www.ICA2014.org

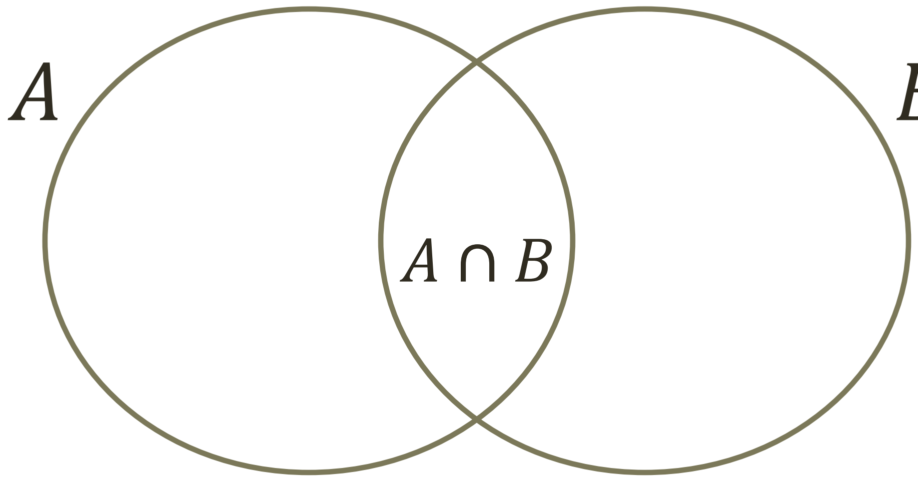
LEARN
INTERACT
GROW

Bayesian Analysis Applications in Actuarial Science Using MCMC Methods: Some Theory

Presented by: Glenn G. Meyers FCAS, CERA
Curtis Gary Dean FCAS, MAAA

Bayes Theorem

$$f(A|B) = \frac{f(A \cap B)}{f(B)} = \frac{f(B|A)f(A)}{f(B)}$$



Bayesian Inference

- n data observations: $X = (X_1, X_2, .., X_n)$
- k model parameters: $\Theta = (\theta_1, \theta_2, .., \theta_k)$

$$f(\Theta|X) = \frac{f(X|\Theta)\pi(\Theta)}{f(X)}$$

- Parameters Θ are random variables
- The distribution of Θ is conditional on data observations X

Likelihood

Prior Distribution

$$f(\Theta|X) = \frac{f(X|\Theta)\pi(\Theta)}{f(X)}$$

Posterior Distribution

$\int f(X|\Theta)\pi(\Theta)d\Theta$

Prior Distribution

- Prior distribution $\pi(\Theta)$
  - Prior knowledge about Θ, or
  - if little or no knowledge, use a diffuse prior.

Likelihood $f(X|\Theta)$

$f(X|\Theta)$: probability (density) of observing $X = (X_1, X_2, .., X_n)$
given parameters $\Theta = (\theta_1, \theta_2, .., \theta_k)$

Frequentist: Find the values of Θ that maximize $f(X|\Theta)$: *maximum likelihood estimation*.

Bayesian Inference: Parameters $\Theta = (\theta_1, \theta_2, .., \theta_k)$ are random variables. Using observations X the probability distribution for Θ is updated using Bayes Theorem.

Bayesian Inference

Posterior → $f(\Theta|X) = \left(\frac{f(X|\Theta)}{f(X)}\right)\pi(\Theta)$ ← **Prior**

- For any Θ, if $f(X|\Theta) > f(X)$ then more probability will be assigned to that Θ in the posterior distribution than in the prior. Note $f(X) = \int f(X|\Theta)\pi(\Theta)d\Theta$ is averaged over all Θ.

- Bayesian inference using data X will shift the distribution, i.e. assign more probability , to values of Θ that are more likely to generate data X

Classic (Tractable) Actuarial Example: Poisson Likelihood, Gamma Prior

- Each risk in population has risk parameter θ
- The number of claims X in one year for risk with parameter θ is Poisson distributed with mean θ

Likelihood

$$\Pr(X = x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}$$

Classic (Tractable) Actuarial Example: Poisson Likelihood, Gamma Prior

- Risk parameter θ is gamma distributed in the population of risks

Prior distribution $$\pi(\theta) = \frac{\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta}}{\Gamma(\alpha)}$$

- For randomly chosen risk from population:

$$E[\theta] = \int_0^{\infty} \theta\pi(\theta)d\theta = \alpha/\beta$$

Classic (Tractable) Actuarial Example: Poisson Likelihood, Gamma Prior

- A risk is randomly chosen from the population

- Without further information about the risk we would infer the expected annual number of claims for risk to be

$$E[X] = E_\theta[E_X[X|\theta]] = E_\theta[\theta] = \alpha/\beta$$

Classic (Tractable) Actuarial Example: Poisson Likelihood, Gamma Prior

- The selected risk is observed to have c claims in one year
- Update distribution of θ using Bayes Theorem

$$f(\theta|X=c) = \frac{\left(\frac{\theta^c e^{-\theta}}{c!}\right)\left(\frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}\right)}{f(c)}$$

Gamma Conjugate Prior for Poisson

- $f(\theta|X=c) = \dfrac{\left(\frac{\theta^c e^{-\theta}}{c!}\right)\left(\frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}\right)}{f(c)}$

$$= constant \cdot \theta^{\alpha+c-1} e^{-(\beta+1)\theta}$$

- Let $\alpha' = \alpha + c$ and $\beta' = \beta + 1$ then

$$f(\theta|X=c) = constant \cdot \theta^{\alpha'-1} e^{-\beta'\theta}$$

- $f(\theta|X=c)$ is gamma so $constant = \beta^{\alpha'}/\Gamma(\alpha')$

Posterior Distribution is Gamma

- $f(\theta|X=c) = \frac{\beta^{\alpha'}\theta^{\alpha'-1}e^{-\beta'\theta}}{\Gamma(\alpha')}$

  and $E[\theta|\text{ c claims in one year}] = \frac{\alpha'}{\beta'} = \frac{\alpha+c}{\beta+1}$

- Given $c = c_1 + \cdots + c_y$ claims in y years

  then $E[\theta|\text{ c claims in y years}] = \frac{\alpha + c}{\beta + y}$

  Prior (α, β) — Data (c, y)

If Prior Distribution is NOT Conjugate Prior for Likelihood

- The posterior distribution is

$$f(\Theta|X) = \left(\frac{f(X|\Theta)}{\int f(X|\Theta)\,\pi(\Theta)d\Theta} \right) \pi(\Theta)$$

- The integral in the denominator must be evaluated.

If Prior Distribution is **NOT** Conjugate Prior for Likelihood

- The posterior mean is:

$$E(\Theta|X) = \int \Theta f(\Theta|X) d\Theta$$

$$= \int \Theta \left(\frac{f(X|\Theta)}{\int f(X|\Theta)\, \pi(\Theta) d\Theta} \right) \pi(\Theta) d\Theta$$

- Numerical integration???

If Prior Distribution is **NOT** Conjugate Prior for Likelihood

- The predictive distribution for future outcomes Y given past outcomes X is

$$f(Y|X) = \int f(Y|\Theta)\left\{\underbrace{\left(\frac{f(X|\Theta)}{\int f(X|\Theta)\,\pi(\Theta)d\Theta}\right)\pi(\Theta)}_{f(\Theta|X)}\right\}d\Theta$$

- How do we perform integrations, especially if there are many parameters?

Posterior Probability Distribution

$$f(\Theta|X) = \left(\frac{f(X|\Theta)}{\int f(X|\Theta)\,\pi(\Theta)d\Theta}\right)\pi(\Theta)$$

In general there is no nice formula for $f(\Theta|X)$ unlike the conjugate prior model.

The integral can be very hard to evaluate, especially if there are multiple parameters in model.

Posterior Probability Distribution

- We want to know properties of the posterior distribution such as its <u>mean</u> or <u>percentiles:</u>

$$\text{Mean: } E(\Theta|X) = \int \Theta f(\Theta|X) d\Theta$$

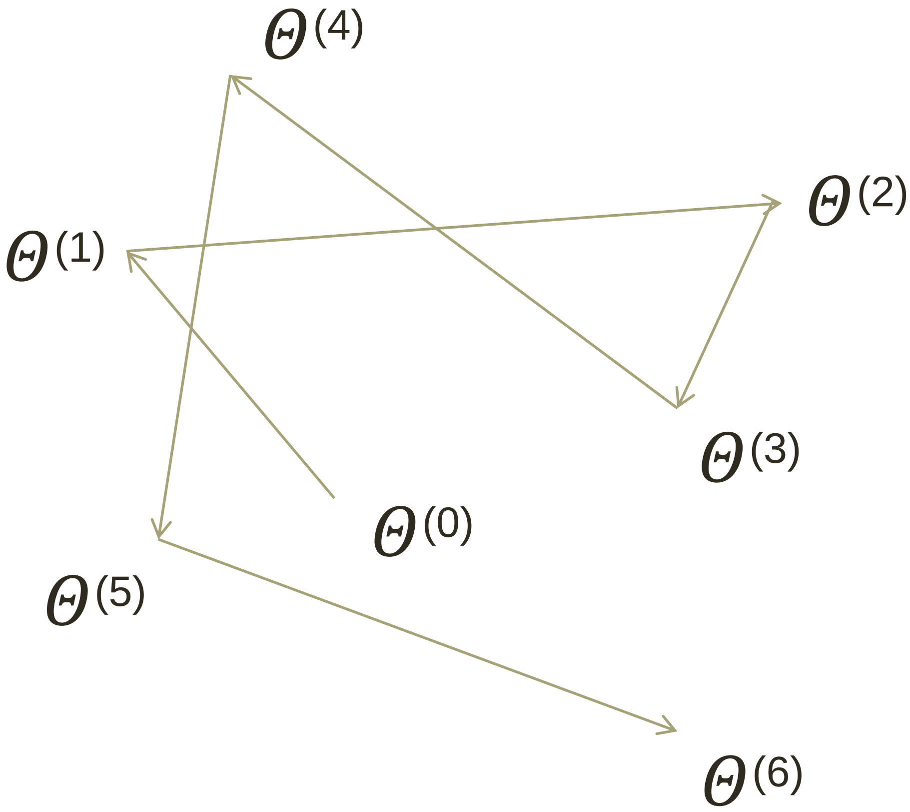
$$(100\,p)^{\text{th}} \text{ percentile } \pi_p\text{: } p = \int_{-\infty}^{\pi_p} f(\Theta|X) d\Theta$$

Markov Chain Monte Carlo (MCMC) to the Rescue

Intuitive definition: A *Markov chain* represents the random motion of a particle moving around in a space S. A *Markov chain* is a sequence of random variables.

- S is the sample space for $\Theta = (\theta_1, \theta_2, .., \theta_k)$. The coordinates for a point in S are $(\theta_1, \theta_2, .., \theta_k)$.
- The random variables that make up the *Markov chain* are the coordinates of the moving particle.
- The particle jumps from one point to another with ticks of the clock.

k – dimensional space: $\Theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_k^{(i)})$



Markov Chain Monte Carlo (MCMC)

- A *Markov chain* is generated using *Monte Carlo* simulation.
- The *Markov chain* will be a sequence of points $\Theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_k^{(i)})$ that represent different values that random variable Θ can have.
- The distribution of the values $\Theta^{(i)}$ in the sequence will approximate the posterior distribution $f(\Theta|X)$

Markov Chain Monte Carlo (MCMC)

Algorithm

1. Select initial values $\Theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)})$
2. Generate $\Theta^{(t+1)}$ from $\Theta^{(t)}$ using transition kernel $P(\Theta^{(t+1)} | \Theta^{(t)})$ appropriate for $f(\Theta|X)$
3. Repeat second step n times to get $\{\Theta^{(0)}, \Theta^{(1)} \ldots, \Theta^{(n)}\}$
4. Drop $\Theta^{(0)}$and next m simulated values. This is the burn in period.
5. $\{\Theta^{(m+1)}, \ldots, \Theta^{(n)}\}$ is our sample for $f(\Theta|X)$

Markov Chain

- The next $\Theta^{(t+1)}$ in the sequence $\left\{\Theta^{(0)}, \dots, \Theta^{(t)}, \dots\right\}$ depends only on the current value $\Theta^{(t)}$ and not on the sequence of values that preceded $\Theta^{(t)}$:

$$\text{Prob}\left(\Theta^{(t+1)} = y \,\middle|\, \Theta^{(t)} = x, \dots, \Theta^{(1)} = x_1, \Theta^{(0)} = x_0\right) =$$

$$\text{Prob}\left(\Theta^{(t+1)} = y \,\middle|\, \Theta^{(t)} = x\right)$$

- The next step depends on where you are now but not how you got here

MCMC: Ergodic Theory

If transition kernel $P(\Theta^{(t+1)} \mid \Theta^{(t)})$ is suitably constructed:

- The limiting (stationary) distribution for $\{\Theta^{(m+1)}, \dots, \Theta^{(n)}\}$ is $f(\Theta|X)$. A big enough sample is representative of the whole population.

- If A is a region in the Θ parameter space then the relative proportion of time that $\Theta^{(t)}$ lands in A is equal to $\int_A f(\Theta|X)d\Theta$

- $\{\Theta^{(m+1)}, \dots, \Theta^{(n)}\}$ can be used to estimate the mean, moments, etc. of $f(\Theta|X)$. In particular,

$$E[h(\Theta)|X] = \int h(\Theta)f(\Theta|X)d\Theta \approx \frac{1}{n-m}\sum_{t=m+1}^{n} h(\Theta^{(t)})$$

Requirements for Transition Kernel

***The Markov chain generated from transition kernel $P(\Theta^{(t+1)} | \Theta^{(t)})$ should be*:**

- Irreducible – the chain can eventually go from any region of the Θ parameter space to any other region
- Recurrent – the chain will return to the current region of parameter space if you wait long enough (finite wait)
- Aperiodic – there is no pattern in the chain returning to the current region. The chain will not get stuck in a cycle.

Generating Markov Chains for Bayesian Analysis

Two commonly used methods to construct the transition kernel $P(\Theta^{(t+1)} | \Theta^{(t)})$

1) **Metropolis-Hastings Algorithm**: does NOT require an explicit expression for the posterior distribution or conditional distributions.

2) **Gibbs Sampler**: the conditional distributions $f(\theta_i \mid \boldsymbol{X}, \theta_1,.., \theta_{i-1}, \theta_{i+1}, \theta_k)$ must be known for each individual parameter in $\Theta = (\theta_1,\theta_2,..,\theta_k)$.
   It's a special case of Metropolis-Hastings.

Metropolis-Hastings Algorithm

1. Select initial values $\Theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$

t = 0 at start

2. Find a <u>candidate</u> for $\Theta^{(t+1)}$, the next point after $\Theta^{(t)}$.

2. Generate <u>candidate</u> $\Theta^* = (\theta_1^*, \ldots, \theta_k^*)$ using a proposal distribution

$$q(\Theta^* | \Theta^{(t)})$$

Metropolis-Hastings Algorithm

- The proposal distribution should be easy to sample from. A multivariate normal distribution is a possibility:

$$q(\Theta^*|\Theta^{(t)}) \sim N(\Theta^{(t)}, \boldsymbol{\sigma})$$

- The proposal distribution $q(\Theta^*|\Theta^{(t)})$ should contain the support of distribution that we are trying to model: $f(\Theta|X)$.

- Note $\boldsymbol{\sigma}$ determines the step size.

Wait a minute!

- We are trying create a sample from $f(\Theta|X)$, not a multivariate normal or some other proposal distribution.

- Where does the posterior distribution come in?

- **The posterior distribution determines whether we take the proposed step! Do we step from $\boldsymbol{\Theta}^{(t)}$ to $\boldsymbol{\Theta}^{(t+1)} = \boldsymbol{\Theta}^{*}$?**

Should I Stay or Should I Go

from $\Theta^{(t)}$ to Θ^* ?

Trivia question: What band had this hit song? (Just the first line!)

Metropolis-Hastings Algorithm

- Define acceptance ratio $r = \frac{f(\Theta^*|X)/q(\Theta^*|\Theta^{(t)})}{f(\Theta^{(t)}|X)/q(\Theta^{(t)}|\Theta^*)}$

- Generate random number $u \sim U(0,1)$

- If $r > u$ then set $\Theta^{(t+1)} = \Theta^*$, else $\Theta^{(t+1)} = \Theta^{(t)}$

How Do We Compute $f(\Theta^*|X)$???

- **We don't because:**

$$\frac{f(\Theta^*|X)}{f(\Theta^{(t)}|X)} = \frac{f(X|\Theta^*)\pi(\Theta^*)/f(X)}{f(X|\Theta^{(t)})\pi(\Theta^{(t)})/f(X)}$$

$$= \frac{f(X|\Theta^*)\pi(\Theta^*)}{f(X|\Theta^{(t)})\pi(\Theta^{(t)})}$$

Metropolis-Hastings Algorithm

- Acceptance ratio

$$r = \frac{f(X|\Theta^*)\pi(\Theta^*)/q(\Theta^*|\Theta^{(t)})}{f(X|\Theta^{(t)})\pi(\Theta^{(t)})/q(\Theta^{(t)}|\Theta^*)}$$

- Generate random number $u \sim U(0,1)$

- If $r > u$ then set $\Theta^{(t+1)} = \Theta^*$, else $\Theta^{(t+1)} = \Theta^{(t)}$

Metropolis-Hastings Algorithm

$$r = \frac{f(X|\Theta^*)\pi(\Theta^*)/q(\Theta^*|\Theta^{(t)})}{f(X|\Theta^{(t)})\pi(\Theta^{(t)})/q(\Theta^{(t)}|\Theta^*)}$$

- The $q(|)$ terms are there to adjust for the behavior of the proposal distribution. Suppose Θ^* gets proposed a lot when at $\Theta^{(t)}$ but $\Theta^{(t)}$ does <u>not</u> get proposed much when at Θ^*. Then Θ^* would show up relatively more than it should in the Markov chain.

Metropolis-Hastings Algorithm

- If the proposal distribution is symmetric

$$q\left(\Theta^*|\Theta^{(t)}\right) = q\left(\Theta^{(t)}|\Theta^*\right),$$

then

$$r = \frac{f(X|\Theta^*)\pi(\Theta^*)}{f(X|\Theta^{(t)})\pi(\Theta^{(t)})}\,.$$

- If $r > 1$ then accept Θ^* because there is higher probability at Θ^*.

- If $r \leq 1$ accept Θ^* with probability r, i.e. generate $u \sim U(0,1)$ and accept Θ^* if $r > u$.

Metropolis-Hastings Algorithm: All Together Now

1. Select initial values $\Theta^{(0)} = (\theta_1^{(0)},\ldots, \theta_k^{(0)})$.

2. Generate candidate $\Theta^* = (\theta_1^*,\ldots, \theta_k^*)$ for $\Theta^{(t+1)}$, using a proposal distribution $q(\Theta^*|\Theta^{(t)})$

3. Define acceptance ratio $r = \dfrac{f(\Theta^*|X)/q(\Theta^*|\Theta^{(t)})}{f(\Theta^{(t)}|X)/q(\Theta^{(t)}|\Theta^*)}$

4. Generate random number $u \sim U(0,1)$. If $r > u$ then set $\Theta^{(t+1)} = \Theta^*$, else $\Theta^{(t+1)} = \Theta^{(t)}$

Metropolis-Hastings Algorithm in Bayesian Analysis

- Generates a random walk through the support of $f(\Theta|X)$ that favors $\Theta's$ with higher probabilities

- Each point will be visited in proportion to its probability

- The Markov chain $\{\Theta^{(m+1)}, \ldots, \Theta^{(n)}\}$ after burn in serves as a sample from $f(\Theta|X)$

Metropolis-Hastings Algorithm: Issues

- What step size should we take?
  - too small, we don't explore distribution
  - too big, we may propose low probability points too often
- Related question: What percentage of the time should we accept Θ^* ?
- We may need to "tune" our proposal distribution.

Gibbs Sampler: Quickly

Define $\boldsymbol{\Theta_{-i}^{(t)} = (\theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t)}, \dots, \theta_k^{(t)})}$

1. Select initial values $\Theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_k^{(0)})$
2. Generate $\Theta^{(t+1)}$ from $\Theta^{(t)}$ one parameter at a time:

$$
\begin{aligned}
f(\theta_1^{(t+1)} | X, \Theta_{-1}^{(t)}) &\xrightarrow{generates} \theta_1^{(t+1)} \\
f(\theta_2^{(t+1)} | X, \Theta_{-2}^{(t)}) &\longrightarrow \theta_2^{(t+1)} \\
&\vdots \\
f(\theta_k^{(t+1)} | X, \Theta_{-k}^{(t)}) &\longrightarrow \theta_k^{(t+1)}
\end{aligned}
$$

3. Repeat second step n times to get $\{\Theta^{(0)}, \Theta^{(1)} \dots, \Theta^{(n)}\}$

Pros and Cons of Gibbs Sampling

Pro

1. No need to tune proposal distribution
2. No rejected proposals (inefficient)

Con

1. Must have conditional distribution for each parameter and efficient method to generate variates
2. Highly correlated parameters can slow down the tour

Examples on the Way!