Posterior Exploration of Markov Chains in a Bayesian Analysis of Discrete Finite Mixture Models

Peter M. Saama
Quantitative Genetics Lab, Michigan State University, East Lansing, MI, USA.

Abstract
Markov Chain Monte Carlo (MCMC) methods make possible the use of flexible Bayesian models that would otherwise be computationally infeasible. In essence, MCMC methods involve sampling from a particular posterior distribution by simulating a Markov Chain with that posterior as its stationary density. However, one must decide when to stop the iterations, or more precisely, judge how close the underlying algorithm is to convergence after a specified number of iterations. Furthermore, an MCMC simulation converges to a target distribution, rather than a target point, and inferences are based on moments of that target distribution. Problems in mixtures arise because the mixing distribution is unknown and, in Bayesian nonparametric analysis, it is considered as a random distribution function which is usually given a Dirichlet process prior. This paper examines the performance of an L^1 distance convergence diagnostic which assesses the convergence of the joint density of a Gibbs Sampler algorithm in a discrete finite mixture model. Basically, the convergence diagnostic method measures the difference between distributions of a fixed number of replications sub-sampled from independent Markov chains by (over)estimating the total variation distance between the densities. Initially, the problem of determining the probability that a given data point is assigned to a given component in a mixture is addressed. Then the convergence diagnostic is illustrated and interpreted using simulated data. It is shown, that this diagnostic has advantages over many existing convergence diagnostics in terms of consistency, applicability, computational expense, and interpretability.

1. Background

Mixture distributions (Everitt, 1981; Titterington et. al, 1985; Ferguson, 1983), are typically used to model data in which each observation is assumed to have arisen from one of a number of distinct sub-populations. In situations where the number of components is unknown, mixture densities of the form

$$\sum_{j=1}^{k} \pi_j \, \mathrm{N}\left(\theta_j, \sigma_j^2\right) \qquad [1]$$

have found their widest applications as a model based clustering procedure; π_j is the probability that observation y_i comes from the j^{th} component of the mixture. Herein θ will denote the set of all unknown parameters and $p(.|.)$ is used to denote a generic conditional probability density function.

A mixture of two normal densities was first considered by Pearson in 1894 with parameter estimates obtained from the method of moments and involved the solution of a ninth-degree polynomial. The seminal paper on the EM algorithm (Dempster, Laird and Rubin, 1977) has greatly stimulated work on finite mixtures of distributions. Applications of mixture models reported by Titterington, Smith and Makov (1985) and McLachlan and Basford (1988) use the Expectation Maximization (EM) algorithm. Its disadvantages include:

1) extreme slowness of convergence when the proportion of missing data is high;

2) absence of standard errors from the information matrix at convergence.

Competitors of EM are Gauss-Newton (Lois, 1982; Aitkin et al, 1994), Fisher Scoring (Rao,

1948), and Differential Evolution (Price and Storn, 1997). The Gauss-Newton (GN) algorithm, is not guaranteed to converge when the log-likelihood is not concave but when it does converge, this rate of convergence is usually quadratic, compared to linear from EM. A hybrid EM-GN was proposed and implemented by Aitkin et al (1994).

A Bayesian analysis of mixture models presents certain advantages over the classical approaches. In theory, quantities of interest are written down as integrals of the form

$$\mathrm{E}\left(G(\Theta)|y^k\right)=\int G(\Theta)\,p(\theta\mid y^k)d\theta \qquad [2]$$

where $G(\Theta)$ is the kernel distribution given the data, $G(\Theta)|y^k$. In practice these integrals cannot be evaluated by traditional numerical methods. When the number of groups is assumed known, Markov Chain Monte Carlo (MCMC) methods such as the Gibbs sampler can be used to perform the integration. These methods rely on the construction of a Markov chain $\left\{\Theta^{(t)}\right\}$ with the property that the *sample path average*

$$F_N=\frac{1}{N}\sum_{t=1}^{N}G\left(\Theta^{(t)}\right) \qquad [3]$$

is a consistent estimator for $\mathrm{E}\left(G(\Theta)|y^k\right)$, in that it converges to $\mathrm{E}\left(G(\Theta)|y^k\right)$ as $N\to\infty$. Such a markov chain can be constructed in situations where it is not possible to sample from $p(\theta\,|y^k)$ directly, as is usually the case in mixture models.

It is a well-known problem in finite mixture models that the parameters are fundamentally not identifiable in that the likelihood parameters corresponding to the k components is unchanged by permutations of the component labels 1, ..., k. In a Bayesian analysis, this typically leads to a joint density of the parameters which is highly multimodal which causes label-switching in the Gibbs sampler output and makes inferences for individual components of the mixture meaningless. A common practice is to impose *identifiability constraints* on the model parameters such as $\sigma_1<\sigma_2<\ldots<\sigma_k$ but this is often not a satisfactory solution (Diebolt and Robert, 1994). Stephens (1997) suggests a general solution which involves permuting samples from the parameter posterior density so as to remove as much multimodality as possible and allows interpretations for groups to be discovered rather than imposed.

Problems with mixture distributions can arise when one combines information on a given trait from several herds which are distributed across a wide range of environments or when, over time, the trait of interest varies greatly when measured on a given animal such as milk yield in a lactation. We certainly should be concerned about mixture distributions when we combine information from several countries. Therefore, for many economically important dairy cattle traits, it seems appropriate to consider the distribution of the data as a mixture of parametric densities such as in [1]. However, this approach is not suitable for overdispersed categorical traits such as culling or survival rates. Such traits can be modeled as a discrete distribution on a finite number of mass points using nonparametric finite mixture models (Escobar, 1995).

In a Bayesian nonparametric analysis, the mixing distribution $G(\Theta)$ is unknown and it is usually given a Dirichlet process prior (Antoniak, 1974; Ferguson, 1983; Petrone, 1997). It has been shown that Markov chains resulting from Gibbs sampling for nonparametric mixtures are uniformly ergodic (Petrone *et al*, 1998). However, in the absence of any general techniques for *apriori* prediction of run lengths, it is necessary to carry out some

form of statistical analysis in order to assess convergence. A number of useful methods for *apriori* exploration of the target distribution have been proposed (see Brooks and Roberts; 1998). Several of the methods have problems with interpretability and are problem specific thus requiring different code to be written for each problem in order to produce the required output.

Brooks et al (1997) suggest an approach to diagnosing the convergence which attempts to obtain the upper bound on the L^1 distance between full dimensional kernel estimates from different chains. They illustrate how it can be applied to continuous distributions and indicate that it also applicable to MCMC algorithms for the analysis of discrete distributions. To the knowledge of the author no such application has been advanced in the literature.

The objective of this study was to examine the performance of an L^1 distance convergence diagnostic in a Bayesian analysis of discrete finite mixture models.

2. Bayesian discrete finite mixture model

The Estimation problem

Practical approaches for implementing Dirichlet process models have been developed by Escobar and West (1995) and MacEachern and Müller (1998). The model applies to data $y_i = y_{i1}, ..., y_{ik}$ which are assumed to be exchangeable, or as being independently drawn from some unknown distribution. The y_i may be multivariate with components that are real-valued or categorical and are seen as realizations of corresponding random variables $Y_i = Y_{i1}, ..., Y_{ik}$. Suppose we have observed or are interest in N of these vectors. We wish to find the predictive distribution for one or more of the unknown attributes given the values of the known attributes.

The Model

Assume that the parameters (θ) of the process generating the data are stable, and that given knowledge of these parameters, the distribution from which the y_i are drawn is a mixture of distributions of the form $F(\theta)$ with the mixing distribution over θ being $G(\Theta)$. Let the prior for this mixing distribution be a Dirichlet process (Ferguson, 1973), with concentration parameter α and base distribution $G_0(\Theta)$. The discrete finite mixture model is (Neal, 1998):

$$\begin{aligned} y_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid G(\Theta) &\sim G(\Theta) \\ G(\Theta) &\sim D(G_0(\Theta),\ \alpha) \end{aligned} \qquad [4]$$

where

$$D(G_0(\Theta),\ \alpha) = \left(\frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_g \phi_g^{(\alpha/k)-1} \right) \bullet \prod_{g,j} \left(\frac{\Gamma(\beta_j)}{\Gamma(\beta_j/N)^N} \prod_{\upsilon} \psi_{g,j,\upsilon}^{(\beta/N)-1} \right) \qquad [5]$$

Here ϕ_g is the probability of a mechanism $g = G_0(\Theta)$ being used, $\psi_{g,j,\upsilon}$ is the probability that a mechanism $g = G_0(\Theta)$ will produce value υ for attribute j, and it is customary to set $\beta_j = N$. Realizations of the Dirichlet process are discrete with probability one (Ferguson, 1983). Integrating over $G(\Theta)$ in model [4], provides a representation of the prior distribution of the θ_i in terms of conditional distributions of the form (Blackwell and MacQueen, 1973):

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} \sim \frac{1}{i-1+\alpha}\sum_{j=1}^{i-1}\delta(\theta_j) + \frac{\alpha}{i-1+\alpha}G_0(\Theta) \quad [6]$$

where $\delta(\theta)$ is the distribution concentrated at the single point θ.

The Gibbs Sampler

Exact computations for a Dirichlet process mixture model is infeasible when the number of observations is large. We can sample from the posterior distribution of the $\theta_i, \ldots, \theta_n$ by repeatedly drawing values of each θ_i from its conditional distribution given both the data and the θ_j for $j \neq i$ (written as θ_{-i}). From [6] the conditional distribution for $\theta_i \mid \theta_{-i}$ is,

$$\theta_i \mid \theta_{-i} \sim \frac{1}{n-1+\alpha}\sum_{j\neq i}\delta(\theta_j) + \frac{n}{n-1+\alpha}G_0(\Theta) \quad [7]$$

The full conditional distribution is,

$$\theta_i \mid \theta_{-i}, y_i \sim \sum_{j\neq i} q_{i,j}\delta(\theta_j) + r_i H_i \quad [8]$$

where H_i is the posterior distribution for θ based on the prior $G_0(\Theta)$ and y_i; Values for $q_{i,j}$ and r_i are defined as

$$q_{i,j} = b\ p(\theta \mid y^k) \quad [9]$$

$$r_i = b\alpha\int p(\theta \mid y^k)dG_0(\Theta) \quad [10]$$

Here b is a normalizing constant such that $\sum_{j\neq i} q_{i,j} + r_i = 1$. Computation of $\int p(\theta \mid y^k)dG_0(\Theta)$ is possible when $G_0(\Theta)$ is the conjugate prior. Neal (1998) proposes the following auxilliary variable Gibbs sampling algorithm:

Let the Markov chain consist of $c_1, \ldots, c_n$ and $\phi = (\phi_1, \ldots, \phi_n)$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: Let k^- be the number of distinct c_j for $j \neq i$, and let $h = k^- + r$; here $r \geq 1$. Label the c_j with values in $\{1, \ldots, k^-\}$. If $c_i = c_j$ for some $j \neq i$, draw values independently from $G_0(\Theta)$ for those ϕ_c for which $k^- < c \leq h$. If $c_i \neq c_j$ for all $j \neq i$, let c_i have the label $k^- + 1$, and draw values independently from $G_0(\Theta)$ for those ϕ_c for which $k^- + 1 < c \leq h$. Draw a new value for c_i using the following probabilities:

$$P(c_i = c \mid c_{-i}, y_i, \phi_1, \ldots, \phi_h) = \begin{cases} b\dfrac{n_{-i,c}}{n-1+\alpha}\, p(\varphi_c \mid y_i) & \text{for } 1 \leq c \leq k^- \\ b\dfrac{\alpha/r}{n-1+\alpha}\, p(\varphi_c \mid y_i) & \text{for } k^- \leq c \leq h \end{cases}$$

where $n_{-i,c}$ is the number of c_j for $j \neq i$ that are equal to c, and b is a normalizing constant. Change the state to contain only those ϕ_c that are currently associated with one or more observations.

- For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ subject to $c_i = c$. Thus, the update to ϕ_c must leave this distribution invariant.

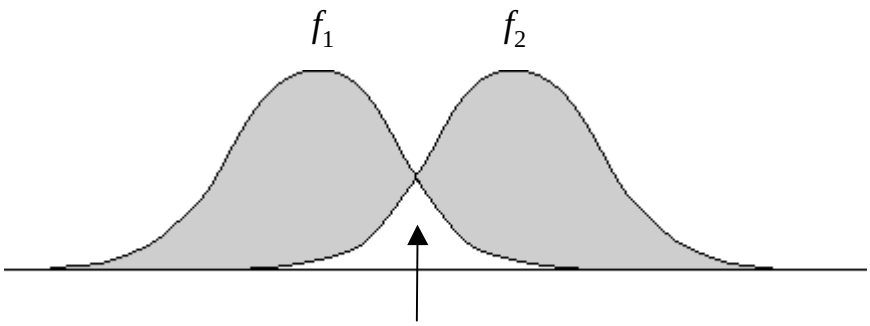

Figure 1. Illustrating the L^1 distance densities f_1 and f_2

3. The total variation convergence diagnostic

The L^1 (total variation) distance between two probability measures f_1 and f_2 is given by (Dellaportas, 1995),

$$|f_1 - f_2| = \frac{1}{2}\int_{-\infty}^{\infty} |f(x) - g(x)|\, dx. \qquad [11]$$

In [11], f_1 and f_2 admit $L^p = L^p[0,1] = L^p([0,1], B, m)$, where B is the Borel σ–algebra and m is Lebesgue measure. This definition covers probability density functions for distributions encountered in dairy cattle breeding research. Figure 1 below provides an illustration of the L^1 distance between two densities f_1 and f_2. The L^1 distance is the area defined by the clear region.

Brooks *et al*. (1997) suggest that we run m independent chains and split each chain into l blocks of n_0 observations, with $\Theta_i^{(t)}$ denoting the state of chain i at time t. For the Gibbs Sampler, the Rao-Blackwell estimator for the density of Θ in l^{th} block and i^{th} chain is,

$$\mathrm{K}_{il}(\Theta) = \sum_{t=(l-1)n_0+1}^{l_{n_0}} \frac{\mathrm{K}\left(\Theta_i^{(t)}, \Theta_i^{(t-1)}\right)}{n_0} \qquad [12]$$

where $\mathrm{K}\left(\Theta_i^{(t)}, \Theta_i^{(t-1)}\right)$ is the one-step transition kernel for the chain moving from state $\Theta_i^{(t)}$ to $\Theta_i^{(t-1)}$. The mean between chain distance is

$$B_l = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} 1 - \min\left(1, \frac{\mathrm{K}_{il}(\Theta)}{\mathrm{K}_{jl}(\Theta)}\right) \qquad [13]$$

A characteristic jump point in B_l value from 1 to 0 in a relatively small number of steps suggests convergence to the stationary distribution. This manifestation is consistent with the "cut-off phenomenon" of Diaconis (1988). Overdispersed starting points for each chain allow for better comparisons between increasing within-sequence variance and the decreasing between-sequence variance.

4. Example

Simulated data

Binary data were simulated from the Binomial distribution. Each case was composed of ten binary attributes. The distribution of these binary vectors was a mixture of four component distributions, in each of which the ten attributes were independent. The four

components each had probability 0.25 of being chosen to generate a case. The probabilities for each component for each binary attributes were as shown in Table 1. Each row gives the probabilities of each of the attributes being '1' for one component of the mixture. The columns are for the ten binary attributes in each case. The vectors generated in this way can be seen as coming from one of four "patterns": 0000011111, 0111100001, 1001100111, and 1110011001, but with each bit of the chosen pattern having a small probability of being switched (ranging from 0.1 to 0.3) in any particular case. Five hundred cases were generated from this distribution. This data structure was designed to give a finite number of components in the mixture distribution. The ten binary attributes can be taken to represent records on ten animals. The four patterns could represent four traits measured on the ten animals.

The simulated data were modeled as a mixture of four components. Thus, each component of the mixture defines a joint probability for the 10 sets of target attributes. Three independent Markov chains were simulated using the priors in Table 2.

The probability that each component of the mixture would give each target attribute a "1" or "0" was determined from the logistic function:

$$\Pr(c_i = 1) = \frac{1}{(1 + \exp(-offset))}$$

The *offset* parameters for each attribute were given vague Gaussian prior distributions with means and standard deviations that were variable hyperparameters to favor convergence to a stable process. Specifically, the priors for the top-level and lower-level standard deviations for the *offset* were from independent $\text{Gamma}(.05, .5)$ and $\text{Gamma}(.05, .2)$, respectively; the prior for the mean component offsets was 10.

In each of the three chains, the 500 cases from the simulated data were used as a training set to generate 1000 states for each of the parameters. Then a Markov of chain of length 5000 was simulated.

Table 1. Probabilities for each of four components for each of ten binary attributes. Data for each component were generated from the Binomial distribution using these probabilities.

	Attribute									
Component	1	2	3	4	5	6	7	8	9	10
1	.1	.2	.2	.2	.2	.8	.8	.8	.8	.7
2	.1	.8	.8	.8	.8	.2	.2	.2	.2	.7
3	.9	.2	.2	.8	.8	.2	.2	.8	.8	.7
4	.9	.8	.8	.2	.2	.8	.8	.2	.2	.7

Table 2. Priors for the concentration parameter of the Dirichlet distribution.

Chain	α	**Interpretation**
1	1/4	unequal mixing proportions
2	2/4	mixing proportions in either direction
3	3/4	equal mixing proportions

Table 3. Posterior means and SD's (in brackets) for model parameters following from three parallel chains.

	$m = 1$	$m = 2$	$m = 3$
θ_1	0.298	0.297	0.297
	(0.022)	(0.020)	(0.021)
θ_2	0.561	0.561	0.561
	(0.021)	(0.020)	(0.022)
θ_3	0.806	0.805	0.806
	(0.019)	(0.018)	(0.018)

5. Results

Under the prior specifications for all model parameters, the corresponding posterior means and standard deviations for three of four parameters are presented in Table 3. Data for one of the components are not presented because the kernel estimates for that parameter stabilized and remained constant at 1.

We note that the posterior means and standard deviations for the parameters from the three parallel chains are very similar. Trace plots of the raw p values and the kernel density plots for the three parameters are shown in Figure 2. The trace plots suggest evidence of mixing in the chains. We observe that posterior marginal distributions were skewed to the left for θ_1 but were skewed to the right for θ_2 and θ_3 thus confirming that the joint posterior distribution was multi-modal.

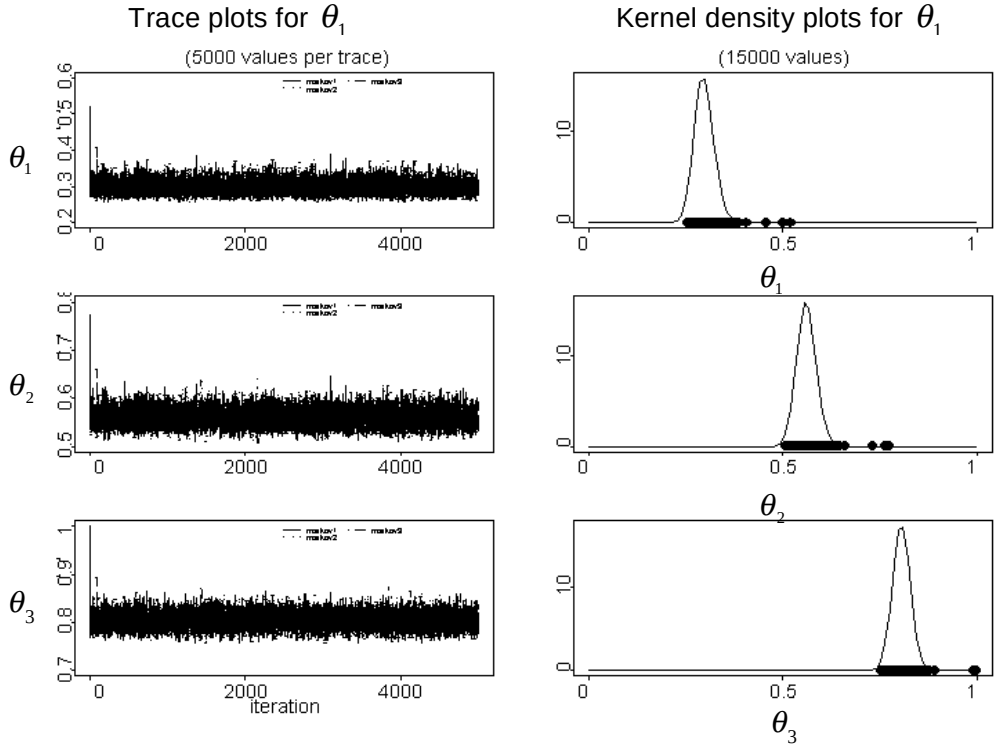

Figure 2. Traces of the raw p values and the kernel density estimates for three components in a discrete finite mixture model.

Figure 3 displays the total variation diagnostic. Brooks *et al.* (1998) point out that if n_0 is small the total variation diagnostic may not indicate convergence when it is achieved. For this example, the sharp drop in B_l at $l \approx 7$ is a clear indication that convergence was achieved by this point. This jump is a manifestation of the "cut-off phenomenon". However, these plots are hard to interpret. Hence, when n_0 is small relative to the length of the parallel chains, trace plots of a widowed mean of the B_l values could provide more interpretable results. The plots in Figure 2 support the total variation diagnostic because the raw *p* values for all 3 chains settled to the same time that the diagnostic indicated convergence.

Figure 4 shows the performance of the diagnostic with $n_0 = 50$. This plot confirms that convergence had been attained at $l < 50$. We also note that values for the diagnostic stabilized at $B_l \approx .16$ for $n_0 = 5$ while values for B_l were $\approx .025$ when $n_0 = 50$. We can see that, in spite of the volatility of the B_l statistic, values for B_l were smaller when the block length was increased.

Most of the existing convergence diagnostics are based on output analysis. The Geweke (1992) diagnostic looks at trace plots of z-scores for the first and last segment of the simulation. Trace plots of the first segment in the states for the 3 parallel chains are presented in Figure 5.

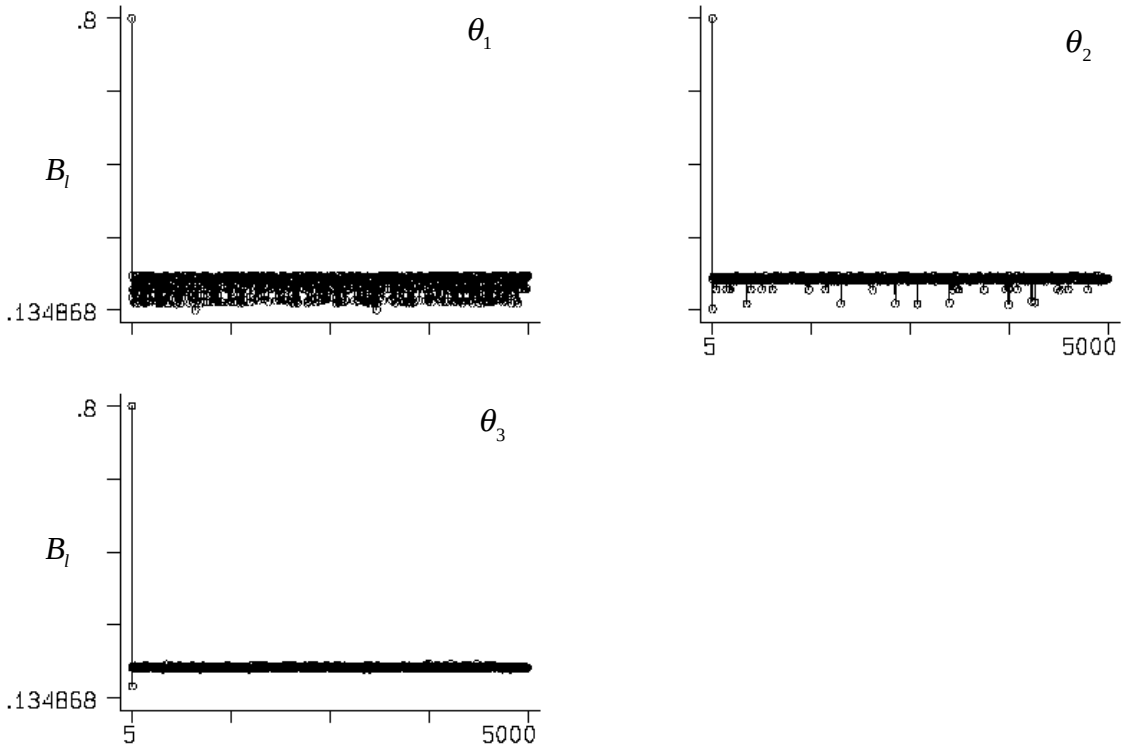

Figure 3. Total variation diagnostic for parameters with a small block length, n_0 = 5.

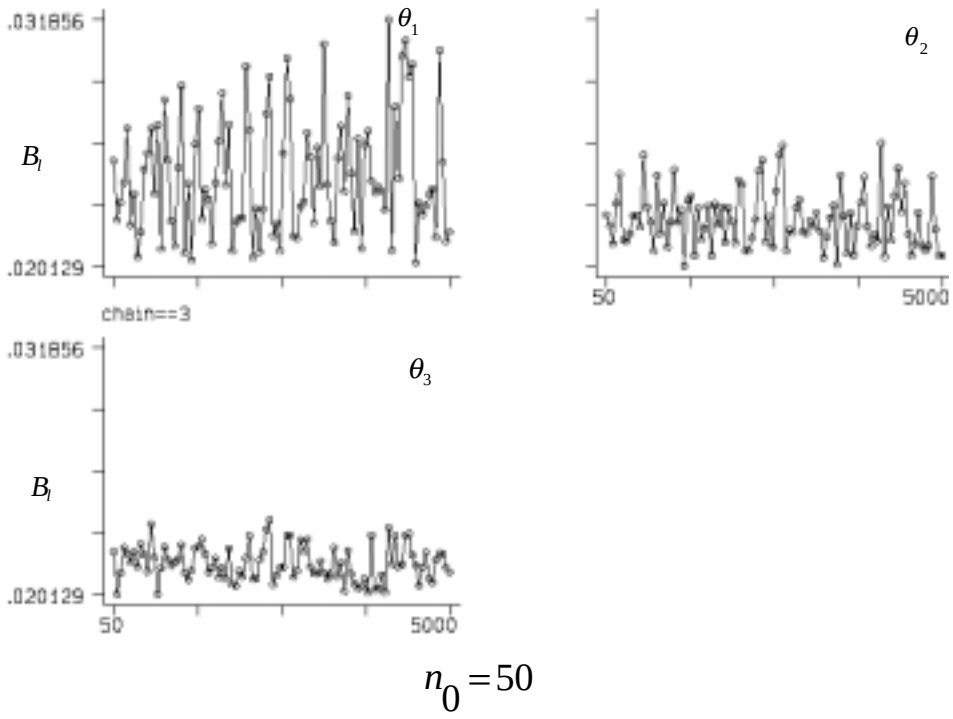

Figure 4. Total variation diagnostic for parameters with a large block length, n_0=50.

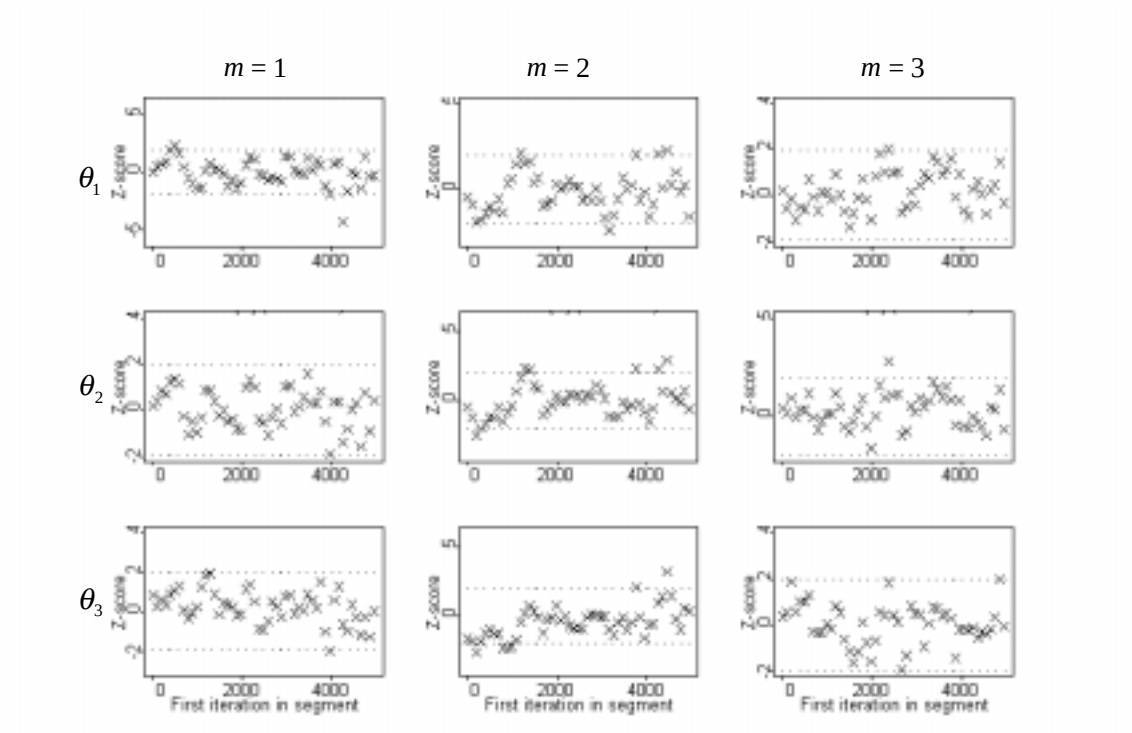

Figure 5. Trace plots for the Geweke (1992) convergence diagnostic.

For these data and the underlying probability model, the Geweke diagnostic does not give consistent results thus making it difficult to diagnose convergence. It is apparent at $l \approx 4040$ that the first chain ($m = 1$) jumped to a new and less stable state. Observe that the output is much harder to interpret and that only part of the available information was used.

In Figure 6, plots for the Gelman and Rubin (1992) convergence diagnostic support the total variation diagnostic. Convergence was apparent at $l \approx 5$ which is consistent with the plots for the B_l statistic in Figure 2. However, this method is based on output analysis.

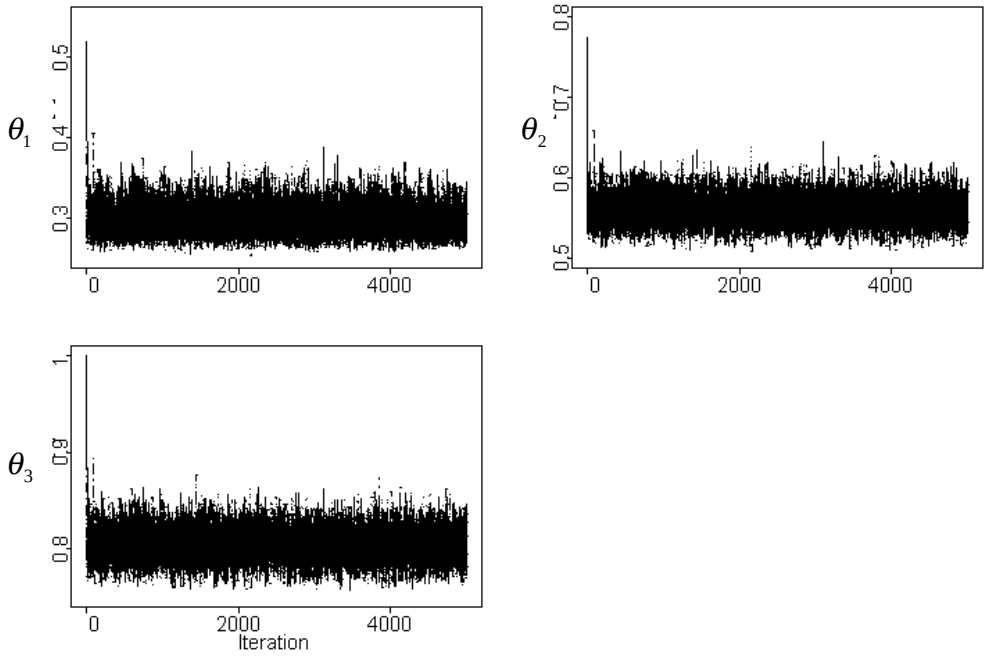

Figure 6. Trace plots for the Gelman and Rubin (1992) convergence diagnostic.

6. Discussion

We have demonstrated that in a Bayesian analysis of discrete finite mixture models, where the posterior surface may be less well understood, sampling from the conjugate priors can lead to stationary posterior distributions. The auxilliary-variable Gibbs sampler provides consistent estimates of the unknown parameters. Factors affecting the performance of the Bayesian discrete finite mixture model include: 1) size of the training sets; larger training items have enough data to force most of the probability to the region near the true parameter values; 2) starting values for the hyperparameters; 3) number of parallel sequences; 4) slow and poor mixing in the markov chains.

The performance of an L^1 distance convergence diagnostic was examined. The

main advantages of this diagnostic are: 1) it provides interpretable output; 2) The "cut-off" makes it easy to assess convergence; 3) it elicits some mathematical interpretation; B_l stabilizes around 0.025; 4) it is based upon multiple replications and a joint density; 5) it makes full use of all available information; 6) it can be easily adapted to provide a convergence diagnostic for the parallel Gibbs Sampler.

Disadvantages of the total variation diagnostic include: 1) computational expense; 2) it requires transition probabilities from the target distribution which is not a problem when full conditionals are available; 3) choice of the block length, n_0 . Experience is required to pick a suitable value.

This paper demonstrates that methods are available to handle overdispersed traits. It is postulated that these methods can be useful in international genetic evaluations where records are from pooled from heterogeneous environments. The convergence diagnostic examined in this paper can be applied during or after an MCMC simulation which involves at least two independent parallel chains. Further work is needed to compare Dirichlet process mixture models with existing approaches for handling overdispersed data.

ACKNOWLEDGEMENTS

Drs. I. L. Mao, B. D. Banks, R. J. Tempelman, T. A. Ferris (Michigan State Univ., USA) who made it possible for me to present this work at Computational Cattle Breeding '99. Dr. Stephen Brooks (Univ. of Cambridge, UK) is cited for providing preprints of the convergence algorithm. The author expresses gratitude towards Dr. Neal (Univ. of Toronto, Ontario) for his assistance in implementing the Gibbs Sampler. Major components of the MSU Quantitative Genetics Lab were purchased from the MSU Animal Initiative Equipment Fund.

References

Aitkin, M., and I. Aitkin. 1994. *Efficient computation of maximum likelihood estimates in mixture distributions, with reference to overdispersion and variance component models*. Proceedings XVIIth International Biometrics Conference, 8-12 August 1994, 123-138.

Antoniak, C. E. 1974. *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*. Annals of Statistics, 2: 1152-1174.

Brooks, S. P, P. Dellaportas, and G. O. Roberts. 1997. *A total variation method for diagnosing convergence of MCMC algorithms*. J. Comp. Graph. Stat. 6: 251-265.

Dellarpotas, P. 1995. *Random variate transformations in the Gibbs sampler: Issues of efficiency and convergence*. Statistics and Computing **5**: 133-140.

Dempster, A. P., Laird, N.M., and, D. B. Rubin. 1977. *Maximum likelihood from incomplete data via the EM algorithm*. J. Roy. Statist. Soc. B 39:1-38.

Diaconis, P. 1988. *Group representations in Probability and Statistics*. vol. 11 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics. pp. 91.

Diebolt, J., and C. P. Robert. 1994. *Estimation of finite mixture distributions through Bayesian Analysis*. J. Roy. Statist. Soc. B 56(2):363-375.

Escobar, M. D., and West, M. 1995. *Bayesian density estimation and inference using mixtures*. J. Amer. Stat. Assoc. 90: 577-587.

Everitt, B. S. and Hand, D. J. 1981. *Finite Mixture Distributions*, London: Chapman and Hall.

Ferguson, T. S. 1973. *A Bayesian analysis of some nonparametric problems*. Annals of Statistics, 1:209-230.

Ferguson, T. S. 1983. *Bayesian density estimation by mixtures of normal*

distributions. Recent Advances in Statistics, H. Rizvi and J. Rustagi. Eds. New York: Academic Press, 287-302.

Gelman, A., and D. Rubin. 1992. *Inference from iterative simulation using multiple sequences*. Statistical Science **7**:547-511.

Geweke, J. 1992. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. In J. M. Bernado, A. F. M. Smith, A. P. Dawid, and J. O. Berger (eds.), Bayesian Statistics 4, Oxford University Press, New York, New York, pp 156-163.

Lois. T. A. 1982. *Finding the observed information when using the EM algorithm*. J. Roy. Statist. Soc. B 44:226-233.

McLachlan, G. J., and K. E. Basford. 1988. *Mixture models*. Marcel Dekker, New York, NY.

MacEachern, S. N., and Müller, P. 1998. *Estimating mixture of Dirichlet process models*. Communications in Statistics: Simulation and Computations, 23:727-741.

Neal, R. M. 1998. *Markov chain sampling methods for Dirichlet process mixture models*. Technical Report No. 9815. Dept of Statistics, Univ. of Toronto, Toronto, Canada.

Petrone, S., and Raftery, A.E. 1997. *A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability*. Statist. Prob. Letters, 36, 69-83.

Petrone, S., G. O. Roberts, and J. R. Rosenthal. 1998. *A note on convergence rates of Gibbs sampling for nonparametric mixtures*. Technical report. Dipartimento di Economia Politica e Metodi Quantitativi, Universita' degli Studi di Pavia.

Price, K., and R. Storn. 1997. *Differential evolution*. Dr. Dobb's Journal. 264: 18-24.

Rao, C. R. 1948. *The utilization of multiple measurements in problems of biological classification*. J. Roy. Statist. Soc. B 10:159-203.

Stephens, M. 1997. *Bayesian Methods for Mixtures of Normal Distributions*. PhD Dissertation. Department of Statistics, University of Oxford.

Titterington, D. M., Smith, A. F. M., and Makov, U.E. 1985. *Statistical Analysis of Finite Mixture Distributions*. Chichester, New York: Wiley.