A Product Partition Model With Regression on Covariates

Peter MÜLLER, Fernando QUINTANA, and Gary L. ROSNER

We propose a probability model for random partitions in the presence of covariates. In other words, we develop a model-based clustering algorithm that exploits available covariates. The motivating application is predicting time to progression for patients in a breast cancer trial. We proceed by reporting a weighted average of the responses of clusters of earlier patients. The weights should be determined by the similarity of the new patient's covariate with the covariates of patients in each cluster. We achieve the desired inference by defining a random partition model that includes a regression on covariates. Patients with similar covariates are a priori more likely to be clustered together. Posterior predictive inference in this model formalizes the desired prediction.

We build on product partition models (PPM). We define an extension of the PPM to include a regression on covariates by including in the cohesion function a new factor that increases the probability of experimental units with similar covariates to be included in the same cluster. We discuss implementations suitable for any combination of continuous, categorical, count, and ordinal covariates.

An implementation of the proposed model as R-package is available for download.

Key Words: Clustering; Nonparametric Bayes; Variable selection.

1. INTRODUCTION

We develop a probability model for clustering with covariates, that is, a probability model for partitioning a set of experimental units, where the probability of any particular partition is allowed to depend on covariates. The motivating application is inference in a clinical trial. The outcome is time to progression for breast cancer patients. The covariates include treatment dose, initial tumor burden, an indicator for menopause, and more. We wish to define a probability model for clustering patients with the specific feature that patients with equal or similar covariates should be a priori more likely to co-cluster than others.

Peter Müller is Professor, Department of Biostatistics, M. D. Anderson Cancer Center, Houston, TX 77230-1402 (E-mail: *pmueller@math.utexas.edu*). Fernando Quintana is Profesor Titular, Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile. Gary L. Rosner is Director, Division of Oncology Biostatistics, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD 21205-2013.

Let $i = 1, \ldots, n$, index experimental units, and let $\rho_n = \{S_1, \ldots, S_{k_n}\}$ denote a partition of the n experimental units into k_n subsets S_j. Let x_i and y_i denote the covariates and response reported for the ith unit. Let $x^n = (x_1, \ldots, x_n)$ and $y^n = (y_1, \ldots, y_n)$ denote the entire set of recorded covariates and response data, and let $x_j^* = (x_i, i \in S_j)$ and $y_j^* = (y_i, i \in S_j)$ denote covariates and response data arranged by clusters. Sometimes it is convenient to introduce cluster membership indicators $e_i \in \{1, \ldots, k_n\}$ with $e_i = j$ if $i \in S_j$, and use $(k_n, e_1, \ldots, e_n)$ to describe the partition. We call a probability model $p(\rho_n)$ a clustering model, excluding in particular purely constructive definitions of clustering as a deterministic algorithm (without reference to probability models). In particular, the number of clusters k_n is itself unknown. The clustering model $p(\rho_n)$ implies a prior model $p(k_n)$. Many clustering models include, implicitly or explicitly, a sampling model $p(y^n | x^n, \rho_n)$. Probability models $p(\rho_n)$ and inference for clustering have been extensively discussed over the past few years. See the article of Quintana (2006) for a recent review. In this article we are interested in adding a regression to replace $p(\rho_n)$ with $p(\rho_n | x^n)$.

We focus on the product partition model (PPM). The PPM (Hartigan 1990; Barry and Hartigan 1993; Crowley 1997) constructs $p(\rho_n)$ by introducing cohesion functions $c(A) \geq 0$ for $A \subseteq \{1, \ldots, n\}$ that measure how tightly grouped the elements in A are thought to be, and defines a probability model for a partition ρ_n and data y^n as

$$p(\rho_n) \propto \prod_{j=1}^{k_n} c(S_j) \qquad \text{and} \qquad p(y^n | \rho_n) = \prod_{j=1}^{k_n} p_j(y_j^\star). \tag{1.1}$$

Model (1.1) is conjugate. The posterior $p(\rho_n | y^n)$ is again in the same product form.

Alternatively, the species sampling model (SSM) (Pitman 1996; Ishwaran and James 2003) defines an exchangeable probability model $p(\rho_n)$ that depends on ρ_n only indirectly through the cardinality of the partitioning subsets, $p(\rho_n) = p(|S_1|, \ldots, |S_{k_n}|)$. The SSM can be alternatively characterized by a sequence of predictive probability functions (PPFs) that describe how individuals are sequentially assigned to either already formed clusters or to start new ones. The choice of the PPF is not arbitrary. One has to make sure that a sequence of random variables that are sampled by iteratively applying the PPF is exchangeable. The popular Dirichlet process (DP) model (Ferguson 1973) is a special case of a SSM. Moreover, the marginal distribution that a DP induces on partitions is also a PPM with cohesions $c(S_j) = M(|S_j| - 1)!$ (Quintana and Iglesias 2003). Here M denotes the total mass parameter of the DP prior.

Model-based clustering (Banfield and Raftery 1993; Dasgupta and Raftery 1998; Fraley and Raftery 2002) implicitly defines a probability model on clustering by assuming a mixture model $p(y_i | \theta, \tau, k_n) = \sum_{j=1}^{k_n} \tau_j p_j(y_i | \theta_j)$, where $\theta = (\theta_1, \ldots, \theta_{k_n})$ and $\tau = (\tau_1, \ldots, \tau_{k_n})$. Together with a prior $p(k_n)$ on k_n and $p(\theta, \tau | k_n)$, the mixture implicitly defines a probability model on clustering. Consider the equivalent hierarchical model

$$p(y_i | e_i = j, k_n, \theta, \tau) = p_j(y_i | \theta_j) \qquad \text{and} \qquad \Pr(e_i = j | k_n, \theta, \tau) = \tau_j. \tag{1.2}$$

The implied posterior distribution on $(e_1, \ldots, e_n)$ and k_n defines a probability model on ρ_n. Richardson and Green (1997) developed posterior simulation strategies for mixtures of

normal models. Green and Richardson (1999) discussed the relationship to DP mixture models.

In this article we build on the PPM (1.1) to define a covariate-dependent random partition model by augmenting the PPM with an additional factor that induces the desired dependence on the covariates. We refer to the additional factor as *similarity function*. Similar approaches are discussed by Shahbaba and Neal (2009) and Park and Dunson (2010). Both effectively included the covariates as part of an augmented response vector. We discuss more details of these two and other alternative approaches in Section 5 where we also carry out a small Monte Carlo study for an empirical comparison.

In Section 2 we state the proposed model and considerations in choosing the similarity function. In Section 3 we show that the computational effort of posterior simulation remains essentially unchanged from PPM models without covariates. In Section 4 we propose specific choices of the similarity function for common data formats. Section 5 reviews some alternative approaches, and we summarize a small Monte Carlo study to compare some of these approaches. In Section 6 we show a simulation study and a data analysis example.

2. THE PPMX MODEL

We build on the PPM (1.1), modifying the cohesion function $c(S_j)$ with an additional factor that achieves the desired regression. Recall that $x_j^* = (x_i, i \in S_j)$ denotes all covariates for units in the jth cluster. Let $g(x_j^*)$ denote a nonnegative function of x_j^* that formalizes similarity of the x_i with larger values $g(x_j^*)$ for sets of covariates that are judged to be similar. We define the model

$$p(\rho_n | x^n) \propto \prod_{j=1}^{k_n} g(x_j^*) \cdot c(S_j) \tag{2.1}$$

with the normalization constant $g_n(x^n) = \sum_{\rho_n} \prod_{j=1}^{k_n} g(x_j^*) c(S_j)$. By a slight abuse of notation we include x behind the conditioning bar even when x is not a random variable. We will later discuss specific choices for the similarity function g. As a default choice we propose to define $g(\cdot)$ as the marginal probability in an auxiliary probability model q, even if x_i are not considered random,

$$g(x_j^\star) \equiv \int \prod_{i \in S_j} q(x_i | \xi_j) q(\xi_j) \, d\xi_j. \tag{2.2}$$

There is no notion of the x_i being random variables. But the use of a probability density $q(\cdot)$ for the construction of $g(\cdot)$ is convenient since it allows for easy calculus. The correlation that is induced by the cluster-specific parameters ξ_j in (2.2) leads to higher values of $g(x_j^\star)$ for tightly clustered, similar x_i, as desired. More importantly, we show below that under some minimal assumptions a similarity function $g(\cdot)$ necessarily is of the form (2.2). The function (2.2) satisfies the following two properties that are desirable for a similarity function in (2.1). First, we require symmetry with respect to permutations of the sample indices i. The probability model must not depend on the order of introducing the

experimental units. This implies that the similarity function $g(\cdot)$ must be symmetric in its arguments. Second, we require that the similarity function scales across sample size, in the sense that $g(x^\star) = \int g(x^\star, x)\, dx$. In words, the similarity of any cluster is the average of any augmented cluster.

Under these two requirements (2.2) is not only technically convenient. It is the only possible similarity function that satisfies these two constraints.

Theorem. *Assume that* (i) *a similarity function* $g(x^\star)$ *satisfies the two constraints;* (ii) $g(x_j^\star)$ *integrates over the covariate space,* $\int g(x_j^\star)\, dx_j^\star < \infty$. *Then* $g(x_j^\star)$ *is proportional to the marginal distribution on* $x_j^\star$ *under a hierarchical auxiliary model* (2.2).

Proof: The result is a direct consequence of De Finetti's representation theorem for exchangeable probability measures, applied to a normalized version of $q(\cdot)$. See, for example, the book by Bernardo and Smith (1994, chapter 4.3). The representation theorem applies for an infinite sequence of variables that are subject to the symmetry constraint. The result establishes that all similarity functions that satisfy the symmetry constraint are of the form (2.2). □

The definition of the similarity function with the auxiliary model $q(\cdot)$ also implies another important property. The random partition model (2.1) is coherent across sample sizes. The model for the first n experimental units follows from the model for $(n + 1)$ observations by appropriate marginalization. Without covariates we would simply require $p(\rho_n) = \sum_{e_{n+1}} p(\rho_{n+1})$, with the summation defined over all possible values of e_{n+1}. With covariates we consider the following condition. Assume that the similarity function is defined by means of an auxiliary model, as in (2.2). The covariate-dependent PPM (2.1) is coherent across sample sizes as formalized by the following relationship of $p(\rho_n|x^n)$ and $p(\rho_{n+1}|x^n, x_{n+1})$:

$$p(\rho_n|x^n) = \sum_{e_{n+1}} \int p(\rho_{n+1}|x^n, x_{n+1})q(x_{n+1}|x^n)\, dx_{n+1} \qquad (2.3)$$

for the probability model $q(x_{n+1}|x^n) \propto g_{n+1}(x^{n+1})/g_n(x^n)$.

We complete the random partition model (2.1) with a sampling model that defines independence across clusters and exchangeability within each cluster. We include cluster-specific parameters θ_j and common hyperparameters η:

$$p(y^n|\rho_n, \theta, \eta, x^n) = \prod_{j=1}^{k_n} \prod_{i \in S_j} p(y_i|x_i, \theta_j, \eta) \qquad \text{and} \qquad p(\theta|\eta) = \prod_{j=1}^{k_n} p(\theta_j|\eta), \quad (2.4)$$

where $\theta = (\theta_1, \ldots, \theta_{k_n})$. We refer to (2.1) together with (2.4) as PPM with covariates, and write PPMx for short. The resulting model extends PPMs of the type (1.1) while keeping the product structure. Note that the sampling model $p(y_i|x_i, \theta_j, \eta)$ in (2.4) can include a regression of the response data on the covariate x_i. For example, Shahbaba and Neal (2009) included a logistic regression for y_i on x_i, in addition to a covariate-dependent partition model $p(\rho_n|x^n)$. Similarly, cluster-specific covariates w_j could be included by replacing $p(\theta_j|\eta)$ by $p(\theta_j|\eta, w_j)$.

3. POSTERIOR INFERENCE

3.1 MARKOV CHAIN MONTE CARLO POSTERIOR SIMULATION

A practical advantage of the proposed default choice for $g(x_j^*)$ is that it greatly simplifies posterior simulation. In words, posterior inference in the PPMx (2.1) and (2.4) is identical to the posterior inference that we would obtain if x_i were part of the random response vector y_i. The opposite is not true. Not every model with augmented response vector is equivalent to a PPMx model.

Formally, and as a computational device only, consider the auxiliary model $q(\cdot)$ defined as

$$q(y^n, x^n | \rho_n, \theta, \eta) = \prod_{j=1}^{k_n} \prod_{i \in S_j} p(y_i | x_i, \theta_j, \eta) q(x_i | \xi_j) \qquad \text{and}$$

$$q(\theta, \xi | \eta) = \prod_{j=1}^{k_n} p(\theta_j | \eta) q(\xi_j),$$

(3.1)

and replace the covariate-dependent prior $p(\rho_n | x^n)$ by the PPM $q(\rho_n) \propto \prod_{j=1}^{k_n} c(S_j)$. The posterior distribution $q(\rho_n | y^n, x^n)$ under the auxiliary model is identical to $p(\rho_n | y^n, x^n)$ under the proposed model (2.4). But posterior simulation under $q(\cdot)$ can be carried out using standard techniques, for example, as discussed by Quintana (2006). An important caveat of this operational trick, though, is that ξ_j and θ_j must be a priori independent. In particular, the prior on θ_j and $q(\xi_j)$ must not include common hyperparameters. If we had $p(\theta_j | \eta)$ and $q(\xi_j | \eta)$ depend on a common hyperparameter η, then the posterior distribution under the auxiliary model (3.1) would differ from the posterior under the original model. Let $q(x^n | \eta) = \sum_{\rho_n} \prod_j \int q(x_j^\star | \xi_j) \, dq(\xi_j | \eta)$. The two posterior distributions would differ by a factor $q(x^n | \eta)$. The implication on the desired inference can be substantial.

However, we do not consider the principled way of introducing the covariates in the PPM to be the main feature of the PPMx. Rather, we argue that the proposed model greatly simplifies the inclusion of covariates including a variety of data formats. It would be unnecessarily difficult to attempt the construction of a joint model for continuous responses and categorical, binary, and continuous covariates. We argue that it is far easier to focus on a modification of the cohesion function. This is illustrated in the data analysis example in Section 6. From a modeling perspective it is desirable to separate the choice of the prior on the partition versus the sampling model conditional on an assumed partition.

3.2 PREDICTIVE INFERENCE

A minor complication arises with posterior predictive inference, that is, when reporting $p(y_{n+1} | y^n, x^n, x_{n+1})$. Using $\tilde{x} = x_{n+1}$, $\tilde{y} = y_{n+1}$, and $\tilde{e} = e_{n+1}$ to simplify notation, we find $p(\tilde{y} | \tilde{x}, x^n, y^n) = \int p(\tilde{y} | \tilde{x}, \rho_{n+1}, x^n, y^n) \, dp(\rho_{n+1} | \tilde{x}, y^n, x^n)$. The integral is simply a sum over all configurations ρ_{n+1}. But it is not immediately recognizable as a posterior integral with respect to $p(\rho_n | x^n, y^n)$. This can easily be overcome by an importance sampling

reweighting step. Let $g(\emptyset) = c(\emptyset) \equiv 1$. The prior on $\rho_{n+1} = (\rho_n, \tilde{e})$ can be written as

$$p(\tilde{e} = \ell, \rho_n | \tilde{x}, x^n) \propto \prod_{j \neq \ell} c(S_j) g(x_j^*) c(S_\ell \cup \{n+1\}) g(x_\ell^*, \tilde{x})$$

$$= p(\rho_n | x^n) \frac{g(x_\ell^*, \tilde{x})}{g(x_\ell^*)} \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)}.$$

Let $q_\ell(\tilde{x} | x_\ell^*) \equiv g(x_\ell^*, \tilde{x}) / g(x_\ell^*)$. The posterior predictive distribution becomes

$$p(\tilde{y} | \tilde{x}, y^n, x^n) \propto \int \sum_{\ell=1}^{k_n+1} p(\tilde{y} | \tilde{x}, y_\ell^*, x_\ell^*, \tilde{e} = \ell) q_\ell(\tilde{x} | x_\ell^*)$$

$$\times \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)} p(\rho_n | y^n, x^n) \, d\rho_n. \tag{3.2}$$

The first factor reduces to $p(\tilde{y} | y_\ell^*, \tilde{e} = \ell)$ when (2.4) does not include a regression on x_i in the sampling model. Sampling from (3.2) is implemented on top of posterior simulation for $\rho_n \sim p(\rho_n | x^n, y^n)$. For each imputed ρ_n, generate $\tilde{e} = \ell$ with probabilities proportional to $w_\ell = q_\ell(\tilde{x} | x_\ell^*) c(S_\ell \cup \{n+1\}) / c(S_\ell)$, and generate $\tilde{y}$ from $p(\tilde{y} | y_\ell^*, \tilde{e} = \ell)$, weighted with w_ℓ. In the special case $\ell = k_n + 1$ we get $w_{k_n+1} = g(\tilde{x}) c(\{n+1\})$.

3.3 POSTERIOR INFERENCE ON CLUSTERS

In many applications inference on the number of clusters, k_n, is of interest. Under the proposed model $p(k_n = r | y^n, x^n) = \sum_{\{\rho_n : |\rho_n| = r\}} p(\rho_n | y^n, x^n)$ is the induced posterior distribution on the number of clusters. The marginal posterior $p(k_n = r | y^n, x^n)$ is reported as a simple summary of posterior Markov chain Monte Carlo (MCMC) simulation. Each iteration of the posterior simulation scheme in Section 3.1 involves imputing a partition ρ_n, and thus k_n. A formal selection of the number of clusters could be carried out using a decision-theoretic approach, that is, a loss function that weighs model complexity and the value of learning about the inferential target. Such formal procedures were discussed, among others, by Quintana and Iglesias (2003) and Lau and Green (2007). The special case of finding the maximum a posteriori clustering in a class of product partition models is considered by Dahl (2009).

Reporting posterior inference for the random partition ρ_n is complicated by the label switching problem. The posterior distribution is invariant under permutations of the cluster labels when $p(\rho_n)$ and $p(\theta)$ are both symmetric. This implies that inference on cluster-specific parameters is meaningless. See the article by Jasra, Holmes, and Stephens (2005) for a recent discussion and review of the literature. Usually the focus of inference in semi-parametric mixture models like (2.4) is on density estimation and prediction, rather than inference for specific clusters and parameter estimation. Predictive inference is not subject to the label switching problem, as the posterior predictive distribution marginalizes over all possible partitions. In the application reported later, when we want to highlight inference about specific clusters we use a simple ad hoc approach that we found particularly meaningful to report inference about the regression $p(\rho_n | x^n)$. We report inference stratified by k_n. For given k_n we first find a set of k_n indices $I = (i_1, \ldots, i_{k_n})$ with high posterior

probability of the corresponding cluster indicators being distinct, that is, $\Pr(D_I|y)$ is high for $D_I = \{e_i \neq e_j \text{ for } i \neq j \text{ and } i, j \in I\}$. To avoid computational complications, we do not insist on finding the k_n-tuple with the highest posterior probability. We then use $i_1, \ldots, i_{k_n}$ as anchors to define cluster labels by restricting posterior simulations to k_n clusters with the units i_j in distinct clusters. Postprocessing MCMC output, this is easily done by discarding all imputed parameter vectors that do not satisfy the constraint. We relabel clusters, indexing the cluster that contains unit i_j as cluster j. Now it is meaningful to report posterior inference on specific clusters. The proposed postprocessing is similar to the pivotal reordering suggested by Marin and Robert (2007, chapter 6.4).

4. SIMILARITY FUNCTIONS

For continuous covariates we suggest as a default choice for $g(x_j^*)$ the marginal distribution of x_j^* under a normal sampling model. Let $N(x; m, V)$ denote a normal model for the random variable x, with moments m and V, and let $\text{Ga}(x; a, b)$ denote a Gamma distributed random variable with mean a/b. We use $q(x_j^*|\xi_j) = \prod_{i \in S_j} N(x_i; m_j, v_j)$, with a conjugate prior for $\xi_j = (m_j, v_j)$ as $p(\xi_j) = N(m_j; m, B) \cdot \text{Ga}(v_j^{-1}; v, S_0)$, with fixed m, B, v, S_0. The main reason for this choice is operational simplicity. A simplified version uses fixed $v_j \equiv v$. The resulting function $g(x_j^\star)$ is the joint density of a correlated multivariate t-distribution, with location parameter m and scaling matrix $B/(vI + B)$. The fixed variance v specifies how strongly we weigh similarity of the x_i. In implementations we used $v = c_1 \widehat{S}$, where $\widehat{S}$ is the empirical variance of the covariate, and $c_1 = 0.5$ is a scaling factor that specifies over what range we consider values of this covariate as similar. The choice between fixed v versus variable v_j should reflect prior judgment on the variability of clusters. Variable v_j allows for some clusters to include a wider range of x values than others. Finally, a sufficiently vague prior for m_j is important to ensure that the similarity is appropriately quantified even for a group of covariate values in the extreme areas of the covariate space. In our implementation we used $B = c_2 \widehat{S}$ with $c_2 = 10$.

When constructing a cohesion function for categorical covariates, a default choice is based on a Dirichlet prior. Assume x_i is a categorical covariate, $x_i \in \{1, \ldots, C\}$. To define $g(x_j^*)$, let $q(x_i = c|\xi_j) = \xi_{jc}$ denote the probability mass function. Together with a conjugate Dirichlet prior, $q(\xi_j) = \text{Dir}(\alpha_1, \ldots, \alpha_C)$, we define the similarity function as a Dirichlet-categorical probability

$$g(x_j^*) = \int \prod_{i \in S_j} \xi_{j, x_i} \, dq(\xi_j) = \int \prod_{c=1}^{C} \xi_{jc}^{n_{jc}} \, dq(\xi_j), \qquad (4.1)$$

with $n_{jc} = \sum_{i \in S_j} I(x_i = c)$. This is a Dirichlet-multinomial model without the multinomial coefficient. For binary covariates the similarity function becomes a Beta–Binomial probability without the Binomial coefficient. The choice of the hyperparameters α needs some care. To facilitate the formation of clusters that are characterized by the categorical

covariates we recommend Dirichlet hyperparameters $\alpha_c < 1$. For example, for $C = 2$, the bimodal nature of a Beta distribution with such parameters assigns high probability to binomial success probabilities ξ_{j1} close to 0 or 1. Similarly, the Dirichlet distribution with parameters $\alpha_c < 1$ favors clusters corresponding to specific levels of the covariate.

For ordinal covariates, a convenient specification of $g(\cdot)$ is an ordinal probit model. The model can be defined by means of latent variables and cutoffs. Assume an ordinal covariate x with C categories. Following Johnson and Albert (1999), consider a latent trait Z and cutoffs $-\infty = \gamma_0 < \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_{C-1} < \gamma_C = \infty$, so that $x = c$ if and only if $\gamma_{c-1} < Z \leq \gamma_c$. We use fixed cutoff values $\gamma_c = c - 1$, $c = 1, \ldots, C - 1$, and a normally distributed latent trait, $Z_i \sim N(m_j, v_j)$. Let $\Phi_c(m_j, v_j) = \Pr(\gamma_{c-1} < Z \leq \gamma_c | m_j, v_j)$ denote the normal quantiles and let $\xi_j = (m_j, v_j)$ denote the cluster-specific moments of the latent score. We define $q(x_i = c | e_i = j, \xi_j) = \Phi_c(m_j, v_j)$. The definition of the similarity function is completed with $q(\xi_j)$ as a normal-inverse gamma distribution as for the continuous covariates. We have $q(x_j^*) = \int \prod_{c=1}^C \Phi_c^{n_{cj}}(m_j, v_j) N(m_j; m, B) \, \mathrm{Ga}(v_j^{-1}; \nu, S_0) \, dm_j \, dv_j$, where $n_{cj} = \sum_{i \in S_j} I\{x_i = c\}$.

Finally, to define $g(\cdot)$ for count-type covariates, we use a mixture of Poisson distributions

$$g(x_j^*) = \frac{1}{\prod_{i \in S_j} x_i!} \int \xi_j^{\sum_{i \in S_j} x_i} \exp(-\xi_j |S_j|) \, dq(\xi_j). \tag{4.2}$$

With a conjugate gamma prior, $q(\xi_j) = \mathrm{Ga}(\xi_j; a, b)$, the similarity function $g(x_j^*)$ allows easy analytic evaluation. As a default, we suggest choosing $a = 1$ and $a/b = c\hat{S}$, where $\hat{S}$ is the empirical variance of the covariate, and a/b is the expectation of the gamma prior.

The main advantage of the proposed similarity functions is computational simplicity of posterior inference. A minor limitation is the fact that the proposed default similarity functions, in addition to the desired dependence of the random partition model on covariates, also include a dependence on the cluster size $n_j = |S_j|$. From a modeling perspective this is undesirable. The mechanism to define size dependence should be the underlying PPM and the cohesion functions $c(S_j)$ in (2.1). However, we argue that the additional cluster size penalty that is introduced through $g(\cdot)$ is negligible in comparison to, for example, the cohesion function $c(S_j) = M(n_j - 1)!$ that is implied by the popular DP prior. It is straightforward to show that the similarity function introduces the following additional cluster size penalty in model (2.1). Consider the case of constant covariates, $x_i \equiv x$, and let n_j denote the size of cluster j. The default choices of $g(x_j^\star)$ for continuous, categorical, and count covariates introduce a penalty for large n_j, with $\lim_{n_j \to \infty} g(x_j^\star) = 0$. But the rate of decrease is ignorable compared to the cohesion $c(S_j)$. Let $f(n_j)$ be such that $\lim_{n_j \to \infty} g(x_j^\star)/f(n_j) \geq M$ with $0 < M < \infty$. For continuous covariates the rate is $f(n_j) = (2\pi)^{-n_j/2} V^{-(n_j-1)/2}(r + n_j)^{1/2}$, with $r = V/B$. For categorical covariates it is $f(n_j) = (A + n_j)^{A-\alpha_x}$, with $A = \sum_c \alpha_c$. For count covariates it is $f(n_j) = C^{-n_j/2}(\alpha + n_j x)^{1/2}$, with $C = 2\pi x \exp(1/6x)$.

5. ALTERNATIVE APPROACHES AND COMPARISON

5.1 OTHER APPROACHES

We propose the PPMx model as a principled approach to defining covariate-dependent random partition models. A related approach that has attracted considerable attention, especially in the recent marketing and psychometrics literature, is clusterwise regression. The basic model is defined in the article by Späth (1979). The model uses a partition of the experimental units (observations) into subsets S_j. For each partitioning subset a different linear regression mean function is defined. A recent implementation of clusterwise regression as an R package *flexmix* was described by Leisch (2004). We use *flexmix* in the comparison below.

Another popular class of models that implements flexible regression based on clustering experimental units are hierarchical mixtures of experts (HME) and related models. In contrast to clusterwise regression the HME model allows for the cluster weights to be dependent on covariates themselves. HME models were introduced by Jordan and Jacobs (1994). Bayesian posterior inference was discussed by Bishop and Svensén (2003). We consider a specific instance of the HME model that we will use for an empirical comparison in Section 5.2. The model is expressed as a finite mixture of normal linear regressions, with mixture weights that themselves depend on the available covariates as well,

$$p(y_i; \eta) = \sum_{j=1}^{k_n} w_{ij}(x_i; \alpha_j) N(y_i; \theta_j^T x_i, \sigma_i^2). \tag{5.1}$$

Here $\eta = (\theta_1, \ldots, \theta_{k_n}, \sigma_1^2, \ldots, \sigma_{k_n}^2, \alpha_1, \ldots, \alpha_{k_n})$ is the full vector of parameters. The size k_n of the mixture is prespecified. The covariate vector x_i includes a 1 for an intercept term. The weights $0 \le w_{ij} \le 1$ are defined as $w_{ij} \propto \exp(\alpha_j^T x_i)$. For identifiability we assume $\alpha_1 \equiv 0$. The model specification is completed by assuming $\theta_j \sim N(\mu_\theta, V_\theta)$, $\alpha_j \sim N(\mu_\alpha, V_\alpha)$, $\sigma_j^{-2} \sim \text{Ga}(a_0, a_1)$, and prior independence of all parameters. The HME formulation defines a highly flexible class of parametric models. But there are some important limitations. The dependence of cluster membership probabilities on the covariates is strictly limited to the assumed parametric form of $w_{ij}(\cdot)$. For example, unless an interaction of two covariates is included in $w_{ij}(\cdot)$, no amount of data will allow clusters that are specific to interactions of covariates. Second, the number of mixture components k_n is assumed known. Ad hoc fixes are possible, such as choosing a very large k_n, or introducing a hyperprior on k_n.

Clusterwise regression and HME do not explicitly include a probability model over random partitions. A model $p(\rho_n)$ is implicitly defined as in (1.2), as model-based clustering. The dependence of cluster membership on covariates remains by definition restricted to the given structure. Alternatively, some recently proposed methods in the Bayesian literature take a perspective similar to the proposed PPMx model and include an explicit probability model for a random partition.

Park and Dunson (2010) considered the special case of continuous covariates. They defined the desired covariate-dependent random partition model as the posterior random

partition under a PPM model for the covariates x_i, that is, (1.1) with covariates x_i replacing the response data y_i. The posterior $p(\rho_n|x^n)$ is used to define the prior for the random partition ρ_n. In other words, they proceeded with an augmented response vector (x_i, y_i), carefully separating the prior for parameters related to the x and y subvectors, as in (3.1). See Section 3.1 for a discussion of the need to separate prior parameters related to x and y.

Another recently proposed approach is that of Shahbaba and Neal (2009). They used a logistic regression for a categorical response variable on covariates together with a random partition of the samples, with cluster-specific regression parameters in each partitioning subset. The random partition is defined by a DP prior and includes the covariates as part of the response vector. The proposed model can be written as (3.1) with $p(y_i|x_i, \theta_j, \eta)$ specified as a logistic regression with cluster-specific parameters θ_j, and $q(x_i|\xi_j)$ as a normal model with moments $\xi_j = (\mu_j, \Sigma_j)$. The cohesion functions $c(S_j)$ are defined by a DP prior.

Dahl (2008) recently proposed another interesting approach to covariate-dependent random clustering. Let $\mathbf{e} = (e_1, \ldots, e_n)$ and let $\mathbf{e}_{-i}$ denote the vector of cluster membership indicators without the ith element. Let $p(e_i = j|\mathbf{e}_{-i})$ denote the conditional prior probabilities for the cluster membership indicators. For the random clustering implied by the DP prior, $G \sim \text{DP}(M, G^\star)$, these conditional probabilities are known as the Polya urn. Let k_n^- denote the number of distinct elements in $\mathbf{e}_{-i}$, $S_j^- = \{h : e_h = j \text{ and } h \neq i\}$, and $n_j^- = |S_j^-|$. The Polya urn specifies $p(e_i = j|\mathbf{e}_{-i}) = n_j^-/(n - 1 + M)$ for $j = 1, \ldots, k_n^-$, and $M/(n - 1 + M)$ for $j = k_n^- + 1$. Dahl (2008) defined the desired covariate-dependent random partition model by modifying the Polya urn. Assume covariates x_i are available and it is desired to modify $p(\mathbf{e})$ such that any two units with similar covariates should have increased prior probability for co-clustering. Assume $d_{ih} = d(x_i, x_h)$ can be interpreted as a distance of x_i and x_h. Let $d^\star = \max_{i<h} d_{ih}$ and define $h_i(S_j) = c \sum_{h \in S_j}(d^\star - d_{ih})$ with c chosen to achieve $\sum_{j=1}^{k_n} h_i(S_j) = n$. Dahl (2008) defined the modified Polya urn

$$p(e_i = j|\mathbf{e}_{-i}, x^n) \propto \begin{cases} h_i(S_j^-), & j = 1, \ldots, k_n^- \\ M, & j = k_n^- + 1. \end{cases} \tag{5.2}$$

This set of transition probabilities defines an ergodic Markov chain. Thus it implicitly defines a joint probability distribution on $\mathbf{e}$ that is informed by the relative distances d_{ij} as desired.

5.2   COMPARISON OF COMPETING APPROACHES

We set up a Monte Carlo study to compare the performance of the proposed PPMx model against four of the described alternative approaches: (i) clusterwise regression using the *flexmix* R package (Leisch 2004); (ii) the HME model (5.1); (iii) the approach of Dahl (2008); and (iv) the model of Park and Dunson (2010).

The study is set up to compare the competing methods as nonlinear regression methods. We thus use mean squared error (MSE) as the criterion for the comparison. MSE is the average squared deviation of the fitted regression mean function and the known simulation truth, averaging with respect to repeated sampling. We report MSE for a set of covariate

combinations including some covariate combinations that are not observed in the data, that is, that require extrapolation. For a fair comparison we use a simulation truth that is not characterized as the assumed model under any of the competing approaches. In particular, we do not generate a simulation truth of assumed clusters. Instead we set up a nonlinear regression of a continuous dependent variable on three independent variables, including two binary variables and one continuous variable.

Let x_1, x_2, x_3 denote the three covariates, $x_1 \in R$ and $x_2 \in \{0, 1\}$, $x_3 \in \{0, 1\}$. The simulation truth is chosen to mimic the example in Section 6. Think of x_2 as an indicator for high dose (HI), x_3 as an indicator for ER positive tumors (ER+), and x_1 as tumor size (TS), with $-1, 0$, and 1 representing small, medium, and large size. There is a strong interaction of x_2 and x_3 with longest survival for $(x_2, x_3) = (1, 1)$, and a sizeable main effect for x_1.

Figure 1 shows the simulation truth. The figure plots the distribution of the univariate response y_i arranged by $x_1 \in \{-1, 0, 1\}$ (the three distributions shown in each panel) and the four combinations of (x_2, x_3) (one panel for each combination). Note the interaction of x_1 and x_2. We simulated $M = 100$ datasets of size $n = 200$ using the described sampling model. The simulated data included all covariate combinations shown in Figure 1 except for $(x_1, x_2, x_3) = (1, 0, 0)$ and $(1, 1, 0)$. Simulation also included intermediate values of the continuous variable x_1. The $M = 100$ datasets were generated by resampling one big set of $N = 1000$ simulated data points. The big dataset and the expectations $E(Y|x_1, x_2, x_3)$
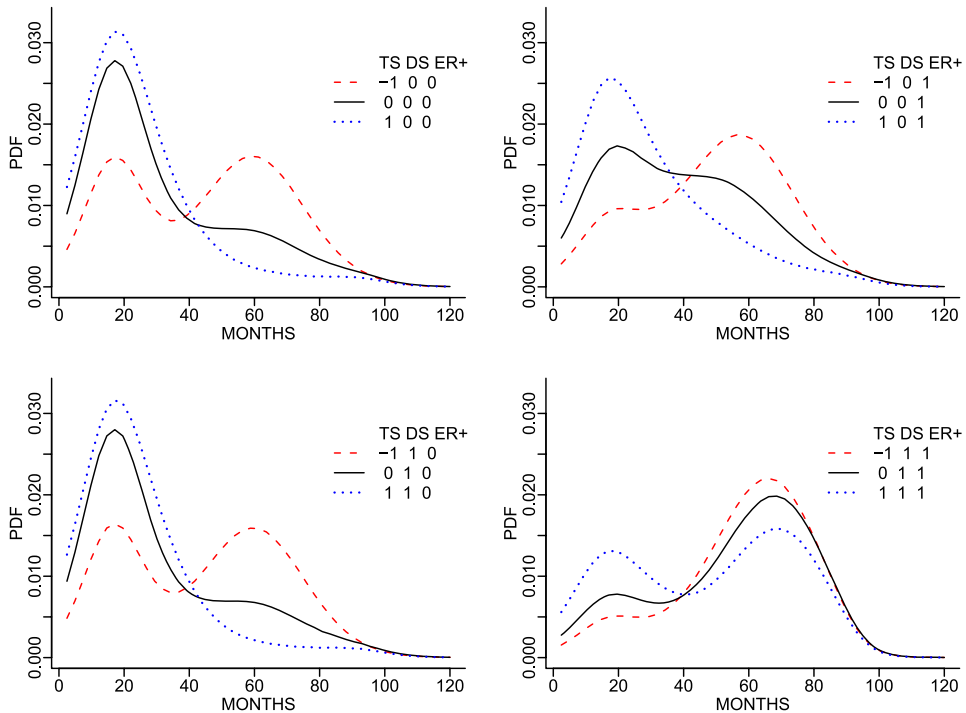

Figure 1.   Simulation truth pdf for the outcome variable for different combinations of the three covariates. The online version of this figure is in color.

Table 1.   Root MSE for estimating $E(y|x_1, x_2, x_3)$ for 12 combinations of (x_1, x_2, x_3) and the four competing models. Models (i) through (iv) and the proposed model are indicated as flexmix, HME, DAHL, P&D, and PPMx, respectively. Covariate combinations that require extrapolation beyond the range of the data are indicated by ⋆.

x_1	x_2	x_3	flexmix	PPMx	P&D	DAHL	HME	
−1	0	0	13.9	7.9	**2.7**	8.6	13.0	
0	0	0	7.7	**3.9**	15.0	4.6	8.3	
1	0	0	5.2	**2.8**	21.5	8.7	8.7	⋆
−1	1	0	7.4	5.4	**2.3**	6.2	5.7	
0	1	0	**4.4**	4.6	15.5	7.9	5.5	
1	1	0	**3.3**	4.0	21.0	12.6	5.3	⋆
−1	0	1	8.1	6.1	**1.8**	9.5	7.5	
0	0	1	6.8	4.2	7.0	**2.4**	3.5	
1	0	1	**3.7**	4.5	17.4	8.2	4.9	
−1	1	1	8.0	9.5	12.1	12.2	**6.5**	
0	1	1	11.5	8.3	8.7	10.2	**5.4**	
1	1	1	**3.1**	6.2	2.4	4.8	3.8	
Average			6.9	**5.6**	10.6	8.0	6.5	

under the simulation truth are available in the supplemental materials or at *http://odin. mdacc.tmc.edu/~pmueller/prog.html#PPMx*.

We compare the proposed PPMx model and the alternative approaches (i) through (iv). As criterion we use the root mean squared error (RMSE) in estimating $E(y|x_1, x_2, x_3)$ for the 12 combinations of the covariates (x_1, x_2, x_3) shown in Figure 1. We evaluated MSE by averaging squared errors over $M = 100$ repeat simulations. Table 1 summarizes the results. Rows $(x_1, x_2, x_3) = (1, 0, 0)$ and $(1, 1, 0)$ report covariate combinations that were not included in the dataset. Results are thus based on extrapolation. The PPMx model performs well for these extrapolation problems. The last three rows correspond to the right lower panel in Figure 1. Performance for these scenarios reflects the adaptation of the model to extreme interaction effects. The HME performs surprisingly well. The PPMx reports reasonable MSE. The clusterwise regression (flexmix) is comparable to PPMx. Overall there is no clear winner in the comparison, although PPMx attains the lowest average RMSE across all covariate combinations. In any case we caution against over-interpreting the results for one example. The main conclusion is that all five approaches are reasonably comparable, and a caveat about extrapolation in the approach by Park and Dunson (2010). The choice of approach should depend on the inference goal. Clusterwise regression and HME are perfectly appropriate if the main objective is flexible regression. If available information about similarities is naturally expressed by pairwise distances d_{ij}, and the main focus is predictive inference, then the approach by Dahl (2008) is attractive. A limitation of the latter three approaches is the lack of specific inference on the random partition. Inference on cluster membership indicators can be reported. But under the HME cluster membership is strictly limited to the functional form of $w(\cdot; \alpha_j)$. Under the approach of Dahl (2008), although a model is implied by the set of conditional distributions (5.2), there is in general no clearly identified prior probability model for ρ_n. The approach of Park and Dunson (2010) is reasonable if the DP prior is chosen for the underlying PPM, the covariates are

continuous, and the covariates can be considered random variables. The proposed PPMx is attractive when the set of covariates includes a mix of different data formats, and if there is specific prior information of how important different covariates should be for the judgment of similarity.

6. EXAMPLE: A SURVIVAL MODEL WITH PATIENT BASELINE COVARIATES

We consider data from a high-dose chemotherapy treatment of women with breast cancer. The data for this particular study have been discussed by Rosner (2005) and come from Cancer and Leukemia Group B (CALGB) Study 9082. They consist of measurements taken from 763 randomized patients, available as of October 1998 (enrollment had occurred between January 1991 and May 1998). The response of interest is the survival time, defined as the time until death from any cause, relapse, or diagnosis with a second malignancy. There are two treatments, one involving a low dose of the anti-cancer drugs, and the other consisting of aggressively high-dose chemotherapy. The high-dose patients were given considerable regenerative blood-cell supportive care (including marrow transplantation) to help decrease the impact of opportunistic infections rising from the severely affected immune system. The number of observed failures was 361, with 176 under high-dose and 185 under low-dose chemotherapy.

The dataset also includes information on the following covariates for each patient: a treatment indicator defined as 1 if a high dose was administered and 0 otherwise (HI); age in years at baseline (AGE); the number of positive lymph nodes found at diagnosis (POS) (the more the worse the prognosis, i.e., the more likely the cancer has spread); tumor size in millimeters (TS), a one-dimensional measurement; an indicator of whether the tumor is positive for the estrogen or progesterone receptor (ER+) (patients who were positive also received the drug tamoxifen and are expected to have better risk), and an indicator of the woman's menopausal status, defined as 1 if she is either perimenopausal or postmenopausal or 0 otherwise (MENO). Two of these six covariates are continuous (AGE, TS), three are binary (ER+, MENO, HI), and one is a count (POS).

6.1 PPMx With One Covariate

First we carried out inference in a model using the indicator for high dose as the only covariate, that is, $x_i = $ HI. We implemented model (2.4) with a similarity function for the binary covariate based on the beta–binomial model (4.1). We used $\alpha = (0.1, 0.1)$ to favor clusters with homogeneous dose assignment. Conditional on an assumed partition ρ_n we use a normal sampling model $p(y_i|\theta_j) = N(\mu_j, V_j)$, with a conjugate normal–inverse gamma prior $p(\theta_j|\eta) = N(\mu_j|m_y, B_y)\,\mathrm{Ga}(V_j^{-1}|s/2, sS_y/2)$ and hyperprior $m_y \sim N(a_m, A_m)$, $S_y \sim \mathrm{Ga}(q, q/R)$. Here $\eta = (a_m, A_m, B_y, q, R)$ are fixed hyperparameters. We use $a_m = \widehat{m}$, the sample average of y_i, $A_m = 100$, $B_y = 100^2$, $s = q = 4$, and $R = 100$. The cohesion functions were chosen as before with $c(S_j) = M(n_j - 1)!$, matching the PPM implied by the DP prior. We include a $\mathrm{Ga}(1, 1)$ hyperprior for the total mass parameter M.

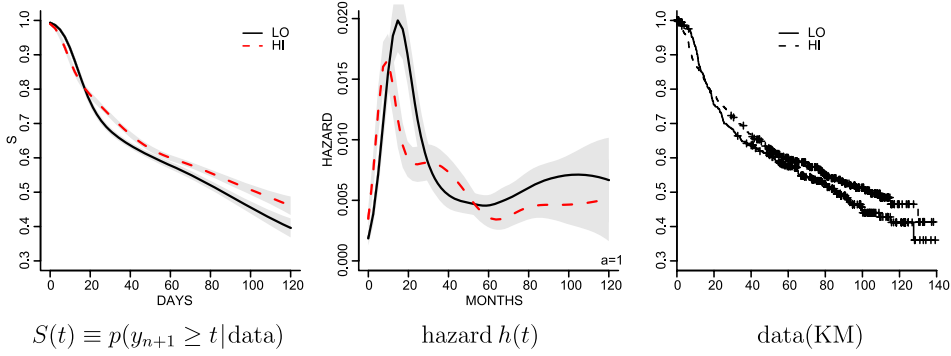$$S(t) \equiv p(y_{n+1} \geq t \mid \text{data}) \qquad \text{hazard } h(t) \qquad \text{data(KM)}$$

Figure 2. Survival example: Estimated survival function (left panel) and hazard (center panel), arranged by $x \in \{\text{HI, LO}\}$. The gray shades show pointwise one posterior predictive standard deviation uncertainty. The right panel shows the data for comparison (Kaplan–Meier curve by dose). The online version of this figure is in color.

Figure 2 shows inference summaries. The posterior distribution $p(k_n \mid \text{data})$ for the number of clusters is shown in Figure 3. The three largest clusters contain 28%, 23%, and 14% of the experimental units.

6.2 PPMx WITH MULTIPLE COVARIATES

Next we extended the covariate vector to include all six covariates. Denote by $x_i = (x_{i1}, \ldots, x_{i6})$ the six-dimensional covariate vector. We implement random clustering with regression on covariates as in model (2.4). The similarity function is defined as

$$g(x_j^*) = \prod_{\ell=1}^{6} g^\ell(x_{j\ell}^*). \qquad (6.1)$$

For each covariate we follow the suggestion in Section 4 to define a factor $g^\ell(\cdot)$ of the similarity function, using hyperparameters specified as follows. The similarity func-
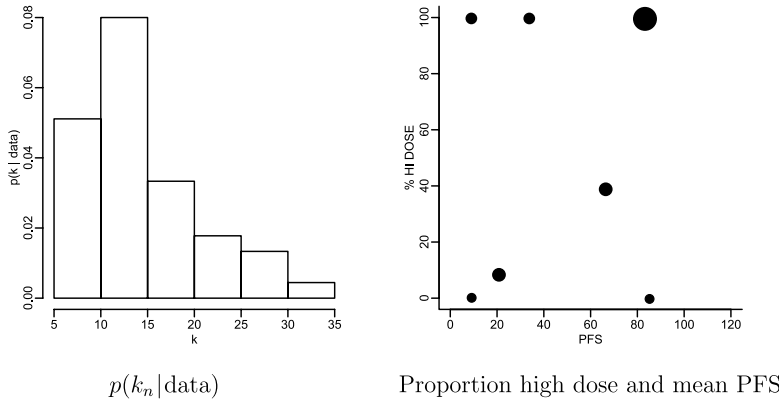

$$p(k_n \mid \text{data}) \qquad \text{Proportion high dose and mean PFS}$$

Figure 3. Survival example: Posterior for the number of clusters k_n (left panel), and proportion of patients with high dose (% HI) and average progression free survival (PFS) by cluster. The size (area) of the bullets is proportional to the average cluster size.

tion for the three binary covariates is defined as in (4.1) with $\alpha = (0.1, 0.1)$ for HI, and $\alpha = (0.5, 0.5)$ for ER+ and MENO. The two continuous covariates AGE and TS were standardized to sample mean 0 and unit standard deviation. The similarity functions were specified as described in Section 4, with fixed $s = 0.25$, $m = 0$, and $B = 1$. Finally, for the count covariate POS we used the similarity function (4.2) with $(a, b) = (1.5, 0.1)$. The sampling model is unchanged from before.

We assume that censoring times are independent of the event times and all parameters in the model. Posterior predictive survival curves for various covariate combinations are shown in Figure 4. In the figure, "baseline" refers to HI = 0, tumor size 38 mm (the empirical median), ER = 0, MENO = 0, average age (44 years), and POS = 15 (empirical mean). Other survival curves are labeled to indicate how the covariates change from baseline, with TS− indicating tumor size 26 mm (the empirical first quartile), TS+ indicating tumor size 50 mm (third quartile), HI referring to high-dose chemotherapy, and ER+ indicating positive estrogen or progesterone receptor status. The inference suggests that treated patients with tumor size below the empirical median and that were positive for estrogen or progesterone receptor have almost uniformly highest predicted survival curves than any other combination of covariates.

Figure 5 summarizes features of the posterior clustering. Interestingly, clusters are typically highly correlated to the postmenopausal status, as seen in the right panel.
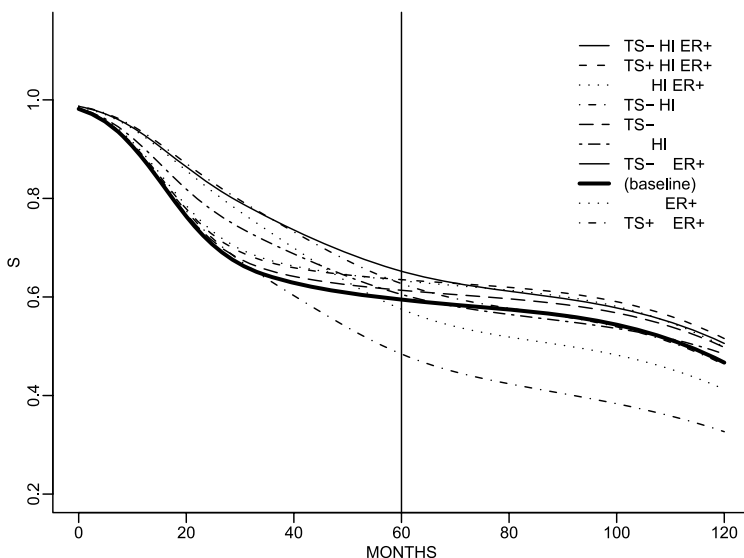

Figure 4.   Survival example: Posterior predictive survival function $S(t|x) \equiv p(y_{n+1} > t|x_{n+1} = x, \text{data})$, arranged by x. The "baseline" case refers to all continuous and count covariates equal to the empirical mean, and all binary covariates equal 0. The legend indicates TS− and TS+ for tumor size equal 26 mm and 50 mm (first and third empirical quartile), HI for HI = 1, and ER+ for ER = 1. The legend is sorted by the survival probability at 5 years, indicated by a thin vertical line.
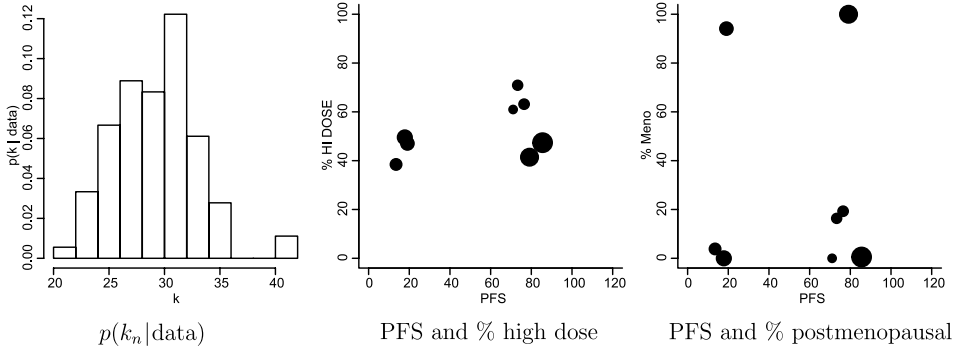
Figure 5. Survival example: Posterior distribution for the number of clusters (left panel), mean PFS and % high-dose patients per cluster (center panel), mean PFS and % postmenopausal patients per cluster (right panel).

6.3 VARIABLE SELECTION

The PPMx model replaces variable selection in a regression model by clustering of experimental units (patients). Consider, for example, inference in Figure 5. Clusters are heterogeneous with respect to HI, but very homogeneous with respect to MENO. In other words, cluster membership and thus predictive inference for a future observation varies more by MENO than by HI.

Similar summaries for all covariates lead us to consider formal variable selection. We explore the use of reduced models by comparing alternative models including subsets of the $p = 6$ covariates. Let $\gamma = (\gamma_1, \ldots, \gamma_p)$ denote a vector of binary indicators and modify (6.1) to include only selected covariates, $g(x_j^\star) = \prod_{\ell:\gamma_\ell = 1} g^\ell(x_{j\ell}^*)$. We complete the probability model with a prior on γ. Without loss of generality we assume $p(\gamma_j = 1) = \pi$, independently, $j = 1, \ldots, p$.

Posterior simulation for γ involves a minor challenge related to the normalization constant $g_n(x^n)$ in (2.1). We write $g_n(x^n|\gamma)$ to highlight that the definition of $g(\cdot)$ is indexed by γ. We implement posterior simulation with a Metropolis–Hastings transition probability to change indicators γ_j, keeping the currently imputed partition ρ and all other parameters unchanged. Let γ^1 and γ^0 denote two vectors of variable selection indicators. Without loss of generality we assume that the two vectors differ only in the first indicator with $\gamma_1^1 = 1$, $\gamma_1^0 = 0$, and $\gamma_j^1 = \gamma_j^0$, $j = 2, \ldots, p$. The acceptance probability includes the ratio

$$\frac{p(\gamma^1|\rho, x^n)}{p(\gamma^0|\rho, x^n)} = \prod_{j=1}^{k_n} g(x_{j1}^\star) \frac{g_n(x^n|\gamma^0)}{g_n(x^n|\gamma^1)}.$$

We use (independent) Monte Carlo integration to compute $g_n(x^n|\gamma)$ for all 2^p possible γ vectors. We do this once, up front, before starting the posterior MCMC simulation. Let $g(x^n|\rho, \gamma) = \prod_j g(x_j^\star|\gamma)$ denote the product of similarity functions under partition ρ and variable selection γ. Recall that $q(\rho) = \prod c(S_j)$ denotes the random partition under the PPM without the additional similarity function. We use

$$g_n(x^n|\gamma) = \sum_\rho g(x^n|\rho, \gamma) q(\rho) \approx \frac{1}{M} \sum_m g(x^n|\rho_m, \gamma^1)$$

for $\rho_m \sim q(\rho)$, iid. We use Rao–Blackwellization to compute the marginal posterior probabilities $p(\gamma|x^n, y^n)$ as ergodic averages of $p(\gamma|\rho, x^n)$, averaging over imputed partitions ρ.

Using the described posterior simulation and the data from the previous example, we find the following posterior probabilities for variable selection. The posterior mode is achieved with $p(\gamma = (1, \ldots, 1)|x^n, y^n) = 0.62$. This is the model with all covariates. The next most likely model is given by $p(\gamma = (1, 1, 0, 1, 1, 1)|x^n, y^n) = 0.38$ for the model without the indicator HI for high dose. It is clinically plausible that the effect of high dose would fade in comparison with initial tumor burden and other important baseline characteristics. One would still continue to include HI as the treatment covariate, that is, the only covariate that can be manipulated by choice. The cumulative posterior probability of all other 62 models is less than 0.001.

In summary, model-based inference confirms the inference reported by summaries such as Figure 5. In general we prefer the use of summaries of cluster homogeneity as in Figure 5, as it readily generalizes to larger numbers of covariates, including possibly interesting interactions.

7. CONCLUSION

We have proposed a novel model for random partitions with a regression on covariates. The model builds on the popular PPM random partition models by introducing an additional factor to modify the cohesion function. We refer to the additional factor as similarity function. It increases the prior probability that experimental units with similar covariates are co-clustered. We provide default choices of the similarity function for popular data formats.

The main features of the model are the possibility to include additional prior information related to the covariates, the principled nature of the model construction, and a computationally efficient implementation.

Among the limitations of the proposed method is an implicit penalty for the cluster size that is implied by the similarity function. Consider all equal covariates $x_i \equiv x$. The value of the similarity functions proposed in Section 4 decreases across cluster size. This limitation could be mitigated by allowing an additional factor $c^\star(|S_j|)$ in (2.4) to compensate the size penalty implicit in the similarity function.

SUPPLEMENTARY MATERIALS

R-package: R-package `PPMx` containing code to carry out inference for the model described in the article. The function `ppmx` implements the proposed covariate-dependent random partition model for an arbitrary combination of continuous, categorical, binary, and count covariates, using a mixture of normal sampling model for y_i. The package is also available for download from *http://www.math.utexas.edu/users/pmueller/prog. html*. (ppmx_1.0.tar.gz)

Simulation data: The data file contains the 1000 simulated data points used for the Monte Carlo study that is described in Section 5.2. (dta.txt)

ACKNOWLEDGMENTS

REFERENCES

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821. [261]

Barry, D., and Hartigan, J. A. (1993), "A Bayesian Analysis for Change Point Problems," *Journal of the American Statistical Association*, 88, 309–319. [261]

Bernardo, J.-M., and Smith, A. F. M. (1994), *Bayesian Theory. Wiley Series in Probability and Mathematical Statistics*, Chichester: Wiley. [263]

Bishop, C. M., and Svensén, M. (2003), "Bayesian Hierarchical Mixtures of Experts," in *2003 Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, eds. U. Kjaerulff and C. Meek, pp. 57–64. [268]

Crowley, E. M. (1997), "Product Partition Models for Normal Means," *Journal of the American Statistical Association*, 92, 192–198. [261]

Dahl, D. B. (2008), "Distance-Based Probability Distribution for Set Partitions With Applications to Bayesian Nonparametrics," in *Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association. [269,271]

———— (2009), "Modal Clustering in a Class of Product Partition Models," *Bayesian Analysis*, 4, 243–264. [265]

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. [261]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [261]

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [261]

Green, P. J., and Richardson, S. (1999), "Modelling Heterogeneity With and Without the Dirichlet Process," technical report, University of Bristol, Dept. of Mathematics. [262]

Hartigan, J. A. (1990), "Partition Models," *Communications in Statistics, Part A—Theory and Methods*, 19, 2745–2756. [261]

Ishwaran, H., and James, L. F. (2003), "Generalized Weighted Chinese Restaurant Processes for Species Sampling Mixture Models," *Statistica Sinica*, 13, 1211–1235. [261]

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005), "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling," *Statistical Science*, 20, 50–67. [265]

Johnson, V. E., and Albert, J. H. (1999), *Ordinal Data Modeling. Statistics for Social Science and Public Policy*, New York: Springer-Verlag. [267]

Jordan, M., and Jacobs, R. (1994), "Hierarchical Mixtures-of-Experts and the EM Algorithm," *Neural Computation*, 6, 181–214. [268]

Lau, J. W., and Green, P. J. (2007), "Bayesian Model-Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558. [265]

Leisch, F. (2004), "FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R," *Journal of Statistical Software*, 11, 1–18. [268,269]

Marin, J.-M., and Robert, C. P. (2007), *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, New York: Springer-Verlag. [266]

Park, J.-H., and Dunson, D. (2010), "Bayesian Generalized Product Partition Models," *Statistica Sinica*, 20, 1203–1226. [262,268,269,271]

Pitman, J. (1996), "Some Developments of the Blackwell–MacQueen Urn Scheme," in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell. Institute of Mathematical Statistics Lecture Notes—Monograph Series*, Vol. 30, eds. T. S. Ferguson, L. S. Shapeley, and J. B. MacQueen, Hayward, CA: Institute of Mathematical Statistics, pp. 245–268. [261]

Quintana, F. A. (2006), "A Predictive View of Bayesian Clustering," *Journal of Statistical Planning and Inference*, 136, 2407–2429. [261,264]

Quintana, F. A., and Iglesias, P. L. (2003), "Bayesian Clustering and Product Partition Models," *Journal of the Royal Statistical Society, Ser. B*, 65, 557–574. [261,265]

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 59, 731–792. [261]

Rosner, G. L. (2005), "Bayesian Monitoring of Clinical Trials With Failure-Time Endpoints," *Biometrics*, 61, 239–245. [272]

Shahbaba, B., and Neal, R. M. (2009), "Nonlinear Models Using Dirichlet Process Mixtures," *Journal of Machine Learning Research*, 10, 1829–1850. [262,263,269]

Späth, H. (1979), "Clusterwise Linear Regression," *Computing*, 22, 93–119. [268]