JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/10618600.2021.1999824

Clustering and Prediction With Variable Dimension Covariates

Garritt L. Page[a,b], Fernando A. Quintana[c,d], and Peter Müller[e]

[a]Department of Statistics, Brigham Young University, Provo, UT; [b]BCAM—Basque Center for Applied Mathematics, Bilbao, Spain; [c]Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile; [d]Millennium Nucleus Center for the Discovery of Structures in Complex Data, Santiago, Chile; [e]Department of Mathematics, The University of Texas at Austin, TX

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 random partition models that seamlessly handles missing covariates while completely avoiding any type of imputation. The method we develop allows in-sample as well as out-of-sample predictions, even if the missing pattern in the new subjects' incomplete covariate vector was not seen in the training data. Any data type, including categorical or continuous covariates are permitted. In simulation studies, the proposed method compares favorably. We illustrate the method in two application examples. Supplementary materials for this article are available here.

ARTICLE HISTORY
Received June 2020
Revised September 2021

KEYWORDS
Bayesian nonparametrics; Dependent random partition models; Indicator-missing; Pattern missing

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 missing covariates. Instead we start with a distribution for 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 exhibiting missing observations, 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 impact on inference for model parameters. The focus of this work is on incomplete covariate vectors and their impact on prediction accuracy. Even though incomplete predictor vectors are also common in practice (White and Carlin 2010) and can have adverse effects on prediction accuracy (destructive if an influential predictor is missing; see e.g. Mercaldo and Blume 2020), the missing observations literature is less developed 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 2020). 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, Daniels, and Winterstein (2016) employed 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) employed 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) used random forests to carry out imputation. Recently, Storlie et al. (2020) built 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 robust. All these and most other multiple imputation type approaches focus on inference for model parameters. If prediction is tangentially considered, then 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). Considering these limitations, our interest lies in developing a procedure that avoids imputation while still providing a good and flexible model for the available data.

CONTACT Garritt L. Page page@stat.byu.edu Department of Statistics, Brigham Young University, Provo, UT.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

The so-called missing indicator approach (Little 1992, Jones 1996, van der Heijden et al. 2006) has been developed to avoid 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; 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 start with a covariate-dependent random partition model that naturally allows for missing values in the covariates, and can accommodate mixed-type covariates. Adding a cluster-specific sampling model to the random partition defines a posterior predictive distribution that makes out-of-sample prediction straightforward. Covariate-dependent random partitions are particularly well suited for prediction, as they permit complex interactions and nonlinear associations between covariates and responses, simply by including corresponding clusters in the partition. Perhaps, the missing data method whose focus is most similar to what we develop is found in Kapelner and Bleich (2015). They use BART to implement 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 a random partition model (rather than deterministic partition creation), which permits more flexibility in how covariates interact and accounts for all uncertainties. Throughout we assume that missing data are missing at random (MAR), with some exceptions. Simple MNAR due to a detection limit, for example, is easily accommodated by introducing an additional binary covariate.

The remainder of the article is organized as follows. In Section 2, we provide background associated with covariate dependent product partition models. Section 3 describes our extension that permits incomplete covariates vectors of varying dimensions. Section 4 contains a simulation study while data applications are described in Section 5. Some concluding remarks are provided in Section 6.

2. A Covariate-Dependent Product Partition Model

We build on a covariate-dependent partition model proposed by Müller, Quintana, and Rosner (2011). We introduce notation by way of a brief review of their approach. For more details see Müller, Quintana, and Rosner (2011), Park and Dunson (2010), or Quintana, Loschi, and Page (2018).

Let $i = 1, \ldots, m$ index m experimental units. Let $\rho_m = \{S_1, \ldots, S_{k_m}\}$ denote a partition (or clustering) of the m units into k_m nonempty and exhaustive subsets so that $\{1, \ldots, m\} = \bigcup_j S_j$, for disjoint subsets S_j. To simplify notation we omit the subscript m for ρ unless explicitly needed. A common alternative representation of ρ introduces cluster membership indicators $c_i = j$ if $i \in S_j$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ denote a $1 \times p$ covariate vector measured on unit i and $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$. Further, let $\boldsymbol{x}_j^\star = \{\boldsymbol{x}_i : i \in S_j\}$ denote covariate vectors arranged by clusters. We will generally use a superscript "$\star$" to mark cluster-specific entities. The covariate-dependent product partition model (PPMx) prior on ρ formalizes the idea that units with similar covariate values are more likely a priori to belong to the same cluster than 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 hyper-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 $g(\boldsymbol{x}_j^\star \mid \boldsymbol{\xi})$ and parameterized by $\boldsymbol{\xi}$, formalizes the "closeness" of the x_i's in a cluster by producing larger values of $g(\boldsymbol{x}_j^\star \mid \boldsymbol{\xi})$ 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, sec. 2.4). With the similarity and cohesion functions, the form of the PPMx prior is the following product:

$$p(\rho \mid \boldsymbol{x}, M, \boldsymbol{\xi}) \propto \prod_{j=1}^{k_m} c(S_j \mid M) g(\boldsymbol{x}_j^\star \mid \boldsymbol{\xi}). \tag{1}$$

The cohesion function we employ in what follows is $c(S_j \mid M) = M \times (|S_j| - 1)!$ for some positive M and $|\cdot|$ denoting cardinality. This cohesion is commonly employed as the corresponding prior $p(\rho)$ is identical to the popular Chinese restaurant process (Aldous 1985; Broderick, Jordan, and Pitman 2013). Regarding possible similarity functions, Müller, Quintana, and Rosner (2011) discussed choices for different covariate data types (continuous, ordinal, or categorical), and suggest using

$$g(\boldsymbol{x}_j^\star \mid \boldsymbol{\xi}) = \int \prod_{i \in S_j} q(\boldsymbol{x}_i \mid \boldsymbol{\zeta}_j) q(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}) d\boldsymbol{\zeta}_j \tag{2}$$

where $q(\boldsymbol{x}_i \mid \boldsymbol{\zeta}_j)$ and $q(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi})$ are a conjugate pair whose form depends on the type of covariate and $\boldsymbol{\xi}$ is a fixed "hyper-parameter". With a conjugate pair, the integral in (2) can be evaluated in closed form. For example, in the numerical experiments in Section 4 and data applications in Section 5 we use $q(\cdot \mid \zeta_j) = N(\cdot; \zeta_j, v_x^2)$ and $q(\zeta_j) = N(\zeta_j; m_x, s_x^2)$. This results in $\boldsymbol{\xi} = (v_x^2, m_x, s_x^2)$ and $g(\boldsymbol{x}_j^\star \mid \boldsymbol{\xi}) = N_{n_j}(m_x \boldsymbol{j}_{n_j}, s_x^2 \boldsymbol{J}_{n_j} + v_x^2 \boldsymbol{I}_{n_j})$ where $\boldsymbol{j}_{n_j}$ is a n_j-dimensional vector of ones, $\boldsymbol{J}_{n_j}$ is a $n_j \times n_j$ matrix of ones, and $\boldsymbol{I}_{n_j}$ is a n_j-dimensional identity matrix. For simplicity, we use (2) for scalar covariates only and construct $g(\cdot)$ for multivariate $\boldsymbol{x}_i$ using separate similarity functions g_ℓ for each covariate and set $g(\boldsymbol{x}_j^\star \mid \boldsymbol{\xi}) = \prod_{\ell=1}^{p} g_\ell(\boldsymbol{x}_{j\ell}^\star \mid \boldsymbol{\xi})$ where $\boldsymbol{x}_{j\ell}^\star = \{x_{i\ell} : i \in S_j\}$. See Page and Quintana (2018) for more discussion on other possible specifications for the similarity function.

For a given cluster arrangement ρ, we complete the model construction with a sampling model for the response y_i by introducing cluster-specific parameters $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_{k_m}^\star)$ and assuming conditional independence at the observation level. Letting y_i denote the ith response and $\boldsymbol{y} = (y_1, \ldots, y_m)$ this leads to the following model:

$$\begin{aligned} p(\boldsymbol{y}, \rho, \boldsymbol{\theta}^\star \mid \boldsymbol{x}, M, \boldsymbol{\xi}) &= p(\boldsymbol{y} \mid \rho, \boldsymbol{\theta}^\star)\, p(\rho \mid \boldsymbol{x}, M, \boldsymbol{\xi}) p(\boldsymbol{\theta}^\star) \\ &\propto \prod_{j=1}^{k_m} \left\{ \left(\prod_{i \in S_j} p(y_i \mid \boldsymbol{\theta}_j^\star) \right) p(\boldsymbol{\theta}_j^\star) \right\} \\ &\quad \times p(\rho \mid \boldsymbol{x}, M, \boldsymbol{\xi}), \end{aligned} \tag{3}$$

where $p(\boldsymbol{\theta}^\star)$ is a prior distribution for $\boldsymbol{\theta}^\star$ whose components are assumed to be independent and identically distributed. The model can be written in hierarchical form using latent cluster membership indicators,

$$\begin{aligned} &y_i \mid \boldsymbol{\theta}^\star, c_i = j \overset{\text{ind}}{\sim} p(y_i \mid \boldsymbol{\theta}_j^\star) \\ &\boldsymbol{\theta}_j^\star \mid \rho \overset{\text{iid}}{\sim} p(\boldsymbol{\theta}_j^\star \mid \rho) \\ &p(\rho = \{S_1, \ldots, S_{k_m}\} \mid \boldsymbol{x}, M, \boldsymbol{\xi}) \propto \textstyle\prod_{j=1}^{k_m} c(S_j \mid M) g(\boldsymbol{x}_j^\star \mid \boldsymbol{\xi}). \end{aligned} \tag{4}$$

The exact form of sampling model $p(y_i \mid \boldsymbol{\theta}_j^\star)$ depends on the type of response. For example, in the data applications in Section 5 we consider both a continuous and binary response. The former naturally leads to using a Normal sampling model while the later to a probit regression-type sampling model. Finally, neither the independence nor the iid assumption in the first two lines of (4) are strictly needed for the upcoming discussion, and could be relaxed.

3. PPMx With Missing Covariates

We extend the PPMx model (4) to allow for variable dimension covariate vectors. In short, we generalize the similarity function $g(\boldsymbol{x}_j^\star)$ in the prior for the random partition to use only available covariates. We introduce this construction next. This will eventually lead to a variable-dimension covariate regression. We refer to the proposed model in general, and the implied variable-dimension covariate regression as VDReg.

3.1. Random Partitions With Variable-Dimension Covariates

To develop an extension of the PPMx model that accommodates missing covariates, denote by $\mathcal{O}_i$ the collection of covariate indices that are observed for subject i. The ith subject's observed covariate vector can be now denoted as $\boldsymbol{x}_i^o = \{x_{i\ell} : \ell \in \mathcal{O}_i\}$ and the collection of observed covariate vectors that belong to the jth cluster is $\boldsymbol{x}_j^{\star o} = \{\boldsymbol{x}_i^o : i \in S_j\} = \{x_{i\ell} : \ell \in \mathcal{O}_i, i \in S_j\}$. Then missing covariates can be accommodated in the PPMx by evaluating the similarity function g_ℓ using only subjects $i \in \mathcal{C}_{j\ell} = \{i : i \in S_j, \ell \in \mathcal{O}_i\}$, i.e., those with observed covariate ℓ. Letting $\boldsymbol{x}_{j\ell}^{\star o} = \{x_{i\ell} : i \in \mathcal{C}_{j\ell}\}$, we define a modified similarity function as

$$\begin{aligned} \tilde{g}(\boldsymbol{x}_j^{\star o} \mid \boldsymbol{\xi}) &\overset{\text{def}}{=} \prod_{\ell=1}^{p} \tilde{g}_\ell(\boldsymbol{x}_{j\ell}^{\star o} \mid \boldsymbol{\xi}_\ell) \overset{\text{def}}{=} \prod_{\ell=1}^{p} \\ &\int \prod_{i \in \mathcal{C}_{j\ell}} q(x_{i\ell} \mid \boldsymbol{\zeta}_{j\ell})\, dq(\boldsymbol{\zeta}_{j\ell} \mid \boldsymbol{\xi}_\ell). \end{aligned} \tag{5}$$

Importantly, in the presence of missing covariates, the similarity function for the ℓth 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.

We note briefly that in the context of variable selection Quintana, Müller, and Papoila (2015) considered similarity functions that are similar in form to (5), but with each cluster selecting a cluster-specific subset of covariates. Importantly, in that application $\mathcal{C}_{j\ell}$ is a random cluster-specific parameter that includes the subset of covariates that were selected for the jth cluster. In that case it is important that $g(x_{j\ell}^\star)$ be scaled such that $g(x_{j\ell}^\star) > 1$ for $x_{j\ell}^\star$ that are judged to be very similar and $g(x_{j\ell}^\star) < 1$ for very diverse $x_{j\ell}^\star$. That is, $g(\cdot)$ needs to be centered around 1, lest it would introduce an inappropriate prior probability for including a variable. Quintana, Müller, and Papoila (2015) introduced an additional factor to ensure such scaling. However, this issue does not arise here, since $\mathcal{C}_{j\ell}$ is fixed, that is, inference is conditioned on the observed covariates.

3.2. Variable Dimension Covariate Regression (VDReg)

An important feature of the PPMx prior on partitions is the flexibility in capturing the role of covariates in the predictive distribution which we now discuss. 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 dataset. To see this, consider the predictive multinomial probabilities that the $(m+1)$st subject belongs to one of the groups $h = 1, \ldots, k_m$ conditional on ρ_m:

$$\begin{aligned} &p(c_{m+1} = h \mid \rho_m, \boldsymbol{x}^o, \boldsymbol{x}_{m+1}^o) \\ &\quad \propto \begin{cases} \dfrac{c(S_h \cup \{m+1\})\tilde{g}(\boldsymbol{x}_h^{\star o} \cup \{\boldsymbol{x}_{m+1}^o\})}{c(S_h)\tilde{g}(\boldsymbol{x}_h^{\star o})} & \text{for } h = 1, \ldots, k_m \\ c(\{m+1\})\tilde{g}(\{\boldsymbol{x}_{m+1}^o\}) & \text{for } h = k_m + 1, \end{cases} \end{aligned} \tag{6}$$

where $\tilde{g}(\boldsymbol{x}_h^{\star o} \cup \{\boldsymbol{x}_{m+1}^o\})$ is computed including $i = m+1$ in S_h. That is, letting $\tilde{\mathcal{C}}_{j\ell} = \{i : \ i \in S_j \cup \{m+1\}$ and $\ell \in \mathcal{O}_i\}$, we define

$$\tilde{g}(\boldsymbol{x}_h^{\star o} \cup \{\boldsymbol{x}_{m+1}^o\}) = \prod_{\ell=1}^{p} \int \left\{ \prod_{i \in \tilde{\mathcal{C}}_{j\ell}} q(x_{i\ell} \mid \boldsymbol{\zeta}_{h\ell}) \right\} dq(\boldsymbol{\zeta}_{h\ell} \mid \boldsymbol{\xi}_\ell). \tag{7}$$

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 always be evaluated. In the extreme case of a "new" subject with an entirely missing covariate vector, then $\boldsymbol{x}_h^{\star o} \cup \{\boldsymbol{x}_{m+1}^o\} = \boldsymbol{x}_h^{\star o}$ implying that $\tilde{g}(\boldsymbol{x}_h^{\star o} \cup \{\boldsymbol{x}_{m+1}^o\}) = \tilde{g}(\boldsymbol{x}_h^{\star o})$ and thus the conditional probabilities for the cluster membership indicator in (11) reduce to those when making predictions using the PPM.

To allow for prediction we add the sampling model (4) (first two lines) to include responses y_i. In the full model, posterior predictive probabilities (11) for the cluster membership c_{m+1} imply a flexible regression for y_{m+1} on $\boldsymbol{x}_{m+1}^o$. In words, the regression is described as a locally weighted mixture of predictions under different clusters, with the local weighting induced by (11) and all being marginalized with respect to posterior uncertainty on the clustering. The local weighting introduces the regression on $\boldsymbol{x}_{m+1}^o$, with the desired feature of allowing variable dimension $\boldsymbol{x}_{m+1}^o$. This is because only observed covariates

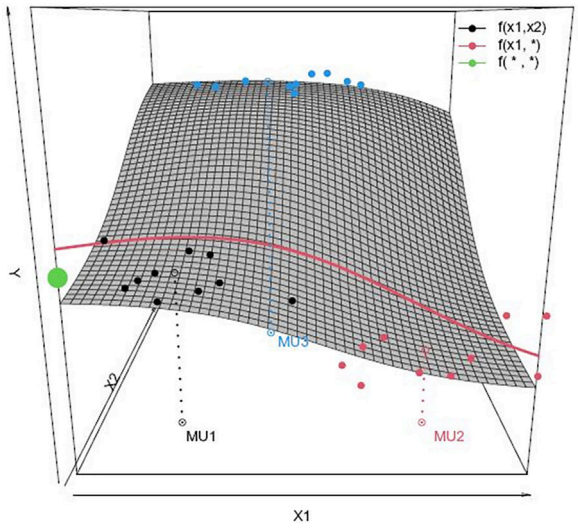

Figure 1. Regression with variable dimension covariates, for data (x_{1i}, x_{2i}, y_i) with black, blue, and red indicating data linked to three imputed clusters. The plot shows the posterior predictive regression $f(x_1, x_2) = E(y_{n+1} \mid x_{n+1,1} = x_1, x_{n+1,2} = x_2$, data) for a future data point with both, (x_1, x_2) observed (gray response surface), $f(x_1, \bullet)$ for a data point $(x_1, \bullet)$ with missing x_2 (red curve) and f for a future observation without available covariates (green bullet on the Y-axis).

are used in (11). Formally, let $w_j(\boldsymbol{x}^o_{m+1};\ \boldsymbol{x}^{\star o}_j) = p(c_{m+1} = h \mid \rho_m, \boldsymbol{x}^o, \boldsymbol{x}^o_{m+1})$, and let $f_j(y_{m+1};\ \boldsymbol{y}^\star_j) = \int p(y_{m+1} \mid c_{m+1} = j, \boldsymbol{y}^\star_j, \boldsymbol{\theta}^\star_j)\, dp(\boldsymbol{\theta}^\star_j \mid \boldsymbol{y}^\star_j)$. We get a locally weighted regression

$$p(y_{m+1} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}_{m+1}, \rho_m) = \sum_{j=1}^{k_m} w_j(\boldsymbol{x}_{m+1};\ \boldsymbol{x}^\star_j)\, f_j(y_{m+1}; \boldsymbol{y}^\star_j), \quad (8)$$

and finally

$$p(y_{m+1} \mid \boldsymbol{x}^o, \boldsymbol{y}, \boldsymbol{x}^o_{m+1}) = E_\rho \left\{ \sum_{j=1}^{k_m} w_j(\boldsymbol{x}_{m+1};\ \boldsymbol{x}^\star_j) f_j(y_{m+1}; \boldsymbol{y}^\star_j) \mid \boldsymbol{y} \right\}. \quad (9)$$

Expression (8) clearly exposes how the variable dimension covariate regression is implemented by using weights w_j that only make use of available covariates. The final equation averages over the unknown partition, with respect to $p(\rho_m \mid \boldsymbol{y}, \boldsymbol{x}^o)$. The regression (9) concisely summarizes the proposed approach to implement variable dimension covariate regression. In summary, by the implied posterior predictive distribution (9) the proposed model defines a variable dimension covariate regression. This is illustrated in Figure 1.

For the upcoming simulation studies and ozone data example, we implement the model for continuous outcomes y_i and continuous covariates $x_{i\ell}$, using a normal sampling model $p(y_i \mid c_i = j, \boldsymbol{\theta}^\star_j = (\mu^\star_j, \sigma^{2\star}_j)) = N(\mu^\star_j, \sigma^{2\star}_j)$ with a conjugate normal prior on the location parameter and a uniform prior on cluster-specific standard deviations $\sigma^\star_j$. For the cohesion function, we use $c(S_j \mid M) = M(|S_j| - 1)!$. The similarity functions are specified using (5), with $q(x_{i\ell} \mid \boldsymbol{\zeta}_{j\ell})$ and $q(\boldsymbol{\zeta}_{j\ell} \mid \boldsymbol{\xi})$ corresponding to the Normal–Normal pair detailed in Section 2 with fixed values for $\boldsymbol{\xi} = (v^2_x, m_x, s^2_x)$ and other sampling model hyperparameters. For later reference, we summarize the complete VDReg model with these choices:

$$\begin{aligned}
&y_i \mid \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*}, c_i = j \sim N(\mu^*_j, \sigma^{2*}_j) \text{ for } i = 1, \ldots, m\\
&(\mu^*_j \mid \rho) \sim N(\mu_0, \sigma^2_0) \text{ for } j = 1, \ldots, k_m\\
&(\sigma^\star_j \mid \rho) \sim \text{Uniform}(0, a_\sigma) \text{ for } j = 1, \ldots, k_m\\
&\mu_0 \sim N(m_0, v^2)\\
&\sigma_0 \sim \text{Uniform}(0, a_{\sigma_0})\\
&p(\rho = \{S_1, \ldots, S_{k_m}\} \mid \boldsymbol{x}^o, M, \boldsymbol{\xi}) \propto \textstyle\prod_{j=1}^{k_m} c(S_j \mid M) \tilde{g}(\boldsymbol{x}^{\star o}_j \mid \boldsymbol{\xi}).
\end{aligned} \quad (10)$$

The uniform prior on cluster-specific standard deviations follows suggestions in Gelman (2006). The function $\tilde{g}(\cdot)$ in the last line is where the model accommodates variable-dimension covariate vectors using (7) to define a similarity function on the basis of available covariates only. This is at the heart of the proposed VDReg model.

3.3. Posterior Computation

Fitting the model detailed in Equation (10) requires an MCMC algorithm that samples from the joint posterior distribution of model parameters. Our approach is to employ a hybrid Gibbs-sampler with Metropolis steps that is based on algorithm 8 of Neal (2000). Based on Algorithm 8, the full conditional probability of c_i for $h = 1, \ldots, k^{-i}_m$ (where k^{-i}_m denotes the number of clusters after having removed the ith observation) is

$$Pr(c_i = h|-) \propto \begin{cases} N(y_i; \mu^\star_h, \sigma^{2\star}_h) \dfrac{c(S^{-i}_h \cup \{i\}) \tilde{g}(\boldsymbol{x}^{\star o(-i)}_h \cup \{\boldsymbol{x}^o_i\})}{c(S^{-i}_h) \tilde{g}(\boldsymbol{x}^{\star o(-i)}_h)} & \text{for } h = 1, \ldots, k^{-i}_m \\ N(y_i; \mu^\star_{new,h}, \sigma^{2\star}_{new,h}) c(\{i\}) \tilde{g}(\{\boldsymbol{x}^o_i\}) & \text{for } h = k^{-i}_m + 1, \end{cases} \quad (11)$$

where $\mu^\star_{\text{new},h}$, and $\sigma^{2\star}_{\text{new},h}$ are drawn from their respective prior distributions. The main difference between the cluster probabilities in (11) for complete data and that for missing data is that an indicator matrix that identifies which covariate values are missing must be carried along. To update the variance components $(\sigma^{2\star}_1, \ldots, \sigma^{2\star}_{k_m}, \sigma^2_0)$, we use a random walk Metropolis step with a Normal proposal density. The means in both levels of model (10) $(\mu\star_1, \ldots, \mu^\star_{k_m}, \mu_0)$ are updated with a Gibbs step as their full conditionals are Normal and can be derived using well-known arguments.

Introducing a missingness indicator in the MCMC algorithm does not seem to adversely affect its behavior relative to an algorithm with no missing covariate values. For example, the rate of convergence and the mixing is comparable, in our experience, to posterior sampling of mixture models when there is no missingness. It is worth mentioning that as p and/or m increase, the algorithm's speed decreases. In fact, for a moderate number of covariates (30–50), it might be worth exploring alternative computing approaches for $m > 1000$. This is actually a problem that is not exclusive of VDReg, but is shared by many related approaches that are based on updating cluster membership indicators in a random partition model.

4. Numerical Experiments

We conduct two simulation studies. The first is carried out to assess how predictions are affected by (a) an increase in the number of covariates and the missingness rate, (b) different types of missingness, and (c) how informative covariates are in forming clusters. Datasets in the first simulation are generated using clusters that are covariate dependent. The second experiment is conducted to study how the VDReg method performs when datasets are not created using covariate informed clusters.

4.1. Simulation Study 1

4.1.1. Simulation Scenarios

We generate data with 100 testing and 100 training observations. We describe next the generation of covariates, responses, and the missing data.

Covariates: datasets are generated with a varying number of covariates, $p \in \{2, 4, 10\}$. Specific values for the covariates are generated using four p-dimensional Gaussian distributions, thus creating $k_m = 4$ covariate dependent clusters. For example, when $p = 2$, we use four bivariate normals, $N(m_j, V_j)$, with $m_j = (1,1), (1,-1), (-1,1)$, and $(-1,-1)$ to generate 200 sets of covariate values $\boldsymbol{x}_i$ (50 in each cluster). Similarly, with $p = 4$, we use $m_j = (1,1,1,1), (1,-1,1,-1), (-1,1,-1,1)$, and $(-1,-1,-1,-1)$, to create $k_m = 4$ clusters with 50 observations each. And lastly, for $p = 10$, we use $m_j = (1,1,1,1,1,1,1,1,1,1), (1,-1,1,-1,1,-1,1,-1,1,-1)$, $(-1,1,-1,1,-1,1,-1,1,-1,1)$, and $(-1,-1,-1,-1,-1,-1,-1,-1,-1,-1)$. Notice that as p increases $k_m = 4$ remains constant, but the clusters become sparser in the covariate space. To study how the overlap among clusters ("cluster noise") affects prediction results, the covariance matrix V_j used to generate covariates is set as $V_j = s^2 \boldsymbol{I}_p$ with $s^2 \in \{0.25^2, 0.5^2, 0.75^2\}$. Under $s^2 = 0.25^2$, the clusters are well separated, while for $s^2 = 0.5^2$ the clusters are adjacent, and for $s^2 = 0.75^2$ the clusters overlap substantially. Figure 1 in the online supplementary material displays the cluster configuration for each of the s^2 when $p = 2$. This describes the generation of the covariates $\boldsymbol{x}_i$.

Responses: y_i are generated as $p(y_i \mid s_i = j) = N(\mu_j^\star, \sigma_j^{2\star})$. We use $\mu_j^\star = -1, -0.5, 0$, and 0.5 for observations in clusters $j = 1$ through 4, respectively. To study how heterogeneity of variances across clusters impacts inference we use two sets of simulations, one with $\sigma_j^{2\star} = 0.25^2$ for all j, and one with $\sigma_j^\star = 0.1, 0.25, 0.5$ and 0.75, respectively for $j = 1, \ldots, 4$.

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, Lugtig and Vink 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, we generate data under simulation truths varying the following factors: (A) *type of missing* (MAR or MNAR), (B) *missing fraction* (0%, 10%, 25%, 50%), (C) *number of covariates* ($p = 2, 4$ and 10), (D) *cluster noise* ($s^2 \in \{0.25^2, 0.5^2, 0.75^2\}$), and (E) *heteroscedasticity* (yes, no).

Comparison. As mentioned, each created dataset is comprised of 200 observations (50 in each cluster). Then each dataset is randomly split into 100 training and 100 test observations. For each simulated dataset we implement inference under the following models and approaches: (a) BART: The method detailed in Kapelner and Bleich (2015) and carried out using the `bartMachine` package (Kapelner and Bleich 2016) in `R`; (b) MI: Based on 10 imputed datasets via the `complete` function of the `mi` package (Su et al. 2011) from the statistical software `R` (R Core Team 2018); (c) PSM: Pattern submodel approach using method in Mercaldo and Blume (2020) and code available at *https://github.com/sarahmercaldo/MissingDataAndPrediction*. (D) VDReg: model (10)

When fitting the VDReg model (10) covariates are standardized to zero mean and unit variance and the similarity parameters are set to $v_x^2 = 0.5$, $m_x = 0$, and $s_x^2 = 1$. Finally, fixed hyperparameters are $m_0 = 0$, $v^2 = 10^2$, $a_\sigma = 1$ and $a_{\sigma_0} = 2$. These values for hyperparameters were selected since the mean response for these data is relatively close to zero (-0.25 approximately) relative to the standard deviation (less that 1). With these prior specifications, we fit model (10) 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 model (10) is carried using the `gaussian_ppmx` function that is part of the `ppmSuite` R-package version `0.1.7` (Page 2021).

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 compare each method's ability to fit the data, we use the MSE (mean squared error), $MSE = \frac{1}{100}\sum_{i=1}^{100}(Y_{oi} - \hat{Y}_{oi})^2$, where i indexes the 100 training observations (Y_o) and $\hat{Y}_{oi}$ is the fitted value for the i observation. MSE quantifies in-sample prediction which may be a type of prediction that is of interest. We also include the MSPE (mean squared prediction error) which measures the out-of-sample predictive performance of the models, $MSPE = \frac{1}{100}\sum_{i=1}^{100}(Y_{pi} - \hat{Y}_{pi})^2$, where i indexes the 100 testing observations (Y_p) and $\hat{Y}_{pi} = E(Y_{pi} \mid \boldsymbol{Y}_o)$.

Results. Before describing simulation results, we note that under a missing rate of 50% and $p = 10$ covariates the software used to fit the PSM model exceeds an internal computational limit and aborts in error. As a result, PSM is not included in this scenario. We found that the simulation results are similar under the various combinations of data being MNAR or MAR, and whether or not we assume heteroscadasticity. We therefore present only results under MNAR and heteroscedasticity, and summarize other results in Section 1 of the supplementary

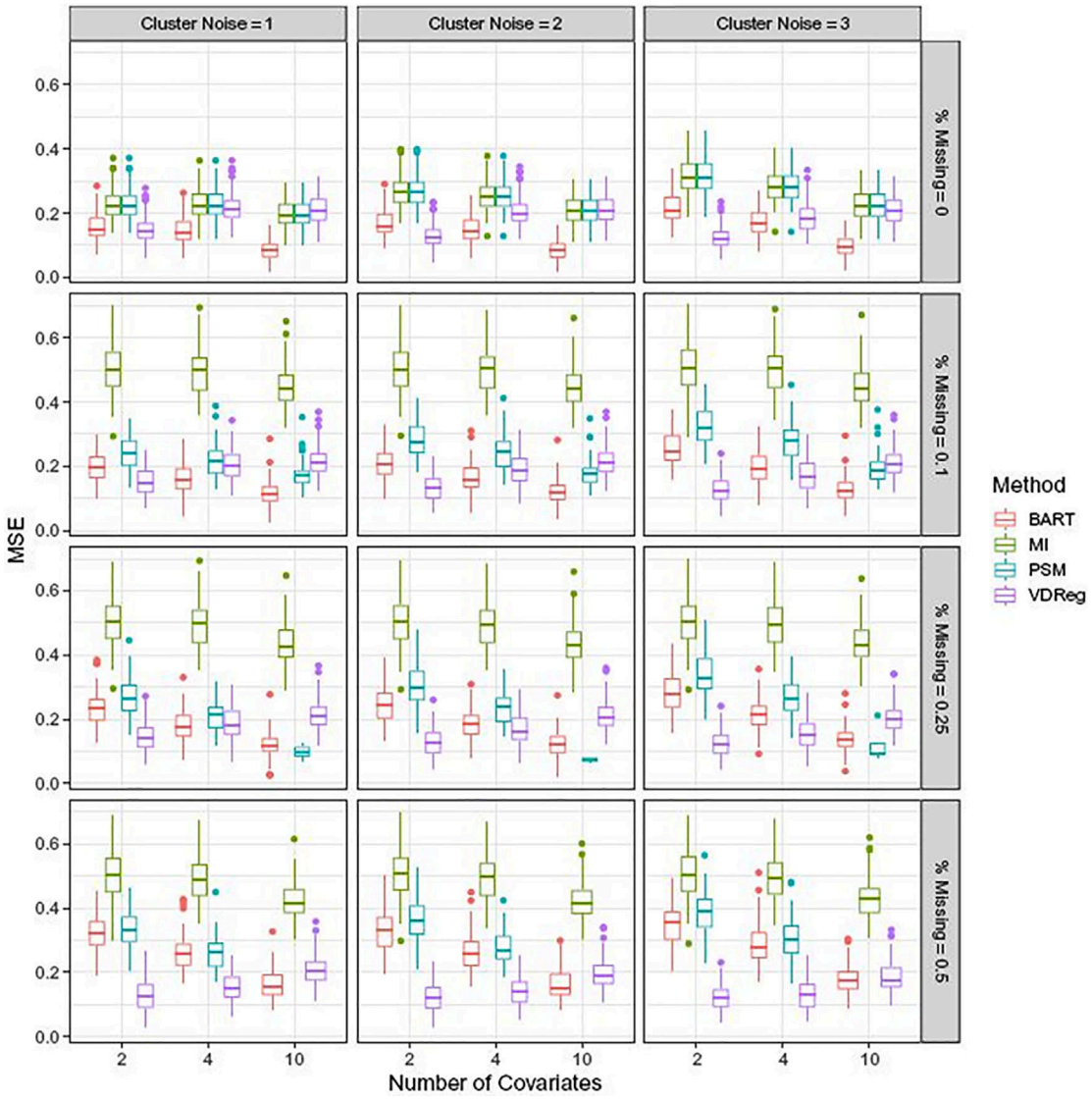

Figure 2. MSE results from simulation study when missing is not at random and s^2 is not constant across clusters. All methods except MI have the smallest MSE in at least on instance. VDReg's MSE tends to decrease as the missing percentage increases or as the number of covariates decreases. BART and PSM tend to fit the data better as the number of covariates increases.

material. Figures 2 and 3 display MSE and MSPE as a function of the number of covariates, missing fraction and cluster noise.

Focusing on the MSE values first, notice that under 0% missing fraction, 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, with increasing missing rate VDReg reports the best model fit, with the differences between procedures increasing with a higher missing fraction, cluster noise and p. Generally speaking, MI tends to perform least favorably (as one might expect of a very generic method).

Regarding MSPE, results are very similar across procedures when there are no missing values, with relative performance under VDReg looking increasingly better with increasing cluster noise and p. With increasing missing rate the prediction accuracy of the PSM and MI degrades the most (which was expected). VDReg and BART generally predict better as the number of covariates increases. Overall, VDReg is least impacted by an increase in the missing fraction and cluster noise. The simulation study indicates that VDReg performs favorably in accommodating missing values relative to BART, MI, and PSM, regardless of the type of missingness (see Figures 2–7 in the supplementary material).

4.2. Simulation Study 2

We now consider a simulation scenario based on the data generating mechanism found in Friedman (1991); see also Chipman, George and McCulloch (2010). These data do not originate from clusters that are covariate informed. As in Section 4.1, we consider the generation of covariates, responses, and the missing data individually.

Covariates: fix $p = 10$ and create covariate values using $x_{i1}, \ldots, x_{ip} \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$.

Responses: y_i are generated use $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ where

$$f(\boldsymbol{x}_i) = 10 \sin(\pi x_{i1} x_{i2}) + 20(x_{i3} - 0.5)^2 + 10x_{i4} + 5x_{i5}.$$

Notice that $x_{i6}, \ldots, x_{i10}$ are noise covariates as they do not contribute to the response value. We consider two different ϵ_i

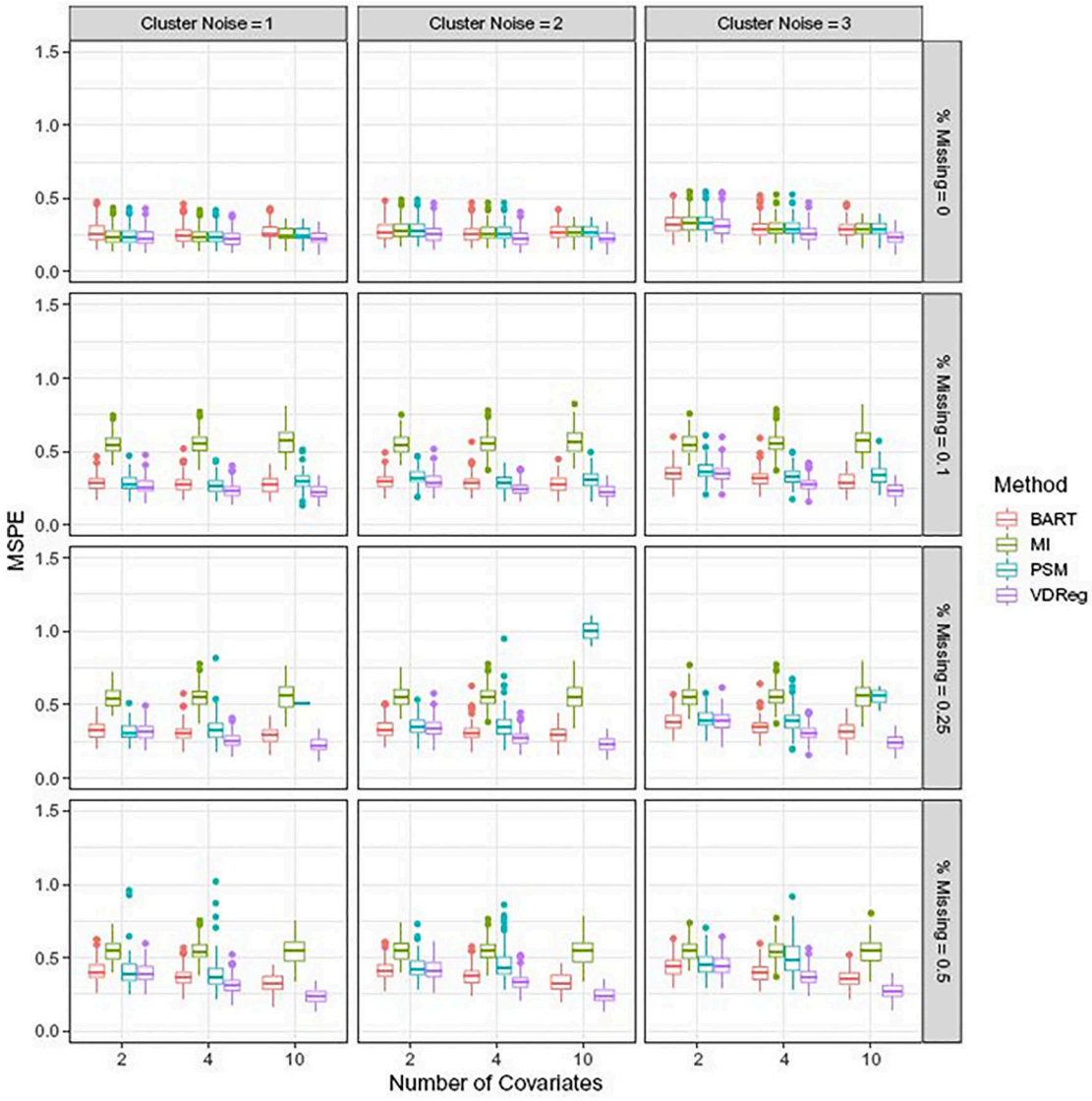

Figure 3. MSPE results from simulation study when missing is not at random and s^2 is not constant across clusters. VDReg and BART's MSPE values decrease as the number of covariates increases with VDReg outperforming BART. PSM seems to break down with more than four covariates with 50% missing. As expected MI performs the worst.

terms. The first is the iid case so that $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$. The second error terms depends on $\boldsymbol{x}_i$ such that $\epsilon_i \overset{\text{ind}}{\sim} N(0, \exp(x_{i1}))$

Missing values: in the covariates are inserted using the same procedure as described in Section 4.1.

Comparison. As in Section 4.1, each synthetic dataset is comprised of 100 training and 100 testing observations and the same competing methods are included. When fitting the VDReg model in this numerical study we do not standardize covariates and as such, we set $v_x^2 = 0.05$, $m_x = 0.5$, and $s_x^2 = 0.05$. In addition, sampling model hyper-parameters are set to $m_0 = 0$, $v^2 = 10^2$, $a_\sigma = 2$, and $a_{\sigma_0} = 2$. The same metrics in Section 4.1 are used to compare the four methods. Results are provided in Figures 4 and 5.

Results. Focusing on the MSE values, it can be easily seen from Figure 4 that VDReg outperforms the other methods. It seems that BART and MI are more severely impacted by missing covariates. That said, PSM is not even available for percent missing more than 25%. Focusing now on the MSPE values in Figure 5, it also appears that MSPE is not influenced much by idiosyncratic noise that depends on x_1 (unlike MSE values). In addition, it seems that PSM and MI perform the best when there are no missing covariates. This is to be expected as the linear model that they employ is the correct sampling model. As the amount of missing increases however, MI and PSM are the most severely impacted. In fact, it appears that the PSM is not a viable option for more than 10% missing. VDReg performs better than BART across the board and is comparable to PSM with 10% missing and better for any percent missing greater than 10%. Among the procedures, MI and BART seem to be impacted least by the type of missingness.

5. Application Examples

5.1. Ozone Data

We consider a small environmental dataset that is publicly available. This dataset consists of 112 measurements of maximum daily ozone in Rennes. In addition, temperature (T), nebulosity (Ne), and projection of wind speed vectors (Vx) were measured

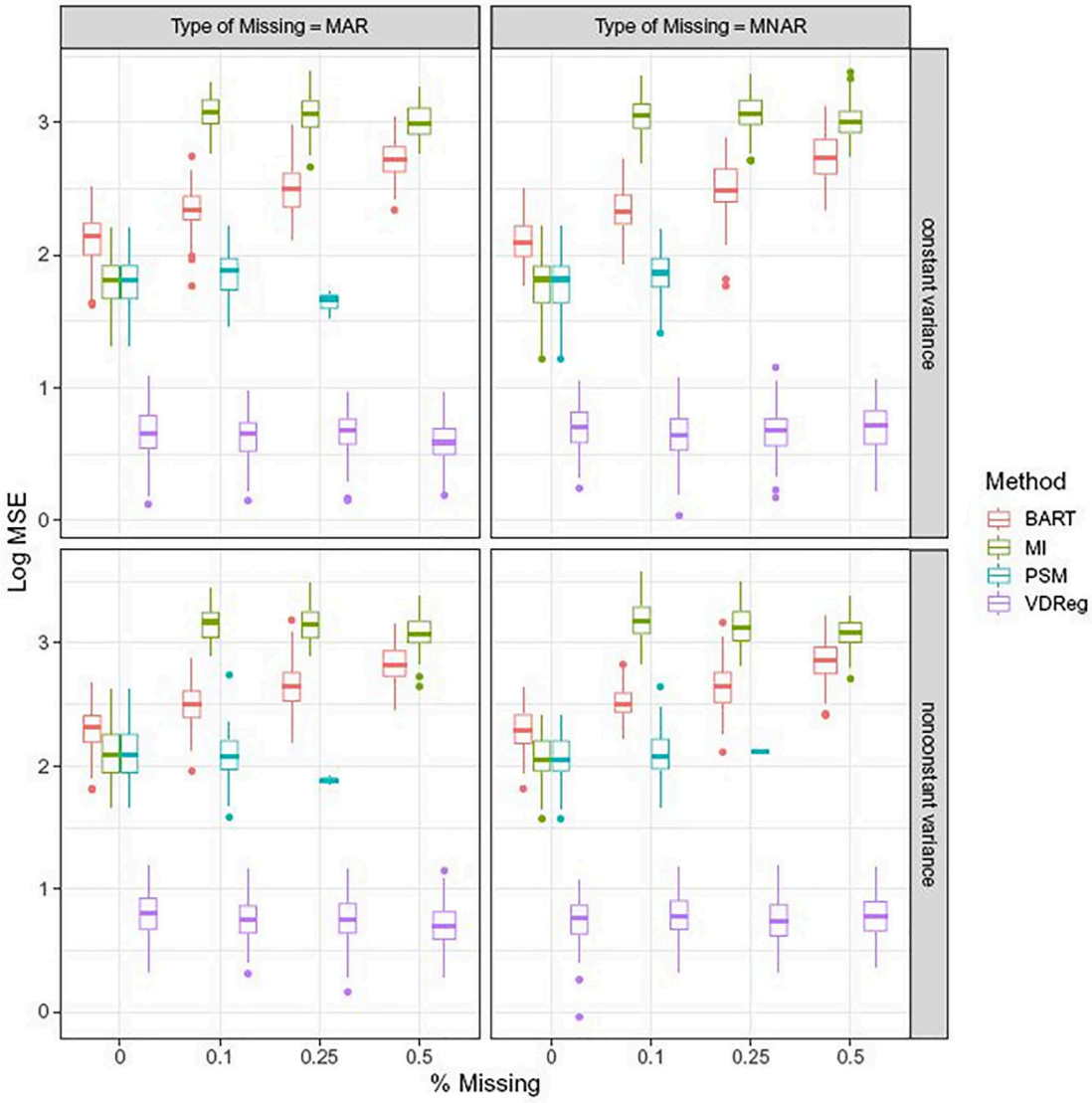

Figure 4. MSE results from the second simulation study. VDReg fit the data better than all the other methods in all scenarios. PSM breaks down if percent missing is greater than 25%.

three times daily (9:00, 12:00, and 15:00 hours) resulting in nine covariates. There are 16 locations for which the 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 6 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 only 14.6% of observations are complete cases.

The 96 observations are divided in training and test datasets by randomly selecting 75 observations as training data and treating the remaining 21 as test data. The procedure of randomly splitting into training and test data is repeated 100 times, and each time we fit the training data and make predictions for the testing data using BART, MI, PSM, and VDReg (see the previous section for a brief description of the methods). For each of the fits MSE and MSPE is calculated. Also, in order to further study how increasing p impacts the out-of-sample prediction performance we repeat the described process again using only $p = 2$ covariates (temperature at 9:00 and 12:00), then $p = 3$ (temperature at 9:00, 12:00 and 15:00), and next sequentially adding nebulosity and then projection of wind speed vectors for each time during the day. Since ozone values range between 42 and 166 we set the sampling model hyper-parameters to $m_0 = 0$, $v^2 = 10^2$, $a_\sigma = 10$, and $a_{\sigma_0} = 10$. The similarity inputs used and the number of MCMC samples collected for each cross-validation dataset correspond to those employed in the simulation study of Section 4.1.

The average MSE and MSPE values over the 100 cross-validation datasets are provided in Figures 7 and 8. From Figure 7 notice that the MSE values for the VDReg model are lower than BART, MI, or PSM regardless of the number of covariates that are considered. In terms of out-of-sample prediction, it seems that VDReg has the lowest MSPE among the five methods regardless of the number of covariates. It seems that the PSM method performed the worst with performance decaying drastically as the number of covariates is increased.

5.2. Prostate Cancer Data

We consider data from a prostate cancer study that was analyzed in Deng et al. (2016), who employ two variations of imputation

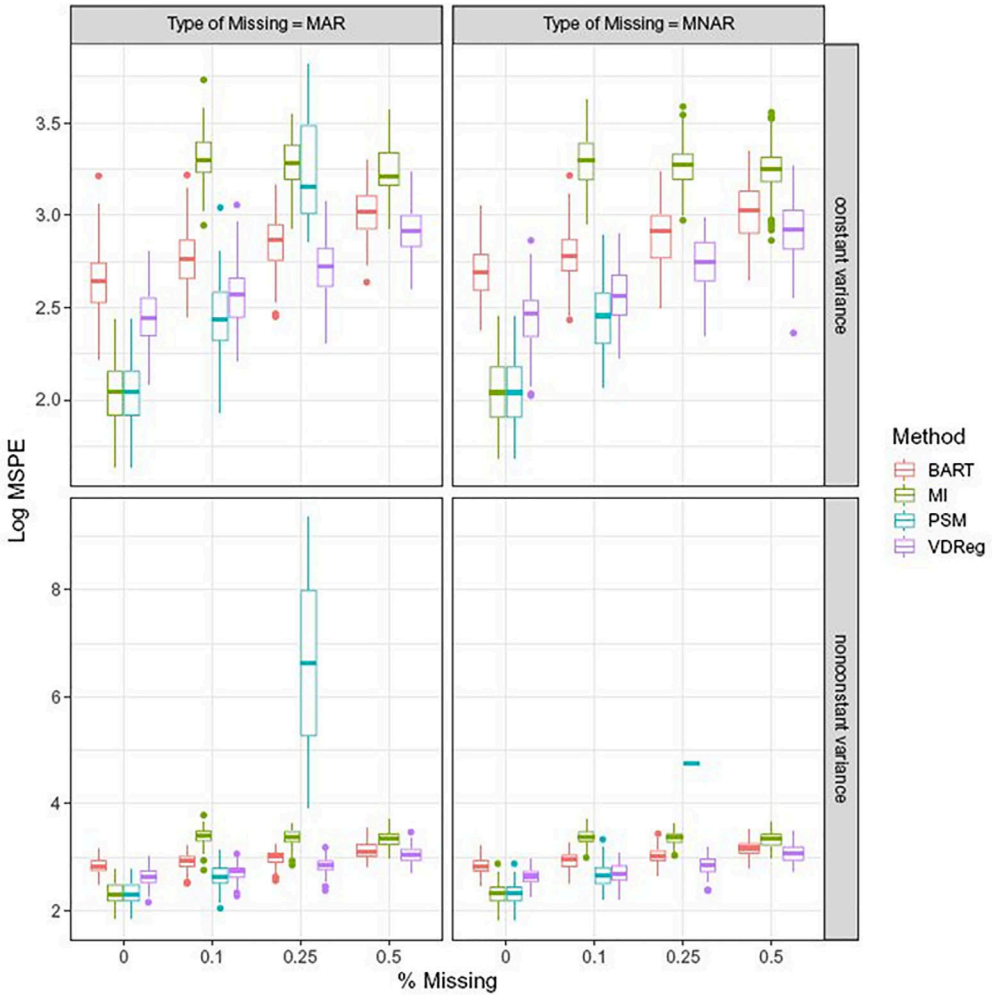

Figure 5. MSPE results from the second simulation study. PSM and MI had the lowest MSPE with no missing covariates. As the percent missing increases, all methods degrade in out-of-sample prediction with VDReg performing the best with percent missing greater than 10%.

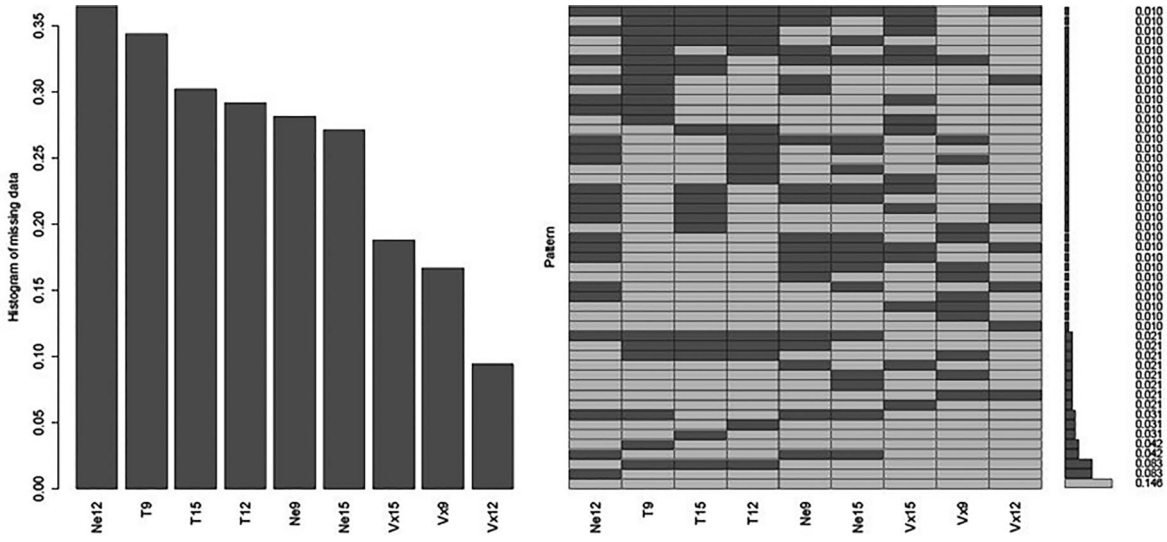

Figure 6. Missing rates and patterns associated with the ozone dataset. The left plot displays the percent missing for each covariate. In the right plot, each row corresponds to a missing pattern with cells colored in dark gray indicating the covariate is missing. The histogram in the right margin of the right plot corresponds to the fraction of observations for each missing pattern.

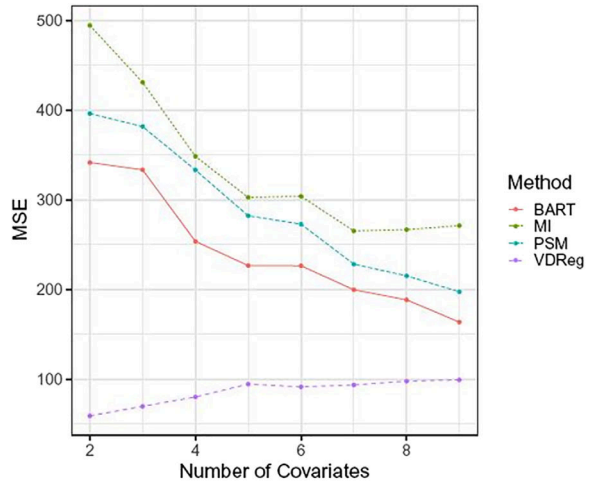

Figure 7. MSE values averaged over 100 cross-validation datasets based on ozone data

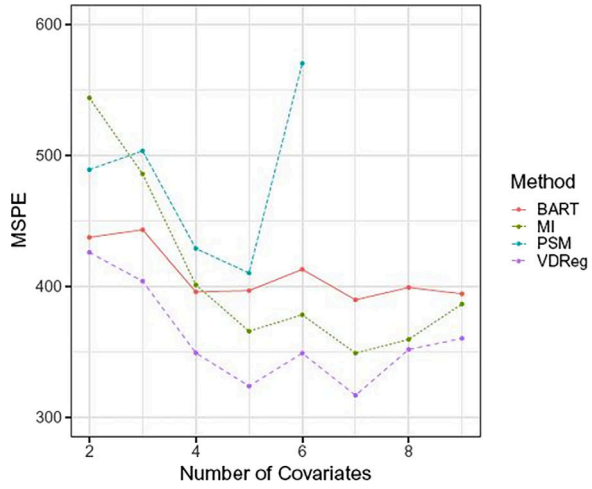

Figure 8. MSPE values averaged over 100 cross-validation datasets based on ozone data (values of PSM for $p = 7 - 9$ are beyond the plotting limits).

to accommodate missing covariates. The dataset is publicly available (GEO GDS3289) and is based on 99 subjects, including 34 benign and 65 malignant epithelium samples, each with 20,000 biomarkers. We code the response as $y_i = 1$ for benign samples and $y_i = 0$ for malignant. Besides a minor adaptation of (10) for the binary response in the sampling model, the VDReg model can be employed without alterations. We use a latent probit score, that is, $p(y_i = 1 \mid c_i = j, \mu^\star) = \Phi(\mu_j^\star)$ (with $\Phi(\cdot)$ denoting a standard normal c.d.f.), and otherwise leave (10) unchanged. Of the 20,000 biomarkers, Deng et al. (2016) focused on three (FAM178A, IMAGE:813259 and UGP2) that are known to be associated with the response. The missing rates of the three covariates are 31.3%, 45.5%, and 26.3%, respectively. Deng et al. (2016) used multiple imputation methods based on 2107 biomarkers that do not have any missing values and then using the imputed datasets, fit a logistic regression model and report estimates of the regression coefficients.

Since our focus is on prediction, we instead split the 99 subjects into 75 training and 24 testing observations and fit the VDReg model using $m_0 = 0, v^2 = 10^2, a_\sigma = 1$, and $a_{\sigma_0} = 2$ and standardized each of the three covariates to have mean zero and standard deviation one. Standardizing the covariates facilitates selecting values for the similarity and we employ $m_x = 0, s_x^2 = 1$ and $v_x^2 = 0.5$. Splitting the dataset into training and testing observations was carried out 100 times and for each split we evaluated within sample and out of sample predictions. It took approximately 55 seconds to sample 150,000 MCMC iterates, of which 1000 were retained after discarding the initial 100,000 as burn-in and thinning by 50. Results are shown in Table 1. In addition to prediction rates, we report Tjur's R^2 (Tjur 2009). This metric compares the average estimated probability of being in the benign group for subjects with benign samples to the average estimated probability of being in the malignant group for subjects with malignant samples. As this number approaches one, it is an indication of superior model fit. For comparison, we also include results under BART and PSM (as in Section 5.1). VDReg compares favorably to the other two methods in terms of in-sample prediction rate and Tjur's R^2 value. For out-of-sample prediction, VDReg does slightly better than the other two methods, but with worse Tjur's R^2.

Table 1. Cross validation results based on the prostate cancer data. Each of the 100 cross-validation datasets were comprised of 75 training and 24 testing observations.

	In Sample Prediction		Out Sample Prediction	
Method	% Correct	Tjur R^2	% Correct	Tjur R^2
BART	0.81	0.29	0.70	0.16
PSM	0.79	0.39	0.70	0.24
VDReg	0.99	0.44	0.71	0.14

NOTE: Results presented in the table are averages over the 100 cross-validation datasets.

Last, by way of comparison with the imputation methods used in Deng et al. (2016), using the estimated logistic regression coefficients reported in Deng et al. (2016), we predicted cancer status for the 26 complete cases found in the dataset. We then fit VDReg to all 99 observations and also predicted the cancer status of the 26 complete cases. Of these 26 predicted outcomes, the VDReg was correct for 88% of them compared to 69% based on the imputation methods.

6. Conclusions

We have extended the PPMx random partition model to allow for missing covariate values without resorting to any imputation or substitution. This is particularly useful when the main inferential target is prediction. The proposed approach facilitates out-of-sample predictions with any subset of covariates.

Some limitations remain, and provide opportunities for further generalizations. In the current form the model does not include any notion of variable selection or transformation. While independent variable selection is straightforward to add, the use of partially missing covariate vectors would complicate any approach that involves dependent priors over variables. Similarly, the use of any transformation or projections of the joint covariate vector is not straightforward in the presence of missing covariates without imputation. 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, Quintana, and Rosner 2021), and afterwards applying our approach based only on those covariates that are useful.

Funding

Garritt L. Page acknowledges support from the Basque Government through the BERC 2018-2021 program, by the Spanish Ministry of Science, Innovation and Universities through BCAM Severo Ochoa accreditation SEV-2017-0718. F. Quintana's research is funded by ANID - Millennium Science Initiative Program—NCN17_059. F. Quintana is also supported by FONDECYT grant 1180034. P. Müller acknowledges partial support fromgrantNSF/DMS 1952679 from the National Science Foundation, and under R01 CA132897 from the U.S. National Cancer Institute.

Supplementary material

The online supplementary material contains a document that provides additional results from the first simulation study and a zipped folder that contains all computer codes that were used to carry out both simulation studies and all data analysis.

References

Aldous, D. J. (1985), "Exchangeability and Related Topics," in *École d'été De Probabilités de Saint-Flour, XIII—1983*, ed. P. L. Hennequin, Vol. 1117 of *Lecture Notes in Math.*, Berlin: Springer, pp. 1–198. [2]

Broderick, T., Jordan, M. I., and Pitman, J. (2013), "Cluster and Feature Modeling From Combinatorial Stochastic Processes," *Statistical Science*, 28, 289–312. [2]

Burgette, L. F., and Reiter, J. P. (2010), "Multiple Imputation for Missing Data Via Sequential Regression Trees," *Practice of Epidemiology*, 172, 1070–1076. [1]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266 – 298. [6]

Daniels, M. J., and Hogan, J. W. (2008), *Missing Data in Longitudinal Studies*, Chapman & Hall/CRC Interdisciplinary Statistics, Boca Raton: CRC Press. [1]

Deng, Y., Chang, C., Ido, M. S., and Long, Q. (2016), "Multiple Imputation for General Missing Data Patterns in the Presence of High-Dimensional Data, *Scientific Reports*, 6, 21689. [8,10]

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1 – 67. [6]

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models" (comment on Browne and Draper), *Bayesian Analysis*, 1, 515–534. [4]

Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., and Moons, K. G. (2012), "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. [2]

Jones, M. P. (1996), "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression," *Journal of the American Statistical Association*, 91, 222–230. [2]

Kapelner, A., and Bleich, J. (2015), Prediction With Missing Data Via Bayesian Additive Regression Trees," *The Canadian Journal of Statistics*, 43, 224–239. [2,5]

——— (2016), "Bartmachine: Machine Learning With Bayesian Additive Regression Trees," *Journal of Statistical Software*, 70, 1–40. [5]

Little, R. J. A. (1992), "Regression With Missing x's: A Review," *Journal of the American Statistical Association*, 87, 1227–1237. [2]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), Hoboken, NJ: Wiley & Sons. [1]

Mercaldo, S. F. and Blume, J. D. (2020), "Missing Data and Prediction: The Pattern Submodel," *Biostatistics*, 21, 236–252. [1,5]

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014), *Handbook of Missing Data Methodology*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, New York: Taylor & Francis. [1]

Moons, K. G., Donders, R. A., Stijnen, T. and Harrell, F. E. (2006),. "Using the Outcome for Imputation of Missing Predictor Values Was Preferred," *Journal of Clinical Epidemiology*, 59, 1092–1101. [1]

Müller, P., Quintana, F., and Rosner, G. L. (2011), "A Product Partition Model With Regression on Covariates," *Journal of Computational and Graphical Statistics*, 20, 260–277. [2]

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265. [4]

Page, G. L. (2021), *ppmSuite: A Collection of Models that Employ a Prior Distribution on Partitions*. R package version 0.1.7. Available at: https://CRAN.R-project.org/package=ppmSuite [5]

Page, G. L., and Quintana, F. A. (2018), "Calibrating Covariate Informed Product Partition Models," *Statistics and Computing*, 28, 1009–1031. [2]

Page, G. L., Quintana, F. A., and Rosner, G. L. (2021), "Discovering Interactions Using Covariate Informed Random Partition Models," *Annals of Applied Statistics*, 15, 1–21. [11]

Park, J.-H., and Dunson, D. B. (2010), "Bayesian Generalized Product Partition Model," *Statistica Sinica*, 20, 1203–1226. [2]

Quintana, F. A., Loschi, R. H., and Page, G. L. (2018), *Bayesian Product Partition Models*, American Cancer Society, pp. 1–15. [2]

Quintana, F. A., Müller, P., and Papoila, A. L. (2015), "Cluster-Specific Variable Selection for Product Partition Models," *Scandinavian Journal of Statistics*, 42, 1065–1077. [3]

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [5]

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley. [1]

Schouten, R. M., Lugtig, P., and Vink, G. (2018), "Generating Missing Values for Simulation Purposes: A Multivariate Amputation Procedure," *Journal of Statistical Computation and Simulation*, 88, 2909–2930. [5]

Stekhoven, D. J., and Bühlmann, P. (2012), "MissForest - Nonparametric Missing Value Imputation for Mixed-Type Data," *Bioinformatics*, 28, 112–118. [1]

Storlie, C. B., Therneau, T. M., Carter, R. E., Chia, N., Bergquist, J. R., Huddleston, J. M., and Romero-Brufau, S. (2020), "Prediction and Inference With Missing Data in Patient Alert Systems," *Journal of the American Statistical Society*, 115, 32–46. [1]

Su, Y.-S., Gelman, A., Hill, J., and Yajima, M. (2011), "Multiple Imputation With Diagnostics (mi) in R: Opening Windows Into the Black Box," *Journal of Statistical Software, Articles*, 45, 1–31. [5]

Sutton, C. D. (2005), *Classification and Regression Trees, Bagging, and Boosting*, Amsterdam, The Netherlands: American Cancer Society, Chapter 11, pp. 303–329. [2]

Tjur, T. (2009), "Coefficients of Determination in Logistic Regression Models — A New Proposal: The Coefficient of Discrimination," *The American Statistician*, 66, 366 – 372. [10]

van Buuren, S. (2012), *Flexible Imputation of Missing Data*, Boca Raton, FL: Chapman & Hall/CRC Interdisciplinary Statistics, Taylor & Francis. [1]

van Buuren, S., and Groothuis-Oudshoorn, K. (2011), "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45, 1–67. [5]

van der Heijden, G. J., Donders, A. R. T., Stijnen, T. and Moons, K. G. (2006), "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. [2]

White, I. R., and Carlin, J. B. (2010), "Bias and Efficiency of Multiple Imputation Compared With Complete-Case Analysis for Missing Covariate Values," *Statistics in Medicine*, 29, 2920–2931. [1]

Xu, D., Daniels, M. J., and Winterstein, A. G. (2016), "Sequential BART for Imputation of Missing Covariates," *Biostatistics*, 17, 589–602. [1]

Xu, Y., Müller, P., Tsimberidou, A. M., and Berry, D. (2019), "A Nonparametric Bayesian Basket Trial Design," *Biometrical Journal*, 61, 1160–1174. [3]