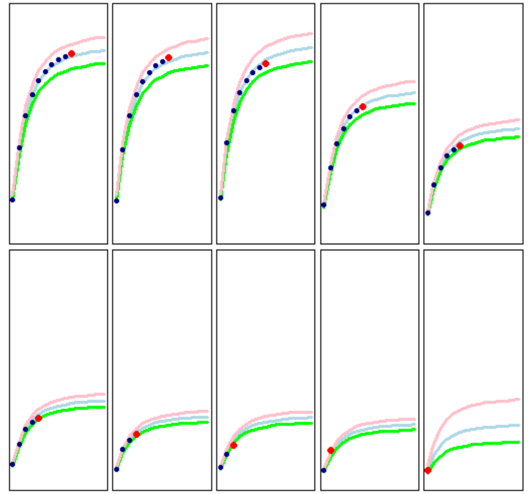


The Future of Loss Reserving

A Bayesian 21st Century

R for Insurance Conference
Amsterdam

Jim Guszcza, PhD, FCAS
Deloitte Consulting LLP
jguszcza@deloitte.com

June 28, 2015

Supp. Adv. Appl. Prob. 7, 106–115 (1975)
Printed in Israel

THE FUTURE OF STATISTICS — A BAYESIAN 21ST CENTURY

D. V. LINDLEY, *University College London and University of Iowa*

The thesis behind this talk is very simple: the only good statistics is Bayesian statistics. Bayesian statistics is not just another technique to be added to our repertoire alongside, for example, multivariate analysis; it is the only method that can produce sound inferences and decisions in multivariate, or any other branch of, statistics. It is not just another chapter to add to that elementary text you are writing; it is that text. It follows that the unique direction for mathematical statistics must be along the Bayesian road.

Motivation:
Why Bayes, Why Now

Probably what we want

"Given any value (estimate of future payments) and our current state of knowledge, what is the probability that the final payments will be no larger than the given value?"

-- Casualty Actuarial Society (2004)
Working Party on Quantifying Variability in Reserve Estimates

I read this as a request for a Bayesian predictive distribution.

Bayes gives us what we want

"Modern Bayesian methods provide richer information, with greater flexibility and broader applicability than 20th century methods. Bayesian methods are intellectually coherent and intuitive. Bayesian analyses are readily computed with modern software and hardware."

-- John Kruschke, Indiana University Psychology

Why Bayes

- "A coherent integration of evidence from different sources"
  - Background information
  - Expert knowledge / judgment (“subjectivity” is a feature, not a bug)
  - Other datasets (e.g. multiple triangles)
  - Shrinkage, “borrowing strength”, hierarchical model structure – all coin of the realm

Why Bayes

- "A coherent integration of evidence from different sources"
  - Background information
  - Expert knowledge / judgment ("subjectivity" is a feature, not a bug)
  - Other datasets (e.g. multiple triangles)
  - Shrinkage, "borrowing strength", hierarchical model structure – all coin of the realm

- Rich output: full probability distribution estimates of <u>all</u> quantities of interest
  - Ultimate loss ratios by accident year
  - Outstanding loss amounts
  - Missing values of any cell in a loss triangle

Why Bayes

- "A coherent integration of evidence from different sources"
  - Background information
  - Expert knowledge / judgment ("subjectivity" is a feature, not a bug)
  - Other datasets (e.g. multiple triangles)
  - Shrinkage, "borrowing strength", hierarchical model structure – all coin of the realm

- Rich output: full probability distribution estimates of <u>all</u> quantities of interest
  - Ultimate loss ratios by accident year
  - Outstanding loss amounts
  - Missing values of any cell in a loss triangle

- Model the process that generates the data
  - As opposed to modeling the data with "procedural" methods
  - We can fit models as complex (or simple) as the situation demands
  - Nonlinear growth patterns, trends, autoregressive, hierarchical, structure, …

Why Bayes

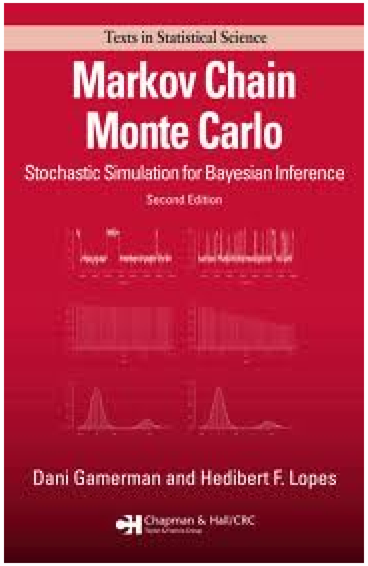
- "A coherent integration of evidence from different sources"
  - Background information
  - Expert knowledge / judgment ("subjectivity" is a feature, not a bug)
  - Other datasets (e.g. multiple triangles)
  - Shrinkage, "borrowing strength", hierarchical model structure – all coin of the realm

- Rich output: full probability distribution estimates of <u>all</u> quantities of interest
  - Ultimate loss ratios by accident year
  - Outstanding loss amounts
  - Missing values of any cell in a loss triangle

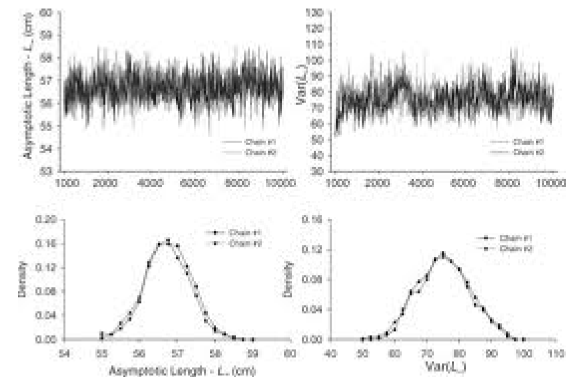
- Model the process that generates the data
  - As opposed to modeling the data with "procedural" methods
  - We can fit models as complex (or simple) as the situation demands
  - Nonlinear growth patterns, trends, autoregressive, hierarchical, structure, …

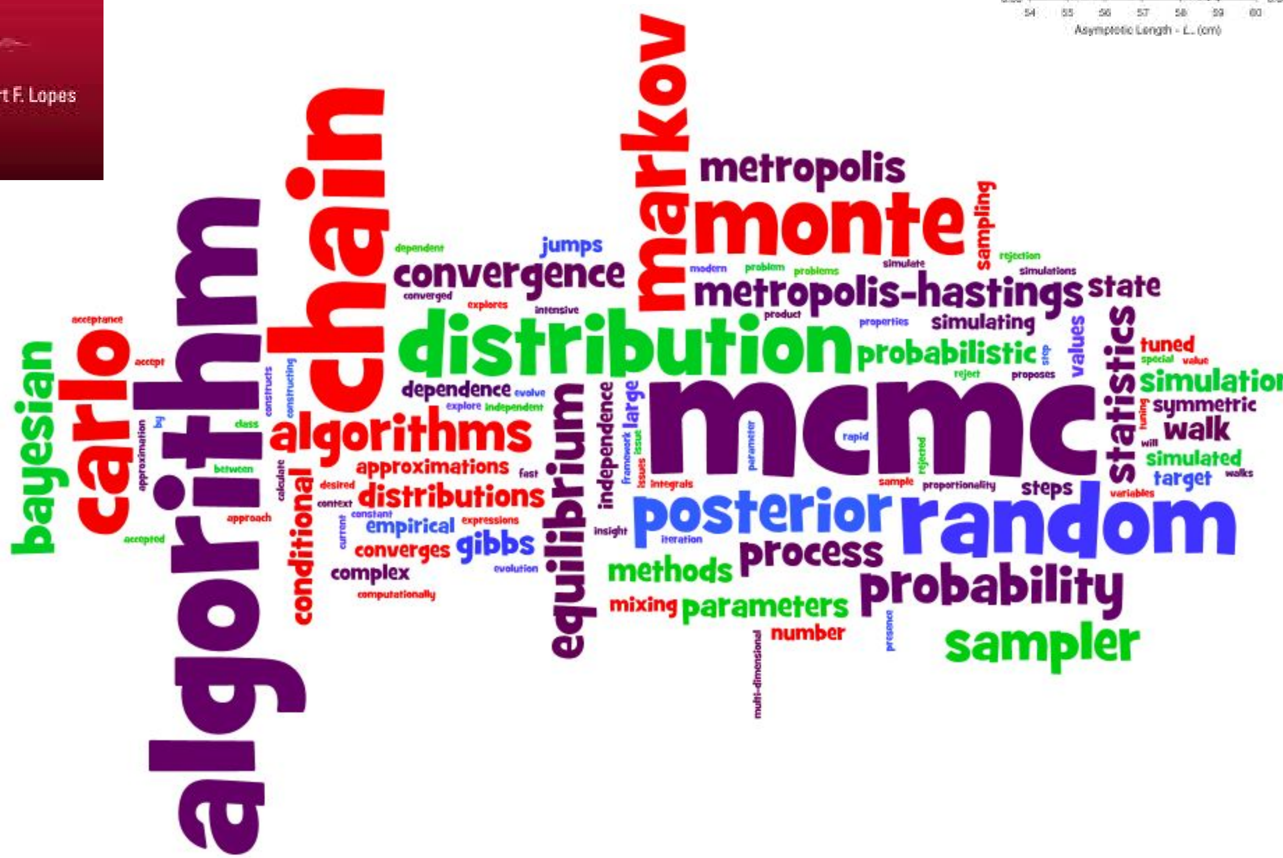
- Conceptual clarity
  - Single-case probabilities make sense in the Bayesian framework
  - Communication of risk: "mean what you say and say what you mean"

Today's Bayes



Is our industry living up to its rich Bayesian heritage?





Bayesian Principles

The Fundamental Bayesian Principle

"For Bayesians as much as for any other statistician, parameters are (typically) fixed but unknown. It is the knowledge about these unknowns that Bayesians model as random…

… typically it is the Bayesian who makes the claim for inference in a particular instance and the frequentist who restricts claims to infinite populations of replications."

-- Andrew Gelman and Christian Robert

The Fundamental Bayesian Principle

"For Bayesians as much as for any other statistician, parameters are (typically) fixed but unknown. It is the knowledge about these unknowns that Bayesians model as random…

… typically it is the Bayesian who makes the claim for inference in a particular instance and the frequentist who restricts claims to infinite populations of replications."

-- Andrew Gelman and Christian Robert

Translation:

- **Frequentist:** Probability models the infinite replications of the data $\boldsymbol{X}$
- **Bayesian:** Probability models our partial knowledge about θ

Updating Subjective Probability

- Bayes' **Theorem** (a mathematical fact):

$$\Pr(H \mid E) = \frac{\Pr(H \wedge E)}{\Pr(E)} = \frac{\Pr(E \mid H)\Pr(H)}{\Pr(E)}$$

- Bayes' **updating rule** (a methodological premise):
- Let *P*(*H*) represents our belief in hypothesis *H* before receiving evidence *E*.
- Let *P**(*H*) represent our belief about *H* *after* receiving evidence *E*.
- **Bayes Rule: *P**(H) = Pr(*H*|*E*)**

$$\boxed{\Pr(H) \underset{E}{\longrightarrow} \Pr(H \mid E)}$$

Learning from data

Suppose Persi tosses a coin 12 times and gets 3 heads.
What is the probability of heads on the 13th toss?

Learning from data

Suppose Persi tosses a coin 12 times and gets 3 heads. What is the probability of heads on the 13th toss?

Frequentist analysis

$X_i \sim_{iid} Bern(\theta)$ ➔ $L(\theta|H=3, n=12) = \prod \theta^3 (1-\theta)^9$ ➔ $\hat{\theta}_{MLE} = \frac{1}{4}$

Learning from data

Suppose Persi tosses a coin 12 times and gets 3 heads.
What is the probability of heads on the 13th toss?

Frequentist analysis

$X_i \sim_{iid} Bern(\theta)$ ➔ $L(\theta|H = 3, n = 12) = \prod \theta^3 (1-\theta)^9$ ➔ $\hat{\theta}_{MLE} = \frac{1}{4}$

Thoughts

- "Parameter risk": 12 flips is not a lot of data ("credibility concerns")
- We've flipped other coins before… isn't that knowledge relevant?
- It would be nice to somehow "temper" the estimate of ¼ or "credibility weight" it with some other source of information
- It would be nice not to just give a point estimate and a confidence interval, but say things like: $Pr(L < \theta < U) = p$

Learning from data

Suppose Persi tosses a coin 12 times and gets 3 heads. What is the probability of heads on the 13th toss?

Bayesian analysis

$\theta \sim Beta(\alpha, \beta) \Rightarrow \theta \sim Beta(\alpha + 3, \beta + 9)$

Learning from data

Suppose Persi tosses a coin 12 times and gets 3 heads.
What is the probability of heads on the 13th toss?

Bayesian analysis
$\theta \sim Beta(\alpha, \beta)$ ➔ $\theta \sim Beta(\alpha + 3, \beta + 9)$

Thoughts

- "Parameter risk": quantified by the posterior distribution
- Prior knowledge: encoded in the choice of $\{\alpha, \beta\}$
- Other data: maybe Persi has flipped other coins on other days… we could throw all of this (together with our current data) into a **hierarchical model**
- Mean what we say and say what we mean: $Pr(L < \theta < U) = p$ is a "**credibility interval**"… it's what most people think confidence intervals say… (but don't!)

Prior distributions: a feature, not a bug

"Your 'subjective' probability is not something fetched out of the sky on a whim; it is what your actual judgment should be, in view of your information to date and other people's information."

-- Richard Jeffrey, Princeton University

Prior distributions: a feature, not a bug

"Your 'subjective' probability is not something fetched out of the sky on a whim; it is what your actual judgment should be, in view of your information to date and other people's information."
-- Richard Jeffrey, Princeton University

- "Subjective" probability is really "judgmental" probability

- The choice of likelihood function is also "subjective" In this sense
  - ODP (or other) distributional form
  - Inclusion of covariates
  - Trends
  - Tail factor extrapolations
  - ….

Bayesian Computation

An intractable problem

Before 1990: this sort of thing was often viewed as a parlor trick because of the need to analytically solve high-dimensional integrals:

$$f(Y \mid X) = \int f(Y \mid \theta) f(\theta \mid X) d\theta = \int f(Y \mid \theta) \left(\frac{f(X \mid \theta)\pi(\theta)}{\int f(X \mid \theta)\pi(\theta) d\theta} \right) d\theta$$

$$f(\theta \mid X) = \frac{f(X \mid \theta)\pi(\theta)}{\int f(X \mid \theta)\pi(\theta) d\theta}$$

Why Everyone Wasn't a Bayesian

Before 1990: this sort of thing was often viewed as a parlor trick because of the need to analytically solve high-dimensional integrals:

$$f(Y \mid X) = \int f(Y \mid \theta) f(\theta \mid X) d\theta = \int f(Y \mid \theta) \left(\frac{f(X \mid \theta)\pi(\theta)}{\int f(X \mid \theta)\pi(\theta) d\theta} \right) d\theta$$

Why Isn't Everyone a Bayesian?

B. EFRON*

Originally a talk delivered at a conference on Bayesian statistics, this article attempts to answer the following question: why is most scientific data analysis carried out in a non-Bayesian framework? The argument consists mainly of some practical examples of data analysis, in which the Bayesian approach is difficult but Fisherian/frequentist solutions are relatively easy. There is a brief discussion of objectivity in statistical analyses and of the difficulties of achieving objectivity within a Bayesian framework. The article ends with a list of practical advantages of Fisherian/frequentist methods, which so far seem to have outweighed the philosophical superiority of Bayesianism.

MCMC makes it practical

After 1990: The introduction of Markov Chain Monte Carlo [MCMC] simulation to Bayesian practice introduces a "new world order":

Now we can simulate Bayesian posteriors.

Sampling-Based Approaches to Calculating Marginal Densities

ALAN E. GELFAND AND ADRIAN F. M. SMITH*

© 1990 American Statistical Association
Journal of the American Statistical Association
June 1990, Vol. 85, No. 410, Theory and Methods

Chains we can believe in

We set up random walks through parameter space that… in the limit… pass through each region in the probability space in proportion to the posterior probability density of that region.

- How the **Metropolis-Hastings sampler** generates a **Markov chain** $\{\theta_1, \theta_2, \theta_3, \ldots\}$:

1. Time t=1: select a random initial position θ_1 in parameter space.
2. Select a **proposal distribution** $p(\theta)$ that we will use to select proposed random steps away from our current position in parameter space.
3. Starting at time t=2: repeat the following until you get convergence:
   a) At step t, generate a proposed $\theta^* \sim p(\theta)$
   b) Also generate $u \sim \text{unif}(0,1)$
   c) If $R > u$ then $\theta_t = \theta^*$. Else, $\theta_t = \theta_{t-1}$.

$$R = \frac{f(\theta^* \mid X)}{f(\theta_{t-1} \mid X)} \cdot \frac{p(\theta_{t-1} \mid \theta^*)}{p(\theta^* \mid \theta_{t-1})}$$

Step (3c) implies that at step t, we accept the proposed step θ^* with probability min(1,R).

Let's go to the Metropolis

- So now we have something we can easily program into a computer.

- At each step, give yourself a coin with probability of heads min(1,*R*) and flip it.

$$R = \frac{f(X \mid \theta^*)\pi(\theta^*)}{f(X \mid \theta_{t-1})\pi(\theta_{t-1})} \cdot \frac{p(\theta_{t-1} \mid \theta^*)}{p(\theta^* \mid \theta_{t-1})}$$

- If the coin lands heads move from $\theta_{t\text{-}1}$ to θ^*

- Otherwise, stay put.

- The result is a Markov chain (step *t* depends only on step *t*-1… not on prior steps). And it converges on the posterior distribution.

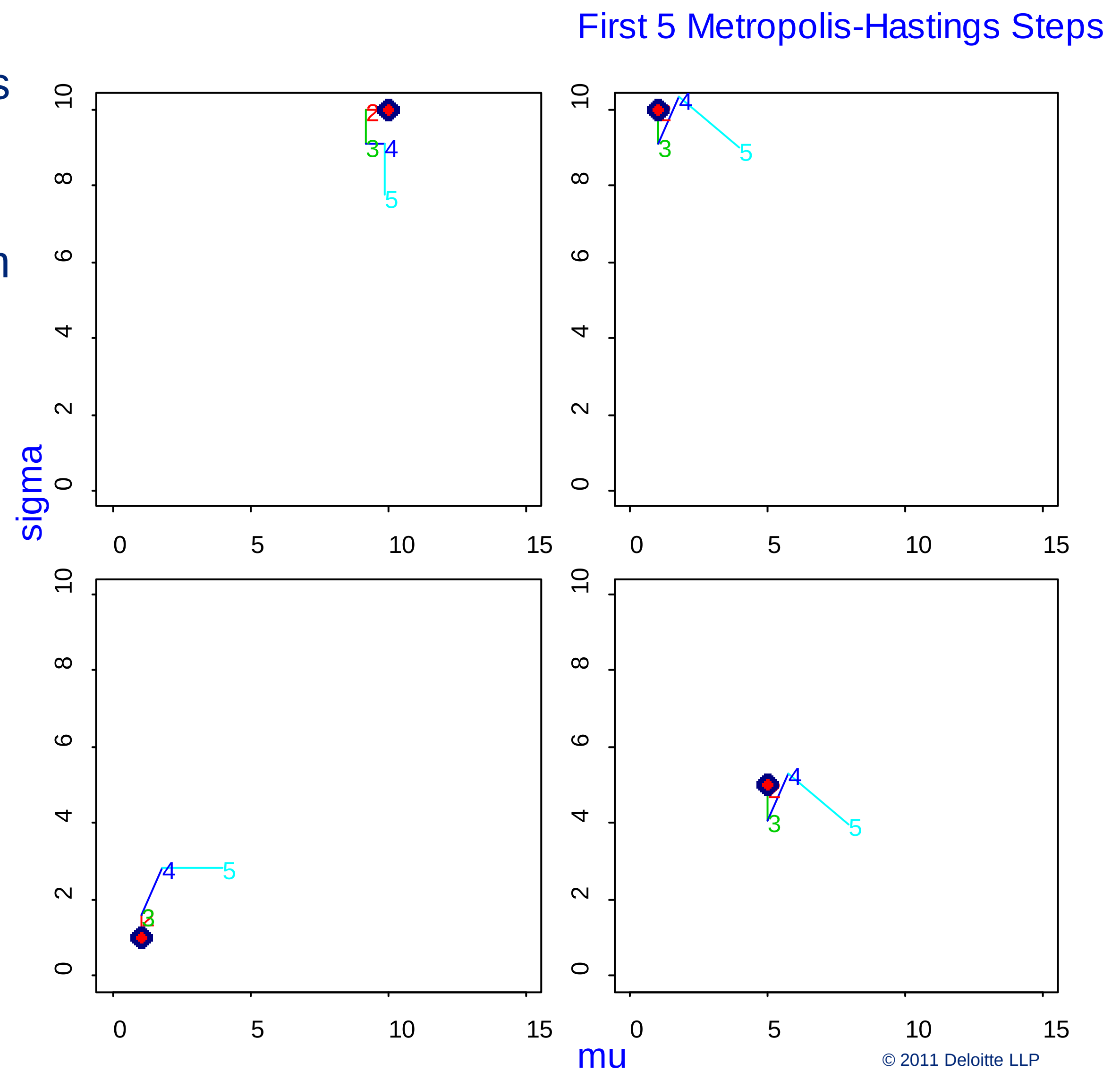
Random walks with 4 different starting points

- We estimate the lognormal density using 4 separate sets of starting values.

- Data: 50 random draws from lognormal(9,2).

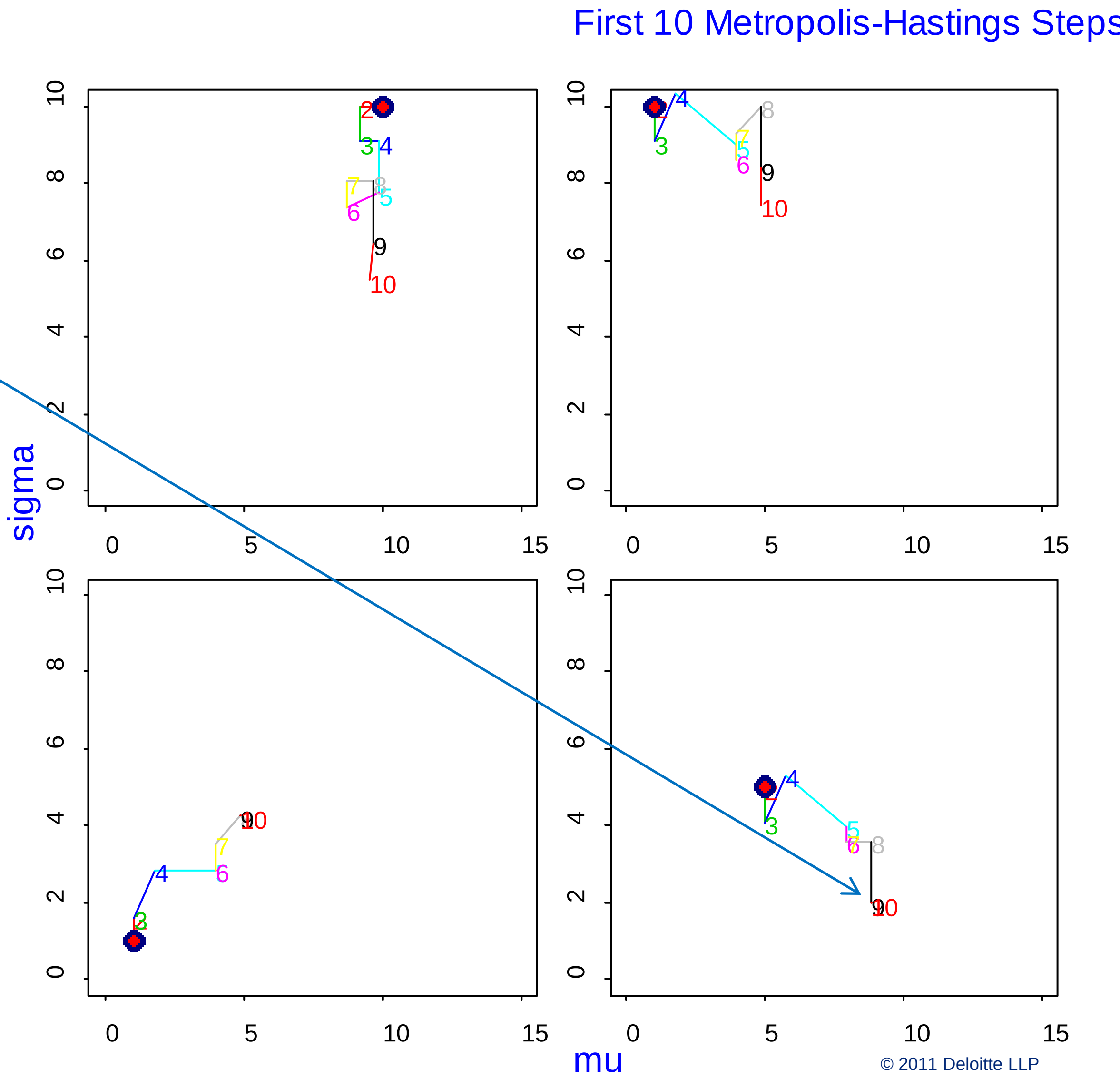
$$f(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-z^2\big/2\right) \quad , \quad z = \frac{\ln(x) - \mu}{\sigma}$$

```
> round(xx)[order(xx)]
 [1]     50    210    443    561    596    779
 [7]   1037   1544   2365   2480   2749   2764
[13]   2865   2947   3007   3440   3599   4226
[19]   4348   4770   4962   5411   6438   6682
[25]   7128   7612   8555   9260   9697   9697
[31]  10486  11380  13630  17910  19014  25840
[37]  28737  35448  38379  50122  60746  78688
[43]  94977  97028  98491 139625 143219 199609
[49] 494979 662527
```



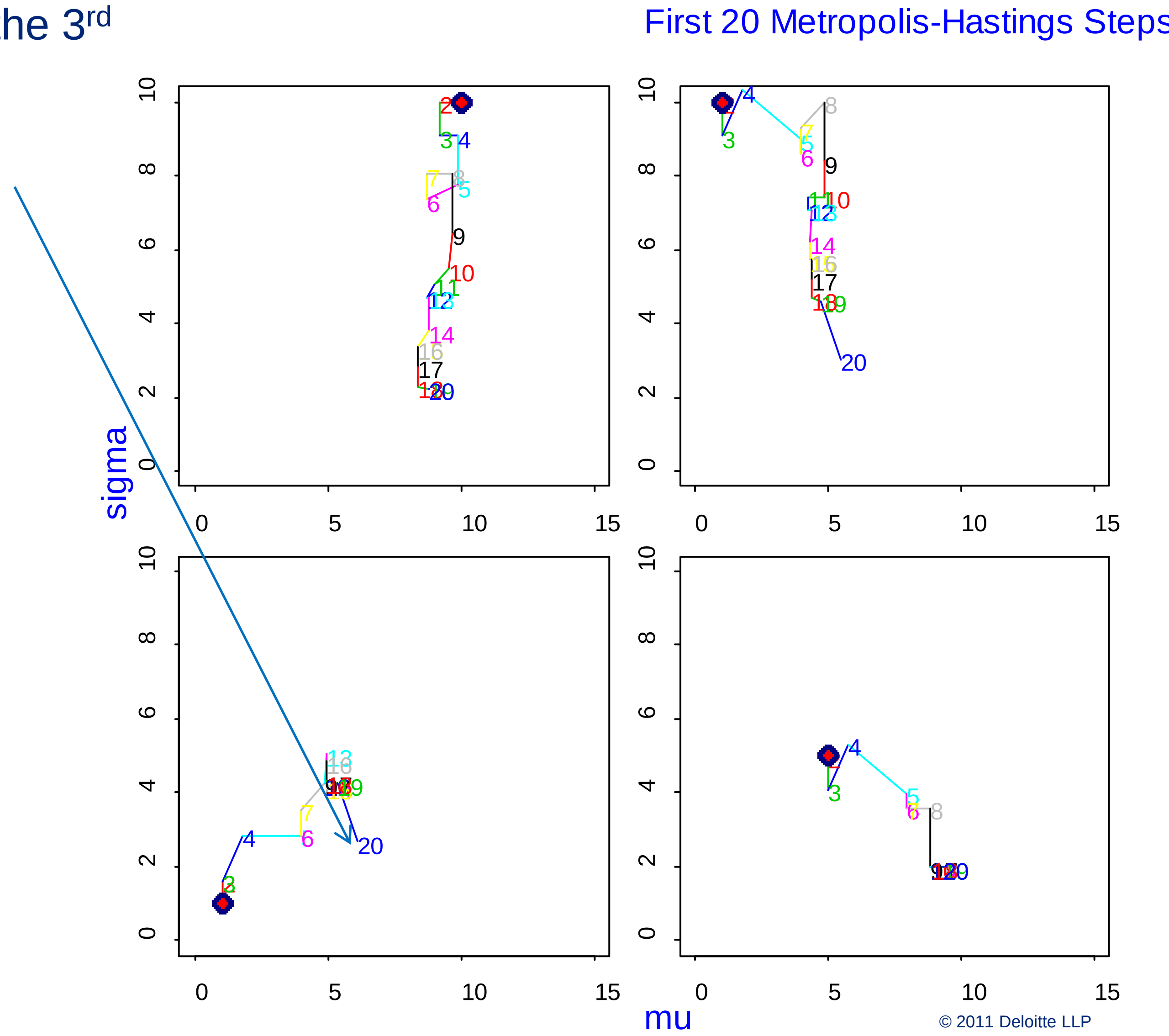
Random walks with 4 different starting points

- After 10 iterations, the lower right chain is already in the right neighborhood.



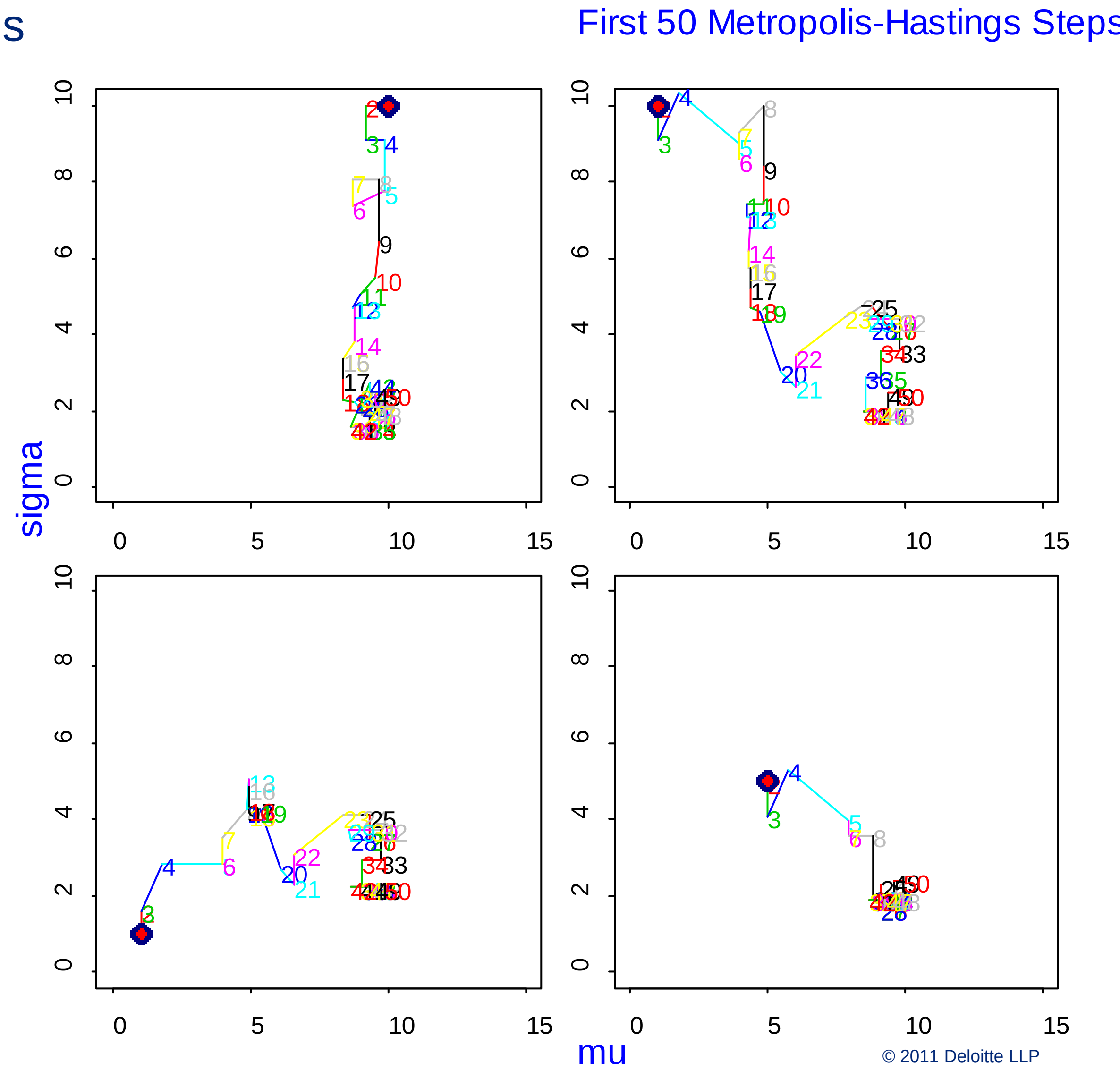
Random walks with 4 different starting points

- After 20 iterations, only the 3rd chain is still in the wrong neighborhood.



Random walks with 4 different starting points

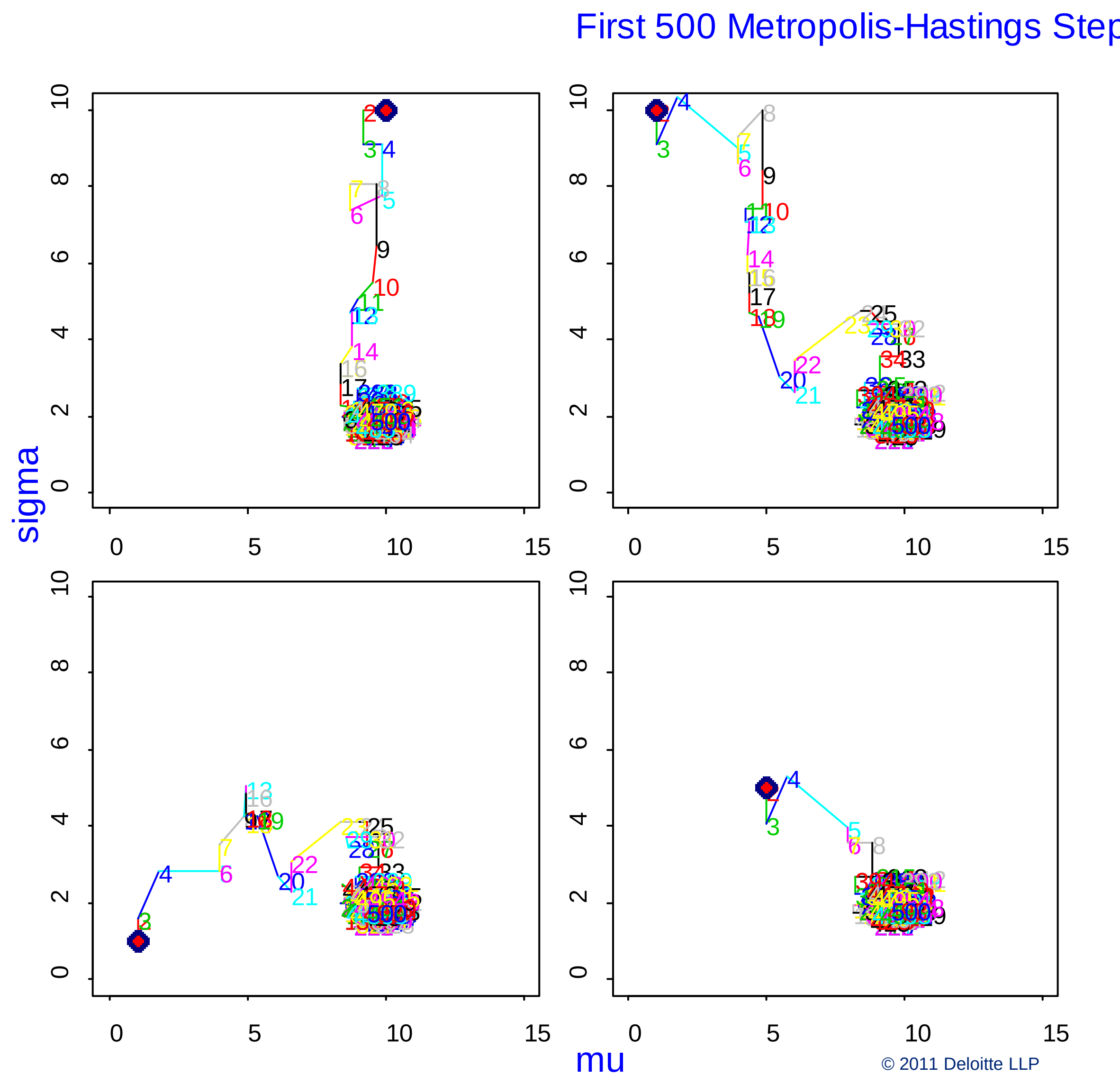
- After 50 iterations, all 4 chains have arrived in the right neighborhood.



Random walks with 4 different starting points

- By 500 chains, it appears that the burn-in has long since been accomplished.
- The chain continues to wander.

The time the chain spends in a neighborhood approximates the posterior probability that (μ,σ) lies in this neighborhood.

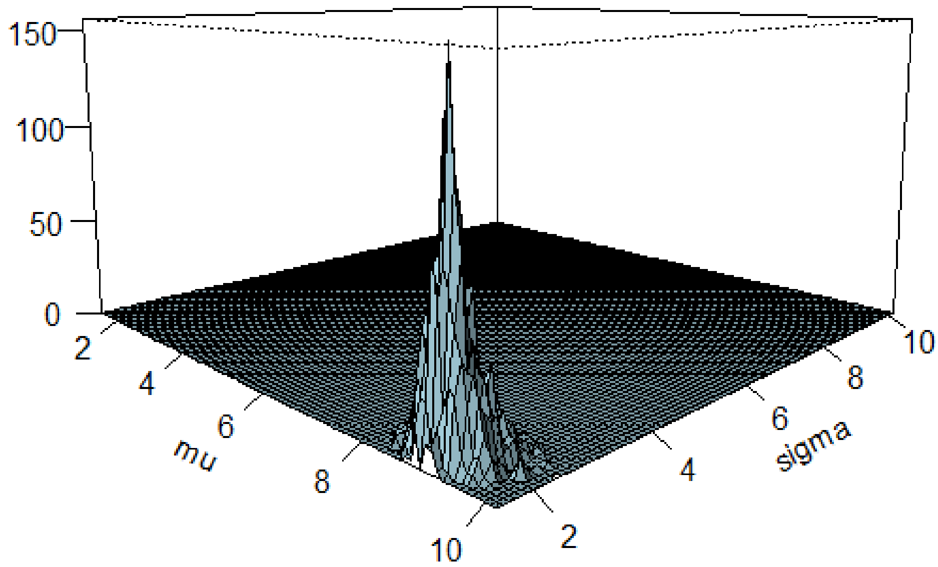


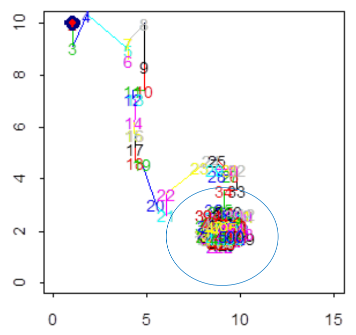
In 3D

Recall the true lognormal parameters are:

μ=9 and σ=2

Metropolis-Hastings Posterior Density Estimate



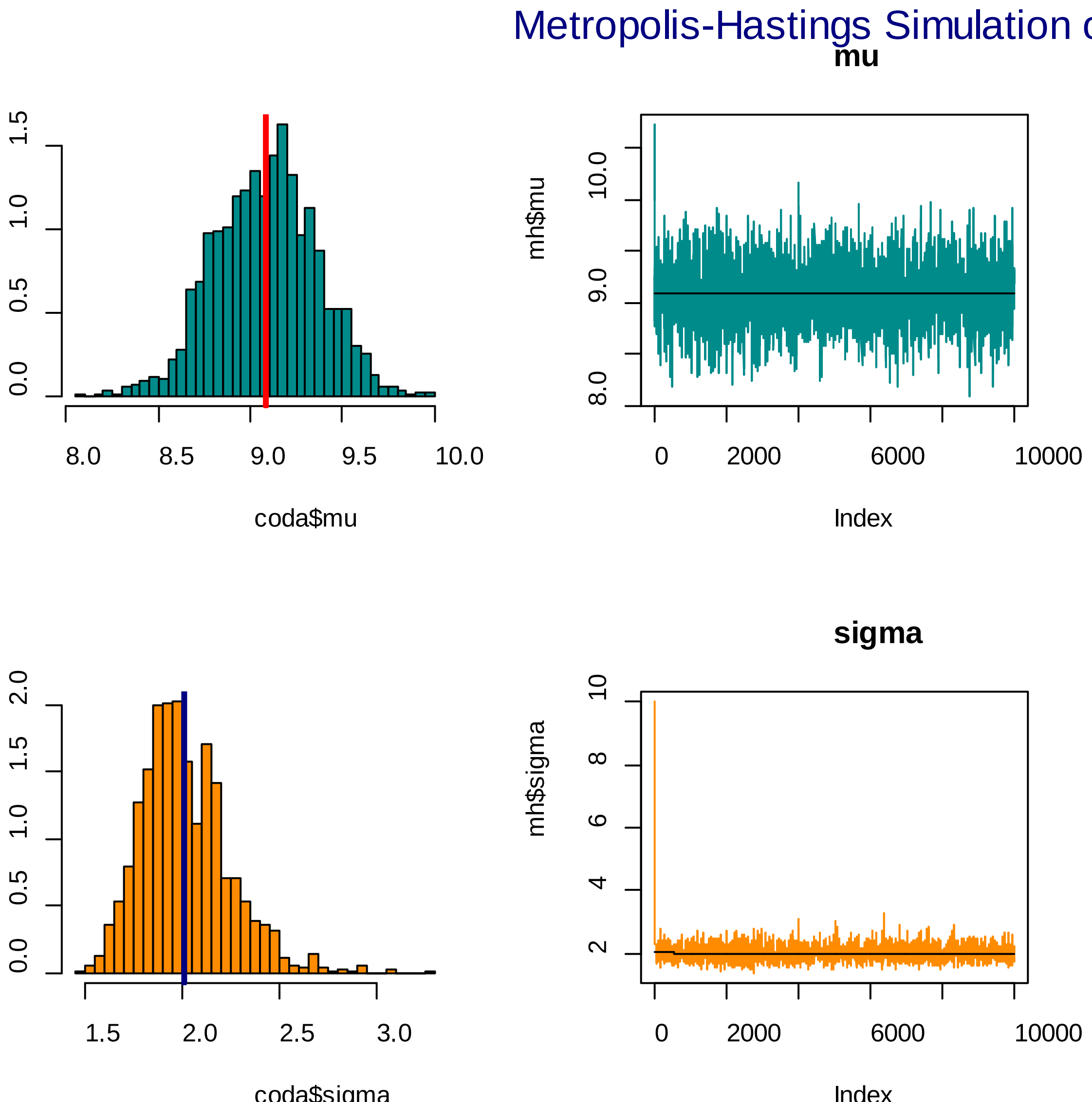


Metropolis-Hastings results

- The MH simulation is gives consistent results:

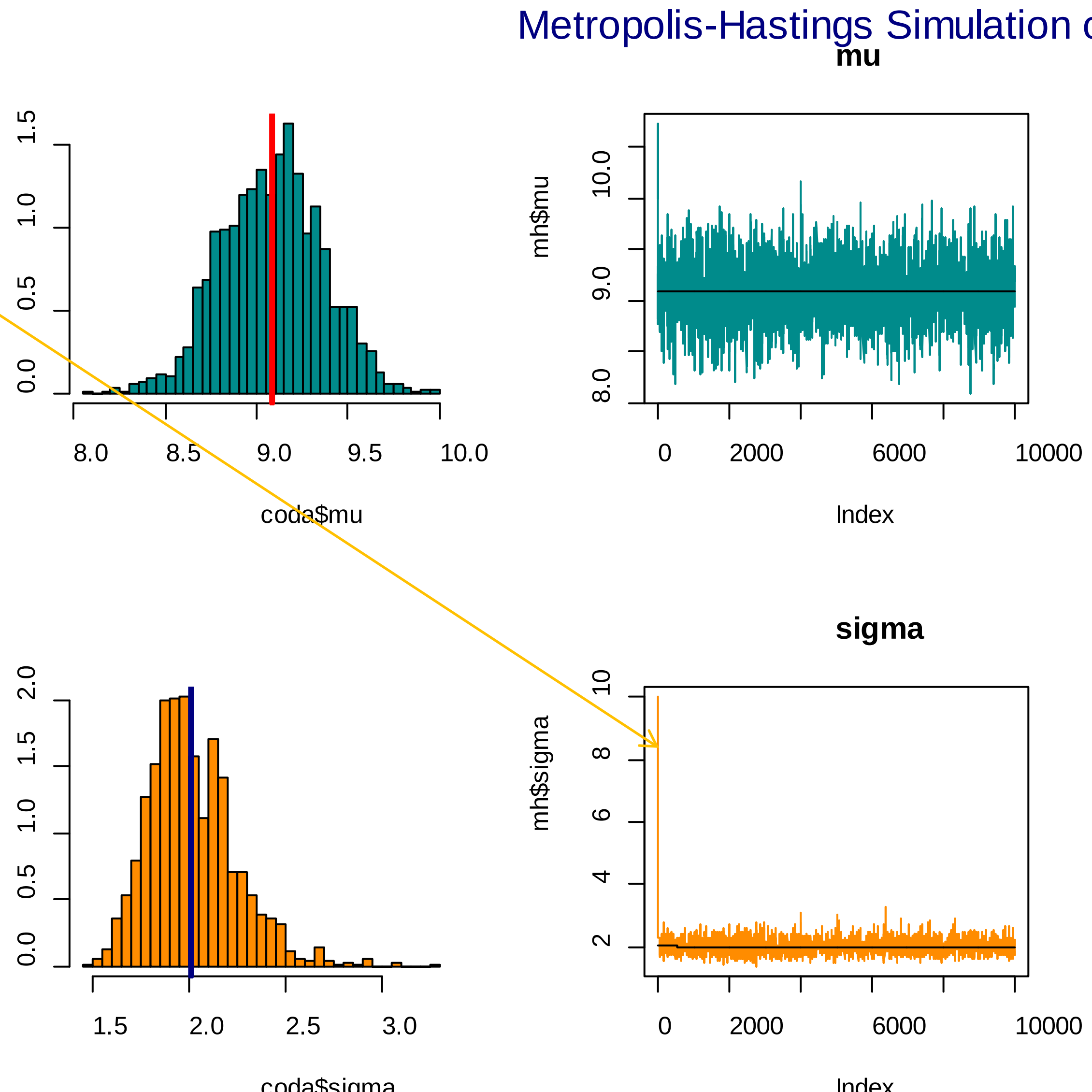
```
> apply(coda, 2, mean)
      mu     sigma
9.077489 2.007377
> apply(coda, 2, sd)
       mu      sigma
0.2741341 0.2247070
```

- Only the final 5000 of the 10000 MH iterations were used to estimate μ,σ



Metropolis-Hastings results

Note the very rapid convergence despite unrealistic initial values.



An easier way to get the same result

Call JAGS from within R

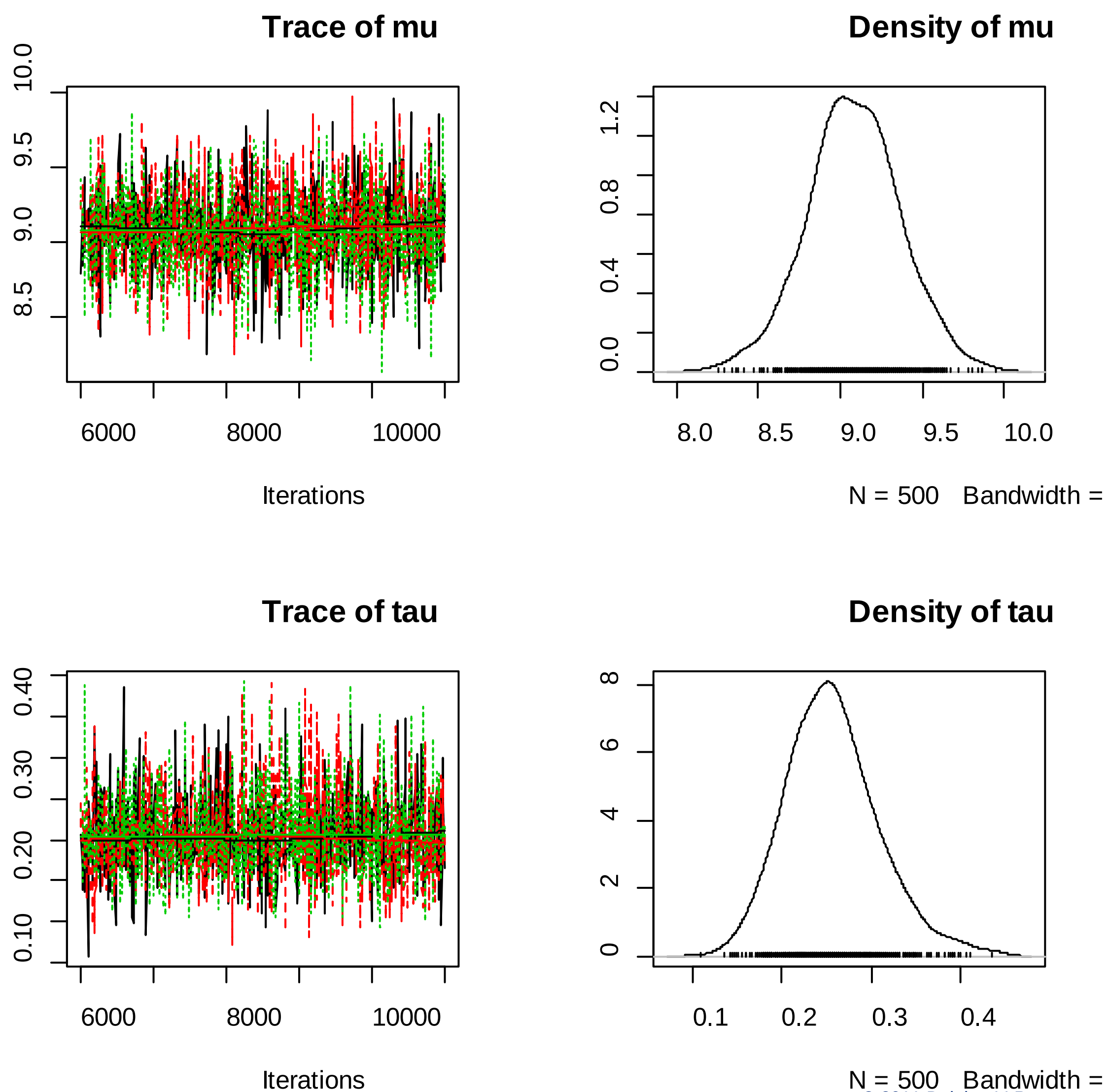
```
model {
for (i in 1:n) {
      x[i] ~ dlnorm( mu, tau )
    }
mu ~ dnorm(0, .0001)
tau ~ dgamma(.0001, .0001)
}
```

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

      Mean      SD Naive SE Time-series SE
mu  9.0830 0.28265 0.007298       0.006878
tau 0.2569 0.05208 0.001345       0.001262

2. Quantiles for each variable:

      2.5%    25%    50%    75%  97.5%
mu  8.5053 8.9020 9.0782 9.2648 9.6409
tau 0.1653 0.2206 0.2535 0.2877 0.3769
```



Bayesian Loss Reserving

Methodology: sophisticated simplicity

*"It is fruitful to start simply and complicate if necessary. That is, it is recommended that an initial, **sophisticatedly simple** model be formulated and tested in terms of explaining past data and in forecasting or predicting new data. If the model is successful... it can be put into use. If not, [it] can be modified or elaborated to improve performance..."*

-- Arnold Zellner, The University of Chicago

Methodology: sophisticated simplicity

*"It is fruitful to start simply and complicate if necessary. That is, it is recommended that an initial, **sophisticatedly simple** model be formulated and tested in terms of explaining past data and in forecasting or predicting new data. If the model is successful… it can be put into use. If not, [it] can be modified or elaborated to improve performance…"*

-- Arnold Zellner, The University of Chicago

This is precisely what Bayesian Data Analysis enables us to do!

Methodology: sophisticated simplicity

"It is fruitful to start simply and complicate if necessary. That is, it is recommended that an initial, ***sophisticatedly simple*** *model be formulated and tested in terms of explaining past data and in forecasting or predicting new data. If the model is successful... it can be put into use. If not, [it] can be modified or elaborated to improve performance..."*

-- Arnold Zellner, The University of Chicago

Start with a simple model and then add structure to account for:

- Other distributional forms (what's so sacred about GLM or exponential family??)
- Negative incremental incurred losses
- Nonlinear structure (e.g. growth curves)
- Hierarchical structure (e.g. fitting multiple lines, companies, regions)
- Prior knowledge
- Other loss triangles ("complement of credibility")
- Calendar/accident year trends
- Autocorrelation
- …

Background: hierarchical modeling from A to B

- Hierarchical modeling is used when one’s data is **grouped** in some important way.
  - Claim experience by state or territory
  - Workers Comp claim experience by class code
  - Claim severity by injury type
  - Churn rate by agency
  - Multiple years of loss experience by policyholder.
  - **Multiple observations of a cohort of claims over time**

- Often grouped data is modeled either by:
  - Building separate models by group
  - Pooling the data and introducing dummy variables to reflect the groups

- Hierarchical modeling offers a “middle way”.
  - Parameters reflecting group membership enter one’s model through appropriately specified **probability sub-models**.

Common hierarchical models

- Classical linear model
  - Equivalently: $Y_i \sim N(\alpha+\beta X_i, \sigma^2)$
  - Same α, β for each data point

$$\boxed{Y_i = \alpha + \beta X_i + \varepsilon_i}$$

- Random intercept model
  - Where: $Y_i \sim N(\alpha_{j[i]}+\beta X_i, \sigma^2)$
  - And: $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$
  - Same β for each data point; but α varies by group j

$$\boxed{Y_i = \alpha_{j[i]} + \beta X_i + \varepsilon_i}$$

- Random intercept and slope model
  - Both α and β vary by group

$$\boxed{Y_i = \alpha_{j[i]} + \beta_{j[i]} X_i + \varepsilon_i}$$

$$\boxed{Y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} \cdot X_i, \sigma^2\right) \quad \textit{where} \quad \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \Sigma\right) \quad , \quad \Sigma = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}}$$

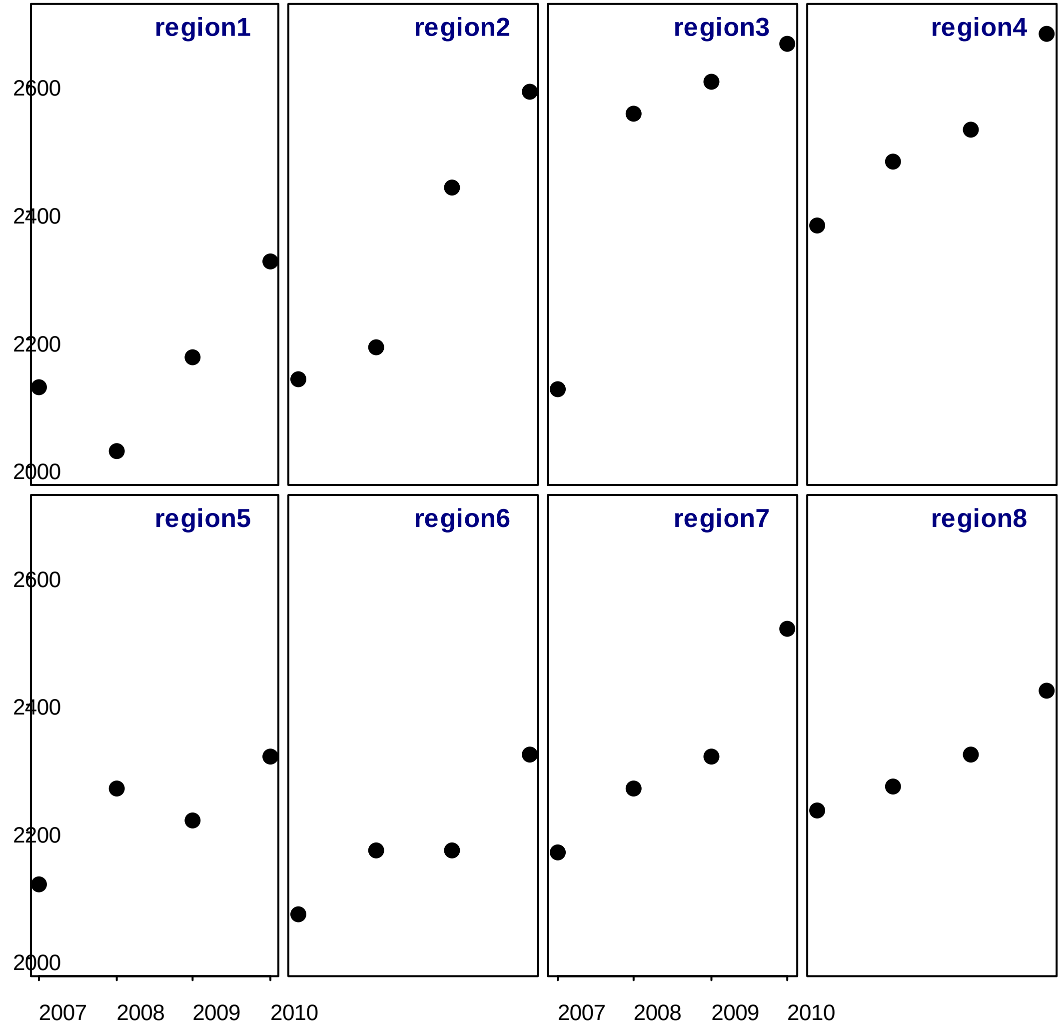
Simple example: policies in force by region

- Simple example: Change in PIF by region from 2007-10
- 32 data points
  - 4 years
  - 8 regions

region	2005	2006	2007	2008
1	2124	2024	2174	2324
2	2138	2188	2438	2588
3	2121	2554	2604	2666
4	2380	2480	2530	2680
5	2118	2268	2218	2318
6	2070	2170	2170	2320
7	2167	2267	2317	2517
8	2232	2272	2322	2422

- But we could as easily have 80 or 800 regions
- Our model would not change
- We view the dataset as a bundle of very short time series

PIF Growth by Region



Classical linear model

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

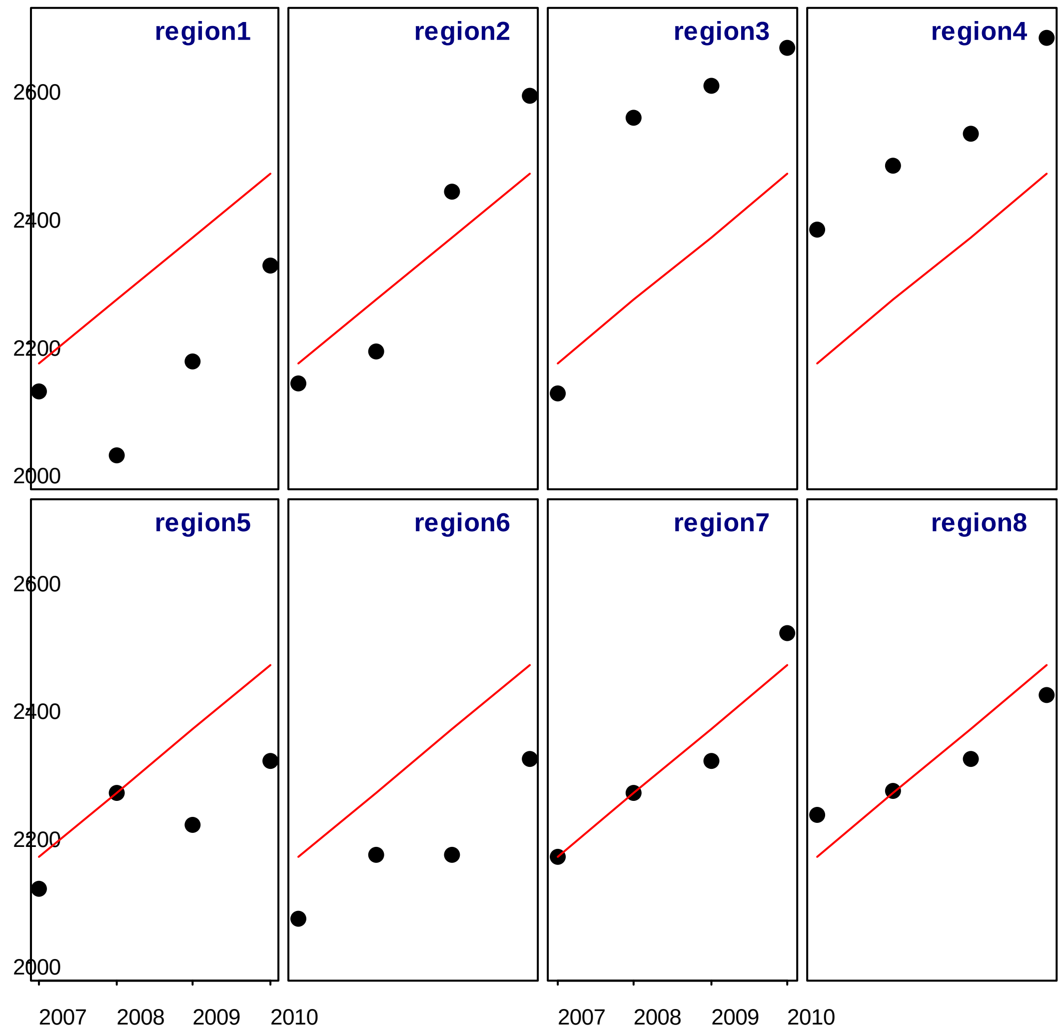
- Option 1: the classical linear model

- Complete Pooling
  - Don't reflect region in the model design
  - Just throw all of the data into one pot and regress

- Same α and β for each region.

- This obviously doesn't cut it.
  - But filling 8 separate regression models or throwing in region-specific dummy variables isn't an attractive option either.
  - Danger of over-fitting
  - i.e. "credibility issues"

PIF Growth by Region



Randomly varying intercepts

$$Y_i \sim N(\alpha_{j[i]} + \beta X_i, \sigma^2) \qquad \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- Option 2: random intercept model

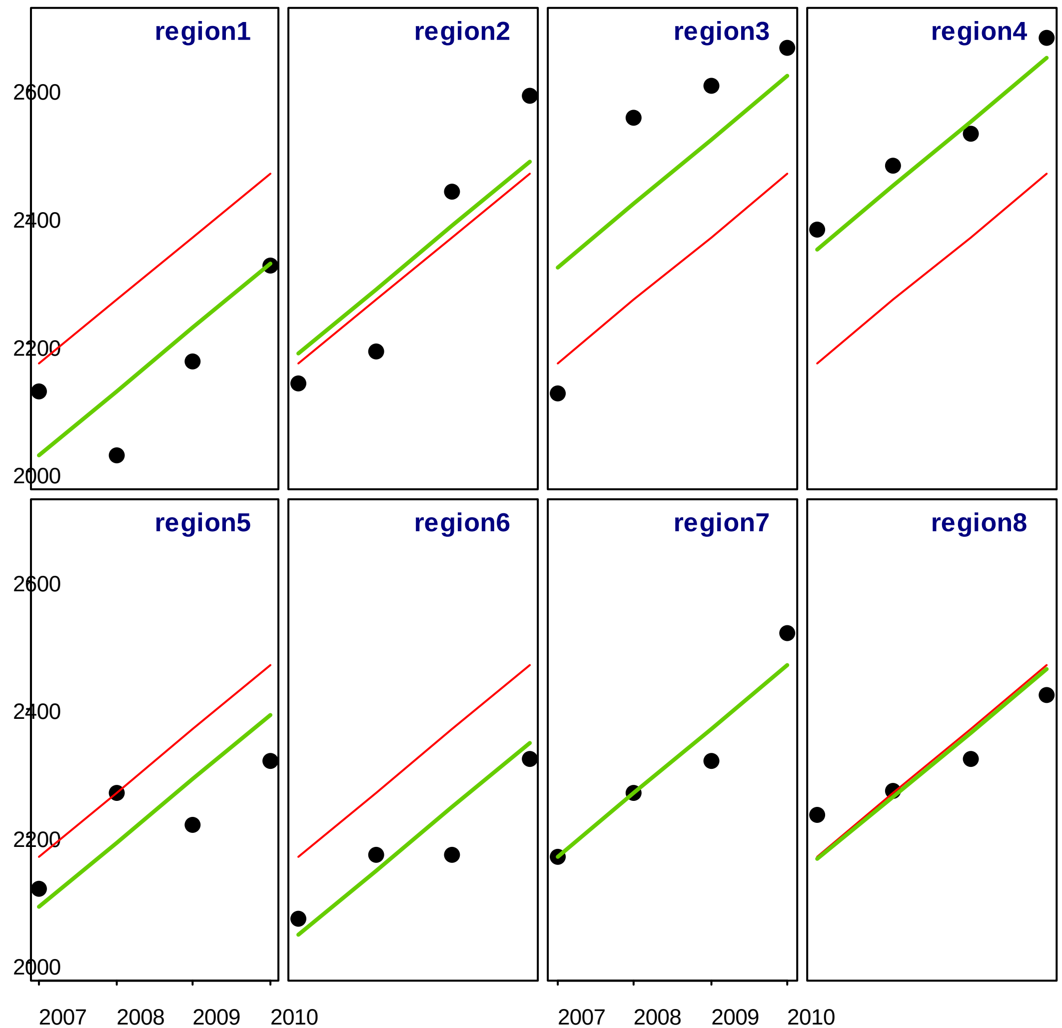
- $Y_i = \alpha_{j[i]} + \beta X_i + \varepsilon_i$

- This model has 9 parameters: $\{\alpha_1, \alpha_2, \ldots, \alpha_8, \beta\}$

- And it contains 4 hyperparameters: $\{\mu_\alpha, \beta, \sigma_\alpha, \sigma\}$

- A major improvement

PIF Growth by Region



Randomly varying intercepts and slopes

$$Y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} \cdot X_i, \sigma^2\right) \quad where \quad \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \Sigma\right) \quad , \quad \Sigma = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}$$

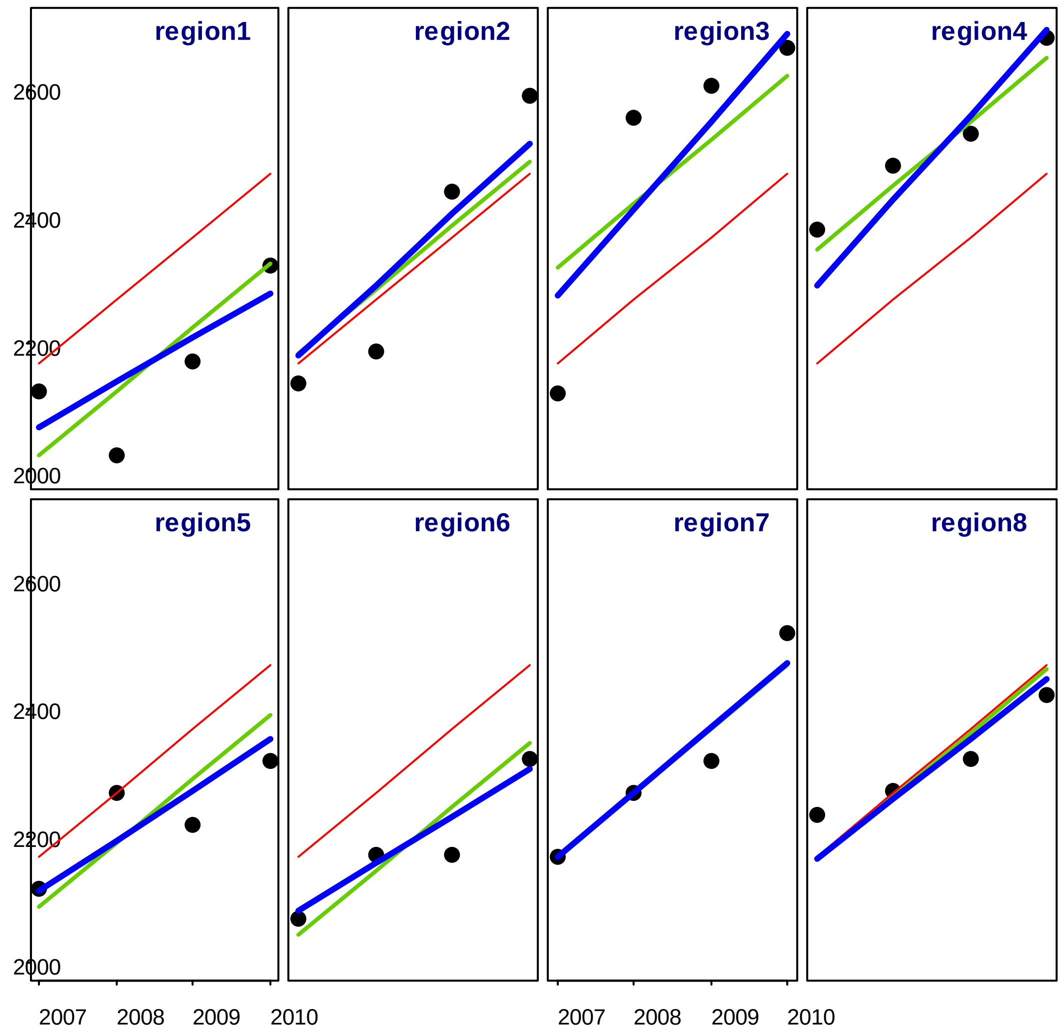
- Option 3: the random slope and intercept model
- $Y_i = \alpha_{j[i]} + \beta_{j[i]} X_i + \varepsilon_i$
- This model has 16 parameters:

  $\{\alpha_1, \alpha_2, \ldots, \alpha_8,$
  $\beta_1, \beta_2, \ldots, \beta_8\}$
  - (note that 8 separate models also contain 16 parameters)
- And it contains 6 hyperparameters:

  $\{\mu_\alpha, \mu_\beta, \sigma, \sigma_\alpha, \sigma_\beta, \sigma_{\alpha\beta}\}$

It'd be the same number of hyperparameters if we had 80 or 800 regions

PIF Growth by Region



A compromise between complete pooling and no pooling

$$PIF = \alpha + \beta t + \varepsilon$$

Complete Pooling

- Ignore group structure altogether

$$\left\{PIF = \alpha^k + \beta^k t + \varepsilon^k\right\}_{k=1,2,..,8}$$

No Pooling

- Estimate a separate model for each group

Compromise

Hierarchical Model

- Estimates parameters using a compromise between complete pooling and no pooling.

$$Y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} \cdot X_i, \sigma^2\right) \quad where \quad \begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim N\left(\begin{bmatrix}\mu_\alpha \\ \mu_\beta\end{bmatrix}, \Sigma\right) \quad , \quad \Sigma = \begin{bmatrix}\sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2\end{bmatrix}$$

A credible approach

- For illustration, recall the random intercept model:

$$\boxed{Y_i \sim N(\alpha_{j[i]} + \beta X_i, \sigma^2) \qquad \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)}$$

- This model can contain a large number of parameters $\{\alpha_1,\ldots,\alpha_J,\beta\}$.

- Regardless of *J*, it contains 4 <u>hyper</u>parameters $\{\mu_\alpha, \beta, \sigma, \sigma_\alpha\}$.

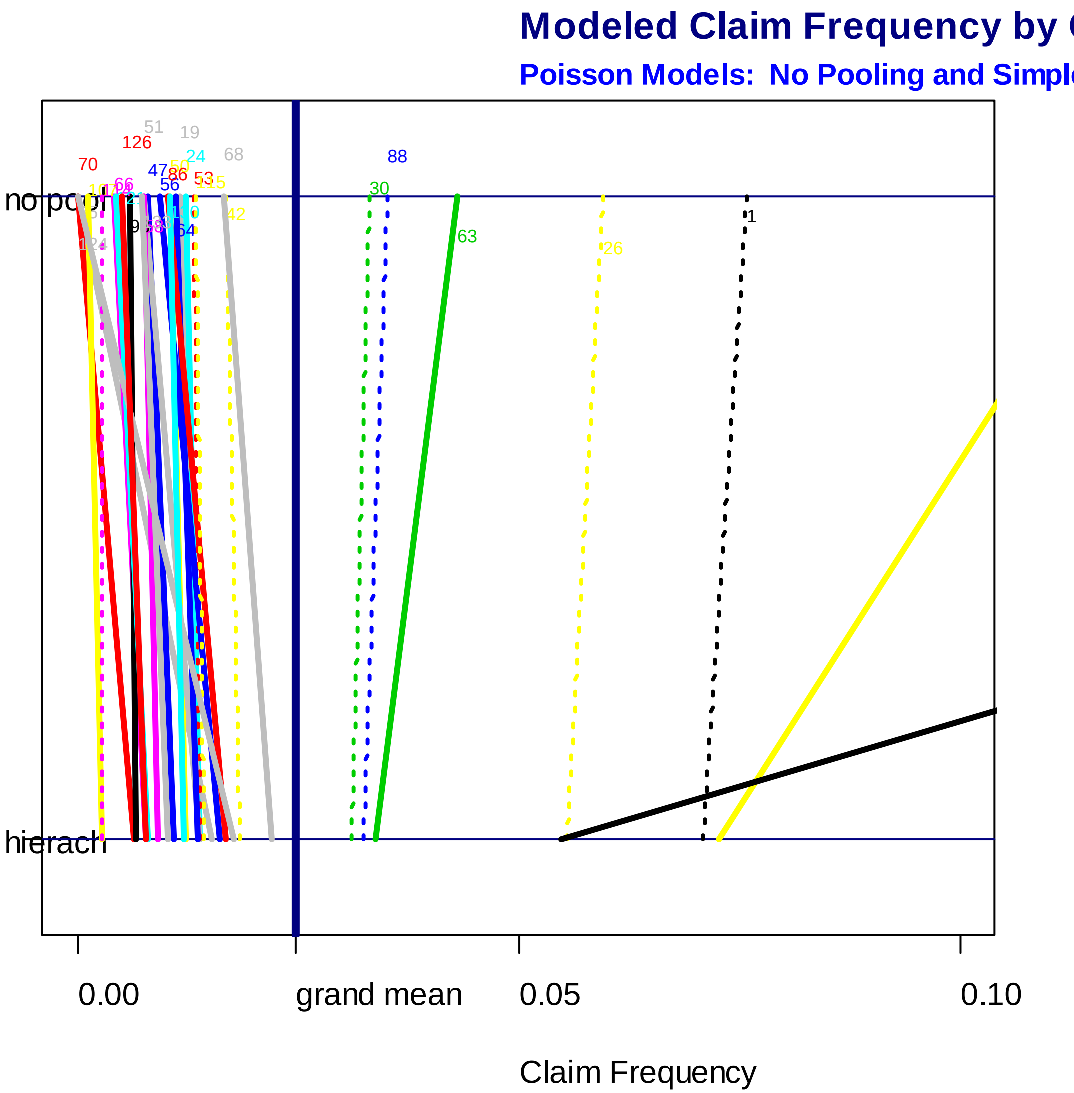
- Here is how the hyperparameters relate to the parameters:

$$\boxed{\hat{\alpha}_j = Z_j \cdot (\bar{y}_j - \beta \bar{x}_j) + (1 - Z_j) \cdot \hat{\mu}_\alpha \quad \textit{where} \quad Z_j = \frac{n_j}{n_j + \sigma^2 / \sigma_\alpha^2}}$$

Bühlmann credibility is a special case of hierarchical models.

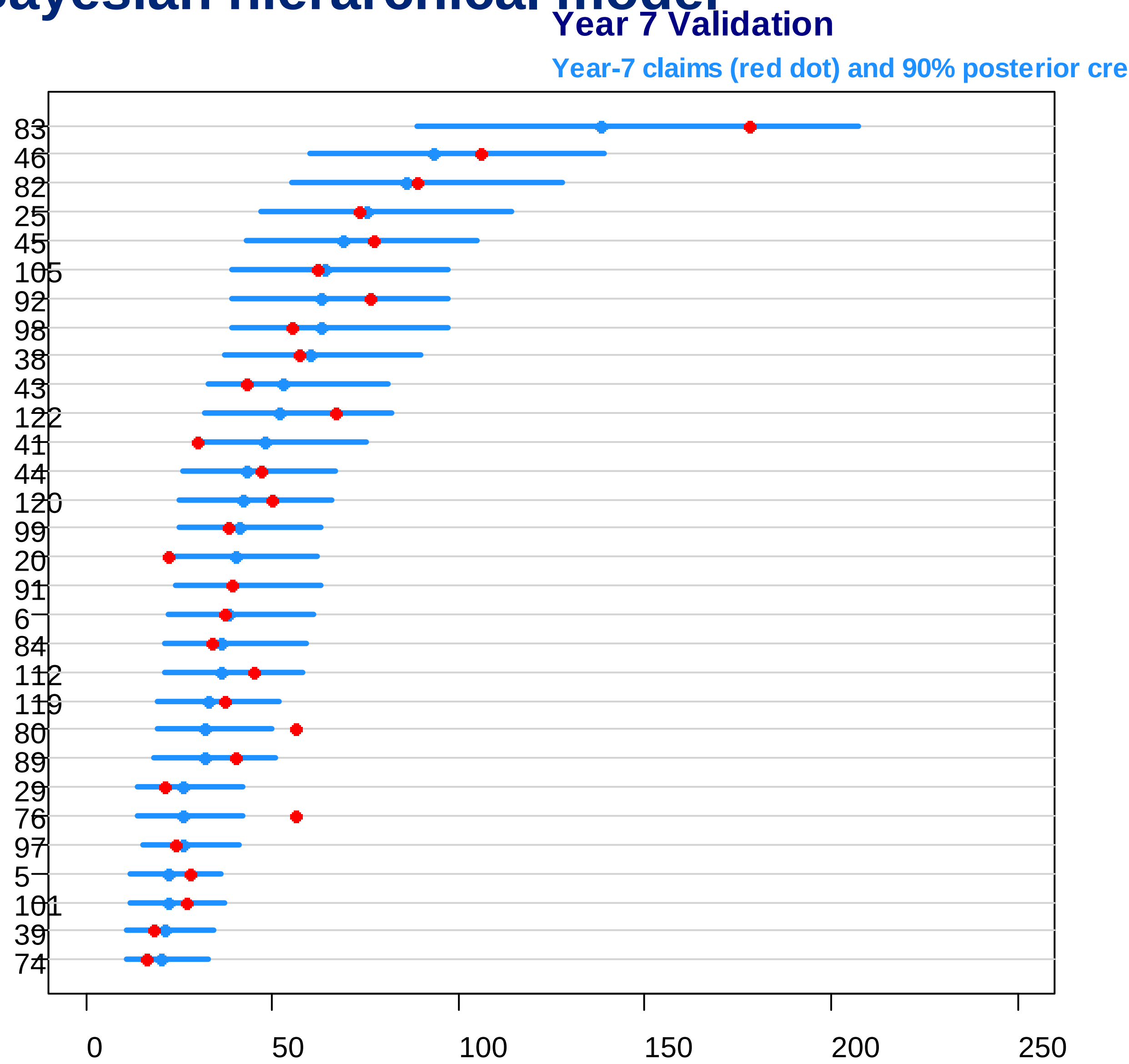
Shrinkage Effect of Hierarchical Models

- Illustration: estimating workers compensation claim frequency by industry class.
- Poisson hierarchical model (aka "credibility model")



Validating the fully Bayesian hierarchical model

Roughly 90% of the claims from the validation time period fall within the 90% posterior credible interval.



Case Study:
A Fully Bayesian Model

Collaboration with Wayne Zhang and Vanja Dukic

Data

A garden-variety Workers Compensation Schedule P loss triangle:

Cumulative Losses in 1000's

AY	*premium*	*12*	*24*	*36*	*48*	*60*	*72*	*84*	*96*	*108*	*120*	*CL Ult*	*CL LR*	*CL res*
1988	2,609	404	986	1,342	1,582	1,736	1,833	1,907	1,967	2,006	2,036	2,036	0.78	0
1989	2,694	387	964	1,336	1,580	1,726	1,823	1,903	1,949	1,987		2,017	0.75	29
1990	2,594	421	1,037	1,401	1,604	1,729	1,821	1,878	1,919			1,986	0.77	67
1991	2,609	338	753	1,029	1,195	1,326	1,395	1,446				1,535	0.59	89
1992	2,077	257	569	754	892	958	1,007					1,110	0.53	103
1993	1,703	193	423	589	661	713						828	0.49	115
1994	1,438	142	361	463	533							675	0.47	142
1995	1,093	160	312	408								601	0.55	193
1996	1,012	131	352									702	0.69	350
1997	976	122										576	0.59	454
chain link		2.365	1.354	1.164	1.090	1.054	1.038	1.026	1.020	1.015	1.000	12,067		1,543
chain ldf		4.720	1.996	1.473	1.266	1.162	1.102	1.062	1.035	1.015	1.000			
growth curve		21.2%	50.1%	67.9%	79.0%	86.1%	90.7%	94.2%	96.6%	98.5%	100.0%			

- Let's model this as a longitudinal dataset.
- Grouping dimension: Accident Year (AY)

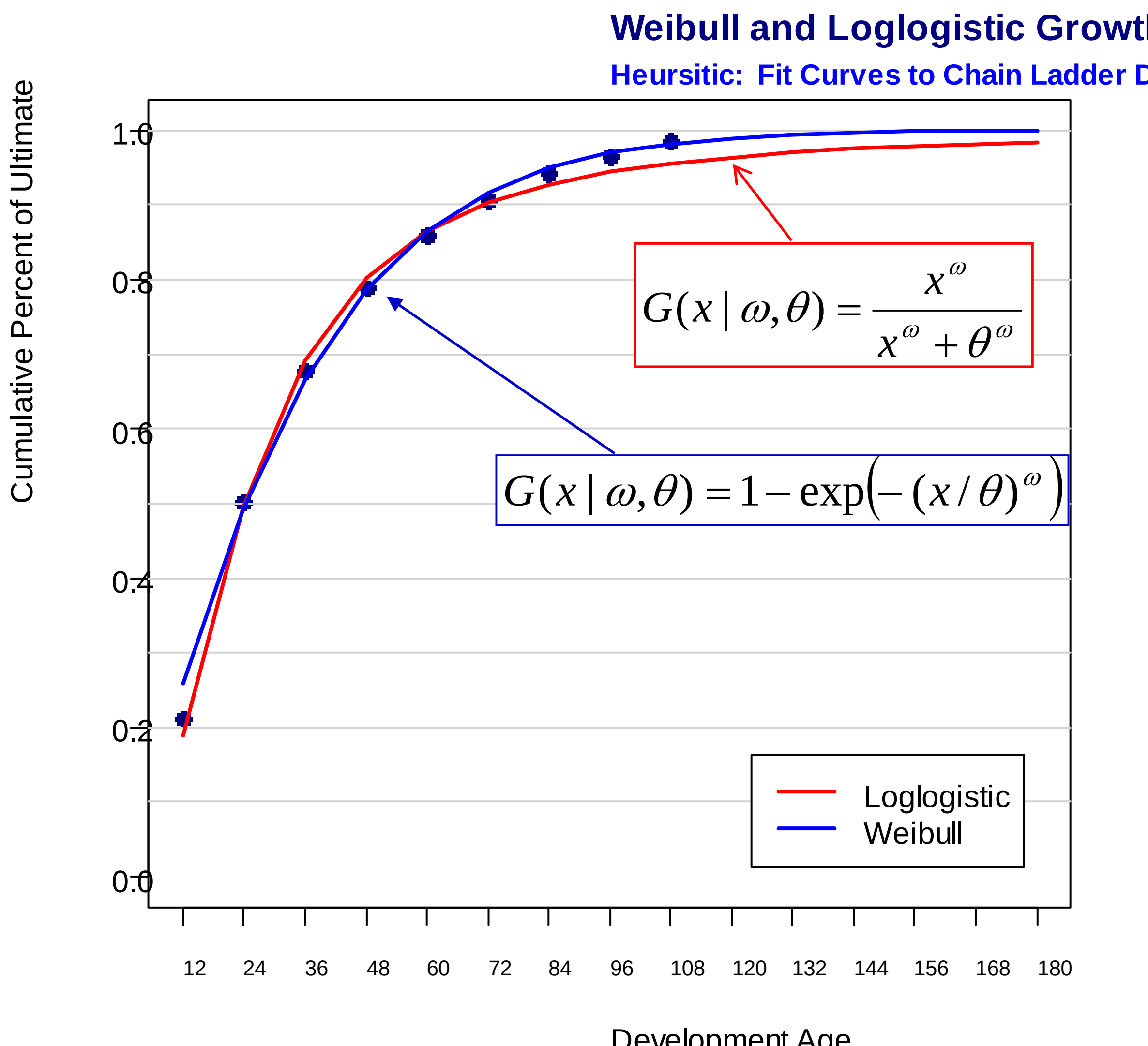
We can build a parsimonious non-linear model that uses random effects to allow the model parameters to vary by accident year.

Growth curves – at the heart of the model

- We want our model to reflect the **non-linear** nature of loss development.
  - GLM shows up a lot in the stochastic loss reserving literature…
  - … but are GLMs natural models for loss triangles?

- Growth curves (Clark 2003)
  - γ = ultimate loss ratio
  - θ = scale
  - ω = shape ("warp")

- Heuristic idea
  - We judgmentally select a growth curve form
  - Let γ vary by year (hierarchical)
  - Add priors to the hyperparameters (Bayesian)



An exploratory non-Bayesian hierarchical model

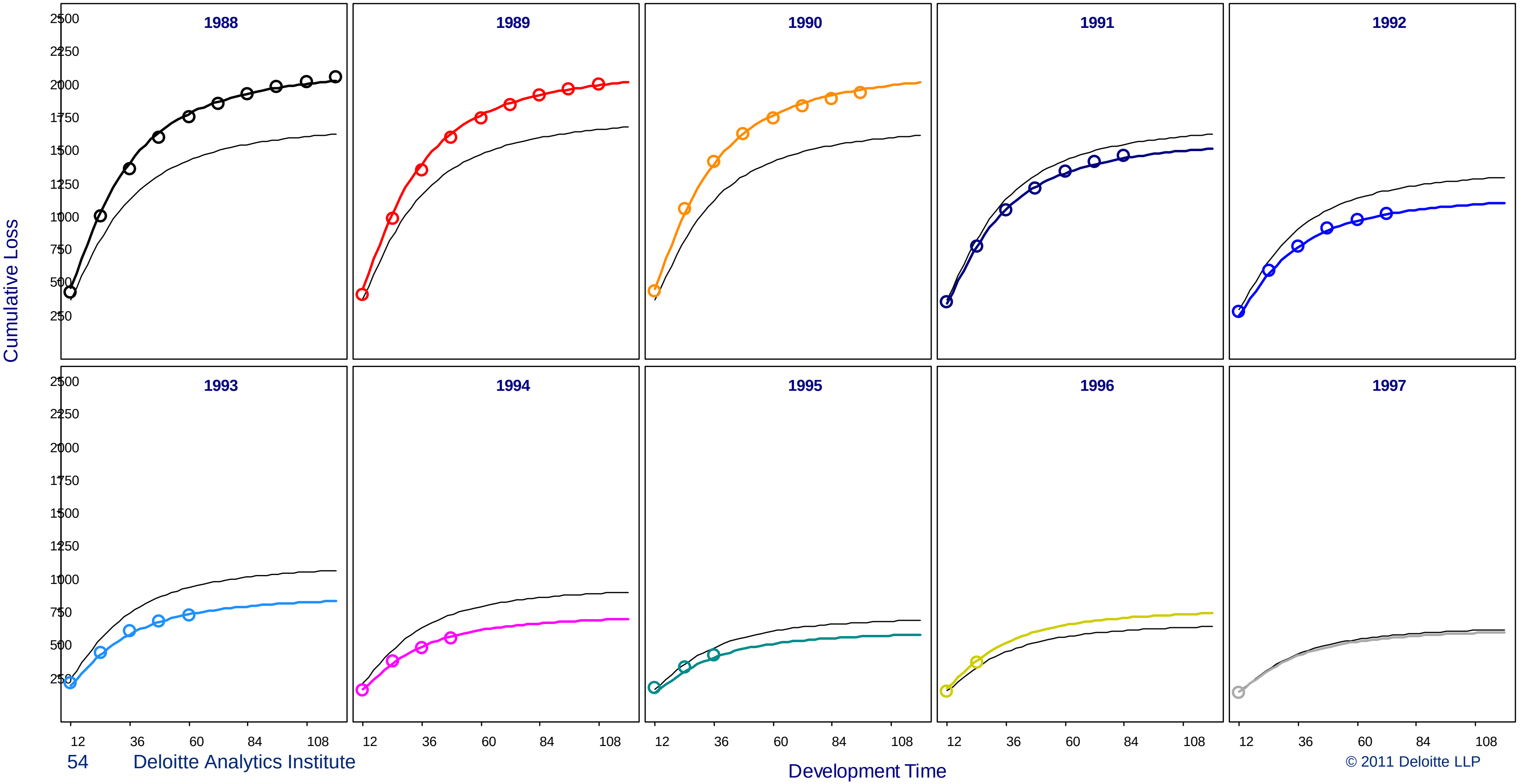
It is easy to fit non-Bayesian hierarchical models as a data exploration step.

$$y_i(t_j) = \gamma_i * p_i * \left(\frac{t^{\omega}}{t^{\omega} + \theta^{\omega}} \right) + \varepsilon_i(t_j)$$

$$\gamma_i \sim N\left(\gamma, \sigma_{\gamma}^2\right)$$

$$\varepsilon_i(t_j) = \rho \varepsilon_i(t_{j-1}) + \delta_i(t_j)$$

Log-Loglistic Hierarchical Model (non-Bayesian)



Adding Bayesian structure

- Our hierarchical model is "half-way Bayesian"
  - On the one hand, we place probability sub-models on certain parameters
  - But on the other hand, various (hyper)parameters are estimated directly from the data.

- To make this fully Bayesian, we need to put probability distributions on **all** quantities that are uncertain.

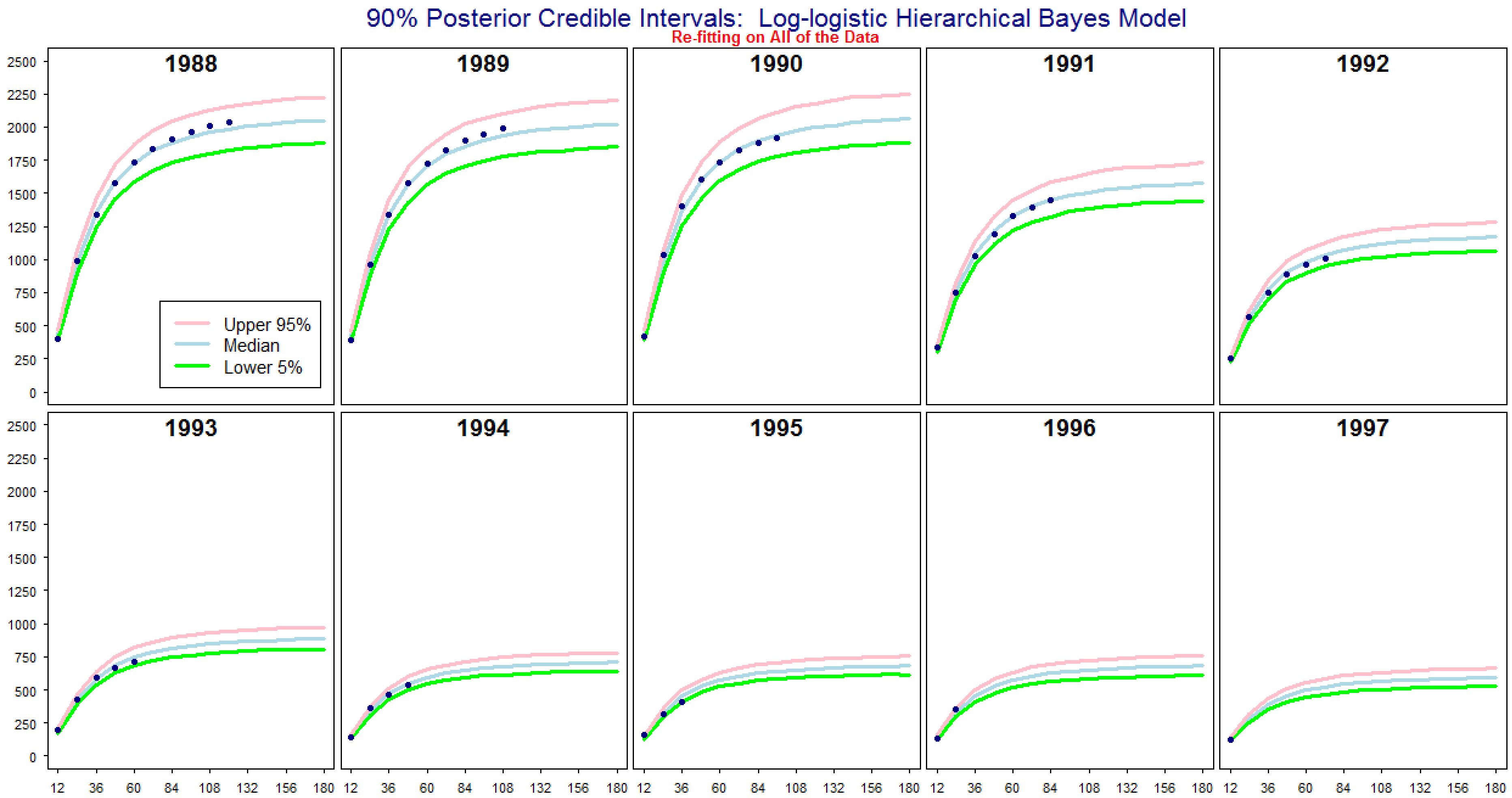
- We then employ Bayesian updating: the model ("likelihood function") together with the prior results in a posterior probability distribution over **all** uncertain quantities.
  - Including ultimate loss ratio parameters and hyperparameters!

  ➔ *We are directly modeling the ultimate quantity of interest.*

- This is not as hard as it sounds:
  - We do **not** explicitly calculate the high-dimensional posterior probability distribution.
  - We **do** use Markov Chain Monte Carlo [MCMC] simulation to sample from the posterior.
  - Technology: JAGS ("Just Another Gibbs Sampler"), called from within R.

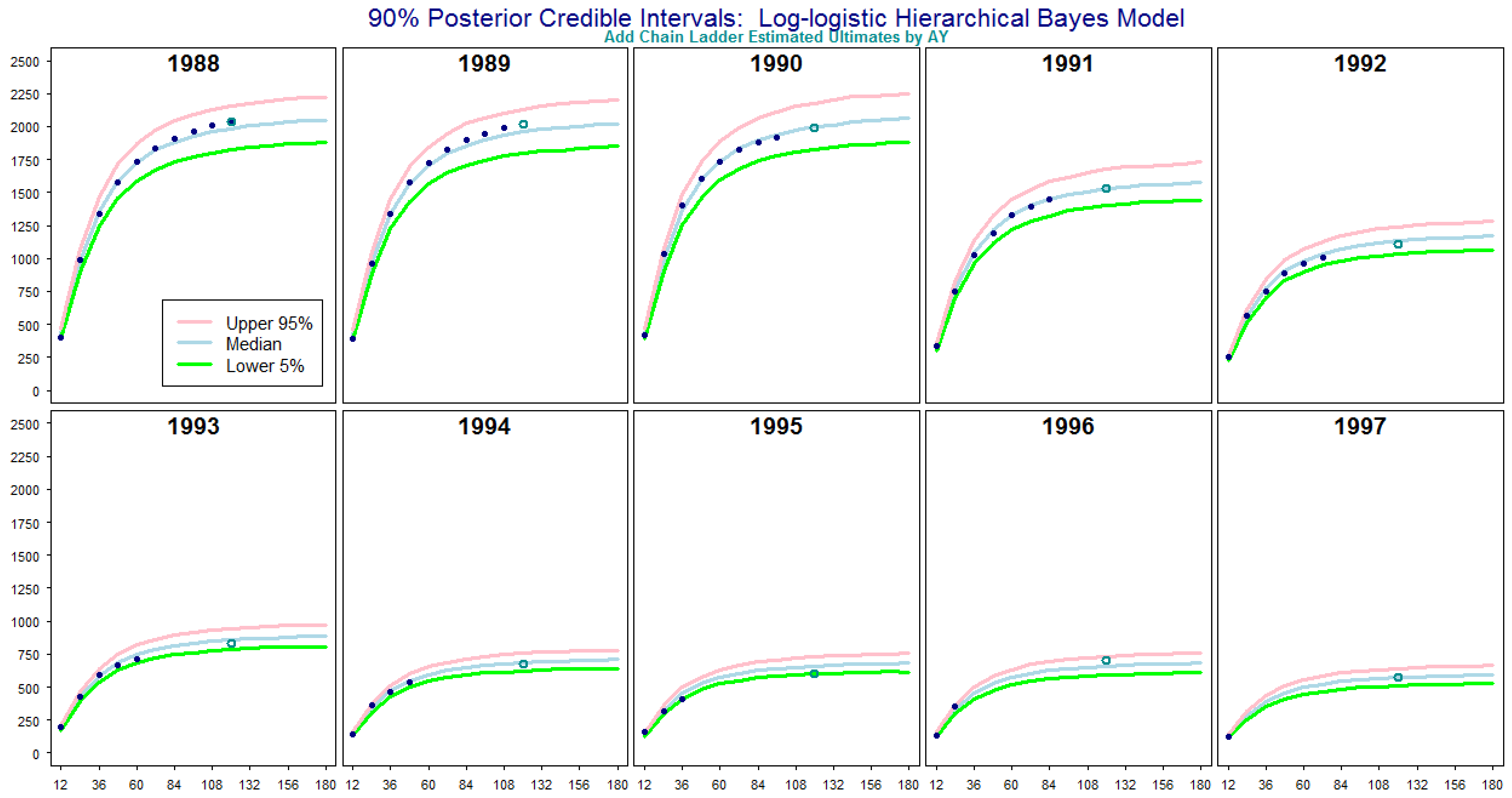
Example *(with Wayne Zhang and Vanja Dukic)*

- Posterior credible intervals of incremental losses – by accident year
  - Based on non-linear hierarchical growth curve model



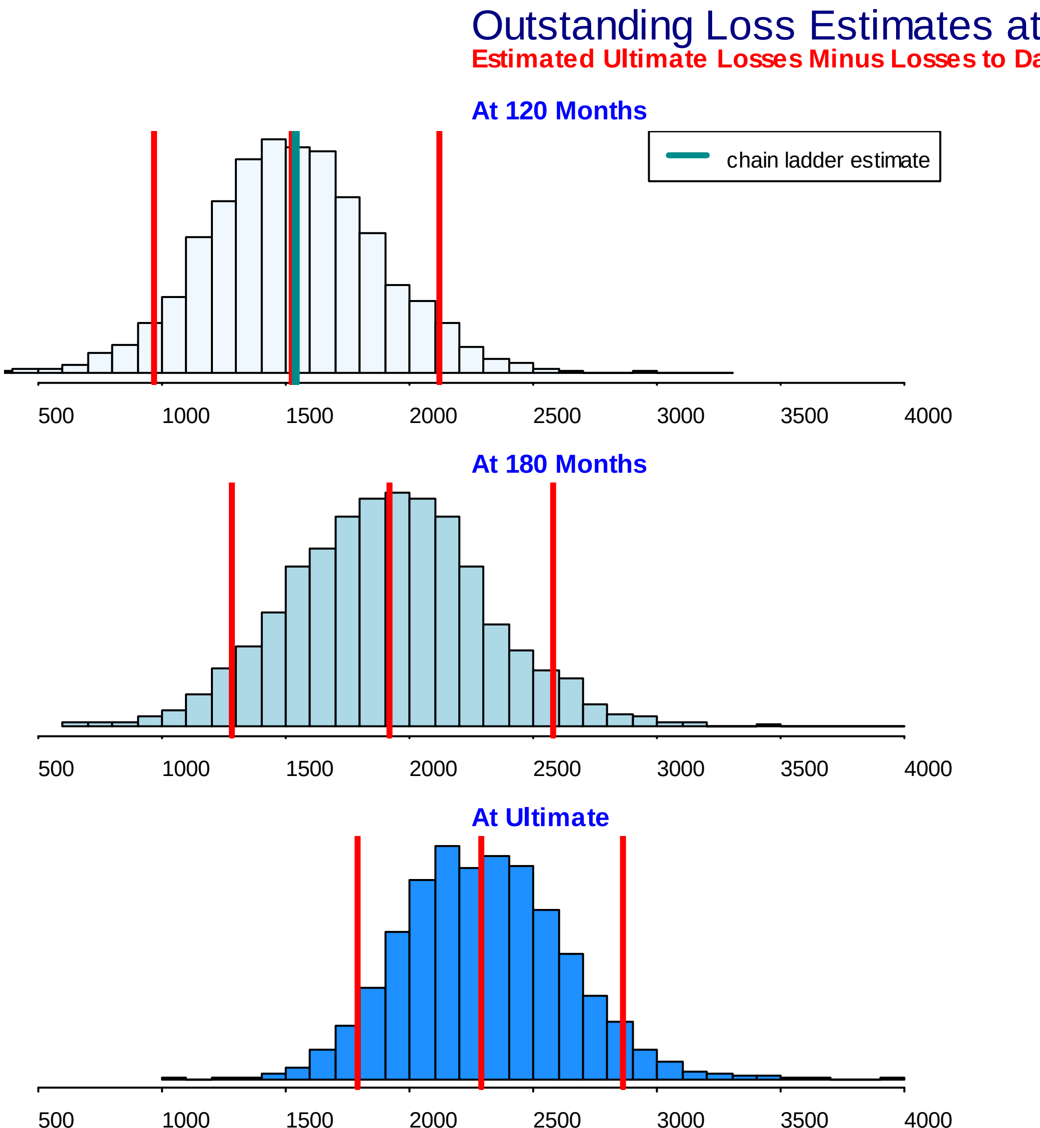
Example (with Wayne Zhang and Vanja Dukic)

- Posterior credible intervals of incremental losses – by accident year
  - Based on non-linear hierarchical growth curve model



Posterior distribution of aggregate outstanding losses

- Non-informative priors were used
  - Different priors tested as a sensitivity analysis
- A full posterior distribution falls out of the analysis
  - No need for boostrapping, ad hoc simulations, settling for a point estimate with a confidence interval
- Use of non-linear (growth curve) model enables us to project beyond the range of the data
  - Choice of growth curves affects the estimates more than the choice of priors!
  - This choice "does the work of" a choice of tail factors



Why Bayes

- "A coherent integration of evidence from different sources"
  - Background information
  - Expert knowledge / judgment ("subjectivity" is a feature, not a bug)
  - Other datasets (e.g. multiple triangles)
  - Shrinkage, "borrowing strength", hierarchical model structure – all coin of the realm
- Rich output: full probability distribution estimates of <u>all</u> quantities of interest
  - Ultimate loss ratios by accident year
  - Outstanding loss amounts
  - Missing values of any cell in a loss triangle
- Model the process that generates the data
  - As opposed to modeling the data with "procedural" methods
  - We can fit models as complex (or simple) as the situation demands
  - Nonlinear growth patterns, trends, autoregressive, hierarchical, structure, …
- Conceptual clarity
  - Single-case probabilities make sense in the Bayesian framework
  - Communication of risk: "mean what you say and say what you mean"

A Parting Thought

Parting thought: our field's Bayesian heritage

"Practically all methods of statistical estimation... are based on... the assumption that any and all collateral information or a priori knowledge is worthless. It appears to be only in the actuarial field that there has been an organized revolt against discarding all prior knowledge when an estimate is to be made using newly acquired data."

-- Arthur Bailey (1950)

Parting thought: our field's Bayesian heritage

"Practically all methods of statistical estimation… are based on… the assumption that any and all collateral information or a priori knowledge is worthless. It appears to be only in the actuarial field that there has been an organized revolt against discarding all prior knowledge when an estimate is to be made using newly acquired data."

-- Arthur Bailey (1950)

... And today, in the age of MCMC, cheap computing, and open-source software...

Parting thought: our field's Bayesian heritage

"Practically all methods of statistical estimation... are based on... the assumption that any and all collateral information or a priori knowledge is worthless. It appears to be only in the actuarial field that there has been an organized revolt against discarding all prior knowledge when an estimate is to be made using newly acquired data."

-- Arthur Bailey (1950)

... And today, in the age of MCMC, cheap computing, and open-source software...

"Scientific disciplines from astronomy to zoology are moving to Bayesian data analysis. We should be leaders of the move, not followers."

-- John Kruschke, Indiana University Psychology (2010)

Appendix:
Some MCMC Intuition

Metropolis-Hastings Intuition

- Let's take a step back and remember why we've done all of this.
- In ordinary Monte Carlo integration, we take a large number of independent draws from the probability distribution of interest and let the sample average of $\{g(\theta_i)\}$ approximate the expected value $E[g(\theta)]$.
- The Strong Law of Large Numbers justifies this approximation.
- But: when estimating Bayesian posteriors, we are generally not able to take independent draws from the distribution of interest.
- Results from the theory of stochastic processes tell us that suitably well-behaved Markov Chains can *also* be used to perform Monte Carlo integration.

Some Facts from Markov Chain Theory

How do we know this algorithm yields reasonable approximations?

- Suppose our Markov chain θ_1, θ_2, … with transition matrix *P* satisfies some "reasonable conditions":
  - Aperiodic, irreducible, positive recurrent (see next slide)
  - Chains generated by the M-H algorithm satisfy these conditions

- **Fact #1 (convergence theorem):** *P* has a unique stationary ("equilibrium") distribution, π. (i.e. $\pi=\pi P$). Furthermore, the chain converges to π.
  - Implication: We can start anywhere in the sample space so long as we through out a sufficiently long "burn-in".

- **Fact #2 (Ergodic Theorem):** suppose $g(\theta)$ is some function of θ. Then:

$$\boxed{\frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)}) \quad \underset{N\to\infty}{\rightarrow} \quad \int g(\theta)\pi(\theta)d\theta = E\left[g(\theta)\right]}$$

  - Implication: After a sufficient burn-in, perform Monte Carlo integration by averaging over a suitably well-behaved Markov chain.
  - The values of the chain are *not* independent, as required by the SLLN.
  - But the Ergodic Theorem says we're close enough to independence to get what we need.

Conditions for Ergodicity

More on those “reasonable conditions” on Markov chains:

- **Aperiodic:** The chain does not regularly return to any value θ in the state space in multiples of some k>1.
- **Irreducible:** It is possible to go from any state θ_i to any other state θ_j in some finite number of steps.
- **Positive recurrent:** The chain will return to any particular state θ with probability 1, and expected return time finite.
- *Intuition:*
  - *The Ergodic Theorem tells us that (in the limit) the amount of time the chain spends in a particular region of state space equals the probability assigned to that region.*
  - *This won’t be true if (for example) the chain gets trapped in a loop, or won’t visit certain parts of the space in finite time.*
- *The practical problem:* use the Markov chain to select a representative sample from the distribution π, expending a minimum amount of computer time.