Latent-Variable Approaches for Accurate Computation in Bayesian Scale-Usage Models

THE OHIO STATE UNIVERSITY

Christopher Hans

Department of Statistics

Joint work with G. Allenby, P. Craigmile, J. Lee, S. MacEachern and X. Xu

Latent Variables 2016
University of South Carolina

October 14, 2016

Customer Satisfaction Surveys

Motivating example: Rossi, Gilula, Allenby (JASA 2001)

On a scale from 1 to 10 where 10, means an "Excellent" performance and 1 means a "Poor" performance, please rate BRAND on the following items:

Q1. Overall value

Price:

Q2. Setting competitive prices.

Q3. Holding price increases to a reasonable minimum for the same ad as last year.

Q4. Being appropriately priced for the amount of customers attracted to your business.

Effectiveness:

Q5. Demonstrating to you the potential effectiveness of your advertising purchase.

Q6. Attracting customers to your business through your advertising.

Q7. Reaching a large number of customers.

Q8. Providing long-term exposure to customers throughout the year.

Q9. Providing distribution to the number of households and/or business your business needs to reach.

Q10. Proving distribution to the geographic areas your business needs to reach.

- Interested in which aspects are associated with Overall value

- Requires individuals to interpret and use a response scale (1, …, 10)

Data Setting and Modeling Goals

- Have responses of N individuals on M items

$$X_{ij} \quad i = 1, \dots, N \quad j = 1, \dots, M$$

- Ordinal response scale

$$X_{ij} \in \{1, \dots, K\}$$

- Interested in:
  - Covariance structure of response vector across items
  - Modeling individual response heterogeneity

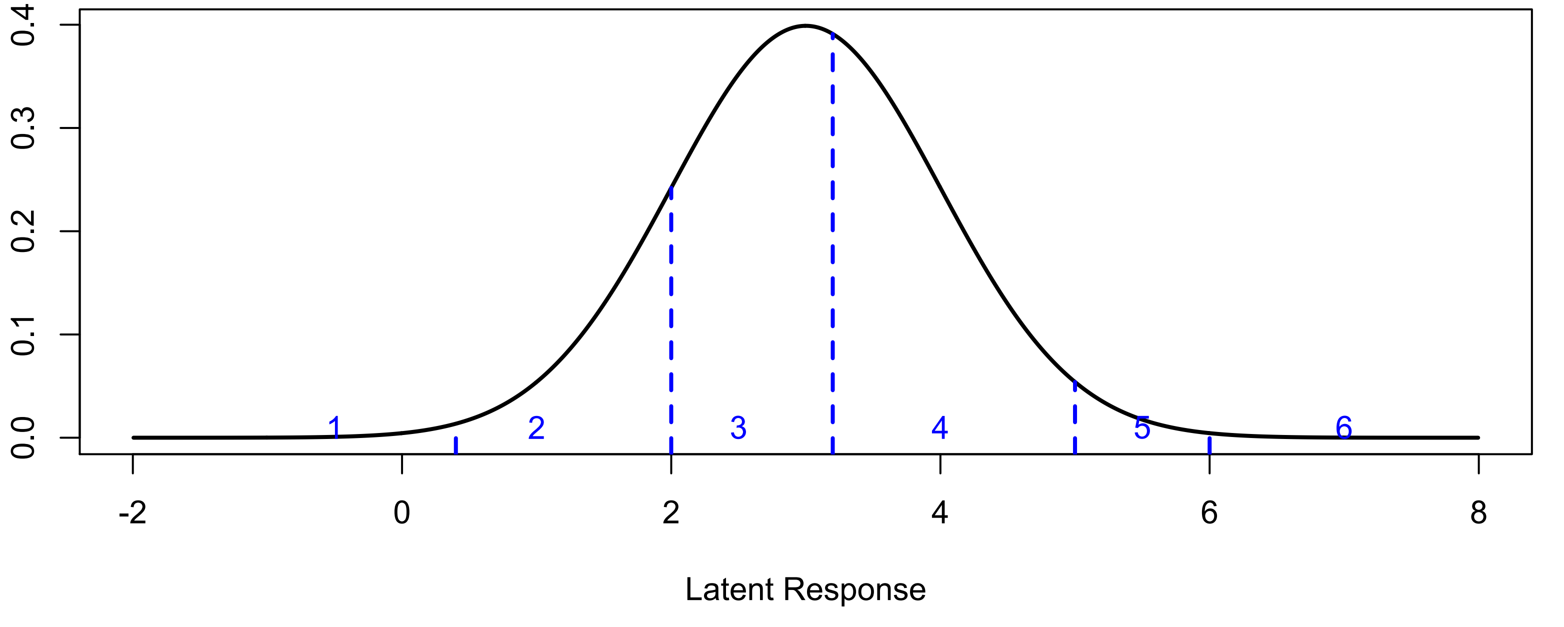
- Develop new computational methods (MCMC)

Multivariate Ordinal Probit Model

The discrete response is viewed as a censored outcome of a continuous latent variable.

$$Y_i \mid \mu, \Sigma \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \Sigma)$$

$$X_{ij} \mid Y_{ij}, c = \{k \;:\; c_{k-1} \leq Y_{ij} < c_k\}$$

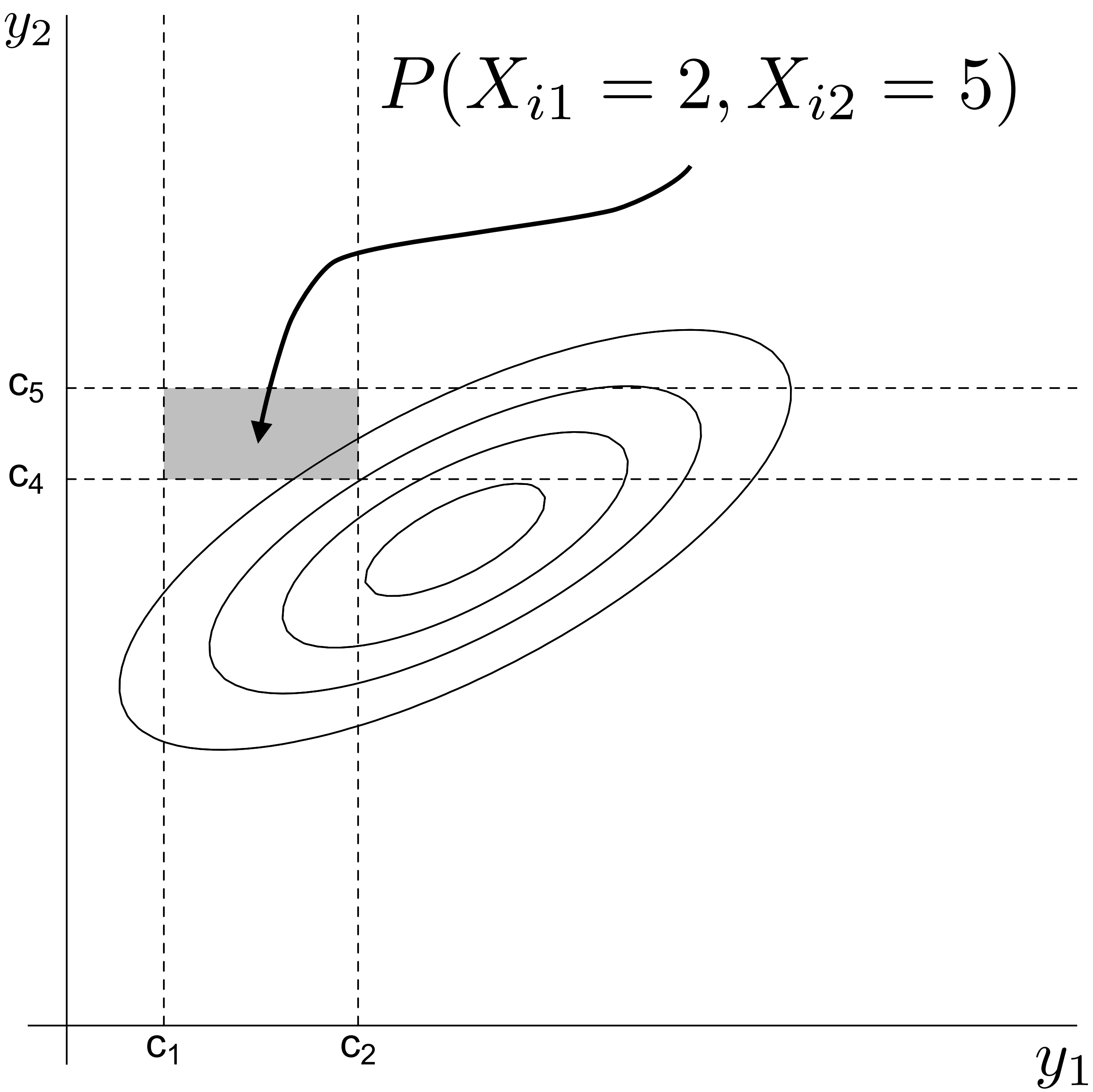


$$-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = \infty$$

Aitchison & Silvey (1957)
McCullagh (1983)
Albert & Chib (1993)
Chen & Dey (2000)

Response Probabilities

Example: M = 2 questions with K = 10 possible responses



Computational Difficulties

$$\int_{R_i(c)} \mathrm{N}(Y_i \mid \mu, \Sigma) dY_i$$

- Dimension of integrals (M)

- Tails of distribution

Patterns of Scale Usage

Different respondents use the response scale differently:

- Location heterogeneity

- Scale heterogeneity

If not appropriately accounted for, scale-usage heterogeneity can bias estimates of correlation between response items.

Explicitly model scale-usage heterogeneity with individual-specific location and scale parameters:

$$Y_i \mid \mu, \Sigma, \tau_i, \sigma_i^2 \overset{\text{ind}}{\sim} \mathrm{N}(\mu + \tau_i \mathbf{1} \ , \ \sigma_i^2 \Sigma)$$

Rossi, Gilula, Allenby (JASA, 2001)

Others: Javaras and Ripley (JASA, 2007)

Build a Bayesian model and base inference on posterior $\pi(\Sigma, \mu \mid X)$

Model Components

Parameters:	μ	Σ	τ_i	σ_i^2	Y_{ij}	c
	10	45	1811	1811	18110	9

Impose constraints on the cutpoints: for $C > 0$,

$$-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = \infty$$

$$c_0 = -C \qquad c_K = C$$

Place prior on *increments* between cutpoints.

This fixes a baseline location and scale for the latent variables.

Prior distributions:

$$E(\sigma_i^2) = 1 \qquad E(\Sigma) = I$$

$$E(\tau_i) = 0 \qquad E(\mu) = 0$$

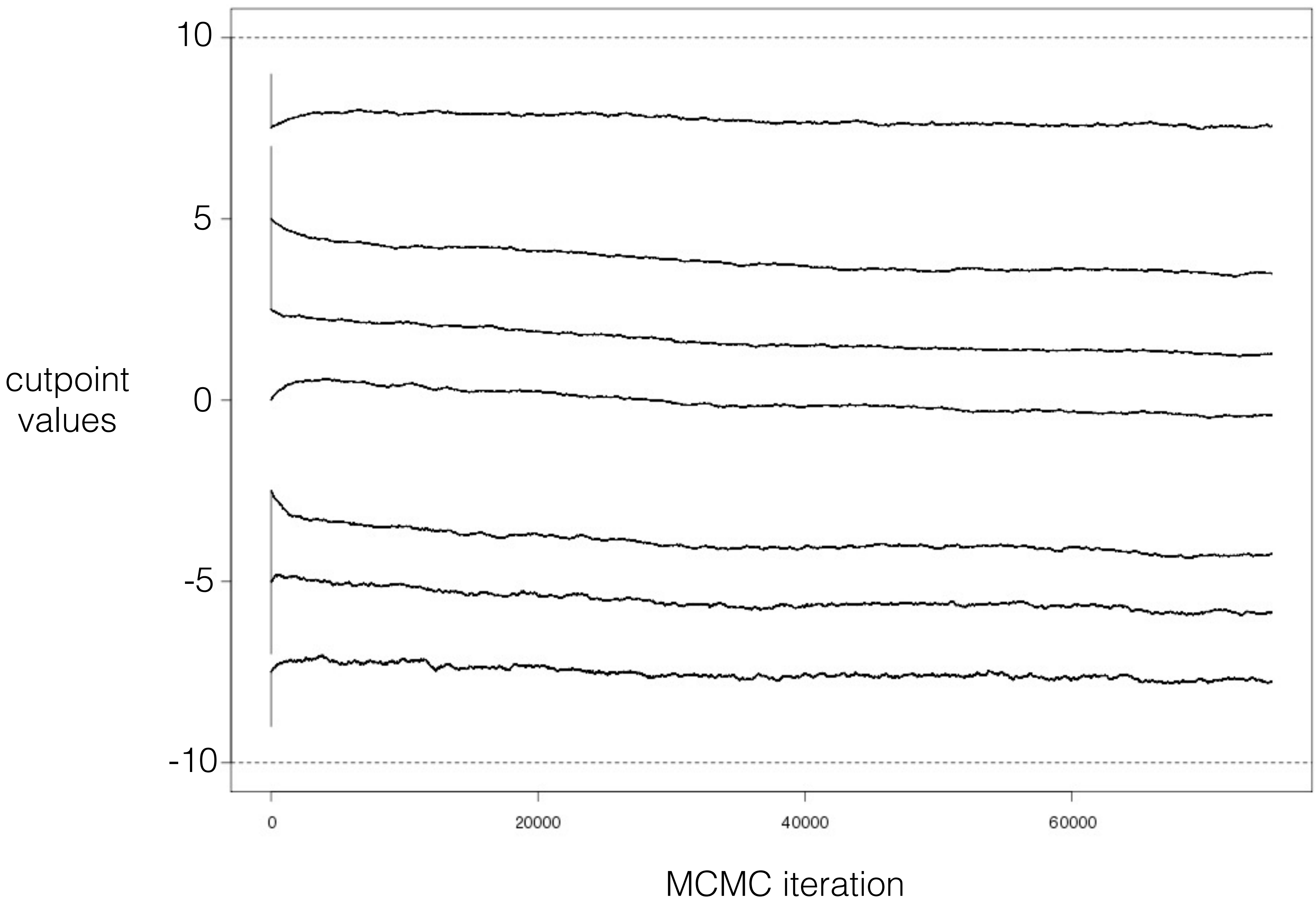
Computational Challenges

- Use MCMC to obtain samples from the posterior
- First consider “naive” Gibbs sampler
  - Primary difficulty lies in sampling the cutpoints
- Next consider improvements motivated by univariate models
  - Cowles (1996)
  - Two problems arise when generalizing to multivariate data
- Reformulate the data augmentation to obtain a sampler that
  - Mixes well
  - Is easy to implement (computational speedy)

Naive Gibbs Sampler

Components	Notes
Step 1: $\theta_k \mid \theta_{(k)}, Y, c, X$	$\theta = (\mu, \Sigma, \tau, \sigma^2)$ "easy"
Step 2: $c_k \mid c_{(k)}, Y, \theta, X$	$X_{ij} \mid Y_{ij}, c = \{k \ : \ c_{k-1} \leq Y_{ij} < c_k\}$ $\max_{(i,j) \ : \ X_{ij}=k} Y_{ij} < c_k < \min_{(i,j) \ : \ X_{ij}=k+1} Y_{ij}$
Step 3: $Y_{ij} \mid Y_{i(j)}, c, \theta, X$	univariate truncated normals

Naive Gibbs Sampler



“Improved” Gibbs Sampler

Components	Notes
Step 1: $\theta_k \mid \theta_{(k)}, Y, c, X$	same as before
Step 2: $c \mid \theta, X$	Requires evaluation of the observed data likelihood: $$\prod_{i=1}^{N} \int_{R_i(c)} \mathrm{N}(Y_i \mid \mu + \tau_i \mathbf{1}\ ,\ \sigma_i^2 \Sigma) dY_i$$
Step 3: $Y_i \mid c, \theta, X$	Requires sampling **directly** from multivariate truncated normal distributions

Problems

Must estimate the integrals $$\prod_{i=1}^{N} \int_{R_i(c)} \mathrm{N}(Y_i \mid \mu + \tau_i \mathbf{1} \ , \ \sigma_i^2 \Sigma) dY_i$$

- At each iteration in the sampler
- Under both the current and proposed cutpoints

Monte Carlo integration is commonly used, however accurate estimates may require many Monte Carlo samples.

Taking too few samples results in an incorrect MCMC update!

Gibbs sampler gives incorrect output.

New Data Augmentation

Decompose: $\Sigma = R + D$

D diagonal, positive definite

R nonnegative definite

New Data Augmentation:

$$Z_i \sim \mathrm{N}(0, \sigma_i^2 R)$$
$$Y_i \mid Z_i \sim \mathrm{N}(\mu + \tau_i \mathbf{1} + Z_i \ , \ \sigma_i^2 D)$$

Keys to the Approach:

- Distribution of the observed data remains the same
- Given Z_i, the Y_{ij} are independent

Covariance Decomposition

Other uses of “orthogonalizing latent variables” in related model settings:

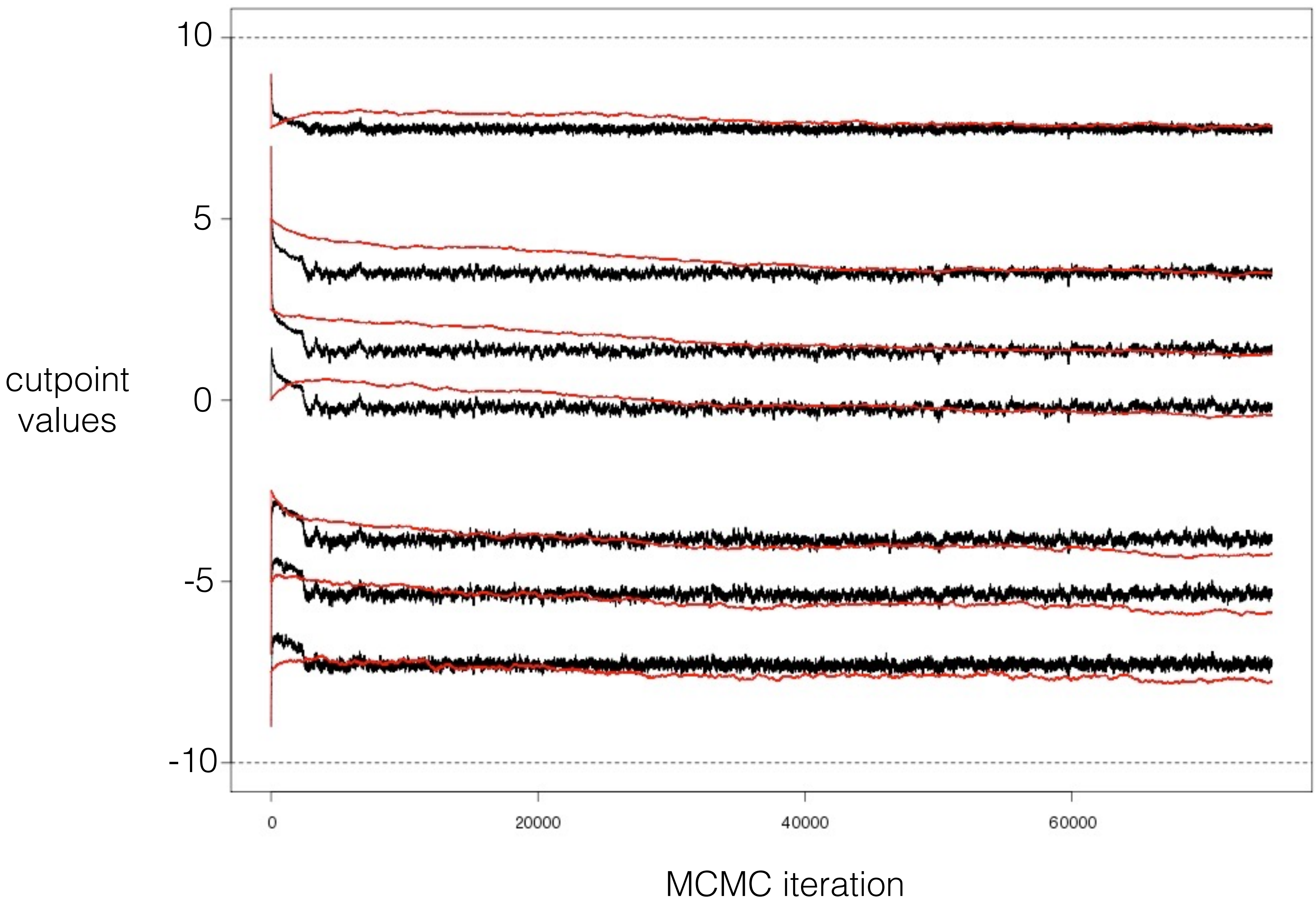
- Stern (1992): Used to construct Monte Carlo estimates of multivariate normal probabilities

- Chib & Jeliazkov (2006): Used in models for dynamic binary longitudinal data to simplify matrix inversion calculations.

New Gibbs Sampler

Components	Notes
Step 1: $\theta_k \mid \theta_{(k)}, Y, c, X$	same as before
Step 2: $Z_i \mid \theta, Y, c, X$	unrestricted normal distributions
Step 3: $c \mid Z, \theta, X$	Need only evaluate: $\prod_{i=1}^{N} \prod_{j=1}^{M} \left[\Phi\left(\frac{c_{x_{ij}} - \mu_j - \tau_i - Z_{ij}}{\sigma_i \sqrt{D_{jj}}} \right) - \Phi\left(\frac{c_{x_{ij}-1} - \mu_j - \tau_i - Z_{ij}}{\sigma_i \sqrt{D_{jj}}} \right) \right]$
Step 4: $Y_{ij} \mid c, Z_{ij}, \theta, X_{ij}$	univariate truncated normals

Markov chain Comparison



Decomposing Σ

Different decompositions yield different Markov chains.

Want to maximize rate of convergence.

Hard! Pass to a simpler problem, use theory as a guide:

$$Y_i^{(t)} \mid Y_i^{(t-1)} \sim$$

$$N\left((R^{-1}+D^{-1})^{-1}D^{-1}(Y_i^{(t-1)}-\mu-\tau_i\mathbf{1})+\mu+\tau_i\mathbf{1}\ ,\ \sigma_i^2 D+\sigma_i^2(R^{-1}+D^{-1})^{-1}\right)$$

Vector autoregressive model (Roberts and Sahu, JRSS-B, 1997):

minimize the largest eigenvalue of $(R^{-1}+D^{-1})^{-1}D^{-1}$

Optimal Decomposition

Property: under optimal decomposition, *R* is singular.

Can derive optimal decomposition when Σ has special structure.

(independence, exchangeable correlation structure, circular correlation structure, reversible correlation structure, block diagonal structure)

Default decomposition: (results in singular *R*)

(diagonal variance matrix) $V = \mathrm{diag}(\Sigma)$

(correlation matrix) $C = V^{-1/2}\Sigma V^{-1/2}$

$D = \lambda_M(C)V$

(smallest eigenvalue of C)

Comparison

Have said that poor estimates of the integrals

$$\prod_{i=1}^{N} \int_{R_i(c)} \mathrm{N}(Y_i \mid \mu + \tau_i \mathbf{1} \ , \ \sigma_i^2 \Sigma) dY_i$$

can change the limiting distribution of the Markov chain.

Here we compare two MCMC algorithms:

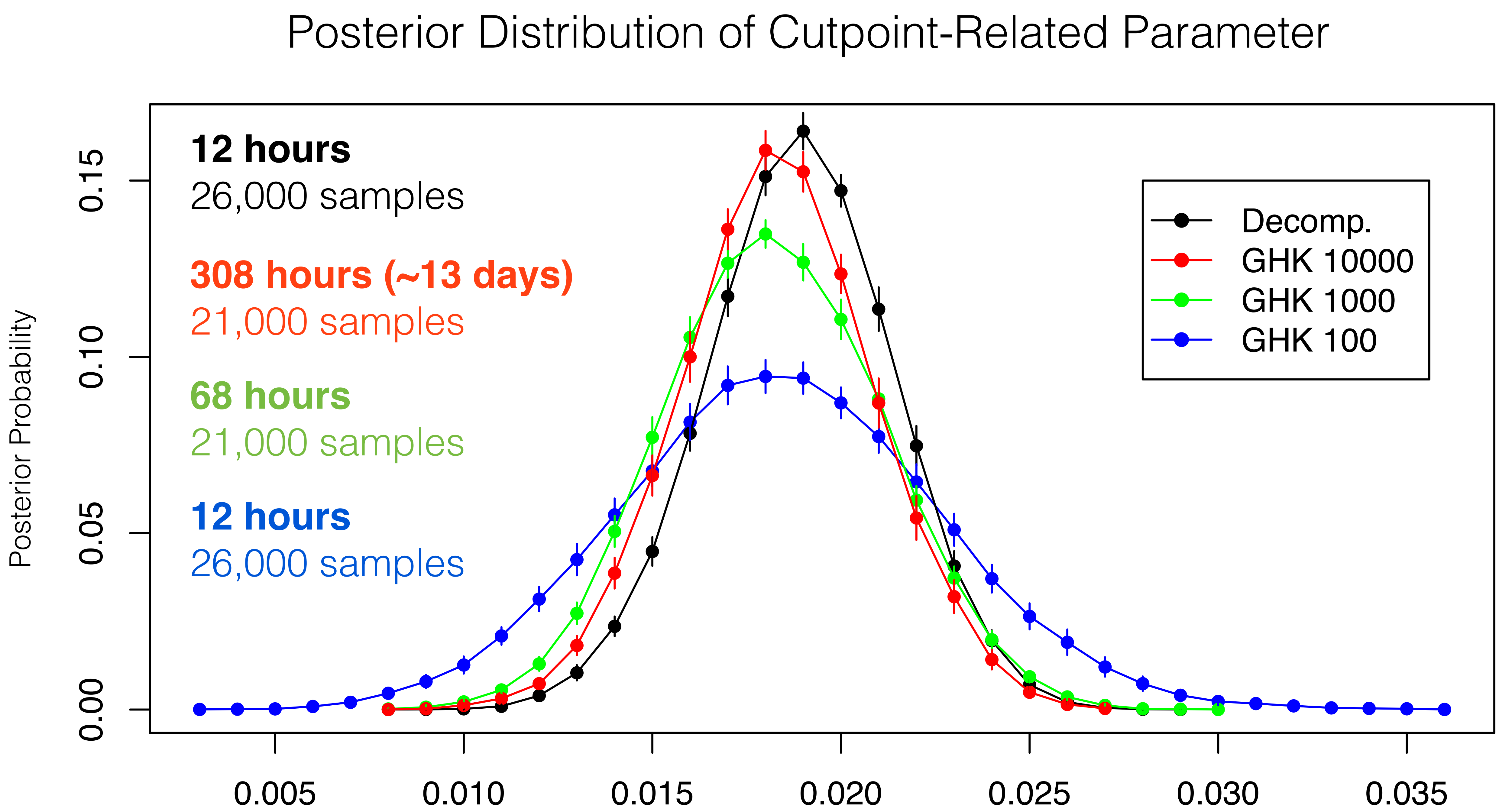
- An algorithm that uses Monte Carlo estimates of the integrals.
- Our algorithm that avoids these integrals entirely via the orthogonalizing latent variables.

Comparison

- The approximation approach uses the “GHK” Monte Carlo method for estimating the integrals.
  - Larger numbers of Monte Carlo samples should give better estimates.
  - Try 100 (default), 1,000 and 10,000 Monte Carlo samples

- Compare the estimated posterior distribution of a parameter related to the cutpoints for each method.

Comparison



Posterior Distribution of Cutpoint-Related Parameter
12 hours
26,000 samples
308 hours (~13 days)
21,000 samples
68 hours
21,000 samples
12 hours
26,000 samples
Decomp.
GHK 10000
GHK 1000
GHK 100
Posterior Probability
0.00
0.05
0.10
0.15
0.005
0.010
0.015
0.020
0.025
0.030
0.035

Summary

Introduced a covariance decomposition (new DA) to improve MCMC convergence rate.

- Carefully constructed latent variables.

Tuning of DA/latent variables driven by theoretical guidance.

Hans, C., Allenby, G.M., Craigmile, P.F., Lee, J., MacEachern, S.N. and Xu, X. (2012). Covariance decompositions for accurate computation in Bayesian scale-usage models. *Journal of Computational and Graphical Statistics*, **21**, 538–557.

Support from NSF grants:

DMS-0604963, DMS-0706948, SES-0437251