Prediction in the Presence of Missing Covariates

Garritt L. Page
Department of Statistics
Brigham Young University, Provo, Utah
Basque Center of Applied Mathematics
page@stat.byu.edu

Fernando A. Quintana
Departamento de Estadística
Pontificia Universidad Católica de Chile, Santiago
and Millennium Nucleus Center for the Discovery of Structures in Complex Data
quintana@mat.puc.cl

Peter Müller
Department of Mathematics
The University of Texas at Austin
pmueller@math.utexas.edu

January 1, 2020

Abstract
In many applied fields incomplete covariate vectors are commonly encountered. It is well known that this can be problematic when making inference on model parameters, but its impact on prediction performance is less understood. We develop a method based on covariate dependent partition models that seamlessly handles missing covariates while avoiding completely any type of imputation. The method we develop is not only able to make in-sample predictions, but out-of-sample as well even if the missing pattern in the new subjects’ incomplete covariate vector was not seen in the training data. Further, both categorical and continuous covariates are permitted. Our proposal fares well when compared to other al-
ternatives based on imputations. We illustrate the method using simulation studies and an ozone dataset.

**Key Words:** Dependent random partition models, indicator-missing, pattern missing, Bayesian nonparametrics

### 1 Introduction

We introduce an approach for prediction with missing covariates, that is, regression with a variable-dimension covariate vector. The proposed model does not require any notion of imputing or substituting or other assumptions about the missing covariates. Instead we start with a random partition based on available covariates, and then add a cluster-specific sampling model for the response. The result is an elegant and uncomplicated variable-dimension regression approach.

Missing observations are regularly encountered in data-driven research (Daniels and Hogan 2008, Molenberghs et al. 2014). Because of this, there is a rich literature dedicated to methods that have been developed to accommodate them. These methods range from being ad-hoc like the complete-case approach which simply deletes subjects/units with a missing observation, to more statistically sound procedures like (multiple) imputation which probabilistically “fills” in the missing values (see Rubin 1987, Little and Rubin 2002, van Buuren 2012 or Molenberghs et al. 2014). Most of the statistical literature dedicated to missing observations is focused on missing response values and their effect on inferences associated with model parameters. The focus of this work is on incomplete covariate vectors and their affect on prediction accuracy. Even though incomplete predictor vectors are ubiquitous and can have adverse effects on prediction accuracy (destructive if an influential predictor is missing), the missing observations literature is much more sparse for this case.

In the presence of missing covariates the complete-case approach is still an option, but often performs poorly when prediction is of interest (Mercaldo and Blume 2018). Some multiple imputation methods that were developed for missing response values can also be employed for missing covariates. Focusing on methods that allow mixed data types, multiple imputation by chained equations
(MICE), which employs conditionally specified models, can be used to impute missing covariates one-at-a-time (van Buuren 2012). This approach is somewhat ad-hoc as there is no guarantee that the conditionally specified models produce a valid joint model for the covariates. To avoid this, Xu et al. (2016) employ Bayesian additive regression trees (BART) to impute missing covariates based on the MICE framework. Although their approach produces a valid joint distribution, the order of the conditional models impacts the imputations. Similarly, Burgette and Reiter (2010) employ classification and regression trees (CART) to impute within a MICE type algorithm which permits more flexibility in the conditional distributions, while Stekhoven and Bühlmann (2012) use random forests to carry out imputation. Recently, Storlie et al. (2019) build a flexible yet complex Bayesian nonparametric model to carry out imputation. Their approach jointly models mixed-type covariates and includes a variable selection component making the procedure more efficient. All these multiple imputation type approaches (and many others not listed) focus on model parameter inference. If prediction is tangentially considered, the complications that arise when predicting based on multiple imputation are not considered. For example, procedures based on multiple imputation are problematic when out-of-sample prediction is desired as it is not possible to connect a response to the vector of covariates when carrying out imputation (a response does not exist). This has been shown to negatively impact predictive performance (Moons et al. 2006). Taking into account all these limitations, our interest lies in developing a procedure that requires absolutely no type of imputation while simultaneously providing a good and flexible model for the available data.

The so called missing indicator approach (Little 1992, Jones 1996, van der Heijden et al. 2006) has been developed to avoid taking on the sometimes unverifiable assumptions of multiple imputation. But these methods must be used with care in practice as they are prone to producing biased estimates, and as a result, poor predictions (see van der Heijden et al. 2006 and Groenwold et al. 2012). Also, under this approach, there is no clear way to handle the case of a new subject in out-of-sample prediction exhibiting a different missing pattern than those found in the training data.
Our approach to incorporating missing covariates in a prediction model stems from a completely different perspective. We focus on extending a covariate dependent random partition model so that missing values are permitted. Covariate dependent random partitions are particularly well suited for the task of prediction, as they permit complex interactions and nonlinear associations between covariates and responses. They are also able to seamlessly handle mixed-type covariates and the availability of a predictive distribution makes out-of-sample prediction straightforward regardless of the missing patterns. Perhaps the missing procedure whose focus is most similar to what we develop is found in Kapelner and Bleich (2015). They use BART to make predictions and employ a missing indicator when constructing trees (i.e., the trees are not used as a tool to impute). Although their motivation is similar to ours, our approach is based on partitions which permits more flexibility in how covariates interact and their associations.

The remainder of the article is organized as follows. In Section 2 we provide background associated with covariate dependent product partition models and detail our extension that permits incomplete observed covariates vectors of varying dimensions. In this section we also detail briefly a method that combines our covariate dependent product partition model with indicator missing approaches. Section 3 contains a simulation study and a data application. Some concluding remarks are provided in Section 4.

2 Missing Values and Covariate Dependent Product Partition Models

The proposed procedure builds on the covariate dependent partition model proposed by Müller et al. (2011). We first provide background for this model and later describe in detail the proposed extensions that permit integrating missing covariates.
2.1 Preliminaries: Covariate Dependent Product Partition Model

We only provide background to covariate dependent product partition models to the extent needed to describe our approach to accommodate missing covariates. A more detailed description of these models can be found in Müller et al. (2011), Park and Dunson (2010), or Quintana et al. (2018).

Let $i = 1, \ldots, m$ index the $m$ experimental units in a designed experiment or $m$ subjects in an observational study. Further, let $\rho_m = \{S_1, \ldots, S_{k_m}\}$ denote a partitioning (or clustering) of the $m$ units into $k_m$ nonempty and exhaustive subsets such that $i \in S_j$ implies that unit $i$ belongs to cluster $j$. To simplify notation we omit the subscript $m$ for $\rho$ unless explicitly needed. A common alternative notation that specifies a partitioning of the $m$ units into $k$ clusters is to introduce $m$ cluster labels $s_1, \ldots, s_m$ such that $s_i = j$ implies $i \in S_j$. Let $x_i = (x_{i1}, \ldots, x_{ip})$ denote a $1 \times p$ covariate vector measured on unit $i$ and $x = (x_1, \ldots, x_m)$ will denote a matrix of “stacked” covariate vectors. Further, let $x_j^* = \{x_i : i \in S_j\}$ denote a “stacked” collection of covariate vectors from all units that belong to cluster $j$. Note that in what follows we will use superscript “$\star$” to denote cluster-specific entities. The covariate dependent product partition model (PPMx) prior on $\rho$ formalizes the idea that units with similar covariate values are more likely a priori to belong to the same cluster relative to units with dissimilar covariate values. The prior consists of two set functions. The first, called a cohesion function and denoted by $c(S_j \mid M) \geq 0$ for $S_j \subset \{1, \ldots, m\}$ and parameter $M$, measures prior belief associated with the co-clustering of the elements of $S_j$. The second, called a similarity function and denoted by $g(x_j^* \mid z)$ and parametrized by $z$, formalizes the “closeness” of the $x_i$’s in a cluster by producing larger values of $g(x_j^* \mid z)$ for $x_i$’s that are more similar. The similarity function in the PPMx plays a similar role to that of the impurity function when building trees using CART (Classification And Regression Trees). See, for example, Sutton (2005, Section 2.4). With the similarity and cohesion functions, the form of the PPMx prior is the following product

$$Pr(\rho \mid x, M, z) \propto \prod_{j=1}^{k_m} c(S_j \mid M)g(x_j^* \mid z).$$  (1)
The cohesion function we employ in what follows is \( c(S_j) = M \times (|S_j| - 1)! \) for some positive \( M \) and \(| \cdot |\) denoting cardinality. This cohesion is commonly employed and its popularity stems from connections the resulting random partition model has to that which is based on a Dirichlet process model (see, e.g., Quintana 2006). Regarding possible similarity functions, Müller et al. (2011) discuss choices or different types of covariates (e.g., continuous, ordinal, or categorical), and suggest using

\[
g(x_j^* \mid z) = \int \prod_{i \in S_j} q(x_i \mid \zeta_j) q(\zeta_j \mid z) d\zeta_j. \tag{2}
\]

The integral can be evaluated in closed form if \( q(x_i \mid \zeta_j) \) and \( q(\zeta_j \mid z) \) are chosen as a conjugate sampling model and prior pair, using the distributions only for easy calculus, rather than any notion of statistical modeling (indeed \( x \) is not even random). They are only used as a means to measure the “agreement” of the covariates from units assigned to cluster \( S_j \). In fact, any function that produces larger values as the entries of \( x_j^* \) become more similar can be considered as a similarity function. For simplicity it is common to define a similarity function \( g_\ell \) for each covariate and set \( g(x_j^* \mid z) = \prod_{\ell=1}^{p} g_\ell(x_{j\ell}^* \mid z_\ell) \) where \( x_{j\ell}^* = \{x_{i\ell} : i \in S_j\} \). See Page and Quintana (2018) for more discussion on other possible specifications for the similarity function.

For a given cluster arrangement \( \rho \), we complete the model construction with a sampling model for the response \( y_i \) by introducing cluster-specific parameters \( \theta^* = (\theta_1^*, \ldots, \theta_{km}^*) \) and assuming conditional independence at the observation level. Letting \( y_i \) denote the \( i \)th response and \( y = (y_1, \ldots, y_m) \) this leads to the following model

\[
p(y, \rho, \theta^*, M, z \mid x) = p(y \mid \rho, \theta^*, x) p(\rho \mid x, M, z) p(\theta^*) \propto \prod_{j=1}^{km} \left\{ \left( \prod_{i \in S_j} p(y_i \mid \theta_j^*) \right) p(\theta_j^*) \right\} Pr(\rho \mid x, M, z), \tag{3}
\]

where \( p(\theta^*) \) is a prior distribution for \( \theta^* \) which are assumed to be independent and identically distributed. The model can be written as a hierarchical model with latent cluster membership
indicators,

\[
y_i \mid \theta^*, s_i, x^\text{ind} \sim p(y_i \mid \theta^*_s, x_i)
\]

\[
\theta^*_j \mid \rho \overset{iid}{\sim} p(\theta^*_j \mid \rho)
\]

\[
Pr(\rho = \{S_1, \ldots, S_{km}\} \mid x, M, z) \propto \prod_{j=1}^{km} c(S_j \mid M) g(x^*_j \mid z).
\]

Notice that \(x\) could potentially appear in two levels of the hierarchical model described in equation (4) (in the top level sampling model or the prior on partitions). The main purpose of this paper is to extend the PPMx in such a way that missingness in \(x\) is accommodated without the need to carry out any type of imputation. Thus, we focus mainly on the case when \(x\) only appears in the prior distribution on partitions. However, for the sake of completeness, we do briefly detail a possible approach that accommodates missingness in \(x\) when this is included in both, top level sampling model and prior on partitions, while still avoiding any type of imputation. The approach we take is similar to the pattern submodel procedure found in Mercaldo and Blume (2018) and the missing indicator approach. Specific details are provided in Section 2.3.2.

2.2 Accommodating Missing Covariates with the PPMx

In a prediction setting, missing values may exist not only among subjects whose observations are employed to fit a particular statistical model, but also when predicting for a new (out-of-sample) subject. An additional challenge arises when the missingness pattern in the predictor profile of the new subject is new, i.e., has not been seen in any of the subjects used to train the model. The methodology we propose naturally accommodates this type of situation.

2.2.1 Missingness When Training Model

To illustrate how the PPMx can be extended to accommodate missing covariates, denote by \(\mathcal{O}_i\) the collection of covariate indices that are observed for subject \(i\). The \(i\)th subject’s observed covariate vector can be now denoted as \(x^o_i = \{x_{i\ell} : \ell \in \mathcal{O}_i\}\) and the collection of observed covariate vectors

\[
\begin{align*}
\end{align*}
\]
that belong to the $j$th cluster is $x_{j^o} = \{x_i^o : i \in S_j\} = \{x_{i\ell} : \ell \in O_i, i \in S_j\}$. Then, missing covariates can be accommodated in the PPMx by evaluating the similarity function based only on those subjects in the $j$th cluster with the $p$th covariate observed. Missing values are accommodated by defining

$$g(x_j^* \mid z) = \tilde{g}(x_{j^o}^* \mid z) = \prod_{\ell=1}^p \tilde{g}(\{x_{i\ell} : \ell \in O_i, i \in S_j\} \mid z_{i\ell}) = \prod_{\ell=1}^p \int \prod_{i \in C_{j\ell}} q(x_{i\ell} \mid \zeta_{j\ell})q(\zeta_{j\ell} \mid z_{i\ell})d\zeta_{j\ell}, \quad (5)$$

where $C_{j\ell} = \{i : i \in S_j, \ell \in O_i\}$. Importantly, in the presence of missing covariates, the similarity function for the $\ell$th covariate is evaluated based only on subjects for which the covariate is measured. In other words, missing values are simply skipped over when evaluating the similarity function. As a result, no imputation (implicit or not) is being employed. [Xu et al. (2019)] used a similar strategy when using the PPMx in a basket trial design, but without any notion of prediction. From a computational viewpoint, the methods needed to fit this model are unchanged with respect to the case with no missing observations save for a matrix of indicators must be carried along. General computational details can be found in [Page and Quintana (2015)] and we provide more specifics in Section ?? of the online supplementary material.

We note briefly that in the context of variable selection [Quintana et al. (2015)] consider similarity functions that are similar in form to (5), but with each cluster selecting a cluster-specific subset of covariates. Importantly, in that application $C_{j\ell}$ is a random cluster-specific parameter that includes the subset of covariates that were selected for the $j$th cluster. In that case it is important that $g(x_{j^o})$ be scaled such that $g(x_{j^o} > 1)$ for $x_{j^o}$ that are judged to be very similar and $g(x_{j^o}) < 1$ for very diverse $x_{j^o}$. A similarity function with $g(x_{j^o} < 1)$ for all $x_{j^o}$ would exclude covariate $\ell$ from ever being selected. [Quintana et al. (2015)] introduce an additional factor to ensure such scaling. However, this is not required here since $C_{j\ell}$ is fixed, i.e., inference is conditioned on the observed
2.2.2 Predicting Future Observations with Missing Entries

An outstanding feature that motivates consideration of the PPMx prior on partitions, is the flexibility in which they capture ways that covariates influence the predictive distribution, making out-of-sample prediction more accurate relative to other models, for example, the PPM. The new similarity function in (5) easily accommodates incomplete covariate vectors when making predictions for “new” individuals even if the pattern of missingness has not been observed among individuals included in the training data set. To see this, consider the following predictive multinomial probabilities that the \((m + 1)\)st subject belongs to one of the groups \(h = 1, \ldots, k_m\)

\[
Pr(s_{m+1} = h \mid \rho_m, \mathbf{x}^o, \mathbf{x}^o_{m+1}) \propto \begin{cases} 
\frac{c(S_h \cup \{m + 1\})\tilde{g}(\mathbf{x}^o_h \cup \{x^o_{m+1}\})}{c(S_h)\tilde{g}(\mathbf{x}^o_h)} & \text{for } h = 1, \ldots, k_m \\
\frac{c(\{m + 1\})\tilde{g}(\{x^o_{m+1}\})}{c(S_h \cup \{m + 1\})\tilde{g}(\mathbf{x}^o_h \cup \{x^o_{m+1}\})} & \text{for } h = k_m + 1.
\end{cases}
\]  

(6)

Here \(\tilde{g}(\mathbf{x}^o_h \cup \{x^o_{m+1}\})\) is computed as

\[
\prod_{\ell = 1}^p \int \prod_{i \in C_h \ell} q(x_{i\ell} \mid \zeta_{h\ell}) \prod_{\ell \in O_{m+1}} q(x_{m+1,\ell} \mid \zeta_{h\ell})q(\zeta_{h\ell} \mid z_\ell)d\zeta_{h\ell}.
\]  

(7)

Thus, any missing covariate for the \((m + 1)\)st subject is handled in (7) by simply skipping over those missing values, and therefore, the similarity can be always evaluated. In the extreme case of a “new” subject with an entirely missing covariate vector, the similarity function for the \(\ell\)th covariate is simply computed using \(\tilde{g}(\mathbf{x}^o_h) = \int \prod_{i \in C_h \ell} q(x_{i\ell} \mid \zeta_{h\ell})q(\zeta_{h\ell} \mid z_\ell)d\zeta_{h\ell}

and the probabilities in (6) resort to those when making predictions using the PPM.

2.3 Data Model Specifications

The choice of the sampling model in (4) depends on the variable type of the response and whether it includes a regression of \(y\) on \(x\), in addition to the regression on \(x\) that is already implied in the
PPMx prior. Since the application in Section 3 is based on ozone measurements, here we provide details for on a continuous $y$ response. We first provide specific details for model (4) when a model for the relationship between $x$ and $y$ is not explicitly included in the top level sampling model, and discuss the case with regression in the following subsection. Recall that our main motivation here is to completely avoid imputation.

2.3.1 Hierarchical Model with Covariates in the PPMx Only

When covariates are not included in the top level sampling model and therefore only appear in the PPMx, we employ the following hierarchical model

$$ y_i \mid \mu^*, \sigma^{2*}, c_i \sim N(\mu^*_c, \sigma^{2*}_c) \text{ for } i = 1, \ldots, m $$

$$ (\mu^*_j, \sigma^*_j) \sim N(\mu_0, \sigma^2_0) \times UN(0, a_\sigma) \text{ for } j = 1, \ldots, k_m $$

$$ (\mu_0, \sigma_0) \sim N(m_0, v^2) \times UN(0, a_{\sigma_0}) $$

$$ Pr(\rho = \{S_1, \ldots, S_{km}\} \mid x^o, M, z) \propto \prod_{j=1}^{km} c(S_j \mid M) \tilde{g}(x^o_j \mid z). $$

The Gaussian prior for the cluster-specific means is commonly used and a uniform prior on cluster-specific standard deviations follows suggestions in Gelman (2006). Prior parameters, denoted here using Latin letters, are user supplied. In what follows we will refer to (8) as “PPMx”.

2.3.2 Covariates in Data Model and PPMx: Pattern Submodel with the PPMx

When covariates are not explicitly included in the top level sampling model of model (4), it is possible to glean information regarding the effect that the covariates have on responses by exploring the posterior distribution of $\rho$ in a post-hoc manner. When covariates are included in the top level sampling model (in addition to the prior on partitions) missingness can still be accommodated without imputation using a pattern submodel (PSM) type approach like that found in Mercaldo and Blume (2018). Combining a pattern submodel with our PPMx approach accommodates missing patterns in out-of-sample prediction for cases where covariate values do not
appear in the training data.

The pattern submodel approach specifies a unique set of regression coefficients for each missing pattern. To this end, let \( \mathbf{\beta} = (\beta_1, \ldots, \beta_G) \) denote the coefficient vectors that correspond to the \( G \) unique missing patterns in the training data. The dimension of \( \beta_g \) for \( g = 1, \ldots, G \) ranges from 0 to \( p \). Note that if an individual has no covariate information, then prediction is carried out using PPM. Now, let \( g_1, \ldots, g_m \) be \( m \) known indicator variables that connects the \( i \)th individual with its corresponding missing pattern such that \( g_i \in \{1, \ldots, G\} \). The following hierarchical model incorporates missing covariate values in the top level sampling model through a combination of the pattern submodel and a PPMx prior on partitions:

\[
y_i \mid \mu^*, \sigma^2, \beta, c_i, g_i, \mathbf{x}_i \sim N(\mu_{c_i}^* + \mathbf{x}_i^\prime \beta_{g_i}, \sigma_{c_i}^2) \quad \text{for} \quad i = 1, \ldots, m
\]

\[
(\mu_j^*, \sigma_j^2) \sim N(m, s^2) \times UN(0, a_\sigma) \quad \text{for} \quad j = 1, \ldots, k_m
\]

\[
\beta_g \sim N_{pg}(b_g, B_g) \quad \text{for} \quad g = 1, \ldots, G
\]

\[
(\mu_0, \sigma_0) \sim N(m_0, v^2) \times UN(0, a_\sigma_0)
\]

\[
Pr(\rho = \{S_1, \ldots, S_{km}\} \mid \mathbf{x}^0, M, \mathbf{z}) \propto \prod_{j=1}^{km} c(S_j \mid M) \tilde{g}(\mathbf{x}_{j}^* \mid \mathbf{z}).
\]

As with model (8) all Latin letters are user supplied. It should be evident that as \( p \) grows the number of missing patterns will increase. In extreme cases, it is possible that a particular \( \beta_g \) is estimated using a single observation. Mercaldo and Blume (2018) detail a few approaches that are able to handle such cases, but for the sake of simplicity we stick to the hierarchical model in (9). In the simulation study detailed in Section 3 we explore the predictive performance of this approach as a function of \( p \). Finally, interpreting the regression coefficients must be done with care as values of the coefficient for a specific covariate may be considerably different for different missingness patterns. In what follows we will refer to (9) as “PPMxPSM".
3 Simulation Study and Data Application

In this section we describe a simulation study designed to explore the performance of our proposed method with regards to model fit and out-of-sample prediction. Then we consider a real-world data set containing ozone measurements.

3.1 Simulation Study

We conduct a simulation study that illustrates the performance of our extension to the PPMx with and without the pattern submodel. The main objectives of the experiment are (1) determine how increasing the number of covariates together with increasing rates of missing affect prediction, (2) determine how different types of missing affect prediction, and (3) determine if partition structure contained in covariates affects prediction.

To carry out the simulation, synthetic data sets with 100 testing and 100 training observations are created. Data sets are generated under a varying number of covariates with \( p \in \{2, 4, 10\} \). Specific values for the covariates are generated using four Gaussian distributions thus creating four covariate dependent clusters. For example, when \( p = 2 \), four bivariate Gaussian distributions centered at \((1,1), (1,-1), (-1,1), \) and \((-1,-1)\) are used to generate 200 sets of covariate values (50 from each cluster). Similarly, when \( p = 4 \), four-dimensional Gaussians with mean vectors equal to \((1,1,1,1), (1,-1,1,-1), (-1,1,-1,1)\), and \((-1,-1,-1,-1)\) are used to create 4 clusters each with 50 observations. And lastly, when \( p = 10 \), four 10-dimensional Gaussians with means equal to \((1,1,1,1,1,1,1,1,1,1,1), (1,-1,1,-1,1,-1,1,-1,1,-1), (-1,1,-1,1,-1,1,-1,1,-1,1,-1), \) and \((-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1)\) are used to generate covariates coming from four clusters each with 50 observations. Notice that as \( p \) increases, the number of clusters remains the same (i.e., four), but they become much more sparse in the covariate space. To study how the cluster noise (i.e., overlap among clusters) affects prediction results, the covariance matrix used to generate covariate values for the scenarios just described is \( s^2 I_p \), where \( s^2 \in \{0.25^2, 0.5^2, 0.75^2\} \). When \( s^2 = 0.25^2 \) the clusters are well separated, while for \( s^2 = 0.5^2 \) the cluster borders meet and
for $s^2 = 0.75^2$ the clusters overlap. To illustrate, Figure ?? in the online supplementary material displays the cluster configuration for each of the $s^2$ when $p = 2$.

In all cases, the response is also generated using Gaussian distributions with means equal to $(-1, -0.5, 0, 0.5)$ (50 values are generated from each Gaussian). To study how constant variance in the response across clusters versus a cluster-specific variance in the response impacts prediction performance, we generate response values in two ways. The first sets the standard deviation of each Gaussian equal to 0.25 and the second uses $(0.1, 0.25, 0.5, 0.75)$ as standard deviations in the four Gaussian distributions.

Finally, missing values in the covariates are inserted as follows. For each covariate a specific fraction (approximately) of values are randomly selected to be classified as missing. We consider two types of missingness. The first is missing at random (MAR) and the second missing not at random (MNAR). Generating both types of missing is facilitated using the ampute function found in the mice R-package [van Buuren and Groothuis-Oudshoorn, 2011]. For MNAR, the ampute function is used for each covariate with the missing probabilities being a function of the covariate value (see Schouten et al. 2018 for specific details regarding the function used to produce probability of missing). The ampute function is also applied separately to each covariate for the MAR case where each covariate entry is equally likely to be classified as missing. In summary, the numerical experiment considers the following factors.

1. type of missing (MAR or MNAR)

2. Fraction of missing covariate values (0%, 10%, 25%, 50%)

3. Number of covariates (2, 4, 10)

4. Noise in partition ($s^2 \in \{0.25^2, 0.5^2, 0.75^2\}$)

5. Constant variance across clusters (Yes, No)

As mentioned, each generated data set consists of 100 training and 100 testing observations and to each the following procedures are fit.
1. PPMx: Model with covariates only in PPMx and detailed in (8)

2. PPMx_PSM: PPMx model coupled with a Pattern Submodel that is detailed in (9)

3. MI: Based on 10 imputed data sets via the complete function of the mi package (Gelman and Hill 2011) from the statistical software R (R Core Team 2018)

4. BART: The method detailed in Kapelner and Bleich (2015) and carried out using the bartMachine package (Kapelner and Bleich 2016) in R

5. PSM: Pattern submodel approach using method in Mercaldo and Blume (2018) and code available at https://github.com/sarahmercaldo/MissingDataAndPrediction

When fitting models (8) and (9), covariates are standardized so that each one has mean 0 and standard deviation 1. For similarity $g(x_j^* \mid z)$ we use equation (2) where for each covariate $q(\cdot \mid \zeta_j) = N(\cdot; \zeta_j, 0.5)$ and $q(\zeta_j \mid z) = N(\zeta_j; 0, 1)$. Finally, for prior values we set $m_0 = 0$, $v^2 = 100^2$, $a_\sigma = 10$, $a_{\sigma o} = 10$, $b_{\ell} = 0_{p \ell}$ and $B_{\ell} = 100I_{p \ell}$. With these prior specifications, we fit models (8) and (9) by collecting 1000 MCMC samples after discarding the first 25,000 as burn-in and thinning by 25 (i.e., 50,000 total MCMC draws are sampled). All computation for models (8) and (9) is carried using the ppmx.missing function that is part of the ppmSuite R-package found on the first author’s webpage.

In order to make out-of-sample predictions using MI, covariates in training and testing data were joined, and imputation was carried out based only on this joined matrix (i.e., the response associated with training data was not included in the imputation). Default parameter values for the BART and PSM procedures are used. To evaluate each method’s ability to make predictions when covariates are missing, we employ the following metrics.

- **MSE**: represents the mean squared error defined as $\frac{1}{100} \sum_{i=1}^{100} (Y_{o i} - \hat{Y}_{o i})^2$ where $i$ indexes the 100 training observations $(Y_{o i})$ and $\hat{Y}_{o i}$ is the fitted value for the $i$ observation. This quantity measures goodness-of-fit.
• MSPE: represents the mean squared prediction error defined as \( \frac{1}{100} \sum_{i=1}^{100} (Y_{pi} - \hat{Y}_{pi})^2 \) where \( i \) indexes the 100 testing observations \( (Y_p) \) and \( \hat{Y}_{pi} = E(Y_{pi} \mid Y_o) \). This quantity measures the predictive performance of the models.

We found that the simulation results are similar under the various combinations of data being MNAR or MAR, and variance associated with response being constant or cluster-specific. Therefore, we present here results for data that are MNAR and with cluster-specific response variance, and present other results in Section ?? of the online supplementary material. Figures 1 and 2 display the MSE and MSPE results as a function of the number of covariates, missing fraction and cluster noise. Focusing on the MSE values first, notice if there is no missing values that BART fits the data best and the other procedures are similar with cluster noise impacting MI and PSM the most (which is to be expected). However, as the missing rate increases, then either the PPMx or PPMx,PSM performs best in model fit and the differences between procedures tend to increase as the missing fraction increases, the clusters become noisier and the number of covariates increase. Generally speaking, MI tends to perform the worst (which is to be expected).

Focusing now on MSPE, results are very similar among the procedures (save PSM) when there are no missing values, with the PPMx and PPMx,PSM predicting better relative to other procedures as the noise in the clusters and the number of covariates increase. We do note that the PSM and MI are at an inherent disadvantage as no linear model was explicitly included in the top level sampling model. When the missing rate increases, the prediction accuracy of the PSM and MI degrades the most (which was expected). However, as the number of covariates increase, the PPMx,PSM procedure tends to break down. This to is expected as the number of missing patterns and as a result the number of linear models increases drastically. However, it seems that the PPMx and BART generally predict better as the number of covariates increase (with the PPMx doing better). It also appears that generally speaking the PPMx procedure is least impacted by increasing the missing fraction and the cluster noise. Overall, the simulation study seems to indicate that our extension to the PPMx does very well in accommodating missing values relative to BART, MI, and PSM, and that this is regardless of the type of missingness (see
Figures ?? - ?? in the online supplementary material).

Figure 1: MSE results from simulation study when missing is not at random and the $s^2$ is not constant across clusters.

### 3.2 Ozone Data Example

We consider a small environmental data set that is publicly available[^1]. This data set consists of 112 measurements of the maximum daily ozone in Rennes. In addition, temperature (T), nebulosity (Ne), and projection of wind speed vectors (Vx) were measured at three times during the day (9:00, 12:00, and 15:00 hours) resulting in nine covariates. There are 16 locations for which the

[^1]: https://github.com/njtierney/user2018-missing-data-tutorial/blob/master/ozoneNA.csv
Figure 2: MSPE results from simulation study when missing is not at random and the $s^2$ is not constant across clusters.
response (maximum daily ozone measurements) is missing. This could be handled with any of the existing methods in the literature focused on missing responses. However, for the sake of simplicity we remove these observations. Figure 3 displays the amount of missing for each covariate and the missing patterns. Notice that there are a number of missing patterns that appear only one time and 14.6% of observations are complete cases.

![Figure 3: Missing rates and patterns associated with the ozone data set. The left plot displays the percent missing for each covariate. In the right plot each row corresponds to a missing pattern with cells colored in dark gray indicating the covariate is missing](image)

The 96 observations are divided in training and testing datasets by randomly selecting 75 observations as training data and treating the remaining 21 as testing data. The procedure of randomly partitioning the data into training and testing observations is repeated 100 times and to each one the PPMx, PPMx_PSM, BART, MI, and PSM procedures are fit. For each of the fits the MSE and MSPE is calculated. Also, in order to further study how increasing the number of covariates (and as a result the number of missing patterns) impacts the out-of-sample prediction performance of our approach, we carry out first the process described previously for only two covariates (temperature at 9:00 and 12:00) and then three (temperature at 9:00, 12:00 and 15:00). This process was repeated by sequentially adding nebulosity and then projection of wind speed.
vectors for each time during the day. The MCMC specifications and prior values for models (8) and (9) are the same as those used in the simulation study of Section 3.1. Also as in the simulation study, the remaining procedures are fit using the default tuning parameter values.

The average MSE and MSPE values over the 100 cross-validation data sets are provided in Figures 4 and 5. From Figure 4 notice that the MSE values for the PPMx and PPMx_PSM model are much lower than BART, MI, or PSM with PPMx_PSM lower than PPMx regardless of the number of covariates that are considered. In terms of out-of-sample prediction, it seems that PPMx has the lowest MSPE among the five methods regardless of the number of covariates. It seems that the PPMx_PSM method’s performance tends to decay as the number of covariates (and as a result the number of missingness patterns) increases.

![Figure 4: MSE values averaged over 100 cross-validation datasets based on ozone data](image)

4 Conclusions

We have extended the PPMx random partition model so that it is able to accommodate missing covariate values without having to adopt any of the assumptions that typically accompany impu-
Figure 5: MSPE values averaged over 100 cross-validation datasets based on ozone data

tation techniques. This is particularly useful when the main inferential target is of a predictive nature, as our approach facilitates making out-of-sample predictions regardless of whether missing patterns are observed in the training data or not. We also considered combining the extended PPMx with a pattern submodel and found in our simulation study that this approach performs well when the number of covariates is not large, and as a result, the number of missingness patterns is small. Incorporating ideas similar to those employed in Mercaldo and Blume (2018) that tend to scale well as the number of covariates increases is the topic of future research.

In preliminary results not shown, we explored the proposed method in the case when the underlying data structure is such that only a small number of covariates inform the partition relative to the total number measured. We found that the PPMx model in these circumstances is not as competitive as the BART approach, as indicated in our simulation study. In the case of a scenario with many covariates, we suggest first employing some dimension reduction or variable selection technique (one option is described in Page et al. 2018), and afterwards applying our approach based only on those covariates that are useful.
Acknowledgments

F. Quintana’s research is supported by Millennium Science Initiative of the Ministry of Economy, Development, and Tourism, grant “Millennium Nucleus Center for the Discovery of Structures in Complex Data”. F. Quintana is also supported by Fondecyt grant 1180034.

References

Lane F. Burgette and Jerome P. Reiter. Multiple imputation for missing data via sequential regression trees. *Practice of Epidemiology*, 172(9):1070–1076, 2010.

Michael J. Daniels and Joseph W. Hogan. *Missing Data in Longitudinal Studies*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, 2008.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 09 2006. doi: 10.1214/06-BA117A. URL [https://doi.org/10.1214/06-BA117A](https://doi.org/10.1214/06-BA117A).

Andrew Gelman and Jennifer Hill. Opening windows to the black box. *Journal of Statistical Software*, 40, 2011.

Rolf H.H. Groenwold, Ian R. White, A. Rogier T. Donders, James R. Carpenter, Douglas G. Altman, and Karel G.M. Moons. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Journal of Clinical Epidemiology*, 184:1265–1269, 2012.

Michael P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, pages 222–230, 1996.

Adam Kapelner and Justin Bleich. Prediction with missing data via Bayesian additive regression trees. *The Canadian Journal of Statistics*, pages 224–239, 2015.

Adam Kapelner and Justin Bleich. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016. doi: 10.18637/jss.v070.i04.

Roderick J. A. Little. Regression with missing x’s: A review. *Journal of the American Statistical Association*, 87:1227–1237, 1992.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley & Sons, 2 edition, 2002.

Sarah Fletcher Mercaldo and Jeffery D. Blume. Missing data and prediction: the pattern submodel. *Biostatistics*, 09 2018. ISSN 1465-4644. doi: 10.1093/biostatistics/kxy040. URL [https://doi.org/10.1093/biostatistics/kxy040](https://doi.org/10.1093/biostatistics/kxy040).

21
Karel G.M. Moons, Rogier A.R.T. Donders, Theo Stijnen, and Frank E. Harrell. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59 (10):1092 – 1101, 2006. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2006.01.009. URL [http://www.sciencedirect.com/science/article/pii/S0895435606000606](http://www.sciencedirect.com/science/article/pii/S0895435606000606).

Peter Müller, Fernando Quintana, and Gary L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–277, 2011.

Garritt L. Page and Fernando A. Quintana. Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates. *Bayesian Analysis*, 10(2):379–410, 06 2015. doi: 10.1214/14-BA919. URL [https://doi.org/10.1214/14-BA919](https://doi.org/10.1214/14-BA919).

Garritt L. Page and Fernando A. Quintana. Calibrating covariate informed product partition models. *Statistics and Computing*, 28:1009–1031, 2018.

Garritt L. Page, Fernando A. Quintana, and Gary L. Rosner. Discovering interactions using covariate informed random partition models. 2018. URL [https://arxiv.org/abs/1810.00121](https://arxiv.org/abs/1810.00121).

Ju-Hyun Park and David B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20:1203–1226, 2010.

Fernando A. Quintana. A predictive view of Bayesian clustering. *J. Statist. Plann. Inference*, 136 (8):2407–2429, 2006.

Fernando A. Quintana, Peter Müller, and Ana Luisa Papoila. Cluster-specific variable selection for product partition models. *Scandinavian Journal of Statistics*, 42:1065–1077., 2015. doi: 10.1111/sjos.12151.

Fernando A. Quintana, Rosangela H. Loschi, and Garritt L. Page. Bayesian Product Partition Models, pages 1–15. American Cancer Society, 2018. ISBN 9781118445112. doi: 10.1002/9781118445112.stat08123. URL [https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08123](https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08123).

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL [https://www.R-project.org/](https://www.R-project.org/).

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & sons, 1987.

Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.

Daniel J. Stekhoven and Peter Bühlmann. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28:112–118, 2012.
Curtis B. Storlie, Terry M. Therneau, Rickey E. Carter, Nicholas Chia, John R. Bergquist, Jeanne M. Huddleston, and Santiago Romero-Brufau. Prediction and inference with missing data in patient alert systems. *Journal of the American Statistical Society*, 00(0):1–28, 2019. doi: 10.1080/01621459.2019.1604359. URL [https://doi.org/10.1080/01621459.2019.1604359](https://doi.org/10.1080/01621459.2019.1604359).

Clifton D. Sutton. *Classification and Regression Trees, Bagging, and Boosting*, chapter 11, pages 303–329. American Cancer Society, 2005. ISBN 9781118445112. doi: 10.1016/S0169-7161(04)24011-1.

Stef van Buuren. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 2012. ISBN 9781439868249. URL [https://books.google.com/books?id=M89TDSml-FoC](https://books.google.com/books?id=M89TDSml-FoC).

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. URL [https://www.jstatsoft.org/v45/i03/](https://www.jstatsoft.org/v45/i03/).

Geert J.M.G. van der Heijden, A. Rogier T. Donders, Theo Stijnen, and Karel G.M. Moons. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology*, 59:1102–1109, 2006.

Dandan Xu, Michael J. Daniels, and Almut G. Winterstein. Sequential BART for imputation of missing covariates. *Biostatistics*, pages 589–602, 2016.

Yanxun Xu, Peter M"uller, Apostolia M. Tsimberidou, and Donald Berry. A nonparametric Bayesian basket trial design. *Biometrical Journal*, 61:1160–1174, 2019.