what Professor Hardy has done, and it seems not unreasonable in this case that it turns out that the hedging error turns out to have lower variance under this option. But for other options the same result is not found. In my view one always needs to consider exactly how funds are invested or capital is financed, and not just assume that one should discount at any given rate.

In Chapter 9 risk measures are discussed, in particular quantile reserves (QR or VaR), and conditional tail expectations (CTE or Tail VaR). The latter have many desirable properties, and Professor Hardy, the Canadian Institute of Actuaries Taskforce and WWY all agree in preferring them to the former. Hardy shows how QR and CTE are related, how in some simple cases they can be calculated analytically, and how confidence intervals can be derived when they are simulated, all with practical examples. One nice feature is that graphs of distribution functions are drawn with the axes transposed (0 to 1 on the x axis, amounts on the y axis) so that a "more risky" distribution appears higher than a less risky one.

Hardy compares the QRs and CTEs found from the static (actuarial) and the dynamic (hedging) approaches, and shows that the latter gives (in her examples) lower extreme quantiles than the former, though the average cost/claim is often higher. This agrees with most of the results in WWY for GAOs, but they found that in some cases hedging gave even higher quantiles than the static approach.

A point not mentioned by Hardy is that CTEs allow an easy method of assessing the costs for individual policyholders as opposed to the costs for the whole portfolio; this is discussed in WWY. But a further point is that, although the CTE is analogous to a stop-loss calculation, being enough to provide a quantile reserves and also pay a "premium" for the average claim in excess of the QR, such insurance could not possibly be obtained at that price, so in effect the CTE, without reinsurance, is just a QR with a higher security level, a higher value of α.

In Chapter 10 the contracts are investigated using emerging cash flow analysis and profit testing, taking capital requirements into account. The distribution of profit using some desired rate of return on the capital required is the focus of interest. This is quite similar to what WWY have done, though the way that it is expressed by the different authors does not make this immediately clear. Hardy assumes the charge as given and calculates the expected profit and distribution of profit at different desired rates of return (risk discount rates).

WWY choose specimen rates of return, and calculate the break-even charge that results. But in both cases it is recognised that prudential reserves, whether these are part of the policy reserves or treated as solvency capital, are required, and the policyholders need to pay the average cost of their benefits, plus a "rent" for the use of this capital. So the premium they pay for the guarantee needs to be enough to cover both parts.

Hardy discusses the development of the prudential reserves (on a 95% QR basis) for specimen simulations, but does not bring out the additional aspect that the "fair value" of the contract, the price at which it could be transferred to another provider, which is what modern accounting principles are working towards, should be calculated on the same principles as the initial premium, as the expected value of the benefits (the "best estimate" perhaps) together with a sum that allows an adequate profit on the required contingency reserves. The fair value does not include the contingency reserves, but prudent reserves do include it. This is discussed by WWY, but whether the prudent reserves are part of the policy reserves or are part of the solvency capital, which in some countries may be of considerable practical significance, e.g. in relation to tax, is not considered, though Hardy mentions this point.

Chapter 11 discusses the important topic of forecast uncertainty (I should not say "important"; all the chapters in this book cover important topics). Professor Hardy attacks this in four steps: first, the errors from the random sampling inherent in Monte Carlo simulation; then variance reduction techniques; then parameter uncertainty; and finally model uncertainty.

Increasing the number of simulations reduces the random sampling errors, and it is useful to try out the convergence when the asymptotic result can be calculated analytically. The number of simulations required depends on the quantity we are estimating; tail values require more simulations than do means. A number of instructive examples are given.

Variance reduction techniques are also discussed, but Professor Hardy concludes that the only one that helps in this context is the control variate method. I had found, long ago, that some variance reduction techniques, such as importance sampling, were more trouble than they were worth, and indeed were sometimes so much slower than the simple method of just increasing the number of simulations was the best technique. The speed of computers has made it easier to do many more simulations. But one small feature that I discovered recently was that to calculate QRs or CTEs one needs to sort the results into order; many sorting routines increase in speed with the square of the number of cases sorted, and I found that the sorting took longer than the simulations had done; further investigation showed that a modern sorting technique (in fact Quiksort) improved my sorting speed over 100-fold, and that to take account of the fact that very many simulations gave guarantee costs of zero improved my sorting speed another 50-fold. Good computer algorithms, and also the source language one uses (complied or interpretive), can still make an enormous difference to computer run times. Looking carefully at your programmes may be a lot better than any variance reduction techniques.

Parameter uncertainty can be dealt with in three ways, of which Professor Hardy discusses only two, the Bayesian MCMC approach, and "stress testing" by using alternative, but perhaps arbitrary, sets of parameters. With a complicated model it is not always easy to see which way one should move the parameters to test for stress, so I favour the "hypermodel" approach, by which I mean choosing, for each simulation, a random set of parameter values from some multivariate distribution for the parameters. Hardy uses the results from the MCMC approach; the alternative is to use the information matrix from the MLE method and to assume that the parameters are multivariate normally distributed. As I have noted above, this may require a careful choice of which parameter one chooses; log σ^2 may be better than just σ. The multivariate normal method requires much less storage than MCMC, and it has the advantage that one can more easily tinker with the hyperparameters (the parameters of the distribution of the simulation parameters), and even splice together estimates from different investigations, which I suppose cannot be done with MCMC.

Model uncertainty is the last topic in this Chapter. Hardy's method is to try out alternative models. I would do just the same.

This ends the book's discussion of performance guarantees. Chapters 12 and 13 discuss, rather briefly, two extra topics: guaranteed annuity options (GAOs), and equity-indexed annuities. It is useful that these are mentioned, but a pity that they could not be fully developed. The paper on GAOs by WWY has 129 pages, and Yang's (2001) thesis is a great deal longer. Hardy gives 16 pages to the topic.

Some aspects of GAOs are similar to equity linked life insurance, in that, like them, the benefit can be defined as $\mathrm{Max}(A, B)$. Large contingency reserves may be required, and the actuarial and the hedging approaches are both possible. But GAOs have many different features. The type discussed by Hardy and by WWY is an equity-linked policy with a GAO at a fixed maturity (retirement) date, but in practice many of the policies issued in the U.K. have been with profits policies, with a range of possible maturity dates. Sometimes the guarantee is simply that a minimum amount of annual annuity will be available, rather than that the fund proceeds can be converted at a guaranteed rate; this is of course much cheaper.

To value GAOs one needs a stochastic model for interest rates, as well as for shares. A full yield curve model would be desirable, but one can do a lot with a model that allows for a level yield curve. Hardy has investigated U.K. data, and suggests two regime switching models, one for the FTSE All-Share index, one (with two autoregressive models) for the (long-term) yields on $2\frac{1}{2}\%$ Consolidated Stock ("Consols"), which is in effect a perpetual (and very old) British Government stock. This section is new material. But it is then shown how the actuarial approach can be applied, assuming that future mortality rates are known.

In practice future mortality rates cannot be forecast with certainty, and Yang (2001) investigates the effect of assuming a stochastic model (or "hypermodel") for forecast mortality rates. This is not just a matter of allowing for the random deaths in a small population of annuitants, but of allowing for the uncertainty of the underlying rates. Yang's method resembles that of Lee & Carter (1992), with some simplifications and some additional features. WWY show that the improvements in mortality in the U. K. since 1985 have been just as important in increasing the cost of GAOs as the falls in interest rates that have occurred. Hardy does not discuss these points.

A further feature of GAOs is that, with a fixed guaranteed rate (Hardy, as Yang and WWY, uses £111 per annum per £1,000, though the actual rates offered by different offices vary considerably) the cost of the guarantee, however measured, varies very much with current interest rates, i.e. how much into or out of the money the guarantee is, whereas (at least under the lognormal model) the value of the equity linked investment guarantee is the same at all starting dates. This means that the uniform monthly charge, useable for equity linked guarantees, is unsuitable here, and an up-front charge, or at least a periodic charge that is fixed in advance and depends on the conditions at commencement is desirable.

GAOs lend themselves to option pricing models. It is convenient, though less realistic, to model the market annuity rate (at age 65) as a lognormal model, as do Yang and WWY. Full yield curve models after retirement have also been proposed, by e.g. Boyle and Hardy (2002). GOAs can be treated like a portfolio of bond "swaptions", as shown by Pelsser (2002). But an extra feature of the equity-linked GAO is that the amount to be converted depends on share price performance, so the option is analogous to a "quanto" option. This makes the option pricing mathematics harder to develop, but the results are not too difficult to understand. However, to hedge one needs to hold the full value of the policy including the option in shares, and then have offsetting long and short amounts, long in a portfolio that would replicate the deferred annuity and short in a zero-coupon bond maturing at the maturity date. The required amounts are the larger the more the option is "in the money". But whether long enough bonds to match the deferred annuity, and whether it is practicable to have large short holdings in zero-coupon bonds (unless they can be "borrowed" from the rest of the life office) are both doubtful.

Thus the dynamic hedging approach for GAOs may be impractical. It is therefore necessary for life office to consider the required contingency reserves, with both the static and the dynamic approaches. Hardy covers the main aspects well, but necessarily leaves much unsaid.

Equity-indexed annuities, covered in Chapter 13, are much simpler, appearing very similar to the equity-linked guarantee, but typically funded as an investment in bonds plus a call option, rather than as an investment in shares plus a put option. The term is typically much shorter, the option risk is often reassured with a third party, and the guarantee depends usually on the share price index, not a total return index. However, there are many interesting features of these contracts, including annual minima and maxima, and the possibility that the share return guaranteed is taken as only a fraction of the actual return. However as Hardy says, these policies are usually tackled as (possibly complicated) option pricing problems, and the actuarial method is normally absent.

This review is rather longer than is usual in *ASTIN Bulletin*, but I have had a lot to say on the subject. But modern reviews do not begin to compete with those of the 19th Century. Macaulay's review in *The Edinburgh Review of Gleig's Memoirs of the life of Warren Hastings* (1841) takes 140 pages in my reprinted (1898) copy. Macaulay's review is perhaps more worth reading nowadays than the book he was reviewing. This is not the case for this article. Read Mary Hardy's excellent book.

DAVID WILKIE

REFERENCES

BOYLE, P.P. and HARDY, M.R. (2002) *Guaranteed annuity options*. Working Paper, University of Waterloo.

HARRIS, G.R. (1999) Markov chain Monte Carlo estimation of regime switching vector autoregressions. A*STIN Bulletin* **29**, 47-80.

LEE, R.D. and CARTER, L. (1992) Modelling and forecasting U.S. mortality, *Journal of the American Statistical Association* **87**, 659-671.

MATURITY GUARANTEES WORKING PARTY (1980) Report of the Maturity Guarantees Working Party. *Journal of the Institute of Actuaries* **107**, 103-209.

PELSSER, A. (2002) *Pricing and hedging guaranteed annuity options via static option replication*. Working paper.

WHITTEN, S.P. and THOMAS, R.G. (1999) A non-linear stochastic asset model for actuarial use. *British Actuarial Journal* **5**, 919-953.

WILKIE, A.D., WATERS, H.R. and YANG, S. (2003) Reserving, pricing and hedging of policies with guaranteed annuity options. *British Actuarial Journal* **9**, forthcoming.

YANG, S. (2001) *Reserving, pricing and hedging for guaranteed annuity options*, PhD Thesis, Heriot-Watt University, Edinburgh.

Report on the
XXXIV International ASTIN Colloquium,
August 24-27, 2003, Berlin

The 34th International ASTIN Colloquium 2003 was held in Berlin from August 24 to 27. More than 300 participants from 34 countries all over the world made it an extremely successful event. Opening addresses were given by Edward Levay and W. James MacGinnitie for the ASTIN Committee, and Elmar Helten, Dieter Köhnlein and Christian Hipp for the local Organizing and Scientific Committee. In contrast to its predecessors, no parallel sessions for presentations were planned this time. Instead, Poster Sessions were organized which offered ample opportunity to the participants for scientific exchange with the authors during extended coffee breaks. This new concept was very much welcomed by the attendants and will most probably be maintained also in future ASTIN conferences.

Social events were a visit to "TIPI – the tent", a varieté show on the evening of the first conference day, a boat trip on the Wannsee on the second day and a closing gala dinner at the post museum in the evening of the final day.

The topics of the invited speakers for this year's Colloquium focused on two major theme groups, one of which was the ongoing and deepening interplay between financial and insurance markets (key words: Financial Risk Management, Securitization, Loss Reserving, Solvency Standards). The second one was devoted to the possible consequences of modern medical genetic research for life and health insurance.

The opening plenary lecture was given by Paul Embrechts (ETH Zürich) on the topic "Insurance analytics: actuarial tools for financial risk management". A key message within this very comprehensive and refreshing survey over recent developments in this area was "actuarial thinking", in particular in connection with financial risk management. As new challenges for actuaries due to regulatory measures in the spirit of Basel I and II were mentioned: premium principles and risk measures – pricing in incomplete markets – stress- and solvency-testing – dynamic financial analysis (DFA) – stochastic dependence structures "beyond the normal distribution" (key word: copulas).

The following presentation of David Mocklow (Chicago) was devoted to the topic "Risk Linked Securities: what's shaking?". Seen from the perspective of a reinsurer, a thorough survey over different products and markets related to ART was given, with a particular emphasis on natural catastrophes and the "value of securitization".

The lecture of Greg Taylor (Sydney and University of Melbourne) closed the plenary lecture of the first day. It was dealing with "Loss reserving techniques: past, present and future". In his lecture, he presented a hierarchical approach to classifying the various known loss reserving techniques, differentiating between macroscopic vs. microscopic models, stochastic vs. deterministic models and models with and without dependence structures. An in-depth-analysis was made for adaptive approaches including Kalman filtering, a technique which is well-known in the area of generalized linear models (GLM).

The second day of the colloquium started with a survey lecture of Søren Asmussen (Aarhus University) on "Some applications of phase-type distributions to insurance and finance". After some introductory remarks on historic developments and fundamental properties of phase-type distributions emphasis was put on the advantages of a rigorous application of matrix calculus in this field. This simplifies not only a lot the classical proofs, but allows also for new results in an elegant way, in particular in ruin theory with a finite time horizon, or for pricing of Russian options.

The subsequent two lectures were devoted to the medical topics outlined above. Jean Lemaire (Wharton School, Philadelphia) and Angus MacDonald (Heriot-Watt University, Edinburgh) shared their presentation on "Genetics, family history, and insurance underwriting: an expensive combination?". A major problem in this context is the question whether life or health insurance companies have the right to use or obtain information on the genetic code of a client in order to fix a risk-adjusted premium. From a legal point of view, this problem is dealt with in very different ways even within Europe, not to speak of the rest of the world. Besides this aspect, it was pointed out that also from the medical perspective, statistically significant prognoses on a possible outbreak of diseases related to gene defects are dubious, in particular if multi-factorial gene disorders have to be considered. Family history is another source of information that can lead to different conclusions here.

These statements were strongly supported by Jens Reich (Humboldt-Universität Berlin) with his lecture on "Living is a risky endeavour – less so through genetic medicine?". The audience was informed in detail about the biological foundations of the human genome and various aspects of "cloning". Special emphasis was given to the legal problems with stem cell research in particular in Germany, where such topics are still discussed in a quite controversial way.

The closing plenary lecture of the last day was presented by Harry Panjer (University of Waterloo, Canada) with the title "Development of international insurance company solvency standards", finding thus a way back to the topic of the opening lecture of the first day. Perspectives for future actuarial activities especially for the development and unification of tools for regulatory authorities were outlined, such as risk measures and their properties (VaR, coherent risk measures, TailVaR), stochastic modeling of dependencies by copulas and an improvement of the "covariance formula" for the valuation of risk based capital (RBC).

Besides the plenary lectures, various other contributions were organized in working sessions. In the order of presentations, these were:

- Techniques for valuation a general insurance company within the framework of IAS standards: some proposals (by Aurélie Despeyroux, Charles Levi, Christian Partrat and Jerôme Vignancour)
- Asbestos: The current situation in Europe (by Laura Salvatori, Alessandro Santoni and Darren Michaels)

- Munich Chain Ladder – Closing the gap between paid and incurred IBNR-estimates (by Gerhard Quarg, with an additional comment by Thomas Mack)
- Capital and Asset Allocation (by René Schnieper)
- Stochastic orders in dynamic reinsurance markets (by Thomas Møller)
- Risk Exchange with distorted probabilities (by Andreas Tsanakas)
- Multidimensional Credibility applied to estimating the frequency of big claims (by Hans Bühlmann, Alois Gisler and Denise Kollöffel)
- Credibility weighted hazard estimation (by Jens Perch Nielsen and Bjørn Lunding Sandqvist)
- Marketing and Bonus-Malus Systems (by Sandra Pitrebois, Michel Denuit and Jean-Francois Walhin)
- Insurance applications of near-extremes (by Enkelejd Hashorva and Jürg Hüsler)
- Windstorm claims dependence and copulas (by Olivier Belguise and Charles Levi)
- Tail distribution and dependence measures (by Arthur Charpentier)
- Robust inference in rating models (by Gilles Dupin, Alain Monfort and Jean-Pierre Verle)
- Basis risk and cat risk management (by Frank Krieter)
- Copula: A new vision for economic capital and application to a four line of business company (by Fabien Faivre)
- Effets de la dépendance entre différentes branches sur le calcul des provisions (by Antonin Gillet and Benjamin Serra)
- Modeling and generating dependent risk processes for IRM and DFA (by Dietmar Pfeifer and Johana Neslehová)
- Capital allocation survey with commentary (by Gary G. Venter).

Poster contributions were (in alphabetical order of the authors):

- Fair Value of Life Liabilities with Embedded Options: An Application to a Portfolio of Italian Insurance Policies (by Giulia Andreatta and Stefano Corradin)
- A Comparison of Strategic Reinsurance Programs (SRP) with Banking Activities and Other Insurance and Reinsurance Activities (by Baruch Berliner)
- Dynamic Asset Liability Management: A Profit Testing Model for Swiss Pension Funds (by Ljudmila Bertschi, Sven Ebeling and Andreas Reichlin)
- Economic Risk Capital and Reinsurance: An Extreme Value Theory's Application to Fire Claims of an Insurance Company (by Stefano Corradin)
- On the Distribution of the Deficit at Ruin and the Surplus Prior to Ruin in the Compound Binomial Model (by Esther Frostig)
- The Impact of Statistical Dependence on Multiple Life Insurance Programs (by Esther Frostig and Benny Levikson)
- Optimal Dividend Payment under a Ruin Constraint: Discrete Time and State Space (by Christian Hipp)
- The Impact of Reinsurance on the Cost of Capital (by Werner Hürlimann)
- Optimality of a Stop-Loss Reinsurance in Layers by Werner Hürlimann)

- The Czeledin Distribution Function (by Markus Knecht and Stefan Küttel)
- On the Loading of Largest Claims Reinsurance Covers (by Erhard Kremer)
- Exposure Rating in Liability Reinsurance (by Thomas Mack and Michael Fackler)
- Credibility Evaluation for Heterogeneous Populations (by Udi E. Makov)
- On a Non-Linear Dynamic Solvency Control Model (by Vsevolod Malinovskii)
- Capital Consumption: An Alternative Methodology for Pricing Reinsurance (by Donald Mango)
- A Stochastic Control Model for Individual Asset-Liability Management (by Sachi Purcal)
- On Error Bounds for the Approximation of Random Sums (by Bero Roos and Dietmar Pfeifer)
- A Risk Charge Calculation Based on Conditional Probability (by David Ruhm and Donald Mango)
- Conditional Risk Charge Demo using DFAIC (by David Ruhm and Donald Mango)
- A Risk Theoretical Model for Assessing the Solvency Profile of a General Insurer (by Nino Savelli)
- On Unknown Accumulations in Accident Insurance: An Upper Bound of the Expected Excess Claim (by Hans Schmitter)
- Actuarial Principles of the Cotton Insurance in Uzbekistan (by Bakhodir Shamsuddinov)
- The Estimation of Market VaR using Garch Models and Heavy Tail Distributions (by Ricardo A. Tagliafichi)
- Fit to a t – Estimation, Application and Limitations of the t-Copula (by Gary G. Venter)
- Une Nouvelle Caractérisation de la Distribution de Pareto, avec Application à la Cadence de Paiement du Réassureur en Excédent de Sinistre (by Jean-Francois Walhin).

The closing ceremony of the Colloquium was performed by Edward Levay and Dieter Köhnlein.

All presentations (invited lectures, working papers and posters) can be downloaded from the Colloquium website at www.astin2003.de.

Dietmar Pfeifer

Report on the International AFIR Colloquium 2003, Maastricht, The Netherlands

From 17 to 19 September, 2003, Maastricht was the venue for the annual AFIR Colloquium. In this report you will find impressions of both the content and the surrounding social activities.

Participation

The number of participants was 154 with 23 accompanying persons. There was a good spread over 22 countries. Apart from The Netherlands, which as a host country not surpisingly had the highest number (52), the Nordic countries stood out with 43 participants. Worthwhile mentioning also is Slovenia with 14 participants and the fact that seven countries were each represented by 1 participant.

Invited lecturers

Two invited lecturers had been invited. Firstly, Mr Luc Henrard, Chief Risk Officer of FORTIS, Belgium, gave a lecture under the title "The management of a financial conglomerate: a challenge for the actuaries?". He made an almost passionate plea for actuaries to broaden their horizon to include more expertise outside their traditional role. Secondly, Mr Roderick Munsters, Chief Investment Officer of PGGM, The Netherlands, made a presentation on "The actuary and the investor on a rollercoaster ride" in which he, referring to PGGM's investment policies, equally challenged the actuaries to intensify their cooperation with the "asset specialists". Both presentations tied in very well with the subjects of of the regular papers and the special paper, mentioned below.

Papers

The number of accepted papers was 35. Most of them were presented by the author (or one of the authors) during 30-minute sessions. The variety of subjects was quite interesting and all papers were classified in one of the three predefined categories, i.e. "Asset Classification", "Market valuation of Liabilities" and "Risk Measurement and Management". They are all accessible and downloadable through the website www.afir2003.nl, which will remain open for some time. We hope that a number of them will be published in the ASTIN Bulletin.

Special paper

One non-scientific paper was presented about "the future of AFIR" and the dual role of dealing with both traditional and new areas. The authors made a plea for a reorientation of AFIR's "mission". In their view, AFIR should become more active in stimulating actuaries to become "risk officers" in a far more general sense. Of course this would include adjustments to the educational syllabuses. The paper also described — on the basis of a survey in 10 countries — the extremely diverse way in which presently AFIR activities are organised around the world.

General meeting and Bob Alting von Geusau Memorial Prize

On Friday afternoon the General Meeting for AFIR members was held to discuss accounts, Committee membership nominations and some other topics of a more formal nature. The Chairman also announced that the discussion on the future of AFIR would have a follow-up during the next Colloquium and that several sub-committees had been appointed to prepare for this. Following the General Meeting, a ceremony took place to present a Prize for the best article, published in the ASTIN Bulletin during the years 2001 and 2002. It was the first time this new annual Prize, named after one of the "founding fathers" of AFIR in the seventies, was awarded. The winner was Shaun Wang for his paper "A framework for pricing financial and insurance risks" in the ASTIN Bulletin 32.2 of November 2002.

Social program

A number of social activities had been arranged for participants and companions to liaise with others "outside the world of formulae". On the opening day (Wednesday) a tour of the surrounding area of Maastricht was made, ending with a reception at Maastricht's Town Hall. There was a companion's program on Thursday and in the evening all guests were invited to an informal dinner in the city. Finally, there was a formal farewell dinner on Friday night in one of the castles in Maastricht's province of Limburg.

Next Colloquium

During the General Meeting the venue for the next Colloquium was disclosed: USA, (probably New York) in November 2004. On behalf of the Organising and Scientific Committees I take pleasure in wishing our successors the very best in preparing for that event!

Hugo Berkouwer
chairman of the Organising Committee, Maastricht 2003

International Conference on "Dependence Modelling: Statistical Theory and Applications to Finance and Insurance" (DEMOSTAFI)
20-22 May 2004 in Quebec City, Canada

This conference is a sequel to the series of conferences on copulas, dependence models and their applications that were held in Rome (1990), Seattle (1993), Prague (1996) and Barcelona (2000).

The purpose of this conference is to bring together researchers interested in modelling stochastic dependence and measuring its effects in statistics, actuarial science and finance. The meeting aims to attract copula specialists and statistical researchers interested in their development and use in characterizing and modelling of dependence (stochastic orderings, distributions with fixed marginals, etc.). We would also like to put emphasis on applications of the relevant concepts and inferential techniques in the fields of actuarial science and finance, which are thriving at present. The "technological transfer" aspect of the conference will be especially important; for example, several survey talks by world specialists have been planned.

For additional information about this meeting, registration material and so on, please visit the conference website at http://www.fsa.ulaval.ca/demostafi/

Etienne Marceau, Ph. D., A.S.A.
Member of the organizing committee
Associate Professor
Ecole d'Actuariat
Laval University
Quebec (Que)
Canada, G1K 7P4
Phone: (418) 656-2013
Fax: (418) 656-7790
email: emarceau@act.ulaval.ca
website: http://hyperion.act.ulaval.ca/~emarceau/

XXXV INTERNATIONAL ASTIN COLLOQUIUM
Sunday June 6 to Wednesday June 9, 2004

The 35^{th} International ASTIN Colloquium will be held in Bergen, Norway. The Norwegian Actuarial Society extends its most sincere welcome to all participants and partners to join us in what will surely become days full of meaning both professionally and socially.

Call for papers

We invite authors to submit papers on any subject covered by ASTIN; see the "Call for papers" entry on the website. Deadline for submitting scientific papers in their final form: **February 1, 2004**.

Scientific Program committee

Erik Bølviken (chairman), Paul Embrechts, Simen Gaarder, Angus MacDonald, Ragnar Nordberg, Mette Rytgaard.

Key notes lectures

Opening by prominent actuary in the insurance industry.
Closing by prominent academic on actuarial science in 21^{th} century.

Organized sessions

Thematic sessions have been organized on

- Insurance fraud
- Genetics and insurance
- Climatic change and its impact on insurance

Coordinated lectures will be given by actuaries and by relevant people from the outside. Details: See conference website

Conference venue

The colloquium will take place in Bergen (Norway's second largest city), located on the west coast at the entry of magnificent fjords and waterfalls. The town, still marked by its hanseatic origin, was for centuries the commercial capital of the country. Today Bergen is the bustling home of two universities, and became a European city of culture three years ago. Karl Borch, co-founder of modern risk theory, resided here.

www.astin2004.no

The website gives you up-to-date information.
Here you will find the preregistration schedule.

The Department of Statistics & Actuarial Science of the University of the Aegean is pleased to host the 3rd Conference in Actuarial Science and Finance, to be held on Samos, on September 2-5, 2004.

This event is jointly organized with the Katholieke Universiteit Leuven (Department of Applied Economics and Department of Mathematics), the Université Catholique de Louvain (Institute of Statistics and Actuarial research group) and the University of Copenhagen (Laboratory of Actuarial Mathematics).

The Conference allows the presentation of the latest works in the area of actuarial science and finance. It is open to all persons interested in actuarial science and finance, be they from universities, insurance companies, banks, consulting firms or regulatory authorities. The conference aims to facilitate the contact and the communication between the practicians and the researchers.

The topics of the sections include:

– Extremes and Large Deviations in Actuarial Science – Chair J. Teugels

– Non-life Insurance – Chair R. Verrall

– Advances in Incomplete Markets – Chair Th. Zariphopoulou

– Modelling Dependence in Actuarial Science – Chair Th. Mikosch

– Risk and Control – Chair S. Asmussen

– Life, Pension and Health Insurance – Chair H. Gerber

There will be four short courses. Two before the conference: 30th of August - 1st of September

– Stochastic Claims Reserving, by R.J. Verrall

– Stochastic Control Applied to Actuarial Problems, by H. Schmidli

and the other two after the conference: 6th September - 8th September

– Risk Measures and Optimal Portfolio Selection (with applications to elliptical distributions), by J. Dhaene and E. Valdez

– Advanced Statistical Methods for Insurance, by M. Denuit

Postgraduate students and young reseachers are specially welcome.

Scientific Committee:

Asmussen Soeren
Denuit Michel
Foss Serguei

Frangos Nicos
Gerber Hans
Goovaerts Marc
Guillen Montserrat
Konstantinides Dimitrios
Makov Udi
Marceau Etienne
Mikosch Thomas
Ng Kai
Papaioanou Takis (Chair)
Tessaromatis Nicos
Teugels Jef
Verrall Richard
Willder Mark
Zariphopoulou Thaleia

Organizing Committee:

Dhaene Jan
Frangos Nicos (Chair)
Konstantinides Dimitrios
Mikosch Thomas
Purcaru Oana
Schmidli Hanspeter
Teugels Jef
Walhin Jean-Francois

Local Committee:

Chatzispyros Spyridon
Giannakopoylos Athanasios
Katsis Athanasios
Konstantinides Dimitrios (Chair)
Milionis Alexandros
Nakas Christos
Poufinas Thomas

For further information, please refer to
http://www.stat.ucl.ac.be/Samos2004/

GEORGIA STATE UNIVERSITY
J. Mack Robinson College of Business

Two Faculty Positions in Risk Management

The Robinson College of Business at Georgia State University invites applications for two tenure-track positions at the assistant professor level to begin in fall 2004. These two hires are the first of six positions for which we anticipate hiring during the next three years for the purpose of forming a cross-disciplinary group charged to conduct fundamental research on the economics of uncertainty and on the management and pricing of risk.

JOB QUALIFICATIONS:

Qualified candidates will be expected to possess a PhD by the time of their appointment. All areas of specialization will be considered provided the candidate has a strong interest in and academic background in some area of risk, broadly defined. We are particularly interested in individuals whose studies involve dynamic asset pricing, computational methods including financial econometrics, statistics or actuarial science, equilibrium theory or the microeconomics of uncertainty, optimal contracting, and the estimation theory for dynamic games and dynamic contracting. The ideal candidate will be expected to publish in the major journals of economics and finance as well as important journals of their specific discipline.

ABOUT THE ENVIRONMENT:

The Department of Risk Management and Insurance houses one of the oldest and most influential risk management programs in the U.S. Beginning in academic year 2003, and continuing in 2004 and 2005, the department intends to recruit six new faculty members trained in the most advanced methods of their disciplines for a coordinated study of risk management problems at their most fundamental levels. Salaries will be competitive and strong research support will be provided. Teaching loads will be low to ensure the group has the opportunity to produce high quality research. A statement outlining the vision the department has for the group is available upon request from the co-chairs of the recruiting committee.

FURTHER INFORMATION AND APPLICATION PROCEDURE:

Applicants should send a current curriculum vita, three letters of recommendation, and recent publications or working papers for review. Applications should be submitted electronically via e-mail as pdf files to rphillips@gsu.edu. Applicants wishing to send their materials via regular mail should forward them to:

Richard D. Phillips, Co-chair
Department of Risk Management and Insurance
Robinson College of Business
Georgia State University
P.O. Box 4036
Atlanta, GA 30302-4036

For further information, contact either co-chair of the search committee: Martin Grace - mgrace@gsu.edu or 404-651-2789; Richard Phillips - rphillips@gsu.edu or 404-651-3397.

Interviews can be scheduled for the 2003 Financial Management Association Annual Meeting or the 2004 Allied Social Sciences Association Annual Meeting. Preference will be given to applications received by **December 1, 2003**.

Georgia State University is an equal opportunity educational institution/affirmative action employer and encourages applications from qualified minorities. All Positions are Subject to Final Approval for Funding.

DIRECTOR, Actuarial Science program
Department of Risk Management and Insurance
Georgia State University

JOB QUALIFICATIONS:

The Department of Risk Management and Insurance invites applications for the position of Director of the Actuarial Science Program. The successful candidate will have an established record of high-quality research in the field of actuarial science, statistics, or related field such as financial mathematics and a demonstrated ability to lead one of the outstanding actuarial science programs. A doctorate in actuarial science, finance, mathematics, statistics, or related field is required. This tenure track position, to be filled effective fall 2004 at the rank of associate or full professor, requires maintenance of a successful research agenda and demonstrated teaching excellence.

Membership in a professional actuarial organization is required. These may include the Casualty Actuarial Society, the Society of Actuaries, the Institute of Actuaries, the Faculty of Actuaries, or the Australian Institute of Actuaries. Associates of these societies or members of other actuarial societies may also meet this requirement if they have exceptional strength in scholarship, teaching, and leadership.

ABOUT THE ENVIRONMENT:

From its founding in 1958, the GSU Actuarial Science Program has been one of the leading programs in North America. It has a distinguished history of serving students, alumni, and the actuarial profession. Hundreds of our graduates have become Fellows or Associates of the Society of Actuaries and/or the Casualty Actuarial Society and many have become leaders in the professional and business communities. The program is housed in the Department of Risk Management and Insurance, regarded as one of the best departments of its type internationally.

The department recently expanded its research and educational mission to include mathematical risk management, offering masters and doctoral degrees oriented at the intersection of actuarial science and mathematical finance. The Mathematical Risk Management and Actuarial Science Programs work closely in student recruitment and placement. Some courses are cross-listed and many actuarial students take mathematical risk management courses, such as financial engineering, as electives.

FURTHER INFORMATION AND APPLICATION PROCEDURE:

Preference will be given to applications received by December 1, 2003. Applicants should send a current curriculum vita, three letters of recommendation, and recent publications or working papers to:

Richard D. Phillips, Chair
Search Committee
Department of Risk Management and Insurance
J. Mack Robinson College of Business
Georgia State University
P.O. Box 4036
Atlanta, GA 30302-4036
Tel: 404-651-3397

Applications may be submitted electronically via e-mail at: rphillips@gsu.edu. For further information, contact the chair of the search committee.

Georgia State University is an equal opportunity educational institution/affirmative action employer and encourages applications from qualified minorities. Position is subject to final approval for funding.

Assistant Professor in Actuarial Science
University of Iowa

Applications are invited for a tenure-track assistant professor in actuarial science starting August 2004. Applicants must show promise for excellence in both teaching and creative research. They must have completed a Ph.D. in a relevant field and at least the first four Society of Actuaries' exams, or equivalent exams in a major actuarial organization. Fellowship or Associateship in a professional actuarial society is preferred. The appointee is expected to conduct research in actuarial science/financial mathematics, to assist in building a Ph.D. program in this area, and to supervise Ph.D. students.

The selection process begins December 1, 2003 and continues until the position is filled.

Please send a curriculum vitae, a transcript for new Ph.D.s, and have three confidential letters of reference sent to:

Actuarial Search Committee
Dept. of Statistics & Actuarial Science
University of Iowa
Iowa City, IA 52242.

Email: actuarial-search@stat.uiowa.edu

The Department currently has seventeen tenure-track faculty who are engaged in various areas of research in statistics, actuarial science, and financial mathematics. The current actuarial faculty members are Jim Broffitt, A.S.A., Gordon Klein, F.S.A., and Elias Shiu, A.S.A. The B.S. and M.S. degrees are offered in both actuarial science and statistics, the Ph.D. in statistics. Actuarial students may earn a Ph.D. in statistics with emphasis in actuarial science/financial mathematics. The number of actuarial science majors is about 45 graduate and 20 undergraduate students. For additional information about the Department, please refer to the website: http://www.stat.uiowa.edu/.

The University of Iowa is nestled in the rolling hills of eastern Iowa along the banks of the Iowa River. Approximately 30,000 students are enrolled in eleven colleges: Liberal Arts, Graduate, Business, Law, Medicine, Public Health, Dentistry, Nursing, Pharmacy, Education, and Engineering. The University is known for its fine arts, and a variety of touring dance, musical, and theatrical groups perform on campus each year. As a member of the Big Ten Conference, Iowa hosts many outstanding athletic events.

Iowa City is a clean, attractive community of approximately 62,000 people. It is noted for its public schools, medical and athletic facilities, attractive business district, parks, and mass transit system. In 1999, editor & Market guide rated Iowa City as the best metropolitan area to live in the USA. Among smaller metropolitan areas, the 2003 Milken Institute Best Performing Cities Institute ranked Iowa City number 1. Iowa City is within 300 miles of Chicago, St. Louis, Kansas City, and Minneapolis.

Women and minorities are encouraged to apply. The University of Iowa is an Affirmative Action Equal Opportunity Employer.

Drake University
College of Business and Public Administration
Des Moines, IA 50311, USA

POSITION:

A tenure track position in actuarial science in the College of Business and Public Administration, to begin August, 2004, pending final budgetary approval. Rank and salary based on qualifications.

DUTIES:

Teach six courses per year; recruit, advise, and place students; conduct scholarly research; and serve the University and the profession.

QUALIFICATIONS:

Ph.D. in actuarial science or a related area along with Associateship or Fellowship in the CAS or SOA is preferred. Candidates with lesser qualifications will be considered if there are compensating factors.

APPLICATIONS:

Submit a curriculum vitae and arrange for three letters of reference to be sent to Professor Stuart Klugman, F.S.A.; CBPA; Drake University; Des Moines, IA 50311. Applications will be accepted until the position is filled. Drake University is an equal opportunity/affirmative action employer and actively seeks applications from women and minority group members who are qualified for this position.

Stuart Klugman, F.S.A., Ph.D.
Principal Financial Group Professor of Actuarial Science Drake University
Des Moines, IA 50311
515-271-4097
E-mail: stuart.klugman@drake.edu

CIBC Chair in Financial Risk Management
UNIVERSITY OF WATERLOO

The University of Waterloo (UW) has one of the most eminent actuarial science programs in the world. The faculty members in this unit have attained a high level of distinction in their research, teaching and professional contributions. The graduates of the Waterloo program are internationally recognized. In recent years, UW has also developed a strong research presence in the field of modern finance, with a special emphasis on computational finance. More than 75 new bachelor's and 20 master's and doctoral graduates enter the insurance and finance industries each year from the undergraduate and graduate actuarial science programs, and the master's program in quantitative finance co-ordinated by the Centre for Advanced Studies in Finance.

The University has recently created the Institute for Quantitative Finance and Insurance (IQFI) to combine the strengths of these two disciplines and provide the vehicle for a major research and teaching thrust in the area of financial risk management broadly defined. This initiative was made possible through support from companies in the insurance and financial services industries, and matching contributions from the Province of Ontario through the Ontario Research and Development Challenge Fund. The Institute will advance research in the financial risk management area and disseminate new knowledge. A generous contribution by the Canadian Imperial Bank of Commerce to UW's capital campaign, "Building a Talent Trust", has enabled the University to inaugurate two CIBC Chairs in Financial Risk Management under the auspices of the IQFI. One of these Chairs will be located in the Department of Statistics and Actuarial Science.

The purpose of this Chair is

- to help the University of Waterloo enhance its leadership role in finance and insurance education and research, and to expand its expertise by building on its present strengths
- to attract students of the highest calibre, and to support and supplement faculty influence on professional education and current practice
- to give prominence and recognition to the interaction between the university and the insurance and financial services industries, particularly the support provided by CIBC
- to support applied research, and the transfer of basic research into current practice in the financial services and insurance industries

The Canadian Imperial Bank of Commerce is a leading North American financial institution. Through its comprehensive electronic banking network, branches and offices across Canada and around the world, CIBC offers a full range of products and services to more than nine million personal banking and business customers.

The appointment is for a period of up to five years, with expectation of renewal. The anticipated start date is July 1, 2004.

The duties of the Chair holder include:

– conducting and overseeing a program of research relevant to financial risk management, actuarial science or insurance
– undergraduate and graduate teaching; leading seminars and colloquia that involve both undergraduate and graduate students
– disseminating applied research through seminars and professional meetings

The Chair holder will be a member of the Department of Statistics and Actuarial Science in the Faculty of Mathematics at the University of Waterloo and play a leadership role in the Institute for Quantitative Finance and Insurance. The ideal candidate will have earned a PhD in an appropriate field of research within the last few years, and will already have a strong record as a researcher, with exceptional promise for distinction in some aspect of actuarial science, finance or closely related discipline. The successful candidate must possess strong communication skills and be an excellent teacher.

Please send applications and nominations, including a recent curriculum vitae, to Professor Alan George, Dean, Faculty of Mathematics, University of Waterloo, Waterloo, ON Canada N2L 3G1 by April 30, 2004.

In accordance with Canadian immigration requirements, citizens and permanent residents of Canada will be considered first for this position. The University of Waterloo encourages applications from all qualified individuals including women, members of visible minorities, native peoples, and persons with disabilities.

GUIDELINES TO AUTHORS

1. Papers for publication should be sent in quadruplicate to one of the Editors:

Andrew Cairns
Department of Actuarial Mathematics
and Statistics
Heriot-Watt University
Edinburgh EH14 4AS, United Kingdom
A.Cairns@ma.hw.ac.uk

Paul Embrechts
Department of Mathematics
ETHZ
CH-8092 Zurich, Switzerland.
embrechts@math.ethz.ch

Submission of a paper is held to imply that it contains original unpublished work and is not being submitted for publication elsewhere.
Receipt of the paper will be confirmed and followed by a refereeing process, which will take about three months.

2. The basic elements of the journal's style have been agreed by the Editors and Publishers and should be clear from checking a recent issue of *ASTIN BULLETIN*. If variations are felt necessary they should be clearly indicated on the manuscript.

3. Papers should be written in English or in French. Authors intending to submit longer papers (e.g. exceeding 30 pages) are advised to consider splitting their contribution into two or more shorter contributions.

4. The first page of each paper should start with the title, the name(s) of the author(s), and an abstract of the paper as well as some major keywords. An institutional affiliation can be placed between the name(s) of the author(s) and the abstract.

5. Footnotes should be avoided as far as possible.

6. References should be arranged alphabetically, and for the same author chronologically. Use a, b, c, etc. to separate publications of the same author in the same year. For journal references give author(s), year, title, journal (in italics, cf. point 9), volume (in boldface, cf. point 9), and pages. For book references give author(s), year, title (in italics), publisher, and city.
Examples
BARLOW, R.E. and PROSCHAN, F. (1975) *Mathematical Theory of Reliability and Life Testing*. Holt, Rinehart, and Winston, New York.
JEWELL, W.S. (1975a) Model variations in credibility theory. In *Credibility: Theory and Applications* (ed. P.M. KAHN), pp. 193-244, Academic Press, New York.
JEWELL, W.S. (1975b) Regularity conditions for exact credibility. *ASTIN Bulletin* **8**, 336-341.
References in the text are given by the author's name followed by the year of publication (and possibly a letter) in parentheses.

7. The address of at least one of the authors should be typed following the references.

8. Italics (boldface) should be indicated by single (wavy) underlining. Mathematical symbols will automatically be set in italics, and need not be underlined unless there is a possibility of misinterpretation. Information helping to avoid misinterpretation may be listed on a separate sheet entitled 'special instructions to the printer'. (Example of such an instruction: Greek letters are indicated with green and script letters with brown underlining, using double underlining for capitals and single underlining for lower case).

9. Contributions must be typewritten on one side of good quality paper, with double spacings and ample margins all round.
Illustrations should be submitted as clear black and white prints. Photocopies are not acceptable. Line thickness and lettering size should be adopted to suit any likely degree of reduction.
Each contribution should obtain the author(s) full address(es), including e-mail and fax numbers when available. Authors should notify the publisher whenever their contact details change.
A copy of text should also be submitted on disk, when available. Please state clearly the type of software used, and note the filename. Contributors should also retain a copy of their article.
Two sets of proofs will be shipped to authors who should ensure that one set plus the manuscript is returned to PEETERS within one week of receipt. Authors may be charged for alterations to the original manuscript. If authors proofs are not returned by the required date, the publisher's own corrected set will be forwarded to the printer.

10. Authors will receive 30 offprints free of charge. Additional offprints may be ordered when returning corrected proofs. A scale of charges will be enclosed when the proofs are sent out.