

# Interpretation and Inference in Mixture Models

John Geweke

University of Technology Sydney and University of Colorado

Workshop in Mixture Estimation and Applications

Edinburgh

March 3, 2010

Interpretation and Inference in Mixture Models:  
Simple MCMC Works

*Computational Statistics and Data Analysis* 51 (2007) 3529 - 3550

## Summary

- For functions of interest invariant to state permutation in mixture models labeling is not a problem. This includes prediction and inference about distributions.
- If the function of interest is not invariant to parameter permutation, e.g. in the interpretation of states, then
  - restrictions that are valid in the true or pseudo-true model eliminate labelling problems asymptotically;
  - restrictions that are not valid in the true or pseudo-true model fail to resolve labelling problems, even asymptotically;
  - in finite sample these restrictions may lead to multimodal likelihood functions whose implications for uncertainty are not transparent.
- The multimodal posterior arising from parameter permutation in mixture models introduces no complications for MCMC posterior simulators. This is true for any function of interest and any sample size.

## Outline

- 1 Introduction
- 2 The labeling problem
- 3 Interpreting the likelihood function
- 4 Bayesian modeling with MCMC
- 5 Searching for problems
- 6 Conclusion

## What is the labeling problem?

In the population ...

- Simple but prototypical mixture model:

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}, A, m) = \sum_{j=1}^m \pi_j f(\mathbf{y}_t \mid \boldsymbol{\theta}_j, A) \quad (t = 1, \dots, T)$$

$$\boldsymbol{\theta}_j \in \Theta_j = \Theta \quad (j = 1, \dots, m)$$

- Then for any permutation  $\rho(j)$  of  $j = 1, \dots, m$ ,

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}, A, m) = \sum_{j=1}^m \pi_j f(\mathbf{y}_t \mid \boldsymbol{\theta}_j, A)$$

$$= \sum_{j=1}^m \pi_{\rho(j)} f(\mathbf{y}_t \mid \boldsymbol{\theta}_{\rho(j)}, A) := p(\mathbf{y}_t \mid \rho(\boldsymbol{\theta}), A, m).$$

## What is the labeling problem?

In the sample ...

- Let  $\mathbf{y}_t^o$  denote the observed value of  $\mathbf{y}_t$  and denote  $\mathbf{Y}^o = [\mathbf{y}_1^o \cdots \mathbf{y}_T^o]$ .
- The likelihood function is

$$L(\boldsymbol{\theta}; \mathbf{Y}^o, A, m) = p(\mathbf{Y}^o \mid \boldsymbol{\theta}, A, m) = \prod_{t=1}^T p(\mathbf{y}_t^o \mid \boldsymbol{\theta}, A, m)$$

- Recalling

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}, A, m) = p(\mathbf{y}_t \mid \rho(\boldsymbol{\theta}), A, m) \quad (t = 1, \dots, T)$$

- the likelihood function shares the same property:

$$L(\boldsymbol{\theta}; \mathbf{Y}^o, A, m) = L(\rho(\boldsymbol{\theta}); \mathbf{Y}^o, A, m).$$

## What is the labeling problem?

In the posterior distribution ...

- If the prior density also shares the same property, namely

$$p(\boldsymbol{\theta} | A) = p(\rho(\boldsymbol{\theta}); A),$$

- then so does the posterior density

$$\begin{aligned} p(\boldsymbol{\theta} | \mathbf{Y}^o, A) &\propto p(\boldsymbol{\theta} | A) L(\boldsymbol{\theta}; \mathbf{Y}^o, A, m) \\ &= p(\rho(\boldsymbol{\theta}) | A) L(\rho(\boldsymbol{\theta}); \mathbf{Y}^o, A, m) \\ &\propto p(\rho(\boldsymbol{\theta}) | \mathbf{Y}^o, A). \end{aligned}$$

- It follows that the posterior density consists of  $m!$  reflected copies of the same surface.

## Some influential reactions to the problem

- Celeux, Hurn and Robert, *JASA* 2000 (121 ISI citations)  
Computational and Inferential Difficulties with Mixture Posterior Distributions:
  - "...almost the entirety of MCMC samplers implemented for mixture models has failed to converge."
- *Journal of Time Series*, November 2009, in a model mixing multivariate normal distributions:
  - "We prefer to update the parameters by random walk Metropolis-Hastings moves, because the Gibbs sampler is less able to traverse the posterior surface and to escape local modes, as pointed out by Celeux et al. (2000) for univariate mixture models."

- Jasra, Holmes and Stephens, *Statistical Science* 2005 (53 ISI citations)

Markov chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling:

- “While MCMC provides a convenient way to draw inference from complicated statistical models, there are many, perhaps under-appreciated, problems associated with the MCMC analysis of mixtures.”
- Conclusion: “The Gibbs sampler is not always appropriate for sampling from a mixture posterior”
- and when it is, “It is then meaningless to draw inference directly from MCMC output using ergodic averaging”

## Interpreting the Likelihood Function

- Illustration: Mixture of two normal distributions

$$p(y | \mu_1, \mu_2, \sigma_1, \sigma_2, \pi, ) = \pi f_N(y | \mu_1, \sigma_1) + (1 - \pi) f_N(y | \mu_2, \sigma_2)$$

- Artificial data with

$$\mu_1 = 1, \mu_2 = 2, \sigma_1^2 = 0.1, \sigma_2^2 = 0.15, \pi = 0.5$$

- Model (but not data) taken from Frühwirth-Schnatter, Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models, *JASA* 2001.

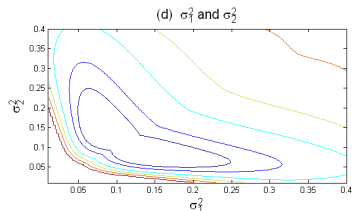
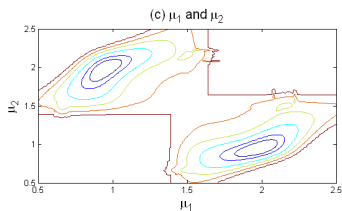
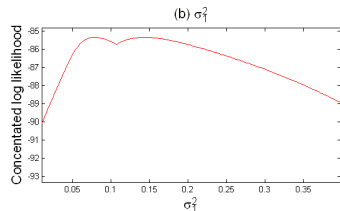
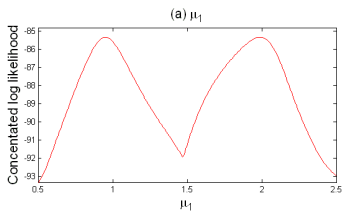


Figure:  $\mu_1 = 1$ ,  $\mu_2 = 2$ ,  $\sigma_1^2 = 0.1$ ,  $\sigma_2^2 = 0.15$ ,  $\pi = 0.5$ ,  $T = 10$

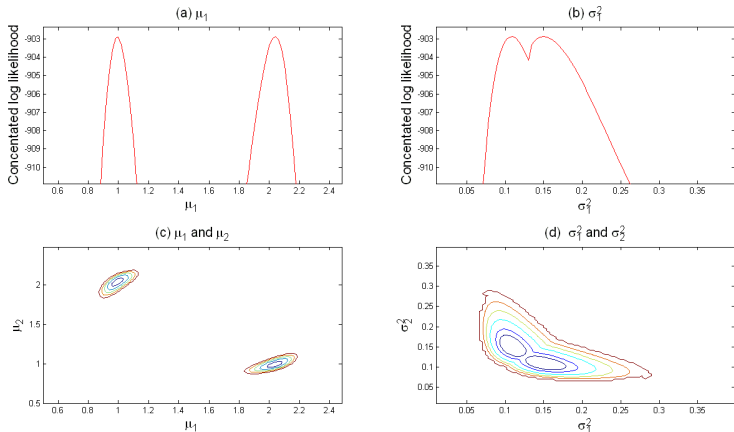


Figure:  $\mu_1 = 1$ ,  $\mu_2 = 2$ ,  $\sigma_1^2 = 0.1$ ,  $\sigma_2^2 = 0.15$ ,  $\pi = 0.5$ ,  $T = 1000$

## Is multimodality a problem?

- Let  $h(\boldsymbol{\theta})$  denote the function of interest.
- There is no multimodality problem from label permutation *per se* if

$$h(\boldsymbol{\theta}) = h(\rho(\boldsymbol{\theta})) \quad \forall \boldsymbol{\theta} \in \Theta, \forall \rho.$$

- Leading example:  $h$  depends only on

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}, A, m) = \sum_{j=1}^m \pi_j f(\mathbf{y}_t \mid \boldsymbol{\theta}_j, A).$$

- Flexible modeling of distributions
- Prediction

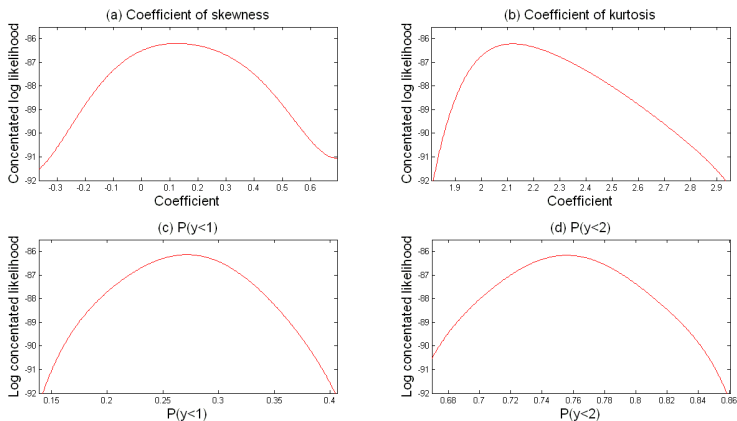


Figure:  $\mu_1 = 1$ ,  $\mu_2 = 2$ ,  $\sigma_1^2 = 0.1$ ,  $\sigma_2^2 = 0.15$ ,  $\pi = 0.5$ ,  $T = 100$

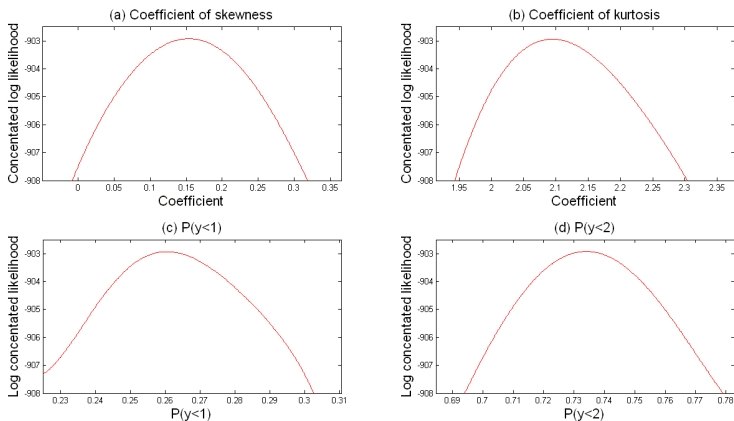


Figure:  $\mu_1 = 1$ ,  $\mu_2 = 2$ ,  $\sigma_1^2 = 0.1$ ,  $\sigma_2^2 = 0.15$ ,  $\pi = 0.5$ ,  $T = 1000$

## Is multimodality a problem?

- It is at least a potential problem if

$h(\boldsymbol{\theta}) \neq h(\rho(\boldsymbol{\theta}))$  for at least some  $\boldsymbol{\theta} \in \Theta$ , for at least some  $\rho$ .

- Leading examples:
  - State classification
  - Properties of states

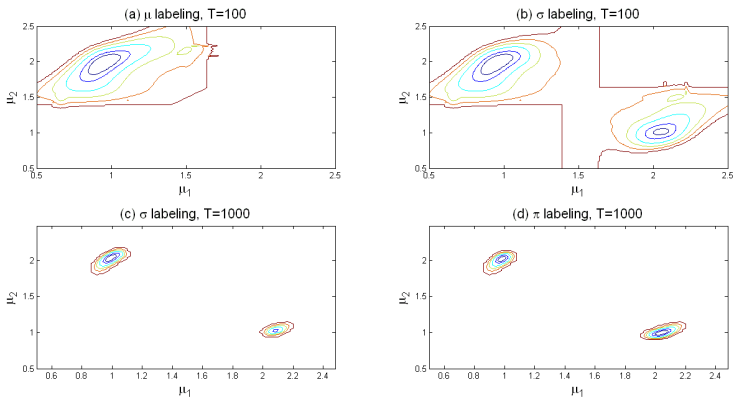


Figure:  $\mu_1 = 1$ ,  $\mu_2 = 2$ ,  $\sigma_1^2 = 0.1$ ,  $\sigma_2^2 = 0.15$ ,  $\pi = 0.5$

## Bayesian modeling with MCMC

- The model:

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}, A, m) = \sum_{j=1}^m \pi_j f(\mathbf{y}_t \mid \boldsymbol{\theta}_j, A) \quad (t = 1, \dots, T)$$

$$p(\boldsymbol{\theta}_i \mid A) = p(\boldsymbol{\theta}_j \mid A) \quad \forall i, j$$

$$\boldsymbol{\theta}_j \in \Theta_j = \Theta \quad (j = 1, \dots, m)$$

- Augment the model with latent states  $\tilde{s}_t$ ,

$$P(\tilde{s}_t = j \mid \tilde{s}_r (r \neq t), \boldsymbol{\theta}, A) = \pi_j \quad (j = 1, \dots, m; t = 1, \dots, T).$$

- Then

$$p(\mathbf{y}_t \mid \tilde{s}_t = j, \boldsymbol{\theta}, A) = f(\mathbf{y}_t \mid \boldsymbol{\theta}_j, A)$$

## Purported problems

- The stated problem: The obvious Gibbs sampler converges very slowly due to multimodality.
- Celeux, Hurn and Robert (2000):
  - “We know that some regions of the parameter space have not been visited by the Markov chain. . . Although we may be somewhat presumptuous, we consider that almost the entirety of MCMC samplers implemented for mixture models has failed to converge!”
- Jasra, Holmes and Stephens (2005):
  - “It was established by Celeux, Hurn and Robert (2000) that the Gibbs sampler is not always appropriate for sampling from a mixture posterior. This is because of the inability of the Gibbs sampler to traverse the support of highly multimodal distributions.”

## Frühwirth-Schnatter (2001)

- At the end of each iteration of the simulator,
- randomly permute the parameter vectors  $\theta_j$ ;
- equivalently, permute the state to which each of the  $m$  vectors is assigned.
- Jasra, Holmes and Stephens (2005):
  - “We note that it is always possible to achieve label switching between the  $m!$  modes by the simple addition of a proposal that suggests a random permutation of the labels. Our insistence on searching for an algorithm that can achieve symmetry without such a move is that any concerns over convergence are not necessarily dealt with by such a strategy, which simply alleviates the most obvious symptom.”

- The JHS (2005) argument is persuasive only to the extent that there are mixing problems beyond labeling.
- But Celeux, Horn and Robert (2000) argue
  - “The main defect of the Gibbs sampler from our perspective is the ultimate attraction of the local modes”;
  - and go on to propose a simulated tempering scheme that makes no use of the permutation invariance of the posterior distribution.
  - CHR (2000) demonstrates (and states) that this approach requires substantial skill, trial and error on the part of the investigator and increases computation time by several orders of magnitude.

## Attractions of Frühwirth-Schnatter (2001)

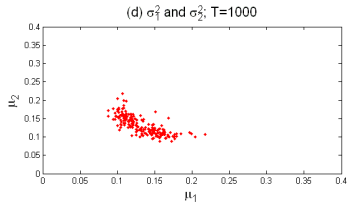
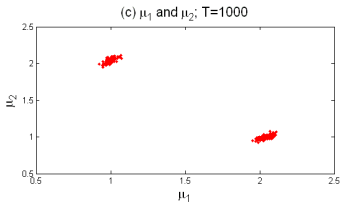
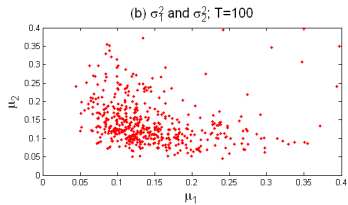
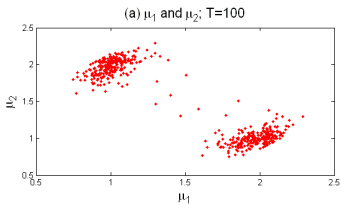
- Exploits the analytical structure of the troublesome posterior
- Simple and effective
- Requires little additional time (or investigator or computer)
- But there is an improvement ...
  - that is simpler and
  - makes it clear that the problem for Gibbs sampling algorithms emphasized in CHR (2000) and JHS (2005) does not exist.

## Improvement on Frühwirth-Schnatter (2001)

- *Core simulator*: Obvious MCMC, alternating between  $\theta_j$ 's states  $\tilde{s}_t$ .
- *Permutation-augmented simulator*: All  $m!$  copies of the core simulator
- If priors and functions of interest are permutation-invariant: then the copies are irrelevant
- If priors have labeling constraints:
  - Regard the permutation-augmented simulator as an importance sampler
  - $\implies$  Pick the copy that satisfies the constraints
  - $\iff$  Proposition 3.1 of Stephens (1997) D.Phil. thesis.

## Permutation-augmented MCMC in the example

- Prior distribution
  - Conditionally conjugate, proper, diffuse
- Fully compiled code
  - Known to be correct (see Geweke (2004), JASA)
  - 11,000 iterations of which first 1,000 discarded
  - Permutation-augmented simulator
  - $T = 100$ : 0.5 seconds;  $T = 1,000$ : 3.0 seconds



## Mixing problems beyond labeling?

- Jasra, Holmes and Stephens (2005):
  - “We note that it is always possible to achieve label switching between the  $m!$  modes by the simple addition of a proposal that suggests a random permutation of the labels. Our insistence on searching for an algorithm that can achieve symmetry without such a move is that any concerns over convergence are not necessarily dealt with by such a strategy, which simply alleviates the most obvious symptom.”
- Do the random permutation simulator (Fruhwirth-Schnatter) and the permutation-augmented simulator (Geweke) obscure other difficulties peculiar to MCMC simulation in mixture models?
- There is no reason to think so, and none has been offered.
- But this question can be examined carefully in specific cases.

## Example 1: Frühwirth-Schnatter (2001) mixture of normals

$$y_t \mid (\tilde{s}_t = j) \sim N \left[ \alpha_0 + \alpha_j, (h_0 h_j)^{-1} \right], \quad P(\tilde{s}_t = j \mid \boldsymbol{\pi}) = \pi_j \quad (j = 1, \dots, m)$$

- Prior distribution:

$$\alpha_0 \sim N(\underline{\alpha}^*, \underline{h}_\alpha^{*-1}),$$

$$\alpha_j \mid h_0 \sim N \left[ 0, (\underline{h}_\alpha h_0)^{-1} \right] \quad (j = 1, \dots, m),$$

$$\underline{s}^{*2} h_0 \sim \chi^2(\underline{\nu}^*),$$

$$\underline{\nu} h_j \sim \chi^2(\underline{\nu}) \quad (j = 1, \dots, m),$$

$$\boldsymbol{\pi} \sim \text{Beta}(\underline{r}, \dots, \underline{r}).$$

- Hyperparameter values:

$$\underline{\alpha}^* = 0, \quad \underline{h}_\alpha^* = 0.01; \quad \underline{s}^{*2} = 0.5, \quad \underline{\nu}^* = 5;$$

$$\underline{r} = 1; \quad \underline{\nu} = 6, \quad \underline{h}_\alpha = 1.$$

## Example 1 (continued)

- Data generation process:  $m = 2$ ,  $\pi_1 = \pi_2 = 0.5$ ,  $\mu_1 = 1$ ,  $\mu_2 = 2$ ,  $\sigma_1^2 = 0.1$ ,  $\sigma_2^2 = 0.15$
- Sample sizes:  $T = 100$ ,  $T = 1000$
- 1000 executions of the permutation-augmented MCMC posterior simulator
  - Initial draw from the prior distribution
  - 1000 burn-in iterations
  - 10,000 following posterior draws retained
  - Execution time 0.5 seconds ( $T = 100$ ), 3.0 seconds ( $T = 1000$ ).

## Example 1 (concluded)

- Posterior moments computed
  - 4 label invariant moments: skewness, kurtosis,  $P(y \leq 1)$ ,  $P(y \leq 2)$
  - 11 label-sensitive moments with  $\sigma^2$  labeling (  $\implies$  difficult interpretation)
- Tests:
  - Numerical standard errors computed within the chain and across chains agree.
  - Standard tests (Kolmogorov-Smirnov, Jarque-Bera) reject normality at rates consistent with test size.

## Example 2: Geweke-Keane (2007) mixture of experts dynamic model

- Global features of the model:
  - Probit link function
  - $m = 3$  components
  - $T = 2628$  daily returns for S&P 500, January 1990 - December 1999
  - Covariates include previous day's return and geometrically declining average of past absolute returns
- 1000 executions of the permutation-augmented MCMC posterior simulator
  - Initial draw from the prior distribution
  - 2000 burn-in iterations
  - 10,000 following posterior draws thinned to 100 and retained
  - Execution time 3 minutes

## Example 2 (concluded)

- Posterior moments computed:
  - 3 label-invariant moments (c.d.f. evaluations)
  - 3 label sensitive moments (Means of  $\sigma_1 < \sigma_2 < \sigma_3$ )
- Tests:
  - Numerical standard errors computed within the chain and across chains agree.
  - Standard tests (Kolmogorov-Smirnov, Jarque-Bera) reject normality at rates consistent with test size.

## Conclusion

- Posterior moments of functions sensitive to state permutation in mixture models may be difficult to interpret.
  - This can be a serious problem but it is entirely independent of the mechanics of posterior simulation.
- The labeling problem, *per se*, **never** presents a problem for MCMC posterior simulators in mixture models.
  - For permutation-sensitive posterior moments, permuted copies should be selected if and as required.
  - For permutation-insensitive posterior moments, these copies are irrelevant.
  - There is no reason to expect further problems specific to mixture models, and no evidence that any exist.
- Simple MCMC works in mixture models.