Distance-Based Probability Distribution for Set Partitions with Applications to Bayesian Nonparametrics*

David B. Dahl†
August 5, 2008

Abstract

Integration of several types of data is a burgeoning field. Some data naturally lead to formal models; others may convey proximity among observations. Clustering methods are typically either model-based or distanced-based, but not both. We propose a method that is simultaneously model-based and distance-based, permitting the use of both types of data. We show the Dirichlet process induces a clustering distribution in which the probability that an item is clustered with another item is uniform across all items. We provide an extension which incorporates distance information to provide a probability distribution for partitions that is indexed by pairwise distances between items. We show how to utilize this new distance-based probability distribution over partitions as a prior clustering distribution in Bayesian nonparametric models. We show an application to ion mobility-mass spectrometry.

Key Words: Bayesian nonparametrics; Clustering distributions; Dirichlet process mixture model; Partition models

1. Introduction

Consider the space $\mathcal{F}$ of all possible partitions of n items. A partition $\pi = \{S_1, \ldots, S_q\}$ in $\mathcal{F}$ of the set $S_0 = \{1, \ldots, n\}$ has the following properties: 1. $S_i \neq \emptyset$ for $i = 1, \ldots, q$ (i.e., non-empty subsets), 2. $S_i \cap S_j = \emptyset$ for $i \neq j$ (i.e., mutually exclusive subsets), and 3. $\cup_{j=1}^{q} S_j = S_0$ (i.e., exhaustive subsets). The number of subsets q for a partition π can range from 1 (i.e., all items belong to the same subset) and n (i.e., each item is in a singleton subset). It is sometimes convenient to represent a partition π as a vector $\boldsymbol{c}$ of cluster labels, where $c_i = j$ if and only if $i \in S_j$, for $i = 1, \ldots, n$ and $j = 1, \ldots, q$. The size of $\mathcal{F}$ grows exponentially according to the Bell number. Note that a probability distribution over $\mathcal{F}$ is discrete, but the size of the space makes exhaustive calculations impossible for all but very small n.

We use the notation $p(\pi)$ to denote a probability distribution for a random partition π in $\mathcal{F}$. Popular models $p(\pi)$ include product partition models (PPM, see Hartigan 1990; Barry and Hartigan 1992), species sampling models (SSM, see Pitman 1995; Pitman and Yor 1997), and model-based clustering (Richardson and Green 1997; Fraley and Raftery 2002). Others include those of Pitman (2003) and Lijoi et al. (2005). Many of these models are reviewed Quintana (2006) and Lau and Green (2007). These random probability models are routinely used in nonparametric Bayesian analysis.

Some recent applications of Bayesian nonparametric methods have implied random probability influenced *a priori* by covariates. These methods are reviewed by Müller and Quintana (2008). This paper is unique in that it provides an explicit probability distribution over set partitions that is indexed by a function of covariates. The proposed probability distribution (described in Section 3) results from modifying the probability distribution from the Dirichlet process (described in Section 2). Section 4 introduces the

*Citation for this paper: Dahl, D. B. (2008), "Distance-Based Probability Distribution for Set Partitions with Applications to Bayesian Nonparametrics," in *JSM Proceedings*, Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association.

†Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A. The author thanks David H. Russell (Dept. of Chemistry, Texas A&M University), his Ph.D. student Lei Tao, and laboratory scientist Kent Gillig for their application described in Section 6. The author also wishes to thank those who have commented on talks of this paper, especially Peter Müller (Dept. of Biostatistics, U. of Texas MD Anderson Center) who gave several helpful suggestions.

c-value to aid in summarizing a probability distribution over set partitions. The application to Bayesian nonparametrics is described in Section 5 and the method is illustrated with an example in ion mobility-mass spectrometry in Section 6.

2. Distribution for Set Partitions Implied by Dirichlet Process

Using the Dirichlet process (Ferguson 1973) as a prior distribution for an unknown mixing distribution in mixture models was first suggested by Antoniak (1974). Its use leads to arguably the most popular Bayesian nonparametric model known as Dirichlet process mixture models. The Dirichlet process implies the Polya urn scheme (Blackwell and MacQueen 1973), a probability model $p(\pi)$ for a partition given by:

$$p(\pi) = \frac{\prod_{S \in \pi} \alpha \Gamma(|S|)}{\prod_{i=1}^{n} (\alpha + i - 1)}, \tag{1}$$

where $\Gamma(x)$ is the gamma function, $|S|$ is the number of items in subset S, and α can be viewed as a precision parameter influencing the number of subsets. Equation (1) can be equivalently factored as:

$$p(\pi) = p(\boldsymbol{c}) = \prod_{i=1}^{n} p(c_i \mid c_1, \ldots, c_{i-1}), \tag{2}$$

where:

$$p(c_i = j \mid c_1, \ldots, c_{i-1}) \;=\; \begin{cases} \sum_{k=1}^{i-1} \mathrm{I}\{c_k = j\}/(\alpha + i - 1) & \text{for } \; j = 1, \ldots, q \\ \alpha/(\alpha + i - 1) & \text{for } \; j = q + 1, \end{cases} \tag{3}$$

with $\mathrm{I}\{c_k = j\}$ being the indicator function.

Equations (2) and (3) provide a simple prescription for sampling a partition π from the clustering distribution $p(\pi)$ implied by the Dirichlet process. Another technique for sampling from this distribution is to run a Markov chain over $\mathcal{F}$ as described below. Because the model is exchangeable with respect to permutations of the indices, (3) implies the full conditional distribution of c_i given all the other labels (denoted $\boldsymbol{c}_{-i}$) is:

$$p(c_i = j \mid \boldsymbol{c}_{-i}) \propto \begin{cases} |S_j^{-i}| & \text{for } \; j = 1, \ldots, q \\ \alpha & \text{for } \; j = q + 1, \end{cases} \tag{4}$$

where $|S^{-i}|$ is the number of items in subset S not counting i. Note that the probability of forming a new cluster is $\alpha/(\alpha + n - 1)$.

By repeatedly sampling from the full conditional distributions for $i = 1, \ldots, n$, we obtain a Markov chain (Taylor and Karlin 1994) that is irreducible, recurrent, and aperiodic, thus satisfying the conditions to have an equilibrium distribution. This Markov chain has a finite state space since the number of possible clusterings (given by the Bell number) is finite. Since every clustering is accessible from every other clustering through repeated application of the transition rule, the Markov chain is irreducible. The Markov chain is aperiodic because, for every clustering, it is possible that an application of the transition rule will lead to the same clustering. Finally, since the weights are strictly positive, there is positive probability of returning to every clustering and the Markov chain is recurrent.

3. Proposed Distribution for Set Partitions

3.1 Relaxing the Uniform Assumption of the Dirichlet Process

The full conditional distribution in (4) implies that the probability that item i is clustered with another item k is uniform across all k. That is:

$$p(c_i = c_k \mid \boldsymbol{c}_{-i}) \propto \; 1 \; \text{ for } \; k = 1, \ldots, i - 1, i + 1, \ldots, n \tag{5}$$

and the probability that i is placed in a new cluster is proportional to α. Seeing the probability distribution induced by the Dirichlet process in this light leads to an obvious extension: Modify (5) to utilize information regarding the co-clustering of items. This has the effect of relaxing the assumption that the probability i is clustered with another item k is uniform across all k. Specifically, replace one in (5) with a weight h_{ik} influencing the probability i is clustered with item k. That is, (5) becomes:

$$p(c_i = c_k \mid \boldsymbol{c}_{-i}) \propto \; h_{ik} \;\text{ for }\; k = 1, \ldots, i-1, i+1, \ldots, n, \tag{6}$$

where h_{ik} are such that $\sum_{i \neq k} h_{ik} = n - 1$. The scaling of h_{ik} is a key feature: It leaves the probability of forming a new cluster $\alpha/(\alpha + n - 1)$ unchanged from that of the Dirichlet process.

The equivalent expression to (4) for the proposed model is:

$$p(c_i = j \mid \boldsymbol{c}_{-i}) \propto \begin{cases} h_i(S_j) & \text{for } \; j = 1, \ldots, q \\ \alpha & \text{for } \; j = q + 1, \end{cases} \tag{7}$$

where $h_i(S_j) = \sum_{k \in S_j} h_{ik}$. Although we cannot explicitly write the joint distribution implied by the full conditions in (7), we have implicitly defined a probability distribution $p(\pi)$ over $\mathcal{F}$ informed by probabilities of co-clustering through the use of a Markov chain with an equilibrium distribution.

3.2 Specifying the Weights from Distances

There is tremendous flexibility in how the weights h_{ik} might be specified. Here we describe one based on pairwise distances. In many situations the information regarding co-clustering of n items is expressed in terms of an $n \times n$ distance matrix of elements d_{ik} giving the distance between items i and k. Items i and k with a large distance d_{ik} should have a small value h_{ik} to induce a low clustering probability. One potential transformation is:

$$h_{ik} \propto (d^* - d_{ik})^t, \tag{8}$$

where d^* is the overall maximum pairwise distance plus a small increment to ensure that all weights are strictly positive. Note that t has the effect of dampening or accentuating the proximity measurements. Recall that, for each $i = 1, \ldots, n$, we scale $h_{i1}, \ldots, h_{in}$ such that $\sum_{k \neq i} h_{ik} = n - 1$.

4. Summarizing a Distribution over Set Partitions

After repeated application of the transition rule in (7), many clusterings $\pi^{(1)}, \ldots, \pi^{(B)}$ will have been sampled from the distance-based clustering distribution. One may desire to summarize the clustering distribution with a typical clustering both numerically and graphically. We do so using the least-squares clustering method of Dahl (2006). The pairwise probability p_{ik} that two items i and k are clustered is estimated by $\hat{p}_{ik}$, the relative frequency among the candidate clusterings that i and k occupy the same cluster. The least-squares clustering selects the candidate clustering closest to the estimated pairwise probabilities in terms of squared distances:

$$\pi^{\text{LS}} = \underset{\pi \in \{\pi^{(1)}, \ldots, \pi^{(B)}\}}{\arg\min} \sum_{i=1}^{n} \sum_{k=1}^{n} (\delta_{ik}(\pi) - \hat{p}_{ik})^2, \tag{9}$$

where $\delta_{ik}(\pi) = 1$ if i and k occupy the sample cluster in partition π, and 0 otherwise. This method minimizes a posterior expected loss of Binder (1978) with equal costs of clustering mistakes. A closely related method is suggested by Lau and Green (2007).

The least-squares estimate represents a point estimate of the clustering, but does not give an indication of clustering uncertainty. For this purpose, we propose using a quantity we call the c-value. The c-value

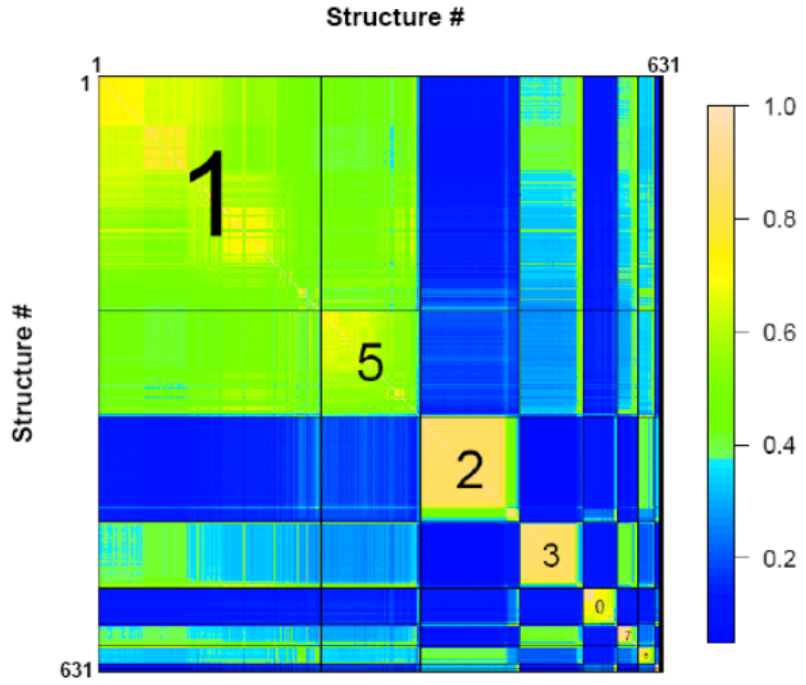


Figure 1: Confidence plot showing the clustering and pairwise probabilities of clustering.

for an item i in cluster S is the average pairwise probability of being clustered with the other items in cluster S, i.e.:

$$c\text{-value}_i = \frac{\sum_{k \in S} \hat{p}_{ik}}{|S|}, \tag{10}$$

where $|S|$ is the size the cluster S. If an item has a high c-value, there is more confidence that the item is clustered appropriately. A low c-value indicates less certainty that the item is placed in the correct cluster. Given a clustering estimate, the items in a cluster can be summarized in a variety of ways. For example, quantitative characteristics of items in a particular cluster can be averaged, although this leads to a cluster summary that may not have been observed. To select an item that represents a particular cluster, we suggest choosing the item with the highest c-value among those items in the cluster.

The clustering uncertainty can also be assessed by plotting a pairwise probability matrix in what we call a confidence plot. Arrange the rows and columns by the least-squares clustering and, within each cluster, order from largest to smallest c-value. Further, make the color of each matrix element indicate the value of the estimated pairwise probability. This plot makes it easy to see which clusters are well defined and which clusters are closely related or very different from other clusters.

Figure 1 gives an example confidence plot for 631 items, where the color of the matrix element (i, j) indicates the probability that item i is clustered with item j. The rows and columns of the matrix have been re-arranged to group items by cluster, with the smallest cluster (labeled 1) in the upper-right hand corner. Notice that clusters 2 and 3 contain items that are especially similar to each other (i.e., tan colored and, hence, high probabilities of being clustered) and very different from other clusters (i.e., the off-diagonal blocks are dark blue or purple). On the other hand, clusters 1 and 5 could relatively easily form one big cluster, since the off-diagonal blocks are similar in color to the colors in clusters 1 and 5.

5. Application to Bayesian Nonparametrics

In this section, we describe how the proposed distribution over set partitions can be used as a prior clustering distribution in Bayesian nonparametric models, leading to a new class of Bayesian nonparametric

models that utilize distance-based information to influence co-clustering of observations. Many Bayesian nonparametric models for data $\boldsymbol{y} = (y_1, \ldots, y_n)$ take the following form:

$$\begin{aligned} y_i \mid \theta_i &\sim p(y_i|\theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim p(G), \end{aligned} \tag{11}$$

where $p(y|\theta)$ is a parametric sampling model indexed by θ and G is an unknown distribution having a prior distribution given by $p(G)$. Note that $y_1, \ldots, y_n$ are independent given $\theta_1, \ldots, \theta_n$, and $\theta_1, \ldots, \theta_n$ are independent given G. The model can be enriched by letting the sampling model depend on other parameters. Further, the assumption that observations within a cluster are exchangeable and observations across clusters are independent can be relaxed.

If $p(G)$ is the Dirichlet process (Ferguson 1973) with mass parameter α and centering distribution G_0, the model in (11) can be rewritten explicitly in terms of the centering distribution G_0 and a set partition $\pi = \{S_1, \ldots, S_q\}$:

$$\begin{aligned} y_i \mid \theta_i &\sim p(y_i|\theta_i) \\ \theta_i &= \sum_{j=1}^{q} \phi_j \mathrm{I}\{i \in S_j\} \\ \phi_j &\sim G_0 \\ \pi &\sim p(\pi), \end{aligned} \tag{12}$$

where $p(\pi)$ is given in (1). Again, the model may be extended by having G_0 dependent on parameters for which prior distributions are specified. A new class of Bayesian nonparametric models is obtained by simply using the proposed distance-based distribution for set partitions implied by (7) for $p(\pi)$.

The substitution of $p(\pi)$ in (1) to that of $p(\pi)$ in (7) involves nearly no modification of existing sampling methods and code to fit Dirichlet process mixture models. A technical point does arise, however, when one is interested in the posterior predictive distribution $p(y_{n+1}|y_1, \ldots, y_n)$ of a new observation y_{n+1}. The issue stems from the fact that the proposed distribution implied by (7) is not an exchangeable product partition model (EPPF, see Pitman 1995) since it does not scale across sample size. That is, using the cluster label notation, the following property does not hold for the proposed distribution:

$$p(\boldsymbol{c}) = \sum_{j=1}^{q+1} p(\boldsymbol{c}, c_{n+1} = j), \tag{13}$$

although it does for the clustering distribution of the Dirichlet process.

All inferences are made from a posterior distribution that conditions on the observed data. For example, inference on the model parameters $\theta_1, \ldots, \theta_n$ is based on the distribution $p(\theta_1, \ldots, \theta_n|y_1, \ldots, y_n)$. Likewise, clustering inference is based on the distribution $p(\pi|y_1, \ldots, y_n)$. In situations where one is interested in posterior predictive distribution $p(y_{n+1}|y_1, \ldots, y_n)$, we suggest fitting the model with $n + 1$ observations, treating the y_{n+1} as a missing observation whose value is updated at each iteration of the Markov chain. Further, we recommend making its distance uniform across all observations, specifically, $h_{i,n+1} = 1$ for all $i = 1, \ldots, n$. Under this choice, (13) holds approximately for distances that would be used in practice.

6. Example: Cross Section Analysis in Ion Mobility-Mass Spectrometry Experiments

There is considerable interest in cluster analysis of candidate peptides from molecular dynamic simulations in the field of ion mobility-mass spectrometry. A peptide's backbone can be represented as the

(x, y, z) coordinates of an amino acid's four heavy atoms (N, C_α, C, and O) along the backbone. The root-mean-square-deviation (RMSD) between two optimally-aligned peptides gives a measure of dissimilarity among peptides. The pairwise distances could be used in a distance-based clustering procedure. Conversely, a protein's backbone can be described by torsion angle pair (ϕ, ψ) along the backbone (Ramachandran et al. 1963). At the i^{th} residue, the ϕ angle describes the torsion around the bond N_i-$C_{\alpha i}$, measuring the angle between the C_{i-1}-N_i and the $C_{\alpha i}$-C_i bonds, while the ψ angle describes the torsion around the bond $C_{\alpha i}$-C_i, measuring the angle between the N_i-$C_{\alpha i}$ and the C_i-N_{i+1} bonds. The similarity of the vectors of torsion angle pairs among peptides can also be used in a model-based clustering procedure.

We present methodology that allows both types of data to influence the clustering of peptides by proposing a Bayesian nonparametric model for the torsion angles along a backbone whose prior clustering distribution is based on the pairwise RMSD distance values. The model-based part of our proposed procedure is a multiple-positional version of the single-position nonparametric Bayesian model presented by Lennox et al. (2009). Unlike Lennox et al. (2009), however, we also use the RMSD distances to influence clustering. Lennox et al. (2009) use the sine-model bivariate von Mises distribution (Singh et al. 2002) for the sampling distribution of a torsion angle at a single position, specifically:

$$p(\phi, \psi \mid \mu, \nu, \kappa_1, \kappa_2, \lambda) = C \exp\{\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \lambda \sin(\phi - \mu) \sin(\psi - \nu)\}, \quad (14)$$

for $\phi, \psi, \mu, \nu \in (-\pi, \pi]$, $\kappa_1, \kappa_2 > 0$, $\lambda \in (-\infty, \infty)$, and

$$C^{-1} = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m I_m(\kappa_1) I_m(\kappa_2). \quad (15)$$

For notational convenience, let Ω be a 2×2 matrix with both off-diagonal elements equal to $-\lambda$ and diagonal elements κ_1 and κ_2. Instead of a single amino acid position, we consider a peptide with M amino acid positions. The proposed model is merely an instance of (12), with $p(\pi)$ being the proposed distance-based distribution on set partitions. Further, the sampling model $p(y_i \mid \theta_i)$ for the vector of torsion angles is the product of sine-model bivariate von Mises distribution, with $y_i = (\phi_{i1}, \psi_{i1}, \ldots, \phi_{iM}, \psi_{iM})$ being a vector of torsion angle pairs at the M positions and $\theta_i = (\mu_{i1}, \nu_{i1}, \Omega_{i1}, \ldots, \mu_{iM}, \nu_{iM}, \Omega_{iM})$ being the parameters of the sampling distribution. The centering distribution G_0 is expressed as a product $H_1 H_2$, where H_1 is a product of M independent bivariate von Mises sine models for the means and H_2 is the product of bivariate Wishart distributions for the precision matrices. We recommend fitting the model with Markov chain Monte Carlo (MCMC) to obtain samples from the posterior distribution for numerical integration. The computational procedure is a natural M amino acid extension of the procedure given by Lennox et al. (2009) for a single amino acid position.

The method was applied to peptide MIFAGIK (residues 80-86 of cytochrome c) in which 631 candidate structures from molecular dynamic simulations were available. The results presented here were obtained by setting $\alpha = 1$ and t set at a value between 15 and 30 to provide a clustering with a few well-separated clusterings. We applied the transition rule about 500,000 times. In repeated application of the algorithm to the same dataset, the resulting clusterings were very similar as measured by the adjusted Rand index (Hubert and Arabie 1985). The least-squares clustering was obtained and the confidence plot in Figure 1 was produced. Further, the c-value was used to select representative peptides (as described in Section 4) and is displayed in Figure 2. Figure 3(a) shows the marginal distribution of position 2. Figure 3(b) shows the conditional distribution of position 2 given position 3 is at its mode $(\phi, \psi) = (-1.40, -1.10)$, marginalizing over positions 0, 1, and 4. Figure 3(c) shows position 2's (marginal) conditional distribution given position 3 is at its other mode. Such conditional distributions can be very helpful in protein structure prediction.

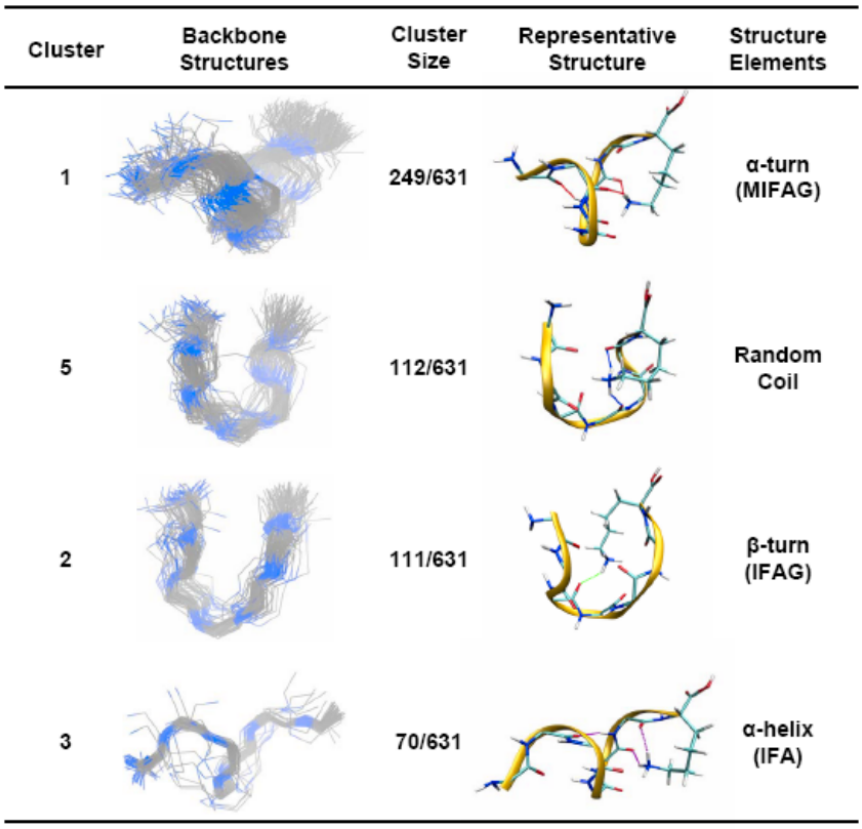


Figure 2: Three-dimensional view of backbones of individual peptides grouped by cluster, together with the backbone that best summaries the cluster. Results from the four largest clusters are shown.

References

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

Barry, D. and Hartigan, J. A. (1992), "Product partition models for change point problems," *The Annals of Statistics*, 20, 260–279.

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65, 31–38.

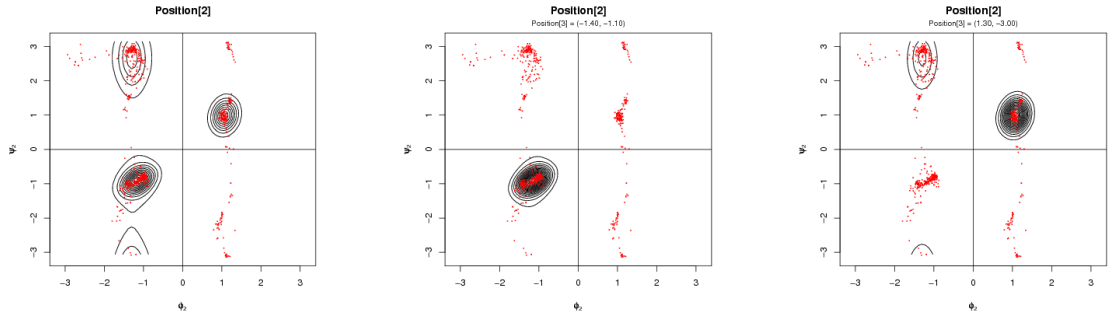
Blackwell, D. and MacQueen, J. B. (1973), "Ferguson Distributions Via Polya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

Dahl, D. B. (2006), "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model," in *Bayesian Inference for Gene Expression and Proteomics*, eds. Do, K.-A., Müller, P., and Vannucci, M., Cambridge University Press, pp. 201–218.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

Fraley, C. and Raftery, A. E. (2002), "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, 97, 611–631.

Hartigan, J. A. (1990), "Partition Models," *Communications in Statistics, Part A – Theory and Methods*, 19, 2745–2756.



(a) Marginal distribution at position 2.

(b) Conditional distribution at position 2 given position 3 is $(-1.40, -1.10)$.

(c) Conditional distribution at position 2 given position 3 is $(1.30, -3.00)$.

Figure 3: Marginal and conditional distributions for selected positions.

Hubert, L. and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218.

Lau, J. W. and Green, P. J. (2007), "Bayesian Model Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558.

Lennox, K. P., Dahl, D. B., Vannucci, M., and Tsai, J. W. (2009), "Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics," *Journal of the American Statistical Association*, in press.

Lijoi, A., Mena, R. H., and Prünster, I. (2005), "Hierarchical Mixture Modeling with Normalized Inverse-Gaussian Priors," *Journal of the American Statistical Association*, 100, 1278–1291.

Müller, P. and Quintana, F. A. (2008), "Random Partition Models with Regression on Covariates," submitted.

Pitman, J. (1995), "Exchangeable and Partially Exchangeable Random Partitions," *Probability Theory and Related Fields*, 102, 145–158.

— (2003), "Poisson-Kingman Partitions," in *Science and statistics: A festschrift for Terry Speed*, IMS Press, pp. 1–34.

Pitman, J. and Yor, M. (1997), "The Two-parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator," *The Annals of Probability*, 25, 855–900.

Quintana, F. A. (2006), "A Predictive View of Bayesian Clustering," *Journal of Statistical Planning and Inference*, 136, 2407–2429.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963), "Stereochemistry of Polypeptide Chain Configurations," *Molecular Biology*, 7, 95–99.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With An Unknown Number of Components," *Journal of the Royal Statistical Society, Series B, Methodological*, 59, 731–758.

Singh, H., Hnizdo, V., and Demchuk, E. (2002), "Probabilistic Model for Two Dependent Circular Variables," *Biometrika*, 89, 719–723.

Taylor, H. M. and Karlin, S. (1994), *An Introduction to Stochastic Modeling*, Academic Press.