

# DIRICHLET PROCESS MIXTURE MODEL WITH APPLICATIONS TO BIOINFORMATICS

KANTI V. MARDIA, JOHN T. KENT, ZHENGZHENG ZHANG AND CHARLES C. TAYLOR

ABSTRACT. Motivated by examples in protein bioinformatics, we study a mixture model of multivariate angular distributions. The distribution treated here (multivariate sine distribution) is a multivariate extension of the well-known von Mises distribution on the circle. The sine distribution has an intractable normalizing constant and here we propose to replace it in the concentrated case by a simple approximation. We study a Dirichlet process mixture (DPM) model of concentrated sine distributions and apply it to practical examples from protein bioinformatics. The various issues arising from the DPM model, e.g. the ‘label switching’ issue, are discussed.

## 1. BACKGROUND

**1.1. Dirichlet process.** A random distribution  $G$  on  $\Omega$  follows a *Dirichlet process* (Ferguson, 1973), denoted by  $DP(\alpha, G_0)$ , if for all natural numbers  $K$  and  $K$ -partitions  $\Omega = \bigcup_{k=1}^K B_k$ ,  $B_j \cap B_k = \emptyset$  for all  $j \neq k$ :

$$(G(B_1), G(B_2), \dots, G(B_K)) \sim \text{Dirichlet}(\alpha G_0(B_1), \alpha G_0(B_2), \alpha G_0(B_K)), \quad (1)$$

where  $\alpha$  is the scaling parameter that is a positive real number,  $G_0$  is the base distribution on an arbitrary space  $\Omega$  and  $\text{Dirichlet}(\beta_1, \beta_2, \dots, \beta_K)$  denotes the distribution on the  $(K-1)$ -dimensional simplex with density at  $(x_1, x_2, \dots, x_K)$  proportional to  $\sum_{k=1}^K x_k^{\beta_k-1}$ . For any measurable set  $B \subset \Omega$ ,  $E[G(B)] = G_0(B)$ .

**1.2. Stick-breaking construction.** Intuitively, a random draw from a Dirichlet process is a weighted sum of point masses. It will lead to a constructive definition of a Dirichlet process, called the stick-breaking construction (Sethuraman, 1994). The process is defined as follows:

$$\text{let } \beta_k \sim \text{Beta}(1, \alpha), \alpha \geq 0, k = 1, 2, \dots; \quad (2)$$

$$\text{let } \omega_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l); \quad (3)$$

$$\text{let } \psi_k^* \sim G_0 \text{ i.i.d.}, k = 1, 2, \dots; \quad (4)$$

$$\text{then } \sum_{k=1}^{\infty} \omega_k = 1, G = \sum_{k=1}^{\infty} \omega_k \delta_{\psi_k^*}, \quad (5)$$

where  $G_0$  is the base distribution on a arbitrary space  $\Omega$ ;  $\psi_k^*$  is a distinct value drawn from  $G_0$  and  $\delta_{\psi_k^*}$  is a point mass located at  $\psi_k^*$ . Clearly  $G$  is random, because the  $\psi_k^*$  are drawn independently from  $G_0$ , and the weights  $\{\omega_k\}$  are drawn from the stick breaking distribution, denoted by  $\text{Stick}(\alpha)$ , with its density of (3).

**1.3. Dirichlet process mixture model.** Suppose that a Dirichlet process mixture model applies to data  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$  which can be drawn from some unknown distribution. The data  $\mathbf{y}_i$ , which can be multivariate, are drawn from a mixture of distributions of the form  $F(\psi)$ , with a Dirichlet process over parameters  $\psi$  being  $G$ .  $G_0$  is a base distribution on a parameter space  $\Omega$ . The model is given as follows:

$$\begin{aligned} \mathbf{y}_i | \psi_i &\sim F(\psi_i); \\ \psi_i | G &\sim G; \\ G | \alpha, G_0 &\sim DP(\alpha, G_0). \end{aligned} \tag{6}$$

In Neal (2000), they integrate over  $G$  in model (6) to obtain a predictive distribution of the  $\psi_i$  in the following form (Blackwell & MacQueen, 1973):

$$\psi_i | \psi_1, \dots, \psi_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{k=1}^{i-1} \delta(\psi_k) + \frac{\alpha}{i-1+\alpha} G_0 \tag{7}$$

where  $\delta(\psi_k)$  is the distribution concentrated at the single point  $\psi_k$ . The weight associated with  $G_0$  is proportional to  $\alpha$ , while the empirical distribution has weight proportional to the number of observations  $n$ . Then,  $\alpha$  can be interpreted as the mass associated with the prior. If we fix  $n$  to a positive integer and allow  $\alpha$  to vary from 0 to infinity, then (7) can be interpreted as follows. As  $\alpha$  goes to 0, the prior  $G_0$  becomes non-informative and then the predictive distribution is dominated by the empirical distribution. On the other hand, as  $\alpha$  goes to infinity, the distribution is simply interpreted by the prior distribution. However, if  $n \gg \alpha$ , i.e.,  $n$  is much greater than  $\alpha$ , then the predictive distribution will be dominated by the empirical distribution.

Equivalently, consider a finite mixture model with  $K$  components. The mixing probabilities  $p_1, \dots, p_K$  are drawn from a Dirichlet distribution with concentration parameter  $\alpha/K$ , and the sum of  $p_1, \dots, p_K$  is 1. Let

$$z_k = \pi(i), \text{ where } i = 1, \dots, n, \quad k = 1, \dots, K,$$

be a labeling function, assigning the  $i$ -th observation  $\mathbf{y}_i$  to the  $k$ -th component. Then, the model is written as follows:

$$\begin{aligned} \mathbf{y}_i | \psi_{z_k} &\sim F(\psi_{z_k}); \\ z_k | p &\sim \text{Discrete}(p_1, \dots, p_K); \\ \psi_{z_k} &\sim G_0; \\ p_1, \dots, p_K &\sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K). \end{aligned} \tag{8}$$

By integrating over the mixing proportions,  $p_1, \dots, p_K$ , we then have the prior for  $z_k$  as follows:

$$p(z_k = \pi(i) | \pi(\setminus i), \alpha, K) = \frac{n_k + \alpha/K}{i - 1 + \alpha} \quad (9)$$

where  $z_k = 1, \dots, K$  are component labels;  $\pi(\setminus i) = \pi(1), \dots, \pi(i-1)$  are all previously observed labels;  $n_k$  is a number of observations having a group label  $z_k$ .

As  $K$  goes to infinity, the prior probability of assigning  $\mathbf{y}_i$  to one of previously observed components,  $\pi(\setminus i)$ , is

$$p(z_k = \pi(i) | z_k = \pi(\setminus i), \alpha) = \frac{n_k}{i - 1 + \alpha}. \quad (10)$$

The prior probability of assigning  $\mathbf{y}_i$  to any new component, as  $K$  goes to infinity, is

$$p(z_k = \pi(i) | z_k \neq \pi(\setminus i), \alpha) = \frac{\alpha}{i - 1 + \alpha}. \quad (11)$$

In other words, the probability of assigning  $\mathbf{y}_i$  to an already seen component is proportional to the number of observations in that components, while the probability of creating a new component is proportional to the concentration parameter  $\alpha$ . Further, the limit of the finite mixture model (8) becomes the Dirichlet process mixture model (6), since the conditional probabilities for  $z_k$  in (10) and (11) correspond to the conditional probabilities for  $\psi_k$  in (7).

Recall that the prior conditional for  $\psi_i | \psi_{\setminus i}$  has the form of (7), where  $\psi_{\setminus i} = \psi_1, \dots, \psi_{i-1}$ . Given  $\mathbf{y}_i$ , the posterior conditional for  $\psi_i | \psi_{\setminus i}$  was written in Neal (2000) as follows:

$$\psi_i | \psi_{\setminus i}, \mathbf{y}_i \sim \sum_{k \neq i} q_{i,k} \delta(\psi_k) + r_i H_i \quad (12)$$

where  $H_i$  is the posterior distribution for  $\psi$  based on the prior  $G_0$  and the likelihood  $F(\mathbf{y}_i, \psi)$  on the single observation  $\mathbf{y}_i$ . The values of  $q_{i,k}$  and of  $r_i$  are defined as

$$q_{i,k} = bF(\mathbf{y}_i, \psi_k) \quad (13)$$

$$r_i = b\alpha \int F(\mathbf{y}_i, \psi) dG_0(\psi) \quad (14)$$

where  $b$  is such that  $\sum_{k \neq i} q_{i,k} + r_i = 1$ . When  $G_0$  is a conjugate prior, the Gibbs sampler method is generally feasible to compute the integral defined  $r_i$  and sample from  $H_i$ .

The posterior conditional probability of assigning  $\mathbf{y}_i$  to a previously observed component  $z_k = \pi(\setminus i) = \pi(1), \dots, \pi(i-1)$  is

$$p(z_k = \pi(i) | z_k = \pi(\setminus i), \mathbf{y}_i, \psi_{\pi(i)}, \alpha) = b \frac{n_k}{i - 1 + \alpha} F(\mathbf{y}_i, \psi_k), \quad (15)$$

while observing a new component  $z_k$  is

$$p(z_k = \pi(i) | z_k \neq \pi(\setminus i), \mathbf{y}_i, \psi_{\pi(i)}, \alpha) = \frac{\alpha}{i - 1 + \alpha} \int F(\mathbf{y}_i, \psi) dG_0(\psi). \quad (16)$$

The number of clusters  $K$  will be bounded by the number of observations  $n$ . Given  $\{\psi_{z_k} : k = 1, \dots, K\}$ ,  $\{\mathbf{y}_i : i = 1, \dots, n\}$ , the log-likelihood function of the mixture model

has the form:

$$\log g(\mathbf{y}; \boldsymbol{\psi}) = \sum_{i=1}^n \log F(\mathbf{y}_i; \boldsymbol{\psi}_{z_k=\pi(i)}), \quad (17)$$

where  $F$  is any given density.

Teh (2010) noticed that for  $i \geq 1$ , the observation  $\mathbf{y}_i$  assign to a new group with probability  $\frac{\alpha}{\alpha+i-1}$  independent of the previously observed components. The number of clusters  $K$  has mean and variance:

$$\begin{aligned} E[K|n] &= \sum_{i=1}^n \frac{\alpha}{\alpha+i-1} = \alpha(\phi(\alpha+n) - \phi(\alpha)) \\ &\simeq \alpha \log\left(1 + \frac{n}{\alpha}\right) \quad \text{for } n, \alpha \gg 0, \end{aligned} \quad (18)$$

$$\begin{aligned} V[K|n] &= \alpha(\phi(\alpha+n) - \phi(\alpha)) + \alpha^2(\phi'(\alpha+n) - \phi'(\alpha)) \\ &\simeq \alpha \log\left(1 + \frac{n}{\alpha}\right) \quad \text{for } n > \alpha \gg 0, \end{aligned} \quad (19)$$

where  $\phi(\cdot)$  is the digamma function. The number of clusters  $K$  grows logarithmically in the number of observations  $n$ . However, as in (19), the variance of  $K$  given  $n$  will also grow logarithmically if  $\alpha$  increases. In the context of the Dirichlet process mixture model, the model has the rich-gets-richer phenomenon. The number of clusters  $K$  has to be smaller than the number of observations  $n$ .  $\alpha$  will then control the number of clusters (see (18)) if  $n$  is given. The larger  $\alpha$  the larger number of clusters a priori.

Appendix A gives a tutorial example to understand Dirichlet process mixture models.

## 2. A DIRICHLET PROCESS MIXTURE MODEL OF MULTIVARIATE CONCENTRATED SINE DISTRIBUTIONS

**2.1. The posterior distributions.** If  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^T$  has a  $p$ -variate sine distribution, denoted by  $\text{VM}_s^p(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda})$ , its probability density function is given by

$$f_s(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = C_p^{-1}(\boldsymbol{\kappa}, \boldsymbol{\Lambda}) \exp\left\{\boldsymbol{\kappa}^T c(\boldsymbol{\theta}, \boldsymbol{\mu}) - \frac{1}{2} s(\boldsymbol{\theta}, \boldsymbol{\mu})^T \boldsymbol{\Lambda} s(\boldsymbol{\theta}, \boldsymbol{\mu})\right\}, \quad (20)$$

where  $-\pi < \theta_j \leq \pi$ ,  $-\pi < \mu_j \leq \pi$ ,  $\kappa_j \geq 0$ ,  $-\infty < \lambda_{jl} < \infty$ ,

$$\begin{aligned} c(\boldsymbol{\theta}, \boldsymbol{\mu}) &= (\cos(\theta_1 - \mu_1), \cos(\theta_2 - \mu_2), \dots, \cos(\theta_p - \mu_p)), \\ s(\boldsymbol{\theta}, \boldsymbol{\mu}) &= (\sin(\theta_1 - \mu_1), \sin(\theta_2 - \mu_2), \dots, \sin(\theta_p - \mu_p)), \\ \boldsymbol{\mu}^T &= (\mu_1, \mu_2, \dots, \mu_p), \quad \boldsymbol{\kappa}^T = (\kappa_1, \kappa_2, \dots, \kappa_p), \\ (\boldsymbol{\Lambda})_{jl} &= \lambda_{jl} = \lambda_{lj}, \quad j \neq l, \quad (\boldsymbol{\Lambda})_{jj} = \lambda_{jj} = 0, \quad j, l = 1, \dots, p \end{aligned}$$

and  $C_p^{-1}(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$  is a normalizing constant.

If the concentration parameters are sufficiently large, the normalizing constant,  $C_p$ , can be approximated by

$$E = D \exp\left\{\sum_j \kappa_j\right\} \quad (21)$$

where

$$D = (2\pi)^{p/2} |\Sigma|^{\frac{1}{2}}, \quad (22)$$

$$(\Sigma^{-1})_{jl} = (\Lambda)_{jl} = \lambda_{jl} = \lambda_{lj}, \quad j \neq l, \quad (\Sigma^{-1})_{jj} = \kappa_j, \quad j, l = 1, \dots, p. \quad (23)$$

Then, (20) is approximated by

$$f^*(\theta; \mu, \Sigma^{-1}) = D^{-1} \exp\left\{-\frac{1}{2}[\kappa^T(2 - 2\mathbf{c}(\theta, \mu)) + \mathbf{s}(\theta, \mu)^T \Lambda \mathbf{s}(\theta, \mu)]\right\} \quad (24)$$

We call this the *concentrated multivariate sine density*. It preserves several key properties of the multivariate sine density, such as modality and periodicity. We now show that the MLE can be then computed easily under the assumption that  $\Sigma$  is positive definite. Note that Mardia & Voss (2011) has shown the multivariate sine density is unimodal if  $\Sigma$  is positive definite.

Consider a  $p$ -variate concentrated sine distribution,  $\text{VM}_p^s(\mu, \Sigma^{-1})$ , and its pdf is written in (24). Let  $\theta = (\theta_1, \dots, \theta_n)$  be a random sample from this sine distribution.

Suppose the mean  $\mu$  is known, and we would like to make inference about a covariance matrix  $\Sigma$ . The prior of  $(\Sigma)$  has an inverse Wishart distribution,  $W^{-1}(\Phi, m)$ , and its pdf is

$$p(\Sigma) = \frac{|\Phi|^{m/2} |\Sigma|^{-(m+p+1)/2} \exp\{-\text{tr}(\Phi \Sigma^{-1})/2\}}{2^{mp/2} \Gamma_p(m/2)} \quad (25)$$

where  $\Phi$  is a  $p \times p$  positive definite matrix, and  $m$  is the degrees of freedom. As the observations  $\theta = (\theta_1, \dots, \theta_n)$  are drawn independently, the posterior density function is of the form

$$\begin{aligned} p(\Sigma | \theta, \mu) &\propto p(\Sigma) l(\Sigma | \theta, \mu) \\ &= p(\Sigma) \prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\theta_i - \mu)^T \Sigma^{-1} (\theta_i - \mu)\right\} \\ &= p(\Sigma) \frac{1}{(2\pi)^{np/2} |\Sigma|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (\theta_i - \mu)^T \Sigma^{-1} (\theta_i - \mu)\right\}. \end{aligned} \quad (26)$$

The sum in (26) can be written as

$$\begin{aligned} &\sum_{i=1}^n \sum_{a=1}^p \sum_{b=1}^p (\theta_{ai} - \mu_a) (\Sigma^{-1})_{ab} (\theta_{bi} - \mu_b) \\ &= \sum_{a=1}^p \sum_{b=1}^p (\Sigma^{-1})_{ab} \sum_{i=1}^n (\theta_{ai} - \mu_a) (\theta_{bi} - \mu_b) \\ &\cong \sum_{a=1}^p \sum_{b=1}^p (\Sigma^{-1})_{ab} \sum_{i=1}^n \sin(\theta_{ai} - \mu_a) \sin(\theta_{bi} - \mu_b) \\ &= \sum_{a=1}^p \sum_{b=1}^p (\Sigma^{-1})_{ab} S(\mu)_{ab} = \text{tr}(\Sigma^{-1} S(\mu)), \end{aligned} \quad (27)$$

where  $S(\boldsymbol{\mu}) = \sin(\boldsymbol{\theta} - \boldsymbol{\mu})^T \sin(\boldsymbol{\theta} - \boldsymbol{\mu})$  is  $n$  times the (circular) sample covariance matrix. Substituting (27) and (25) into (26), we have

$$\begin{aligned} p(\boldsymbol{\Sigma}|\boldsymbol{\theta}, \boldsymbol{\mu}) &\propto \frac{|\boldsymbol{\Phi}|^{m/2} |\boldsymbol{\Sigma}|^{-(m+p+1)/2} \exp\{-\text{tr}(\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-1})/2\}}{2^{mp/2} \Gamma_p(m/2)} \\ &\times \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\{-\text{tr}(\boldsymbol{\Sigma}^{-1}S(\boldsymbol{\mu}))/2\} \\ &\propto |\boldsymbol{\Sigma}|^{-(m+n+p+1)/2} \exp\{-\text{tr}((\boldsymbol{\Phi} + S(\boldsymbol{\mu}))\boldsymbol{\Sigma}^{-1})/2\}, \end{aligned} \quad (28)$$

where  $\Gamma_p(\cdot)$  is a multivariate gamma function. Then the posterior for  $\boldsymbol{\Sigma}$  is also an inverse Wishart distribution,  $W^{-1}(S(\boldsymbol{\mu}) + \boldsymbol{\Phi}, n + m)$ , and its p.d.f is

$$p(\boldsymbol{\Sigma}|\boldsymbol{\theta}, \boldsymbol{\mu}) = \frac{|\boldsymbol{\Phi} + S(\boldsymbol{\mu})|^{(m+n)/2} |\boldsymbol{\Sigma}|^{-(m+n+p+1)/2} \exp\{-\text{tr}((\boldsymbol{\Phi} + S(\boldsymbol{\mu}))\boldsymbol{\Sigma}^{-1})/2\}}{2^{(m+n)p/2} \Gamma_p((m+n)/2)}. \quad (29)$$

Hence the precision matrix  $\boldsymbol{\Sigma}^{-1}$  has a Wishart distribution, i.e.,

$$\boldsymbol{\Sigma}^{-1}|\boldsymbol{\theta}, \boldsymbol{\mu} \sim W((S(\boldsymbol{\mu}) + \boldsymbol{\Phi})^{-1}, n + m). \quad (30)$$

From (Mardia *et al.*, 1979, p85) we know that the mean of the posterior is

$$E(\boldsymbol{\Sigma}|\boldsymbol{\theta}, \boldsymbol{\mu}) = \frac{S(\boldsymbol{\mu}) + \boldsymbol{\Phi}}{n + m - p - 1}, \quad (31)$$

while the mode (O'Hagan & Forster, 2004, p406) is

$$\frac{S(\boldsymbol{\mu}) + \boldsymbol{\Phi}}{n + m + p + 1}. \quad (32)$$

Note that the inverse Wishart distribution is conjugate to the multivariate normal since the prior and posterior distributions are the same family.

Suppose the covariance matrix  $\boldsymbol{\Sigma}$  is known, we then make inference about mean. The prior  $p(\boldsymbol{\mu})$  has a multivariate concentrated sine distribution,

$$\boldsymbol{\mu} \sim \text{VM}_p^s(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^{-1}). \quad (33)$$

If the observations  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$  are drawn independently, then the posterior density function for  $\boldsymbol{\mu}$  is of the form

$$\begin{aligned} p(\boldsymbol{\mu}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) &\propto p(\boldsymbol{\mu})l(\boldsymbol{\mu}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \\ &\cong p(\boldsymbol{\mu}) \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (\boldsymbol{\theta}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu})\right\} \\ &= p(\boldsymbol{\mu}) \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1}S(\boldsymbol{\mu}))\right\} \end{aligned} \quad (34)$$

where

$$\begin{aligned}
S(\mu) &= \sum_{i=1}^n (\theta_i - \mu)(\theta_i - \mu)^T \\
&= \sum_{i=1}^n (\theta_i - \bar{\theta})(\theta_i - \bar{\theta})^T + n(\bar{\theta} - \mu)(\bar{\theta} - \mu)^T \\
&= S_0(\mu) + n(\bar{\theta} - \mu)(\bar{\theta} - \mu)^T
\end{aligned} \tag{35}$$

and  $\bar{\theta}$  is the circular sample mean. Substituting (35) into (34) leads to the following

$$\begin{aligned}
p(\mu|\theta, \Sigma) &\propto \exp\left\{-\frac{1}{2}(\mu - \mu_0)^T \Sigma_0^{-1}(\mu - \mu_0) - \frac{n}{2}(\mu - \bar{\theta})^T \Sigma^{-1}(\mu - \bar{\theta})\right\} \\
&= \exp\left\{-\frac{1}{2}(\mu - (\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\theta})(\Sigma_0^{-1} + n\Sigma^{-1})^{-1})^T (\Sigma_0^{-1} + n\Sigma^{-1})\right. \\
&\quad \left. (\mu - (\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\theta})(\Sigma_0^{-1} + n\Sigma^{-1})^{-1})\right\}
\end{aligned} \tag{36}$$

So the posterior of  $\mu$  is

$$\mu \sim N_p((\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\theta})(\Sigma_0^{-1} + n\Sigma^{-1})^{-1}, (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}). \tag{37}$$

**2.2. The Dirichlet process mixture model.** Lennox *et al.* (2009) used a Dirichlet process mixture (DPM) model together with bivariate sine distributions to study backbone dihedral angles. In this section, we extend this to fit a mixture of multivariate concentrated sine distributions. The proposed model is:

$$\begin{aligned}
\theta_i | \mu_{\pi(i)}, \Sigma_{\pi(i)}^{-1} &\sim p(\mu_{\pi(i)}, \Sigma_{\pi(i)}^{-1}), \quad i = 1, 2, \dots, n; \\
(\mu_{\pi(i)}, \Sigma_{\pi(i)}^{-1}) | G &\sim G; \\
G &\sim \text{DP}(\alpha, H_1 H_2),
\end{aligned} \tag{38}$$

where  $p(\mu_{\pi(i)}, \Sigma_{\pi(i)}^{-1})$  is a multivariate concentrated sine distribution. Its density has the form of (24). The variable  $G$  is a random realization from  $\text{DP}(\alpha, H_1 H_2)$ , a Dirichlet process with mass parameter  $\alpha$  and centering distribution  $H_1 H_2$ .  $H_1$  is a multivariate concentrated sine distribution for the mean  $\mu$ , whereas  $H_2$  is a multivariate Wishart distribution for the precision matrix  $\Sigma^{-1}$ .

The posterior predictive distribution can be obtained through Monte carlo Methods. We chose the Auxiliary Gibbs sampler of (Neal, 2000), providing an MCMC update of allocation of objects to clusters. The method was also used in Lennox *et al.* (2009). The Auxiliary Gibbs sampler requires an update scheme for the parameters,  $\mu$  and  $\Sigma^{-1}$ . As described previously in Section 2.1, the full conditional distribution of the mean  $\mu$  is (37), whereas the full conditional distribution of the precision matrix  $\Sigma^{-1}$  is given in (30).

The computational procedure is described as follows:

(1) Initialize the parameter values:

(a): Start from either one cluster or many clusters for all observations.

- (b): For the initial cluster, initialize the value of the parameters  $\mu$  and  $\Sigma^{-1}$  by sampling from the priors (25) and (33).
- (2) Sample from the posterior distribution by repeating the following:
  - (a): Given the mean and precision values, update the clustering configuration using one scan of the Auxiliary Gibbs sampler of Neal (2000).
  - (b): Given the clustering configuration and precision values, update the values of  $\mu$  for each cluster using the full conditional distribution using (37).
  - (c): Given the clustering configuration and mean values, update the precision matrix  $\Sigma^{-1}$  for each cluster using (30).

2.3. **Example 1.** In this example, we carry out a simulation study assessing the performance of the proposed mixture model. Consider a mixture of six four-variate sine distributions where the precision matrices,  $\Sigma_j^{-1}, j = 1, \dots, 6$ , of all the components are all equal, but the mean parameters,  $\mu_j, j = 1, \dots, 6$ , of all the components are different from each other; See Table 1. The means of group 4 and 5 are similar whereas for the rest four groups, the means are far apart; Also see Figure 1, in particular, the scatter plot for  $(\psi, \chi_2)$  where the four angles are denoted by  $(\phi, \psi, \chi_1, \chi_2)$ . A random sample of 2000 data points is then simulated from this known mixture. We fit a mixture of six four-variate concentrated distributions to the sample. The fitting procedure is implemented, as in 2.2, with some pre-defined priors as follows: the mean prior is drawn from a wrapped normal distribution such that  $\mu_0 = \mathbf{0}$  and  $\Sigma_0$  is a diagonal matrix with diagonal elements  $1/\pi^2$ . The small concentrations leads to a noninformative prior for  $\mu$ . For the Wishart prior, we set the scale matrix  $\Phi$  to be a diagonal matrix with diagonal elements 0.25 and used  $m = 2$  degrees of freedom. This also provides a noninformative prior for  $\Sigma$ . For this sample, the scaling parameter,  $\alpha$ , of the Dirichlet process is taken to be 1.

Convergence is diagnosed using several criteria described by Green & Richardson (2001). In particular, they measured equality of allocation using entropy, defined as,

$$-\sum_j (n_j/n) \log(n_j/n), \quad j = 1, \dots, K,$$

where  $K$  is number of components in DPM.

Figures 10 and 11 give various trace plots of the MCMC simulation. We take the first 1000 sweeps as burn-in. All the criteria indicate rapid convergence after this burn-in period.

After running the program, we take the parameter estimates from the last sweep as an example. To see the performance of this DPM model, 2000 data points are simulated from the fitted model. Figure 1 gives a scatter plot of the sample on the top, and the simulated data on the bottom. It can be seen from the figure that the data simulated from the DPM reveals the patterns as shown in the sample, i.e., the model has a good fit to the sample. In Figure 2, the first three histograms give empirical distributions of the number of data points in the three largest groups at all sweeps after the burn-in period. The largest group is likely to have around 776 data points,

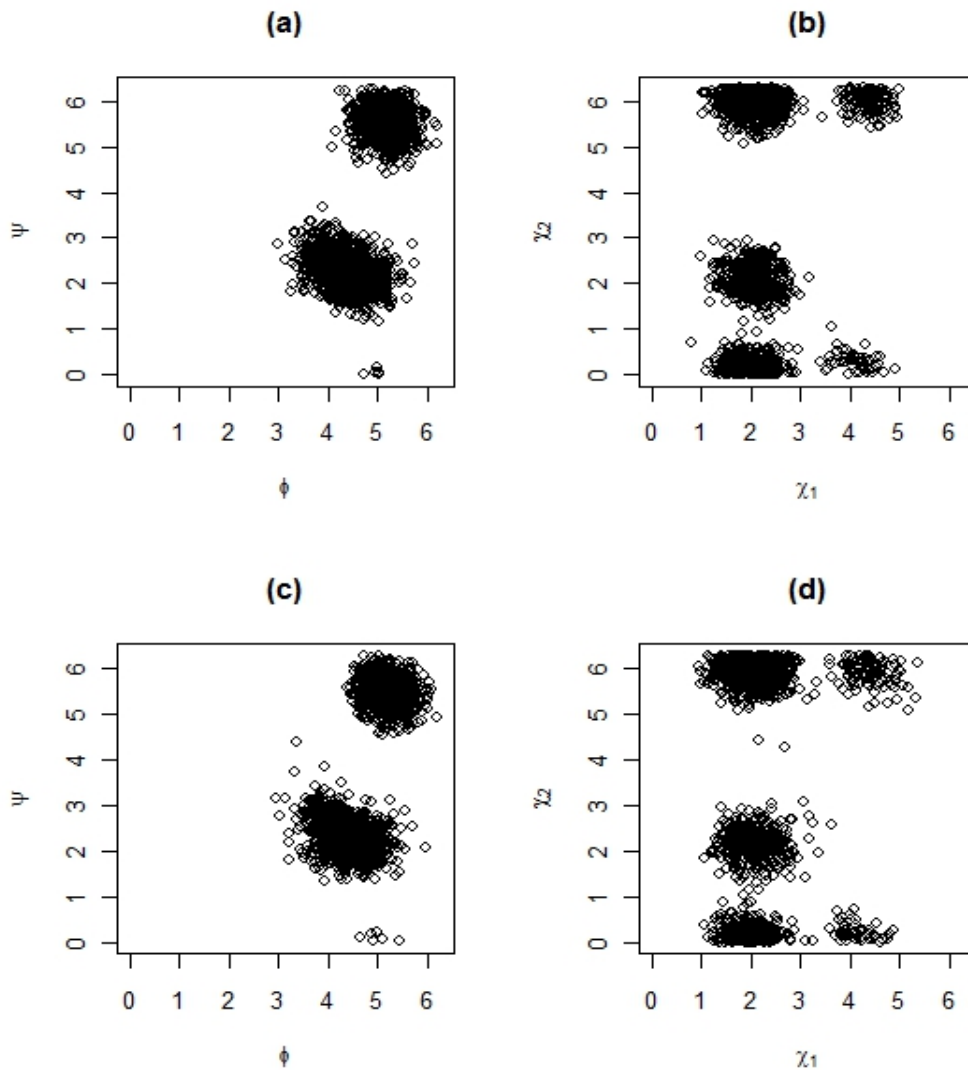


FIGURE 1. The random sample of 2000 data points is drawn on the left panel. Data is simulated using the parameters in the fitted model as shown on the right.

while 639 data points are appearing in the second largest. The last histogram gives the distribution of the number of groups after removing the burn-in period. The DPM with 12 components occurs with a chance around 0.1735, whereas the DPM with 11 components are round 0.1687. We then take all estimates, of which the mixture models have 12 components, to make a summary statistic. The parameter estimates are given in Table 2(a), with the standard deviations in Table 2(b). The smallest six components

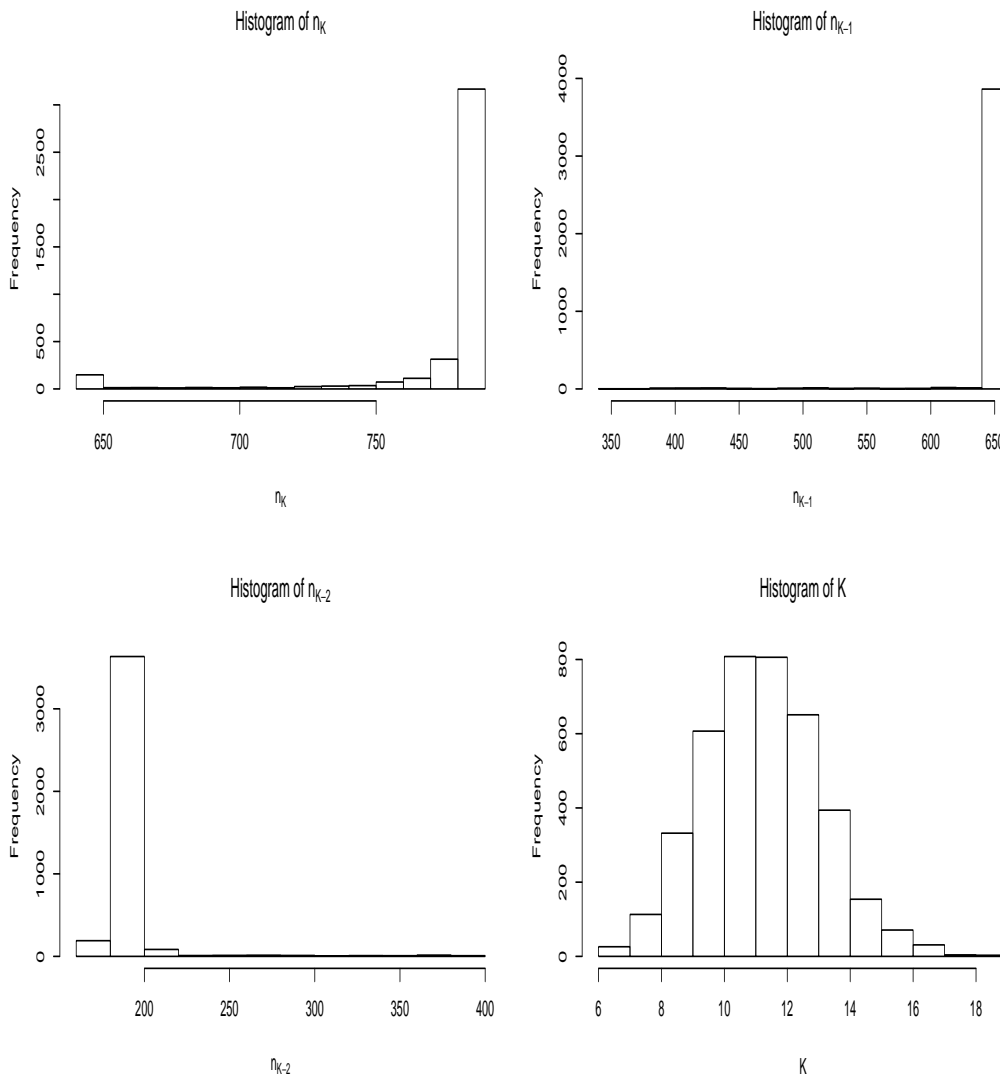


FIGURE 2. The first graphs are histograms of observations in the first three largest groups over all sweeps. The last one gives a histogram of the number of groups over all sweeps.

of the DPM model count for less than 4 percent of the total, so we exclude them in Table(s) 2.

On comparison of Table 2(a) and Table 1, the parameter estimates obtained from the DPM are compared to the true parameters. The components of the DPM model do not have one to one correspondence to the true components. The fifth component is possibly matching up with the sixth true component, and the parameter estimates have little difference from the true parameter of the component. The fourth component of

TABLE 1. the true parameters of the mixture model.

	%	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$	$\kappa_1$	$\kappa_2$	$\kappa_3$	$\kappa_4$	$\lambda_{12}$	$\lambda_{13}$	$\lambda_{14}$	$\lambda_{23}$	$\lambda_{24}$	$\lambda_{34}$
1	8.11	4.03	2.80	4.22	6.16	10	10	10	10	2	2	2	2	2	2
2	8.96	5.08	5.53	2.04	2.07	10	10	10	10	2	2	2	2	2	2
3	10.13	4.71	2.17	2.14	2.08	10	10	10	10	2	2	2	2	2	2
4	19.56	4.62	2.14	2.08	6.14	10	10	10	10	2	2	2	2	2	2
5	20.71	4.18	2.22	2.09	6.08	10	10	10	10	2	2	2	2	2	2
6	32.54	5.18	5.50	1.97	6.08	10	10	10	10	2	2	2	2	2	2

TABLE 2. (a) A summary statistic of the parameter estimates obtained from the all mixtures with 12 components. For the mean parameters, the sample circular mean directions are given. (b) Standard deviations are calculated for the estimates in (a). For the mean parameters, the sample circular standard deviations, defined in (Mardia & Jupp, 1999, pp19), are given. The 6 smallest groups are not shown in the tables.

(A)

	%	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$	$\kappa_1$	$\kappa_2$	$\kappa_3$	$\kappa_4$	$\lambda_{12}$	$\lambda_{13}$	$\lambda_{14}$	$\lambda_{23}$	$\lambda_{24}$	$\lambda_{34}$
1	2.65	4.58	2.25	2.22	2.13	13.3	15.3	12.4	13.4	1.9	1.5	1.1	2.1	0.7	0.7
2	5.44	4.52	2.29	2.34	2.01	12.8	15.0	12.2	12.4	2.3	1.5	1.3	1.7	1.0	1.2
3	8.18	4.11	2.74	4.17	6.22	9.9	12.1	11.3	10.7	1.8	2.0	1.1	1.9	2.3	3.6
4	9.17	4.99	5.50	2.04	2.12	11.5	10.8	10.5	12.0	2.7	2.9	3.0	2.8	1.9	1.8
5	32.00	5.19	5.47	1.97	6.08	9.9	9.8	10.3	11.6	1.9	1.4	1.4	1.6	1.9	1.3
6	38.87	4.41	2.19	2.08	6.11	7.1	10.3	10.5	10.4	1.8	1.4	1.8	0.8	1.8	1.8

(B)

	%	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$	$\kappa_1$	$\kappa_2$	$\kappa_3$	$\kappa_4$	$\lambda_{12}$	$\lambda_{13}$	$\lambda_{14}$	$\lambda_{23}$	$\lambda_{24}$	$\lambda_{34}$
1	2.65	0.25	0.22	0.25	0.37	5.8	6.1	5.4	6.2	3.6	3.1	3.0	3.3	3.5	3.4
2	5.44	0.25	0.25	0.61	0.68	4.0	4.2	3.7	3.9	2.4	2.0	2.1	2.2	2.4	2.3
3	8.18	0.26	0.49	0.64	0.65	1.4	1.7	1.4	1.5	1.0	0.9	1.1	0.9	1.1	1.2
4	9.17	0.15	0.50	0.16	0.28	1.5	1.4	1.3	1.5	0.9	0.9	1.0	1.0	0.9	0.9
5	32.00	0.12	0.31	0.02	0.02	0.7	0.6	0.6	0.7	0.4	0.4	0.4	0.4	0.4	0.5
6	38.87	0.12	0.31	0.02	0.02	0.6	0.6	0.6	0.7	0.3	0.3	0.4	0.4	0.4	0.4

the DPM model corresponds to the second true component, whereas the third component of the model matches up with the first true component. The mixing proportion for the sixth component of the model is approximately sum of the proportions for both the fourth and fifth true components. Note that  $\kappa_1$  of the sixth component of the model is 7.1 that is smaller than 10. This means that the data are widely spread over this direction. And also, the mean parameters of the fourth true component is closed to the ones of the fifth true component. Further, the third true component may include both

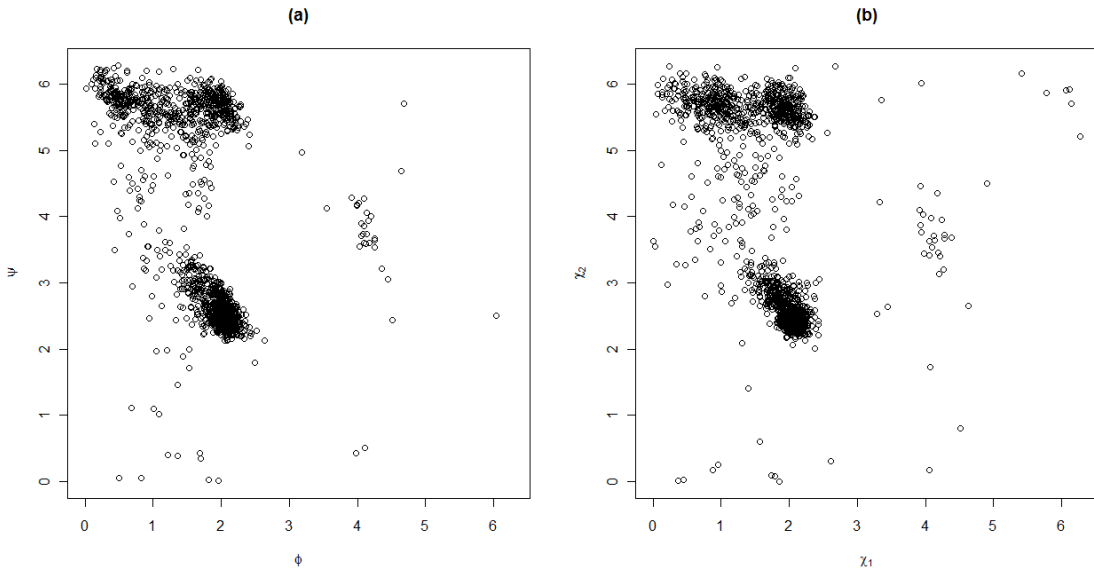


FIGURE 3. The random sample of 2000 data points is drawn on the left panel. Data is simulated using the parameters in the fitted model as shown on the right.

the first and second components of the model. In summary, the DPM model does not produce exact components as in Table 1, but it gives a good approximation to these true components overall.

**2.4. Example 2.** We take the backbone dihedral angles,  $\phi$  and  $\psi$ , of alanine extracted from Kinemage database, as example. The data consists of 8979 pairs of backbone dihedral angles,  $\phi$  and  $\psi$ . It has a dimension of  $8979 \times 2$ . We then take a random sample with a size of 2000 data points. Then, a Dirichlet process mixture model is fitted to the sample. The program is implemented, as described in 2.2, with some predefined priors as follows: the mean prior is drawn from a wrapped normal distribution with  $\mu_0 = \mathbf{0}$  and  $\Sigma_0$  a diagonal matrix with diagonal elements  $1/\pi^2$ . The small concentrations leads to a noninformative prior for  $\mu$ . For the Wishart prior, we set the scale matrix  $\Phi$  to be a diagonal matrix with diagonal elements 0.25 and used  $m = 2$  degrees of freedom. This also provides a noninformative prior for  $\Sigma$ . For this sample, the scaling parameter,  $\alpha$ , of the Dirichlet process is taken to be 0.5.

Figures 12 and 13 show trace plots for the amino acid Alanine. We take the first 1000 sweeps as burn-in. All the criteria indicate rapid convergence after this burn-in period.

Let us take the parameter estimates from the 5000-th sweep, and then 2000 data points are simulated from the fitted model using the simulation method. Figure 3 gives a scatter plot of the sample on the left, and the simulated data on the right. It can be

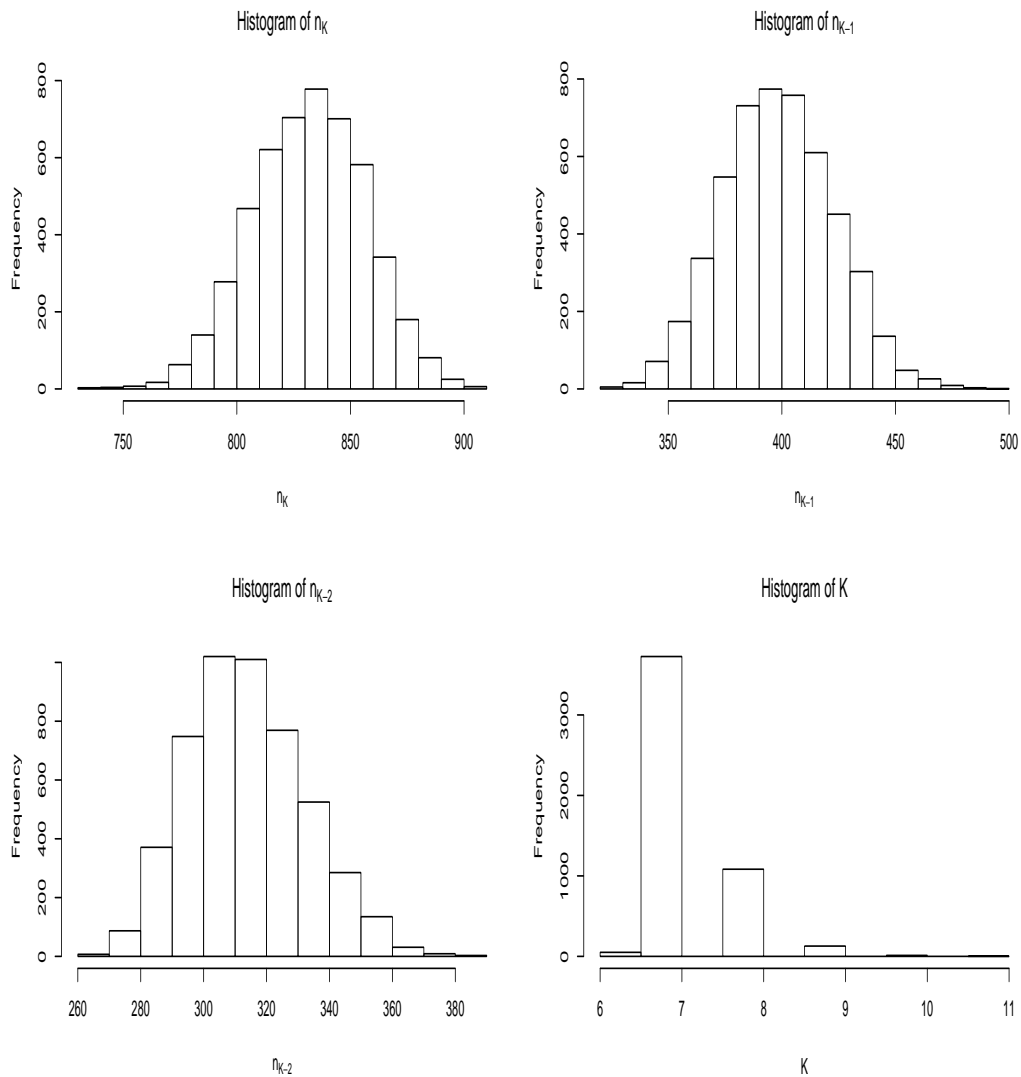


FIGURE 4. The first three graphs are histograms of observations in the first three largest groups over all sweeps. The last one gives a histogram of the number of groups over all sweeps.

seen from the figure that the data simulated from the DPM recovers the patterns in the sample, i.e., the model has a good fit to the sample. In Figure 4, the first three histograms give empirical distributions of the number of data points in the first three largest groups at all sweeps. The largest group is likely to have around 830 data points, while 400 data points are appearing in the second largest. The last histogram gives an distribution of the number of groups at the sweeps after removing the burn-in period. The DPM with 7 components occurs with a chance around 0.75, whereas the DPM

TABLE 3. A summary statistic of the parameter estimates obtained from the all mixtures with 7 components. For  $\mu_1$  and  $\mu_2$ , the sample circular mean directions are given.

Group	%	$\kappa_1$	$\kappa_2$	$\lambda$	$\mu_1$	$\mu_2$
1	0.43	3.02	0.80	0.03	-2.40	-2.26
2	1.08	47.35	12.32	11.56	-2.16	-2.50
3	7.89	5.34	0.94	-0.13	1.30	-1.77
4	13.32	28.50	31.93	11.22	1.87	-0.61
5	15.72	8.80	22.03	6.68	0.87	-0.59
6	19.98	29.91	23.07	16.04	1.91	2.72
7	41.59	160.69	124.43	30.49	2.05	2.43

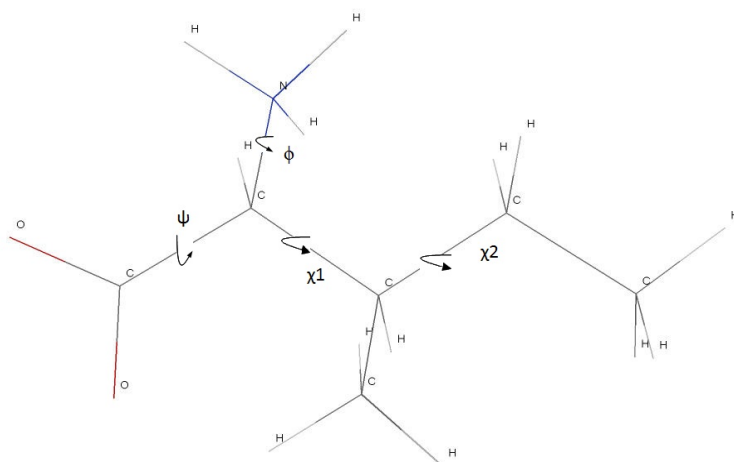
TABLE 4. Standard deviations are calculated for the estimates in Table 3. For  $\mu_1$  and  $\mu_2$ , the sample circular standard deviations, defined in (Mardia & Jupp, 1999, pp19), are given.

Group	$\kappa_1$	$\kappa_2$	$\lambda$	$\mu_1$	$\mu_2$
1	4.33	1.05	0.97	0.537	0.921
2	17.30	4.39	6.19	0.085	0.099
3	0.88	0.12	0.22	0.050	0.218
4	5.90	4.37	3.34	0.245	0.017
5	3.67	3.14	1.44	0.234	0.089
6	2.91	2.08	1.95	0.045	0.089
7	11.19	11.77	6.33	0.003	0.005

with 8 components are round 0.22. We then take all estimates, of which the mixture models have 7 components, to make a summary statistic. The mean values of the estimates are given in Table 3, with the standard deviations in Table 4. The first two largest groups have the smallest standard deviations around the mean directions.

**2.5. Example 3.** We consider here a data for isoleucine which is an amino acid. It has four key dihedral angles  $\phi, \psi, \chi_1$  and  $\chi_2$  as shown in Figure 5. The isoleucine (ILE) data, used in Harder *et al.* (2010), is of 21077 observations with 4 angles; a random sample of 3000 data points are drawn from the data. Hence, we fit a mixture of four-variate sine distributions.

We follow the computational procedures as described in 2.2, but some Dirichlet process priors need to be defined. Our mixture model, written as (38), was used with the mean prior from a wrapped normal distribution of  $\mu_0 = \mathbf{0}$  and  $\Sigma_0$  that was a diagonal matrix with diagonal elements  $1/\pi^2$ . The small concentrations leads to a noninformative prior for  $\mu$ . For the Wishart prior, we set the scale matrix  $\Phi$  to be a diagonal matrix with diagonal elements 0.25 and used  $m = 4$  degree of freedom. This also provides a



Jmol

FIGURE 5. A 3-D wireframe of isoleucine is created using Jmol. The four dihedral angles,  $\phi$ ,  $\psi$ ,  $\chi_1$  and  $\chi_2$  are labeled, and each dihedral is defined by the successive four atoms on both the backbone and the side-chain.

noninformative prior for  $\Sigma$ . The scaling parameter,  $\alpha$ , of the Dirichlet process is taken to be 1.

Figures 14 and 15 show trace plots of isoleucine. We take the first 1000 sweeps as burn-in. All the criteria indicate rapid convergence after this burn-in period. Table 5 gives a summary of estimates obtained at the last sweep. The last two columns give values of  $\sqrt{|\Sigma^{-1}|}$  and  $\sqrt{\kappa_1 \kappa_2 \kappa_3 \kappa_4}$ , respectively, for each group. The largest group consists of 20.77% of the data, and the second largest group counts for 18.30%. The 12-th group has mean values,  $\mu = (1.34, 4.66, 4.92, 2.09)$ , and concentration parameters,  $\kappa_1 = 4.74$ ,  $\kappa_2 = 0.91$ ,  $\kappa_3 = 0.75$ ,  $\kappa_4 = 0.91$ . The estimates for  $\kappa_3$ ,  $\kappa_4$  are diffuse and we did not observe this pattern from the sample in Figure 6. Thus, these estimates are poor because the concentrated sine distribution is inadequate to model the data with relatively low concentrations. In a summary, the model has a good fit in general, but not in great details. Small groups at the right and left bottom are not found in the scatter plot of simulations from the DPM, due to the richer-get-rich phenomenon (Teh, 2010).

In Figure 9, the first three histograms give empirical distributions of the number of data points in the first three largest groups at all sweeps. The largest group is likely to have around 580 data points, while around 480 data points are appearing in the second largest group. The last one gives a histogram of the number of groups at the sweeps after removing the burn-in period. The DPM with 19 components occurs with a chance around 0.189, whereas the DPM with 18 components are round 0.188. The

TABLE 5. A summary of the estimates of concentration parameters obtained from the DPM. The groups are sorted in increasing size of the mixing proportion.

group	%	$n_i$	$\kappa_1$	$\kappa_2$	$\kappa_3$	$\kappa_4$	$\lambda_{12}$	$\lambda_{13}$	$\lambda_{14}$	$\lambda_{23}$	$\lambda_{24}$	$\lambda_{34}$	$\sqrt{ \Sigma^{-1} }$	$\sqrt{\kappa_1 \kappa_2 \kappa_3 \kappa_4}$
1	0.03	1	6.38	10.37	8.48	44.88	2.77	-1.15	-4.63	-7.83	3.76	-7.19	86	159
2	0.03	1	16.38	31.68	2.20	36.40	5.88	-0.33	-1.57	1.08	-7.24	2.34	181	204
3	0.43	13	35.68	15.29	37.56	20.64	2.51	-13.27	7.24	3.30	7.66	-4.12	511	650
4	1.60	48	60.40	45.42	55.10	93.30	12.12	19.44	4.98	20.26	4.42	-11.54	3115	3755
5	1.73	52	69.91	42.18	74.84	27.74	12.07	12.53	3.35	4.99	5.21	2.94	2340	2474
6	1.87	56	77.57	54.35	71.61	60.38	23.63	22.15	-9.68	-3.72	8.96	-19.76	3490	4269
7	2.43	73	52.00	48.05	72.35	54.58	11.89	3.46	-7.67	-23.00	5.92	-4.45	2707	3141
8	3.53	106	3.47	73.70	102.97	50.22	-1.97	5.02	0.47	15.75	4.43	-5.21	1062	1150
9	3.80	114	15.23	16.61	97.03	55.60	5.31	-13.95	4.05	5.84	-6.04	-5.71	933	1168
10	3.83	115	5.31	68.93	97.47	67.63	6.63	-9.34	1.56	-42.92	-2.32	6.66	1175	1553
11	5.73	172	16.36	73.06	78.63	62.23	-3.59	16.29	3.26	-0.45	-1.56	-21.98	1945	2418
12	5.80	174	4.74	0.91	0.75	0.91	0.59	0.03	0.10	0.05	0.02	-0.07	2	2
13	7.53	226	8.56	31.43	70.24	63.03	1.30	-2.98	1.66	-4.62	-6.29	4.19	1059	1092
14	8.93	268	10.63	34.58	92.76	85.85	3.75	3.65	-7.40	17.30	-10.39	-21.03	1503	1711
15	13.63	409	28.67	49.07	74.43	51.38	14.60	9.25	5.36	16.03	1.02	1.29	2026	2319
16	18.30	549	10.90	32.96	128.64	78.40	-3.98	5.25	8.97	22.76	6.49	0.64	1565	1904
17	20.77	623	200.13	203.12	213.90	93.23	62.96	28.30	5.32	40.50	-16.34	-7.13	26156	28472

probabilities of occurrence for the DPM with 17 and 20 components are 0.17 and 0.14 respectively.

We now take all estimates, of which the mixture models have 19 components, to make a summary statistic. The mean values of the estimates of  $\mu^{(j)} = (\mu_1, \dots, \mu_4)^{(j)}$ ,  $\kappa^{(j)} = (\kappa_1, \dots, \kappa_4)^{(j)}$  are given in Table 7, together with the standard deviations in Table 8 for each component  $j = 1, \dots, 19$ . Note that these parameter estimates are in the increasing order of weights. There are many large sample circular standard deviations for the mean parameters, some of which could result from ‘label switching’ (Richardson & Green, 1997). For example, the sample circular standard deviations for the mean parameters of  $\psi$  in the first two largest groups are 1.99 and 1.51, the largest ones in Table 8. This is because these two groups interchange over all sweeps, that is, at some sweeps the group with  $\hat{\mu}_2 = 2.37$  approximately is the largest, while the group with  $\hat{\mu}_2 = 5.28$  approximately is the largest at the other sweeps (see Figure 7). Therefore, some estimates with large sample circular standard deviations could result from this mixing phenomenon.

We could also order the groups at each sweep in terms of some linear combinations of the mean parameter estimates. First of all, let us consider the sweeps consisting of 19 components, and their mean parameter estimates  $\hat{\mu}^{(j)}, j = 1, \dots, 19$  are pooled together. The resulting matrix has a dimension  $14364 \times 4$ . We perform a circular principal component analysis on the resulting matrix,  $\mathbf{x}^{[3]}$ , using the following transformation:

$$x_{rj}^{[3]} = (\sin(\theta_{rj}), \cos(\theta_{rj})), \quad r = 1, \dots, n, \quad j = 1, \dots, p. \quad (39)$$

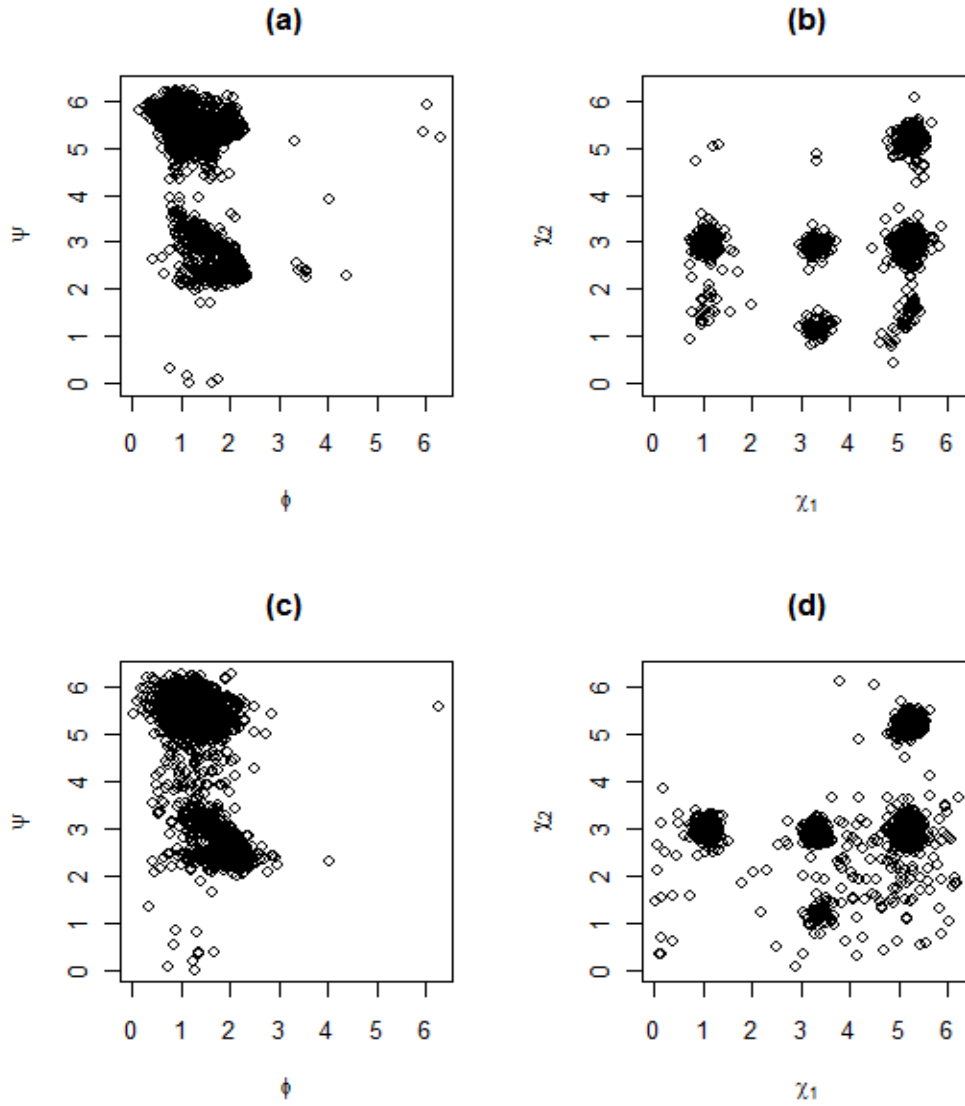


FIGURE 6. The random sample is drawn on the top panel: (a)  $\phi$  vs  $\psi$ , (b)  $\chi_1$  vs  $\chi_2$ ; 3000 data points are simulated using the parameters in the fitted model as shown on the bottom: (c)  $\phi$  vs  $\psi$ , (d)  $\chi_1$  vs  $\chi_2$ .

A spectral decomposition of the sample covariance matrix of  $\mathbf{x}^{[3]}$  yields principal components

$$\begin{aligned}
 y_1 &= 0.395x_1 - 0.411x_2 - 0.320x_3 - 0.092x_4 - 0.006x_5 + 0.750x_6 - 0.034x_7 + 0.029x_8, \\
 y_2 &= -0.098x_1 + 0.173x_2 - 0.534x_3 - 0.222x_4 - 0.097x_5 - 0.153x_6 - 0.683x_7 + 0.357x_8, \\
 y_3 &= 0.075x_1 + 0.176x_2 - 0.646x_3 - 0.326x_4 - 0.058x_5 - 0.222x_6 + 0.546x_7 - 0.299x_8, \\
 y_4 &= 0.031x_1 - 0.015x_2 + 0.103x_3 - 0.230x_4 - 0.019x_5 - 0.022x_6 + 0.443x_7 + 0.859x_8, \\
 y_5 &= 0.295x_1 + 0.039x_2 - 0.290x_3 + 0.823x_4 - 0.284x_5 - 0.162x_6 + 0.099x_7 + 0.184x_8, \\
 y_6 &= 0.447x_1 + 0.688x_2 + 0.262x_3 - 0.177x_4 - 0.410x_5 + 0.228x_6 - 0.060x_7 - 0.056x_8, \\
 y_7 &= 0.439x_1 - 0.526x_2 + 0.180x_3 - 0.273x_4 - 0.412x_5 - 0.482x_6 - 0.123x_7 - 0.078x_8, \\
 y_8 &= -0.591x_1 - 0.134x_2 - 0.013x_3 - 0.004x_4 - 0.754x_5 + 0.232x_6 + 0.096x_7 - 0.041x_8
 \end{aligned}$$

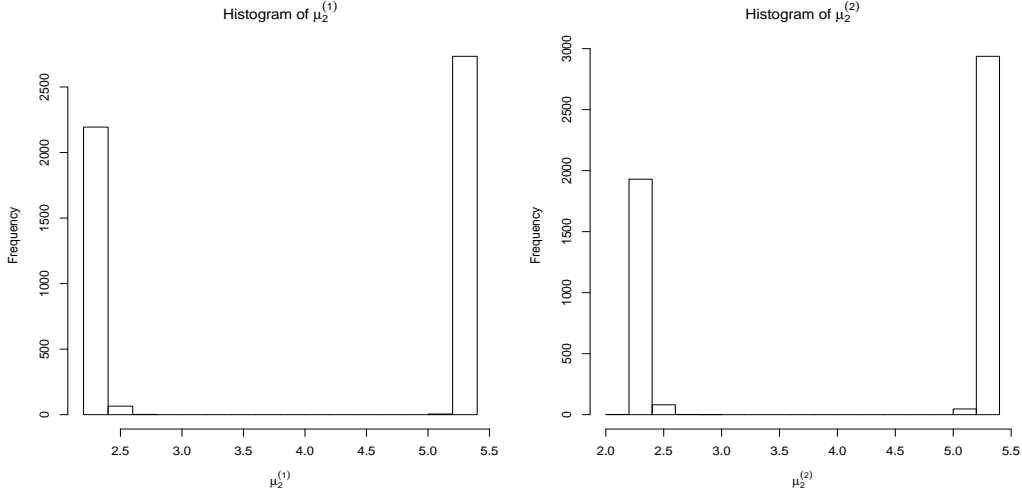


FIGURE 7. The left histogram shows the posterior distribution of  $\hat{\mu}_2$  in the largest group, whereas the right histogram gives the posterior distribution of  $\hat{\mu}_2$  in the second largest group.

with eigenvalues 1.0004, 0.5224, 0.3644, 0.2806, 0.1519, 0.0795, 0.0566 and 0.0191, respectively. The first principal component has the largest variance which contributes to 40.42 % of the total variance, and the last seven components count for 21.11%, 14.72%, 11.34%, 6.14%, 3.21%, 2.29% and 0.77%, respectively. The first four components are informative as they contribute to 87.59% of the total variance. Figure 8 gives a scatter plot on the first two principal components that likely separate out 10 clusters. At the same time, we observe two clusters in 6(c), and five clusters in 6(d), so there are 10 combinations all together. This coincides with what we observe from Figure 8.

We now order the groups by  $\text{atan2}(y_1, y_2)$  measured in  $[-\pi, \pi]$ . We can see from Figure 16 that the posterior distributions of  $\mu_1^{(j)}$  are all unimodal, but bimodality is observed from some posterior distributions of  $\mu_3^{(j)}$  in Figure 17. Therefore, this bimodality is due to ‘labeling switching’. However, we minimize the ‘labeling switching’ issue in this way on comparison to ordering by weights.

Further, we order by determinants of the precision matrices,  $|\hat{\Sigma}_j^{-1}|$ . However, we fail to reduce the ‘labeling switching’ issue using this ordering criterion.

The various values of  $\alpha = 1, 1.2, 1.5, 5$ , which are much smaller than the number of observations,  $n = 3000$ , are considered within a DPM model. The average numbers of components  $K$ , after the burn-in period, are 18.55, 18.40, 18.80 and 20.51, respectively. Slow growth of number of components  $K$  may be seen. As  $\alpha$  increases, the model is likely to introduce some groups with few observations. However, if the number of observations  $n \gg \alpha$ , the predictive distribution in (7) is dominated by the likelihood, and so is the number of components  $K$ .

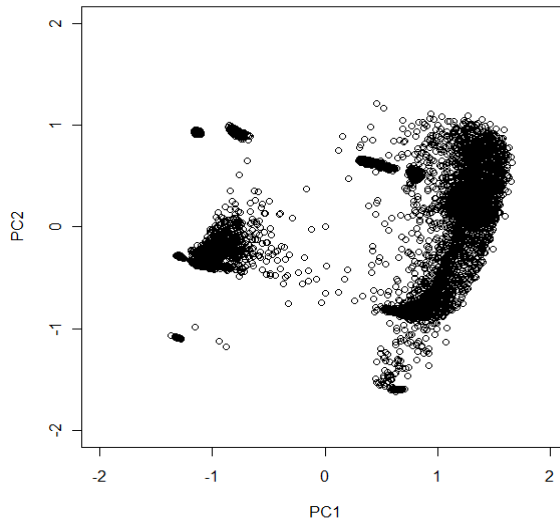


FIGURE 8. The scatter plot on the first two principal components.

### 3. DISCUSSION AND CONCLUSIONS

Mardia *et al.* (2008) have proposed a multivariate sine distribution which has various attractive properties. However, the use is somewhat hampered beyond the bivariate case as the normalizing constant is intractable. We introduce here concentrated multivariate sine distribution. The main idea is to approximate its normalizing constant when the sine distribution is concentrated. This approach simplifies calculation of the maximum likelihood estimates. We then consider mixtures of sine densities by replacing components by concentrated multivariate sine distribution. Indeed, this helps in computation when we are dealing with moderately large number of variables and components of mixtures.

It can be seen from Figure 3 that the simulations from the DPM model (on the right) recover the patterns observed from the sample (on the left). The DPM model also revealed a small group centred at around (4,4). Moreover, it can be seen from Figure 6 that most of the potential groups picked by human eye are also the groups for the DPM model. However, we also observed from the figure that some potential groups involving few observations was not captured by the DPM model. As the data sets are concentrated, the DPM models worked well for fitting a mixture of concentrated sine distributions.

In Example 2, we compared the parameter estimates for the EM model to the ones for the DPM model. We found that these parameter estimates did not have one to one correspondence in general. In Example 3, only the first three largest groups in both the models were matched up to each other. In general, we should not expect that a local

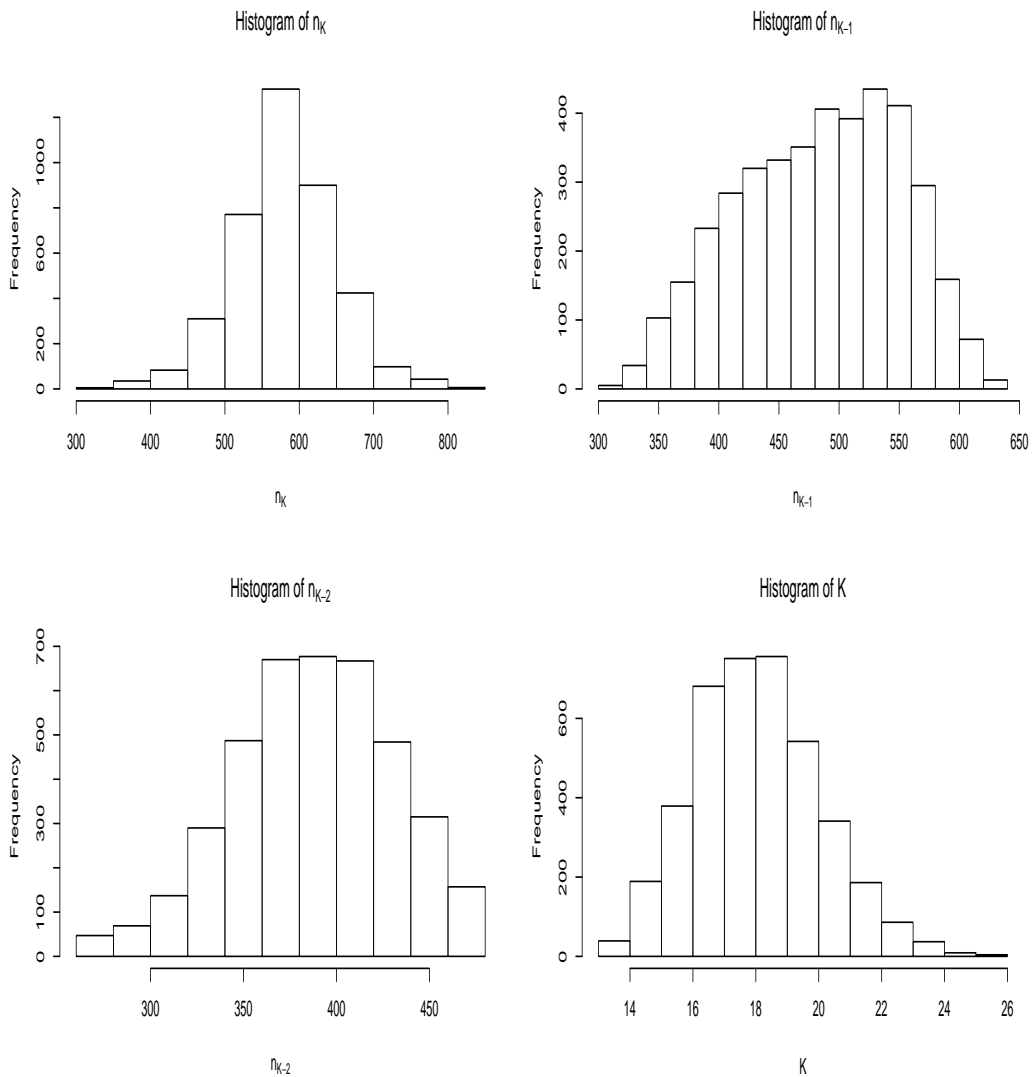


FIGURE 9. The first three histograms give empirical distributions of the number of observations in the first three largest groups over all sweeps. The last one gives a histogram of the number of groups over all sweeps.

solution from an EM mixture model coincides with a solution of the DPM model at a sweep.

Some potential groups with few observations may not be picked up as a group by the DP approach due to the rich-gets-richer phenomenon (Teh, 2010). Further, the DP approach does not need any initial values for parameters to be estimated, but the prior densities and the scaling parameter,  $\alpha$ , have to be given in advance. If these quantities

are badly chosen, this will lead to a slow convergence to the posterior distributions of the parameters. In particular, as in Example 3, we took a variety of values of  $\alpha$ . We found that different  $\alpha$  has a small impact on the posterior distribution of the number of components,  $K$ , if the sample size is large. Lastly, an assessment of convergence is difficult in high dimensions. An assessment of convergence simultaneously for all parameters is especially challenging, and sometimes we have to perform a long run.

More importantly, labeling and post-processing MCMC output become necessary when density estimates or other summaries of the posterior distribution of the parameters of each component are required (Richardson & Green, 1997). In our algorithm, we assign each new group an unique label. If a group becomes empty in a sweep, then this group label and relative parameter estimates will be removed, and the label will not be used again in the following sweeps. Although this preserve uniqueness of each label, there are too many of them. After that, we found that some groups with different group labels may be merged into one component. For all the three examples, the simplest and efficient way for labeling at each sweep is chosen to order on weights (or mixing proportions). However, as in Example 3, ‘label switching’ occurs when summarizing the posterior distribution of the parameters of each component. ‘Label switching’ may be minimized by choosing to order on some combination of mean, precision, and weight parameters as suggested in Richardson & Green (1997).

Our implementation of the DPM in R is much slower than the EM for mixture. If we have a random sample of 3000 data points as in Example 3, it takes 5 days to obtain 5000 sweeps of the parameter estimates. We prefer the EM for mixtures in terms of computing time. However, the speed may be improved by implementing this algorithm in C.

#### ACKNOWLEDGEMENTS

We would like to thank Peter Green for his useful comments on various issues arising from our Dirichlet process mixture model. This work is funded by a Dorothy Hodgkin postgraduate award co-sponsored between BBSRC and GlaxoSmithKline.

## REFERENCES

- BLACKWELL, D. & MACQUEEN, J. (1973). Ferguson distribution via pólya urn schemes. *Annals of Statistics*, **1**, 353–355.
- FERGUSON, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.
- GREEN, P. & RICHARDSON, S. (2001). Modelling heterogeneity with and without the dirichlet process. *The Sandinavian Journal of Statistics*, **28**, 355–375.
- HARDER, T., BOOMSMA, W., PALUSZEWSKI, M., FRELLSEN, J., ENJOHANSSON, K. & HAMELRYCK, T. (2010). Beyond rotamers: A generative, probabilistic model of side chains in proteins. *BMC Bioinformatics*, **11**, 306.
- LENNOX, K., DAHL, D., VANNUCCI, M. & TSAI, J. (2009). Density estimation for protein conformation angles using a bivariate von mises distribution and bayesian nonparametrics. *Journal of the American Statistical Association*, **104**, 586–596.
- MARDIA, K. & JUPP, P. (1999). *Directional Statistics*. WileyBlackwell.
- MARDIA, K. & VOSS, J. (2011). Properties of the multivariate sine distribution. *Preprint*.
- MARDIA, K., KENT, J. & BIBBY, J. (1979). *Multivariate Analysis*. Academic Press.
- MARDIA, K., HUGHES, G., TAYLOR, C. & SINGH, H. (2008). A multivariate von mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, **36**, 99–109.
- NEAL, R. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.
- O’HAGAN, A. & FORSTER, J.J. (2004). *Kendall’s Advanced Theory of Statistics: Bayesian Inference*. 2B (2 ed.). Arnold.
- RICHARDSON, S. & GREEN, P. (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.
- SETHURAMAN, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, **4**, 639–650.
- TEH, Y. (2010). Dirichlet processes. *Encyclopedia of Machine Learning*.

APPENDIX A. A MIXTURE OF NORMALS

Consider a mixture of normals with variances  $\sigma^2 = 1$ ,

$$x_i | \mu_{\pi(i)} \sim \mathbf{N}(\mu_{\pi(i)}, 1), \quad i = 1, 2, \dots, n, \quad (40)$$

where

$$\mu_{\pi(i)} | G \sim G; \quad (41)$$

$$G | \alpha, G_0 \sim \mathcal{DP}(\alpha, G_0); \quad (42)$$

$$G_0 = \mathbf{N}(\mu_0, 1). \quad (43)$$

Let  $\mu_k^* \in \{\mu_1^*, \dots, \mu_K^*\}$  be distinct values of  $\mu_{\pi(i)}$  for all  $i \in \{1, 2, \dots, n\}$ , and  $\pi(i) \in \{1, 2, \dots, K\}$  be a labeling function for each  $i$ . Given  $\mu_{\pi(1)}, \mu_{\pi(2)}, \dots, \mu_{\pi(n-1)}$ , the predictive distribution of  $\mu_{\pi(i)}$  can be written as follows:

$$\mu_{\pi(i)} = \begin{cases} \mu_k^* & \text{with prob. } \frac{\{i: \pi(i)=k\}}{n-1+\alpha}; \\ \mu, \mu \sim G_0, & \text{with prob. } \frac{\alpha}{n-1+\alpha}, \end{cases} \quad (44)$$

where  $|i: \pi(i) = k|$  is the number of times the value  $\mu_k^*$  occurs in  $\{\mu_1, \dots, \mu_{n-1}\}$ . More explicitly,

- $i = 1, \pi(i) = 1$  with prob. 1.
- $i = 2, \begin{cases} \text{either } \pi(i) = 2 \text{ with prob. } \frac{\alpha}{1+\alpha}; \\ \text{or } \pi(i) = 1 \text{ with prob. } \frac{1}{1+\alpha}. \end{cases}$
- $i = 3, \begin{array}{|c|c|} \hline \pi(1) = 1, \pi(2) = 1 & \pi(1) = 1, \pi(2) = 2 \\ \hline P(\pi(i) = 1) = \frac{2}{2+\alpha}; & P(\pi(i) = 1) = \frac{1}{2+\alpha}; \\ P(\pi(i) = 2) = \frac{\alpha}{2+\alpha}. & P(\pi(i) = 2) = \frac{1}{2+\alpha}; \\ & P(\pi(i) = 3) = \frac{\alpha}{2+\alpha}. \\ \hline \end{array}$
- ...

In a summary, the aim is to draw  $x_i \sim \mathbf{N}(\mu_{(i)}, 1)$ , where  $i = 1, 2, \dots, n$  (and  $\mu_{(i)}$  is not labeled), and to find  $k = \pi(i)$  the labeling function. The mechanism can be explained as follows:

- Draw  $\mu_k \sim \mathbf{N}(\mu_0, 1)$ ,  $k = 1, 2, \dots$
- Choose  $\pi(i) = k$  with probabilities given by the predictive distribution. Take  $i = 3$ , shown in the table above, as example. There are only two possibilities, either

$$\pi(1) = \pi(2) = 1 \quad \text{or} \quad \pi(1) = 1, \pi(2) = 2$$

to reach  $\pi(3)$  are displayed. For each way, we then have probabilities for which  $x_i$  draws from one of possible groups, namely,  $k = 1, 2$  or  $3$ . The expected number of clusters  $K$  depends on the number of observations  $n$  and the scaling parameter  $\alpha$  (See (18)),  $\alpha \geq 0$ . If  $n$  is fixed, then the larger  $\alpha$  will imply a larger number of clusters a priori.

- Then draw  $x_i \sim \mathbf{N}(\mu_{(i)}, 1)$ ,  $i = 1, 2, \dots, n$ .

TABLE 6. For  $i = 4$

$\pi(1) = 1, \pi(2) = 1,$ $\pi(3) = 1$	$\pi(1) = 1, \pi(2) = 1,$ $\pi(3) = 2$	$\pi(1) = 1, \pi(2) = 2,$ $\pi(3) = 1$	$\pi(1) = 1, \pi(2) = 2,$ $\pi(3) = 2$	$\pi(1) = 1, \pi(2) = 2,$ $\pi(3) = 3$
$P(\pi(i) = 1) = \frac{3}{3+\alpha};$ $P(\pi(i) = 2) = \frac{\alpha}{3+\alpha}.$	$P(\pi(i) = 1) = \frac{2}{3+\alpha};$ $P(\pi(i) = 2) = \frac{1}{3+\alpha};$ $P(\pi(i) = 3) = \frac{\alpha}{3+\alpha}.$	$P(\pi(i) = 1) = \frac{2}{3+\alpha};$ $P(\pi(i) = 2) = \frac{1}{3+\alpha};$ $P(\pi(i) = 3) = \frac{\alpha}{3+\alpha}.$	$P(\pi(i) = 1) = \frac{1}{3+\alpha};$ $P(\pi(i) = 2) = \frac{2}{3+\alpha};$ $P(\pi(i) = 3) = \frac{\alpha}{3+\alpha}.$	$P(\pi(i) = 1) = \frac{1}{3+\alpha};$ $P(\pi(i) = 2) = \frac{1}{3+\alpha};$ $P(\pi(i) = 3) = \frac{1}{3+\alpha};$ $P(\pi(i) = 4) = \frac{\alpha}{3+\alpha}.$

## APPENDIX B. FIGURES AND TABLES

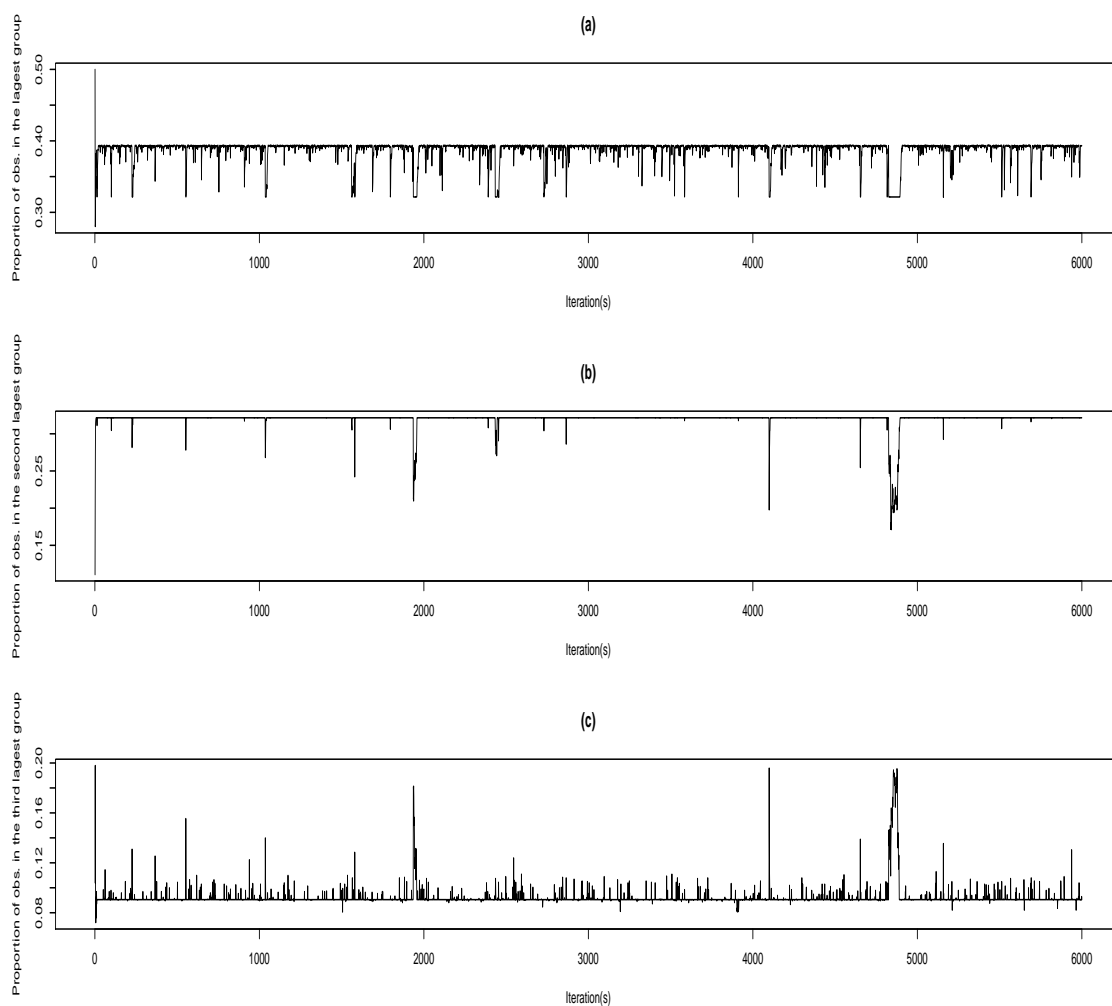


FIGURE 10. (a), (b), (c) plot proportion of observations in the first three largest groups over all iterations.

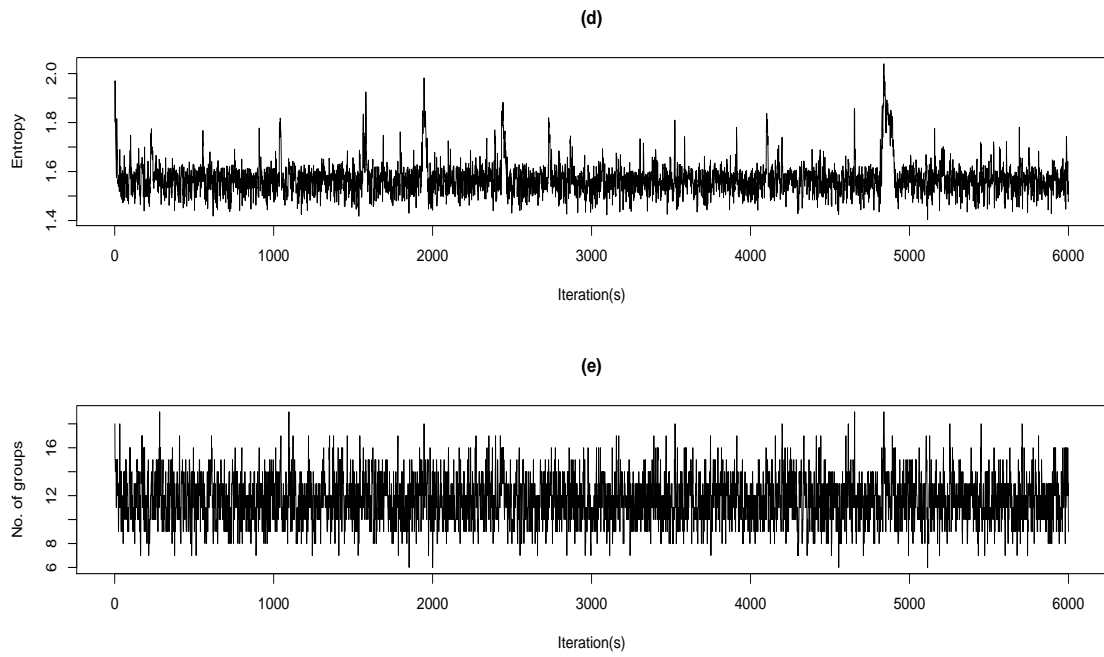


FIGURE 11. Entropy is calculated at each iteration shown in (d), whereas (e) shows number of groups at each iteration.

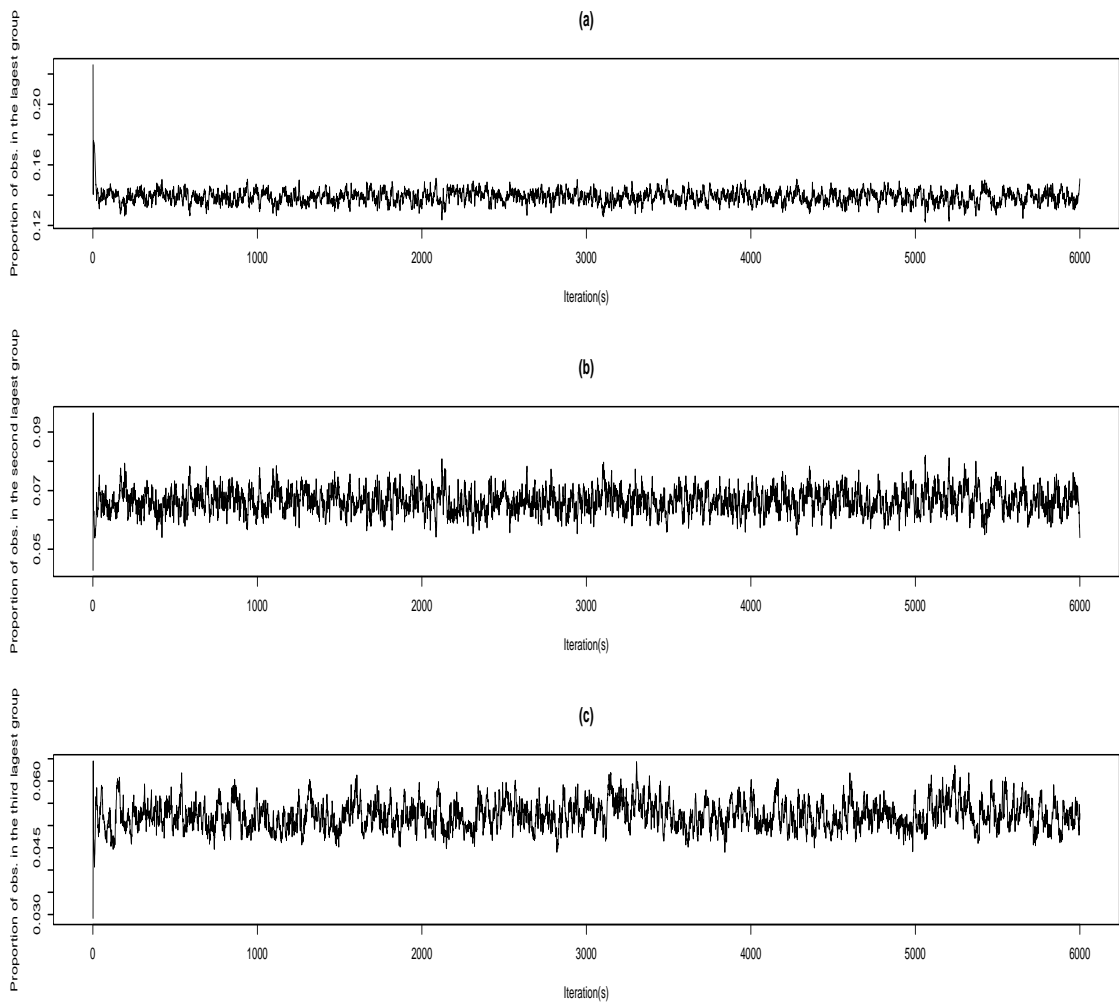


FIGURE 12. (a), (b), (c) plot proportion of observations in the first three largest groups over all iterations.

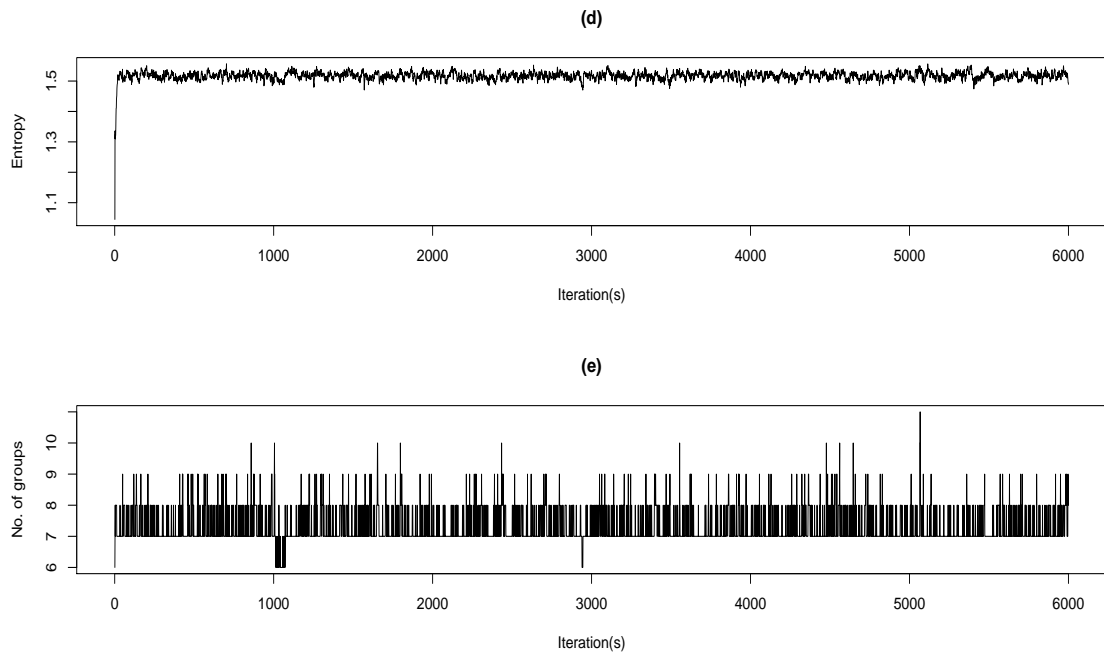


FIGURE 13. Entropy is calculated at each iteration shown in (d), whereas (e) shows number of groups at each iteration.

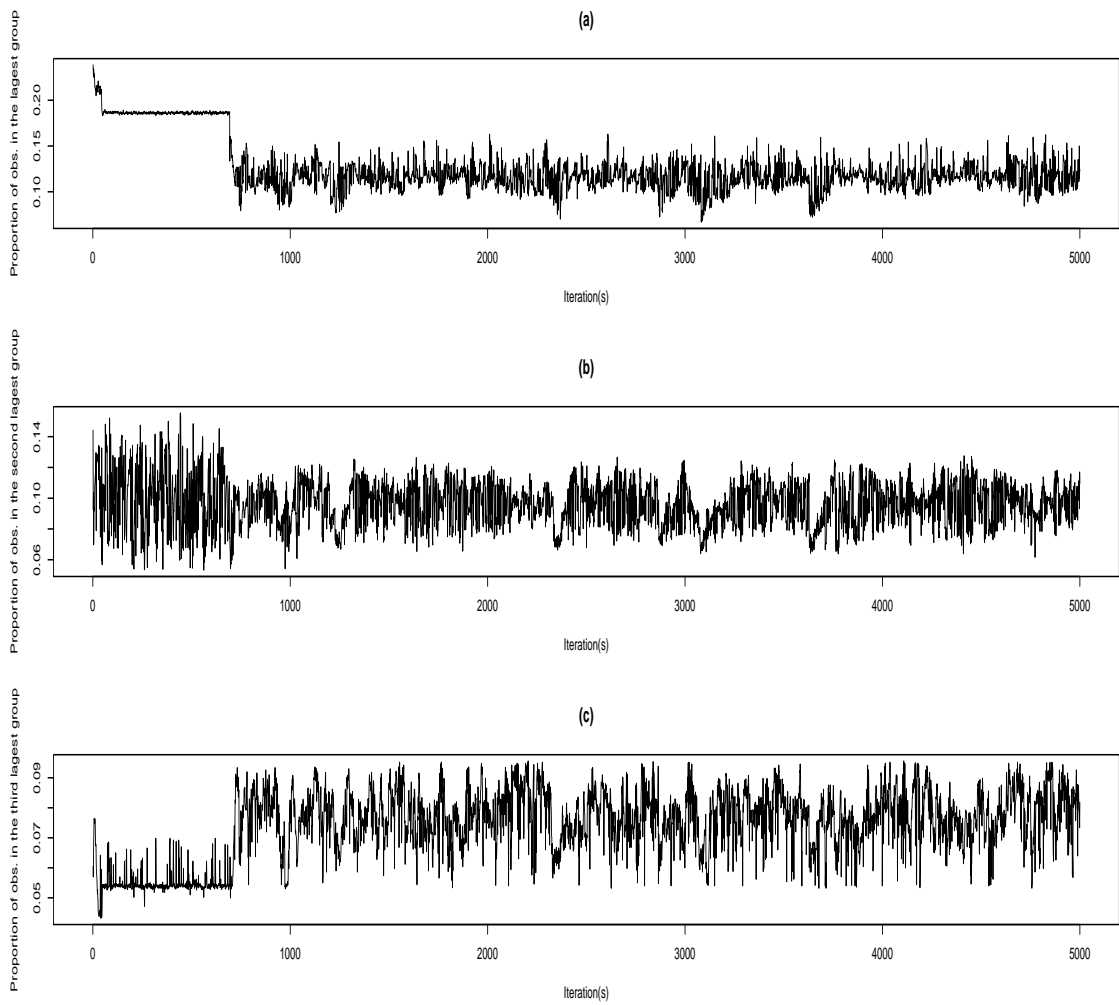


FIGURE 14. (a), (b), (c) plot proportion of observations in the first three largest groups over all iterations.

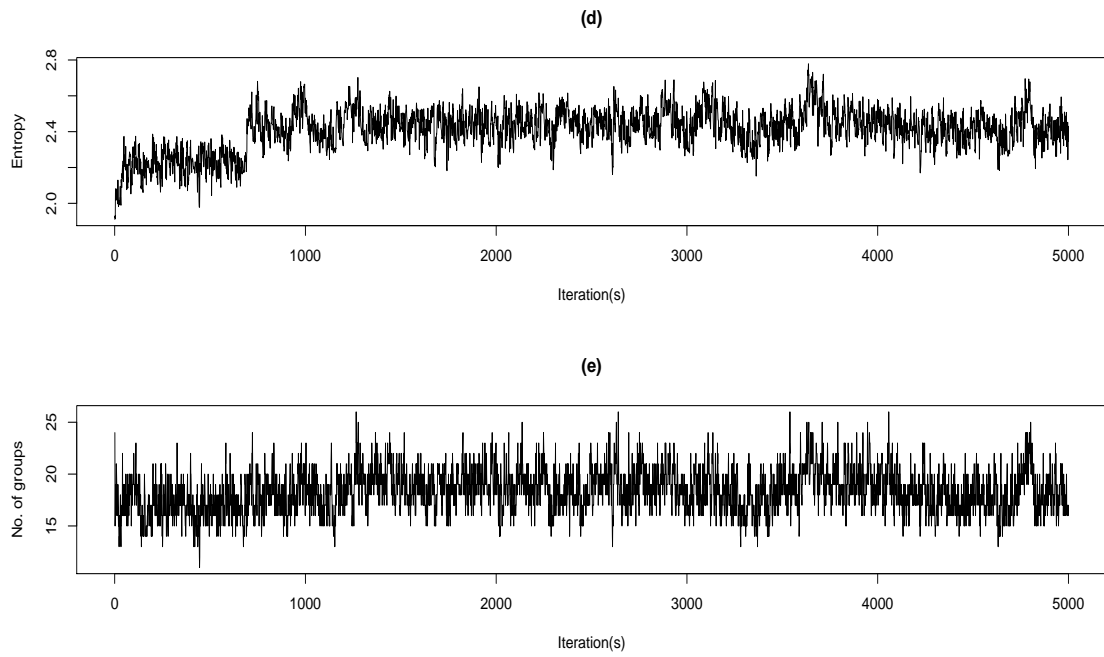


FIGURE 15. Entropy is calculated at each iteration shown in (d), whereas (e) shows number of groups at each iteration.

TABLE 7. A summary statistic of the parameter estimates obtained from all mixtures with 19 components. For the mean parameters, the sample circular mean directions are given.

%	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$	$\kappa_1$	$\kappa_2$	$\kappa_3$	$\kappa_4$
0.07	2.78	2.86	-2.94	2.97	17.29	21.74	14.18	20.68
0.18	2.54	2.73	-2.71	2.84	19.38	26.05	18.72	22.01
0.39	2.29	2.66	-2.50	2.74	25.66	34.53	25.07	24.91
0.71	2.12	2.62	-2.32	2.66	34.61	41.41	33.08	28.00
1.09	2.03	2.60	-2.16	2.65	42.40	48.05	48.53	39.72
1.49	1.98	2.61	-2.66	2.49	52.86	53.39	63.07	50.68
1.79	1.95	2.61	-2.57	2.52	51.08	53.87	68.96	55.98
2.14	1.94	2.57	-2.34	2.76	45.12	54.23	78.58	61.43
2.73	1.78	2.54	-1.85	2.90	34.04	56.92	86.75	66.81
3.58	1.36	-2.27	-2.58	2.91	21.00	48.75	90.92	72.79
4.14	1.27	-1.88	-3.11	2.84	17.37	38.95	84.12	61.35
4.84	1.44	-2.41	-1.08	2.65	20.09	32.39	61.20	39.21
5.50	1.63	-3.11	-1.14	-2.84	19.90	43.36	67.65	42.52
6.40	1.64	2.25	-0.72	-2.52	28.85	61.64	95.06	62.86
7.89	1.25	-0.48	0.62	-3.11	21.89	44.96	93.38	69.68
9.19	1.46	-1.01	-1.02	-1.27	29.67	50.45	102.36	79.07
12.89	1.26	-1.00	-1.08	2.94	54.29	72.75	107.40	63.76
15.86	1.56	-1.20	-1.11	2.96	71.26	94.02	140.46	78.50
19.12	1.71	-1.87	-1.12	2.97	78.47	103.80	154.76	84.77

TABLE 8. Standard deviations are calculated for the estimates in Table 7. For the mean parameters, the sample circular standard deviations, defined in (Mardia & Jupp, 1999, pp19), are given.

$\%$	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$	$\kappa_1$	$\kappa_2$	$\kappa_3$	$\kappa_4$
0.07	0.60	0.42	0.74	0.53	13.49	14.83	14.23	15.78
0.18	0.59	0.38	0.77	0.58	16.11	17.11	19.79	18.06
0.39	0.48	0.36	0.85	0.59	21.64	21.89	24.16	21.52
0.71	0.38	0.35	1.05	0.67	25.33	20.83	30.70	26.12
1.09	0.26	0.36	1.27	0.74	29.58	21.67	32.03	29.01
1.49	0.17	0.25	1.59	0.89	30.57	21.07	27.69	27.16
1.79	0.17	0.31	1.37	0.92	29.78	21.11	25.20	25.68
2.14	0.17	0.36	1.19	0.67	30.32	22.94	25.24	23.39
2.73	0.33	0.80	1.19	0.36	33.48	26.25	25.11	20.51
3.58	0.40	1.90	1.56	0.23	30.17	27.87	27.87	22.96
4.14	0.30	1.49	1.94	0.40	27.74	33.16	39.70	30.75
4.84	0.33	1.39	1.32	0.86	30.74	39.31	53.11	34.01
5.50	0.34	1.38	0.71	1.88	29.82	41.94	53.69	33.40
6.40	0.41	1.68	1.09	1.36	41.38	43.19	40.79	22.64
7.89	0.32	0.80	1.04	0.78	42.49	38.39	32.67	13.34
9.19	0.25	0.85	0.10	0.81	47.39	48.53	35.49	12.12
12.89	0.39	0.96	0.06	0.16	59.58	55.77	48.47	21.46
15.86	0.39	1.51	0.07	0.04	82.63	79.35	55.19	22.15
19.12	0.32	1.99	0.07	0.04	83.27	81.41	52.14	17.32

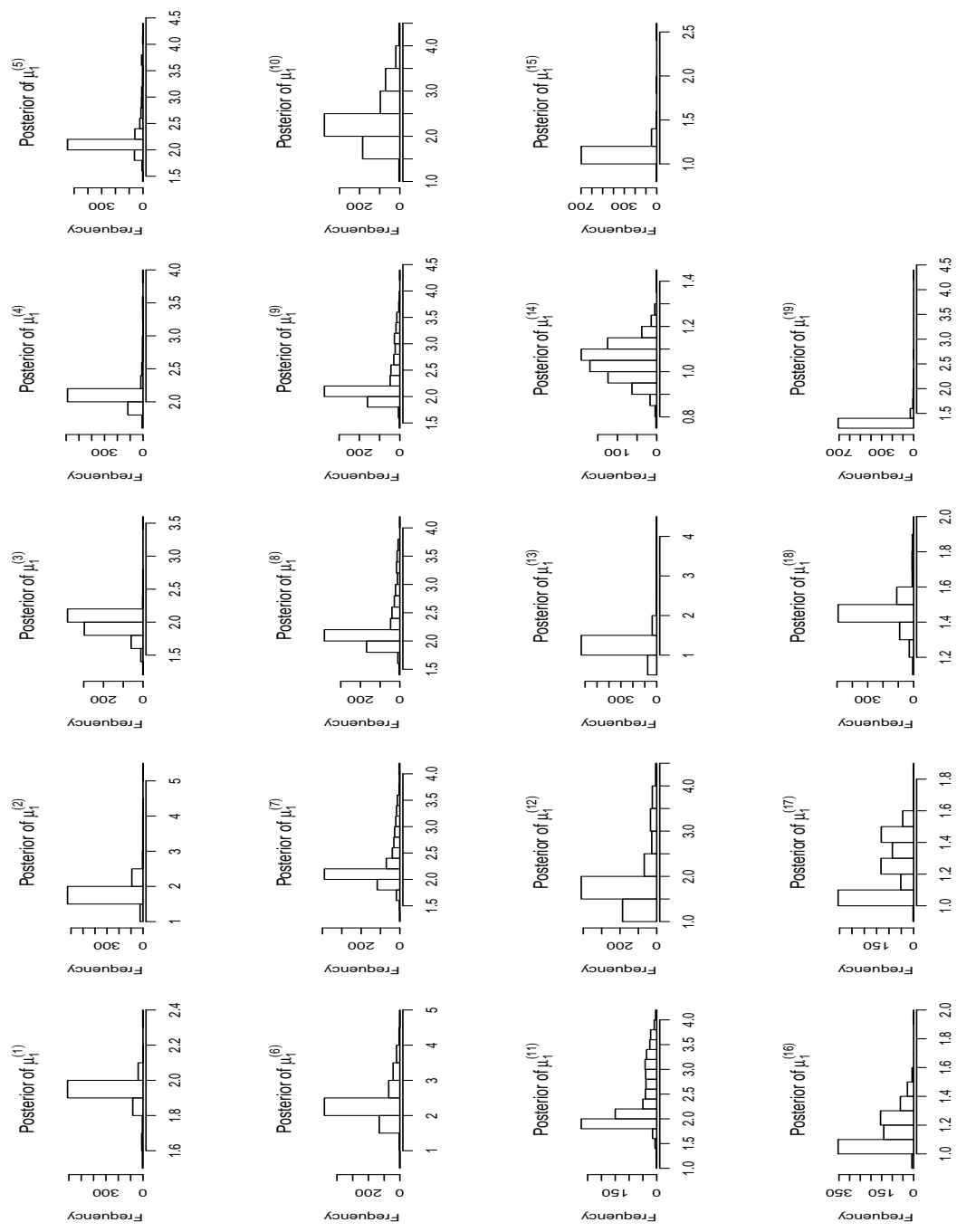


FIGURE 16

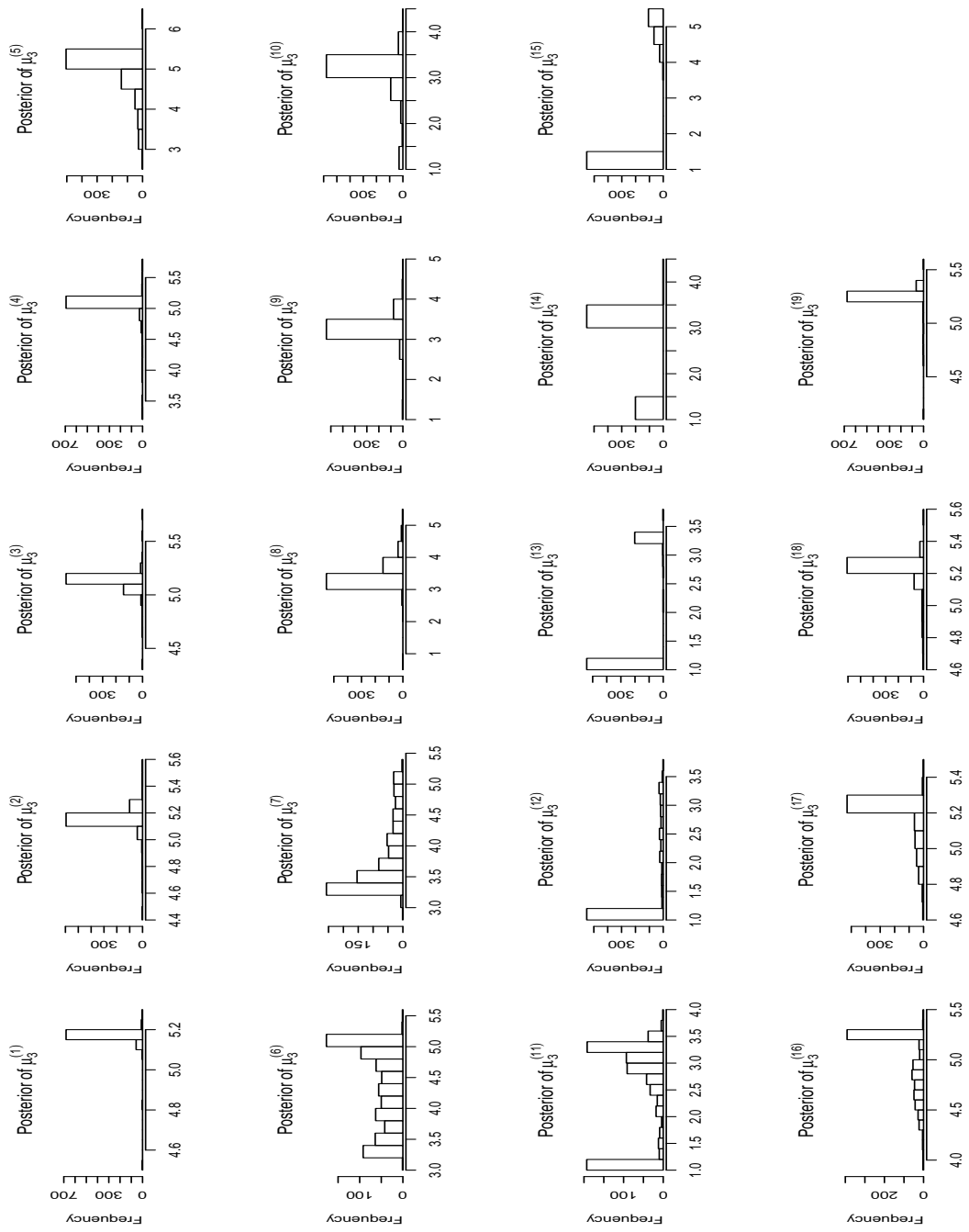


FIGURE 17