

---

# Beyond MCMC (Back to the Future?)

---

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

# Outline

- A context: Model-based Psychometrics
- Review of standard approaches including MCMC
- What did MCMC replace/extend? E.g.
  - Gauss-Seidel, N-R, E-M, Quadrature & MC Integration, Search
- How is MCMC being replaced/extended? E.g.
  - MC-EM, Simulated Annealing, Particle Filtering
- A variational interpretation of E-M leads to E-M like algorithms that are faster than MCMC and nearly as simple to implement
- Fully model-based approaches run into computational bottlenecks...
  - Computational data manipulation motivated by (sufficient statistics of) relevant statistical models may offer a way out

---

# Entry Points into the Literature

- Merin, Mengersen & Robert (2004). Bayesian modeling and inference on mixtures of distributions. *Handbook of Statistics 25*, D. Dey and C.R. Rao (eds). Elsevier. (<http://www.ceremade.dauphine.fr/~xian>)
- Andrieu, de Freitas, Doucet & Jordan (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50, 5-43. (<http://www.springerlink.com/content/100309/>)
- Tom Minka's review and tutorial papers, at <http://research.microsoft.com/~minka/>

---

# A Context: Model-Based Psychometrics

$$y_{ij} \sim f(y|\theta_i, \beta_j, \psi)$$

- $i=1, \dots, N$ : subjects, persons, students, respondents
- $j=1, \dots, J$ : stimuli, items, questions, tasks
- $\theta_i$ : characterize persons' behavior across stimuli
- $\beta_j$ : characterize stimuli across people
- $\psi$ : any incidental structural parameters

# Many Examples...

- Factor Analysis, e.g.

$$y_{ij} \sim N(\lambda_j^T \theta_i, \psi_i), \quad \theta_i \in \mathcal{R}^K, \quad \beta_j = \lambda_j \in \mathcal{R}^K$$

- Discrete-response models

$$y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad p_{ij} = g(\theta_i, \beta_j)$$

- Item Response Theory, e.g.

$$g(\theta_i, \beta_j) = 1 / \{1 + \exp[a_j(\theta_i - b_j)]\}, \quad \theta_i \in \mathcal{R}, \beta_j = (a_j, b_j)$$

- Latent Class Models, e.g.

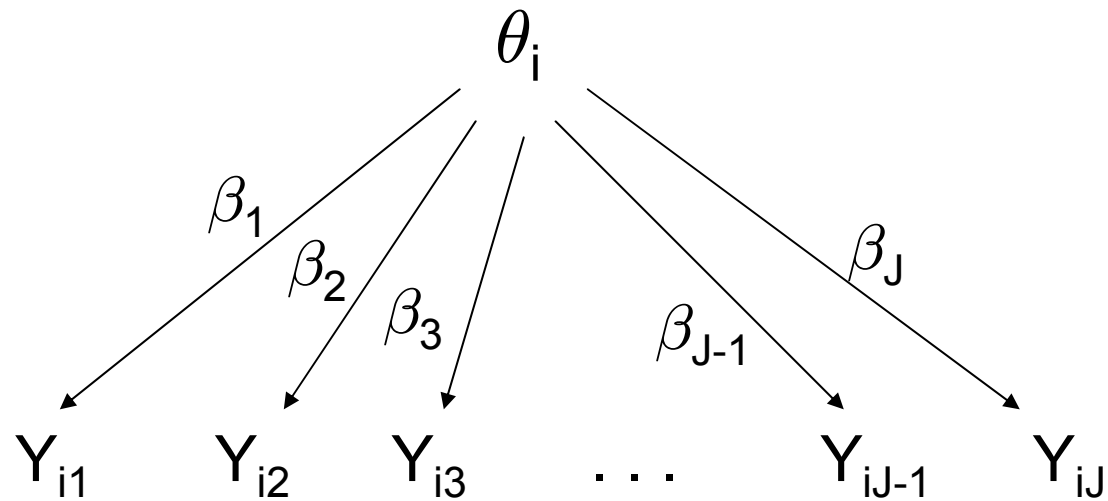
$$g(\theta_i, \beta_j) = \beta_j \theta_j, \quad \theta_i \in \{1, \dots, C\}, \quad \beta_j = (\beta_{j1}, \dots, \beta_{jC})$$

- Cognitive Diagnosis Models, e.g. “RedRUM”

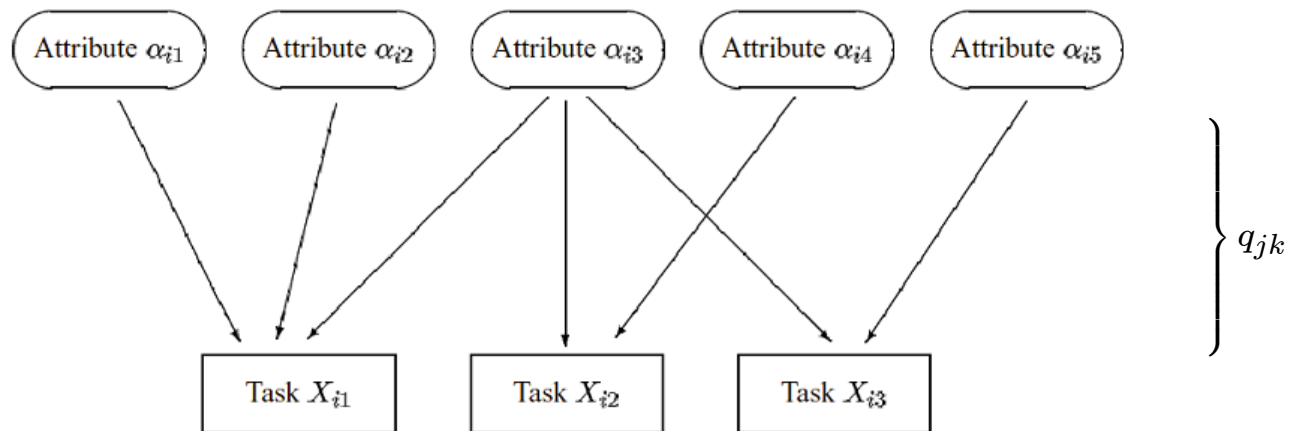
$$g(\theta_i, \beta_j) = \pi_j \prod_{k=1}^K r_{jk}^{(1-\theta_{ik})q_{jk}}, \quad \theta_i \in \{0, 1\}^K, \beta_j = (\pi_j, r_{jk} \text{'s}), \psi = Q = [q_{jk}]$$

- ...and combinations/extensions of these

# Typical Model Structure...



Or



# Likelihood of the Data

$$L(Y|\theta, \beta, \psi) = \prod_{i=1}^N \prod_{j=1}^J f(y_{ij}|\theta_i, \beta_j, \psi)$$

- As N and J grow, both sets of parameters ( $\theta$ 's,  $\beta$ 's) grow
- Leads to asymptotic inconsistency problem
  - Neyman & Scott (1948, *Econometrica*); Andersen (1970, *JASA*)
- If we control relative rates of N and J, can recover...
  - Haberman (1977, *Ann Stat*); Douglas (1997, *Pmka*)
  - The rate  $[N=O(J^q), 1.5 < q < \infty]$  is difficult to verify in practice

# Ignoring the Problem: JML

$$L(Y|\theta, \beta, \psi) = \prod_{i=1}^N \prod_{j=1}^J f(y_{ij}|\theta_i, \beta_j, \psi)$$

## ■ Joint Maximum Likelihood

- Easy to set up Newton-Raphson routines
- Can calculate  $I(\theta, \beta) = [-\partial^2 \log L / \partial \theta_i \partial \beta_j]$ , but no guarantees that it approximates inverse variance

## ■ Joint Bayesian Inference

- Put priors on all parameters
- Posterior distribution gives justifiable finite-sample characterization of uncertainty

# Making the Problem Go Away: CML

$$L(Y|s(Y), \theta, \beta, \psi) = \frac{L(Y|\theta, \beta, \psi)}{L(s(Y)|\theta, \beta, \psi)} \equiv L_{CML}(Y|\beta, \psi)$$

- $s(Y) = (s_1(Y), \dots, s_N(Y))$ ,  $s_i(Y)$  sufficient for  $\theta_i$
- Works in models that are exponential-family in  $\theta$ 
  - Standard in Rasch IRT analysis (e.g. Fischer & Molenaar, 1995, *Rasch Models* book)
  - Apply fairly standard N-R style optimization (BFGS) with some combinatorial cost for computing  $L(s(Y)|\theta, \beta, \psi)$

# Making the Problem Go Away: MML

$$\begin{aligned}L_{MML}(Y|\beta, \psi) &= \int L(Y|\theta, \beta, \psi)p(\theta|\psi)d\theta \\ &= \prod_{i=1}^N \int \prod_{j=1}^J f(y_{ij}|\theta_i, \beta_j, \psi)p(\theta_i|\psi)d\theta_i\end{aligned}$$

- $p(\theta|\psi) = \prod_{i=1}^N p(\theta_i|\psi)$  can be thought of as
  - Prior distribution
  - Population distribution for  $\theta$ , completely missing
- Holland (1990, *Pmka*) shows that CML and JML can be interpreted as approximations to MML

# MML and Expectation-Maximization

$$L_{MML}(Y|\beta, \psi) = \int L(Y|\theta, \beta, \psi)p(\theta|\psi)d\theta$$

- Can attack directly with gradient, N-R, and other methods, e.g. Bock & Lieberman (1970, *Pmka*) for IRT
- E-M (DLR, 1977, *JRSSB*; Bock & Aitken, 1981, *Pmka*) common

**E-step:**  $Q(\beta, \psi|\beta^{(m)}, \psi^{(m)}) = E[\log L(Y, \theta|\beta, \psi)|Y, \beta^{(m)}, \psi^{(m)}]$

**M-step:**  $(\beta, \psi)^{(m+1)} = \operatorname{argmax}_{\beta, \psi} Q(\beta, \psi|\beta^{(m)}, \psi^{(m)})$

- $L(Y, \theta|\beta, \psi) = L(Y|\theta, \beta, \psi)p(\theta|\psi)$
- For posterior modes, add  $\log p(\beta, \psi)$  inside  $E[\ ]$  for  $Q(\ )$ .
- In exponential family & related models, E-step often reduces to mean imputation of the missing data  $\theta$ .

# Empirical Bayes, MML and JML

- At some point we are interested in

$$p(\theta|Y) = \int p(\theta, \beta, \psi|Y) d\theta d\beta d\psi$$

$$p(\theta, \beta, \psi|Y) \propto L(Y|\theta, \beta, \psi)p(\theta, \beta, \psi)$$

- An E.B. approximation is

$$p(\theta|Y) \approx p(\theta, \hat{\beta}, \hat{\psi}|Y) \propto L(Y|\theta, \hat{\beta}, \hat{\psi})p(\theta, \hat{\beta}, \hat{\psi})$$

- Plug-in MML  $\hat{\beta}, \hat{\psi}$  is standard in empirical Bayes
- Laplace's method (Tierney et al., 1989, *JASA*) gives a better estimate by clever normal approximation
- Can also get  $p(\theta|Y)$  directly by integrating  $p(\theta, \beta, \psi|Y)$ , if we knew how to do Joint Bayesian Inference

# Markov Chain Monte Carlo (MCMC): Generic Joint Bayesian Inference

- Start with joint posterior (dropping  $\psi$  for simplicity)

$$p(\theta, \beta|Y) = \frac{L(Y|\theta, \beta)p(\theta)p(\beta)}{p(Y)} \propto L(Y|\theta, \beta)p(\theta)p(\beta)$$

- Standard Markov Chain Theory: successive samples from the “complete conditionals”

$$p(\theta|\beta, Y) \propto L(Y|\theta, \beta)p(\theta)$$

$$p(\beta|\theta, Y) \propto L(Y|\theta, \beta)p(\beta)$$

have joint posterior as their stationary distribution;  
approx any posterior functional this way.

- Analogous to exponential-family E-M:  
Impute  $\theta$ , estimate  $\beta$ , repeat...

# MCMC Basics: Sampling the Complete Conditionals

- The *Metropolis-Hastings (M-H)* step:

- To obtain a “draw” from  $p(\beta|\theta, Y) \propto L(Y|\theta, \beta)p(\beta)$  :
- Draw  $\beta^*$  from  $q(\beta^*|\beta^{(m)})$
- Accept  $\beta^{(m+1)} = \beta^*$  with probability

$$A(\beta^{(m)}, \beta^*) = \min \left\{ 1, \frac{p(\beta^{(m)}|\theta^{(m)}, Y)q(\beta^*|\beta^{(m)})}{p(\beta^*|\theta^{(m)}, Y)q(\beta^{(m)}|\beta^*)} \right\}$$

else  $\beta^{(m+1)} = \beta^{(m)}$

- Success depends on  $q(\beta^*|\beta^{(m)})$ :

- *Gibbs step*: Best if  $q(\beta^*|\beta^{(m)}) = p(\beta^*|\theta^{(m)}, Y)$
- Random walk M-H:  $q(\beta^*|\beta^{(m)}) = N(\beta^{(m)}, \sigma^2_q)$
- Independence M-H:  $q(\beta^*|\beta^{(m)})$  doesn't depend on  $\beta^{(m)}$
- “Adaptive” methods try to learn shape of  $p(\beta^*|\theta^{(m)}, Y)$  on fly

# MCMC Basics: Improving the algorithm

## ■ Algorithm Reformulations

- ❑ Blocking, e.g.:  $\beta = (\beta_1, \beta_2, \dots, \beta_J) = (\beta_{\text{block 1}}, \beta_{\text{block 2}}, \dots, \beta_{\text{block B}})$
- ❑ Intermixing kernels (Gibbs/M-H/etc.)
- ❑ Scan order

## ■ Model Reformulations

- ❑ Pushing parameters into priors
- ❑ Data augmentation (esp to create conditional independence)

## ■ Standard resources

- ❑ Tierney (1994, *Annals of Statistics*)
- ❑ Chib & Greenberg (1995, *American Statistician*)
- ❑ Gilks et al, 1996 *MCMC In Practice*)
- ❑ Andrieu, Freitas, Doucet & Jordan (2003, *Machine Learning*)

---

Aside: Tends to work well in IRT, FA, LCA, CDM, and other “local independence” models

■ E.g. for a 2PL IRT model:

$$p(a_j | rest) \propto \prod_{i=1}^N g(\theta_i; a_j, b_j)^{y_{ij}} [1 - g(\theta_i; a_j, b_j)]^{1-y_{ij}} \cdot p_a(a_j)$$

$$p(b_j | rest) \propto \prod_{i=1}^N g(\theta_i; a_j, b_j)^{y_{ij}} [1 - g(\theta_i; a_j, b_j)]^{1-y_{ij}} \cdot p_b(b_j)$$

$$p(\theta_i | rest) \propto \left\{ \prod_{j=1}^J g(\theta_i; a_j, b_j)^{y_{ij}} [1 - P(\theta_i; a_j, b_j)]^{1-y_{ij}} \right\} p_{\theta}(\theta_i)$$

**Patz & Junker (1999a,b, *JEBS*), Johnson & Albert (2001, *Ordinal Data Modeling*)**

# MCMC: Pro's

- **Sample from the joint posterior  $L(\theta, \beta, \psi | Y)$  can be used for**
  - MC Integration (posterior means, variances, marginal posteriors)
  - MC Search/Optimization
  - Model fit assessment through BF's, AIC/BIC/DIC, PPP, etc.
- **Relatively easy to get up and running for complex models**
  - Hierarchical elaborations of FA, IRT, LCA, CDM's
  - Bayes networks with multiple data types and hidden nodes
  - Lee (IMPS08, Bayesian Graphical Modeling); WinBUGS
- **Feasible for hidden process models**
  - ACT-R (Anderson & Lebiere, 1998, *Atomic Components of Thought*) can be conceived of as a kind of hidden Markov model running from “task statement” to “observable action”;
  - Weaver (to appear, *Cognitive Science*) uses MCMC to estimate sub-symbolic parameters and compare models in ACT-R setting.
  - MML like E-M experience combinatorial explosion (e.g. Seltman, 2001, *Case Studies in Bayesian Statistics*, 5)

---

# MCMC Con's

- **Easy to get a bad sampler running**; hard to “tune” sampler to get good mixing, fast convergence
  - What data augmentation variables to introduce?
  - How to rewrite model to push parameters into priors?
  - How to find/choose efficient  $q(\beta^*|\beta)$ 's?
- **Need not scale well** as parameters, model complexity grows
  - Johnson and Jenkins (2004, *ETS TR*) Bayes/MCMC implementation of full NAEP model
    - runs too slowly for operational use;
    - used as benchmark for adjusting operational MML/E-M implementation.
  - Ayers, Nugent & Dean (2008, *First Int'l Conf. on Educ. Data Mining*) find MCMC scoring of students in a CDM runs about 700 times slower than clustering methods.
- **Tries to estimate entire posterior distribution**, rather than point-estimate-plus-uncertainty
  - inefficient to use for optimization search

---

# MCMC Extensions / Alternatives

- There are many...
  - Perfect sampling eliminates burn-in
  - Adaptive MCMC dynamically updates the proposal distribution to improve “Gibbsness”
  - Auxiliary variable samplers, slice samplers, etc.
  - Combinations of MCMC and E-M extend E-M’s reach and speed up MCMC.
  - Simulated Annealing more efficient for finding maxima with MCMC samples
  - Particle filtering designed for hidden process models
  
- We will briefly examine the last three...

# MCMC Alternatives/Extensions:

## MCMC and EM

- Recall that for E-M, we must compute

$$\begin{aligned} Q(\beta, \psi | \beta^{(m)}, \psi^{(m)}) &= E[\log L(Y, \theta | \beta, \psi) | Y, \beta^{(m)}, \psi^{(m)}] \\ &= \int \log L(Y, \theta | \beta, \psi) p(\theta | Y, \beta^{(m)}, \psi^{(m)}) d\theta \end{aligned}$$

- The integral is usually calculated using (Gaussian) quadrature
- Curse of dimensionality => computational complexity of the integral grows exponentially with latent dimension
  - In many settings MCMC complexity grows “only” polynomially with latent dimension.
- So, use MCMC to get sample from  $p(\theta | Y, \beta^{(m)}, \psi^{(m)})$
- Then estimate the integral as

$$\sum_{MCMC \text{ draws}} \log L(Y, \theta^{(MCMC \text{ draw})}, \beta, \psi)$$

- Gilks et al (1996, *MCMC in Prac*); Andrieu et al (2003, *Mach. Learn.*)

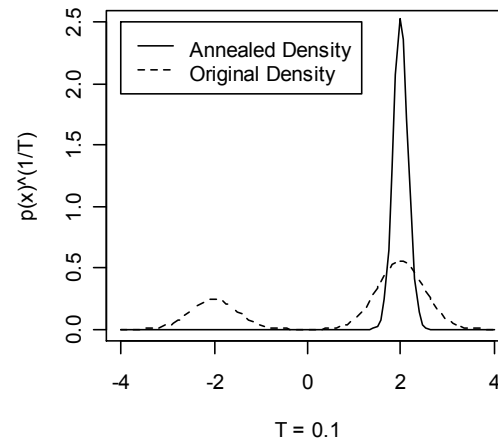
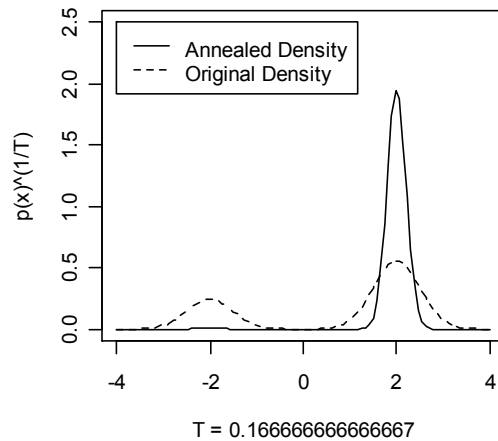
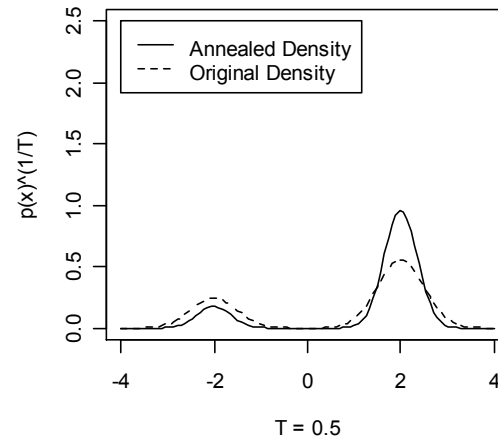
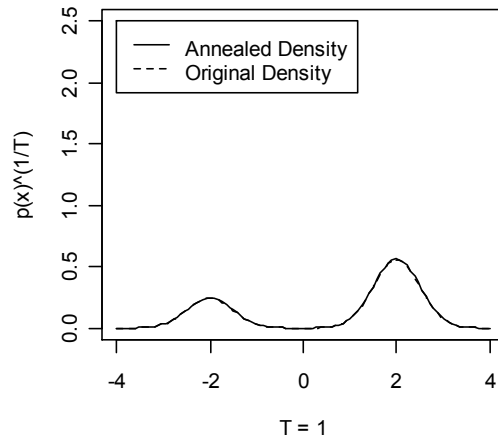
---

# MCMC Alternatives/Extensions:

## Simulated Annealing

- Goal is to maximize the function  $p(\beta)$  [=  $p(\beta|Y)$ , e.g.]
- MCMC from  $p(\beta)$  provides a sample for random search, but typically spends too much time away from the mode
- Instead build MCMC sample from  $p_T(\beta) \propto p(\beta)^{1/T}$ 
  - As  $T \rightarrow 0$ ,  $p(\beta)^{1/T}$  will “peak up” at the maximum
  - M-H is a natural method here
  - Finding appropriate proposal distributions, intermixing MCMC with  $T \rightarrow 0$  can be an art
    - RJMCMC (Green, 1995, *Biometrika*)
    - Bridge, Path Sampling (Gelman & Meng, 1998, *Stat Sci*)
- See Andrieu et al (2003, *Machine Learning*) for a survey of current methods & applications

# Simulated Annealing for Mode Finding



---

# MCMC Alternatives/Extensions: Particle Filtering

- Consider a hidden stochastic process model

$$p(\theta_0)$$

$$p(\theta_t | \theta_{0:t-1}, y_{1:t-1}) \quad , t \geq 1$$

$$p(y_t | \theta_{0:t}, y_{1:t-1}) \quad , t \geq 1$$

for latent variables  $\theta_t$  and observables  $y_t$

- Goal: obtain a sample from the posterior  $p(\theta_{0:t} | y_{1:t})$
- Classic particle filter uses an importance sampling idea to generate the sample

# MCMC Alternatives/Extensions: Particle Filtering

*Sequential importance sampling step*

- For  $m = 1, \dots, M$ , sample

$$\tilde{\theta}_t^{(m)} \sim q_t(\theta_t | \theta_{0:t-1}^{(m)}, y_{1:t}) \quad (\text{proposal distribution})$$

and set  $\tilde{\theta}_{0:t}^{(m)} = (\theta_{0:t-1}^{(m)}, \tilde{\theta}_t^{(m)})$ .

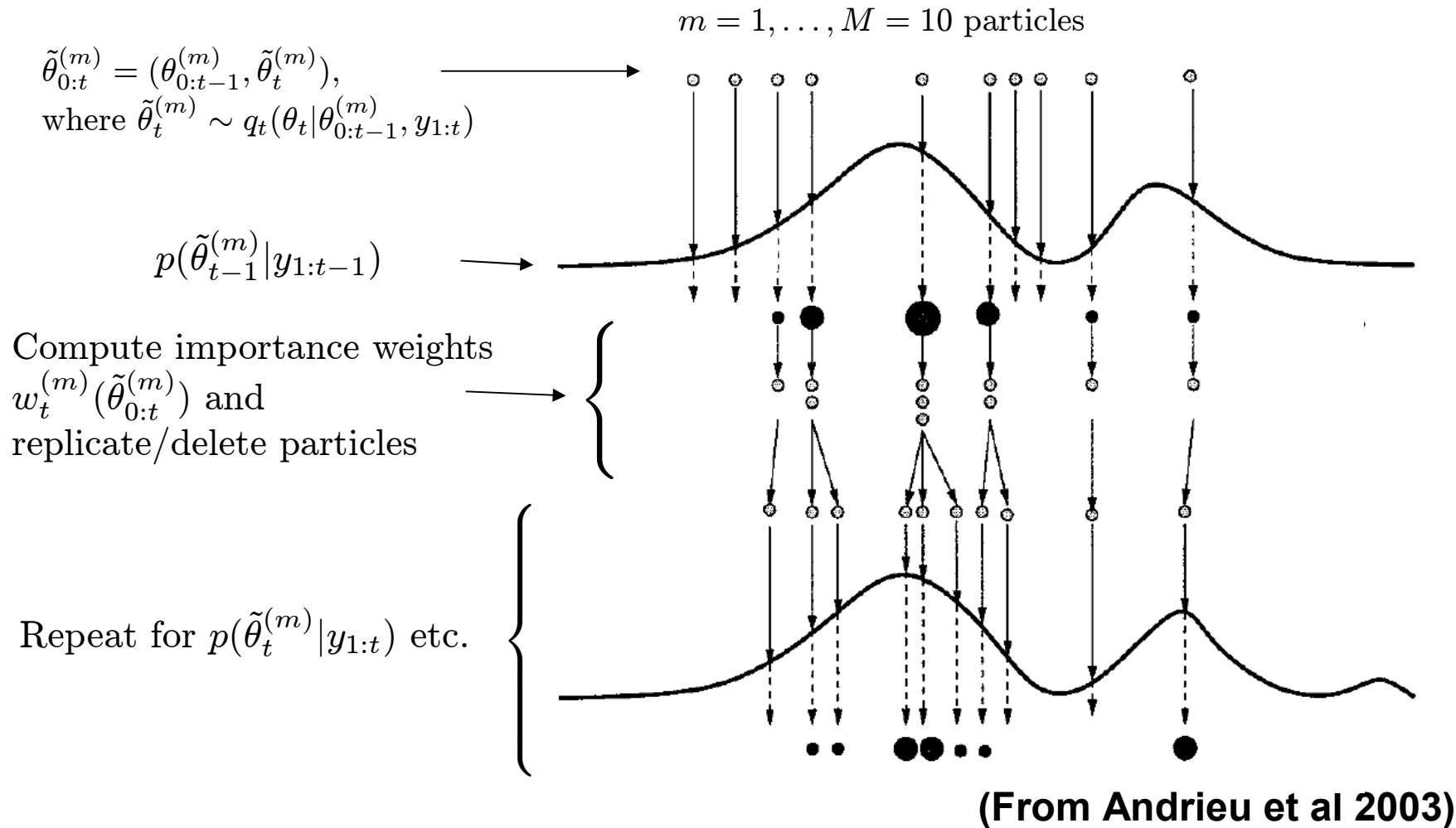
- For  $m = 1, \dots, M$ , evaluate the importance weights

$$w_t^{(m)}(\tilde{\theta}_{0:t}^{(m)}) \propto \frac{p(y_t | \tilde{\theta}_t^{(m)}) p(\tilde{\theta}_t^{(m)} | \theta_{0:t-1}^{(m)}, y_{1:t-1})}{q_t(\tilde{\theta}_t^{(m)} | \theta_{0:t-1}^{(m)}, y_{1:t})}$$

*Selection Step*

- Replicate “provisional” particles  $\tilde{\theta}_{0:t}^{(m)}$  with high  $w_t^{(m)}(\tilde{\theta}_{0:t}^{(m)})$ , discard particles with low  $w$ 's (e.g. via *sampling-importance-resampling, S-I-R*), to obtain  $M$  “final” particles  $\theta_{0:t}^{(m)}$ .

# Particle Filtering Process



---

# MCMC Alternatives/Extensions: Particle Filtering

- PF's work when the data come "sequentially"; since the update itself is sequential in time.
  - A useful MC method for Computerized Adaptive Testing?
- Should be competitive or better than MCMC for hidden process models
  - E.g. Weaver's (2008, *Cog Science*) ACT-R model.
- In practice sometimes PF's are faster, sometimes modified MCMC is faster (e.g. Marthi, 2008, *Google Tech Talks* on youtube.com)
- Again, Andrieu et. al. (2003, *Mach Learn*) good start

---

# Some Connections with Other Methods

- **M-H MCMC and PF's are related to Genetic Algorithms**  
(Back, 1996, *Evolutionary Algorithms in Theory & Practice*; Davis & Principe, 1993, *J Evol Computation*)
  - Innovations are randomly generated, evaluated for “fitness” to survive (acceptance ratio or importance weights)
  - Innovations are generated with the optimization criterion in mind
- **Coordinatewise MCMC is a kind of stochastic Gauss-Seidel “stairstepping” algorithm**
  - G-S originated as a way to solve linear systems but is a metaphor for all kinds of one-dimension-at-a-time search and optimization algorithms today, e.g. variants of gradient, N-R, ALS, etc.
  - Many methods, such as E-M, can be interpreted as generalized Gauss-Seidel algorithms as well
- **Other methods deserve another look**, because MCMC is slow relative to N-R, E-M, etc.

# E-M as a Variational Algorithm

- Standard E-M is efficient to state, yet can be difficult to implement since

$$Q(\beta, \psi | \beta^{(m)}, \psi^{(m)}) = E[\log L(Y, \theta | \beta, \psi) | Y, \beta^{(m)}, \psi^{(m)}]$$

may be difficult to calculate (numerical integration) or optimize (differentiation, non closed-form station. pt).

- Approximating Q from a more convenient family of objective functions can be useful.
- To see what to do, we revisit the proof that E-M “works”.
- The arguments below follow Minka (1998, *Microsoft Research*) and Beal & Ghahramani (2003, *Bayesian Statistics 7*), and originated with Neal & Hinton (1993, *Univ of Toronto CS TR*)

# E-M Lower-Bounding

- We wish to maximize

$$L(Y|\beta) = \int L(Y, \theta|\beta) d\theta$$

with respect to  $\beta$  (dropping  $\psi$  again for simplicity).

- Letting  $q(\theta, \beta)$  be any function, we can write

$$\begin{aligned} \log L(Y|\beta) &= \log \int L(Y, \theta|\beta) d\theta = \log \int \frac{L(Y, \theta|\beta)}{q(\theta, \beta^*)} q(\theta, \beta^*) d\theta \\ &\geq \int \log \frac{L(Y, \theta|\beta)}{q(\theta, \beta^*)} q(\theta, \beta^*) d\theta \equiv Q(\beta|\beta^*) \end{aligned}$$

**for all  $\beta$** , as long as  $q(\theta, \beta^*)$  is a density in  $\theta$ , by Jensen's Inequality.

# E-M Lower-Bounding

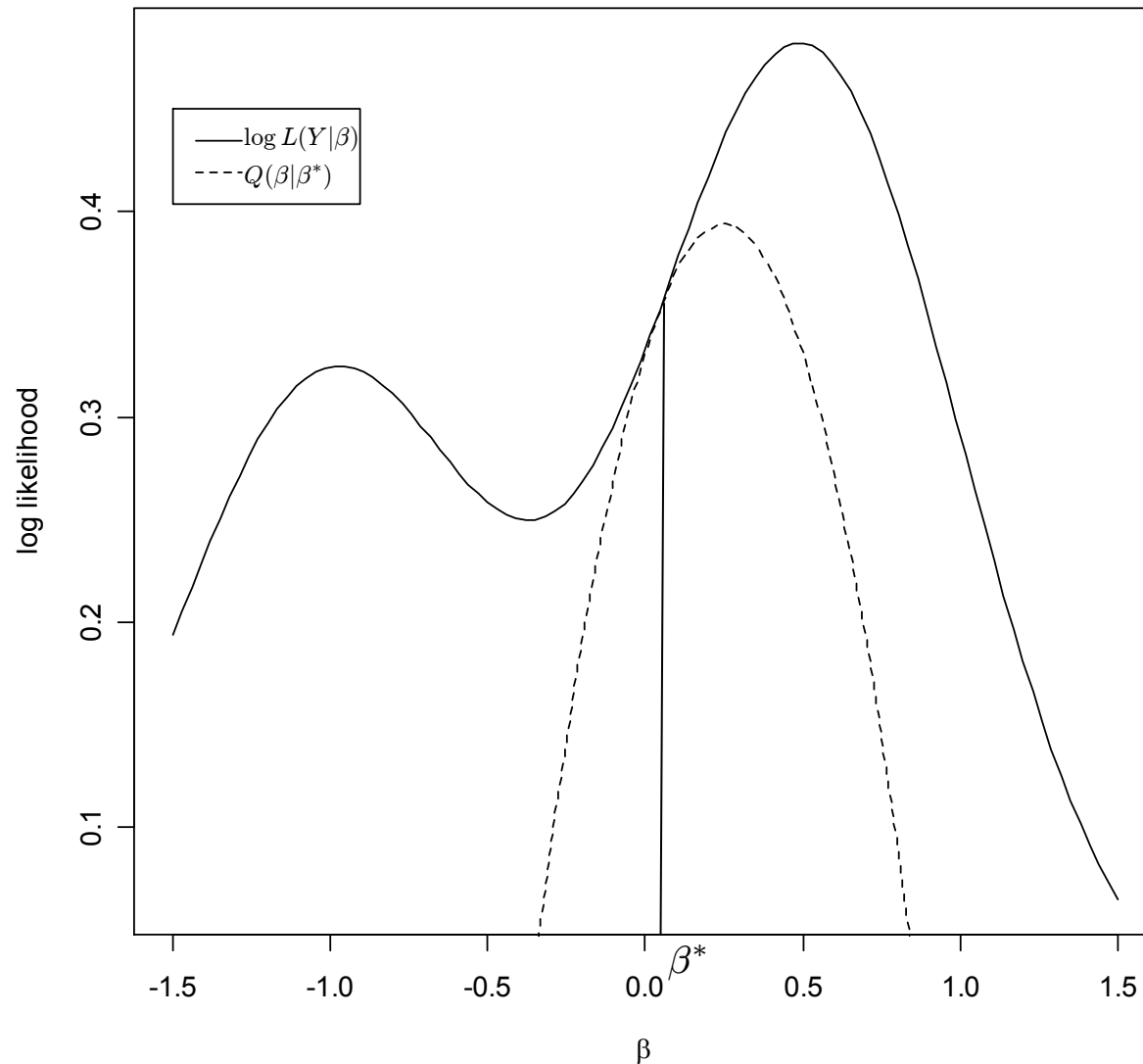
$$\log L(Y|\beta) \geq \int \log \frac{L(Y, \theta|\beta)}{q(\theta, \beta^*)} q(\theta, \beta^*) d\theta \equiv Q(\beta|\beta^*), \quad \forall \beta$$

- There should be a whole family of lower-bound functions  $Q(\beta|\beta^*)$
- As long as we can choose  $Q(\beta|\beta^*)$  so that

$$\log L(Y|\beta^*) = Q(\beta^*|\beta^*)$$

then maximizing  $Q(\beta|\beta^*)$  in  $\beta$  must increase  $\log L(Y|\beta)$  as well.

# Lower-Bounding: $\log L(Y | \beta) \geq Q(\beta | \beta^*)$



## Aside: Quadratic Lower-Bound Principle

(Lange, 1999, *Numerical Analysis for Statisticians*)

Let  $\ell(\beta) = \log L(Y|\beta)$ ; Taylor's theorem gives

$$\ell(\beta) = \ell(\beta^{(m)}) + (\beta - \beta^{(m)})^T \nabla \ell(\beta^{(m)}) + \frac{1}{2} (\beta - \beta^{(m)})^T H(\beta^*) (\beta - \beta^{(m)})$$

where  $H(\beta)$  is the Hessian (matrix of second partial derivatives). If we let

$$Q(\theta|\beta^{(m)}) = \ell(\beta^{(m)}) + (\beta - \beta^{(m)})^T \nabla \ell(\beta^{(m)}) + \frac{1}{2} (\beta - \beta^{(m)})^T B (\beta - \beta^{(m)})$$

where  $B$  and  $H(\beta) - B$  are both positive-definite for all  $\theta$ , then

$$\log L(Y|\beta) - Q(\beta|\beta^{(m)}) \geq 0 = \log L(Y|\beta^{(m)}) - Q(\beta^{(m)}|\beta^{(m)})$$

Since  $Q(\beta|\beta^{(m)})$  is quadratic, we can maximize it in a single Newton step,

$$\beta^{(m+1)} = \beta^{(m)} - B^{-1} H(\beta^{(m)})$$

This is a kind of quasi-Newton/gradient method that trades speed of completing one iteration for more iterations (Böning & Lindsay, 1988).

# Optimal E-M via Variational Calculus

$$\log L(Y|\beta) \geq \int \log \frac{L(Y, \theta|\beta)}{q(\theta, \beta^*)} q(\theta, \beta^*) d\theta \equiv Q(\beta|\beta^*), \quad \forall \beta$$

- Want to choose  $q(\theta, \beta^*)$  to maximize  $Q(\beta|\beta^*) = G(\theta, q)$
- Can solve as a Lagrange Multiplier problem using the Gateaux derivative

$$H(t, \lambda, v) = G(\theta, q(\theta, \beta^*) + tv(\theta)) + \lambda \left( 1 - \int [q(\theta, \beta^*) + tv(\theta)] d\theta \right)$$

- If  $q$  is the max, then  $\left. \frac{\partial H}{\partial t} \right|_{t=0} = 0$  for all  $v(\theta)$ ;
- It follows that  $q(\theta, \beta^*) = p(\theta|Y, \beta^*)$ , which gives the std E-M algorithm Q function (up to a constant).

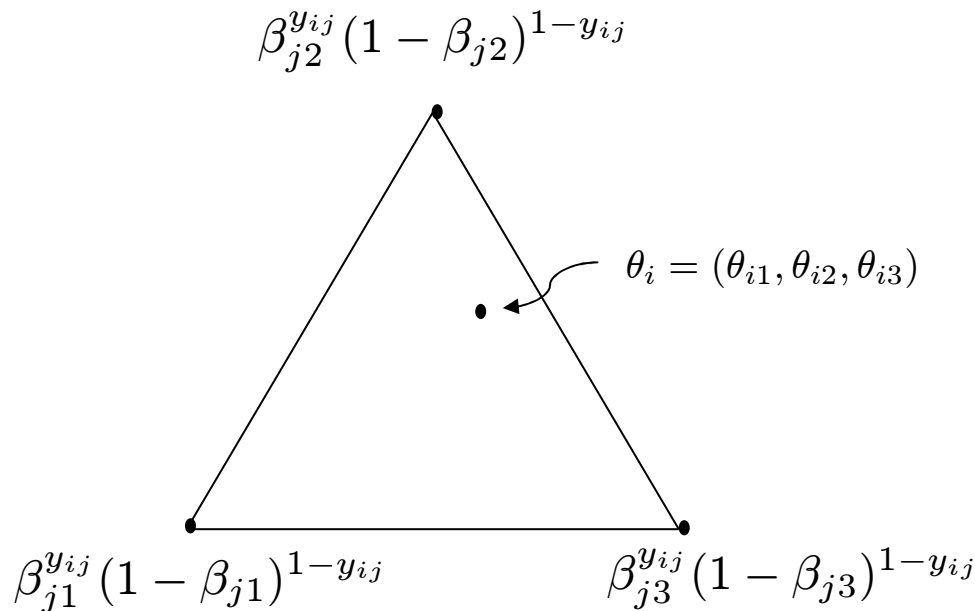
# Variational E-M Comments

- The variational approach recasts E-M as a kind of Gauss-Seidel algorithm:
  - **E step:** Find  $Q(\beta, \beta^{(m)})$  by maximizing over  $q(\theta, \beta^{(m)})$
  - **M step:** Find  $\beta^{(m+1)}$  by maximizing  $Q$  over  $\beta$ .
- The E-step amounts to finding  $q(\theta, \beta^{(m)})$  minimizing

$$D_{KL}(q(\theta, \beta^{(m)}) || P(\theta | Y, \beta^{(m)}))$$

- If we maximize over unconstrained  $q(\theta, \beta^{(m)})$ , we get standard E-M back
- If we maximize over a more convenient set of  $q(\theta, \beta^{(m)})$ 's, we get a new “EM-like” algorithm that may be easier.

# Example: Grade of Membership (GoM) Model



- $\theta_i \sim \text{Multi}(\mathbf{1}, \psi_1, \psi_2, \psi_3)$   
→ **LCA model**
- $\theta_i \sim \text{Dir}(\psi_1, \psi_2, \psi_3)$   
→ **GoM model**

- Woodbury et al, 1978, *Comp & Biomed Research*, 11
- Erosheva et al, 2007, *Ann Applied Statistics*

$$P[Y_{ij} = 1 | \theta_i, \beta_j] = \sum_{k=1}^K [\beta_{jk}^{y_{ij}} (1 - \beta_{jk})^{1-y_{ij}}] \theta_{ik}$$

# Example: GoM (cont'd)

- Basic formulation

$$\theta_i \sim \text{Dir}(\theta_i | \psi) \propto \prod_{k=1}^K \theta_{ik}^{\psi_k}$$

$$P[Y_{ij} = 1 | \theta_i, \beta_j] = \sum_{k=1}^K \beta_{jk}^{y_{ij}} (1 - \beta_{jk})^{1-y_{ij}} \theta_{ik}$$

- Reformulate as per-item LCA model (Erosheva et al, 2007, *Ann Appl Stat*; Blei et al, 2003, *J Machine Learning Research*) or “Latent Dirichlet Allocation (LDA)” model

$$\theta_i \sim \text{Dir}(\theta_i | \psi) \propto \prod_{k=1}^K \theta_{ik}^{\psi_k}$$

$$(z_{ij1}, \dots, z_{ijK}) \sim \text{Multi}(1, \theta_{i1}, \dots, \theta_{iK}) \propto \prod_{k=1}^K \theta_{ik}^{z_{ijk}}$$

$$P[Y_{ij} = 1 | \theta_i, \beta_j, z_{ij\cdot}] = \prod_{k=1}^K \left[ \beta_{jk}^{y_{ij}} (1 - \beta_{jk})^{1-y_{ij}} \right]^{z_{ijk}}$$

# Example: GoM (cont'd)

- Marginal model is then

$$P(Y|\beta, \psi) = \prod_{i=1}^N \int \sum_{k=1}^K \left[ \prod_{j=1}^J p(y_{ij} | z_{ijk} = 1, \beta_j) p(z_{ijk} = 1 | \theta_i) \right] p(\theta_i | \psi) d\theta$$

where  $p(y_{ij} | z_{ijk} = 1, \beta_j) = \beta_{jk}^{y_{ij}} (1 - \beta_{jk})^{1-y_{ij}}$

$$p(z_{ijk} = 1 | \theta_i) = \theta_{ik}$$

$$p(\theta_i | \psi) \propto \prod_k \theta_{ik}^{\psi_k}$$

- this is clearly a model for which we might try
  - MCMC (Erosheva et al, 2007; Griffiths, IMPS08)
  - E-M (but the integral is a difficult multivariate one)
  - Variational methods (Blei et al, 2003; Erosheva et al, 2007)

# Example: GoM (cont'd)

- Following the E-M/variational approach, wish to find posterior mode by maximizing

$$\begin{aligned}\log p(\beta, \psi|Y) &= \log \int \left[ \frac{p(\beta, \psi, \theta, z|Y)}{q(\theta, z, \beta^*, \psi^*)} \right] q(\theta, z, \beta^*, \psi^*) d\theta dz \\ &\geq \int \log \left[ \frac{p(\beta, \psi, \theta, z|Y)}{q(\theta, z, \beta^*, \psi^*)} \right] q(\theta, z, \beta^*, \psi^*) d\theta dz\end{aligned}$$

- Finding maximal  $q()$  equivalent to minimizing

$$D_{KL}( q(\theta, z, \beta^*, \psi^*) || p(\theta, z|Y, \beta^*, \psi^*) )$$

- Unconstrained  $\rightarrow q(\theta, z, \beta^*, \psi^*) = p(\theta, z|Y, \beta^*, \psi^*)$  and standard E-M

## Example: GoM (cont'd)

- $p(\theta, z|Y, \beta^*, \psi^*) = p(\theta, z, Y|\beta^*, \psi^*) / p(Y|\beta^*, \psi^*)$  and the denom. couples  $\theta$  and  $z$  in complex ways
- Try to approximate as

$$\begin{aligned} p(\theta_i, z_i|Y_i, \beta^*, \psi^*) &\approx q(\theta_i, z_i, \beta^*, \psi^*) \\ &= \underbrace{q(\theta_i|\gamma)}_{Dirichlet} \prod_{j=1}^J \underbrace{q(z_{ij}|\phi_i)}_{Multinomial} \end{aligned}$$

- where  $\gamma, \phi_i$  are free params to minimize  $D_{KL}()$ .

---

# Example: GoM (cont'd)

- This leads to
  - E-step: replace numerical integration with closed-form estimates for  $\gamma, \phi_i$
  - M-step: closed-form estimates for  $\beta$ 's; N-R for  $\phi$
- Comparison with E-M
  - Each E-step less efficient (more E-M steps needed)
  - Each E-step less complex (faster, more numerically stable)
- Comparison with MCMC
  - Faster because only want point estimates
  - Approaches simplicity of computing complete conditionals
  - Selection of  $q(\cdot)$  class is an art, like design of complete conditionals and algorithm refinements in MCMC

---

# Other Examples

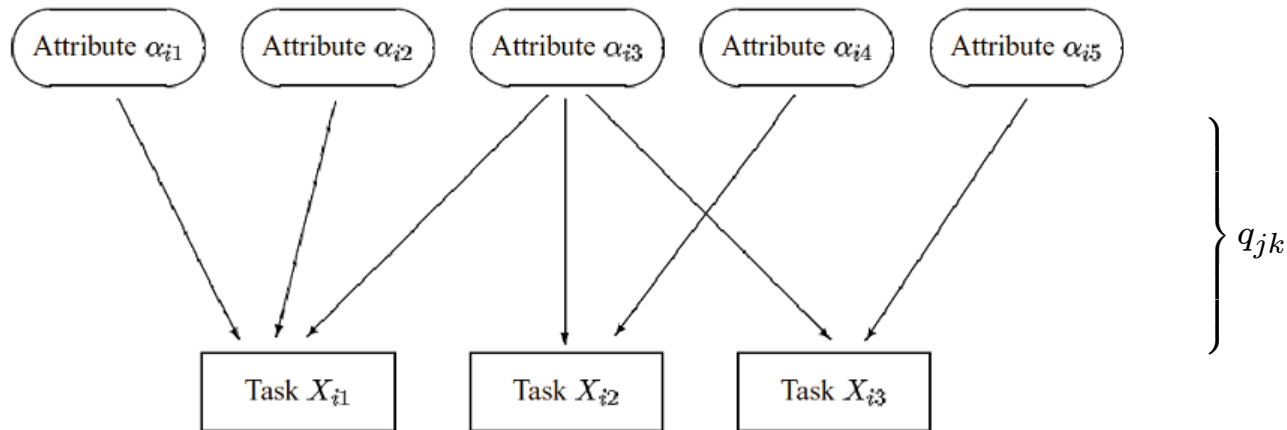
- Jaakola (2000, *Neural Information Processing Systems Conference*) discusses applications to graphical models
- de Freitas, Hojen-Sorensen, Jordan & Russell (2001, *Uncertainty in AI*, 26) apply variational methods to approximate proposal distributions and complete conditionals in MCMC.
- Beal & Ghahramani (2003, *Bayesian Statistics 7*) use the  $q(\theta, z) = q_{\theta}(\theta)q_z(z)$  trick to develop a generic E-M algorithm for lower-bounding or maximizing the log-marginal distribution in Bayesian problems with complex structure and incomplete data

# Computational Bottleneck

- Full model analysis: Flexibility & insight, vs. speed
  - N-R has quadratic convergence near mode
  - E-M / Variational has linear convergence
  - MCMC typically slower than E-M / Variational!

MCMC is trying to do more: Whole posterior!
- Johnson & Jenkins (2001, *NCME*): fully Bayesian, MCMC-based inference for US NAEP
  - Good benchmark but too slow for operational use!
- Anozie (2007, *NCME*): MCMC estimation of DINA in Computer Based Tutoring
  - Approximate limits of roughly 300 tasks, 100 skills, 600 students (1-3 skills/task, 20-40 tasks/student)

# Example: Scoring Students in Cognitive Diagnosis Models



- Henson, Templin & Douglas (2007, *J Ed Meas*): Given  $Q=[q_{jk}]$ , compute

$$W_i = (W_{i1}, \dots, W_{iK}) \text{ where } W_{ik} = \sum_{j:i \text{ saw } j} x_{ij} q_{jk}$$

& cluster these to find scorable skill patterns

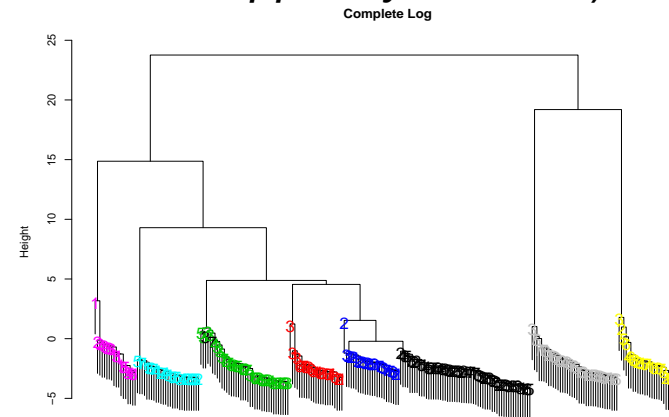
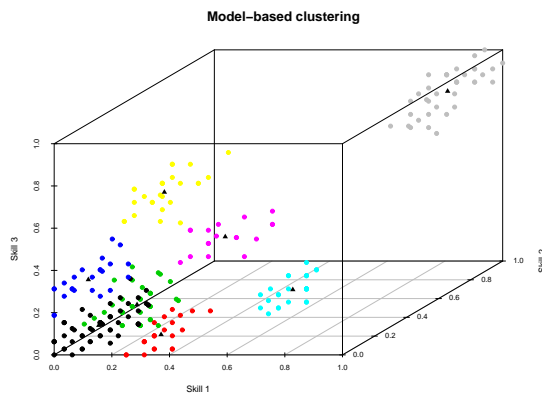
# Ex: Scoring Students in CDMs (cont'd)

- Ayers, Nugent & Dean (2008, *First Int'l Conf on Educ. Data Mining*): Clustering and visualization to aid scoring, using

$$D_i = (D_{i1}, \dots, D_{iK})$$

where  $D_{ik} = W_{ik} / J_{ik}$  where  $J_{ik}$  is the number of tasks involving skill  $k$  seen by student  $i$

- Clustering is up to 700 times faster than MCMC fit of a Cognitive Diagnosis model
- In a submodel of the RedRUM model called the NIDA model, we can verify that  $W_{ik}$  are (partially) “sufficient statistics” for each skill indicator  $\alpha_{ik}$
- The NIDA model has an explicit credit/blame mechanism that may be helpful here (see Junker & Sijtsma, 2001, *Appl Psych Meas*)



# Discussion

- MCMC is here to stay as one tool in the computational toolkit (Griffiths, IMPS08; Lee, IMPS08; etc. etc.)
- It is good to place it in context to see
  - What MCMC replaced/extended, e.g.
    - Gauss-Seidel, N-R, E-M, Quadrature & MC Integration, Search
  - How MCMC is being replaced/extended, e.g.
    - MC-EM, Simulated Annealing, Particle Filtering
- A variational interpretation of E-M leads to E-M like algorithms that are faster than MCMC and nearly as simple to implement
- Fully model-based approaches run into computational bottlenecks as data and model complexity grows...
  - Computational data manipulation motivated by (sufficient statistics of) relevant statistical model may offer a way out