A generic approach to MCMC for parametric IRT[a]

Educational Testing Service
Princeton, NJ
July 8, 1998

Brian W. Junker
Department of Statistics
Carnegie Mellon University
Pittsburgh PA 15213
brian@stat.cmu.edu

[a]Part of this paper is based on Richard J. Patz's Ph.D. dissertation at Carnegie Mellon University, which was supported by a Spencer Foundation Dissertation Fellowship in Educational Research. Junker's work was supported by NSF grant #DMS-97.05032.

Talk Outline:

- Introduction
- Being practical with Bayes
- Example: The hierarchical beta-binomial model
- How it all works for IRT models
- Examples of what you get
- Some other applications

Introduction

- IRT is an incredibly successful statistical modeling and data analysis methodology.
  - Well-known measurement and scaling properties.
  - Greatly extended the data analytic reach of psychometricians, social scientists, and educational measurement specialists.
  - Possible to estimate group effects and to incorporate covariates and other structure.
- One model can go from students' raw responses all the way up to hierarchical multigroup/multisite structure of educational survey data.
  - Mislevy (1985); Mislevy and Sheehan (1989); Algina (1992); Zwick (1992); Johnson, Mislevy and Thomas (1994); Adams, Wilson and Wu (1997).

- Also applied in the analysis of item-level data in controlled experiments in medicine and psychology.
  - Carlin et al. (1992), DiBello et al. (1995), Fischer and Ponocny (1994), Huguenard, et al., (1996), Kelderman and Rijkes (1994) and Legler and Ryan (1997), in addition to the many examples in Fischer and Molenaar (1995) and van der Linden and Hambleton (1997).
- As the applications become more complex, so too become the IRT based models.
  - Complex item- and person- sampling designs in large scale educational assessment surveys.
  - Experimental conditions, subject covariates, item covariates, in psychological experiments.
  - Many sources of uncertainty and fairly complex dependency structures.

- Bayes "used to be" only for simple problems because the computational challenges for complex problems have been too severe.

- Applications of IRT in educational testing itself have successfully pushed E-M very far.

- Bayesian computation has recently undergone a revolution through Markov Chain Monte Carlo (MCMC) methods.

- Models of almost any complexity are conceptually easy to set up for MCMC. This makes MCMC ideal for exploring newly-developed models.

- MCMC produces more than, and is usually slower than, E-M: Whole distribution vs. mean and standard deviation.

- "Large" problems probably require a mix of MCMC, E-M and other methods, to be fast enough to be practical.

Being practical with Bayes

Data: X

[maybe responses, $X = (X_{i1}, \ldots, X_{iJ} : i = 1, \ldots, N)$];

Parameters: τ

[maybe item and person parameters, $\tau = (\theta\text{'s}, \beta\text{'s})$].

Basic identities:

- $p(X, \tau) \quad = \quad p(X|\tau) \cdot p(\tau)$

- $$\begin{aligned} p(\tau|X) &= \frac{p(X|\tau) \cdot p(\tau)}{\int p(X|t) \cdot p(t)\ dt} \\ &\propto p(X|\tau) \cdot p(\tau) \quad = \quad p(X, \tau) \end{aligned}$$

We always want to know something about $p(\tau|X)$:

- EAP: $argmax_\tau\, p(\tau|X)$
- MAP: $E[\tau|X] = \int \tau\, p(\tau|X)\, d\tau$

  (or some other risk-minimizer)
- CI: Find a set A such that $p(\tau \in A|X) = 0.95$
- Graph or "shape" of $p(\tau|X)$

Fundamental difficulty:

All of these operations require that you know a lot about the shape of $p(\tau|X)$.

> *How do you learn enough about the shape of $p(\tau|X)$?*

- If $p(\tau)$ is *conjugate* to $p(X|\tau)$, then $p(\tau|X)$ will belong to the same family as $p(\tau)$.
- If not, then you have a numerical integration problem. These problems are difficult because they depend on the "interesting" features of $p(\tau|X)$.
  - Difficult but do-able if $p(\tau|X)$ is *smooth* and *low-dimensional.*
  - Even if $p(\tau|X)$ is smooth, space becomes too large, as $\dim(\tau)$ grows, for deterministic methods. *Curse of dimensionality.*

- Monte Carlo (MC) methods have long been used to help with the curse. With a random sample $\tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(n)}$, can plot histograms (shape), compute averages (integration), etc.
  - Improves convergence to $O(1/\sqrt{n})$
  - Less smoothing of interesting features as n grows
  - Curse returns: how do you sample from $p(\tau|X)$?
- Markov Chain Monte Carlo (MCMC) improves on MC:
  - Replace draws from $p(\tau|X)$ with a sequence of low-dimensional draws from $p_1(\tau_1)$, $p_2(\tau_2)$, etc.
  - For Bayes, often can exploit "partial" conjugacy to simplify problem of drawing from $p(\tau_{(k)})$.
- Cost:
  - τ's from MC are iid. Easy statistics, "fast" convergence.
  - τ's from MCMC are dependent. Harder statistics, "slow" convergence.

MCMC: The Cliff's Notes

The problem:

- Learn about $\pi(\tau) = \pi(\theta, \beta, \tau)$, where $\tau = (\theta, \beta, \lambda)$ is some high-dimensional set of variables (parameters).

The essential idea:

- Define a (stationary) Markov chain $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots$ with states $\mathcal{M}_k = (\theta^k, \beta^k, \lambda^k)$.

- Simulate the $\mathcal{M}_k$'s; under regularity conditions (e.g., Tierney, 1994), $\mathcal{M}_k$ will converge in distribution to a *stationary distribution* $\pi(\theta, \beta, \lambda)$:

$$\begin{aligned} &\pi(\theta^1, \beta^1, \lambda^1) \\ &\quad = \int P[\mathcal{M}_{k+1} = (\theta^1, \beta^1, \lambda^1) \mid \mathcal{M}_k = (\theta^0, \beta^0, \lambda^0)] \\ &\qquad\qquad \times \ \pi(\theta^0, \beta^0, \lambda^0)\, d(\theta^0, \beta^0, \lambda^0) \end{aligned}$$

- For Bayes, design chain so that $\pi(\theta, \beta, \lambda)$ turns out to be the posterior distribution $p(\theta, \beta, \lambda | X)$.

MCMC: The Cliff's Notes

The magic:

- Write $p(\theta, \beta, \lambda | X) = p(\tau_1, \tau_2, \ldots, \tau_M | X)$, where $(\tau_1, \ldots, \tau_M)$ is a disjoint partition of the original model parameters (θ, β, λ).

- General theory of MCMC (e.g., Tierney, 1994; Chib and Greenberg, 1995): construct state $\mathcal{M}_k = (\tau_1^{(k)}, \ldots, \tau_M^{(k)})$, by sampling each τ_m from its "complete conditional" distribution:

  > *To step from* $\mathcal{M}_{k-1} = (\tau_1^{(k-1)}, \ldots, \tau_M^{(k-1)})$ *to* $\mathcal{M}_k = (\tau_1^{(k)}, \ldots, \tau_M^{(k)})$*:*
  >
  > *1.* $\tau_1^{(k)} \sim p(\tau_1 | \tau_2^{(k-1)}, \ldots, \tau_M^{(k-1)}, X)$;
  >
  > *2.* $\tau_2^{(k)} \sim p(\tau_2 | \tau_1^{(k)}, \tau_3^{(k-1)}, \ldots, \tau_M^{(k-1)}, X)$;
  >
  > *3.* $\tau_3^{(k)} \sim p(\tau_3 | \tau_1^{(k)}, \tau_2^{(k)}, \tau_4^{(k-1)}, \ldots, \tau_M^{(k-1)}, X)$;
  >
  > $\vdots \quad \vdots \quad \vdots \quad \vdots$
  >
  > *M.* $\tau_M^{(k)} \sim p(\tau_M | \tau_1^{(k)}, \tau_2^{(k)}, \tau_3^{(k)}, \ldots, \tau_{M-1}^{(k)}, X,)$;

  This is the basic idea behind all so-called "Gibbs samplers."

MCMC: The Cliff's Notes

Some algorithmic details:

- Once a sampling scheme is decided upon:
  - *If you can* sample τ_m directly from its complete conditional distribution, do so;
  - *Otherwise* perform a "Metropolis step":
    * Sample a "proposal value" τ_m^* from any convenient proposal distribution $q_k(\tau_m|\tau_m^{(k-1)})$
    * Compute the "acceptance probability" α_m^* from $q_k(\tau_m|\tau_m^{(k-1)})$ and the complete conditional $p(\tau_m|rest) \equiv p(\tau_m \mid \tau_1^{(k)}, \ldots, \tau_{m-1}^{(k)}, \tau_{m+1}^{(k-1)}, \ldots, \tau_M^{(k-1)}, X)$:

      $$\alpha_n^* = \min\left\{ \frac{p(\tau_n^*|rest)q_k(\tau_n^{(k-1)}|\tau_n^*)}{p(\tau_n^{(k-1)}|rest)q_k(\tau_n^*|\tau_n^{(k-1)})}, 1 \right\};$$

    * Accept $\tau_n^{(k)} = \tau_n^*$ with probability α_n^*; otherwise set $\tau_n^{(k)} = \tau_n^{(k-1)}$.

This is "Metropolis-Hastings within Gibbs"

Key observation: (IRT motivation; true in general!)

Note that when $\tau = (\theta\text{'s}, \beta\text{'s})$ only,

$$p(X|\beta\text{'s}) = \int p(X|\theta, \beta\text{'s}) \cdot p(\theta|\beta\text{'s})\ d\theta$$

is used for MML estimates of the β's.

A standard "IRT textbook practice" would be to estimate β's by maximizing $p(X|\beta\text{'s})$ and then use $p(\theta|X, \beta)$ for, e.g., EAP estimates of the θ's.

Suppose the θ's and β's have prior distributions $p(\theta)$, $p(\beta)$, and we want EAP estimates of the θ's given the β's. We get

$$\begin{aligned} p(\theta|X, \beta) &= \frac{p(X|\theta, \beta)p(\theta)\ p(\beta)}{\int p(X|t, \beta)p(t)p(\beta)\, dt\, d\beta} \\ &\propto p(X|\theta, \beta)p(\theta) \end{aligned}$$

The shape of $p(\theta|X, \beta)$ is determined by just the parts of the likelihood that depend explicitly on θ.

Example: The hierarchical beta-binomial model

Basic beta-binomial model, for examinees $i = 1, \ldots .N$, each getting y_i of n_i questions right. $[X = \underline{y} = (y_1, \ldots, y_N)]$

1. $y_i|\theta_i \sim Bin(\theta_i, n_i)$;
2. $\theta_i|\alpha, \beta \sim Beta(\alpha, \beta)$

Full model $p(\underline{y}, \underline{\theta}|\alpha, \beta)$ is

$$\prod_{i=1}^{N} \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}$$

Use the "key observation" to calculate

$$p(\theta_i|y_i, \alpha, \beta) \propto \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}$$

i.e. $\theta_i|y_i \sim Beta(\alpha + y_i, \beta + n_i - y_i)$.

Choose fixed values for α, β, then use $Beta(\alpha + y_i, \beta + n_i - y_i)$ for EAP, posterior CI's, or whatever.

No need for anything fancy yet.

What if we want to estimate α, β (estimate the shape of the population / latent distribution)? Add a third modeling assumption:

3. $\alpha, \beta \sim p(\alpha, \beta)$

Now the full model $p(\underline{y}, \underline{\theta}, \alpha, \beta)$ is

$$\prod_{i=1}^{N} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i} \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1} (1-\theta_i)^{\beta-1} p(\alpha, \beta)$$

Can again use the "key observation" to calculate

$$\begin{aligned} p(\theta_i | y_i, \alpha, \beta) &= Beta(\theta_i | \alpha + y_i, \beta + n_i - y_i) \\ p(\alpha, \beta | \underline{y}) &\propto p(\alpha, \beta) \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)} \end{aligned}$$

Gelman et al. (1995, pp. 131–133) suggest computing $p(\alpha, \beta | \underline{y})$—or actually $p(\log(\alpha/\beta), \log(\alpha+\beta) | \underline{y})$—on a grid, using trial and error to place the grid over the "interesting" part of the density.

An MCMC solution:

From the full model $p(\underline{y}, \underline{\theta}, \alpha, \beta)$ is

$$\prod_{i=1}^{N} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i} \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1} (1-\theta_i)^{\beta-1} p(\alpha, \beta)$$

the "key observation" gives the complete conditionals

$$\begin{aligned}
p(\theta_i | rest) &= Beta(\theta_i | \alpha + y_i, \beta + n_i - y_i) \\
p(\alpha | rest) &\propto \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \right]^N \prod_{i=1}^{N} \theta_i^{\alpha} p(\alpha, \beta) \\
p(\beta | rest) &\propto \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} \right]^N \prod_{i=1}^{N} (1 - \theta_i)^{\beta} p(\alpha, \beta)
\end{aligned}$$

- Canned routines sample $\theta_i \sim Beta(\cdots)$
- Can set up Metropolis steps for α and β using Normal proposal draws ("random walk M-H"). Normal variances are "tuning parameters".

How it all works for IRT models

A hierarchical IRT setup:

The raw materials:

- $\{X_{ij} : \ i = 1, \ldots, N; \ j = 1, \ldots, J\}$ item responses (I subjects, J items)
- θ_i latent variables
- Latent distribution param's $\lambda = (\lambda_1, \ldots, \lambda_L)$
- β_j item parameters
- Subject covariates $Y = (Y_1, \ldots, Y_I)$
- Item covariates $Z = (Z_1, \ldots, Z_J)$

The model:

- $X_{ij} \sim P(X_{ij} = 1 | \theta_i, \beta_j)$
- $\theta_i \sim f(\theta_i | \lambda, Y_i)$
- $\lambda \sim p(\lambda | Y)$
- $\beta_j \sim p(\beta_i | Z_j)$

An MCMC framework for this setup:

The full model $p(X, \theta, \lambda, \beta | Y, Z)$ is

$$\prod_{i=1}^{N} \prod_{j=1}^{J} P(\theta_i, \beta_j)^{x_{ij}} (1 - P(\theta_i, \beta_j))^{1-x_{ij}}$$
$$\times \quad p(\beta_j | Z_j)\, p(\theta_i | \lambda, Y_i)\, p(\lambda | Y)$$

So a rough-and-ready set of complete conditionals is

$$\begin{aligned}
p(\theta_i | rest) &\sim \prod_{j=1}^{J} P(\theta_i, \beta_j)^{x_{ij}} (1 - P(\theta_i, \beta_j))^{1-x_{ij}} p(\theta_i | \lambda, Y_i) \\
p(\beta_j | rest) &\sim \prod_{i=1}^{N} P(\theta_i, \beta_j)^{x_{ij}} (1 - P(\theta_i, \beta_j))^{1-x_{ij}} p(\beta_j | Z_j) \\
p(\lambda | rest) &\sim \prod_{i=1}^{N} p(\theta_i | \lambda, Y_i) p(\lambda | Y)
\end{aligned}$$

M-H within Gibbs:

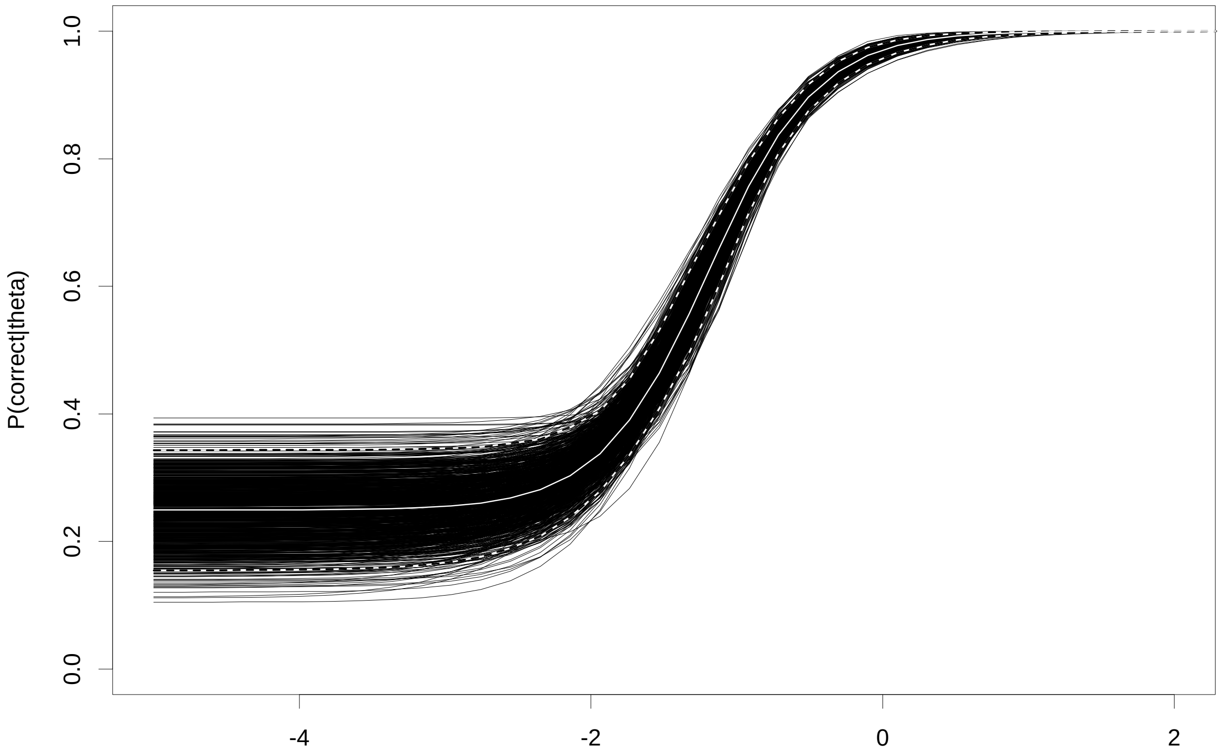
- θ_i, β_j are usually Metropolis steps.
- Congugacy can sometimes be exploited for λ. E.g. $\lambda = \sigma^2 \Rightarrow$ try Normal/Inverse-Gamma.

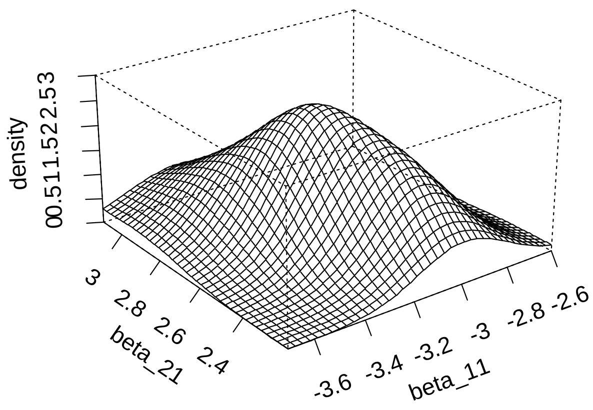
Other methods:

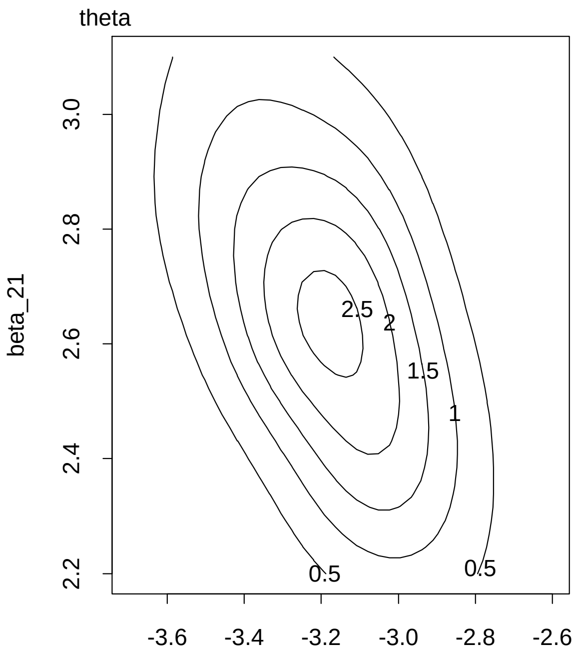
- Albert (1992): Normal Ogive with "data augmentation" leads to simpler complete conditionals; no Metropolis needed.
- BUGS (Spiegelhalter et al., 1996): estimate logistic IRT models like this, using "adaptive rejection" in place of Metropolis.

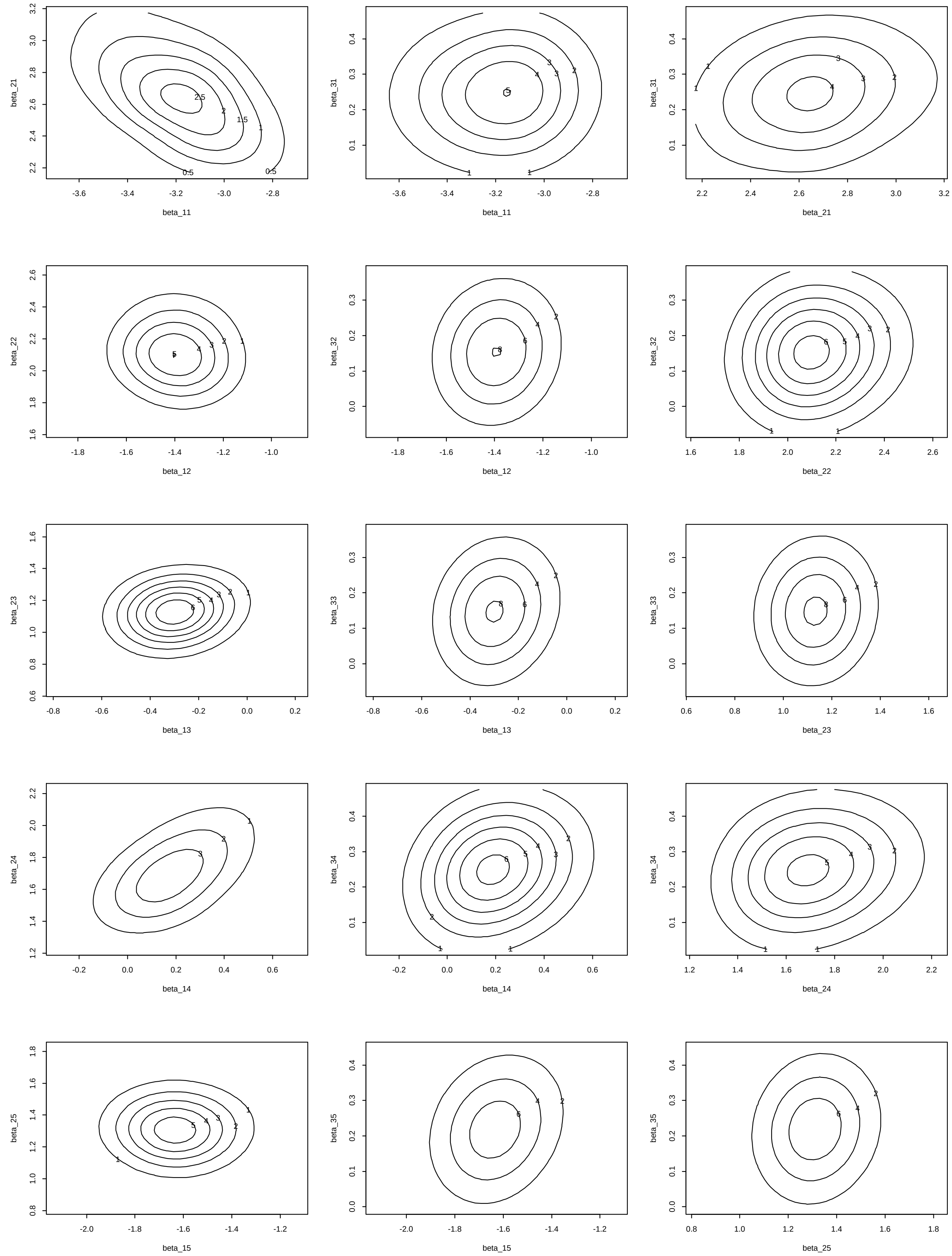
Examples of what you get

(Data extract from NAEP 1992 Trial State Reading Assessment)

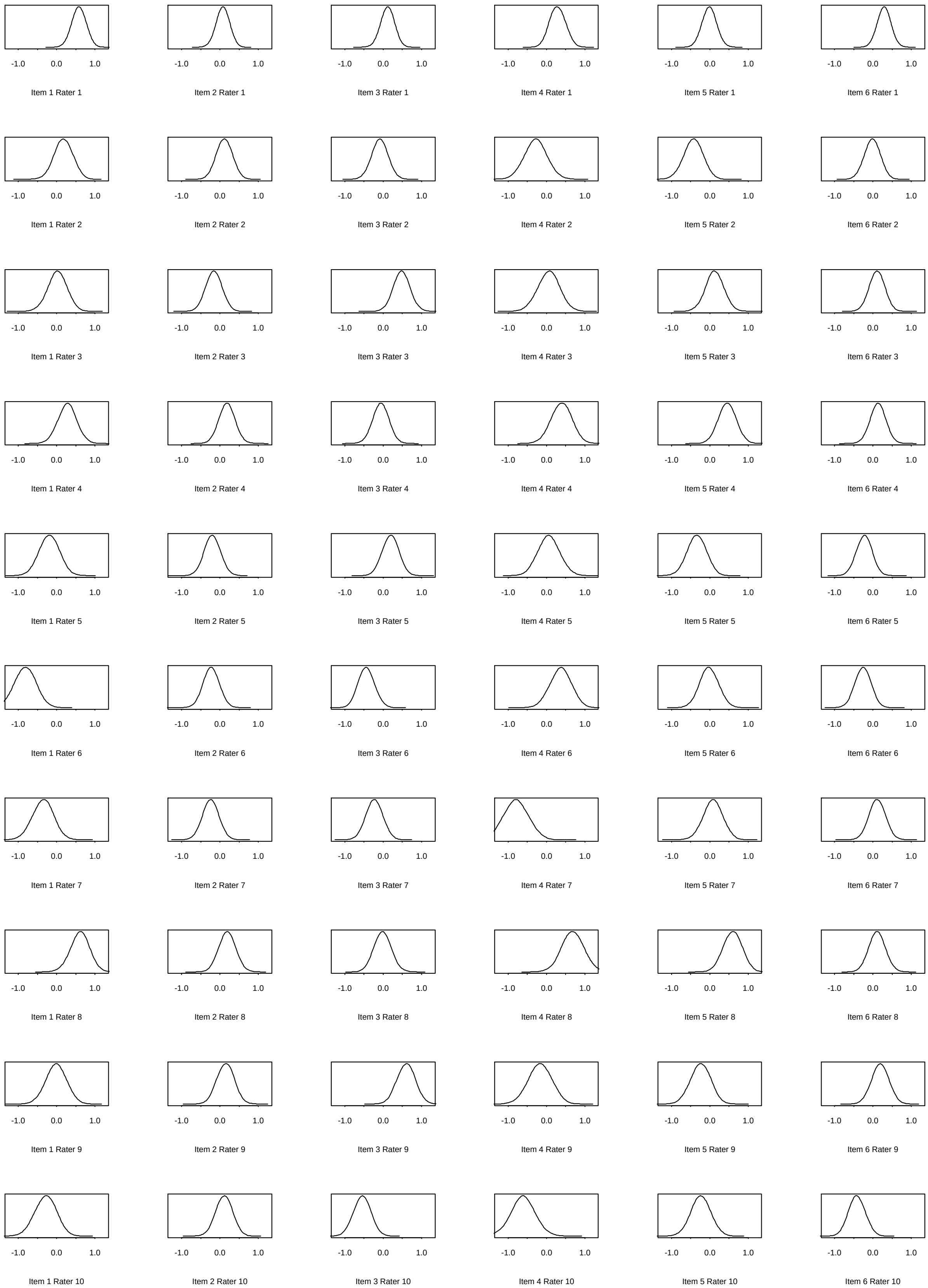








beta_21
beta_11
beta_31
beta_11
beta_31
beta_21
beta_22
beta_12
beta_32
beta_12
beta_32
beta_22
beta_23
beta_13
beta_33
beta_13
beta_33
beta_23
beta_24
beta_14
beta_34
beta_14
beta_34
beta_24
beta_25
beta_15
beta_35
beta_15
beta_35
beta_25


-1.0
0.0
1.0
Item 1 Rater 1
Item 2 Rater 1
Item 3 Rater 1
Item 4 Rater 1
Item 5 Rater 1
Item 6 Rater 1
Item 1 Rater 2
Item 2 Rater 2
Item 3 Rater 2
Item 4 Rater 2
Item 5 Rater 2
Item 6 Rater 2
Item 1 Rater 3
Item 2 Rater 3
Item 3 Rater 3
Item 4 Rater 3
Item 5 Rater 3
Item 6 Rater 3
Item 1 Rater 4
Item 2 Rater 4
Item 3 Rater 4
Item 4 Rater 4
Item 5 Rater 4
Item 6 Rater 4
Item 1 Rater 5
Item 2 Rater 5
Item 3 Rater 5
Item 4 Rater 5
Item 5 Rater 5
Item 6 Rater 5
Item 1 Rater 6
Item 2 Rater 6
Item 3 Rater 6
Item 4 Rater 6
Item 5 Rater 6
Item 6 Rater 6
Item 1 Rater 7
Item 2 Rater 7
Item 3 Rater 7
Item 4 Rater 7
Item 5 Rater 7
Item 6 Rater 7
Item 1 Rater 8
Item 2 Rater 8
Item 3 Rater 8
Item 4 Rater 8
Item 5 Rater 8
Item 6 Rater 8
Item 1 Rater 9
Item 2 Rater 9
Item 3 Rater 9
Item 4 Rater 9
Item 5 Rater 9
Item 6 Rater 9
Item 1 Rater 10
Item 2 Rater 10
Item 3 Rater 10
Item 4 Rater 10
Item 5 Rater 10
Item 6 Rater 10

In general MCMC gives you more information about the fit:

- "Exact" joint distribution of (θ's, β's, λ)
- "Exact" marginal distribution of any parameter(s)

(not just point estimates and standard errors!)

But it takes much longer

- In unpublished work, Chad Schafer shows that the typical time per iteration grows approximately like

$$\begin{aligned} & 8 \times [3 \cdot (\text{total number of parameters}) \\ & \quad + \quad (\text{items} + \text{examinees})] \quad \approx \quad O(NJ) \end{aligned}$$

- Schafer also shows that for 1000 examinees and 10 items, M-H/Gibbs takes:

Code	Machine	Time
Splus	Sun Sparcstation	1165 sec
C++	166MHz Linux PC	86 sec

- BUGS runs intermediate between these, but can take hours to deduce the complete conditionals.

So large-scale problems may require shortcuts such as...

- Mixture of E-M and MCMC
- Normal approximation to some complete conditionals

Some other applications

- Evaluating the educational impact of an introductory college statistics course.
- Modeling dependence in multiple-recapture census problems.
- Survey of 1489 firms in 72 US Industries to assess R&D behavior and efforts to protect and appropriate profits from R&D "products".
- Designing automatic data collection and analysis methods to help a computer-based cognitive tutor learn about errors in its student model from errors that students make.
- Confirming a cognitive model of algebra word problem problem solving.
- Novel models of raters for rated items.