

Bayesian Clustering with Regression

Peter Müller

University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, U.S.A.

Fernando Quintana

Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile.

Gary Rosner

University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, U.S.A.

Summary. 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.

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.

Let $i = 1, \dots, n$, index experimental units, and let $\rho_n = \{S_1, \dots, S_{k_n}\}$ 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 i -th unit. Let $x^n = (x_1, \dots, x_n)$ and $y^n = (y_1, \dots, y_n)$ 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 $e_i \in \{1, \dots, k\}$ with $e_i = j$ if $i \in S_j$, and use (k, e_1, \dots, 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). Many clustering models include, implicitly or explicitly, a sampling model $p(y | \rho)$. Probability models $p(\rho_n)$ and inference for clustering have been extensively discussed over the past few years (see Quintana, 2006, for a recent review). In this paper we are interested in adding a regression to replace $p(\rho_n)$ with $p(\rho_n | x^n)$.

We focus on the product partition models (PPM). The PPM (Hartigan, 1990; Barry and Hartigan, 1993) constructs $p(\rho_n)$ by introducing cohesion functions $c(A) \geq 0$ for $A \subset \{1, \dots, n\}$ 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

$$p(\rho_n) \propto \prod_{j=1}^{k_n} c(S_j) \quad \text{and} \quad p(y^n | \rho_n) = \prod_{j=1}^{k_n} p_j(y_j^*) \quad (1)$$

Model (1) is conjugate since 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, $p(\rho_n) = p(|S_1|, \dots, |S_k|)$. The SSM can be alternatively characterized by a sequence of predictive probability functions (PPF) 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; Antoniak, 1974) is a special case of a SSM. Moreover, the marginal distribution that a DP induces on partitions is also a PPM with cohesions $c(S_j) = M \times (|S_j| - 1)!$ (Quintana and Iglesias, 2003; Dahl, 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 $p(y_i | \eta, k) = \sum_{j=1}^k \tau_j p_j(y_i | \theta_j)$, where $\eta = (\theta_1, \dots, \theta_k, \tau_1, \dots, \tau_k)$ are the parameters of a size k mixture model. Together with a prior $p(k)$ on k and $p(\theta, \tau | k)$, the mixture implicitly defines a probability model on clustering. Consider the equivalent hierarchical model

$$p(y_i | e_i = j, k, \eta) = p_j(y_i | \theta_j) \quad \text{and} \quad Pr(e_i = j | k, \eta) = \tau_j. \quad (2)$$

The implied posterior distribution on (e_1, \dots, e_n) and k defines a probability model on ρ_n . Richardson and Green (1997) develop posterior simulation strate-

gies for mixtures of normal models. Green and Richardson (1999) discuss the relationship to DP mixture models.

Especially in the context of spatial data, a popular clustering model is based on Voronoi tessellations (Green and Sibson, 1978; Okabe et al., 2000; Kim et al., 2005). For a review see, for example, Denison et al. (2002). Although not usually considered as a probability model on clustering, it can be defined as such by means of the following construction. We include it in this brief review of clustering models because of its similarity to the model proposed in this paper. Assume x_i are spatial coordinates for observation i . Let $d(x_1, x_2)$ denote a distance measure, for example Euclidean distance in \mathbb{R}^2 . Clusters are defined with latent cluster centers T_j , $j = 1, \dots, k$, by setting $e_i = \arg \min_j d(x_i, T_j)$. That is, a partition is defined by allocating each observation to the cluster whose center T_j is closest to x_i . Conditional on T , cluster allocation is deterministic. Marginalizing with respect to a prior on T we could define a probability model $p(\rho_n | k)$.

In this paper we build on the PPM (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*. Focusing on continuous covariates a similar approach building on PPMs is being proposed in current work by Park and Dunson (2007). Their approach starts with a PPM for the covariates, and uses the posterior random partition given the covariates as the prior model for inference with the response data. Shahbaba and Neal (2007) introduce a related model for categorical outcomes, i.e., classification. They include the covariates as part of an augmented response vector. This leads to a similar covariate-dependent random partition model. 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. In the context of the simulation study we contrast the proposed approach to an often used ad-hoc implementation of including the covariates in an augmented response vector.

2. The PPMx Model: Clustering with Covariates

We build on the PPM (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 j -th cluster. Let $g(x_j^*)$ denote a non-negative 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

$$p(\rho_n | x) \propto \prod_{j=1}^{k_n} g(x_j^*) \cdot c(S_j) \quad (3)$$

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,

$$g(x_j^*) \equiv \int \prod_{i \in S_j} q(x_i | \xi_j) q(\xi_j) d\xi_j. \quad (4)$$

We use the probability model $q(\cdot)$ to define the desired similarity function $g(x_j^*)$. 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 (4) leads to higher values of $g(x_j^*)$ 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 (4). The function (4) satisfies the following two properties that are desirable for a similarity function in (3). 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 $\int g(x^*) = \int g(x^*, x) dx$. In words, the similarity of any cluster is the average of any augmented cluster.

Under these two requirements (4) is not only technically convenient. It is the only possible similarity function that satisfies these two constraints.

Proposition 1: Assume that (i) a similarity function $g(x^*)$ satisfies the two constraints; (ii) $g(x_j^*)$ integrates over the covariate space, $\int g(x_j^*) dx_j^* < \infty$. Then $g(x_j^*)$ is proportional to the marginal distribution on x_j^* under a hierarchical auxiliary model (4).

The result follows from De Finetti's representation theorem for exchangeable probability measures, applied to a normalized version of $q(\cdot)$. See, for example, 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 (4).

The definition of the similarity function with the auxiliary model $q(\cdot)$ also

implies another important property. The random partition model (3) 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.

Corollary: Assume that the similarity function is defined by means of an auxiliary model, as in (4). The covariate dependent PPM (3) is coherent across sample sizes as formalized by the following relationship of $p(\rho_n | x^n)$ and $p(\rho_{n+1} | x^n, x_{n+1})$:

$$p(\rho_n | x^n) = \sum_{e_{n+1}} \int p(\rho_{n+1} | x^n, x_{n+1}) q(x_{n+1} | x^n) dx_{n+1}, \quad (5)$$

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 (3) 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 . Let $\theta = (\theta_1, \dots, \theta_{k_n})$.

$$p(y^n | \rho_n, \theta, \eta, x^n) = \prod_{j=1}^{k_n} \prod_{i \in S_j} p(y_i | x_i, \theta_j, \eta) \quad \text{and} \quad p(\theta | \eta) = \prod_j p(\theta_j | \eta). \quad (6)$$

We refer to (3) together with (1) as PPM with covariates, and write PPMx for short. The resulting model extends PPMs of the type (1) while keeping the product structure. Note that the sampling model $p(y_i | x_i, \theta_j, \eta)$ in (6) can include a regression of the response data on the covariate x_i . For example, Shahbaba and Neal (2007) includes a logistic regression for y_i on x_i , in addition to a covariate-dependent partition mode $p(\rho_n | x^n)$. 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 (3) and (6) 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, define an auxiliary model $q(\cdot)$ by replacing (6) with

$$q(y^n, x^n \mid \rho_n, \theta, \eta) = \prod_j \prod_{i \in S_j} p(y_i \mid x_i, \theta_j, \eta) q(x_i \mid \xi_j) \text{ and } q(\theta, \xi \mid \eta) = \prod_j p(\theta_j \mid \eta) q(\xi_j) \quad (7)$$

and replace the covariate-dependent prior $p(\rho_n \mid x^n)$ by the PPM $q(\rho_n) \propto \prod c(S_j)$. We continue to use $q(\cdot)$ for the auxiliary probability model that is introduced as a computational device only, and $p(\cdot)$ for the proposed inference model. The posterior distribution $q(\rho_n \mid y^n, x^n)$ under the auxiliary model is identical to $p(\rho_n \mid y^n, x^n)$ under the proposed model. An important caveat 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 $p(\theta_j \mid \eta)$ and $q(\xi_j \mid \eta)$ depend on a common hyperparameter η , then the posterior distribution under the auxiliary model (7) would differ from the posterior under the original model. Let $q(x^n \mid \eta) = \sum_{\rho_n} \prod_j \int q(x_j^* \mid \xi_j) dq(\xi_j \mid \eta)$. The two posterior distributions would differ by a factor $q(x^n \mid \eta)$. The implication on the desired inference can be substantial. See the results in Section 6.1 for an example and more discussion. However, the restriction that θ_j and ξ_j be independent is natural. The function $g(x_j^*)$ defines similarity of the experimental units in one cluster and is chosen by the user. There is no notion of learning about the similarity.

3.2. Predictive Inference

A minor complication arises with posterior predictive inference, i.e., 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 $p(\tilde{y} | \tilde{x}, x^n, y^n) = \int p(\tilde{y} | \tilde{x}, \rho_{n+1}, y^n) dp(\rho_{n+1} | \tilde{x}, y^n, x^n)$. 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 re-weighting 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

$$p(\tilde{y} | \tilde{x}, y^n, x^n) \propto \int \sum_{\ell=1}^{k_n+1} p(\tilde{y} | \tilde{x}, y_\ell^*, x_\ell^*, \tilde{e} = \ell) q_\ell(\tilde{x} | x_\ell^*) \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)} p(\rho_n | y^n, x^n) d\rho_n. \quad (8)$$

The first factor reduces to $p(\tilde{y} | y_\ell^*, \tilde{e} = \ell)$ when (6) does not include a regression on x_i in the sampling model. Sampling from (8) is implemented on top of posterior simulation for $\rho_n \sim p(\rho_n | y^n)$. For each imputed ρ_n , generate $\tilde{e} = \ell$ with probabilities proportional to

$$w_\ell = q_\ell(\tilde{x} | x_\ell^*) \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)},$$

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. Reporting Posterior Inference on Clusters

Reporting posterior inference for random partition models is complicated by problems related to the label switching problem. See, Jasra et al. (2005) for

a recent summary and review of the literature. Usually the focus of inference in semi-parametric mixture models like (6) 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. However, in some examples we want to highlight the choice of the clusters. We briefly comment on the implementation of such inference. We use an approach that we found particularly meaningful to report inference about the regression $p(\rho_n | x^n)$. We report inference stratified by k_n , the number of clusters. For given k_n we first find a set of k_n indices $I = (i_1, \dots, i_k)$ with high posterior probability of the corresponding cluster indicators being distinct, i.e., $Pr(D_I | y)$ 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, \dots, i_k as anchors to define cluster labels by restricting posterior simulations to k_n clusters with the units i_j in distinct clusters. Post-processing MCMC output, this is easily done by discarding all imputed parameter vectors that do not satisfy the constraint. We re-label 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 post-processing is similar to the pivotal re-ordering suggested in Marin and Robert (2007, chapter 6.4). An alternative loss function based method to formalize inference on partitions is discussed in Lau and Green (2007).

The outlined approach to cluster identification can be considerably improved. Several recently proposed methods allow efficient inference about the number of clusters. Marin and Robert (2008) propose a clever method for evaluating the marginal $p(k | y)$ for general mixture models. The approach is based on treating the additional randomness that is introduced by the label switching problem as

an ancillary statistic, and reporting a Rao-Blackwellized estimate by averaging over all possible permutations. McCullagh and Yang (2008) discuss identification of clusters under a specific parametric model, the Dirichlet clustering model. The approach could be used to identify a reasonable choice for k_n .

4. Similarity Functions

Continuous covariates: 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 $p(\xi_j) = N(m_j; m, B) \cdot \text{Ga}(v_j^{-1}; \nu, S_0)$, with fixed m, B, ν, S_0 . The main reason for this choice is operational simplicity. A simplified version uses fixed $v_j \equiv v$. The resulting function $g(x_j^*)$ 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 strong we weigh similarity of the x . 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 judgement 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 $c_2 = 10$.

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

$$g(x_j^*) = \int \prod_{i \in S_j} \xi_{j,x_i} dq(\xi_j) = \int \prod_{c=1}^C \xi_{jc}^{n_{jc}} dq(\xi_j), \quad (9)$$

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.

Ordinal covariates: A convenient specification of $g(\cdot)$ for ordinal covariates 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 $-\infty = \gamma_0 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_{C-1} < \gamma_C = \infty$, so that $x = c$ if and only if $\gamma_{c-1} < Z \leq \gamma_c$. We use fixed cutoff values $\gamma_c = c - 1$, $c = 1, \dots, C - 1$, and a normally distributed latent trait, $Z_i \sim N(m_j, v_j)$. Let $\Phi_c(m_j, v_j) = \Pr(\gamma_{c-1} < Z \leq \gamma_c \mid m_j, v_j)$ denote the normal quantiles and let $\xi_j = (m_j, v_j)$ denote the cluster-specific moments of the latent score. We define $q(x_i \mid e_i = j, \xi_j) = \Phi_c(m_j, v_j)$. 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

$$q(x_j^*) = \int \prod_{c=1}^C \Phi_c^{n_{cj}}(m_j, v_j) N(m_j; m, B) \text{Ga}(v_j^{-1}; \nu, S_0) dm_j dv_j,$$

where $n_{cj} = \sum_{i \in S_j} I\{x_i = c\}$.

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

$$g(x_j^*) = \frac{1}{\prod_{i \in S_j} x_i!} \int \xi_j^{\sum_{i \in S_j} x_i} \exp(-\xi_j |S_j|) dq(\xi_j). \quad (10)$$

With a conjugate gamma prior, $q(\xi_j) = \text{Ga}(\xi_j; a, b)$, 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 (3). However, we argue that the additional cluster size penalty that is introduced through $g(\cdot)$ can be ignored in comparison to, for example, the cohesion function $c(S_j) = M(n_j - 1)!$ that is implied by the popular DP prior. The DP prior is biased towards one large cluster and lots of small clusters. This strong a priori preference for large clusters is not desirable for most applications. Thus the additional penalty could be considered a conservative correction in the desired direction.

Proposition 2: The similarity function introduces an additional cluster size penalty in model (3). Consider the case of constant covariates, $x_i \equiv x$, and

let n_j denote the size of cluster j . The default choices of $g(x_j^*)$ for continuous, categorical and count covariates introduce a penalty for large n_j , with $\lim_{n_j \rightarrow \infty} g(x_j^*) = 0$. But the rate of decrease is ignorable compared to the cohesion $c(S_j)$. Let $f(n_j)$ be such that $\lim_{n_j \rightarrow \infty} g(x_j^*)/f(n_j) \geq M$ with $0 < M < \infty$. For continuous covariates the rate is $f(n_j) = (2\pi)^{-\frac{n_j}{2}} V^{-\frac{n_j-1}{2}} (r + n_j)^{\frac{1}{2}}$ 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 $f(n_j) = C^{-\frac{n_j}{2}} (\alpha + n_j x)^{\frac{1}{2}}$ with $C = 2\pi x \exp(1/6x)$.

Proof: see appendix. \square

Another important concern besides the dependence of $g(\cdot)$ on n_j is the dependence on $q(\xi_j)$ in the auxiliary model in (4). In particular, we focus on the dependence on possible hyperparameters ϕ that index $q(\xi_j)$. We write $q(\xi_j | \phi)$ when we want to highlight the use of such (fixed) hyperparameters. The auxiliary models $q(\cdot)$ for all proposed default similarity functions include such hyperparameters. Model (3) and (4) implies conditional cluster membership probabilities

$$p(e_{n+1} = j | x_{n+1}, x^n, \rho_n) \propto \frac{c(S_j \cup \{n+1\})}{c(S_j)} \underbrace{\frac{g(x_j^*, x_{n+1})}{g(x_j^*)}}_{\equiv q_j(x_{n+1} | x_j^*, \phi)} \quad (11)$$

$j = 1, \dots, k_n$, with the convention $g(\emptyset) = c(\emptyset) = 1$. The cluster membership probabilities (11) are asymptotically independent of ϕ . Noting that $q_j(x_{n+1} | x_j^*, \phi)$ can be written as a posterior predictive distribution in the auxiliary model, the statement follows from the asymptotic agreement of posterior predictive distributions.

Proposition 3: Consider any two similarity functions $g^h(x_j^*)$, $h = 1, 2$, based on an auxiliary model (4) with different probability models $q^h(\xi_j)$, $h = 1, 2$, but the same model $q(x_i | \xi_j)$. For example, $q^h(\xi_j) = q(\xi_j | \phi_h)$. Assume that both auxiliary models satisfy the regularity conditions of Schervish (1995, Section 7.4.2) and assume $q(x_i | \xi) \leq K$ is bounded. Let $q_j^h(x_{n+1} | x_j^*)$ denote

the un-normalized cluster membership probability in equation (11) under $q^h(\xi_j)$.

Then

$$\lim_{n_j \rightarrow \infty} q_j^1(x_{n+1} | x_j^*) - q_j^2(x_{n+1} | x_j^*) = 0 \quad (12)$$

The limit is a.s. under i.i.d. sampling from $q(x_i | \xi_0)$.

See the appendix for a straightforward proof making use of asymptotic posterior normality only. The statement is about asymptotic cluster membership probabilities for a future unit only and should not be over-interpreted. The marginal probability for a set of elements forming a cluster does depend on hyperparameters. An example are the expressions for $g(x_j^*)$ in the proof of proposition 2.

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. The proposed model class includes some recently proposed alternatives as special cases, even when these models were not motivated as a modification of the PPM.

Park and Dunson (2007) consider the special case of continuous covariates. They define the desired covariate-dependent random partition model as the posterior random partition under a PPM model for the covariates x_i , i.e., (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 proceed with an augmented response vector (x_i, y_i) , carefully separating the prior for parameters related to the x and y sub-vectors, as in (7). See Section 3.1 for a discussion of the need to separate prior parameters related to x and y .

Another recently proposed approach is Shahbaba and Neal (2007). They use 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 (7) 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.

Another popular class of models that implements similar covariate-dependent clustering are hierarchical mixtures of experts (HME) and related models. HME models were introduced in Jordan and Jacobs (1994). Bayesian posterior inference is discussed in Bishop and Svensén (2003). An implementation for binary outcomes is developed in Wood et al. (2008). We refer the reader to McLachlan and Peel (2000) for a review of mixture of experts models. 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:

$$p(y_i; \eta) = \sum_{j=1}^K w_{ij}(x_i; \alpha_j) N(y_i; \theta_j^T x_i, \sigma_i^2). \quad (13)$$

Here $\eta = (\theta_1, \dots, \theta_K, \sigma_1^2, \dots, \sigma_K^2, \alpha_1, \dots, \alpha_K)$ is the full vector of parameters. The size K of the mixture is pre-specified. The covariate vector x_i includes a 1 for an intercept term. The weights $0 \leq w_{ij} \leq 1$ 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 $\theta_j \sim N(\mu_\theta, V_\theta)$, $\alpha_j \sim N(\mu_\alpha, V_\alpha)$, $\sigma_j^{-2} \sim Ga(a_0, a_1)$, 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 is assumed known. Ad-hoc fixes are possible, such as choosing a very large K , or introducing a hyperprior on K . One could then use the marginal posterior $p(K | y)$ to guide the choice of a fixed K .

Similar to the HME, clusterwise regression models also define mixture models with separate regression functions within each of K subpopulations (DeSarbo and Cron, 1988). Posterior inference for such models and extensions to generalized linear models is summarized in Lenk and DeSarbo (2000). These models are similar to HME models but without the covariate dependent weights.

Dahl (2008) recently proposed another interesting approach to covariate-dependent random clustering. Let $\mathbf{e} = (e_1, \dots, e_n)$ and let \mathbf{e}_{-i} denote the vector of cluster membership indicators without the i -th element. Let $p(e_{n+1} = j | \mathbf{e})$, $j = 1, \dots, k_n + 1$ denote the conditional prior probabilities for the cluster membership indicators. For the random clustering implied by the DP prior, $G \sim DP(M, G^*)$, these conditional probabilities are known as the Polya urn. Let k^- 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 $j = 1, \dots, k^-$, and $M / (n - 1 + M)$ for $j = k^- + 1$. Dahl (2008) defines 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 $d^* = \max_{i < h} d_{ih}$ and define $h_i(S_j) = c \sum_{h \in S_j} (d^* - d_{ih})$ with c chosen to achieve $\sum_{j=1}^{k_n} h_i(S_j) = n$. Dahl (2008) defines the modified Polya urn

$$p(e_i = j | \mathbf{e}_{-i}, x^n) \propto \begin{cases} h_i(S_j^-) & j = 1, \dots, k \\ M & j = k + 1. \end{cases} \quad (14)$$

This set of transition probabilities define 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 in comparison with three of the described alternative approaches: (i) the HME model (13); (ii) the approach of Dahl (2008); and (iii) the model in Park and Dunson (2007).

We set up a simulation study with 3 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 6.2. 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 , visible by comparing the effects of x_1 (top left versus bottom left panels) and x_2 (top left versus top right) with the combined effect of (x_1, x_2) (top left versus bottom right panel). We simulated $M = 100$ data sets 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 $M = 100$ data sets were generated by resampling one big set of $N = 1000$ simulated data points. The big data set and the expectations $E(Y | x_1, x_2, x_3)$ under the simulation truth are available at <http://odin.mdacc.tmc.edu/~pm/>.

We compare the proposed PPMx model and the alternative approaches (i)

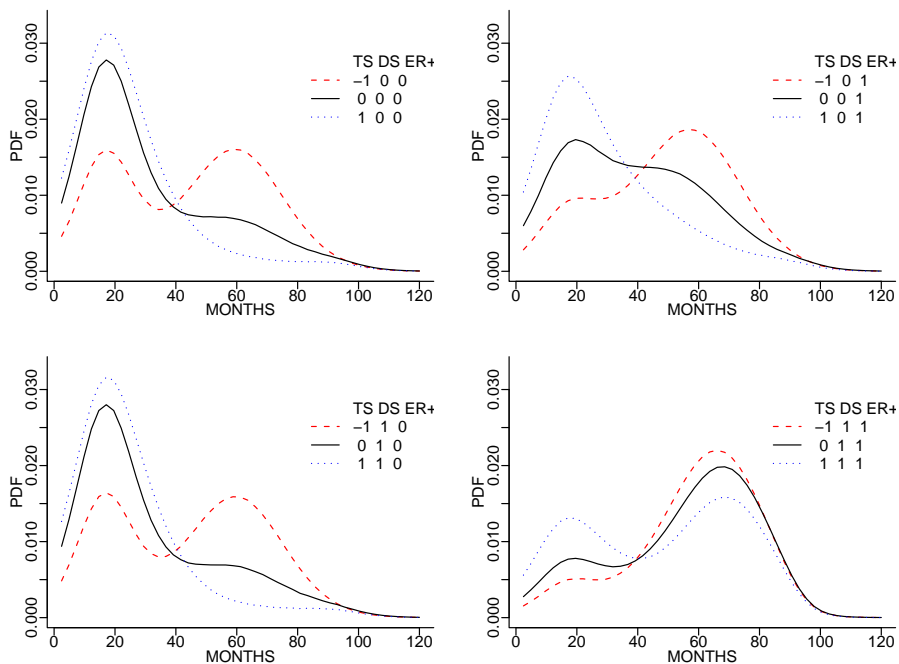


Fig. 1. Simulation truth p.d.f. for the outcome variable for different combinations of the three covariates.

through (iii). 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 data set. 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. Overall there is no clear winner in the compar-

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

x_1	x_2	x_3	RMSE				
			PPMx	P&D	DAHL	HME	
-1	0	0	7.9	2.7	8.6	13.0	
0	0	0	3.9	15.0	4.6	8.3	
1	0	0	2.8	21.5	8.7	8.7	\star
-1	1	0	5.4	2.3	6.2	5.7	
0	1	0	4.6	15.5	7.9	5.5	
1	1	0	4.0	21.0	12.6	5.3	\star
-1	0	1	6.1	1.8	9.5	7.5	
0	0	1	4.2	7.0	2.4	3.5	
1	0	1	4.5	17.4	8.2	4.9	
-1	1	1	9.5	12.1	12.2	6.5	
0	1	1	8.3	8.7	10.2	5.4	
1	1	1	6.2	2.4	4.8	3.8	

ison. In any case we caution against over-interpreting the results for this one example. The main conclusion is that all four approaches are reasonably comparable, and a caveat about extrapolation in the approach by Park and Dunson (2007). The choice of approach should depend on the inference goal. The HME is 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 two 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 Dahl (2008), although a model is implied by the set of conditional distributions (14), there is in general no clearly identified prior probability model for ρ_n . The approach of Park and Dunson (2007) 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 judgement of similarity.

6. Examples

6.1. Simulation Study

We set up a simulation study with a 6-dimensional response y_i , and a continuous covariate x_i , and for $n = 47$ data points (y_i, x_i) , $i = 1, \dots, n$. The simulation is set up to mimic responses of blood counts over time for patients undergoing chemo-immunotherapy. Each data point corresponds to one patient. The six responses for each patient could be, for example, outcomes at six key time points over the course of one cycle of the therapy. The covariate is a treatment dose.

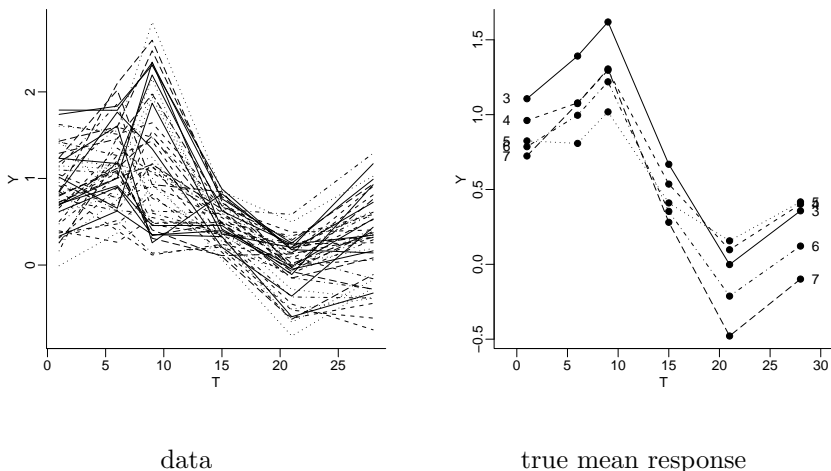


Fig. 2. Simulation example. Data (left panel) and simulation truth for the mean response $E(y_i | x_i)$ as a function of the covariate x_i (right panel).

In the simulation study we sampled x_i from 5 possible values, $x_i \in \{3, 4, 5, 6, 7\}$. Although we only include 5 distinct values in the simulation, we define x_i as a continuous covariate. The reason is the nature of x_i as a continuous dose in the motivating application. For each value of x we defined a different mean vector. Figure 2 (right panel) shows the 5 distinct mean vectors for the 6-dimensional response (plotted against days T). The bullets indicate the six responses. Adding a 6-dimensional normal residual we generated the data shown in Figure 2 (left panel).

We then implemented the PPMx model for covariate-dependent clustering to define a random partition model with a regression on the covariate x_i . Let $N(x; m, s)$ indicate a normal distribution for the random variable x , with moments (m, s) , let $W(V; \nu, S)$ denote a Wishart prior for a random matrix V with scalar parameter ν and matrix parameter S , and let $Ga(s; a, b)$ denote a gamma distribution with mean a/b . Let $\theta_j = (\mu_j, V_j)$. We used a multivariate normal

model $p(y_i | \theta_j) = N(\mu_j, V_j)$, with a conditionally conjugate prior for θ_j , i.e., $p(\theta_j) = N(\mu_j; m_y, B_y) W(V_j^{-1}; s_y, S_y^{-1})$. Let $\text{diag}(x)$ denote a diagonal matrix with x on the diagonal. The model is completed with conjugate hyperpriors for S_y , $S_y \sim W(q, R/q)$ with $q = 50$ and $R = \text{diag}(0.25, \dots, 0.25)$. Let \bar{x} and \widehat{S} denote the empirical mean and covariance matrix of the 6-dimensional response. The hyperparameters for the sampling model are fixed as $m_y = \bar{x}$, $B_y = 4 \cdot \widehat{S}$ and $s_y = 50$. The similarity function is a normal kernel. We use

$$g(x_j^*) = \int \prod_{i \in S_j} N(x_i; m_j, v) N(m_j; m, B) dm_j$$

The hyperparameters m, B, S_0 are fixed. As cohesion function $c(\cdot)$ in (3) we use $c(S_j) = M(n_j - 1)!$, i.e. the random clustering model implied by a DP prior. Hyperparameters for the similarity function are $m = 5$, and $B = 0.1$, and we assume a $\text{Ga}(1, 1)$ hyperprior for the total mass parameter M .

Figure 3 summarizes posterior inference by the posterior predictive distribution for y_{n+1} arranged by x_{n+1} . The marginal posterior distribution for the number of clusters assigns posterior probabilities $p(k = 3 | \text{data}) = 76\%$, $p(k = 4 | \text{data}) = 20\%$ and $p(k = 5 | \text{data}) = 3\%$. We used the procedure described at the end of section 3 to report cluster-specific summaries. Conditional on $k = 3$, the three clusters are characterized by $x_i = 3.4(0.2)$ for the first cluster, $x_i = 5(0)$ for the 2nd cluster and $x_i = 7(0)$ for the third cluster (posterior means and standard deviations of x_i assigned to each cluster). The average cluster sizes are 24, 11 and 11.

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. In this case, the inference is by construction the same for all covariate values.

A technically convenient ad-hoc solution to include covariates in a sampling model is to proceed as if the covariates were part of the response vector. This

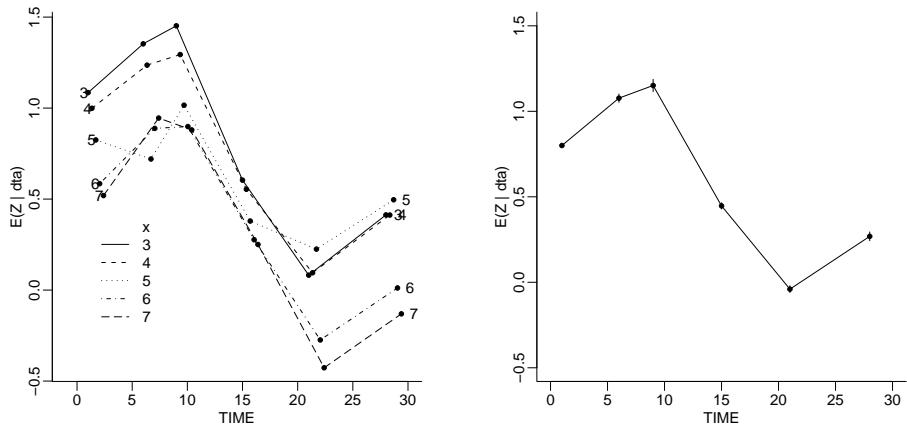


Fig. 3. Simulation example. Estimated mean response under the proposed covariate regression (left panel) and without (right panel).

approach is used, for example in Mallet et al. (1988) or Müller et al. (2004). Doing so introduces an additional factor in the likelihood. Let y generically denote the response, x the covariate, and θ the model parameters. Treating the covariate as part of the response vector is equivalent to replacing the sampling model $p(y | x, \theta)$ by $p(y | x, \theta) \cdot p(x | \theta)$. Equivalently, one can interpret the additional factor $p(x | \theta)$ as part of the prior. The reported posterior inference is as if we had changed the prior model from the original $p(\theta)$ to $\tilde{p}(\theta) \propto p(\theta)p(x | \theta)$. Wong et al. (2003) use this interpretation for a similar construction in the context of prior probability models for a positive definite matrix. Compare with the discussion in Section 3.1. Note that in $p(y | x, \theta) \cdot p(x | \theta)$ the parameters of the original sampling model $p(y | \cdot)$ and $p(x | \cdot)$ are not necessarily separated. In the example below they share the parameter S_y .

The implied modification of the prior probability model could be less innocuous than what it seems. We implemented inference under the PPM prior model (1), without covariates, and a sampling model as before, but now for an

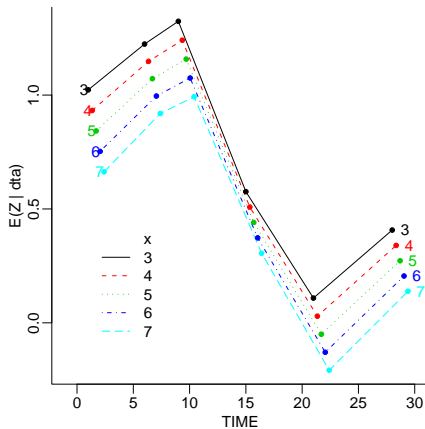


Fig. 4. Prediction for \tilde{y} arranged by \tilde{x} for the covariate response model with the additional factors $p(x_j^n | \theta_j, \eta)$ in the prior.

extended response vector augmented by the covariate x_i . We assume a multivariate normal sampling model $p(y_i, x_i | \theta_j) = N(\mu_j, V_j)$, with a conditionally conjugate prior for θ_j , i.e., $p(\theta_j) = N(\mu_j; m_y, B_y) W(V_j^{-1}; s, S_y^{-1})$. As before the model is completed with a conjugate hyperpriors for S_y . The hyperparameters are chosen exactly as before. The additional 7-th row and column of the prior means for B_y are all zero except for the (7,7) diagonal element which we fix at v to match the moments in the auxiliary model $q(\cdot)$ from before. Similarly, we fix $m_{y7} = m$. We refer to the model as *augmented response* model. Let $p(x_j^* | S_y)$ denote the marginal distribution of x_j^* under the augmented response model. The augmented response model differs from the PPMx by an additional factor $\prod_j p(x_j^* | S_y)$ in the augmented response model. The parameter S_y plays the role of η in the general discussion in 3.1. Figure 4 shows inference under this modified model. In this example, the augmented response model leads to a much reduced size of the partition. With posterior probability 97% the par-

tition includes only one cluster, $p(n_1 = 47 \mid data) = 0.97$. This in turn leads to essentially a simple linear regression implied by the dominating multivariate normal $p(x_i, y_i \mid \theta_1)$, as clearly seen in Figure 4. In contrast, under the proposed PPMx model the size of the largest cluster is estimated between 16 and 26 (not shown). The augmented response model could be modified to better match the simulation truth. This could be achieved, for example, by fixing the hyperparameters in a way that avoids correlation of the first 6 and the last dimension of μ_j . The resulting model would have exactly the format of (3).

We thus 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 and that this does not imply restricting the scope of the proposed models. This is illustrated in the following data analysis example.

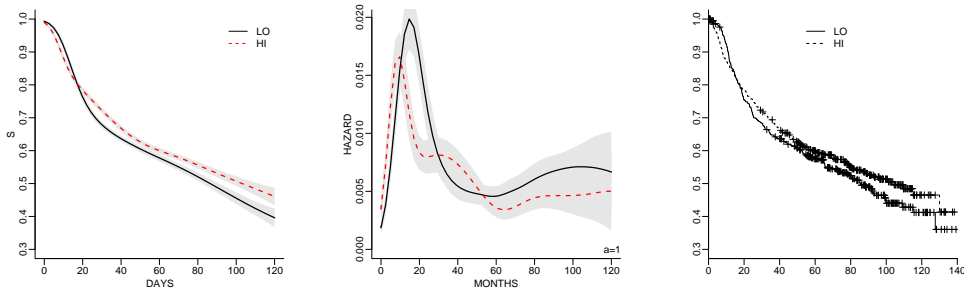
6.2. A Survival Model with Patient Baseline Covariates

We consider data from a high-dose chemotherapy treatment of women with breast cancer. The data for this particular study have been discussed in Rosner (2005) and come from Cancer and Leukemia Group B (CALGB) Study 9082. It consists of measurements taken from 763 randomized patients, available as of October 1998 (enrollment had occurred between January 1991 to 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 con-

siderable regenerative blood-cell supportive care (including marrow transplantation) to help decreasing 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).

First we carried out inference in a model using the indicator for high-dose as the only covariate, i.e. $x_i = \text{HI}$. We implemented model (6) with a similarity function for the binary covariate based on the beta-binomial model (9). 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 $p(\theta_j | \eta) = N(\mu_j | m_y, B_y) \text{Ga}(V_j^{-1} | s/2, sS_y/2)$ and hyperprior $m_y \sim N(a_m, A_m)$, $S_y \sim \text{Ga}(q, q/R)$. Here $\eta = (a_m, A_m, B_y, q, R)$ are fixed hyperparameters. We use $a_m = \hat{m}$, 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 $\text{Ga}(1, 1)$ hyperprior for the total mass parameter M .



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

hazard $h(t)$

data (KM)

Fig. 5. *Survival example: Estimated survival function (left panel) and hazard (center panel), arranged by $x \in \{HI, LO\}$. The grey shades show point-wise one posterior predictive standard deviation uncertainty. The right panel shows the data for comparison (Kaplan-Meier curve by dose).*

Figure 5 shows inference summaries. The posterior distribution $p(k_n \mid \text{data})$ for the number of clusters is shown in Figure 6. The three largest clusters contain 28%, 23%, and 14% of the experimental units.

Next we extended the covariate vector to include all six covariates. Denote by $x_i = (x_{i1}, \dots, x_{i6})$ the 6-dimensional covariate vector. We implement random clustering with regression on covariates as in model (6). The similarity function is defined as

$$g(x_j^*) = \prod_{\ell=1}^6 g^\ell(x_{j\ell}^*). \quad (15)$$

The multiplicative nature of (15) implies that there are no interactions of covariates in the prior probability model $p(\rho_n \mid x^n)$. A posteriori interactions are possible and likely.

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 function for the three binary covariates is defined as in (9) with $\alpha =$

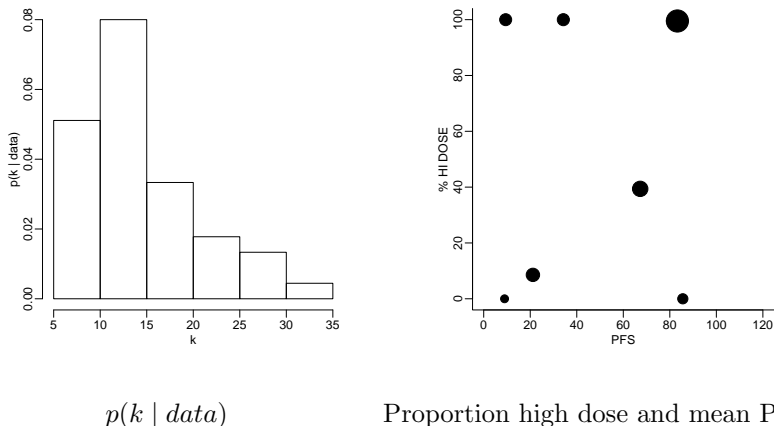


Fig. 6. *Survival example: Posterior for the number of clusters k (left panel), and proportion of patients with high dose (%HI) and average progression free survival (PFS) by cluster. The size (area) of the bullets is proportional to the average cluster size.*

(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 (10) 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 7. In the figure, “baseline” refers to HI = 0, tumor size 38mm (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 26mm (the empirical first quartile), TS+ indicating tumor size 50mm (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 8 summarizes features of the posterior clustering. Interestingly, clusters are typically highly correlated to the postmenopausal status, as seen in the right panel. The high-dose indicator is also seen to be positively correlated to the progression-free survival (PFS).

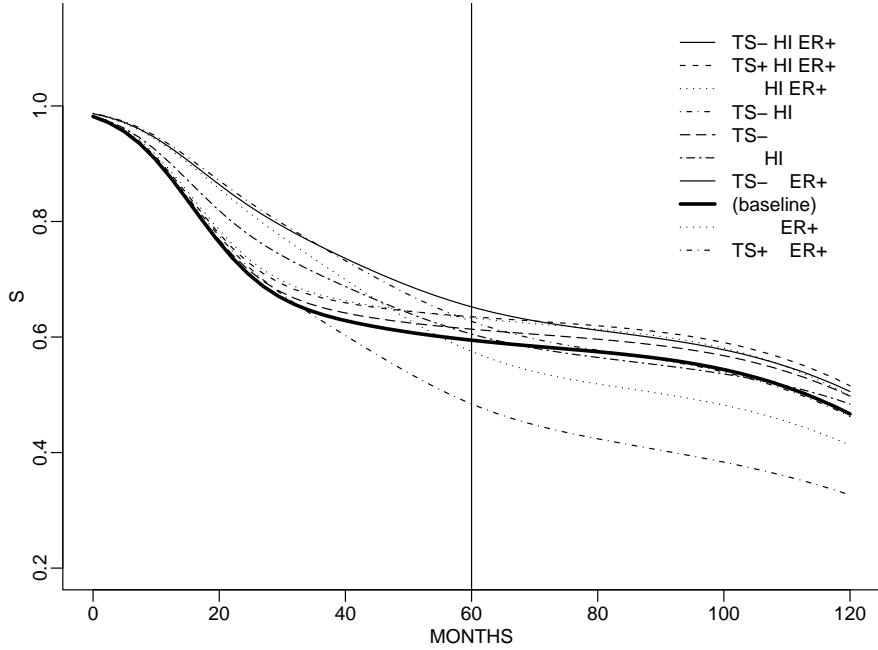
7. Conclusion

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 $c^*(|S_j|)$ in (6) to compensate the size penalty implicit in the similarity function.

The programs are available as a function in the R package PPMx at <http://odin.mdacc.tmc.edu/~pm/prog.html> The function `PPMx(.)` imple-



$$S(t | x)$$

Fig. 7. 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 26mm and 50mm (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.

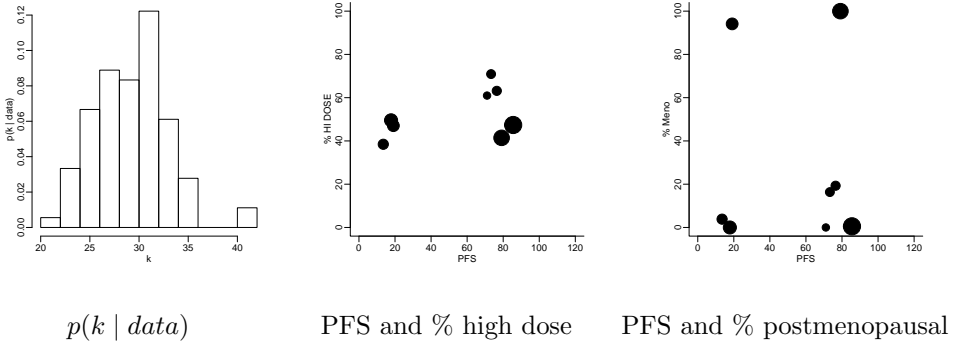


Fig. 8. *Survival example: Posterior distribution for the number of clusters (left panel), mean PFS and % high dose patients per cluster (center panel), mean PFS and % postmenopausal patients per cluster (right panel).*

ments 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 .

Acknowledgment

Research was partially supported by NIH under grant 1R01CA75981, by FONDECYT under grant 1060729 and the Laboratorio de Análisis Estocástico PBCT-ACT13. Most of the research was done while the first author was visiting at Pontificia Universidad Católica de Chile.

Appendix

Proof of Proposition 2

For simplicity we drop the j index in n_j , x_j^* etc., relying on the context to prevent ambiguity. Let $\mathbf{1}$ denote a $(n \times 1)$ vector of all ones. For continuous covariates,

evaluation of the $g(x^*)$ gives

$$g(x^*) = (2\pi)^{-\frac{n}{2}} [(V + nB)V^{n-1}]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(x - m)^2 \frac{1}{V} \mathbf{1}' \left(I - J \frac{B}{V + nB} \right) \mathbf{1} \right\}$$

which after some further simplification becomes $(2\pi V^{\frac{n-1}{n}})^{-\frac{n}{2}} (\rho + n)^{\frac{1}{2}} M(n)$ with $\lim_{n \rightarrow \infty} M(n) = M$, $0 < M < \infty$. For categorical covariates, Stirling's approximation and some further simplifications gives

$$g(x^*) = \frac{\Gamma(A)}{\prod \Gamma(\alpha_c)} \frac{\prod \Gamma(\alpha_c) \frac{\Gamma(\alpha_x + n)}{\Gamma(\alpha_x)}}{\Gamma(A + n)} \geq \frac{1}{A + n^{A - \alpha_x}} M(n).$$

Similarly, for count covariates we find

$$g(x^*) = \left(\frac{1}{x!} \right)^n \frac{b^a}{\Gamma(a)} \frac{\Gamma(a) + nx}{(b + n)^{a + nx}} \geq (2\pi x)^{-\frac{n}{2}} e^{-\frac{n}{12x}} (\alpha + nx)^{\frac{1}{2}} \cdot M(n).$$

Proof of Proposition 3

The ratio $g^h(x_j^*, x_{n+1})/g^h(x_j^*)$ defines the conditional probability $q_j^h(x_{n+1} | x_j^*)$ under i.i.d. sampling in the auxiliary model, and thus $q_j^h(x_{n+1} | x_j^*) = \int q(x_{n+1} | \xi_j) q^h(\xi_j | x_j^*) d\xi_j$.

The result follows from asymptotic normality of $q^h(\xi_j | x_j^*)$, as $n_j \rightarrow \infty$. Let $\widehat{\xi}_j$ denote the m.l.e. for ξ_j based on n_j observations in cluster j . Let $LL' = [-I(\widehat{\xi}_j)]^{-1}$ denote a Choleski decomposition of the negative inverse of the observed Fisher information matrix (Schervish, 1995, equation 7.88), and let $\psi_n = L(\xi_j - \widehat{\xi}_j)$. Theorem 7.89 of Schervish (1995) implies for any $\epsilon > 0$ and any compact subset B of the parameter space:

$$\lim_{n_j \rightarrow \infty} P_{\xi_{j0}} \underbrace{\left(\sup_{\psi_n \in B} |q^1(\psi_n | x_j^*) - q^2(\psi_n | x_j^*)| > \epsilon \right)}_{\pi_n} = 0.$$

The limit is in the cluster size n_j . The probability is under an assumed true sampling model $q(x_i | \xi_j^o)$, and the supremum is over B .

Consider a sequence of reparametrizations of $q(x_i | \xi_j)$ to $q(x_i | \psi_n)$. Let $B_M = \{\xi_j : |\xi_j^o - \xi_j| < M\}$ be an increasing sequence of compact sets, and recall the definition of $\pi_n = P_{\xi_j^o}(\sup_B \dots > \epsilon)$. Then

$$\lim_n \int q(x_{n+1} | \xi_j) (q^2(\xi_j | x_j^*) - q^1(\xi_j | x_j^*)) d\xi_j \leq \lim_M \lim_n \epsilon K(1 - \pi_n) + \pi_n K = \epsilon K,$$

for any $\epsilon > 0$.

References

- Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2**, 1152–1174.
- Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.
- Barry, D. and Hartigan, J. A. (1993) A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, **88**, 309–319.
- Bernardo, J.-M. and Smith, A. F. M. (1994) *Bayesian theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Chichester: John Wiley & Sons Ltd.
- Bishop, C. M. and Svensén, M. (2003) Bayesian hierarchical mixtures of experts. In *2003 Proceedings Nineteenth Conference on Uncertainty in Artificial Intelligence* (eds. U. Kjaerulff and C. Meek), 57–64.
- Dahl, D. B. (2003) Modal clustering in a univariate class of product partition models. *Tech. Rep. 1085*, Department of Statistics, University of Wisconsin.
- (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.
- Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002) *Bayesian methods for nonlinear classification and regression*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd.
- DeSarbo, W. S. and Cron, W. L. (1988) A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**, 249–282.
- Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.
- Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, **97**, 611–631.
- Green, P. J. and Richardson, S. (1999) Modelling Heterogeneity with and without the Dirichlet Process. *Tech. rep.*, University of Bristol, Department of Mathematics.
- Green, P. J. and Sibson, R. (1978) Computing Dirichlet tessellations in the plane. *Comput. J.*, **21**, 168–173.
- Hartigan, J. A. (1990) Partition models. *Communications in Statistics, Part A – Theory and Methods*, **19**, 2745–2756.
- Ishwaran, H. and James, L. F. (2003) Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica*, **13**, 1211–1235.
- Jasra, A., Holmes, C. C. and Stephens, D. A. (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.*, **20**, 50–67.

- Johnson, V. E. and Albert, J. H. (1999) *Ordinal data modeling*. Statistics for Social Science and Public Policy. New York: Springer-Verlag.
- Jordan, M. and Jacobs, R. (1994) Hierarchical mixtures-of-experts and the em algorithm. *Neural Computation*, **6**, 181–214.
- Kim, H.-M., Mallick, B. K. and Holmes, C. C. (2005) Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, **100**, 653–668.
- Lau, J. W. and Green, P. J. (2007) Bayesian Model-Based Clustering Procedures. *Journal of Computational and Graphical Statistics*, **16**, 526–558.
- Lenk, P. J. and DeSarbo, W. S. (2000) Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, **65**, 93–119.
- Mallet, A., Mentré, F., Gilles, J., Kelman, A., Thomson, A., S.M., Bryson and Whiting, B. (1988) Handling covariates in population pharmacokinetics with an application to gentamicin. *Biomedical Measurement Informatics and Control*, **2**, 138–146.
- Marin, J.-M. and Robert, C. (2008) Approximating the marginal likelihood in mixture models. *Bulletin of the Indian Chapter of ISBA*, to appear.
- Marin, J.-M. and Robert, C. P. (2007) *Bayesian Core. A Practical Approach to Computational Bayesian Statistics*. New York: Springer-Verlag.
- McCullagh, P. and Yang, J. (2008) How many clusters? *Bayesian Analysis*, **03**, 101 – 120.
- McLachlan, G. and Peel, D. (2000) *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

- Müller, P., Quintana, F. and Rosner, G. (2004) A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **66**, 735–749.
- Okabe, A., Boots, B., Sugihara, K. and Chiu, S. N. (2000) *Spatial tessellations: concepts and applications of Voronoi diagrams*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd., second edn. With a foreword by D. G. Kendall.
- Park, J.-H. and Dunson, D. (2007) Bayesian generalized product partition models. *Tech. rep.*, Duke University.
- Pitman, J. (1996) Some Developments of the Blackwell-MacQueen Urn Scheme. In *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell* (eds. T. S. Ferguson, L. S. Shapley and J. B. MacQueen), 245–268. Hayward, California: IMS Lecture Notes - Monograph Series.
- Quintana, F. A. (2006) A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, **136**, 2407–2429.
- Quintana, F. A. and Iglesias, P. L. (2003) Bayesian Clustering and Product Partition Models. *Journal of The Royal Statistical Society Series B*, **65**, 557–574.
- 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, Series B*, **59**, 731–792.
- Rosner, G. L. (2005) Bayesian monitoring of clinical trials with failure-time endpoints. *Biometrics*, **61**, 239–245.
- Shahbaba, B. and Neal, R. M. (2007) Nonlinear Models Using Dirichlet Process Mixtures. *Tech. rep.*, University of Toronto.

Wong, F., Carter, C. K. and Kohn, R. (2003) Efficient estimation of covariance selection models. *Biometrika*, **90**, 809–830.

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.