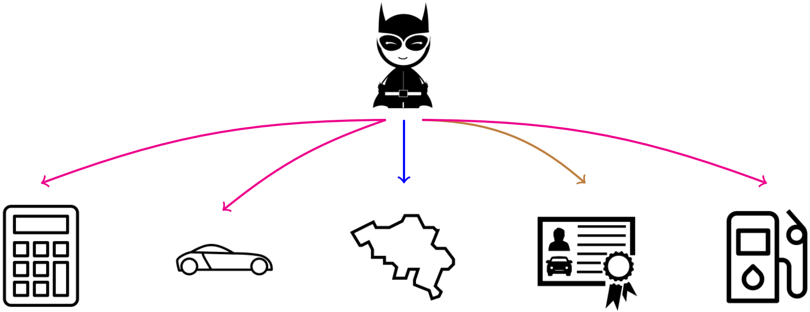
Using risk factors in insurance analytics: data driven strategies

ICA 2018, Berlin

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

June 4, 2018

Motivation



Claim frequency and claim severity

as function of

nominal / numeric ∼ ordinal / spatial

features

Research questions

- Generalized Linear Models (GLMs) for frequency ($\sim$ Poisson) and severity ($\sim$ gamma).
- How to:
  (1) select risk factors or features?
  (2) cluster (or bin or fuse) levels within a risk factor?
      age groups / postal code clusters / clusters of car models
- Procedure should be data driven, scalable to large (big) data.
- End product is interpretable, within actuarial comfort zone.

Research questions - rephrased

- Generalized Linear Models (GLMs) for frequency ($\sim$ Poisson) and severity ($\sim$ Gamma).
- How to:
  (1) avoid overfitting with too many risk factors or levels?
  (2) avoid underfitting with a priori binning/selection?
- Procedure should be data driven, scalable to large (big) data, and automatic!
- End product is interpretable, within actuarial comfort zone.

A data driven strategy for the construction of insurance tariff classes

Henckaerts, Antonio, Clijsters & Verbelen, 2018, Scandinavian Actuarial Journal

A step-by-step solution

Henckaerts, Antonio, et al., 2018 (Scandinavian Actuarial Journal)

Stepwise procedure

1. Do an exhaustive search through variables to find best GAM.
2. Use well-chosen clustering algorithm to bin 2D spatial effect.
3. Use evolutionary trees to bin 1D continuous effects and interactions.
4. Fit GLM with bins and clusters obtained in previous steps.

R packages: `mgcv`, `classInt`, `evtree`, `rpart`

GAM as a starting point

- Generalized Additive Model with predictor:

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}^d + \sum_{j=1}^{q} f_j(x_{ij}^c) + \sum_{j=1}^{r} f_j(x_{ij}^s, y_{ij}^s).$$

- Information criteria:

$$\begin{aligned} \text{AIC} &= -2 \cdot \log \mathcal{L} + 2 \cdot \text{EDF} \\ \text{BIC} &= -2 \cdot \log \mathcal{L} + \log(n) \cdot \text{EDF}, \end{aligned}$$

  balancing goodness-of-fit and complexity.

- Best subset selection strategy!

GAM as a starting point

- MTPL data set from Denuit & Lang (2004), 163 231 records.
- Lowest BIC among exhaustive search with 1 024 fitted models:

$$\begin{aligned}\log(\mathrm{E}(\texttt{nclaims})) = &\log(\texttt{expo}) + \beta_0 + \beta_1 \texttt{coverage}_{PO} + \beta_2 \texttt{coverage}_{FO} + \beta_3 \texttt{fuel}_{diesel} + \\ &f_1(\texttt{ageph}) + f_2(\texttt{power}) + f_3(\texttt{bm}) + f_4(\texttt{ageph}, \texttt{power}) + f_5(\texttt{long}, \texttt{lat}).\end{aligned}$$

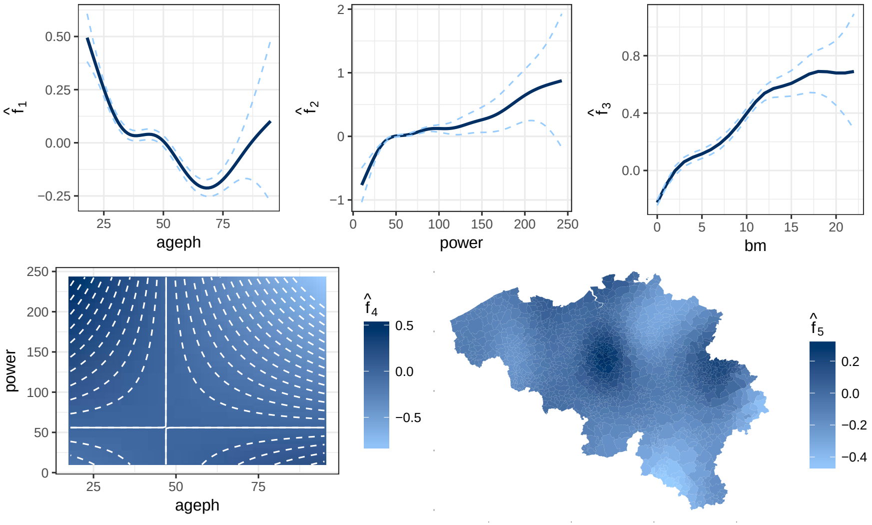
which combines offset and

categorical $\sim$ nominal continuous $\sim$ ordinal

interactions spatial

risk factors.

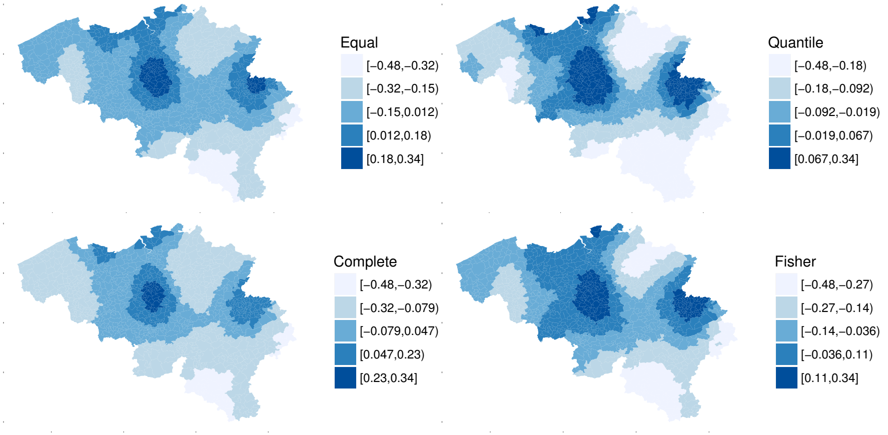
GAM as a starting point



Bin smooth GAM effects - spatial

- Bin or cluster $\hat{f}_5(\texttt{long}_i, \texttt{lat}_i)$ for $i \in \{1, ..., 1\ 146\}$.
- We use: (see `classint` package in `R`)
  - equal intervals;
  - quantile binning;
  - complete linkage (see Kaufman & Rousseeuw, 1990)
  - Fisher's natural breaks (see Fisher, 1958 and Slocum et al., 2005).
- Tune number of clusters towards lowest AIC for the GAM with a binned spatial effect

Bin smooth GAM effects - spatial



Bin smooth GAM effects - continuous

- Bin $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$.
- Use evolutionary trees (`evtree`, Grubinger et al, 2014) e.g. on:

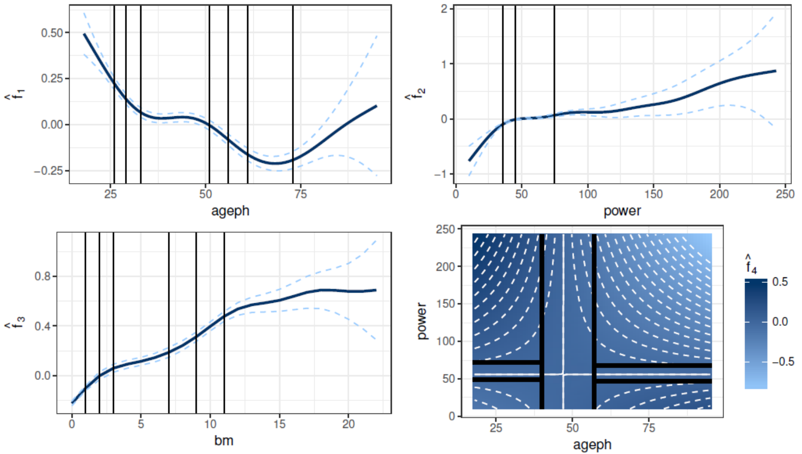
Covariate: ageph	Response: $\hat{f}_1(\texttt{ageph})$	Weight: w
18	0.495	16
19	0.459	116
20	0.424	393

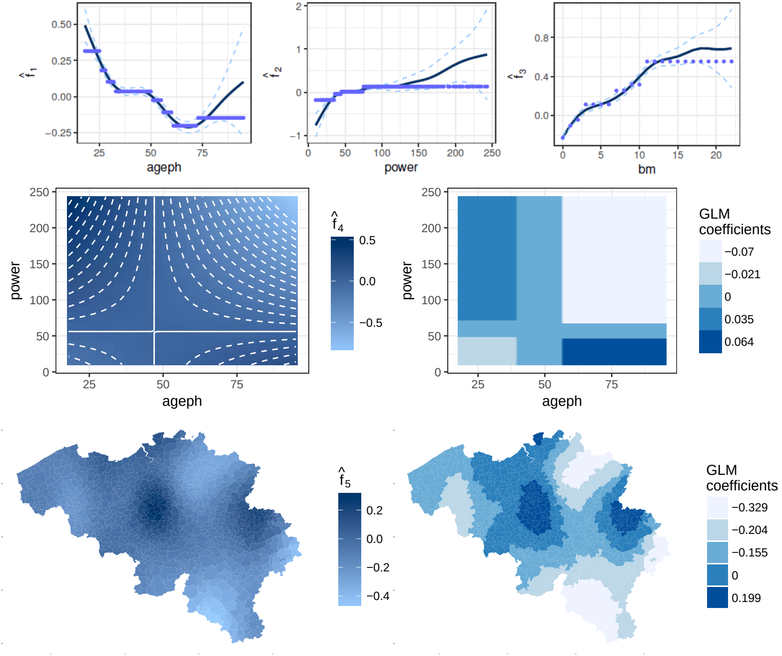
- Evaluate tree with wMSE $+ \alpha \cdot$ complexity penalty,

$$\text{MSE} = \frac{\sum_{i=\min(\texttt{ageph})}^{\max(\texttt{ageph})} w_{\texttt{ageph}_i} (\hat{f}_1(\texttt{ageph}_i) - \hat{f}_1^b(\texttt{ageph}_i))^2}{\sum_{i=\min(\texttt{ageph})}^{\max(\texttt{ageph})} w_{\texttt{ageph}_i}}$$

and α tuning parameter.

Bin smooth GAM effects - continuous





$\hat{f}_1$
ageph
$\hat{f}_2$
power
$\hat{f}_3$
bm
power
ageph
$\hat{f}_4$
0.5
0.0
−0.5
power
ageph
GLM coefficients
−0.07
−0.021
0
0.035
0.064
$\hat{f}_5$
0.2
0.0
−0.2
−0.4
GLM coefficients
−0.329
−0.204
−0.155
0
0.199

Sparsity with multi-type lasso penalties

Devriendt, Antonio, Frees, Reynkens & Verbelen, 2018 (in progress)

Lasso

- Less is more: (Hastie, Tibshirani & Wainwright, 2015)

  a sparse model is easier to estimate and interpret than a dense model.

- Regularize (with budget constraint t, or regularization parameter λ):

$$\min_{\beta_0, \boldsymbol{\beta}} \{-\log \mathcal{L}\} \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq t,$$

  or equivalenty

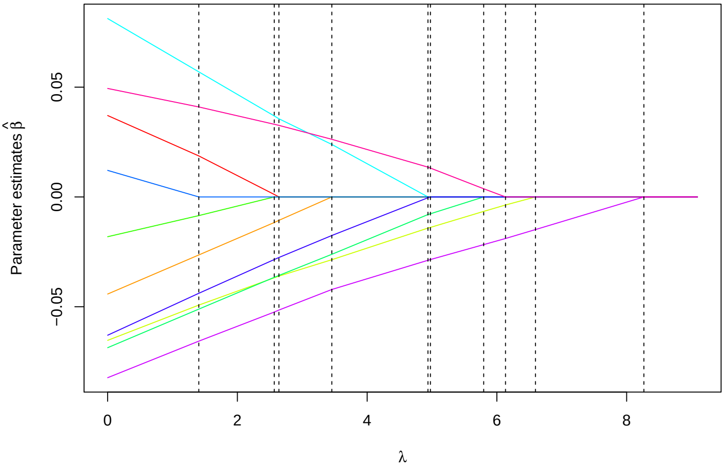
$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ -\log \mathcal{L} + \lambda \cdot \sum_{j=1}^{p} |\beta_j| \right\}.$$

  Shrinks coefficients and even sets some to zero.

A typical lasso plot

`R` package `glmnet`

overfitting $\longleftarrow$ λ $\longrightarrow$ underfitting



Lasso and friends

- Adjust lasso regularization to the type of risk factor:
  - Determine type (nominal / numeric $\sim$ ordinal / spatial)
  - Allocate logical penalty.
- Thus, for J risk factors, each with convex regularization term $g_j(.)$, we want to optimize:

$$-\log \mathcal{L}(\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J) + \lambda \cdot \sum_{j=1}^{J} g_j(\boldsymbol{\beta}_j).$$

A multi-type regularized predictive model!

Multi-type regularization penalty

- Continuous or binary risk factors: lasso

$$g_{\text{Lasso}}(\boldsymbol{\beta}_j) = \sum_i w_{j,i}|\beta_{j,i}|.$$

- Ordinal risk factors: fused lasso

$$g_{\text{fLasso}}(\boldsymbol{\beta}_j) = \sum_i w_{j,i}|\beta_{j,i+1} - \beta_{j,i}|.$$

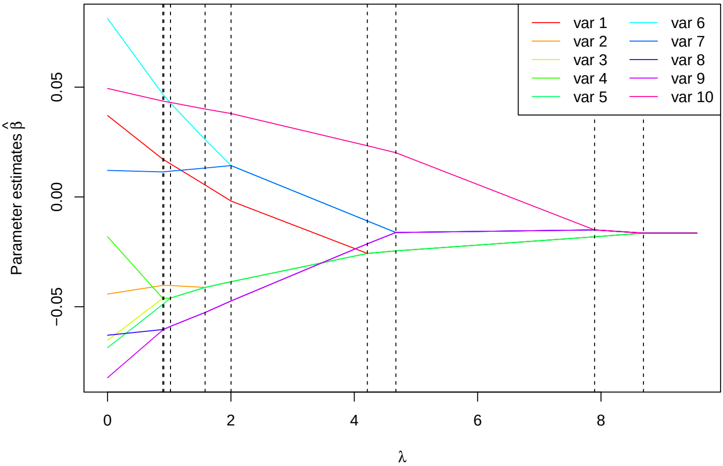
- Nominal risk factors: generalized fused lasso

$$g_{\text{gflasso}} = \sum_{(i,l)\in\mathcal{G}} w_{j,il}|\beta_{j,i} - \beta_{j,l}|.$$

Fused Lasso

`R` package `genlasso`

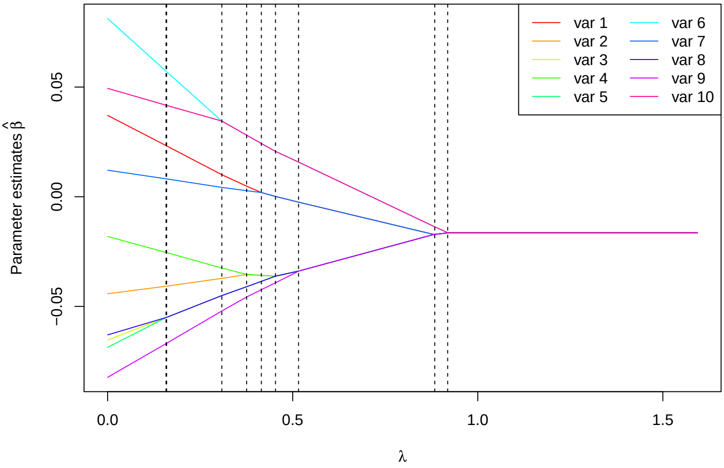
overfitting $\longleftarrow \quad \lambda \quad \longrightarrow$ underfitting



Generalized Fused Lasso

R package `genlasso`

overfitting $\longleftarrow$ λ $\longrightarrow$ underfitting



GLM framework with multi-type penalty

- Gertheiss & Tutz (2010) and Oelker & Gertheiss (2017):
  - GLMs with various penalties.
  - R package available: `gvcm.cat` (not maintained).
- Uses local quadratic approximations of penalties and PIRLS:
  - non-exact selection or fusion;
  - computationally intensive.

Multi-type penalized proximal gradient algorithm

- Our contribution:
  - implements an efficient algorithm (with proximal operators)
    - code bottleneck in C++ (`Rcpp`)
    - efficient linear algebra (`RcppArmadillo`)
    - parallel computations (`parallel`)
  - scalable to big data (splits into smaller sub-problems);
  - flexible regularization
    - penalty takes type of risk factor into account;
    - works for all popular penalties.
- Unifies penalty-specific (machine learning) literature with statistical (or: actuarial) literature!

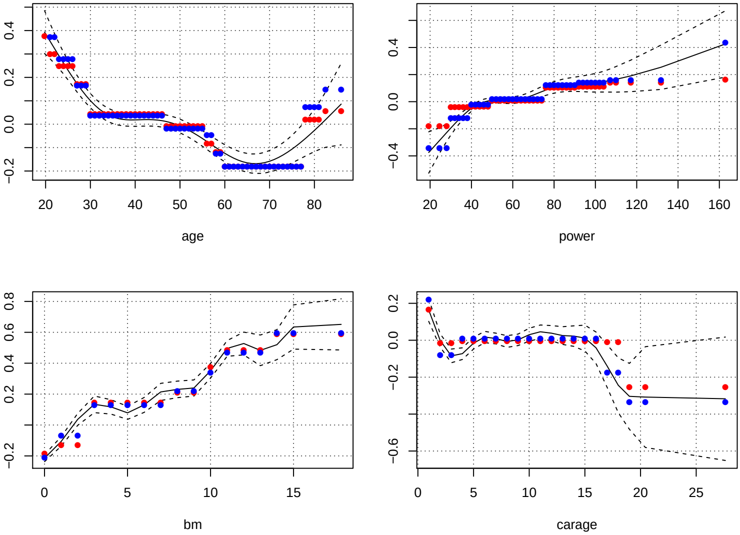
MTPL data: Poisson with multi-type penalty

- Claim frequencies for $n = 163,231$ policyholders.
- 14 risk factors: binary, ordinal, nominal and spatial.
- Exposure included as offset.
- Fit Poisson GLM:
  - $N_i \sim \text{POI}(\mu_i)$
  - $\log(\mu_i) = \log(\text{expo}_i) + \beta_0 + \sum_{j=1}^{14} \boldsymbol{x}_{i,j}\boldsymbol{\beta}_j$
  - $-\mathcal{L}\left(\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{14}\right) + \lambda \cdot \sum_{j=1}^{14} g_j\left(\boldsymbol{\beta}_j\right)$

MTPL data: Poisson with multi-type penalty

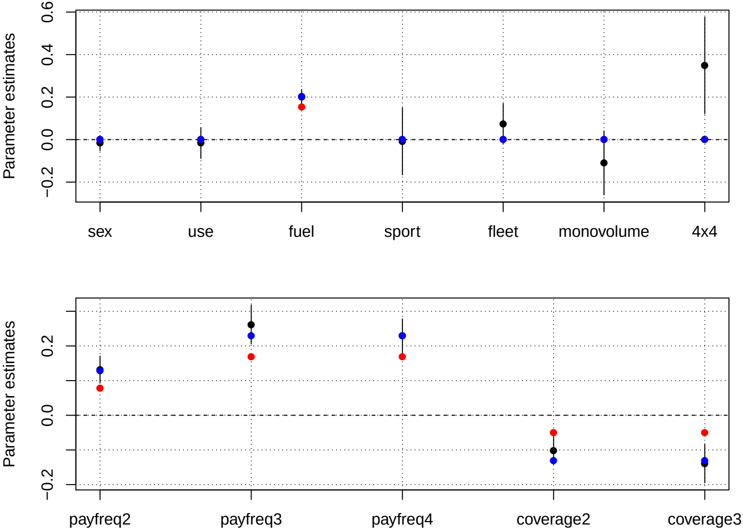
- Settings:
  - incorporate adaptive (GLM) and standardization weights for better consistency and predictive performance.
  - tune λ with out-of-sample deviance ($\hat{\lambda} = 129.36$)
- Re-estimate the final sparse GLM with standard GLM routines (from 266 to 66 params.).

MTPL data: Poisson with multi-type penalty



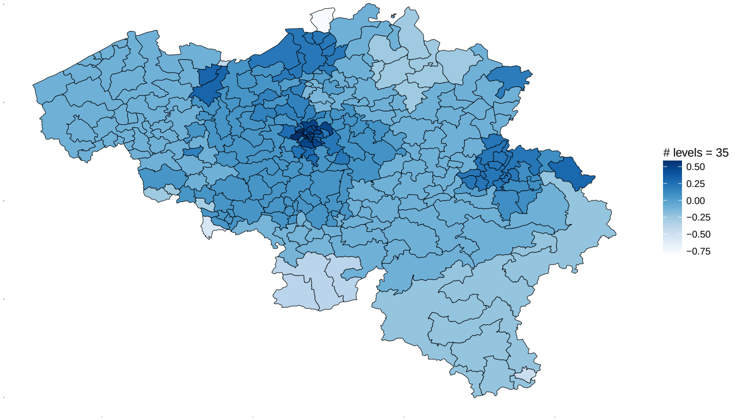
GAM fit, penalized GLM fit, GLM refit with new clusters.

MTPL data: Poisson with multi-type penalty



GAM fit, penalized GLM fit, GLM refit with new clusters.

MTPL data: Poisson with multi-type penalty



Spatial effect, GLM refit with new clusters.

Wrap-up

- From multi-step (published in SAJ, R code upon request) to less is more.
- Flexible regularization can help predictive modeling.
- R package and working paper forthcoming.

More information

For more information, please visit:

- LRisk website, www.lrisk.be
- my homepage, www.feb.kuleuven.be/katrien.antonio.

Thanks to

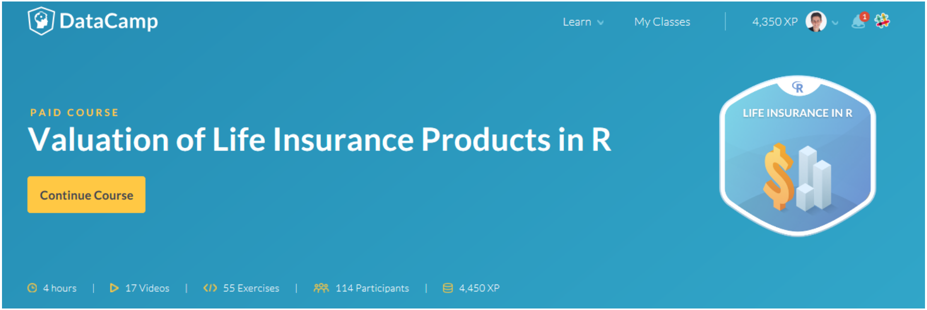
Ageas Continental Europe

Argenta

4th conference of the European Actuarial Journal, in Leuven, September 10 and 11, 2018

with academic and CEO keynotes

www.kuleuvencongres.be/eaj2018



Online course with DataCamp on Valuation of Life Insurance Products in R

designed by Katrien Antonio & Roel Verbelen

http://www.datacamp.com/courses/2333

References

Henckaerts, R., Antonio, K., Clijsters, M. and Verbelen, R. (2018)
A data driven strategy for the construction of insurance tariff classes.
Scandinavian Actuarial Journal

Wood, S. (2006)
Generalized additive models: an introduction with R.
Chapman and Hall/CRC Press.

Gertheiss, J. and Tutz, G. (2010).
Sparse modeling of categorial explanatory variables.
The Annals of Applied Statistics, 4(4), 2150-2180.

Oelker, M. and Gertheiss, J. (2017).
A uniform framework for the combination of penalties in generalized structured models.
Advances in Data Analysis and Classification, 11(1),97-120.

References

Grubinger, T., Zeileis, A., and Pfeiffer, K.-P. (2014).
evtree: Evolutionary learning of globally optimal classification and regression trees in R.
Journal of Statistical Software, 61(1), 1-29.

Bivand, R. (2015).
classInt: Choose Univariate Class Intervals.
R package version 0.1-23.

Parikh, N. and Boyd, S. (2013).
Proximal algorithms.
Foundations and Trends in Optimization, 1(3):123-231.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015)
Statistical learning with sparsity: the Lasso and generalizations.
Chapman and Hall/CRC Press.