Journal of Statistical Planning and Inference 140 (2010) 2801–2808

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# Random partition models with regression on covariates

Peter Müller [a,*], Fernando Quintana [b]

[a] M.D. Anderson Cancer Center, Houston, TX, USA
[b] Pontificia Universidad Catolica, Santiago, Chile

ARTICLE INFO

*Article history:*
Received 1 September 2008
Accepted 3 October 2009
Available online 7 March 2010

*MSC:*
62G08
62C10
60G57

*Keywords:*
Clustering
Non-parametric Bayes
Product partition model

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 $S=\{1,\ldots,n\}$ denote a set of experimental units. A partition is a family of subsets $S_1,\ldots,S_k$ with $S=S_1\cup\ldots\cup S_k$, $S_j\cap S_{j'}=\emptyset$. We write $\rho_n=\{S_1,\ldots,S_k\}$. 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 $i\in S_j$, $i=1,\ldots,n$. Let $\mathbf{s}_n=(s_1,\ldots,s_n)$. 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 $\pi=(\pi_1,\ldots,\pi_n)$ denote a permutation of $S$, and let $\mathbf{s}_\pi=(s_{\pi_1},\ldots,s_{\pi_n})$ describe the clusters implied by re-labeling experimental unit $i$ by $h=\pi_i^{-1}$, i.e., $\pi_h=i$. We require

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

* Corresponding author.
*E-mail address:* pmueller@mdanderson.org (P. Müller).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.03.002

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

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

We refer to these two properties as symmetry and scalability. 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 $p(n_{n1},\ldots,n_{nk})$.

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 $\mathbf{y} = (y_1,\ldots,y_n)$ 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 $\xi_j$:

$$p(\mathbf{y}|\rho_n) = \prod_{j=1}^{k} \int \prod_{i \in S_j} p(y_i|\xi_j^\star)\, \mathrm{d}p(\xi_j^\star). \tag{1}$$

For example, $p(y_i|\xi_j^\star)$ could be a normal model $N(\xi_j^\star, S)$, and the prior $p(\xi_j^\star)$ 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. 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 the random partition implied by the Dirichlet process (DP) prior (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 $G \sim \mathrm{DP}(\alpha, G^\star)$. The base measure parameter $G^\star$ defines the prior mean, $E(G) = G^\star$. 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, $\xi_i \sim G, i = 1,\ldots,n$. The discrete nature of $G$ implies positive probabilities for ties among the $\xi_i$. Let $\{\xi_1^\star,\ldots,\xi_k^\star\}$ denote the unique values among the $\xi_i$ and define $S_j = \{i : \xi_i = \xi_j^\star\}$. The implied probability model on $\rho_n = (S_1,\ldots,S_k)$ is known as the Polya urn scheme. Let $[x]_m = x \cdot (x+1) \cdot \ldots \cdot (x+m-1)$ denote the Pochhammer symbol. The Polya urn defines

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

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

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

It is easily verified that the Polya urn defines indeed an EPPF. Models that are characterized by a sequence of PPFs $\{p_j(\rho_n), j = 1,\ldots,k \text{ and } n = 1,2,\ldots\}$ 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 three-dimensional structure. Structure is recorded as a sequence of seven 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 $\mathbf{x}_n = (x_1,\ldots,x_n)$. 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 2–5 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.