

# Random Partition Models with Regression on Covariates

Peter Müller<sup>1</sup> and Fernando Quintana<sup>2</sup>

<sup>1</sup> M.D. Anderson Cancer Center, Houston, TX

<sup>2</sup> Pontificia Universidad Católica, Santiago, Chile

## Abstract

Many recent applications of nonparametric Bayesian inference use random partition models, i.e. probability models for clustering a set of experimental units. We review the popular basic constructions. We then focus on an interesting extension of such models. In many applications covariates are available that could be used to *a priori* inform the clustering. This leads to random clustering models indexed by covariates, i.e., regression models with the outcome being a partition of the experimental units. We discuss some alternative approaches that have been used in the recent literature to implement such models, with an emphasis on a recently proposed extension of product partition models. Several of the reviewed approaches were not originally intended as covariate-based random partition models, but can be used for such inference.

## 1 Introduction

We review probability models for random partitions. In particular we are interested in random partition models in the presence of covariates. In other words, we discuss regression models where the outcome is an arrangement of experimental units in clusters.

Let  $S = \{1, \dots, n\}$  denote a set of experimental units. A partition is a family of subsets  $S_1, \dots, S_k$  with  $S = S_1 \cup \dots \cup S_k$ ,  $S_j \cap S_{j'} = \emptyset$ . We write  $\rho_n = \{S_1, \dots, S_k\}$ . The random number of clusters,  $k$ , is part of  $\rho_n$ . When the sample size  $n$  is understood from the context we drop the subindex and write  $\rho$ . Sometimes it is technically more convenient to describe a partition by a set of cluster membership indicators  $s_i$  with  $s_i = j$  if  $i \in S_j$ ,  $i = 1, \dots, n$ . Let  $\mathbf{s}_n = (s_1, \dots, s_n)$ . Finally, let  $k_n$  denote the number of clusters. Again, we drop the index  $n$  if the sample size is understood. The number of clusters  $k$  is implicitly coded in  $\mathbf{s}_n$  and  $\rho_n$ . We write  $n_{nj} = |S_j|$  for the size of the  $j$ -th cluster. Again, we drop the subscript  $n$  if the underlying sample size is understood from the context.

A random partition model is a probability model  $p(\rho_n)$ . Two basic properties are desirable for random partition models. The model should be exchangeable with respect to permutations of the indices of the experimental units. Let  $\pi = (\pi_1, \dots, \pi_n)$  denote a permutation of  $S$ , and let  $\mathbf{s}_\pi = (s_{\pi_1}, \dots, s_{\pi_n})$  describe the clusters implied by re-labeling experimental unit  $i$  by  $h = \pi_i^{-1}$ , i.e.,  $\pi_h = i$ . We require

$$p(\mathbf{s}) = p(\mathbf{s}_\pi)$$

for all partitions  $\pi$ . A second important property is that the model should scale across sample sizes. We want

$$p(\mathbf{s}_n) = \sum_{j=1}^{k_n+1} p(\mathbf{s}_n, s_{n+1} = j).$$

A probability model on  $\rho_n$  that satisfies the two conditions is called an exchangeable product partition function (EPPF) (Pitman, 1996). Exploiting the invariance with respect to relabeling the EPPF can be written as  $p(n_{n1}, \dots, n_{nk})$ .

Several probability models  $p(\rho_n)$  are used in the recent literature, including product partition models (PPM), species sampling models (SSM) and model based clustering (MBC).

The SSM and MBC satisfy the requirements of symmetry and scalability by definition, but not all PPMs do. See, for example, Quintana (2006) for a recent review.

Usually the model is completed with a sampling model for observed data  $\mathbf{y} = (y_1, \dots, y_n)$  given  $\rho_n$ . A typical sampling model defines independent sampling across clusters and exchangeability within clusters. In the following discussion we assume that this is the case. We do so for the benefit of a more specific discussion, but without loss of generality. We represent exchangeability within clusters as independent sampling given cluster specific parameters  $\xi_j$ :

$$p(\mathbf{y} \mid \rho_n) = \prod_{j=1}^k \int \prod_{i \in S_j} p(y_i \mid \xi_j^*) dp(\xi_j^*). \quad (1)$$

For example,  $p(y_i \mid \xi_j^*)$  could be a normal model  $N(\xi_j^*, S)$ , and the prior  $p(\xi_j^*)$  could be a conjugate normal prior. In the following discussion we focus on the prior model  $p(\rho_n)$ , and assume (1) when a specific sampling model is required. But little changes in the discussion if the sampling model is of a different form.

The most popular choice for  $p(\rho_n)$  in the recent Bayesian literature is the special case of Dirichlet process (DP) priors (Ferguson, 1973; Antoniak, 1974). DP priors are probability models for unknown distributions  $G$ , i.e., the DP is a probability model on probability models. We write  $G \sim \text{DP}(\alpha, G^*)$ . The base measure parameter  $G^*$  defines the prior mean,  $E(G) = G^*$ . The total mass parameter  $\alpha$  is a precision parameter. One of the important properties is the a.s. discrete nature of  $G$ . This property can be exploited to define a random partition by considering a sequence of i.i.d. draws,  $\xi_i \sim G$ ,  $i = 1, \dots, n$ . The discrete nature of  $G$  implies positive probabilities for ties among the  $\xi_i$ . Let  $\{\xi_1^*, \dots, \xi_k^*\}$  denote the unique values among the  $\xi_i$  and define  $S_j = \{i : \xi_i = \xi_j^*\}$ . The implied probability model on  $\rho_n = (S_1, \dots, S_k)$  is known as the Polya urn scheme. Let  $[x]_m = x \cdot (x+1) \cdot \dots \cdot (x+m-1)$  denote the Pochhammer symbol. The Polya urn defines

$$p(\rho_n) = \frac{\alpha^k \prod_{j=1}^k (n_j - 1)!}{[\alpha]_n}. \quad (2)$$

Model (2) can be written as  $p(\rho_n) \propto \prod_{j=1}^k c(S_j)$ , with  $c(S_j) = \alpha(n_j - 1)!$  Models of the form  $p(\rho_n) \propto \prod c(S_j)$  for general  $c(S_j)$  are known as PPMs (Hartigan, 1990; Barry and Hartigan, 1993).

Equivalently the Polya urn can be characterized by the predictive probability function (PPF), that is

$$p_j(\rho_n) \equiv p(s_{n+1} = j \mid s_1, \dots, s_n) \propto \begin{cases} n_j & j = 1, \dots, k_n \\ \alpha & j = k_n + 1 \end{cases} \quad (3)$$

It is easily verified that the Polya urn defines indeed an EPPF. Models that are characterized by a sequence of PPFs  $\{p_j(\rho_n), j = 1, \dots, k \text{ and } n = 1, 2, \dots\}$  and that satisfy the symmetry and scalability requirements are known as SSMs (Pitman, 1996).

Probability models for random partitions are now routinely used in Bayesian data analysis. In this article we discuss an extension to probability models for random partitions indexed with covariates. An interesting example is reported in Dahl (2008). Proteins are clustered on the basis of 3-dimensional structure. Structure is recorded as a sequence of 7 characteristic angles of the backbone. Let RMSD denote the (root) minimum Euclidean distance between any two proteins, after optimally aligning the two molecules. Dahl (2008) argues that proteins with small RMSD should be *a priori* more likely to co-cluster than others. In other words the prior probability model on clustering should be indexed with covariates.

Let  $x_i$  denote the covariates that are specific to experimental unit  $i$  and write  $\mathbf{x}_n = (x_1, \dots, x_n)$ . We consider models of the form  $p(\rho_n | \mathbf{x}_n)$ . But more generally, the covariates need not be indexed by experimental units. Several of the following models only require that covariates can be grouped by cluster. For example, in Dahl (2008) the covariates are RMSD and are specific to any pair of proteins. Partition models with covariates are useful in many applications, but for a relative lack of standard methods are not currently used extensively.

The rest of this article is organized as follows. In sections 2 through 5 we discuss models for random partitions with covariates based on several alternative approaches, including augmented response vectors, dependent DP models, and hierarchical mixture of experts models. In section 6 we review in more detail an approach based on extending the product partition model.

## 2 Clustering with Augmented Response Vectors

Probability models  $p(\rho_n)$  can systematically be generalized to  $p(\rho_n | \mathbf{x})$  by the following device. Assume for a moment that covariates are random, even when they are fixed by design, using a sampling model as in (1),  $p(\mathbf{x} | \rho_n) = \prod_j \int p(x_i | \eta_j^*) dp(\eta_j^*)$ . The implied conditional distribution  $p(\rho_n | \mathbf{x})$  defines a probability model for  $\rho_n$ , indexed by covariates, as desired. The approach is particularly convenient when the model is combined with the sampling model (1) for observed data  $\mathbf{y}$ . This is implemented in combination with the DP model among others, in Müller et al. (1996) or more recently in Shahbaba and Neal (2007) and in Dunson and Park (2008). The combined model becomes

$$p(\mathbf{y} | \rho_n) p(\rho_n | \mathbf{x}) \propto \prod_{j=1}^n \int p(x_i | \eta_j^*) dp(\eta_j^*) \prod_{j=1}^n \int p(y_i | \xi_j^*) dp(\xi_j^*) p(\rho_n) = \prod_{j=1}^n \left[ \int p(x_i | \eta_j^*) p(y_i | \xi_j^*) dp(\eta_j^*) dp(\xi_j^*) p(\rho_n) \right]. \quad (4)$$

If  $p(\rho_n)$  is the Polya urn (2) implied by the DP prior, then the model can be rewritten as independent sampling of  $(y_i, x_i)$  given latent variables  $(\xi_i, \eta_i)$  that in turn are i.i.d. samples

from a random probability measure with DP prior. Let  $\xi = (\xi_1, \dots, \xi_n)$  and similar for  $\eta$ . Conditional on a random measure  $G$  consider:

$$p(\mathbf{y} \mid \xi) p(\xi, \eta \mid \mathbf{x}, G) \propto \prod_i p(y_i \mid \xi_i) p(x_i \mid \eta_i) G(\xi_i, \eta_i).$$

The model is completed with a DP prior,  $G \sim DP(G^*, \alpha)$  with base measure  $G^*(\eta_j^*, \xi_j^*) = p(\eta_j^*) p(\xi_j^*)$ .

The latent variables  $(\xi_i, \eta_i)$  implicitly code the partition  $\rho_n$  by  $\xi_{i_1} = \xi_{i_2}$  if and only if  $s_{i_1} = s_{i_2}$ . Marginalizing with respect to  $G$  is (4) with the Polya urn prior  $p(\rho_n)$ .

There are two important limitations to this approach. First, the approach requires the specification of a probability model for the covariates  $x_i$ , even if these are not random quantities. For long lists of covariates with a mix of different data formats it can be a challenging task to define suitable  $p(x_i \mid \eta_i)$ . Such situations are quite common for many biomedical applications where  $x_i$  could be a long list of subject-specific characteristics, including treatment history, age, ethnicity, insurance coverage, health literacy, location, and many more.

Second, the approach includes a trap. It is usually defined for the special case with  $x_i \in R^p$  and  $y_i \in R^q$  as vectors of continuous variables. Assume multivariate normal kernels  $p(x_i \mid \eta_i) = N(\eta_i, T)$  and sampling model  $p(y_i \mid \xi_i) = N(\xi_i, S)$  and conjugate normal priors  $p(\eta_j^*)$  and  $p(\xi_j^*)$ . Here  $S$  and  $T$  are (possibly fixed) hyperparameters. It is tempting to combine factors and write the model as one joint model with augmented response vector  $(x_i, y_i)' \sim N((\eta_j^*, \xi_j^*)', \Sigma)$ , where  $\Sigma$  is a  $(p + q) \times (p + q)$  covariance matrix. However, care needs to be taken to enforce a block diagonal structure with  $\Sigma_{i_1, i_2} = 0$  for any two indices  $i_1 \leq p$  and  $i_2 > p$ . Otherwise the equivalence with (4) is lost and it becomes difficult to understand the implied model.

The same comment is true for the base measure  $G^*$  and possible hyperpriors.

### 3 DDP and Related Models

Dependent Dirichlet process (DDP) models are another class of probability models that can generate covariate-dependent random partitions. DDP models were introduced in MacEachern (1999). Although not usually presented as probability models for covariate dependent random partitions, the special case of DDP models with varying weights can be used to achieve this purpose. Recall the DP model. Let  $\delta(m)$  denote a point mass at  $m$ . A random probability measure (RPM)  $G$  has a DP prior,  $G \sim DP(\alpha, G^*)$  if it can be written as

$$G = \sum_{h=1}^{\infty} \pi_h \delta(m_h) \tag{5}$$

with  $m_h \sim G^*$ , i.i.d., and  $\pi_h = V_h \prod_{f < h} (1 - V_f)$  where  $V_h \sim \text{Be}(\alpha, 1)$ , i.i.d. (Sethuraman, 1994).

The DDP model was introduced to generalize the DP prior to a probability model for a family of dependent RPMs  $\{G_x, x \in X\}$  indexed by  $x$ . The marginal model remains a

DP prior,  $G_x \sim DP(\cdot)$  marginally. Implementations of the DDP vary in the way how the dependence of  $G_x$  across  $x$  is defined. One variation is to replace the weights  $\pi_h$  in the DP model by covariate-dependent weights  $\pi_h(x)$ . This is implemented, for example, in Griffin and Steel (2006).

An alternative implementation of DDP models introduces the desired dependence across  $x$  by replacing  $m_h$  in (5) with  $m_{xh}$ , and defining a dependent prior  $p(m_{xh}, x \in X)$ . For example, a Gaussian process prior model. The weights remain unchanged across  $x$ . Although this fixed weight implementation is the more popular one, it is not of interest in the current discussion. It does not naturally lead to covariate dependent random partitions.

The varying weights DDP model can be used to define a random partition model  $p(\rho_n | \mathbf{x})$  indexed by covariates. This is assuming constant locations  $m_h$ , unchanged across the level of  $x$ . The approach is simple. Similar to the definition of the Polya urn we consider independent samples  $\xi_i \sim G_{x_i}$  with a DDP prior on  $\{G_x, x \in X\}$ . The discrete nature of the underlying DP priors implies a positive probability of ties among the  $\xi_i$ . We use this to define clusters. Let  $\{\xi_1^*, \dots, \xi_k^*\}$  denote the  $k \leq n$  unique values of  $\xi_i$ . We define  $i \in S_j$  if  $\xi_i = \xi_j^*$ . The simple form of (3) is lost. But the model still allows straightforward posterior simulation.

A related process is the *weighted mixture of DPs* (WMDP) proposed in Dunson et al. (2007). The WMDP defines a probability distribution for a family  $\{G_x, x \in X\}$  of discrete RPMs. Like the varying weight DDP the construction is such that the random measures  $G_x$  share common locations  $m_h$  for the point masses, but the weights  $\pi_{hx}$  vary across locations. The varying weights are defined by multiplying with a kernel that is a function of the covariates. The construction starts with a grid of covariate values  $\{x_\ell^o, \ell = 1, \dots, L\}$  and associated RPMs  $G_\ell^o$ . The RPM  $G_\ell^o$  characterizes an RPM  $G_x$  with covariate  $x = x_\ell^o$ . For covariate values between the grid points the RPM  $G_x$  is defined as a weighted mixture of the  $G_\ell^o$  measures.

Let  $f(x, x')$  denote a (fixed) kernel. The WMDP assumes

$$G_x = \sum_{\ell=1}^L \left( \frac{\gamma_\ell f(x, x_\ell^o)}{\sum_{k'} \gamma_{k'} f(x, x_{k'}^o)} \right) G_\ell^o \text{ and } G_\ell^o \sim DP(\alpha, G^*).$$

As usual the construction can be exploited to define a covariate-dependent random partition model. Consider  $\xi_i \sim G_{x_i}$ ,  $i = 1, \dots, n$ , and let  $\{\xi_j^*, j = 1, \dots, k\}$  denote the  $k \leq n$  unique values of  $\xi_i$ . The indicators  $s_i = j$  if  $\xi_i = \xi_j^*$  define a random partition. The implied model  $p(\rho_n | \mathbf{x})$  is a random partition model indexed by covariates, as desired.

A similar construction is the *probit stick-breaking process* (PSBP) that was recently introduced in Chung and Dunson (2008). They construct a probability model for related RPMs  $\{F_x, x \in X\}$  by replacing the weights  $\pi_h$  in the DP prior by a probit model.

$$\pi_h(\mathbf{x}) = \Phi(\eta_h(\mathbf{x})) \prod_{\ell < h} \{1 - \Phi(\eta_\ell(\mathbf{x}))\}. \quad (6)$$

Here  $\eta_h(\mathbf{x}) = \alpha_h - \psi_h |x - \Gamma_h|$  is a linear predictor for the probit model and  $\Phi(\cdot)$  is the standard normal cdf. Chung and Dunson (2008) define the model for a multivariate covariate vector  $\mathbf{x}$ . They complete the model with independent normal priors for the probit parameters.

## 4 Hierarchical Mixture of Experts

A widely used method for flexible regression is the hierarchical mixture of experts (HME) model (Jordan and Jacobs, 1994). It is extended to the Bayesian HME in Bishop and Svensén (2003). An implementation for binary outcomes is developed in Wood et al. (2008).

The HME is a flexible regression model for an outcome  $y_i$  and covariate  $\mathbf{x}_i$ . It can be written as

$$p(y_i | \mathbf{x}_i, \eta, \xi) = \sum_{j=1}^k \pi_j(\mathbf{x}_i; \eta_j) p(y_i | \mathbf{x}_i, \xi_j) \quad (7)$$

where  $\eta$  parametrizes the weight for the  $j$ -th component. Bishop and Svensén (2003) use products of logistic regressions to specify the weights  $\pi_j$ . The parameters  $\xi$  parametrize the sampling model for the  $j$ -th component, for example a normal linear regression for a continuous outcome, or a probit binary regression for a binary outcome as in Wood et al. (2008). The component models are also known as the experts. Thus the name of the model.

Mixture sampling models like (7) can be used to define random partition models. Partition or clustering schemes induced by such models are known as model based clustering (Dasgupta and Raftery, 1998). The approach is based on writing the mixture model (7) as a hierarchical model with latent indicators  $s_i$ :

$$p(y_i = j | s_i, \xi) = p(y_i | \mathbf{x}_i, \xi_j) \quad \text{and} \quad p(s_i = j | \mathbf{x}_i, \eta) = \pi_j(\mathbf{x}_i; \eta_j).$$

The indicators  $s_i$  define the partition and  $p(\rho_n | \mathbf{x}_n)$  is a multinomial distribution.

The approach is based on the finite parametric model (7). It inherits all advantages and limitations of a parametric approach. An important advantage is parsimony of the parametrization and interpretability of the parameters. An important limitation is the restriction to the specific parametric form of  $\pi_j$ . For example, if the logistic regression for  $\pi_j$  does not include interaction terms, then no amount of data and evidence will be able to introduce such terms in the posterior predictive inference. Another limitation is the fixed number of clusters  $k$ . This limitation can easily be addressed by adding a hyperprior  $p(k)$ . But in either case the number of clusters would not be adaptive to the sample size, as is the case in other random partition models. The same comments apply to any other model with explicitly parameterized weight function, such as the PSBP in (6).

## 5 Other Approaches

In the previous sections we have reviewed approaches and models that can be used to define covariate-dependent random partitions. A common theme of these approaches is that they were introduced in the literature for different purposes, and just happen to imply covariate-dependent random partition models. The reason for including these approaches in the review is simply the relative lack of literature concerned with covariate-dependent random partition models.

One exception is Dahl (2008). We mentioned the motivating application in the introduction. The desired covariate-based random clustering model is constructed by a smart modification of the Polya urn scheme (3). Let  $d_{i_1 i_2}$  denote the RMSD between any two proteins  $i_1$  and  $i_2$ . Dahl (2008) defines a random partition on the set of proteins. The prior co-clustering probabilities are modified based on  $d_{i_1 i_2}$ . Let  $d^*$  denote the maximum recorded distance. For each cluster  $S_j \subset \{1, \dots, n\}$  and protein  $i$ , Dahl (2008) defines a propensity score  $h_i(S_j) = c_i \sum_{f \in S_j} (d^* - d_{if})$ . Without loss of generality assume  $i = n + 1$ . The scores  $h_{n+1}(S_j)$  are standardized such that  $\sum_{j=1}^k h_{n+1}(S_j) = n$ . Dahl (2008) then modifies (3) to

$$p(s_{n+1} \mid s_1, \dots, s_n, d_{n+1, i}, i = 1, \dots, n) \propto \begin{cases} h_i(S_j) & j = 1, \dots, k_n \\ \alpha & j = k_n + 1. \end{cases} \quad (8)$$

The standardization of  $h_i$  to  $\sum h_{n+1} = n$  serves an important purpose. It leaves the conditional probability of opening a new cluster unchanged. Dahl (2008) recognizes that the modified Polya urn (8) is not necessarily a legitimate PPF anymore, i.e., a predictive probability function that defines a species sampling model (Pitman, 1996). But (8) does define transition probabilities for an ergodic Markov chain. Thus it implicitly defines a joint probability distribution on  $\mathbf{s}$  that is informed by the relative distances  $d_{i_1 i_2}$ , as desired.

Another interesting recent discussion is Monni and Tadesse (2008). Although they do not define a covariate-based random partition model, we still include the approach in this review since we consider it closely related. The approach can be characterized as a joint probability model for  $\rho_n$  and a latent feature corresponding to each cluster. Monni and Tadesse (2008) consider the problem of defining regression models for many response variables  $y_i, i = 1, \dots, n$  on a large set of possible covariates  $\{x_j, j = 1, \dots, p\}$ . The application is to a regression of comparative genomic hybridization (CGH) on mRNA expression levels, with  $n = 261$  and  $p = 3291$ . Their approach is based on a partition  $\rho_n$  of  $S = \{1, \dots, n\}$ . For each partitioning subset  $S_j \subset S$  they define a subset  $R_j \subset \{1, \dots, p\}$  of regressors. For  $i \in S_j$  the model includes then a normal linear regression of  $y_i$  on  $\{x_f, f \in R_j\}$ . The model is completed with a prior  $p(\rho_n, R_1, \dots, R_k)$ . The prior is constructed to favor small partitioning subsets. Let  $n_j = |S_j|$  and  $m_j = |R_j|$ . Monni and Tadesse (2008) define

$$p(\rho_n, R_1, \dots, R_k) \propto \prod_j r^{m_j n_j},$$

for some  $0 < r < 1$  parameter that controls the clustering, with smaller values favoring small clusters and a few active covariates. The model demonstrates the restrictive nature of the Polya urn scheme (2) and the endless possibilities in defining alternative random partition models – in this case jointly with the associated sets  $R_j$ . Of course the flexibility comes at a cost. We lose scalability. But this might not be a concern when there is no notion of predictive inference.

## 6 The Cluster-X Model

### 6.1 A Covariate-Dependent Random Partition Model

In Müller et al. (2008) we develop a model for random partitions with covariates that is based on the product partition model (PPM). The PPM is a model for random partitions defined by

$$p(\rho_n) \propto \prod_{j=1}^k c(S_j),$$

where  $c(\cdot)$  is known as the cohesion function. The Polya urn (2) defines a model that falls within this class by using  $c(S_j) = \alpha(|S_j| - 1)!$  Compare with the discussion following (2).

In words, the proposed model is an extension of the PPM, defined by augmenting the cohesion function  $c(S_j)$  by a factor that favors clusters with similar covariates and discourages clusters with highly dissimilar covariates. Similarity is formalized by a similarity function. The argument of the similarity function are the covariate values recorded in a given cluster. The value is a non-negative number. Specifically, let  $x_{S_j}^* = \{x_i, i \in S_j\}$  denote the set of covariate values for experimental units grouped in the  $j$ -th cluster. We define a similarity function  $g(x_{S_j}^*)$  with  $g(\cdot) \geq 0$  such that large values of  $g$  indicate high similarity of the  $x_i$ ,  $i \in S_j$ . We define

$$p(\rho_n | \mathbf{x}) \propto \prod_j c(S_j) g(x_{S_j}^*).$$

The main component of the proposed model is the similarity function. As a default we propose to define

$$g(x_{S_j}^*) = \int q(x_i | \eta_j) dq(\eta_j), \quad (9)$$

for an auxiliary probability model  $q(x | \eta)$  and a conjugate prior  $q(\eta)$ .

Let  $N(x; m, s^2)$  denote a normal distribution with location  $m$  and scale  $s$ . For example, for a continuous covariate one could use a normal kernel  $q(x_i | \eta_j) = N(x_i; \eta_j, s^2)$  and  $q(\eta) = N(0, \tau)$ . The similarity function would then be the marginal  $n_j$ -dimensional normal distribution of the set  $x_{S_j}^*$ , with larger values for more tightly grouped  $x_i$  values. We propose similar default models for categorical and count covariates using a multinomial/Dirichlet and a Poisson/gamma model. We argue that under some reasonable assumptions on  $g(\cdot)$  the default model (9) is not only a convenient default similarity function. It is the only possible form for a similarity function. The two assumptions are symmetry of  $g(\cdot)$  with respect to permutations of the indices of  $x_i$ , and  $\int g(x_{S_j}^*, x) dx = g(x_{S_j}^*)$ .

For a covariate vector with a mix of different data formats we propose to define a joint similarity function as a product of similarity functions for each covariate or subvector of covariates with common data format.

The proposed model can be considered a special case of (4), with the submodel for  $x_i$  defined by the kernel  $q(x_i | \eta_j)$  that appears in (9). But in contrast to approaches that work with an augmented response vector  $(x_i, y_i)$  the sampling and prior models are already separated by construction. See the discussion following (4).

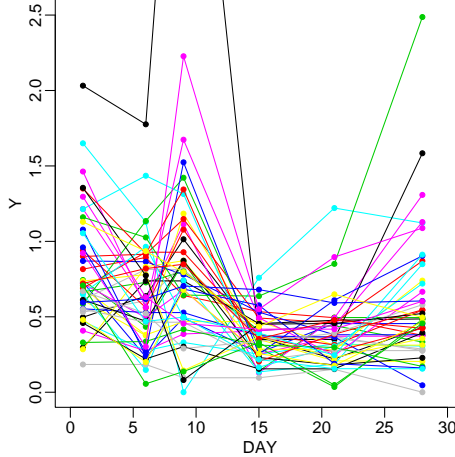


Figure 1: Data. Monocyte count versus day of the first cycle chemotherapy for  $n = 47$  patients.

The interpretation of the model as a special case of (4) is important for posterior simulation. Exploiting the representation as a special case of a DP mixture one can use any of the many posterior Markov chain Monte Carlo (MCMC) schemes that have been proposed for these models.

## 6.2 Example: Semiparametric Mixed Effects Model

Wang et al. (2005) describe a clinical trial of carboplatinum, a chemotherapy agent, with additional immunotherapy using  $\gamma$ -interferon and GM-CSF (colony stimulating factor). One of the desired effects of the immunotherapy is to boost monocyte counts. Monocytes are a subpopulation of white blood cells. All patients are administered the same level of  $\gamma$ -interferon and GM-CSF. Carboplatinum dose varies across patients. The data reports monocyte counts for  $n = 47$  patients. For each patient we use 6 repeat measurements over the first cycle of chemotherapy, using the measurements closest to days 1, 6, 9, 15, 21 and 28. Day 1 is the day of the actual chemotherapy. Day 28 is the day just before the second dose of chemotherapy. Let  $y_{ij}$  denote the  $j$ -th observation for patient  $i$ , and let  $x_i$  denote the dose of carboplatinum.

Figure 1 plots the data for the 47 patients. The initial decline is in response to the chemotherapy treatment on day 1. The first immunotherapy leads to the peak around day 9. The second increase reflects the next set of immunotherapy given just before the next chemotherapy dose  $x_i$ . We implemented the proposed model for covariate-dependent clustering to define a random partition model with a regression on the chemotherapy dose. Let  $N(x; m, s)$  indicate a normal kernel for the random variable  $x$ , with moments  $(m, s)$ , let  $W(V; \nu, S)$  denote Wishart prior for the random matrix  $V$  with scalar parameter  $\nu$  and matrix variate parameter  $S$ , and let  $\text{Ga}(s^{-1}; \nu, \nu s_o)$  denote a gamma distribution (parametrized to have mean  $s_o^{-1}$  and variance  $s_o^{-2}/\nu$ ). Let  $\theta_j = (\mu_j, V_j)$ . We used a multivariate normal

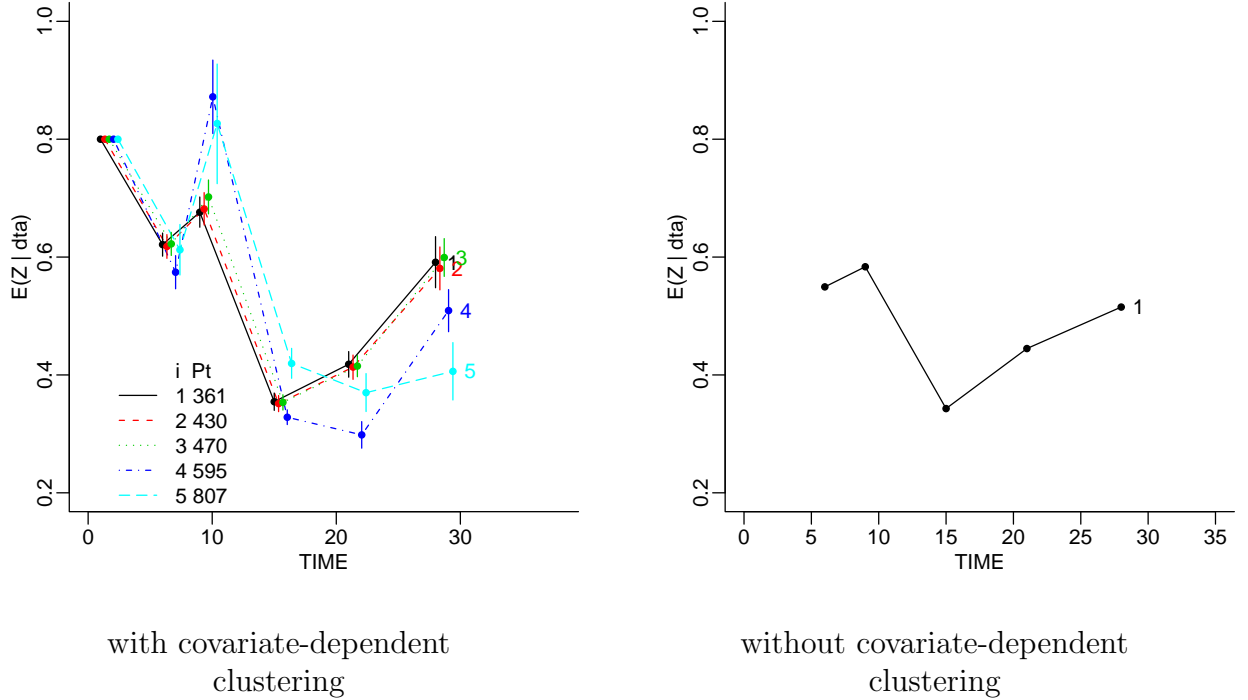


Figure 2: Prediction for  $\tilde{y}$  arranged by  $\tilde{x}$  (left panel) from lowest (“1”) to highest (“5”) level of carboplatin. Without covariates (right panel) prediction is identical for all patients.

model  $p(y_i | \theta_j) = N(\mu_j, V_j)$ , with a conditionally conjugate prior for  $\theta_j$ , i.e.,  $p(\theta_j) = N(\mu_j; m, B) W(V_j^{-1}; s, S^{-1})$ . The model is completed with conjugate hyperpriors for  $m, B$  and  $S$ . The similarity function is a normal kernel. We use

$$g(x_j^*) = \int \prod_{i \in S_j} N(x_i; m_j, s_j) N(m_j; a_0, A_0) \text{Ga}(s_j^{-1}; \nu, \nu S_0) dm_j ds_j.$$

The hyperparameters  $a_0, A_0, S_0$  are fixed.

Posterior inference in the model provides the desired prediction for a future patient,  $i = n + 1$ . Figure 2a shows the prediction arranged by dose level (with increasing doses labeled 1 through 5). For comparison, the right panel of the same figure shows posterior predictive inference in a model using a PPM prior on clustering, without the use of covariates.

For comparison, we also considered an implementation based on (4). For a fair comparison all hyperparameters were left unchanged. In particular, the prior means of the Wishart priors on all precision matrices are block diagonal with zeroes in the elements corresponding to elements of the  $x$  and  $y$  subvectors. The resulting inference is shown in Figure 3. Under the augmented probability model the number of clusters is sharply reduced. Almost all probability mass is concentrated on only one cluster (not shown). The resulting inference is thus a simple normal linear regression of the mean outcomes on the covariate and misses the complicated structure suggested by the data.

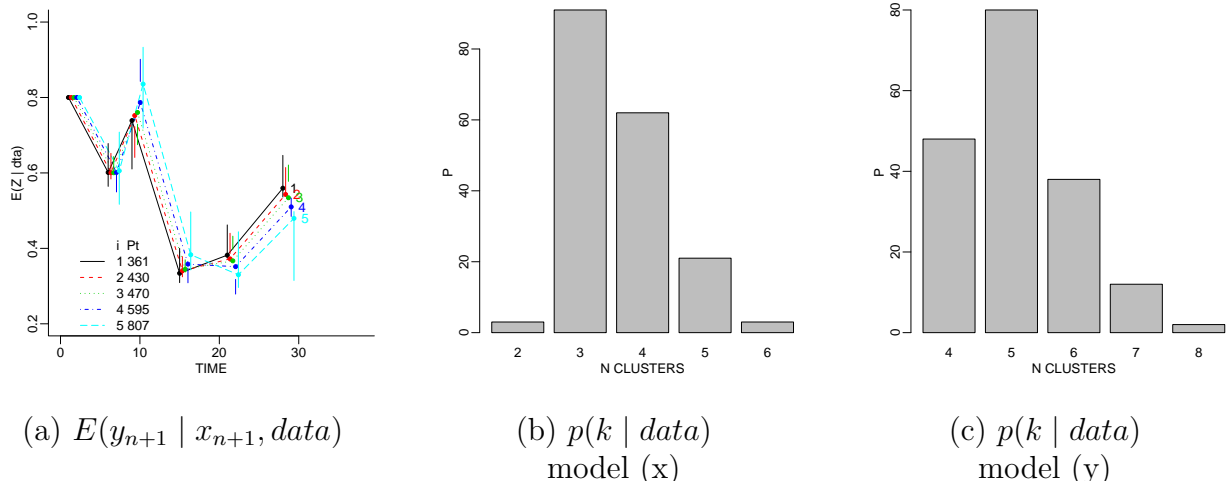


Figure 3: Augmented response vector. Panel (a) shows inference like in Figure 2a, but under a corresponding augmented response vector approach. Panels (b) and (c) show the posterior distribution on the number of clusters  $k$ .

An implementation of the model is provided in an R package that will soon be available at <http://odin.mdacc.tmc.edu/~pm/prog.html>.

## 7 Conclusion

We have discussed probability models for random clustering. We focused on the problem of defining models that can use covariate information to change *a priori* clustering probabilities. Such models are suitable for any application where similar experimental units are more likely to co-cluster. Similarity is defined in terms of available baseline covariates. We argued that several recently proposed methods for dependent RPMs and flexible regression can be exploited to define such models.

The purpose of the review was to establish the context to introduce a new method developed in Müller et al. (2008). The proposed model is discussed in section 6. It is based on an extension of the PPM model by adding an additional factor  $g(x_j^*)$  to the cohesion function. The strengths of the proposal are the conceptually clear way of introducing the desired covariate-dependence, the possibility to include long lists of covariates with a mix of different data formats, and finally the still easy implementation.

The main limitations are (i) the fact that the similarity function can implicitly add a penalty for the cluster size, (ii) the need to elicit parameters for the similarity function, and (iii) the close resemblance to simpler augmented response models.

From a data analysis point of view it is desirable that a cluster size penalty should only be introduced through the cohesion function  $c(S_j)$ . While mathematically possible, it is undesirable to include an additional cluster size penalty in the similarity function. The same statement in different words, for all equal covariates  $x_i \equiv x$ , the model should reduce to the underlying PPM. But this is not the case. It is only approximately true. See Müller et al.

(2008) for more discussion.

The elicitation of parameters for the similarity function is a problem and a feature at the same time. On one hand one always wants to limit the amount of required prior elicitation. On the other hand such prior elicitation can prompt the investigator to clearly articulate available prior information. As an opt out it is always possible to use default choices. Alternatively one could extend the model with hyperpriors on these parameters. We recommend against the latter.

In conclusion, we feel that there are important gaps in the literature for random partition models. The current practice is almost entirely dominated by the Polya urn model that is implied by the DP prior. The reason is usually purely technical convenience. The present discussion of covariate-based random partition models identifies one important direction of generalization. Another important research direction is the consideration of alternative underlying models.

## Acknowledgments

Peter Müller was partially supported by Grant NIH/NCI R01CA75981. Fernando Quintana was partially supported by Grant Fondo Nacional de Desarrollo Científico y Tecnológico Fondecyt 1060729.

## References

- Antoniak, C. E. (1974), “Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems,” *The Annals of Statistics*, 2, 1152–1174.
- Barry, D. and Hartigan, J. A. (1993), “A Bayesian analysis for change point problems,” *Journal of the American Statistical Association*, 88, 309–319.
- Bishop, C. M. and Svensén, M. (2003), “Bayesian Hierarchical Mixtures of Experts,” in *2003 Proceedings Nineteenth Conference on Uncertainty in Artificial Intelligence*, eds. Kjaerulff, U. and Meek, C., pp. 57–64.
- Chung, Y. and Dunson, D. B. (2008), “Nonparametric Bayes Conditional Distribution Modeling with Variable Selection,” Tech. rep., Department of Statistical Science, Duke University.
- 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.
- Dasgupta, A. and Raftery, A. E. (1998), “Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering,” *Journal of the American Statistical Association*, 93, 294–302.

- Dunson, D. B. and Park, J.-H. (2008), “Kernel stick-breaking processes,” *Biometrika*, 95, 307–323, doi:10.1093/biomet/asn012.
- Dunson, D. B., Pillai, N., and Park, J.-H. (2007), “Bayesian Density Regression,” *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 69, 163–183.
- Ferguson, T. S. (1973), “A Bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, 1, 209–230.
- Griffin, J. E. and Steel, M. F. J. (2006), “Order-based Dependent Dirichlet Processes,” *Journal of the American Statistical Association*, 101, 179–194.
- Hartigan, J. A. (1990), “Partition Models,” *Communications in Statistics, Part A – Theory and Methods*, 19, 2745–2756.
- Jordan, M. and Jacobs, R. (1994), “Hierarchical Mixtures-of-Experts and the EM Algorithm,” *Neural Computation*, 6, 181–214.
- MacEachern, S. N. (1999), “Dependent Nonparametric Processes,” in *ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association*.
- Monni, S. and Tadesse, M. (2008), “A Stochastic Partitioning Method to Associate High-dimensional Responses and Covariates,” Tech. rep., Department of Biostatistics and Epidemiology, University of Pennsylvania.
- Müller, P., Muller, P., Mueller, P., Erkanli, A., and West, M. (1996), “Bayesian Curve Fitting Using Multivariate Normal Mixtures,” *Biometrika*, 83, 67–79.
- Müller, P., Quintana, F., and Rosner, G. (2008), “Bayesian Clustering with Regression,” Tech. rep., M.D. Anderson Cancer Center, Houston, TX.
- Pitman, J. (1996), “Some Developments of the Blackwell-MacQueen Urn Scheme,” in *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, eds. Ferguson, T. S., Shapeley, L. S., and MacQueen, J. B., Hayward, California: IMS Lecture Notes - Monograph Series, pp. 245–268.
- Quintana, F. A. (2006), “A predictive view of Bayesian clustering,” *Journal of Statistical Planning and Inference*, 136, 2407–2429.
- Sethuraman, J. (1994), “A constructive definition of Dirichlet priors,” *Statistica Sinica*, 4, 639–650.
- Shahbaba, B. and Neal, R. M. (2007), “Nonlinear Models Using Dirichlet Process Mixtures,” Tech. Rep. 0707, Department of Statistics, University of Toronto.

Wang, E., Ngalame, Y., Panelli, M. C., Nguyen-Jackson, H., Deavers, M., Müller, P., Hu, W., Savary, C. A., Kobayashi, R., Freedman, R. S., and Marincola, F. M. (2005), “Peritoneal and subperitoneal stroma may facilitate regional spread of ovarian cancer,” *Clinical Cancer Research*, 11, 113–122.

Wood, S. A., Kohn, R., Cottet, R., Jiang, W., and Tanner, M. (2008), “Locally Adaptive Nonparametric Binary Regression,” *Journal of Computational and Graphical Statistics*, 17, 352–372.