Constructing support vector machines with missing data

Thomas G. Stewart[*], Donglin Zeng[†], Michael C. Wu[‡]

Article Type:

Focus Article

Abstract

Support vector machine classification (SVM) is a statistical learning method which easily accommodates large numbers of predictors and can discover both linear and non-linear relationships between the predictors and outcomes. A common challenge is constructing an SVM when the training set includes observations with missing predictor values. In this paper, we identify when missing data can bias an SVM classifier. Because the missing data mechanisms which bias SVMs differ from the traditional framework of missing-at-random and missing-not-at-random, we argue for an SVM specific framework for understanding missing data. Further, we compare a number of missing data strategies for SVMs in a simulation study and real data example, and we make recommendations for SVM users based on the simulation study.

[*]Department of Biostatistics, Vanderbilt University School of Medicine
[†]Department of Biostatistics, University of North Carolina at Chapel Hill
[‡]Public Health Sciences Division, Fred Hutchinson Cancer Research Center

INTRODUCTION

Support vector machine classification (SVM) is a statistical method that offers modeling flexibility by allowing for both linear and non-linear relationships between the predictors and outcomes. SVMs have been used within a wide range of different applications, including gene expression analysis[8], fMRI analysis[13], and other biomedical computer vision tasks. SVMs represent a natural approach for building a decision rule because of the modeling flexibility. Although SVMs are an attractive option when constructing a classifier, SVMs do not easily accommodate missing covariate information.

Similar to other prediction and classification methods, in-attention to missing data when constructing an SVM can impact the accuracy and utility of the resulting classifier. This is particularly true when the likelihood that a covariate is missing is a function of the outcome class or the covariate itself. Because missing data are ubiquitous, users of SVMs should consider the impact of their chosen missing data strategy. This focus article is organized in two parts. In the first part we address the question: In what settings does missing data bias the SVM classifier? We suggest that the traditional framework for missing data commonly used in regression settings does not apply to SVMs, and we provide an alternative framework for missing data specific to SVMs.

In the second part, we provide an overview of a number of missing data strategies, some of which are general purpose strategies and others which are specific to SVMs. Then, we compare the performance of a number of the methods in a simulation study. We conclude by providing concrete suggestions for missing data strategies when constructing an SVM classifier.

SUPPORT VECTOR MACHINE BASICS

The SVM classifier[3,6,20] is a supervised learner, which is a family of classifiers built from a training set in which the outcome and covariates are known. For example, consider the task of constructing a classifier to predict the presence or absence of disease from a large array of blood chemistry measures. The training set (denoted $\mathcal{T}_n$) is n observations for which

the disease status (denoted $y_i \in \{-1, +1\}$) and chemistry measures (denoted $x_i \in \mathbb{R}^p$) are known. Specifically, let $y_i = -1$ for absence and $y_i = +1$ for presence of disease; p is the number of covariate chemistry measures.

The SVM classifier is a basis expansion model, which means it is a linear combination of scalar parameters and basis functions. Specifically, for x_o, an input p-vector of covariates, the SVM classifier is

$$f_{\mathcal{T}_n}(x_o) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \kappa_i(x_o) \tag{1}$$

where the basis functions are kernel functions of the form $\kappa_i(x_o) = \kappa(x_o, x_i)$ and the scalar parameters α_i are estimated during the fitting process. To predict the outcome for covariates x_o, one calculates

$$\hat{y}_o = \text{sign}\left[f_{\mathcal{T}_n}(x_o)\right] = \begin{cases} -1 & f_{\mathcal{T}_n}(x_o) < 0 \\ +1 & f_{\mathcal{T}_n}(x_o) \geq 0 \end{cases}$$

The set of covariates where $f_{\mathcal{T}_n}(x_o) = 0$ is called the boundary because it separates regions classified as +1 from those classified as -1. Different kernel functions lead to different flavors of SVM classifiers, and common choices of kernel include the linear and Gaussian kernels. The linear kernel generates decision rules with a linear boundary; the Gaussian kernel generates both linear and nonlinear boundaries. Other sources describe in greater detail the mathematical underpinnings of the SVM and kernel functions[16,19]. For the purposes of this article, let $\mathcal{H}$ be the set of all possible functions described in equation (1) for a specific choice of kernel function.

The SVM classifier is constructed as the solution to a convex optimization problem

$$f_{\mathcal{T}_n} = \arg\min_{f \in \mathcal{H}} \frac{1}{2}||f||^2_{\mathcal{H}} + \frac{C}{n} \sum_{i=1}^{n} \max[0, 1 - y_i f(x_i)] \tag{2}$$

which can be described heuristically as minimizing the combination of two-types of penalties: $||f||^2_{\mathcal{H}}$ is a model-complexity penalty and $\max[0, 1 - y_i f(x_i)]$ is a surrogate miss-classification penalty. The parameter C controls the relative importance of the two types of penalties. Relatively more complex models may have better in-sample miss-classification rates but may have higher out-of-sample miss-classification rates; whereas, less complex models will have similar in-sample and out-of-sample miss-classification rates. Often, the value of C is selected by a grid search coupled with cross-validated in-sample model performance.

The SVM classifier is a margin classifier, and the covariate region where $|f(x_o)| < 1$ is called the SVM margin. It is the region surrounding the boundary where observations contribute to the surrogate miss-classification penalty, even if the observations are accurately classified. Observations not properly classified always contribute to the surrogate miss-classification penalty. Note that observations that do not contribute to the surrogate miss-classification penalty could be removed from the training set without impacting the resulting SVM classifier, which is a primary feature of the method as it represents a potentially large reduction in the classifier's complexity. Observations in the training set that do impact the solution are known as the support vectors.

The SVM classifier is constructed from the training set, $\mathcal{T}_n$; the choice of kernel function (including any kernel-specific tuning parameters); and a choice of tuning parameter C. Note that constructing an SVM classifier requires estimation of $n + 1$ parameters, regardless of the number of covariates p. Thus, an SVM classifier can be constructed even when $p > n$. Routines are available in many commonly used statistical software packages (R, Matlab, Python) and stand-alone programs (libsvm) to construct an SVM classifier.

A FRAMEWORK FOR MISSING DATA AND SVMS

To understand the predictive performance of the SVM and how missing data may affect its performance, it is helpful to consider the infinite-population setting or when n is very large. Returning to equation (2), note that the surrogate miss-classification component is an empirical approximation of the infinite-population surrogate miss-classification rate,

$$\frac{1}{n}\sum_{i=1}^{n}\max[0, 1 - y_i f(x_i)] \rightarrow^{as} E_{Y\times X}\{\max[0, 1 - Yf(X)]\}.$$

Let $\dot{f}$ denote the infinite-population SVM solution. As shown in Lin[11], the SVM solution that minimizes the infinite-population surrogate miss-classification rate also minimizes the true infinite-population miss-classification rate, which is why the SVM is asymptotically an optimal classifier, which is also called a Bayes classifier. To see how missing data may affect classification, let $r_i = 1$ denote observations with at least one missing covariate and let $r_i = 0$

if no covariates are missing. The above expression can be rewritten as

$$E_{Y\times X}\{\max[0,1-Yf(X)]\} = \underbrace{E_{Y\times X|R=0}\{\max[0,1-Yf(X)]\}}_{A} P(R=0)+ \underbrace{E_{Y\times X|R=1}\{\max[0,1-Yf(X)]\}}_{B} P(R=1) \tag{3}$$

The finite-sample strategy of removing observations with $r_i = 1$ from the training set, called the complete case strategy, results in minimizing the slightly different infinite-population expression labeled A in equation (3). If A and B in the same equation are relatively similar, or if $P(R = 1)$ is small, the complete case strategy is a reasonable missing data solution. However, when the classifier that minimizes A and the classifier that minimizes $A+B$ differ, the missing data matters. To get a better handle on the specific setting where missingness is an issue, note that

$$E_{Y\times X|R=0}\{\max[0,1-Yf(X)]\}P(R=0) = E_{Y\times X}\{\max[0,1-Yf(X)]\} - E_{Y\times X}\{\max[0,1-Yf(X)]\cdot P(R=1|Y,X)\}$$

which indicates what conditions must exist for the complete case solution to differ from the full data solution. If the covariate region where $P(R = 1|Y, X)$ is appreciable falls outside the margin of $\dot{f}$, then the complete case strategy will not be biased. Conversely, missing data matters when it occurs differentially between classes within the margin or differentially between classes in regions that violate the infinite-population solution.

Missing on the boundary vs traditional missing data mechanisms

This perspective in the infinite-population setting provides some guidance on missing data in the finite-sample setting. The following general purpose ideas can be shown false in specific, pathological missing data settings; however, the ideas will be applicable to a wide variety of situations. First, because SVMs only rely on a subset of observations, they tend to be more robust to missing data than methods that rely on all the data. One may allow for more observations with missing data with an SVM than with a regression, for example. Second, because observations have more local influence on the margin with non-linear kernels, missing data is less problematic for linear SVMs than for non-linear SVMs. And third, a statement

that may be true for most methods, missing data is less problematic when the outcome class can be predicted with greater accuracy.

Demonstration of SVMs in MAR and MNAR settings

As noted above, the SVM classifier is inherently more robust to missing data because it is a margin classifier. Furthermore, the complete case strategy for missing data can generate unbiased SVM classifiers in settings where other common methods like regression might not. To illustrate this point, we performed a simulation study of the linear SVM classifier in which we considered the number of predictor variables, the degree of separation between outcome classes, the percent of observations with missing data, and the mechanism for generating missing data. For a training set of 500 observations, we generated datasets with 2, 20, and 200 predictors with a mild, banded correlation structure. The degree of separation was calibrated in terms of the optimal prediction error, also known as the Bayes risk. The degree of separation was chosen so that the best classifier could predict the outcome class with .4, .25, and .1 error. We considered four missing data mechanisms:

$$\text{logit P}(X_1 \text{ missing } | X_1, X_2, Y) = \begin{cases} \alpha + \beta X_2 & \text{MAR X} \\ \alpha + \beta X_2 * Y & \text{MAR XY} \\ \alpha + \beta X_1 & \text{MNAR X} \\ \alpha + \beta X_1 * Y & \text{MNAR XY.} \end{cases}$$

Commonly used notation in the missing data literature and introduced by Rubin[14], MAR denotes *missing at random* which indicates that the probability a covariate is missing is a function of the covariates or outcome class that is observed in the dataset. Here, MAR X denotes a mechanism where missingness in X_1 is only a function of the observed X_2 covariate; in contrast, MAR XY is a mechanism that depends on both X_2 and Y. The acronym MNAR denotes *missing not at random* which occurs when the likelihood that a covariate is missing is a function of the covariate itself and/or the other observed covariates and outcome classes. This might occur when a covariate is missing because the unobserved value exceed a threshold. We labeled the last two mechanisms MNAR X and MNAR XY because they are functions of X_1 and $X_1 * Y$, respectively. Lastly, MCAR or *missing completely*

at random is the setting when the missingness probability is not a function of either the unobserved covariate, the observed covariates, or the outcome class. This might occur, for example, when a random selection of blood chemistry measures are lost due to equipment malfunction. Note that the four missing data mechanisms simplify to MCAR mechanisms when $\beta = 0$. Thus, as β moves from zero, the missing data mechanisms transition from MCAR to mild MAR/MNAR to extreme MAR/MNAR. In our simulation study, we set the value of β to -5, -1, 0, 1, and 5. To put these values in context, note that $\beta = 2$ corresponds to an odds ratio ≈ 2.7 and $\beta = 5$ corresponds to an odds ratio ≈ 148. The value of α controls the overall proportion of observations with missing covariates. We selected α so that the missing proportion was .1, .4, and .7.

In summary, the simulation considers 3 numbers of predictors, 3 degrees of separation, 4 missing data mechanisms, 3 proportions of overall missing data, and 5 values of β to provide a spectrum of MCAR to MAR/MNAR missing data mechanisms. For each combination of settings, we generated 250 observations from each class, generated probabilities of missing X_1, then sampled which observations to mark as missing based on the probability. Furthermore, for each combination of settings, we generated an out-of-sample validation set (without missing data). We constructed a linear SVM from the training set omitting observations with missing values, then we constructed the oracle SVM, i.e., the SVM generated from the training set with no missing data. The out-of-sample prediction error of both classifiers was calculated using the validation set. We combined both measures of classifier performance by calculating the difference between the two, a summary measure we call the *above oracle prediction error* (AOPE) because it represents the increase in out-of-sample mis-classification due to missing data. Values of AOPE near zero represent minimal impact of missing data while large values represent a larger impact. The simulation for each combination of settings was repeated 100 times, from which we calculate the median AOPE and inter-quartile range. The cost parameter was selected by grid search and cross-validation.

The results from every setting are reported in the supplementary materials, and we highlight a few results here. In table 1, we show the results with Bayes risk at 0.25 and missing proportion at 0.4. The first noticeable feature is that MAR X and MNAR X had little impact on the accuracy of the complete case SVM. However, the XY missing data

Bayes risk	Missing prop	beta	Predictors	Missing type			
				MAR X	MAR XY	MNAR X	MNAR XY
				0 50	0 50	0 50	0 50
0.25	0.4	-5	2				
			20				
			200				
		-1	2				
			20				
			200				
		0	2				
			20				
			200				
		1	2				
			20				
			200				
		5	2				
			20				
			200				

Table 1: Median [IQR] of out-of-sample *above oracle prediction error* (AOPE) on the percentage-points scale

mechanisms did have some impact, particularly at the most extreme values of β. The XY mechanisms were constructed to have differential, class-specific patterns of missing data in the covariate space, which is related to the asymmetry between $\beta = 5$ and -5 in the MNAR XY case. In the first, the complete case solution performs poorly. In the later, there is only mild increase in prediction error. The asymmetry is a consequence of the distinct covariate regions where the missing data occurs. In the first, the missingness falls in the margin, near the boundary. In the later, the missingness falls away from the margin. Despite being an MNAR data and despite the very strong missingness signal ($\beta = 5$), the missing data only increased less than 5 percentage points for 2 and 20 predictors and less than 1 percentage point for 200 predictors.

Table 1 also illustrates the consequence of increasing the number of predictive covariates while holding the missing percentage constant. Heuristically, missingness in a covariate is less impactful as the relative importance of the single covariate decreases. It is a pattern

observed at the extremes of β. The opposite pattern is observed at the less extreme values of β, even in table 1 where the MCAR setting shows an ever-so-small increase as the number of predictors increases. The complete case strategy is unbiased in the MCAR setting, and the small increase in miss-classification is due to the added variability of the smaller sample size. Thus, a missingness mechanism closer to the MCAR end of the spectrum will have small but growing impact as the number of predictors increases. A mechanism closer the extreme MAR/MCAR XY end of the spectrum can (but may not) be consequential, especially for smaller numbers of predictors.

				Missing type			
Bayes risk	Predictors	beta	Missing prop	MAR X	MAR XY	MNAR X	MNAR XY
				0 50	0 50	0 50	0 50
0.1	2	-5	0.1				
			0.4				
			0.7				
		0	0.1				
			0.4				
			0.7				
	20	-5	0.1				
			0.4				
			0.7				
		0	0.1				
			0.4				
			0.7				
	200	-5	0.1				
			0.4				
			0.7				
		0	0.1				
			0.4				
			0.7				

Table 2: Median [IQR] of out-of-sample *above oracle prediction error* (AOPE) on the percentage-points scale

In table 2, we contrast the MCAR setting ($\beta = 0$) with the extreme case ($\beta = -5$) at increasing levels of overall missing proportion and increasing number of predictors. As ex-

pected, the AOPE is smaller for the MCAR settings, even less so as the number of predictors gets smaller. The missing proportion has a noticeable impact at the extreme level, more so in the MNAR settings. Perhaps what is most striking in table 2 is the rather small effect of moderate levels of the missing proportion even at extreme levels of β.

It is worth repeating that the missing proportions of 0.7 or missing mechanisms where the odds ratio of missing a covariate is greater than a 100 for a unit increase in the same or another covariate is a purposely extreme scenario. It is striking to see that even in the extreme settings, the complete case still performed well when the missing mechanism was MAR X. With that mechanism, the small decrease in accuracy is comparable to the loss of efficiency demonstrated when $\beta = 0$. Except for a few settings when the missing proportion was 0.7, the same can be said of the complete case SVM in the MNAR X setting. In the MNAR/MAR XY settings, the complete case strategy is not as robust.

This simulation study brings greater context to the issue of missing data and SVMs. It illustrates that the framework for missing data commonly used in likelihood or regression settings may not apply to SVM and other margin classifiers. As noted above, the framework for missing data with SVMs centers on the empirical miss-classification penalty which is a finite sample approximation of the infinite-population miss-classification rate. When wondering if a particular missingness mechanism might lead to a biased complete case SVM, the question is not whether or not the mechanism is MCAR, MAR, or MNAR, but rather the question is whether the mechanism biases the empirical miss-classification penalty.

In situations where maximum efficiency is important and an alternative missing data strategy is needed, the results from this simulation study suggest that users should be conscientious about differential missing data mechanisms among outcome classes. If data processing steps lead to outcome class-specific missing data issues, users should implement a strategy that incorporates outcome class.

Lastly, we acknowledge that missing data mechanisms are not known and cannot be known in practice. The user is faced with trade-offs of efficiency, unbiasedness, and ease-of-implementation all while not knowing the degree of efficiency loss or potential bias imposed by the missing data. In light of these challenges, it is still true, however glib, that the best strategy for missing data is not to have any.

A COMPARISON OF GENERAL PURPOSE AND SVM-SPECIFIC MISSING DATA STRATEGIES

Because users of SVM will want to apply the method in settings with missing data, we transition to a discussion of specific missing data strategies. Note that the different missing data methods are strategies for incorporating information from the incomplete observations in addition to the information already provided by the complete observations. Let x_i^o denote the subset of covariates that are observed, and let x_i^m denote the subset of missing covariates for subject i. If the missing data mechanism is such that the probability of missingness is only a function of the observed covariates, $P(R|Y,X) = P(R|Y,X^o)$, then incorporating information from the observations with missing covariates can minimize the impact of the missing data.

Imputation, single and multiple

A common strategy for missing data with SVMs and many analysis methods is imputation, in which missing covariates are filled-in with reasonable values. In its simplest form, the replacement value could be the covariate mean (mean imputation) or simply a random draw from the set of observed covariate values. In more sophisticated settings, the replacement value is drawn from different estimates of $P(X^m|Y = y_i, X^0 = x_i^o)$. For example, one might estimate the conditional distribution under the assumption that $P(X^m|Y = y_i, X^0 = x_i^o)$ is normally distributed and then draw a replacement value from it. A non-parametric approach is K-nearest neighbor where the conditional distribution is approximated by the distribution of X^m in K complete cases with similar values of x_i^o and y_i. There are many algorithms for selecting the replacement value for the missing covariate. Once the missing values are replaced with reasonable surrogates, one constructs the SVM with the newly completed training set.

A second level of sophistication is to repeat the imputation process multiple times[15] with an algorithm that replaces missing values with a draw from a conditional distribution instead of its mean or median. Then, one constructs an SVM classifier with each filled-in training

set, and then combines all the results into a single classifier. The advantage of repeating the imputation process multiple times is that it provides an indication of the increased variation due to the missing covariate data. If the set of SVM classifiers result in similar boundaries, then the missing data had little impact. Disparate boundaries would indicate greater impact.

As a general purpose solution, imputation does not require specialized software. The strategy implements a pre-processing step that allows all down-stream calculations to proceed without alteration. Because it is a straight-forward solution for missing data, it has been widely studied and widely implemented in practice. For example, Acuña and Rodriguez[1] encouraged the use of KNN imputation and Garcia and Hruschka[9] suggested Naive Bayes type imputation. In Farhangfar et al.[7], the authors consider the Gaussian SVM with hot deck imputation, Naive Bayes, mean imputation, and regression-based imputation. Other imputation methods, popular in parametric analyses, include Sequences of Regression Models (SRM)[12] and Multivariate Imputation by Chained Equations (MICE)[4].

Pattern mixture models

In contrast to the imputation solution described above which generates a single solution that makes predictions from a set of fully known covariates, some missing data strategies anticipate that the classifier will be implemented with data that also has missing covariates. Further, if the missing data pattern—the set of covariates that are missing—is informative for predicting the outcome, then models which condition on the missing data pattern can be effective, especially when prediction or classification is the primary objective instead of statistical inference. This family of missing data strategies are known as pattern mixture models because the infinite-population is conceptualized as a mixture of sub-populations characterized by a unique pattern of missing covariates.

The optimal implementation of this approach is an area of current research. The primary advantage of this strategy is its straight-forward application to future data with missing covariates. It is especially useful in settings where data are extracted from distinct sources and the missing data patterns are reflective of source-specific processing which are likely to continue in the future. Furthermore, it does not require specialized software to implement. The primary criticism of this strategy is the potentially large number of missing data patterns

and the small number of observations within each pattern.

SVM-specific strategies

There are a number of SVM-specific strategies for missing data that take advantage of computational details (purposely omitted from this manuscript) or the fact that the SVM is a margin classifier or other properties of the SVM. See the review by García-Laencina et al.[10] for a review of several strategies. Some of the solutions are computationally intensive, constrained by sample size, restricted to specific kernel functions, or require specialized computational routines. These barriers to implementation likely contribute to the relatively fewer examples of these methods in practice compared to imputation.

The probability constraint techniques of Shivaswamy et al.[17] take advantage of the constrained, convex optimization problem that defines the SVM. They note that the covariates only enter into the problem through the constraint involving the surrogate miss-classification. The method replaces the constraint with one which limits the conditional probability of (surrogate) miss-classification. While an intuitive solution for missing data, the strategy suffers from the barriers listed earlier.

Chechik et al[5] proposed a missing data method built on the geometric perspective of SVMs; the basic idea is to define the margin in terms of non-missing covariates. The geometric max margin method redefines the margin uniquely for each observation. This geometric solution performs reasonably well when the missingness is unrelated to the covariates or the outcome. In that setting, there may be some advantages in computation time over imputation. However, computation requires non-convex optimization when the two groups are not separable, which is a very common setting.

Smola et al.[18] provides a principled missing data method based on a connection between SVM and exponential family distributions. The authors note a connection between the sufficient statistics of $P(Y|X)$ and the SVM solution with a specific kernel. The fitted SVM is also the solution which maximizes the likelihood ratio. The method proposes that for missing data, one replace the loss function with one constructed with sufficient statistics from the distribution $P(Y|X^o)$. The solution requires a number of "thorny" computational issues, and it is limited in workable size.

One SVM-specific strategy with strong connections to multiple imputation was proposed by Anderson and Gupta[2]. Because covariate data enters the SVM objective function via the kernel matrix, the $n \times n$ matrix of $\kappa(x_i, x_j)$, the authors propose replacing the kernel matrix with its expectation under user specified assumptions of the covariate distribution. Note that the multiple imputation estimate of the kernel matrix generates an expected kernel under a set of reasonable assumptions and is easily calculated. Rather than using the multiple imputation expected kernel, the authors instead opt for closed form solutions under the assumption that the covariate data is normally distributed.

Simulation study of missing data strategies

In this section, we compare the performance of some of the SVM strategies described above: Complete Case (CC), Mean Imputation (Imp), Multiple Imputation (MI), K-Nearest Neighbor Imputation (KNN), and Probability constraint (PC). As before, we also built the oracle classifier, the SVM classifier built with no missing data. Because so many imputation methods exist, we selected the method most likely to succeed with the data generated in the simulation. Specifically, we chose imputation based on the assumption that the data are normally distributed. (Specifics of the data generation are described below.) Thus, we address the question of imputation models by attempting to select one that should work as well as any other.

The PC method[17], as noted earlier, takes advantage of the constrained optimization problem which characterizes computation of the SVM solution. Without missing data, the constrained optimization problem recasts equation (2) as

$$f_{svm} = \arg\min_{f \in \mathcal{H}} \; ||f||^2_{\mathcal{H}} + \frac{C}{n} \sum_{i}^{n} \xi_i$$

$$\text{Such that } 1 - y_i f(x_i) \leq \xi_i \text{ and } \xi_i \geq 0 \text{ for } \; i = 1, \ldots, n$$

The PC method replaces the constraints involving missing covariates with

$$P(1 - y_i f(x_i) \leq \xi_i | x_i^o, y_i) \geq 1 - v_i \quad v_i \in (0, 1]; \quad i = 1, \ldots, n.$$

The authors show that Chebyshev inequalities applied to the constraint allow the optimization problem to be recast as a second order cone programming problem. We construct linear

classification rules with this method.

Our simulation study is organized as a factorial experiment in which we consider (a) linear and non-linear boundaries, (b) two missing data models, and (c) linear and Gaussian kernel classifiers. For each combination, we generated 100 datasets with $p = 20$ covariates drawn from the normal distribution with mean centered at 10.

In scenarios with a linear boundary, the outcome was generated from the linear log odds model

$$\text{logit } P(Y = +1|X) = \gamma \mathbb{1}_p^t (X - 10 \cdot \mathbb{1}_p). \tag{4}$$

In scenarios with a non-linear boundary, the outcome was generated from the model

$$\text{logit } P(Y = +1|X) = \gamma \left[\delta + X_p - \sum_{i=1}^{p-1} (X_i - 10)^2 \right]. \tag{5}$$

In both the linear and non-linear boundary models, the parameter γ calibrates signal strength, and is set so that the oracle classifier achieves a 15% classification error rate. The parameter δ is set to achieve balanced group proportions.

We generate missing values according to two missing data models, MAR X and MAR XY, constructed similar to the previous simulation. In contrast to the earlier simulation in which missing data only occur in the first predictor, in this simulation, missing values are allowed in all but the last predictor. Further, missingness in X_k is a function of the value X_{k+1}. In both missing data scenarios, we consider a wide range of values for β in order to judge classification performance in varying degrees of missing data model signal strength. Specifically, β = -6, -2, 0, 2, and 6. Note that negative and positive values of β represent considerably different missing data patterns, and methods for missing data can have non-symmetric performance along the range of βs. The parameter α is calibrated so that the percentage of observations with missing covariates is 70% because we want a scenario where complete case shows degraded performance. As before, the specific value of α depends of the value of β.

As in the previous simulation, AOPE represents the loss in accuracy due to missing data in the covariates. Table 3 reports the AOPE for each competing method in each scenario and is based on a validation data set of 100,000 observations.

Boundary	Miss type		β				
			-6	-2	0	2	6
Linear	MAR X	CC	6.1 [4.7, 7.8]	5.6 [4.7, 7.2]	4.7 [3.4, 6.5]	5.9 [3.9, 7.5]	6.5 [5.0, 8.1]
		Imp	0.3 [-0.1, 0.8]	0.3 [-0.2, 0.7]	0.2 [-0.1, 0.7]	0.2 [-0.2, 0.7]	0.2 [-0.2, 0.8]
		KNN	0.5 [-0.0, 0.9]	0.4 [-0.1, 0.9]	0.5 [-0.1, 1.0]	0.4 [-0.1, 0.8]	0.5 [-0.2, 0.9]
		MI	0.2 [-0.1, 0.7]	0.3 [-0.2, 0.6]	0.1 [-0.4, 0.5]	0.2 [-0.2, 0.7]	0.2 [-0.2, 0.5]
		PC	0.3 [-0.2, 0.9]	0.4 [-0.2, 0.8]	0.3 [-0.1, 0.8]	0.3 [-0.1, 1.0]	0.5 [-0.0, 0.9]
	MAR XY	CC	6.3 [5.3, 7.9]	5.8 [4.0, 7.1]	4.7 [3.4, 6.5]	8.1 [6.4, 11.6]	11.3 [9.5, 13.5]
		Imp	0.3 [-0.1, 0.7]	0.2 [-0.2, 0.7]	0.2 [-0.1, 0.7]	0.3 [-0.2, 0.7]	0.2 [-0.2, 0.6]
		KNN	0.5 [-0.1, 0.9]	0.5 [0.0, 1.1]	0.5 [-0.1, 1.0]	0.2 [-0.1, 0.7]	0.3 [-0.1, 0.8]
		MI	0.3 [-0.2, 0.6]	0.3 [-0.1, 0.7]	0.1 [-0.4, 0.5]	0.1 [-0.3, 0.5]	0.2 [-0.2, 0.7]
		PC	0.5 [0.1, 1.3]	0.5 [-0.2, 1.1]	0.3 [-0.1, 0.8]	0.2 [-0.3, 0.6]	0.3 [-0.3, 0.7]
Nonlinear	MAR X	CC	8.2 [2.6, 17.3]	5.3 [2.3, 16.8]	2.6 [1.6, 4.1]	4.0 [2.3, 9.9]	5.3 [2.6, 12.0]
		Imp	0.6 [0.0, 1.2]	0.6 [-0.0, 1.1]	0.4 [-0.0, 0.9]	0.7 [-0.0, 1.3]	0.7 [0.1, 1.3]
		KNN	0.2 [-0.3, 0.6]	0.3 [-0.2, 0.7]	0.2 [-0.2, 0.6]	0.3 [-0.1, 0.9]	0.3 [-0.2, 0.7]
		MI	0.3 [-0.1, 1.1]	0.3 [-0.1, 0.7]	0.3 [-0.2, 0.9]	0.2 [-0.2, 0.8]	0.5 [0.0, 1.1]
	MAR XY	CC	22.2 [17.8, 25.0]	17.3 [12.6, 22.4]	2.6 [1.6, 4.1]	16.5 [12.7, 19.8]	20.0 [16.8, 21.5]
		Imp	0.7 [0.1, 1.2]	0.6 [-0.0, 1.2]	0.4 [-0.0, 0.9]	0.6 [-0.2, 1.1]	1.0 [0.4, 1.7]
		KNN	0.5 [-0.0, 1.1]	0.4 [-0.1, 1.1]	0.2 [-0.2, 0.6]	0.3 [-0.3, 0.8]	0.4 [0.0, 1.2]
		MI	0.4 [-0.0, 1.1]	0.3 [-0.4, 0.9]	0.3 [-0.2, 0.9]	0.4 [-0.1, 1.0]	0.4 [-0.4, 0.9]

Table 3: Median [IQR] of out-of-sample *above oracle prediction error* (AOPE) when the number of predictors is 20 (percentage-points scale)

The most striking result of this simulation is that all the missing data strategies performed relatively well except for complete case. The setting was selected so that complete case would perform poorly. Reported in the percentage-point scale, the largest AOPE was single, mean imputation in the non-linear setting with the rather extreme $\beta = 6$. The median added error was 1.0 percentage points. Despite the large number of observations with missing data and the MAR missing data mechanisms, none of the imputation or PC missing data strategies showed consistently better performance. These results suggest that when large numbers of observations have missing data, implementing imputation or PC can improve the SVM performance over complete case.

Application to HCV-TARGET data

We apply the missing data strategies to a subset of patients from the HCV-TARGET database. HCV-TARGET is a consortium of North American academic and community medical centers performing a longitudinal observational study of patients undergoing treat-

ment for hepatitis C. Specifically, we consider previously treated, non-cirrhotic, female patients treated with Telaprevir. One reason for choosing this subset is that its cure rate is lower than some of its counterparts.

Cure, the outcome of interest, is defined as undetected hepatitis C virus 12 weeks after ending treatment. The training set (N=112) are those patients treated prior to 1 January 2012. The validation set (N=38) are those treated after. The predictors of interest include age and five blood assays at baseline: total bilirubin (BILI), creatinine (CRE), hemoglobin (HGB), hepatitis C viral load (LOGHCV), and absolute neutrophil count (ANC). We selected these predictors from the set of baseline measures in consultation with members of the HCV-TARGET team. In the training set, 37% of patients have incomplete data. Of those with incomplete data, 71% are missing one predictor, 20% are missing two predictors, and the remaining 9% are missing three predictors.

We applied complete case, mean imputation, multiple imputation, k-nearest neighbor, and probability constraint methods to the construction of decision rules with linear and Gaussian kernels. The out-of-sample prediction error for each classification rule is reported in Table 4. The most striking feature of this example is the dominance of the non-linear kernel within each strategy, which suggests non-linear associations between the predictors and the outcome. Each kernel method performed better with the Gaussian kernel than the linear kernel. The multiple imputation method was best (23.7%).

While SVM decision rules are difficult to interpret in terms of a single predictor or even two predictors, inspection of partial-effect plots did demonstrate potentially important non-linear boundaries between outcome groups in the predictor variables. We observed, for example, that a number of relationships indicate one group near the mean and the other group near both tails. AGE is one such predictor with this trend. Others like BILI tend to have a more linear-type relationship, where higher values of BILI do not favor cure.

CONCLUSIONS

As a margin classifier, the SVM has a natural robustness to missing data when compared to other prediction and classification methods. Only missingness among the support vectors and

Method	Linear Kernel	Gaussian Kernel
Complete Case	44.74	31.58
KNN	44.74	31.58
Mean Imputation	47.37	28.95
Multiple Imputation	39.47	23.68
Probability Constraint	31.58	

Table 4: Prediction error of various missing data strategies when applied to HCV-TARGET Data

observations that would be miss-classified have the potential to impact the SVM solution. If the missingness is differential between outcome classes, then a complete case solution may be biased. Unfortunately, the missing data mechanism is not discoverable from the data, so we recommend implementing a missing data strategy when the percentage of observations with missing covariates is large. The exact cut-off between large and small percentages, of course, is ambiguous, but the simulations that report increasing levels of missingness show little advantage of a missing data strategy over the complete case solution when the percentage is less than 15% and even 30%.

While there are SVM-specific missing data strategies, barriers to implementation have likely hindered their wide spread use, whereas; general purpose solutions like imputation are popular because they require a simple pre-processing step and tend to be effective. In light of this, implementation continues be a key consideration for future SVM-specific missing data methods, which is why our recommendation to developers of new SVM missing data strategies is to provide ready-to-implement software. For users of SVM, we recommend imputation because it is easy to implement and many statistical programming languages already have developed routines and packages that will provide single and multiple imputation algorithms.

Acknowledgements

The authors would like to thank Dr. Michael Fried and HCV-TARGET for access to the data analyzed above.

References

1. Acuña E, Rodriguez C. The Treatment of Missing Values and its Effect on Classifier Accuracy. Classification, Clustering, and Data Mining Applications 2004;(1995):639–647.

2. Anderson HS, Gupta MR. Expected kernel for missing features in support vector machines. 2011 IEEE Statistical Signal Processing Workshop SSP 2011;p. 285–288.

3. Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory 1992;p. 144–152. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.3818.

4. van Buuren S, Oudshoorn K. Flexible Multivariate Imputation by MICE; 1999.

5. Chechik G, Heitz G, Elidan G. Max-margin classification of incomplete data. Advances in Neural ... 2007;http://ai.stanford.edu/~galel/papers/ElidanMaxMarginNIPS.pdf.

6. Cortes C, Vapnik S. Support Vector Networks. Machine Learning 1995;20(3):273–297. http://link.springer.com/10.1007/BF00994018.

7. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition 2008 12;41(12):3692–3705. http://linkinghub.elsevier.com/retrieve/pii/S003132030800201X.

8. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16(10):906–914. http://bioinformatics.oxfordjournals.org/content/16/10/906%5Cnhttp://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/16.10.906.

9. Garcia AJT, Hruschka ER. Naïve Bayes as an imputation tool for classification problems. Proceedings - HIS 2005: Fifth International Conference on Hybrid Intelligent Systems 2005;2005:497–499.

10. García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR. Pattern classification with missing data: a review. Journal of Comput & Applic 2010;19:263–282.

11. Lin Y. Support Vector Machines and the Bayes Rule. Data Mining and Knowledge Discovery 2002;6(3):259–275.

12. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 2001;27(1):85–95.

13. Rehme AK, Volz LJ, Feis DL, Bomilcar-Focke I, Liebig T, Eickhoff SB, et al. Identifying neuroimaging markers of motor disability in acute stroke by machine learning techniques. Cerebral Cortex 2015;25(9):3046–3056.

14. Rubin DB. Inference and missing data. Biometrika 1976 12;63(3):581–592. http://biomet.oxfordjournals.org.libproxy.lib.unc.edu/content/63/3/581.

15. Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc.; 1987.

16. Scholkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press; 2002. http://dl.acm.org/citation.cfm?id=559923.

17. Shivaswamy PK, Bhattacharyya C, Smola AJ. Second Order Cone Programming Approaches for Handling Missing and Uncertain Data. Journal of Machine Learning Research 2006;7(1):1283–1314. http://portal.acm.org/citation.cfm?id=1248547.1248594.

18. Smola AJ, Vishwanathan S, Hoffman T. Kernel methods for missing variables. In: Proceedings of the tenth international workshop on artificial intelligence and statistics; 2005. p. 325–332.

19. Steinwart I, Christmann A. Support Vector Machines. Information Science and Statistics, New York, NY: Springer New York; 2008. http://link.springer.com/10.1007/978-0-387-77242-4.

20. Vapnik VN. The Nature of Statistical Learning Theory. New York, NY, USA: Springer-Verlag New York, Inc.; 1995.

Supplementary Material

			Missing type							
Bayes risk	Missing prop	β	MAR X		MAR XY		MNAR X		MNAR XY	
0.1	0.1	-5	0.0	[-0.1, 0.1]	1.0	[0.7, 1.5]	0.0	[-0.1, 0.0]	0.0	[-0.1, 0.0]
		-1	0.0	[-0.0, 0.1]	0.1	[-0.0, 0.3]	0.0	[-0.1, 0.1]	0.0	[-0.0, 0.1]
		0	0.0	[-0.1, 0.1]	0.0	[-0.0, 0.1]	0.0	[-0.0, 0.1]	0.0	[-0.0, 0.1]
		1	-0.0	[-0.1, 0.0]	0.0	[-0.0, 0.1]	0.0	[-0.1, 0.0]	0.2	[0.0, 0.4]
		5	0.0	[-0.1, 0.1]	0.0	[-0.0, 0.1]	-0.0	[-0.1, 0.1]	1.2	[0.8, 1.6]
	0.4	-5	0.1	[0.0, 0.4]	10.2	[8.6, 12.0]	0.1	[-0.0, 0.3]	3.3	[1.8, 4.4]
		-1	0.1	[-0.0, 0.2]	1.4	[1.0, 2.0]	0.1	[-0.0, 0.2]	0.8	[0.4, 1.3]
		0	0.1	[-0.0, 0.1]	0.0	[-0.1, 0.2]	0.0	[-0.0, 0.2]	0.1	[-0.0, 0.2]
		1	0.0	[-0.1, 0.2]	1.1	[0.6, 1.6]	0.1	[0.0, 0.2]	2.0	[1.5, 2.9]
		5	0.1	[-0.0, 0.4]	6.5	[5.3, 8.3]	0.1	[0.0, 0.4]	9.9	[9.2, 10.7]
	0.7	-5	0.6	[0.2, 1.3]	13.6	[11.6, 14.8]	39.7	[39.3, 39.9]	39.9	[39.7, 40.2]
		-1	0.3	[-0.0, 0.7]	3.4	[2.4, 4.6]	0.6	[0.1, 1.3]	7.8	[4.7, 11.1]
		0	0.2	[-0.0, 0.4]	0.2	[-0.0, 0.5]	0.2	[0.0, 0.4]	0.1	[-0.0, 0.3]
		1	0.3	[0.1, 0.6]	4.5	[2.8, 6.4]	0.3	[0.1, 0.7]	4.5	[2.7, 6.1]
		5	0.7	[0.2, 2.3]	37.0	[34.4, 40.2]	39.7	[8.2, 40.0]	19.7	[18.3, 22.1]
0.25	0.1	-5	0.0	[-0.0, 0.1]	1.3	[0.9, 1.8]	0.0	[-0.0, 0.1]	0.1	[-0.0, 0.2]
		-1	0.0	[-0.1, 0.1]	0.3	[0.1, 0.6]	0.0	[-0.1, 0.1]	0.0	[-0.0, 0.2]
		0	0.0	[-0.1, 0.1]	0.0	[-0.1, 0.1]	0.0	[-0.0, 0.1]	0.0	[-0.0, 0.1]
		1	0.0	[-0.0, 0.1]	0.2	[-0.0, 0.4]	0.0	[-0.1, 0.1]	0.2	[0.0, 0.4]
		5	0.0	[-0.1, 0.2]	0.6	[0.2, 1.0]	0.0	[-0.1, 0.1]	1.0	[0.6, 1.2]
	0.4	-5	0.1	[-0.0, 0.4]	8.8	[7.7, 9.7]	0.2	[-0.0, 0.8]	23.9	[19.4, 24.8]
		-1	0.1	[-0.0, 0.2]	2.3	[1.7, 2.8]	0.1	[-0.0, 0.4]	2.9	[2.0, 5.0]
		0	0.1	[-0.0, 0.2]	0.0	[-0.0, 0.2]	0.1	[-0.1, 0.2]	0.1	[-0.0, 0.2]
		1	0.1	[-0.1, 0.4]	3.6	[2.5, 4.9]	0.1	[-0.0, 0.3]	1.5	[1.1, 1.9]
		5	0.2	[0.0, 0.4]	17.2	[15.6, 18.9]	0.3	[0.0, 0.9]	4.7	[4.3, 5.2]
	0.7	-5	1.0	[0.3, 2.4]	10.9	[10.1, 11.9]	24.7	[24.1, 25.0]	38.6	[37.2, 39.7]
		-1	0.4	[0.1, 1.0]	4.7	[3.5, 6.0]	0.4	[0.1, 0.9]	21.2	[16.1, 24.8]
		0	0.4	[0.1, 0.8]	0.3	[0.1, 0.7]	0.2	[-0.0, 0.5]	0.3	[0.0, 0.8]
		1	0.4	[0.1, 0.9]	11.3	[9.2, 13.9]	0.5	[0.2, 1.1]	2.8	[2.2, 3.6]
		5	0.8	[0.1, 2.1]	29.3	[27.6, 30.7]	24.6	[24.2, 25.0]	7.6	[6.8, 8.7]
0.4	0.1	-5	0.0	[-0.0, 0.2]	1.6	[1.0, 2.1]	0.1	[-0.0, 0.3]	4.0	[1.7, 7.6]
		-1	0.0	[-0.1, 0.1]	0.5	[0.1, 0.9]	0.1	[-0.1, 0.2]	0.5	[-0.0, 1.0]
		0	0.1	[-0.1, 0.2]	0.0	[-0.1, 0.1]	0.1	[-0.1, 0.2]	0.0	[-0.1, 0.2]
		1	0.1	[-0.1, 0.2]	0.7	[0.1, 1.5]	0.0	[-0.1, 0.1]	0.2	[-0.0, 0.4]
		5	0.0	[-0.1, 0.3]	3.2	[2.4, 4.4]	0.1	[-0.1, 0.2]	0.5	[0.2, 0.7]
	0.4	-5	0.4	[0.0, 1.1]	4.2	[3.8, 4.8]	1.8	[0.2, 9.2]	15.8	[15.0, 16.5]
		-1	0.3	[0.0, 0.7]	2.5	[1.8, 3.0]	0.3	[-0.0, 0.9]	10.7	[9.3, 12.6]
		0	0.2	[-0.0, 0.4]	0.2	[-0.0, 0.5]	0.3	[-0.1, 0.6]	0.2	[-0.1, 0.6]
		1	0.3	[-0.1, 0.7]	6.7	[5.8, 7.8]	0.2	[-0.0, 0.8]	0.8	[0.4, 1.2]
		5	0.4	[0.0, 1.0]	11.1	[10.5, 12.0]	2.5	[0.3, 9.6]	1.4	[1.0, 1.8]
	0.7	-5	2.9	[0.5, 9.1]	5.2	[4.8, 5.8]	9.4	[8.9, 9.8]	16.5	[15.9, 17.3]
		-1	1.0	[0.3, 2.7]	3.3	[2.8, 4.2]	2.2	[0.6, 9.4]	14.5	[13.8, 15.9]
		0	0.8	[0.2, 2.2]	0.6	[0.0, 1.9]	0.6	[0.1, 1.5]	0.7	[0.1, 1.5]
		1	1.0	[0.3, 8.5]	9.9	[8.6, 10.9]	1.9	[0.4, 9.1]	1.3	[0.7, 1.6]
		5	3.6	[0.8, 9.0]	12.5	[11.7, 13.1]	9.3	[8.5, 9.7]	1.8	[1.5, 2.3]

Table 5: Median [IQR] of out-of-sample *above oracle prediction error* (AOPE) when the number of predictors is 2 (percentage-points scale)

			Missing type							
Bayes risk	Missing prop	β	MAR X		MAR XY		MNAR X		MNAR XY	
0.1	0.1	-5	0.1	[-0.2, 0.3]	0.9	[0.4, 1.2]	0.1	[-0.1, 0.3]	0.1	[-0.1, 0.2]
		-1	0.1	[-0.0, 0.3]	0.2	[-0.0, 0.5]	0.0	[-0.1, 0.2]	0.1	[-0.1, 0.2]
		0	0.1	[-0.1, 0.3]	0.1	[-0.1, 0.2]	0.1	[-0.1, 0.3]	0.1	[-0.1, 0.3]
		1	0.1	[-0.2, 0.3]	0.1	[-0.1, 0.3]	0.1	[-0.1, 0.3]	0.3	[0.0, 0.6]
		5	0.1	[-0.1, 0.3]	0.2	[0.0, 0.5]	0.0	[-0.1, 0.2]	1.1	[0.7, 1.6]
	0.4	-5	0.8	[0.3, 1.5]	8.6	[7.1, 10.5]	0.8	[0.5, 1.4]	4.2	[2.8, 6.0]
		-1	0.6	[0.2, 1.0]	1.6	[1.1, 2.2]	0.8	[0.4, 1.1]	1.4	[1.1, 2.0]
		0	0.6	[0.3, 1.0]	0.6	[0.3, 1.1]	0.7	[0.3, 1.0]	0.5	[0.3, 1.1]
		1	0.8	[0.3, 1.1]	1.6	[1.1, 2.1]	0.7	[0.4, 1.1]	2.2	[1.5, 3.0]
		5	0.7	[0.3, 1.2]	7.9	[6.3, 9.2]	0.9	[0.5, 1.5]	8.6	[7.2, 9.7]
	0.7	-5	3.3	[2.2, 4.8]	12.4	[11.0, 14.3]	10.3	[4.7, 38.1]	39.7	[38.5, 41.2]
		-1	2.2	[1.4, 3.0]	4.3	[3.4, 6.0]	3.2	[2.0, 4.8]	9.4	[6.8, 12.7]
		0	2.0	[1.6, 3.1]	1.9	[1.3, 2.8]	2.2	[1.4, 2.8]	2.2	[1.4, 3.2]
		1	2.1	[1.3, 3.0]	6.9	[5.3, 8.7]	2.8	[1.8, 4.1]	6.2	[4.3, 7.8]
		5	3.6	[2.1, 5.2]	32.5	[30.5, 34.8]	11.0	[4.9, 37.4]	18.2	[15.9, 19.8]
0.25	0.1	-5	0.1	[-0.1, 0.3]	1.1	[0.7, 1.8]	0.1	[-0.1, 0.3]	0.4	[0.1, 0.7]
		-1	0.2	[-0.1, 0.4]	0.1	[-0.1, 0.5]	0.1	[-0.1, 0.4]	0.4	[0.1, 0.8]
		0	0.1	[-0.1, 0.3]	0.1	[-0.1, 0.3]	0.2	[0.0, 0.4]	0.1	[-0.1, 0.3]
		1	0.2	[-0.1, 0.4]	0.3	[0.1, 0.6]	0.1	[-0.0, 0.4]	0.1	[-0.2, 0.4]
		5	0.2	[-0.0, 0.4]	0.8	[0.5, 1.2]	0.1	[-0.1, 0.3]	0.5	[0.3, 0.9]
	0.4	-5	1.0	[0.5, 1.5]	7.8	[6.7, 8.9]	1.0	[0.5, 2.0]	18.4	[16.6, 19.9]
		-1	1.0	[0.5, 1.4]	2.2	[1.7, 2.9]	0.9	[0.5, 1.4]	4.9	[4.0, 6.3]
		0	0.8	[0.3, 1.3]	1.0	[0.5, 1.4]	0.9	[0.5, 1.6]	0.9	[0.5, 1.4]
		1	0.9	[0.3, 1.4]	4.7	[4.0, 5.6]	0.8	[0.3, 1.4]	1.7	[1.3, 2.2]
		5	1.2	[0.6, 1.9]	15.8	[14.7, 17.3]	1.3	[0.7, 2.1]	4.6	[4.1, 5.2]
	0.7	-5	3.9	[2.2, 5.7]	10.8	[9.7, 12.0]	6.5	[3.8, 18.9]	34.3	[32.7, 35.2]
		-1	3.2	[2.1, 4.6]	5.4	[4.2, 6.7]	3.1	[2.2, 4.5]	17.0	[13.5, 18.6]
		0	2.9	[1.8, 4.3]	3.0	[2.1, 3.9]	3.0	[2.2, 4.0]	2.8	[2.0, 3.8]
		1	3.1	[2.0, 4.3]	12.1	[9.8, 13.5]	3.1	[2.0, 4.6]	3.6	[2.9, 5.1]
		5	3.9	[2.7, 5.3]	25.4	[23.8, 27.0]	6.3	[2.8, 20.7]	6.7	[6.0, 7.6]
0.4	0.1	-5	0.3	[-0.0, 0.6]	0.4	[-0.0, 1.0]	0.2	[-0.0, 0.5]	3.6	[2.5, 4.6]
		-1	0.2	[-0.1, 0.5]	-0.0	[-0.4, 0.4]	0.2	[-0.1, 0.5]	1.2	[0.6, 1.8]
		0	0.1	[-0.1, 0.5]	0.3	[-0.0, 0.5]	0.1	[-0.2, 0.4]	0.2	[-0.1, 0.6]
		1	0.1	[-0.2, 0.4]	0.8	[0.5, 1.2]	0.2	[-0.2, 0.5]	-0.4	[-0.6, 0.0]
		5	0.2	[-0.1, 0.6]	2.2	[1.8, 2.9]	0.2	[-0.1, 0.5]	-0.4	[-0.9, 0.0]
	0.4	-5	1.2	[0.4, 1.7]	3.2	[2.4, 3.8]	1.2	[0.3, 2.1]	12.4	[11.7, 13.1]
		-1	0.8	[0.2, 1.6]	1.5	[0.7, 2.1]	1.1	[0.3, 1.8]	7.6	[6.4, 8.6]
		0	1.1	[0.4, 1.5]	1.1	[0.5, 1.7]	0.8	[0.4, 1.6]	1.0	[0.4, 1.5]
		1	1.1	[0.3, 2.0]	4.8	[4.0, 5.6]	1.1	[0.5, 1.9]	0.0	[-0.6, 0.6]
		5	1.0	[0.5, 1.7]	8.6	[7.8, 9.2]	1.2	[0.5, 2.2]	0.3	[-0.3, 1.0]
	0.7	-5	2.9	[2.1, 4.0]	3.6	[2.9, 4.4]	4.4	[2.1, 6.4]	13.4	[12.6, 14.5]
		-1	2.3	[1.4, 3.6]	2.6	[1.9, 3.3]	2.9	[1.6, 4.0]	10.5	[9.4, 11.6]
		0	2.7	[1.3, 3.9]	2.4	[1.3, 3.5]	2.4	[1.5, 3.4]	2.5	[1.6, 3.8]
		1	2.5	[1.5, 3.6]	7.1	[6.4, 8.3]	2.2	[1.3, 3.3]	0.7	[-0.0, 1.4]
		5	3.0	[2.1, 4.0]	9.6	[8.7, 10.3]	3.1	[2.0, 5.4]	0.8	[0.3, 1.4]

Table 6: Median [IQR] of out-of-sample *above oracle prediction error* (AOPE) when the number of predictors is 20 (percentage-points scale)

			Missing type							
Bayes risk	Missing prop	β	MAR X		MAR XY		MNAR X		MNAR XY	
0.1	0.1	-5	0.4	[-0.3, 1.0]	0.6	[0.0, 1.6]	0.2	[-0.3, 0.8]	0.3	[-0.2, 0.8]
		-1	0.6	[0.3, 1.3]	0.7	[-0.1, 1.3]	0.4	[-0.1, 1.1]	0.5	[-0.0, 1.0]
		0	0.5	[-0.1, 1.2]	0.7	[-0.1, 1.1]	0.8	[0.1, 1.4]	0.6	[-0.0, 1.1]
		1	0.8	[0.1, 1.6]	0.6	[0.1, 1.3]	0.5	[0.0, 1.1]	0.5	[-0.2, 1.1]
		5	0.6	[-0.1, 1.1]	1.0	[0.3, 1.4]	0.4	[-0.2, 1.0]	0.0	[-0.7, 0.8]
	0.4	-5	3.9	[2.7, 4.9]	4.9	[4.0, 6.1]	4.1	[2.9, 5.2]	8.3	[6.8, 9.6]
		-1	3.5	[2.3, 4.4]	3.4	[2.3, 4.5]	3.4	[2.5, 4.5]	4.8	[3.8, 6.0]
		0	3.5	[2.8, 4.3]	3.6	[2.3, 4.4]	3.1	[2.2, 4.2]	3.9	[2.6, 4.9]
		1	3.6	[2.4, 4.6]	5.3	[4.2, 6.1]	3.8	[2.5, 4.6]	3.7	[2.4, 4.6]
		5	3.7	[2.5, 4.8]	10.4	[8.8, 11.4]	4.1	[3.0, 5.4]	6.0	[4.8, 7.0]
	0.7	-5	10.3	[8.2, 12.4]	12.4	[10.8, 13.5]	20.2	[17.8, 25.3]	32.0	[31.4, 32.9]
		-1	8.4	[7.4, 9.4]	8.5	[7.2, 9.8]	10.0	[8.3, 11.5]	15.3	[13.7, 17.4]
		0	8.2	[6.8, 9.7]	8.6	[7.4, 9.6]	7.8	[6.7, 9.6]	8.2	[7.1, 9.4]
		1	8.6	[7.2, 10.1]	14.1	[12.5, 15.3]	10.1	[8.0, 12.7]	11.3	[9.5, 13.2]
		5	10.2	[7.9, 12.3]	24.9	[23.8, 26.0]	21.4	[17.6, 24.9]	23.4	[21.6, 25.8]
0.25	0.1	-5	0.3	[-0.2, 1.1]	-0.0	[-0.9, 0.6]	0.6	[-0.1, 1.2]	1.3	[0.4, 2.0]
		-1	0.7	[-0.2, 1.5]	-0.0	[-0.7, 0.8]	0.6	[-0.1, 1.4]	1.0	[0.3, 1.7]
		0	0.8	[-0.0, 1.5]	0.4	[-0.2, 1.2]	0.6	[-0.2, 1.3]	0.7	[-0.1, 1.5]
		1	0.4	[-0.3, 1.2]	0.8	[0.3, 1.6]	0.5	[-0.1, 1.2]	-0.0	[-1.0, 0.6]
		5	0.5	[-0.2, 1.2]	1.3	[0.6, 2.0]	0.6	[-0.3, 1.6]	-1.3	[-2.1, -0.5]
	0.4	-5	2.7	[1.7, 3.8]	2.3	[1.6, 3.3]	3.0	[1.7, 4.2]	12.4	[11.3, 13.6]
		-1	2.6	[1.4, 3.7]	2.1	[1.2, 2.8]	2.6	[1.8, 3.6]	6.1	[5.0, 7.1]
		0	2.9	[1.8, 3.6]	2.8	[2.0, 3.7]	2.5	[1.7, 3.6]	2.9	[2.1, 3.6]
		1	3.0	[1.6, 3.7]	4.6	[3.6, 5.6]	3.1	[2.4, 4.4]	0.9	[-0.1, 1.9]
		5	3.0	[1.8, 3.9]	9.5	[8.8, 11.3]	2.8	[1.9, 4.2]	-0.1	[-1.2, 0.8]
	0.7	-5	5.4	[3.9, 7.1]	6.1	[4.9, 7.1]	6.7	[4.9, 8.3]	19.8	[18.7, 20.9]
		-1	6.0	[4.4, 7.0]	4.9	[3.9, 6.1]	5.7	[4.6, 6.8]	11.2	[10.2, 12.6]
		0	5.8	[4.7, 7.0]	5.7	[4.4, 6.7]	5.7	[4.7, 6.9]	5.9	[4.9, 7.1]
		1	5.4	[4.4, 6.9]	9.4	[7.8, 10.5]	6.3	[5.0, 7.7]	4.2	[3.1, 5.2]
		5	5.9	[4.9, 7.1]	12.9	[12.2, 14.0]	6.9	[5.4, 8.1]	5.0	[3.8, 6.0]
0.4	0.1	-5	0.2	[-0.2, 0.5]	-0.5	[-0.9, 0.0]	0.2	[-0.2, 0.7]	1.4	[1.0, 1.8]
		-1	0.3	[-0.1, 0.5]	-0.1	[-0.6, 0.2]	0.2	[-0.2, 0.7]	0.8	[0.2, 1.3]
		0	0.1	[-0.3, 0.6]	0.2	[-0.3, 0.6]	0.0	[-0.2, 0.5]	0.1	[-0.2, 0.5]
		1	0.2	[-0.3, 0.6]	0.6	[0.3, 1.0]	0.1	[-0.2, 0.6]	-0.4	[-0.8, 0.0]
		5	0.3	[-0.2, 0.6]	0.9	[0.5, 1.3]	0.2	[-0.1, 0.7]	-1.3	[-1.7, -0.9]
	0.4	-5	0.6	[0.1, 1.1]	-0.4	[-1.1, 0.1]	0.8	[0.1, 1.3]	5.5	[5.0, 6.0]
		-1	0.7	[0.2, 1.1]	-0.0	[-0.7, 0.6]	0.7	[0.1, 1.1]	2.8	[2.3, 3.3]
		0	0.5	[0.0, 1.1]	0.8	[0.4, 1.2]	0.7	[0.2, 1.3]	0.8	[0.2, 1.4]
		1	0.6	[0.2, 1.3]	1.8	[1.1, 2.4]	0.8	[0.2, 1.3]	-1.0	[-1.5, -0.4]
		5	0.7	[0.1, 1.2]	3.3	[2.7, 3.9]	0.7	[-0.2, 1.2]	-2.4	[-2.8, -1.9]
	0.7	-5	1.3	[0.5, 2.1]	0.5	[-0.1, 1.1]	1.4	[0.7, 2.1]	5.9	[5.2, 6.5]
		-1	1.4	[0.8, 2.0]	0.7	[0.1, 1.5]	1.1	[0.4, 2.1]	4.1	[3.3, 4.6]
		0	1.3	[0.8, 2.0]	1.3	[0.4, 1.9]	1.3	[0.6, 2.0]	1.4	[0.7, 1.8]
		1	1.3	[0.7, 1.8]	2.5	[1.8, 3.2]	1.4	[0.7, 2.0]	-0.4	[-1.2, 0.1]
		5	1.2	[0.4, 1.9]	3.2	[2.6, 3.6]	1.1	[0.5, 2.2]	-1.4	[-1.8, -0.7]

Table 7: Median [IQR] of out-of-sample *above oracle prediction error* (AOPE) when the number of predictors is 200 (percentage-points scale)

Bayes risk	Missing prop	β	Missing type: MAR X	MAR XY	MNAR X	MNAR XY
			0 50	0 50	0 50	0 50
0.1	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				
0.25	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				
0.4	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				

Table 8: Graphical median [IQR] of out-of-sample *above oracle prediction error* (AOPE) when the number of predictors is 2 (percentage-points scale)

Bayes risk	Missing prop	β	Missing type: MAR X	MAR XY	MNAR X	MNAR XY
			0 50	0 50	0 50	0 50
0.1	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				
0.25	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				
0.4	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				

Table 9: Graphical median [IQR] of out-of-sample *above oracle prediction error* (AOPE) when the number of predictors is 20 (percentage-points scale)

Bayes risk	Missing prop	β	Missing type: MAR X	MAR XY	MNAR X	MNAR XY
			0 50	0 50	0 50	0 50
0.1	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				
0.25	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				
0.4	0.1	-5				
		-1				
		0				
		1				
		5				
	0.4	-5				
		-1				
		0				
		1				
		5				
	0.7	-5				
		-1				
		0				
		1				
		5				

Table 10: Graphical median [IQR] of out-of-sample *above oracle prediction error* (AOPE) when the number of predictors is 200 (percentage-points scale)