Chapter 8

Random Partition Models

8.1. Introduction

In earlier chapters we discussed nonparametric Bayesian priors $p(G)$ for random probability measures. The most commonly used model is the DP prior and its variations and extensions. One of the many interesting properties of the DP is the almost sure discrete nature of a random probability measure G with DP prior, $G \sim \mathsf{DP}(M, G_0)$. The discrete nature of G naturally induces a prior on random partitions, as we have seen many times before in earlier chapters. Consider a random sample, $x_i \mid G \sim G$, $i = 1, \ldots, n$, generated from a probability model with DP prior, $G \sim \mathsf{DP}(M, G_0)$. The discreteness of G implies a positive probability of ties among the x_i. We can use these ties to define a partition of the experimental units $\{1, \ldots, n\}$ as

$$\{1, \ldots, n\} = \bigcup_{j=1}^{k} \underbrace{\{i : \; x_i = x_j^\star\}}_{S_j}$$

defined by the unique values $x_1^\star, \ldots, x_k^\star$. In other words, the DP prior induces a prior on clusters defined by the $k \leq n$ unique values of the random sample.

Many applications of nonparametric Bayesian models focus on this implied clustering. The inference on the unknown probability measure G is often of less interest than the implied clustering. In this chapter we focus on this aspect of nonparametric Bayes models and introduce some alternative models for random partitions. We start by introducing useful notation. Let $S = \{1, \ldots, n\}$ denote the experimental units that are being clustered. Let $\rho_n = \{S_1, \ldots, S_k\}$ denote a partition with non-overlapping subsets S_j that cover S. When the sample size n is obvious from the context we drop the subindex $_n$. Sometimes it is technically more convenient to use alternative equivalent notation with cluster membership indicators $s_i = j$ if $i \in S_j$. Let $y_j^\star = (y_i, \; i \in S_j)$ denote outcomes arranged by clusters. For some models we will make use of available covariates x_i and use $x_j^\star$ to denote covariates arranged by clusters. In this chapter we discuss probability models $p(\rho_n)$ for random partitions and extensions of such models that include a regression on covariates by defining $p(\rho_n \mid \boldsymbol{x})$.

8.2. Random Partition Models

Product Partition Model

Hartigan (1990), Barry and Hartigan (1993), and Crowley (1997) propose and develop the product partition model (PPM) for random partitions. In contrast to the prior on clustering that is implied by the DP prior, the PPM explicitely defines a probability distribution $p(\rho_n)$ over alternative partitions. The PPM uses a

non-negative function $c(S_j)$, known as the cohesion function to define a product partition probability

$$p(\rho_n) = K \prod_{j=1}^{k} c(S_j). \tag{8.1}$$

Conditional on a given partition, the PPM assumes independent sampling across clusters,

$$p(\boldsymbol{y} \mid \rho) = \prod_j p(y_j^\star \mid \mu_j^\star), \tag{8.2}$$

where $\mu_j^\star$ are cluster specific parameters. Applications of the PPM often use exchangeability of y_i across $i \in S_j$ by assuming that y_i, $i \in S_j$ are independent given $\mu_j^\star$. One of the attractions of the PPM is the conjugate nature. The posterior $p(\rho_n \mid \boldsymbol{y})$ is again a product partition model, with updated cohesion functions $c(S_j)p(y_j^\star)$, where $p(y_j^\star)$ is the marginal sampling model for y_i, $i \in S_j$ under partition ρ_n.

The Pólya urn implied by the DP prior, $\mathsf{DP}(M, G_0)$, is a special case of a PPM, with cohesion function $c(S_j) = M(n_j - 1)!$. Another example of a PPM are Gibbs type priors. Recall from §7.3 the format of the EPPF for Gibbs type priors. Let $n_j = |S_j|$ denote the cardinality of the $j-$th cluster. The EPPF of a Gibbs type prior takes the form (8.1) with cohesion function $c(S_j) = (1-\sigma)_{n_j-1}$. Here $a_k = \Gamma(a+k)/\Gamma(a)$ denotes a rising factorial.

Some applications use constrained partition models. For example, when observations are ordered in time it might be desireable to restrict clusters to contiguous sequences of objects (Barry and Hartigan, 1993; Monteiro *et al.*, 2010; Yao, 1984).

Species Sampling Model

We already discussed the species sampling model (SSM) as a large class of prior models for random distributions (Ishwaran and James, 2003; Pitman, 1996) that includes many popular models as special cases. One of the characterizations of the SSM is through the EPPF, the implied probability model for the induced partition of $\{1, \ldots, n\}$. Recall that the EPPF is a symmetric function $f(n_1, \ldots, n_k)$, symmetric in its arguments,

$$p(\rho_n) = f(\boldsymbol{n}).$$

Again the partition that is implied by i.i.d. sampling from a random probability measure with DP prior is a special case with $f(\boldsymbol{n}) \propto \prod_{j=1}^{n} M(n_j - 1)!$.

Model-Based Clustering

In data analysis, when formal probability models are used for clustering, perhaps the most commonly used approach is model-based clustering. Model-based clustering defines a prior $p(\rho_n)$ implicitly through a mixture model for the observed data. Let y_i, $i = 1, \ldots, n$, denote responses for n experimental units. A mixture model $p(y_i \mid k, (\theta_j), (\pi_j)) = \sum_{j=1}^{k} \pi_j f_j(y_i \mid \theta_j)$ can be equivalently written as a hierarchical model with latent indicators $s_i \in \{1, \ldots, k\}$,

$$p(y_i \mid k, (\theta_j), s_i = j) = f_j(y_i \mid \theta_j), \qquad \mathsf{Pr}(s_i = j) = \pi_j. \tag{8.3}$$

When the latent indicators s_i are interpreted as cluster membership indicators, then (8.3) implicitely defines $p(\rho_n)$. Inference for such models is discussed, among others, in Fraley and Raftery (2002), Richardson and Green (1997) and Green and Richardson (2001).

Pólya Urn

Recall the predictive rule for cluster allocation under i.i.d. sampling $x_i \mid G \sim G$ from a random probability measure G with a DP prior, $G \sim \mathsf{DP}(M, G_0)$. Let $s_i = j$ when the i-th observation is equal to the j-th unique value, i.e., when $x_i = x_j^\star$. The Pólya urn (Chinese restaurant process) specifies

$$p(s_{n+1} \mid s_1, \ldots, s_n) = \begin{cases} n_h & \text{with prob } 1/(M+n) \\ k_n + 1 & \text{with prob } M/(M+n). \end{cases} \tag{8.4}$$

The prior $p(\rho_n)$ implied by (7.1) is a special case of the PPM, a special case of the SSM, as well as a special limiting case of model-based clustering. We already mentioned the earlier two special cases. The Pólya urn arises as a limiting case of model-based clustering when $p(\pi_1, \ldots, \pi_k)$ is assumed as a symmetric Dirichlet distribution, $\mathsf{Dir}(\delta, \ldots, \delta)$ and one considers the limiting case $\delta \to 0$ and $k \to \infty$ subject to $k\delta \to M$ (Green and Richardson, 2001). The nature of the DP as a special case of many other models is one of the reasons for the undying popularity of the model.

8.3. Covariate-Dependent Clustering

Covariate-Dependent PPM

The previously discussed clustering models are useful for inference about clusters and subpopulations in observed data, but of little use for predictive inference. In Example 24 we are interested in predicting overall survival time y_{n+1} for a future patient $i = n + 1$ on the basis of data for $n = 763$ patients in a clinical trial. Let $\boldsymbol{y} = (y_1, \ldots, y_n)$. Clustering patients on the basis of the outcome would allow us to predict survival time for a future patient in the same population of patients who were eligible for this trial as

$$p(y_{n+1} \mid \boldsymbol{y}) = \int p(y_{i+1} \mid s_{n+1}, \rho_n, \boldsymbol{y}) dp(s_{n+1} \mid \rho_n) dp(\rho_n \mid \boldsymbol{y}), \tag{8.5}$$

where integration with respect to s_{n+1} is simply averaging over the $k_{n_1} + 1$ possible choices and integration with respect to ρ_n is averaging with respect to the posterior distribution on possible cluster arrangements of the first n patients. This is density estimation for the survival time of women in this population. We would report the same inference for any future patient, independently of the patient's baseline characteristics. This limits the use of (8.5) for prediction in this scenario.

More relevant would be inference of overall survival for a woman with particular baseline covariates x_i. A convenient and often used implementation is to consider an augmented outcome vector $\mathbf{z}_i = (y_i, x_i)$, implement clustering on the basis of $\mathbf{z}_i$ and report

$$p(y_{n+1} \mid x_{n+1}, \boldsymbol{y}, \boldsymbol{x}) \tag{8.6}$$

as the desired inference. The problem is that the covariate vector x_i often involves a mix of data formats, complicating the specification of a sampling model. Also, some of the covariates such as treatment assignment are not random at all, making it awkward to model a distribution for these variables.

Müller *et al.* (2011) propose to instead use a model $p(\rho_n \mid \boldsymbol{x})$, together with a sampling model $p(\boldsymbol{y} \mid \rho_n)$. Here $p(\rho_n \mid \boldsymbol{x})$ is a regression of the random partition ρ_n on the known covariates $\boldsymbol{x}$. The idea is to specify a probability model for random partitions that favors clusters that are homogeneous in the covariates x_i. Predicting the outcome for a future subject is then based on averaging over all clusters, with the weights determined by the respective probability of cluster membership $p(s_{n+1} \mid x_{n+1}, \rho_n, \boldsymbol{x})$. In words, the prediction weighs clusters of earlier patients with similar covariates higher than others.

Formally, let $x_j^\star = (x_i,\ i \in S_j)$ denote covariates of experimental units in the j-th cluster, and let $g(x^\star)$ denote a non-negative function that formalizes homogeneity of a cluster with covariates $x^\star$. For example, $g(x^\star)$ could be the determinant of the empirical precision matrix of the x_i. For a categorical covariate x the similarity function could be related to the number of distinct values in a cluster. For example, a cluster with all women with the same prior treatment history is clinically more meaningful than a cluster that includes a large diversity of prior treatment histories. A simple application of the PPM provides the desired random partition model

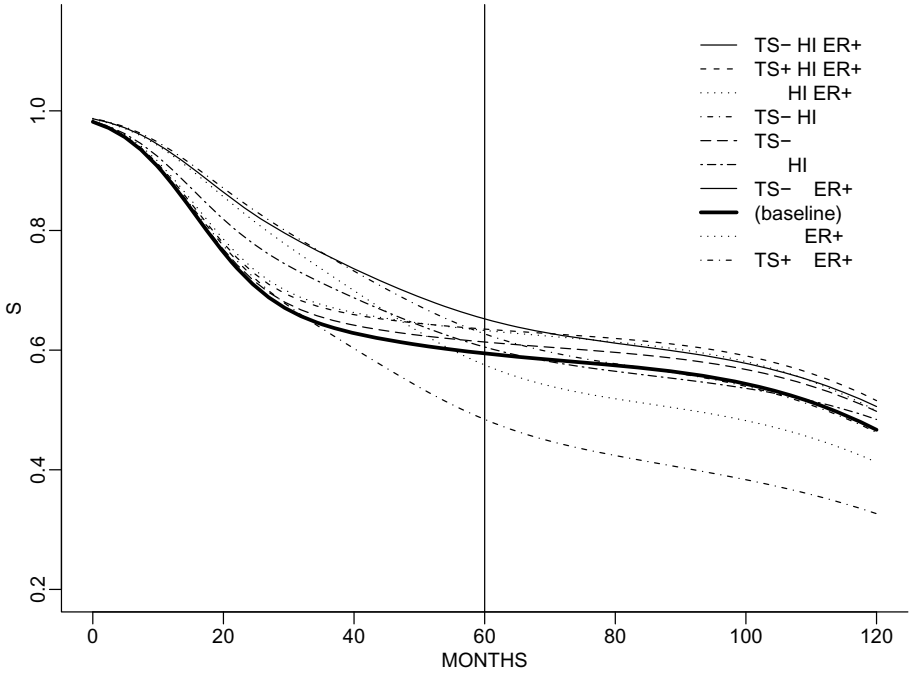
$$p(\rho_n \mid \boldsymbol{x}) \propto \prod_{j=1}^{k} g(x_j^\star) c(S_j). \tag{8.7}$$

The choice of the similarity function depends on the application. As a generic choice, Müller *et al.* (2011) define $g(x^\star)$ on the basis of an auxiliary probability model $q(\cdot)$:

$$g(x_j^\star) \equiv \int \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j) d\xi_j.$$

Chosing $q(x_i \mid \xi_j)$ and $q(\xi_j)$ as a conjugate pair simplifies analytic evaluation of $g(x^\star)$.

Example 24 (Survival Time Model with Clustering) *Müller* et al. *(2011) consider data from a high-dose chemotherapy treatment of $n = 763$ women with breast cancer. The response of interest is overall survival y_i. Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote the observed data. There are six patient-specific covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i6})$, including a binary indicator for high dose chemotherapy, age in years, number of positive lymph nodes, tumor size, indicator for estrogen or progesterone receptor positive tumor, and an indicator for the woman's menopausal status. indicator. Let $x_{j,\ell}^\star = (x_{i\ell}, i \in S_j)$ denote the values for the ℓ-th covariate in cluster j. Müller* et al. *(2011) define a similarity function $g(x_j^\star) = \prod_{\ell=1}^{6} g_\ell(x_{j,\ell}^\star)$ using default similarity function g_ℓ for each data format, including a beta-binomial for the binary covariates, a normal-normal for continuous covariates and a poisson-gamma model for the count covariate. Figure 8.1 summarizes prediction for a future patient as a function of baseline covariates.*



$S(t \mid x)$

FIG 8.1. *Posterior predictive summarized by survival functions* $S(t \mid x) \equiv p(y_{n+1} \geq t \mid x_{n+1} = x, data)$. *In the legend TS− and TS+ indicates tumor size equal to the first and third empirical quantile, ER+ indicates ER-positive tumor, and HI indicates high dose.*

.

Alternative Constructions

Model-based clustering (8.3) allows an easy extension to include covariates in the implied prior on random partitions. Consider

$$y_i \mid x_i \sim \sum_{j=1}^{k} \pi_j(x_i;\, \alpha_j)\, f_j(\cdot \mid \theta_j).$$

The generalization is the explicit inclusion of covariates in the weights of the component models. As before, rewriting the mixture as a hierarchical model with latent indicators defines the desired covariate-dependent random partition model:

$$p(y_i \mid k, (\theta_j), s_i = j) = f_j(y_i \mid \theta_j), \qquad p(s_i = j \mid x_i) = \pi_j(x_i;\, \alpha_j). \tag{8.8}$$

The regression $\pi_j(x_i;\, \alpha_j)$ could be, for example, a logistic regression. This is essentially the hierarchical mixture of experts model (Bishop and Svensén, 2003; Jordan and Jacobs, 1994). The model is very useful for flexible non-parametric regression, especially when the focus is prediction. The limiations of the approach are the use of a fixed number of component models k, which becomes an upper bound for the number of clusters, and the restriction of covariate dependence to the particular parametric form chosen for $\pi_j(x_i;\, \alpha_j)$.

Dahl (2008) defines another interesting probability model for covariate-based clustering. Let $s_{-i} = (s_\ell,\, \ell \neq i)$ denote the partition of all but the i-th object. He defines the desired $p(\rho_n \mid \boldsymbol{x})$ by modifying the complete conditional probabilities $p(s_i \mid s_{-i}, \boldsymbol{x})$. The modification of complete conditionals needs care to assure the existence of a well defined probability model.

Clustering with DDP and Related Models

We earlier introduced the dependent DP model as prior for families of random probability measures $\mathcal{G} = \{G_x,\ x \in X\}$. In particular, recall the definition of the DDP, here with common locations and variable weights

$$G_x = \sum_{h=1}^{\infty} \pi_{hx} \delta_{m_h}. \tag{8.9}$$

Note that here the locations m_h of the point masses are common across x, and only the weights vary.

Similar to how the DP induces implicitely a prior for random partitions $p(\rho_n)$, the DDP can be used to implicitely define a prior $p(\rho_n \mid \boldsymbol{x})$ for random partitions with a regression on covariates. In particular, assume $y_i \mid x_i = x, \mathcal{G} \sim G_x$, independently for experimental units $i = 1, \ldots, n$, with known covariates x_i. In addition, let $y_j^\star$, $j = 1, \ldots, k$ denote the $k \leq n$ unique values among the y_i and define clusters $S_j = \{i :\ y_i = y_j^\star\}$. The construction is almost identical to before, when we used the DP to define a random partition. However, the probabilities for cluster membership now depend on x, as desired. To our knowledge, this construction itself, i.e., the clustering implied by sampling from a DDP family of random probability measures, has not be used in the literature before. However, several proposed approaches can be interpreted as approximations to this natural construction. Note that other variants of dependent stick-breaking prior such as the probit stick-breaking process (see §5.7.1) and the kernel stick-breaking process (see §5.7.2) can be used to generate prior on dependence partitions in a similar fashion.