Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 42: 1065–1077, 2015
doi: 10.1111/sjos.12151

Cluster-Specific Variable Selection for Product Partition Models

FERNANDO A. QUINTANA

Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile

PETER MÜLLER

Department of Mathematics, The University of Texas at Austin

ANA LUISA PAPOILA

CEAUL and Faculdade de Ciências Médicas da Universidade Nova de Lisboa

ABSTRACT. We propose a random partition model that implements prediction with many candidate covariates and interactions. The model is based on a modified product partition model that includes a regression on covariates by favouring homogeneous clusters in terms of these covariates. Additionally, the model allows for a cluster-specific choice of the covariates that are included in this evaluation of homogeneity. The variable selection is implemented by introducing a set of cluster-specific latent indicators that include or exclude covariates. The proposed model is motivated by an application to predicting mortality in an intensive care unit in Lisboa, Portugal.

Key words: clustering, non-parametric regression, random partition model

1. Introduction

We discuss a novel approach for variable selection in random partition models in an application to predict mortality in intensive care units (ICUs). We build on the covariate-dependent product partition models (PPMs) (PPMx) of Müller *et al.* (2011). We first simplify the regression problem by partitioning the population into more homogeneous subsets (clusters) and then consider a different mean response or simplified regression model within each subset. The random partitions are defined such that subsets of experimental units with similar covariates are more likely to co-cluster. This is achieved with the use of similarity functions that evaluate the homogeneity of covariate values for experimental units in a cluster. In the application to ICU mortality, we modify the PPMx by adding cluster-specific inclusion of covariates in the similarity function.

The motivating application is the prediction of mortality for ICU patients. Over the past years, it has been demonstrated that this is not an easy task. One of the main issues concerns a noticeable lack of fit for some subgroups of patients. Our main inference goal for the upcoming discussion is thus prediction. We are concerned with predicting at-risk patients. Inference for variable inclusion and cluster-specific summaries are informative but do not constitute the main inference goal. The novelty of this article is the use of cluster-specific variable selection and the case study with the ICU data. Variable selection is critical for the application to the ICU data with a moderately large number of candidate covariates.

The rest of this article is organized as follows. Section 2 reviews the PPMx model and describes in detail the proposed variable selection scheme and some of the options that are available to define its components. Section 3 illustrates the procedure using mortality data for an ICU at Centro Hospitalar de Lisboa Central in Lisboa, Portugal. In Section 4, two alternative approaches are presented. We conclude with some final remarks in Section 5. Online

supporting information includes further comparison and a simulation study that validates the predictive and fitting capabilities of the modelling approach.

2. Variable selection in the PPMx model

2.1. The PPMx model

Let $\{(y_i, x_i) : i = 1, \ldots, n\}$ denote the available data, where y_i represents the outcome for experimental unit (subject) i and x_i are subject-specific covariates. In many applications, it is reasonable and useful to think of the population as being partitioned into subpopulations that are more homogeneous than the overall population. One possible modelling framework that follows this structure is given by the class of PPMs (Hartigan, 1990). Let ρ_n denote a partition of a set of indices $i = 1, \ldots, n$ into k_n disjoint non-empty subsets or *clusters*, $\rho_n = \{S_1, \ldots, S_{k_n}\}$. Let $y^n = (y_1, \ldots, y_n)$ and $x^n = (x_1, \ldots, x_n)$ denote the complete set of observed responses and recorded covariates, and let $y_j^\star = (y_i,\ i \in S_j)$ and $x_j^\star = (x_i,\ i \in S_j)$ denote responses and covariates arranged by cluster. In the case of a multivariate covariate vector $x_i = (x_{i1}, \ldots, x_{ip})$, we use $x_{j\ell}^\star = (x_{i\ell},\ i \in S_j)$ for the ℓth coordinate. The PPM assumes that

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

where $c(S)$ is the *cohesion function* that describes how likely the elements of S are thought to be grouped together *a priori*. Choosing $c(S) = M \times (|S| - 1)!$ for some number $M > 0$, where $|S|$ is the cardinality of S, implies the same prior on ρ_n that is implied by the ties arising under sampling from a (discrete) random probability measure with a Dirichlet process prior (Ferguson, 1973). The model is completed with a sampling model $p(y^n \mid \rho_n) = \prod_{j=1}^{k_n} p_j\left(y_j^\star\right)$, where each $p_j(\cdot)$ may itself depend on $x_j^\star$. For more discussion, see Müller & Quintana (2010). If $p\left(y_j^\star\right)$ is assumed exchangeable, we write $p(y^\star)$ as $\prod_{i \in S_j} p\left(y_i \mid \theta_j^\star\right)$ with a prior model $\prod_j p\left(\theta_j^\star\right)$. In many applications, the sampling model includes a regression on $x_j^\star$. In such cases, the PPM includes dependence on covariates only through the sampling model and is therefore not meaningful for prediction as there is no notion of pairing a future observation with similar historical ones.

An extension discussed in the study of Müller *et al.* (2011), the PPMx, addresses this limitation by introducing dependence on covariates in the prior distribution on partitions. The goal is a model where two subjects are more likely to co-cluster *a priori* if their corresponding covariate values are similar. Define the *similarity function* $g(x_j^\star)$ as any non-negative function such that larger values are associated to sets of covariates that are deemed similar. The PPMx defines a probability model for partitions as

$$p(\rho_n \mid x^n) \propto \prod_{j=1}^{k_n} g\left(x_j^\star\right) \cdot c(S_j), \tag{2}$$

with normalization constant $\sum_{\rho_n} \prod_{j=1}^{k_n} g\left(x_j^*\right) c(S_j)$ and a likelihood model as before.

Müller *et al.* (2011) considered defining $g(\cdot)$ in terms of an auxiliary probability model $q(\cdot)$, that is, $g\left(x_j^\star\right) = \int \prod_{i \in S_j} q\left(x_i \mid \xi_j^\star\right) q\left(\xi_j^\star\right) d\xi_j^\star$. For multivariate $x_i = (x_{i\ell}, \ell = 1, \ldots, p)$, we use $g\left(x_j^\star\right) = \prod_\ell g_\ell\left(x_{j\ell}^\star\right)$ with covariate-specific $g_\ell\left(x_{j\ell}^\star\right) = \int \prod_{i \in S_j} q\left(x_{i\ell} \mid \xi_{j\ell}^\star\right) q\left(\xi_{j\ell}^\star\right) d\xi_{j\ell}^\star$, constructed as before. A conjugate pair for $q(x \mid \xi^\star)$ and

$q(\xi^\star)$ facilitates easy analytic evaluation of $g\left(x_j^\star\right)$. Our implementation uses a minor variation of this procedure. We define $g(x_j^\star)$ as the posterior predictive distribution of $x_j^\star$ in cluster S_j

$$g\left(x_j^\star\right) = \int \prod_{i \in S_j} q\left(x_i \mid \xi_j^\star\right) q\left(\xi_j^\star \mid x_j^\star\right)\, d\xi_j^\star, \tag{3}$$

with $q\left(\xi_j^\star \mid x_j^\star\right) \propto \prod_{i \in S_j} q\left(x_i \mid \xi_j^\star\right) q\left(\xi_j^\star\right)$. We refer to (3) as the 'double-dipping' similarity function. Here, we stress that $g(\cdot)$ is chosen to reflect prior and preferences for the formation of clusters. It need not be a probability model. There is no notion of modelling a joint distribution (exposure model) of the covariates. This allows us choices like (3), which would be inappropriate as a probability model because of the double use of the covariate values.

2.2. Variable selection

For the analysis of the ICU mortality data, we modify the model by adding cluster-specific inclusion of covariates in the similarity function. The goal is to allow different clusters to be characterized by cluster-specific choices of covariates. Characterizing clusters with small sets of covariates facilitates an interpretation of clusters as meaningful patient subpopulations.

Bayesian variable selection has a long history. The special case of linear models has been extensively studied, and many methods have been proposed and compared. Perhaps, the most widely used method is the stochastic search variable selection of George & McCulloch (1993). For a concise review, see Chipman *et al.* (2001). In non-parametric Bayesian applications, Chung & Dunson (2009) discussed variable selection for continuous predictors in the weights of a probit stick-breaking model. This induces a dependence of the partition structure on covariates, but the corresponding probability model for partitions is only indirectly implied. Perhaps, closest to the proposed model, Hoff (2005) proposes a modification of the popular Polya urn to include a selection of a subset of a high dimensional response vector, but the proposed variable selection is for the response vector rather than covariates.

Our approach builds on the PPMx models described in Section 2.1. The idea is similar to classification and regression trees (Chipman *et al.*, 1998; Denison *et al.*, 1998) in that we characterize homogeneous subsets of experimental units (patients, in our case) by some subset of covariates. However, instead of defining subsets deterministically by covariate thresholds, we use PPMx to define random subsets. We modify the prior clustering structure by introducing binary indicators $\gamma_{j\ell}^*$ for the jth cluster and ℓth covariate. Specifically, we assume that

$$p(\rho_n \mid \boldsymbol{\gamma}, x^n) \propto \prod_{j=1}^{k_n} \left\{ c(S_j) \times \left(\prod_{\ell=1}^{p} g\left(x_{j\ell}^\star\right)^{\gamma_{j\ell}^*} \right) \right\}. \tag{4}$$

Model (4) defines a prior distribution on partitions, indexed by covariates with cluster-specific sets of covariates, that retains the product form. Covariates can be active $\left(\gamma_{j\ell}^* = 1\right)$ or inactive $\left(\gamma_{j\ell}^* = 0\right)$. In the special case that all the $\gamma_{j\ell}^*$ indicators are zero, (4) reduces to the product distribution (1). Model (4) provides great flexibility. The PPMx formulation allows for partitions to be driven by covariate values, with more homogeneous covariate values being more likely to be grouped together, as formalized by the similarity function. Additionally, each cluster may be described by different subsets of covariates, which allows for parsimony at a local level. This formalizes the notion that meaningful subpopulations of patients might be characterized by different subsets of attributes.

The prior is structured to share information across clusters. We specify the prior for the binary indicators hierarchically. Let $z_{j\ell}^\star = \text{logit}\Pr\left(\gamma_{j\ell}^\star = 1\right)$ denote the logit of the inclusion probability for covariate ℓ in cluster j. We assume a hierarchical prior of the form

$$z^*_{j\ell} \mid \mu_{z,j} \overset{\text{iid}}{\sim} N\left(\mu_{z,j}, S_z^2\right), \qquad \mu_{z,j} \overset{\text{iid}}{\sim} N\left(\mu_{z,0}, S_{z,0}^2\right), \tag{5}$$

so that information about important covariates can be shared across clusters. Here, hyperparameters $\mu_{z,0}$, S_z^2 and $S_{z,0}^2$ are assumed known.

Adding covariate selection in the prior (4) requires that the similarity function be calibrated such that $g_\ell\left(x^\star_{j\ell}\right) > 1$ for a set of covariates $x^\star_{j\ell}$ that are judged to be similar. A prior model with $g_\ell\left(x^\star_{j\ell}\right) < 1$ for all $x^\star_{j\ell}$ would *a priori* always favour no covariate selection and fail to formalize any prior preference for homogeneous clusters. We therefore use

$$\tilde{g}_\ell\left(x^*_{j\ell}\right) = \frac{g_\ell\left(x^*_{j\ell}\right)}{\prod_{i \in S_j} q\left(x_{i\ell} \mid \bar{\xi}_\ell\right)}, \tag{6}$$

for some conveniently chosen value of $\bar{\xi}_\ell$ such as the maximum likelihood estimation of ξ_ℓ given the complete set of covariate values x^n under the auxiliary model $q(x^n \mid \xi)$. Posterior simulation is greatly facilitated by the choice of (6), which has an important computational advantage. Candidate's formula, that is, the equation for the marginal model that is implied by Bayes theorem, gives

$$\tilde{g}\left(x^*_j\right) = \frac{q\left(\bar{\xi}\right)}{q\left(\bar{\xi} \mid x^*_j\right)}. \tag{7}$$

The equality is true for any $\bar{\xi}$ on the right-hand side. The important advantage over direct evaluation of the marginal is that the dimension of ξ does not vary with cluster size. For efficient implementation of posterior Markov chain Monte Carlo (MCMC) simulation, it is also important to note that the PPMx is a special case of PPM, as can be immediately seen from (2). In particular, we can easily adapt Algorithm 8 in the study of Neal (2000) for the step that consists of the resampling of configurations and cluster-specific parameters.

3. Analysis of the ICU mortality data

3.1. Data

Evaluating severity of illness and predicting mortality are a major concern in ICUs. Predictive scoring systems have been developed to measure the severity of disease and the prognosis of ICU patients. Three validated severity scoring systems have emerged: the acute physiologic and chronic health evaluation system (Zimmerman & Kramer, 2006), the simplified acute physiologic score (SAPS) (Moreno *et al.*, 2005) and the mortality probability models score (Lemeshow *et al.*, 1993). However, several studies revealed a poor prognostic performance of these systems, and several approaches have been used to try to improve their predictive ability (Metnitz *et al.*, 2005). This is why a clustering approach, like the one proposed in this study, seems promising.

We analyse data from $n = 996$ patients, consecutively admitted to a Portuguese mixed (medical and surgical) ICU. The study is discussed by Papoila *et al.* (2013). All SAPS II data were collected during the first 24 hours after ICU admission. The SAPS II is a severity of illness score that includes 17 variables: 12 physiology variables (heart rate, systolic blood pressure, body temperature, the ratio Pao2/Fio2 for ventilated patients, urinary output, serum urea level, white blood cell count, serum potassium, serum sodium level, serum bicarbonate level, bilirubin level and Glasgow coma score), age, type of admission (scheduled surgical, unscheduled

surgical or medical) and three underlying disease variables [acquired immunodeficiency syndrome (AIDS), metastatic cancer and haematologic malignancy]. We coded the covariates as 1–17, as shown in Table 1. Covariates include a variety of different data formats. Covariates 1–12 are continuous, 13–14 are categorical, with three levels each, and 15–17 are binary. The median age of the patients was 64.0 years (interquartile range: 48–73.25) with a median SAPS II score of 41 (interquartile range: 20–60) and a hospital mortality of 36%. All variables are recorded in a vector of $p = 17$ covariates $x_i = (x_{i,1}, \ldots, x_{i,17}), i = 1, \ldots, n$.

3.2. *Model specification*

Let y_i denote a binary indicator of death for the ith patient admitted to the ICU, $i = 1, \ldots, n$, with $y_i = 1$ if the patient died. The proposed joint probability model for cluster-specific variable selection includes a sampling model expressed as a logistic regression for mortality with cluster-specific parameters. We recode each categorical covariate using two binary indicators and add an extra intercept term, so that the resulting design vector $\tilde{x}_i$ has dimension 20. In the upcoming discussion, it is convenient to use an alternative parametrization of partitions with cluster membership indicators $e_i \in \{1, \ldots, k_n\}$ with $e_i = j$ if $i \in S_j$, so that $(k_n, e_1, \ldots, e_n)$ describes the partition. Given a partition ρ_n, thus represented as $(k_n, e_1, \ldots, e_n)$, we assume that

$$\text{logit}\left\{p\left(y_i = 1 \mid e_i = j, \beta_j^{\star}\right)\right\} = \beta_j^{*'} \tilde{x}_i. \tag{8}$$

We refer to (8) as cluster-specific logistic regression. The inclusion of regression in the cluster-specific sampling model leads to a more parsimonious partition, with fewer and larger clusters. The subsets defined by the partition are only needed to accommodate non-linear and interaction effects that are not included in (8). The prior model includes a hierarchical prior on the regression coefficients

$$p\left(\beta_1^*, \ldots, \beta_{k_n}^* \mid \beta_0, B_\beta\right) = N(\beta_0, B_\beta), \qquad \beta_0 \sim N(\beta_{00}, B_{\beta 0}), \tag{9}$$

for known β_{00} and $B_{\beta 0}$. The model is completed by the product distribution (4) with cohesion function $c(S)$ and the variable selection prior (5). The full joint probability model is

Table 1. *Covariates considered in the intensive care unit application*

Code	Variable description
1	Body temperature
2	Blood pressure
3	Heart rate
4	Serum potassium level
5	Serum sodium level
6	Urinary output
7	Glasgow coma score
8	Age
9	Serum urea level
10	Serum bicarbonate level
11	Bilirubin level
12	White blood cell count/1000
13	Type of admission
14	Ventilation risk
15	Haematologic malignancy
16	Metastatic cancer
17	Acquired immunodeficiency syndrome

$$p(y^n, \rho_n, \beta^\star, z^\star \mid x^n) \propto \prod_{j=1}^{k_n} \left\{ c(S_j) N\left(\mu_{z,j}; \mu_{z,0}, S^2_{z,0}\right) N\left(\beta^*_j; \beta_0, B_\beta\right) \times \prod_{i \in S_j} \left[\left\{ \frac{\exp\left(y_i \beta^{*\prime}_j \tilde{x}_i\right)}{1+\exp\left(\beta^{*\prime}_j \tilde{x}_i\right)} \right\} \prod_{\ell=1}^{p} \left(\tilde{g}_\ell\left(x^\star_{j\ell}\right)^{\gamma^*_{j\ell}} \left\{ \frac{\exp\left(\gamma^*_{j\ell} z^*_{j\ell}\right)}{1+\exp\left(z^*_{j\ell}\right)} \right\} N\left(z^*_{j\ell}; \mu_{z,j}, S^2_z\right) \right) \right] \right\} \tag{10}$$

where the $\tilde{g}_\ell$ similarity functions are as in (6), β^*_j and $\gamma^*_{j\ell}$ are cluster-specific and covariate-specific auxiliary model and variable selection parameters and $N(\cdot; \mu, S)$ stands for the normal density with mean μ and covariance matrix S. The ratios $\exp\left(y_i \beta^*_j \tilde{x}_i\right) / \left(1+\exp\left(\beta^*_j \tilde{x}_i\right)\right)$ are the mortality probabilities $p\left(y_i \mid x_i, e_i = j, \beta^*_j\right)$. The ratios $\exp\left(\gamma^*_{j\ell} z^*_{j\ell}\right) / \left(1+\exp\left(\gamma^*_{j\ell} z^*_{j\ell}\right)\right)$ are the variable selection probabilities $p\left(\gamma^*_{j\ell} \mid z^\star_{j\ell}\right)$.

3.3. Results

We apply model (10) using the covariates described earlier, with similarity function (3). We complete the model specification with a cohesion function $c(S)$. We use $c(S) = M \times (|S|-1)!$, that is, the cohesion function that is implied by the Dirichlet process model. In the online supporting information, we report a comparison with alternative choices of similarity and cohesion function.

For comparison, we also consider a version of model (10) where the likelihood model is simply a Bernoulli distribution, that is, when (8) is replaced by

$$\text{logit}\left[\Pr\left(y_i = 1 \mid e_i = j, p^*_j\right)\right] = p^*_j, \tag{11}$$

together with a conjugate prior, $p^\star_j \sim \text{Be}(a, b)$. We refer to (11) as cluster-specific Bernoulli. Covariate dependence is included only in the prior distribution through the similarity functions $\tilde{g}_\ell\left(x^\star_{j\ell}\right)$. Using a desktop computer with a Core 2 Duo central processing unit (Dell in Austin, Texas, US) 2.53 GHz × 2 and 2.0 GiB of random access memory, a typical single-chain MCMC run to fit model (10) takes about 0.86 seconds per scan. In contrast, using model (11), it takes about 0.08 seconds per MCMC scan.

Figure 1 shows the posterior distribution for the number of clusters under model (8). The number of clusters is a bit elevated from the prior mean $k_n \approx M \log(n) = 7$ that would be expected under a PPM model with the same cohesion function but without the covariate dependence. The prior preference for homogeneous clusters, homogeneous in some of the 17 covariates, increases the *a priori* expected number of clusters.

Figure 2 shows posterior estimated mortality for several combinations of the covariates, under model (8) and the simplified model (11). The estimation of mortality probabilities is the main inference target in this application. The plots show posterior estimated mortality as a function of the continuous covariates. The binary and categorical covariates are held fixed throughout. The specific combination of binary and categorical covariates is the one that is most common in the sample with 330 cases: patients without a diagnosis of AIDS at admission, no metastatic cancer, no haematologic malignancy, with low ventilation risk and with scheduled surgery. The plots show estimated mortality as a function of heart rate, Glasgow score,

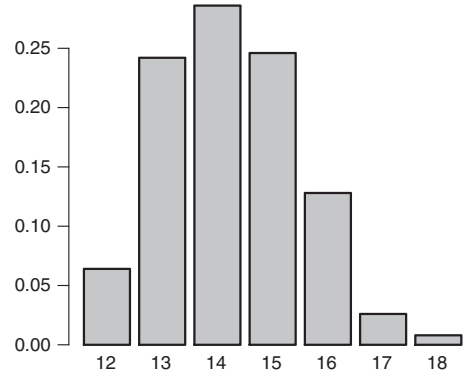

Fig. 1. $p(k \mid data)$ under model (8).

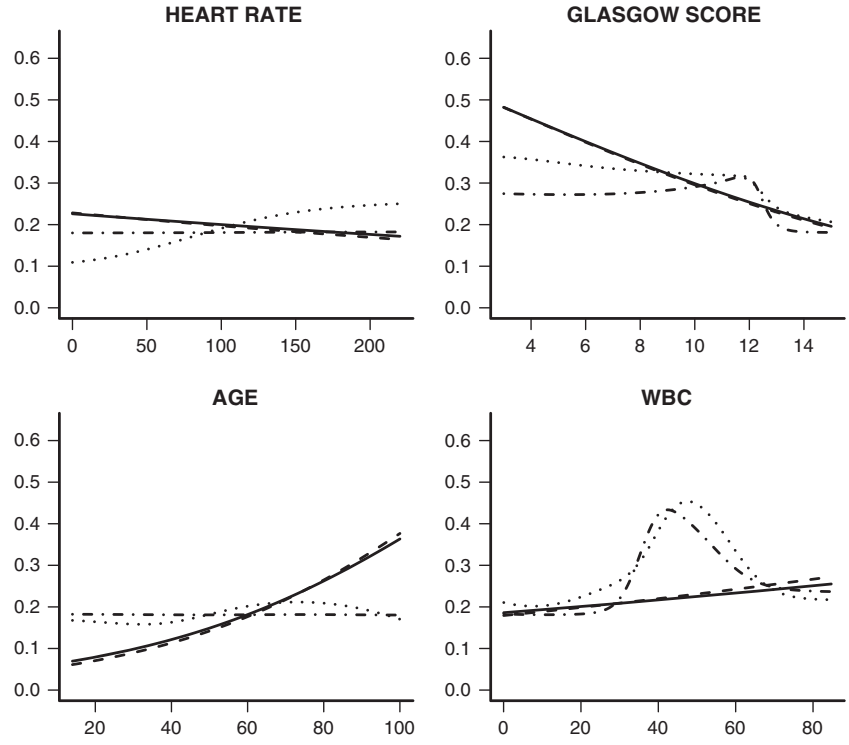

Fig. 2. Posterior estimated mortality for various covariate combinations. The predictions are for patients with scheduled surgery who had no acquired immunodeficiency syndrome, no metastatic cancer, no haematologic malignancy and with low or no risk for ventilation. The plots show estimated mortality as a function of heart rate, Glasgow score, age and white blood cell (WBC) count. Each plot shows predictions for model (8) with (solid line '—-') and without (dashed line '– – –') variable selection and for the simplified model (11) with (dotted line '·····') and without (semi-dashed '— · — · —') variable selection.

age and white blood cell count, keeping all remaining continuous covariates at their empirical median values. Predictions were computed for model (8) using variable selection (solid line) and without variable selection (dashed line). Similarly, we considered prediction for the simplified model (11) using variable selection (dotted line) and without it (semi-dashed line). Predictions

differ considerably under the two models. The lack of monotonicity of some of the estimated regressions under (11) is counter intuitive and points to over-fitting.

Leaving-one-out cross validation confirms this and clearly selects (8) over (11). We set up a model comparison by considering leave-one-out cross validation, predicting each time the binary response y_i that was left out by means of $\widehat{p}_i \equiv p\left(y_i \mid x^n, y^n_{-i}\right)$. Here, $y^n_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. Thresholding $\widehat{p}_i$, we obtain a classification into predicted positives (when $\widehat{p}_i > c$) and negatives ($\widehat{p}_i < c$). Varying the threshold c produces a receiver operating characteristic (ROC) curve. We use the area under the ROC curve (AUC) as a model comparison criterion. This is a relevant one in the ICU data, where the goal is to flag high-risk patients. In what follows, we compute AUCs and 95% confidence intervals (CI) using the `pROC` package (Robin *et al.*, 2011) in `R`. Under the proposed model, including the similarity function (3), the cluster-specific logistic regression (10) and variable selection in the similarity function, we find an AUC of 0.8396, with CI (0.8141 and 0.8649). The AUC drops slightly to 0.8376 with CI (0.8122 and 0.8630) for the same model *without* variable selection. In other words, the parsimony under variable selection comes at practically no price in predictive accuracy. Under the simplified model (11), the AUC drops considerably to 0.8074 with CI (0.7799 and 0.8348).

The sampling model with the cluster-specific logistic regression leads to much smoother curves for the estimated mortality as a function of the four shown covariates. The decreasing and increasing patterns for Glasgow score and white blood cell counts were *a priori* expected, although not formally enforced in the prior model. Finally, we note that under (8), the estimated mortality curves remain almost invariant when adding or not the variable selection.

An important feature of the model is that it can meaningfully predict mortality for any covariate combination, including extrapolation. Inference remains meaningful because it is based on matching a hypothetical new patient with clusters of observed patients. This mitigates the problem of extrapolation under a regression model, when prediction for covariate values beyond the range of the data can lead to meaningless inference. As an illustration, we repeat the predictions in Fig. 3 but assuming now a patient with AIDS at admission time, no metastatic cancer, no haematologic malignancy, with high ventilation risk and with medical type of admission. This particular combination of covariates was not observed in the sample. The results are shown in Fig. 3. The change in baseline covariates causes an overall increase in estimated mortality.

Turning our attention now to the variable selection, the model allows for cluster-specific variable selection. Figure 4 shows the selection probabilities $\Pr\left(\gamma^*_{j\ell} = 1 \mid \text{data}\right)$ for the top eight clusters and all covariates $\ell = 1, \ldots, 17$, numbered according to the codes given in Table 1. For

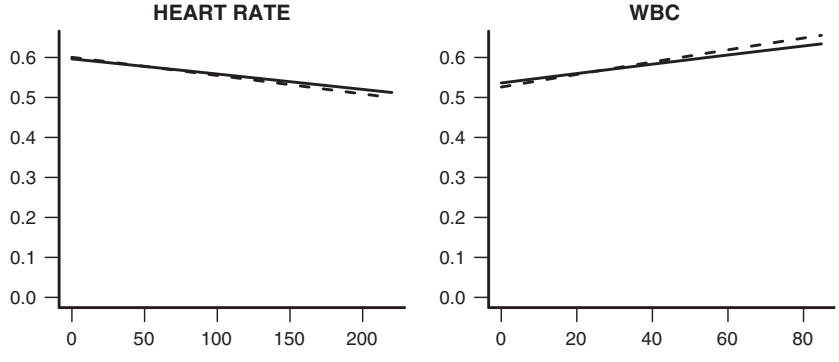

Fig. 3. Estimated mortality rates. Same as Fig. 2 but for patients with medical admission who had acquired immunodeficiency syndrome, no metastatic cancer, no haematologic malignancy and with high ventilation risk. WBC, white blood cell.

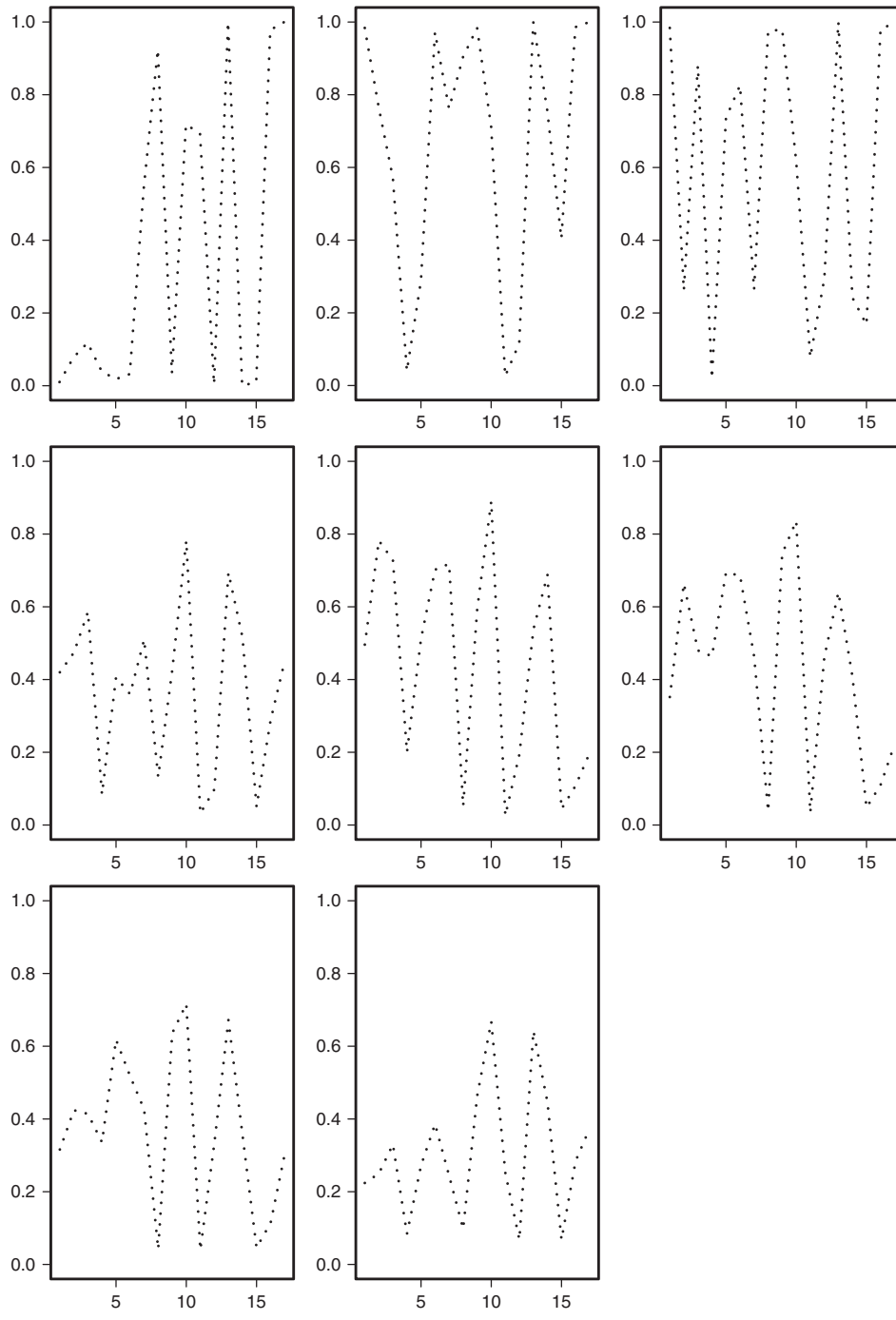

Fig. 4. Variable inclusion probabilities for the top eight clusters. The horizontal axis represents covariate number, and the vertical axis shows the inclusion probabilities. Continuous covariates are numbered 1–12, categorical covariates are 13–14, followed by the binary covariates 15–17.

easier comparison, in the following figures, all selection probabilities under the same setup are connected in a line across variables. For Fig. 4, we had to address the label switching problem. Cluster indices are arbitrary, and thus, posterior inference is symmetric under any permutation of the cluster labels. To facilitate meaningful reports of cluster-specific summaries, we have to add some identification of clusters. We first found a set of eight observations ('anchors') $(i_1, \ldots, i_8)$ with high probability $\Pr(e_{i_r} \neq e_{i_s} \mid y^n)$ for any $r \neq s$ (using high posterior probability instead of maximum posterior probability to avoid a difficult optimization problem). Cluster indices 1 through 8 were then defined as the clusters containing anchors i_1 through i_8. This enables us to report cluster-specific summaries.

The top cluster is characterized by covariates 8, 13, 16 and 17, each with selection probability of at least 80%. For an overall summary of covariate inclusion probabilities, we computed the proportion of clusters, among the top eight clusters, that select each covariate with probability of at least 80%. The covariates with highest average selection probabilities were 8, 13, 16 and 17, with high posterior inclusion probabilities in three out of eight clusters for all four covariates.

As another summary of covariate inclusion probabilities, we computed the weighted average (across clusters) of posterior means of $\gamma^*_{j\ell}$. The average is weighted by the corresponding sizes of the top 10 clusters. The results are displayed in Fig. 5. Variables 8, 10, 13, 16 and 17 have at least 70% of weighted average.

Clusters that are very homogeneous in one or a combination of covariates allow conclusions about important covariates and covariate combinations. This can be carried out as postprocessing of posterior inference. Figure 6 shows an example. In the left panel, the large circle in the left lower corner corresponds to a cluster of patients with no AIDS and no metastatic cancer at admission. The creation of this large cluster indicates an interaction of no AIDS and no metastatic cancer. Similarly, there is an interaction shown in the right panel for clusters with low serum urea level and high Glasgow score.

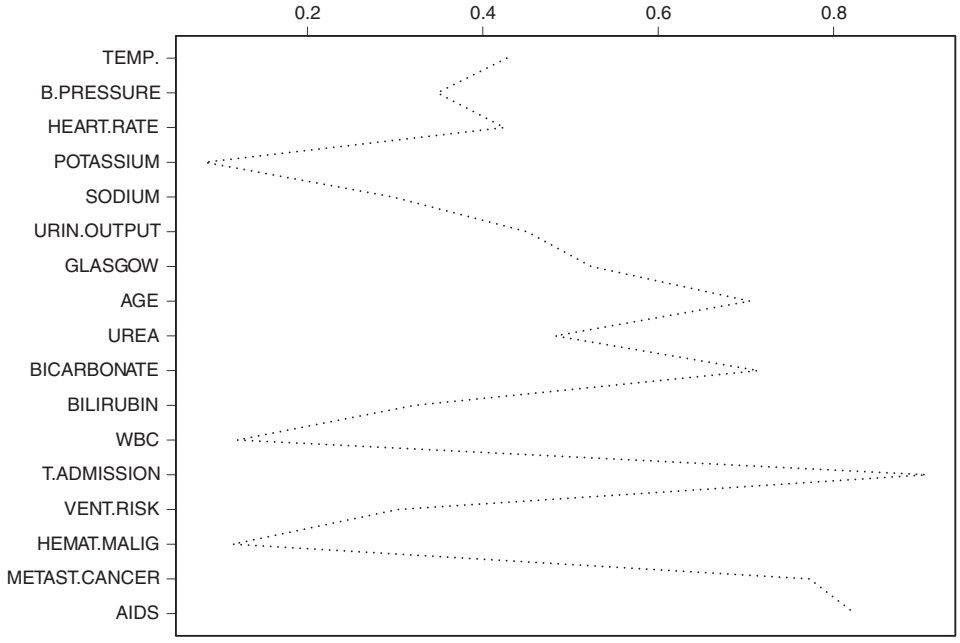

Fig. 5. Weighted average of variable selection probabilities for the top 10 clusters. The averages are weighted with respect to the cluster sizes. Continuous covariates are numbered 1–12, categorical covariates are 13–14, followed by the binary covariates 15–17. WBC, white blood cell.

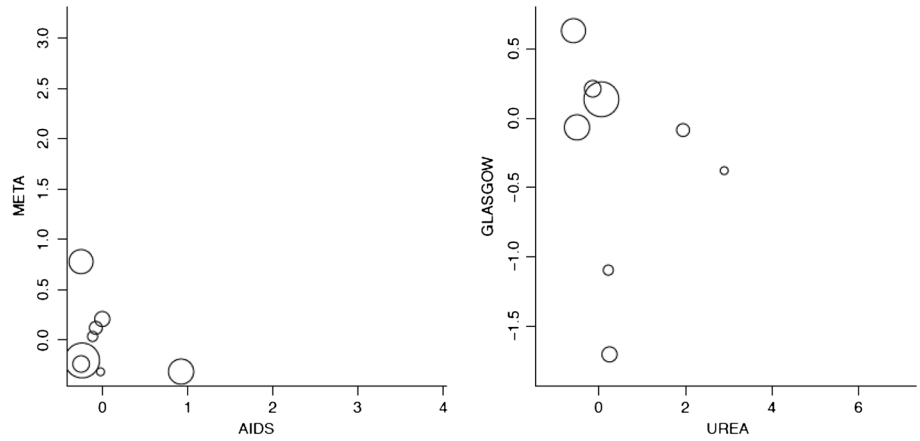

Fig. 6. Eight largest clusters. The circles plot the averages of the indicated variables over each of the clusters. The left panel shows metastatic cancer versus acquired immunodeficiency syndrome (AIDS), and the right panel represents Glasgow score versus serum urea. All variables (including the binary variables) are standardized to mean 0 and variance 1. The size of the circle is proportional to the cluster sizes.

4. Alternative approaches

4.1. Bayesian additive regression trees

For a comparison, we considered the sum of trees implemented in the Bayesian additive regression trees (BART) approach by Chipman *et al.* (2010). This class of models can be described as

$$y(x) = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2),$$

where T_j is a binary tree consisting of a set of interior node decision rules and a set of terminal nodes, M_j denotes a set of parameter values associated with each of the terminal nodes of T_j and g is the function that assigns parameters in M_j to x. BART is a very flexible non-parametric regression model, especially for large values of m. We implemented the leave-one-out cross validation using the `BayesTree` package for `R`, freely available from `CRAN`. We obtained AUC = 0.8319, with CI (0.8062 and 0.8577), slightly less than with the proposed models. While the difference in AUC is practically negligible, we argue for the PPMx model because of the easier interpretability of posterior inference.

4.2. Generalized additive models

As a further comparison, we also considered the generalized additive model (GAM) approach (Hastie & Tibshirani, 1990) as implemented in the `gam` package for `R`. See also Hastie *et al.* (2001). This class of models assumes the mean value of a response variable y, with a distribution in the exponential family, to be expressed in terms of p predictor variables as

$$g(E(y \mid x_1, \ldots, x_p)) = \alpha + \sum_{j=1}^{p} f_j(x_j),$$

where $f_1, \ldots, f_p$ are 'smooth' functions of the corresponding predictors and $g(\cdot)$ is the link function. In our ICU application, we use $g(\mu) = \text{logit}(\mu)$ and $\mu = \mu(x) = \Pr(y = 1 \mid x)$. The smooth functions f_j can be specified in various ways. The `gam` package provides local

regression and splines. We thus repeated the leave-one-out cross validation procedure, obtaining AUC = 0.8215 with CI (0.7948 and 0.8481) and AUC = 0.8192 (using 6 degrees of freedom), with CI (0.7922 and 0.8461), respectively. The model has a complexity that is comparable with the simple linear specification (8) we had previously chosen. As a final comparison, we implemented the same calculation using the `gam` function from the `R` package `mgcv`, which features an estimation of the degrees of freedom for each smooth term. We obtained AUC = 0.8267, with CI (0.8005 and 0.8529).

5. Summary

We have proposed a method for carrying out cluster-specific variable selection in the context of the PPMx model. The model includes covariate-specific similarity functions in the prior for the random partition with a random selection indicator. Depending on these selection indicators, the resulting prior probability model for the random partitions includes as special cases PPMx-style dependence on the covariate and no dependence at all. We considered an application to ICU mortality data. The application involved a logistic cluster-specific sampling model with cluster-specific parameters for the response in terms of covariates and also a simplified version, with only subject-specific death probabilities. The first model performed better in terms of AUC when carrying out a leave-one-out cross validation. For several variations of similarity and cohesion functions, we found slightly better AUC than for alternative models based on BART or GAM models, either fixing or not the degrees of freedom of smooth terms.

We carried out additional comparisons with BART and GAM models as part of a simulation study included in the online supporting information. Because our primary goal is prediction, we focused our efforts on comparing prediction against the simulation truth and reported results in terms of mean squared errors for in-of-sample and out-of-sample predictions. In summary, the proposed method compares favourably in all scenarios, because of the ability of supporting interactions through the prior on partitions, even if they are not explicitly included in the sampling model.

Finally, there are several limitations and opportunities for further research. Perhaps, the most important limitations are the reliance on MCMC posterior simulation and the need for hyperparameter choices in the similarity function. In the examples for this paper, we used data sets with around $n = 1000$ observations and up to 20 covariates. Inference can be implemented for moderate size data sets without a problem. However, MCMC would not be feasible for truly big data sets. The other big limitation is the choice of the similarity functions. Ideally, the model should be able to learn about hyperparameters in the definition of the similarity functions. For example, if a common cluster-specific sampling model with parameters $\beta_j^\star$ is only sustainable for covariate values in a certain range, then posterior inference should be allowed to adjust hyperparameters accordingly.

Acknowledgements

The authors would like to thank the medical team of UCIP from the Centro Hospitalar de Lisboa Central (Lisboa), in particular to Dr Eduardo Silva and Dr Miguel Robalo, for their collaboration in data collection. Ana Luisa Papoila's work was partially supported by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal—FCT under the project PEst-OE/MAT/UI0006/2014. Peter Müller's work was partially supported by grants NIH R01CA157458 and R01CA075981. Fernando A. Quintana's work was partially supported by grant FONDECYT 1100010.

References

Chipman, H. A., George, E. I. & McCulloch, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93**, (443), 935–948.

Chipman, H. A., George, E. I. & McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. In *Model selection*, vol. 38, IMS Lecture Notes Monogr. Ser. Inst. Math. Statist., Beachwood, OH; 65–134. With discussion by M. Clyde, Dean P. Foster, and Robert A. Stine, and a rejoinder by the authors.

Chipman, H. A., George, E. I. & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, (1), 266–298.

Chung, Y. & Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc.* **104**, (488), 1646–1660.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85**, (2), 363–377.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

Hartigan, J. A. (1990). Partition models. *Comm. Statist. Theory Methods* **19**, 2745–2756.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*, Monographs on Statistics and Applied Probability, vol. 43, Chapman and Hall Ltd., London.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning*, Springer Series in Statistics, Springer-Verlag, New York. Data mining, inference, and prediction.

Hoff, P. D. (2005). Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics* **61**, (4), 1027–36.

Lemeshow, S., Teres, D., Klar, J., Avrunin, J.-S., Gehlbach, S.-H. & Rapoport, J. (1993). Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *J. Am. Med. Assoc.* **270**, (20), 2478–2486.

Metnitz, P. G. H., Moreno, R. P., Almeida, E., Jordan, B., Bauer, P., Abizanda-Campos, R., Iapichino, G., Edbrooke, D., Capuzzo, M. & Le Gall, J.-R. (2005). SAPS 3. From evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. *Intens. Care Med.* **31**, 1336–1344.

Moreno, R. P., Metnitz, P. G. H., Almeida, E., Jordan, B., Bauer, P., Abizanda-Campos, R., Iapichino, G., Edbrooke, D., Capuzzo, M. & Le Gall, J.-R. (2005). SAPS 3. From evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. *Intens. Care Med.* **31**, 1345–1355.

Müller, P. & Quintana, F. A. (2010). Random partition models with regression on covariates. *J. Statist. Plann. Inference* **140**, (10), 2801–2808.

Müller, P., Quintana, F. A. & Rosner, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20**, (1), 260–278.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, (2), 249–265.

Papoila, A., Rocha, C., Geraldes, C. & Xufre, P. (2013). Generalized linear models, generalized additive models and neural networks: comparative study in medical applications. In *Advances in regression, survival analysis, extreme values, Markov processes and other statistical applications* (eds J. Lita da Silva, F. Caeiro, I. Natário & C. A. Braumann), 317–324. Studies in Theoretical and Applied Statistics, Springer, Berlin, Heidelberg.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77.

Zimmerman, J. E. & Kramer, A. A. (2006). Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Critical Care Medicine* **34**, 1297–1310.

Received January 2013, in final form February 2015

Fernando A. Quintana, Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile.
E-mail: quintana@mat.puc.cl

Supporting information

Additional supporting information may be found in the online version of this article at the publishers web site.