

# Prediction with Missing Data

Dimitris Bertsimas

Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA.  
ORCID: 0000-0002-1985-1003  
dbertsim@mit.edu

Arthur Delarue

Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA,  
ORCID: 0000-0002-1687-7013  
adelarue@mit.edu

Jean Pauphilet

London Business School, London, UK,  
ORCID: 0000-0001-6352-0984  
jpauphilet@london.edu

Missing information is inevitable in real-world data sets. While imputation is well-suited and theoretically sound for statistical inference, its relevance and practical implementation for out-of-sample prediction remains unsettled. We provide a theoretical analysis of widely used data imputation methods and highlight their key deficiencies in making accurate predictions. Alternatively, we propose *adaptive* linear regression, a new class of models that can be directly trained and evaluated on partially observed data, adapting to the set of available features. In particular, we show that certain adaptive regression models are equivalent to impute-then-regress methods where the imputation and the regression models are learned simultaneously instead of sequentially. We validate our theoretical findings and adaptive regression approach with numerical results with real-world data sets.

*Key words:* Missing data imputation; Linear regression; Adaptive optimization. *Area of review:* Machine Learning and Data Science.

---

## 1. Introduction

Real-world data sets usually combine information from multiple sources, with different measurement units, data encoding, or structure, leading to a myriad of inconsistencies. In particular, they often come with some partially observed features. In contrast, the supervised learning literature largely assumes an ideal setting with a fixed set of features available for every data point. Though helpful to develop new methods, this assumption has become increasingly divorced from the modern data reality.

Recent significant progress on the question of “missing data” has largely focused on statistical inference. For prediction tasks, practitioners often first impute missing values, then train a model on the imputed dataset (impute-then-regress). While this approach is

viable offline for model calibration, it does not designed for computing predictions out-of-sample and its empirical performance is not guaranteed.

*Contributions and outline.* Our work challenges the present state-of-practice from a theoretical and practical standpoint. On the theory side, we study the asymptotic consistency of common imputation rules. Surprisingly, we find that an imputation method leads to consistent predictors (see Definition 2) if the downstream machine learning model can almost surely undo the imputation, or de-impute. This result (Theorem 1 and Corollary 1) constitutes the major theoretical contribution of this paper and questions the need for sophisticated imputation methods. Accordingly, we propose a fundamentally different approach for handling missing data in practice. Instead of imputing missing values and then building a predictive model, our novel regression framework accounts for partially observed data directly. In some special cases, we show that adaptive models are equivalent to learning the imputation and prediction models jointly instead of sequentially. We develop algorithms to train these adaptive linear regression models efficiently, and evaluate their performance in practice.

Section 2 reviews missing data mechanisms and imputation methods, and highlights key differences between inference and prediction settings. We analyze the theoretical (in)consistency of simple missing data imputation rules and the practical implications in Section 3. We introduce and develop the adaptive linear regression framework in Section 4, and numerically compare it with impute-then-regress in Section 5.

*Notation.* Nonbold lowercase characters ( $x$ ) denote scalars and bold lowercase characters ( $\mathbf{x}$ ) vectors. Uppercase letters designate random variables, e.g.  $X$  and  $\mathbf{X}$  respectively denote a random scalar and vector. The symbol  $\perp$  designates independent random variables. For any positive integer  $n$ , we let  $[n] = \{1, \dots, n\}$ .

## 2. Motivating literature review

We review the literature on missing data, noting that both the taxonomy of missing patterns and many imputation techniques have been designed with inference in mind and might not be well suited for prediction.

### 2.1. Missingness mechanisms

Missingness mechanisms answer the question: why is the data missing? In the context of parameter estimation via likelihood maximization, Rubin (1976) introduced three missing

data mechanisms. It is worth noting that in a prediction setting, these classical definitions do not distinguish the dependent or target variable  $Y$  from the input vector  $\mathbf{X}$ .

Consider  $n$  i.i.d. samples  $(\mathbf{x}_i, \mathbf{m}_i)$ ,  $i \in [n]$ , where  $\mathbf{x}_i \in \mathbb{R}^d$  is the vector of covariates and  $\mathbf{m}_i \in \{0, 1\}^d$  is a vector indicating the missing covariates, i.e.,  $m_{ij} = 1$  if  $x_{ij}$  is missing, 0, otherwise. For every data point  $i$ ,  $\|\mathbf{m}_i\|_0 := \sum_j m_{ij}$  covariates are missing. We refer to  $\mathbf{m}_i$  as the missingness *indicator* or missingness *pattern* of sample  $i$ .

We further denote  $\mathbf{o}(\mathbf{x}_i, \mathbf{m}_i)$  the  $(d - \|\mathbf{m}_i\|_0)$ -dimensional vector of **o**bserved covariates (Seaman et al. 2013). Symmetrically,  $\mathbf{o}(\mathbf{x}_i, \mathbf{1} - \mathbf{m}_i)$  is the vector of unobserved ones. With these notations,  $(\mathbf{x}_i, \mathbf{m}_i)_{i \in [n]}$  corresponds to the data set of *realizations* while  $(\mathbf{o}(\mathbf{x}_i, \mathbf{m}_i), \mathbf{m}_i)_{i \in [n]}$  is the data set of *observations*. Let us assume that the realized data set is sampled i.i.d. from a distribution in  $\mathcal{F} = \{g(\mathbf{x}; \boldsymbol{\theta})h(\mathbf{m}|\mathbf{x}; \boldsymbol{\phi}) : (\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Omega\}$ . Here, the density of the fully observed data,  $g(\cdot; \boldsymbol{\theta})$ , is parametrized by the vector  $\boldsymbol{\theta}$ , while  $\boldsymbol{\phi}$  parametrizes the density function of the missingness patterns  $\mathbf{m}$  conditioned on  $\mathbf{x}$ , denoted by  $h(\cdot|\mathbf{x}; \boldsymbol{\phi})$ . Then, the likelihood of the observed data is given by integrating over the missing values

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^n \int g(\mathbf{x}; \boldsymbol{\theta})h(\mathbf{m}_i|\mathbf{x}; \boldsymbol{\phi})\delta_{\mathbf{o}(\cdot, \mathbf{m}_i)=\mathbf{o}(\mathbf{x}_i, \mathbf{m}_i)}(\mathbf{x})d\mathbf{x}, \quad (1)$$

where  $\delta$  is the standard Dirac measure. The parameter  $\boldsymbol{\theta}$ , which controls the data generating process, is often more of interest than the missingness generating process and  $\boldsymbol{\phi}$ . However, in general, the two parameters need to be estimated jointly. Rubin (1976) noted that joint estimation is much easier under a mechanism called Missing At Random:

**DEFINITION 1.** Missing data are called *missing at random* (MAR) if, conditioned on the observed covariates  $\mathbf{o}(\mathbf{X}, \mathbf{M})$ , the missingness indicator  $\mathbf{M}$  is independent of the unobserved covariates  $\mathbf{o}(\mathbf{X}, \mathbf{1} - \mathbf{M})$ .

Indeed, under MAR, the term  $h(\mathbf{m}_i|\mathbf{x}; \boldsymbol{\phi})$  in (1) only depends on  $\mathbf{o}(\mathbf{x}, \mathbf{m}_i) = \mathbf{o}(\mathbf{x}_i, \mathbf{m}_i)$ . Hence,  $\boldsymbol{\theta}$  can be estimated independently of  $\boldsymbol{\phi}$  by maximizing the partial likelihood, i.e., (1) without the  $g(\mathbf{m}_i|\mathbf{x}; \boldsymbol{\phi})$  term. A stronger assumption than MAR is Missing Completely At Random (MCAR), under which  $\mathbf{M}$  is assumed independent of the covariates  $\mathbf{X}$ . MCAR simply assumes that the marginal density function  $g(\mathbf{m}|\mathbf{x}; \boldsymbol{\phi})$  does not depend on  $\mathbf{x}$ . Finally, we say that the missing data is *Not Missing At Random* (NMAR) if MAR is not satisfied.

At this point, we emphasize two limitations of the above definitions in a predictive setting:

1. These definitions apply to a single data set with no distinction between training and test sets, which is a common distinction in supervised learning tasks.
2. The target response  $Y$  is not explicitly considered when defining the missingness mechanisms. At best, we can consider  $Y$  as part of the input  $\mathbf{X}$ . However,  $Y$  is a unique variable - for example, it cannot be missing - and we argue it deserves specific treatment.

## 2.2. Inference with missing data and multiple imputation

Techniques to deal with missing data in inference settings can be divided into two independent directions. For maximum likelihood estimation tasks, such as the one presented in Section 2.1, Dempster et al. (1977) proposed a variant of the EM algorithm which provides unbiased estimators even in presence of missing data, under the MAR assumption. To derive confidence intervals, one can then estimate the variance of these estimates using the supplemented EM algorithm (Meng and Rubin 1991) or Louis’ formula (Louis 1982). However, implementation of these techniques are not straightforward and dedicated algorithms must be designed even for the most common estimation tasks, such as generalized linear models (Ibrahim et al. 2005) or logistic regression (Jiang et al. 2020).

An imputation method is a procedure that takes as input a data set of observations  $(\mathbf{o}(\mathbf{x}_i, \mathbf{m}_i), \mathbf{m}_i)_{i \in [n]}$  and returns an imputed data set  $(\tilde{\mathbf{x}}_i)_{i \in [n]}$ , where  $\tilde{\mathbf{x}}_i \in \mathbb{R}^d$ . The most accurate imputation method (minimizing mean-squared error) satisfies

$$\tilde{\mathbf{x}}_i = \mathbb{E} \left[ \mathbf{X}_i \mid \{ \mathbf{o}(\mathbf{X}_{i'}, \mathbf{m}_{i'}) = \mathbf{o}(\mathbf{x}_{i'}, \mathbf{m}_{i'}) \}_{i' \in [n]} \right].$$

In particular, observed features are left unchanged, i.e.,  $\mathbf{o}(\tilde{\mathbf{x}}_i, \mathbf{m}_i) = \mathbf{o}(\mathbf{x}_i, \mathbf{m}_i)$ . In the literature, a myriad of imputation techniques have been proposed based on mean and mode imputation (Little and Rubin 2019),  $k$ -nearest neighbors (Troyanskaya et al. 2001, Brás and Menezes 2007), least square regression (Bø et al. 2004, Kim et al. 2005, Cai et al. 2006, Zhang et al. 2008), support vector machine/regression (Wang et al. 2006, Bertsimas et al. 2018), decision trees (Bertsimas et al. 2018), neural networks (Yoon et al. 2018), or factor analysis and other dimension reduction techniques (Mohamed et al. 2009, Husson et al. 2019).

The objective in inference is to estimate a parameter  $\theta$  (e.g., a particular coefficient in a linear regression model) based on observations. Given access to a valid estimation procedure for the full data set  $(\mathbf{x}_i)_{i \in [n]}$ , i.e., the ability to compute  $\hat{\theta}(\{\mathbf{x}_i\}_{i \in [n]}) = \mathbb{E}[\theta \mid \{\mathbf{X}_i = \mathbf{x}_i\}_{i \in [n]}]$ , the objective in a missing data setting is to compute

$$\hat{\theta}(\{\mathbf{o}(\mathbf{x}_i, \mathbf{m}_i)\}_{i \in [n]}) = \mathbb{E}[\hat{\theta}(\{\mathbf{X}_i\}_{i \in [n]}) \mid \{\mathbf{o}(\mathbf{X}_{i'}, \mathbf{m}_{i'}) = \mathbf{o}(\mathbf{x}_{i'}, \mathbf{m}_{i'})\}_{i' \in [n]}]. \quad (2)$$

However, a simple impute-then-estimate procedure yields

$$\hat{\theta}(\{\tilde{\mathbf{x}}_i\}_{i \in [n]}) = \hat{\theta}\left(\left\{\mathbb{E}[X_i \mid \{\mathbf{o}(\mathbf{X}_{i'}, \mathbf{m}_{i'}) = \mathbf{o}(\mathbf{x}_{i'}, \mathbf{m}_{i'})\}_{i' \in [n]}]\right\}_{i \in [n]}\right)$$

which is not equivalent to the expectation of  $\theta$  conditioned on the observed data (2) due to the inverted order of the estimation step  $\hat{\theta}(\cdot)$  and the expectation step.

To avoid the pitfalls of single imputation, Rubin (1987) proposed the use of *multiple imputed* data sets. A multiple imputation method returns not one but a collection of imputed data sets  $(\tilde{\mathbf{x}}_i^{(j)}, i \in [n])$  for  $j \in [k]$ , which can be interpreted as  $k$  independent samples from the distribution of

$$\{\mathbf{X}_i\}_{i \in [n]} \mid \{\mathbf{o}(\mathbf{X}_{i'}, \mathbf{m}_{i'}) = \mathbf{o}(\mathbf{x}_{i'}, \mathbf{m}_{i'})\}_{i' \in [n]}$$

By estimating  $\theta$  on each imputed data set individually and combining the results using Rubin's rule, we indeed recover unbiased estimates of the parameter  $\theta$  and its variance. In practice, multiple imputation can be performed using sequential regression (Raghunathan et al. 2001), sequential decision trees (Burgette and Reiter 2010) or random forests (Stekhoven and Bühlmann 2012).

### 2.3. Inference vs. prediction

In supervised learning, the goal is to predict a response  $Y$  based on the vector of covariates  $\mathbf{X}$ . The prediction task can be reduced to an inference setting, by considering predictions of the form  $f(\mathbf{x}; \boldsymbol{\theta})$ , where  $\boldsymbol{\theta} \in \Omega$  and  $\{f(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega\}$  is a fixed class of learners, and solving an empirical risk minimization problem

$$\hat{\boldsymbol{\theta}} \in \arg \min_{\boldsymbol{\theta} \in \Omega} \sum_{i=1}^n \ell(y_i, f(\mathbf{x}_i; \boldsymbol{\theta})) + R(\boldsymbol{\theta}),$$

where  $\ell(\cdot, \cdot)$  is a loss function measuring the prediction error and  $R(\cdot)$  is a regularization term. The loss function is chosen so as to produce asymptotically consistent estimators (see definition below).

DEFINITION 2. A predictor  $f$  trained on complete data is said to be *consistent* if, for any covariates  $\mathbf{x}$ ,  $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ .

Based on the previous discussion in Section 2.2, we argue that an impute-then-predict approach, though widely common in practice, would incur a bias in parameter estimation. Alternatively, one can perform multiple imputations of the training set and compute a family of models  $f(\cdot, \hat{\theta}^{(1)}), \dots, f(\cdot, \hat{\theta}^{(k)})$ . Having multiple models, however, might lead to issues related to storage, tractability, and interpretability.

While these strategies address the issue of training a predictive model on data with missing entries, they do not provide guidance on how to perform out-of-sample prediction. Indeed, predictive models trained on imputed data can later only be applied to a fully observed vector  $\mathbf{x}$ . Given a predictor  $f$  trained on complete or imputed data and a new observation  $(\mathbf{o}(\mathbf{x}, \mathbf{m}), \mathbf{m})$ , one must first compute an imputed version  $\tilde{\mathbf{x}}$  of  $(\mathbf{o}(\mathbf{x}, \mathbf{m}), \mathbf{m})$  and then predict  $f(\tilde{\mathbf{x}})$ . Though this approach is widely implemented in practice, it carries several deficiencies: First, as for parameter estimation, applying the predictor rule  $f$  to a single imputed version of the observation leads to biased prediction, because it obliterates our uncertainty about the missing features (Josse et al. 2019). On the other hand, performing multiple imputations of the test set and averaging predictions leads to a consistent estimator (Josse et al. 2019, Theorem 2), but requires additional computational time and storage resources. This violates the assumption that computing predictions from a previously trained model is extremely fast, computationally efficient, and can be performed in an online and/or embedded fashion — an assumption which is critical for many applications of machine learning.

In short, though multiple imputation is the “gold standard” for statistical inference with missing data, applying it in a prediction setting can be computationally challenging. If data are missing in both the training and testing set, it may be necessary to train multiple models and evaluate each one on multiple imputed versions of each out-of-sample data point. It is therefore of interest to study the impact of missing data on the predictive task directly and design methods that do not require any imputation.

For the latter case, a key obstacle to overcome is that the vector  $\mathbf{o}(\mathbf{x}, \mathbf{m})$  does not have a fixed dimension. One can instead consider the extended vector  $\mathbf{e}(\mathbf{x}, \mathbf{m}) \in (\mathbb{R} \cup \{\text{NA}\})^d$  defined as  $\mathbf{e}(\mathbf{x}, \mathbf{m})_i = \text{NA}$  if  $m_i = 1$ ,  $\mathbf{e}(\mathbf{x}, \mathbf{m})_i = x_i$ , otherwise. However, the function  $f$  needs to be able to account for the half-discrete nature of its argument  $\mathbf{e}(\mathbf{x}, \mathbf{m})$ . To the best

of our knowledge, only tree-based methods are currently able to cope with such features. The Missingness Incorporated as Attribute (MIA) technique (Twala et al. 2008) constructs splits that apply directly on the extended vector and is the state-of-the-art for treating missing values in decision trees (see Josse et al. 2019, for a numerical comparison).

### 3. Particularities of missing data in prediction

In this section, we investigate the validity of the missing data literature presented in Section 2 to predictive settings. In particular, we explore theoretically the quality of simple rules like mean or mode-imputation for prediction in Section 3.1 and 3.2, we question the MAR assumption in Section 3.3, and discuss the issue of inconsistencies between the training and test set in Section 3.4. As in Josse et al. (2019), we consider a simplified setting in  $d$  dimensions where only the first covariate  $X_1$  is missing. In this setting, the optimal (consistent) predictor is predicting

$$\begin{cases} \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, M_1 = 0], & \text{if } m_1 = 0, \\ \mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1], & \text{if } m_1 = 1. \end{cases} \quad (3)$$

Here, we concisely denote  $\mathbf{x}_{2:d}$  the  $(d-1)$ -dimensional vector  $(x_2, \dots, x_d)$ .

#### 3.1. Consistency of generic imputation rules

In practice, impute-then-regress strategies are prevalent due to their ease of implementation. Hence, understanding how and whether they negatively impact the learned prediction rule are crucial questions to answer. We study the common practice of imputing a deterministic value  $\mu(\mathbf{x}_{2:d})$  on the training set and its impact on the downstream prediction task. This model captures mean and mode imputation (for which  $\mu(\mathbf{x}_{2:d})$  does not depend on the other covariates  $\mathbf{x}_{2:d}$ ), conditional mean and mode imputation, and more generally any method that deterministically imputes a value for  $X_1$  as a function of the other covariates  $\mathbf{x}_{2:d}$ . We restrict our analysis to imputation rules  $\mu$  that are almost surely continuous in  $\mathbf{X}_{2:d}$  in the sense of Definition 3.

**DEFINITION 3.** Given a random variable  $\mathbf{X}$  taking values in  $\mathcal{X}$ , a function  $f : \mathcal{X} \mapsto \mathbb{R}$  is *almost surely continuous in  $\mathbf{X}$*  if  $\mathbb{P}_{\mathbf{x} \sim \mathbf{X}}(\{\mathbf{x} : f \text{ continuous at } \mathbf{x}\}) = 1$ .

Constant rules (e.g., mean and mode imputation), piece-wise constant rules (e.g., conditional mean and mode imputation) or linear models satisfy this condition.

**THEOREM 1.** *Consider a universally consistent learning algorithm when trained on any fully observed data set. Consider a vector  $\mathbf{X} = (X_1, X_2, \dots, X_d)$  such that  $(X_2, \dots, X_d)$  has a continuous density  $g > 0$ , and a response  $Y$  such that  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$  is continuous. Further assume that the missingness pattern satisfies  $X_1 \perp\!\!\!\perp M_1 | \mathbf{X}_{2:d}$  (MAR) and that  $\mathbf{x}_{2:d} \mapsto \mathbb{P}(M_1 = 1 | \mathbf{X}_{2:d} = \mathbf{x}_{2:d})$  is continuous. Given an almost surely continuous imputation function  $\mu$ , systematically imputing  $\mu(\mathbf{x}_{2:d})$  for  $X_1 | \mathbf{X}_{2:d} = \mathbf{x}_{2:d}$  on the training set and training a predictor on the imputed data set asymptotically leads to the following prediction rule, denoted  $f_{\mu\text{-impute}}(\mathbf{x})$  and equal almost everywhere to*

$$\begin{cases} \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, M_1 = 0], & \text{if } x_1 \neq \mu(\mathbf{x}_{2:d}), \\ \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, M_1 = 0], & \text{if } x_1 = \mu(\mathbf{x}_{2:d}) \text{ and } \eta(\mathbf{x}) = 0, \\ \alpha(\mathbf{x})\mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1] + (1 - \alpha(\mathbf{x}))\mathbb{E}[Y|X_1 = \mu(\mathbf{x}_{2:d}), \mathbf{X}_{2:d} = \mathbf{x}_{2:d}], & \text{otherwise,} \end{cases}$$

where

- $\eta(\mathbf{x}) = \mathbb{P}(M_1 = 1 | \mathbf{X}_{2:d} = \mathbf{x}_{2:d})$  is the probability that  $X_1$  is missing given  $\mathbf{X}_{2:d}$ ,
- $p_\mu(\mathbf{x}) = \mathbb{P}(X_1 = \mu(\mathbf{x}_{2:d}) | \mathbf{X}_{2:d} = \mathbf{x}_{2:d})$  is the probability for the true  $X_1$  (before imputation) to take the imputed value  $\mu(\mathbf{x}_{2:d})$  given the observed covariates,
- and  $\alpha(\mathbf{x}) = \frac{\eta(\mathbf{x})}{\eta(\mathbf{x}) + p_\mu(\mathbf{x})(1 - \eta(\mathbf{x}))}$  is the posterior probability that  $X_1$  was missing before imputation, given that it takes the value  $\mu(\mathbf{x}_{2:d})$  after imputation, i.e.,  $\mathbb{P}(M_1 = 1 | X_1^{\text{imputed}} = \mu(\mathbf{x}_{2:d}), \mathbf{X}_{2:d} = \mathbf{x}_{2:d})$ .

We defer the proof to Appendix A. Intuitively, in the asymptotic regime, only the training data in the neighborhood of  $\mathbf{x}$  affect the prediction. In the first two cases,  $\mathbf{x}$  is locally surrounded by points with no missing entries (almost surely). In the third case however, the points in the neighborhood of  $\mathbf{x}$  come from a mixture of two distributions: either  $(\mu(\mathbf{x}_{2:d}), \mathbf{X}_{2:d}) | M_1 = 1$  in the case where  $X_1$  is missing, or  $(X_1, \mathbf{X}_{2:d}) | M_1 = 0$ . So, the predicted outcome is a weighted average of both conditional expectations, with  $\alpha(\mathbf{x})$  being the proper weighting factor.

**REMARK 1.** In the case where  $X_1$  is continuous,  $p_\mu(\mathbf{x}) = \mathbb{P}(X_1 = \mu(\mathbf{x}_{2:d}) | \mathbf{X}_{2:d} = \mathbf{x}_{2:d}) = 0$  so  $\alpha(\mathbf{x}) = 1$  and the third case becomes  $\mathbb{E}[Y | \mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1]$ . From the proof of Theorem 1, we also note that the MAR assumption is not needed in this case.

**REMARK 2.** Observe that  $\alpha(\mathbf{x}) \neq \eta(\mathbf{x}) = \mathbb{P}(M_1 = 1 | \mathbf{X}_{2:d} = \mathbf{x}_{2:d})$  in general. Although  $X_1$  and  $M_1$  are conditionally independent by the MAR assumption, imputation induces correlation between  $X_1$  and  $M_1$  after imputation.



We now apply Theorem 1, to study the out-of-sample predictions from a learner trained on  $\mu$ -imputed data. For a new observation  $(\mathbf{o}(\mathbf{x}, m_1), m_1)$ , we apply  $\mu$ -imputation if needed and then predict according to  $f_{\mu\text{-impute}}(\mathbf{x})$ . If  $m_1 = 0$ , the impute-then-predict rule yields

$$\begin{cases} \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, M_1 = 0], & \text{if } x_1 \neq \mu(\mathbf{x}_{2:d}), \\ \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, M_1 = 0], & \text{if } x_1 = \mu(\mathbf{x}_{2:d}) \text{ and } \eta(\mathbf{x}) = 0, \\ \alpha(\mathbf{x})\mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1] + (1 - \alpha(\mathbf{x}))\mathbb{E}[Y|X_1 = \mu(\mathbf{x}_{2:d}), \mathbf{X}_{2:d} = \mathbf{x}_{2:d}], & \text{otherwise.} \end{cases}$$

For this rule to agree with the Bayes-consistent predictor,  $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}, M_1 = 0]$ , almost everywhere, we need the third case to happen with probability zero. If  $X_1$  is continuous, then  $\mathbb{P}(X_1 = \mu(\mathbf{x}_{2:d})) = 0$  so this condition is satisfied by design. On the other hand, if  $X_1$  is discrete, choosing  $\mu(\mathbf{x}_{2:d})$  outside of the original support of  $X_1$ , i.e.,  $\mathbb{P}(X_1 = \mu(\mathbf{x}_{2:d})) = 0$ , provides consistency.

If  $m_1 = 1$ , then  $x_1$  is replaced by its imputed value  $\mu(\mathbf{x}_{2:d})$  and necessarily we have  $\eta(\mathbf{x}) > 0$  (because the event  $M_1 = 1$  occurred) so the impute-then-predict rule yields

$$\alpha(\mathbf{x})\mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1] + (1 - \alpha(\mathbf{x}))\mathbb{E}[Y|X_1 = \mu, \mathbf{X}_{2:d} = \mathbf{x}_{2:d}],$$

which agrees with the Bayes-consistent estimator  $\mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1]$  if  $\alpha(\mathbf{x}) = 1$ . Again,  $\alpha(\mathbf{x})$  corresponds to the posterior probability that  $X_1$  was missing in the original observation given that  $X_1 = \mu(\mathbf{x}_{2:d})$  in the final observation. In other words, Theorem 1 indicates that consistency is achieved as long as the predictor can almost surely de-impute, that is properly guess after imputation whether  $X_1$  was originally missing or not. This discussion can be summarized by the following corollary:

**COROLLARY 1.** *Under the assumptions and notations of Theorem 1,  $\mu$ -imputation-then-regress asymptotically leads to Bayes-consistent estimates at  $\mathbf{x}$  if and only if  $\alpha(\mathbf{x}) = 1$ .*

Consequently, Theorem 1 is highly counter-intuitive. Indeed, one could think that a good imputation method should produce data sets that are plausible, i.e., where imputed and non-imputed observations are undistinguishable ( $\alpha(\mathbf{x}) = 0$ ). On the contrary, as far as predictive power is concerned, Theorem 1 speaks in favor of imputation methods that can be surely de-imputed ( $\alpha(\mathbf{x}) = 1$ ), at least in the asymptotic regime.

### 3.2. (In)consistency of simple imputation rules

In this section, we materialize the previous discussion on two simple and widely used examples: mean imputation for continuous variables and mode imputation for discrete ones. In both cases,  $\mu(\mathbf{x}_{2:d}) = \mu$  is independent of the  $d - 1$  other features.

*Mean imputation.* If  $X_1$  is continuous, the probability that it takes any specific value is 0 ( $p_\mu(\mathbf{x}) = 0$ ). As a result, systematically imputing  $\mu$  for  $X_1$  creates a discontinuity in the posterior (after imputation) distribution of  $X_1$  and the events  $\{X_1 = \mu\}$  and  $\{M_1 = 1\}$  are equal almost surely. In other words,  $\alpha(\mathbf{x}) = 1$ , and according to Theorem 1, mean-imputation is asymptotically consistent. Josse et al. (2019) made a similar observation.

*Mode imputation.* When  $X_1$  is discrete, however, choosing  $\mu$  as the mode of the distribution of  $X_1|M_1 = 0$  does not provide consistency. Indeed,  $p_\mu(\mathbf{x}) > 0$  so  $\alpha(\mathbf{x}) < 1$ . We illustrate this deficiency on a simple example:

EXAMPLE 1. Assume  $X_1$  and  $M_1$  are two independent Bernoulli random variables with parameter  $2/3$  and  $1/2$  respectively. Assume also that  $Y = X_1 + \varepsilon$  with  $\varepsilon$  a centered noise independent from  $X_1$  and  $M_1$ . If we impute  $X_1$  with its mode (i.e., 1), Theorem 1 applies with  $\eta(\mathbf{x}) = 1/2$  and  $p_\mu(\mathbf{x}) = 2/3$  so that  $\alpha(\mathbf{x}) = 3/5$ . Out-of-sample, impute-1-then-predict leads to

$$\begin{aligned} &0, \text{ if } m_1 = 0 \text{ and } x_1 = 0, \\ &4/5, \text{ if } m_1 = 0 \text{ and } x_1 = 1, \\ &4/5, \text{ if } m_1 = 1. \end{aligned}$$

while the Bayes-consistent estimator is

$$\begin{aligned} &x_1, \text{ if } m_1 = 0, \\ &2/3, \text{ if } m_1 = 1. \end{aligned}$$

Conversely, choosing  $\mu$  outside of the original support of  $X_1$ , i.e., encoding missingness as a new value, provides consistency.

In short, simple deterministic imputation rules (mean imputation for continuous variables and encoding missingness as a new category for discrete variables) can asymptotically lead to consistent predictions. However, these consistency guarantees are only valid if the training and test sets are imputed similarly; if the downstream predictor is universally consistent; and in the asymptotic regime with infinite amount of data.

### 3.3. Revisiting Not Missing at Random

A key feature of the MAR assumption is its guarantee that the distributions of  $X_1|\mathbf{X}_{2:d}, M_1 = 1$  and  $X_1|\mathbf{X}_{2:d}, M_1 = 0$  are identical. Since missing data imputation involves learning a model of the first distribution using data from the second distribution, the MAR assumption is critical to most of the missing data imputation literature, and many algorithms break down when the assumption does not hold.

However, we argue that, from the perspective of the downstream prediction task, data not missing at random is a blessing, not a curse. Indeed, such data carries more information about the outcome  $Y$  than if missing at random, and is thus more valuable for prediction. Recall that the optimal Bayes-consistent estimator of  $Y$  is

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}, M_1 = 0] \mathbb{1}[M_1 = 0] + \mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1] \mathbb{1}[M_1 = 1],$$

and the associated mean square error (risk) is

$$\mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}, M_1 = 0])^2] \mathbb{P}(M_1 = 0) + \mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}_{2:d}, M_1 = 1])^2] \mathbb{P}(M_1 = 1). \quad (4)$$

Alternatively, consider another missing pattern  $M'_1$  that is independent of  $Y$ , the average risk of the Bayes-consistent estimator simplifies into

$$\mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}])^2] \mathbb{P}(M'_1 = 0) + \mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}_{2:d}])^2] \mathbb{P}(M'_1 = 1). \quad (5)$$

By definition of the conditional expectation,  $\mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}, M_1 = 0])^2] \leq \mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}])^2]$ , so that the mean square error with independent missingness patterns  $M'_1$  (5) is higher than in the general case (4), as formalized in the following proposition.

**PROPOSITION 1.** *Consider two missingness mechanisms  $M_1$  and  $M'_1$  leading to the same proportion of missing entries:  $\mathbb{P}(M_1 = 1) = \mathbb{P}(M'_1 = 1)$ . Further assume that, conditioned on  $\mathbf{X}_{2:d}$ ,  $M'_1$  is independent of  $Y$ . Then the optimal Bayes-consistent prediction rule achieves a lower prediction error under  $M_1$  than under  $M'_1$ .*

*Proof* Since  $\mathbb{P}(M_1 = 1) = \mathbb{P}(M'_1 = 1)$ , (4)  $\leq$  (5). □

It is easy to find examples where this inequality is strict. For instance, let  $d = 1$ ,  $X_1$  be a Bernoulli random variable with parameter  $1/2$ , and  $Y = X_1$ . Consider  $M_1 = X_1$  and  $M'_1 \perp\!\!\!\perp X_1$  another Bernoulli random variable with parameter  $1/2$ . Then, the optimal predictor under  $M_1$  is  $Y = X_1$  when  $M_1 = 0$  and  $Y = M_1 = 1$  when  $M_1 = 1$ , with empirical risk 0. Meanwhile, the Bayes-optimal predictor under  $M'_1$  is  $Y = X_1$  when  $M'_1 = 0$  and  $Y = 1/2$  when  $M'_1 = 1$ , with empirical risk  $1/8$ .

### 3.4. Pitfalls when missing data is different between the training and test set

So far, we have assumed that data was equally missing in both the train and test sets, and considered the optimal Bayes-consistent estimator (3) based on the observation  $(\mathbf{o}(\mathbf{X}, \mathbf{M}), \mathbf{M})$  as a benchmark. We now analyze the difficulties arising when missing data impacts the training and the test set differently.

Differences in sources and mechanisms of missingness between the training and test sets can occur naturally in real-life settings, and can be considered a form of distributional shift. These differences can be created artificially by using different imputation methods for out-of-sample prediction than for training: for example, when average values for mean-imputation are dynamically updated based on newly observed data, but without re-training the prediction model. Indeed, in Section 3.1 and 3.2, we discussed the (in)consistency of impute-then-regress methods when the same  $\mu$ -imputation method is used for training and out-of-sample prediction. If a different imputation rule  $\mu'(\cdot)$  is used for out-of-sample imputation, however, our conclusions no longer hold. Consider, for instance, that we apply  $\mu'$  imputation for a new observation with  $m_1 = 1$ . If  $\mu'(\mathbf{x}_{2:d}) \neq \mu(\mathbf{x}_{2:d})$ , then  $f_{\mu\text{-impute}}$  predicts  $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}, M_1 = 0]$  instead of the Bayes-consistent estimator  $\mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1]$ .

To analyze the challenges stemming from regime changes between training and testing, we consider a stylized situation where data is never missing in the training data but  $X_1$  can be missing in the test set. In this case, the optimal Bayes-consistent estimator of  $Y$  learned on the train set is simply  $f(x_1, \mathbf{x}_{2:d}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ . For a new observation with  $M_1 = 1$ , it would be Bayes-optimal to predict  $\mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1]$ . However, this estimator cannot be learned since we only observe  $M_1 = 0$  on the training data. At best, one can predict  $\mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}]$ , which only matches the Bayes-optimal prediction under the stringent and unverifiable assumption that  $Y \perp\!\!\!\perp M_1|\mathbf{X}_{2:d}$ . We now consider an impute-then-regress strategy. Denote  $\mu(\mathbf{x}_{2:d})$  the imputed value for  $X_1$  given  $\mathbf{X}_{2:d} = \mathbf{x}_{2:d}$  and  $\varepsilon = X_1 - \mu(\mathbf{X}_{2:d})$  the imputation error. Then,

$$\begin{aligned} \mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}] &= \mathbb{E}_\varepsilon[\mathbb{E}[Y|X_1 = \mu(\mathbf{x}_{2:d}) + \varepsilon, \mathbf{X}_{2:d} = \mathbf{x}_{2:d}]] \\ &= \mathbb{E}_\varepsilon[f(\mu(\mathbf{x}_{2:d}) + \varepsilon, \mathbf{x}_{2:d})] \\ &= f(\mu(\mathbf{x}_{2:d}), \mathbf{x}_{2:d}) + f^{(1)}(\mu(\mathbf{x}_{2:d}), \mathbf{x}_{2:d})\mathbb{E}[\varepsilon] + \frac{1}{2}f^{(2)}(\mu(\mathbf{x}_{2:d}), \mathbf{x}_{2:d})\mathbb{E}[\varepsilon^2] + o(\mathbb{E}[\varepsilon^2]), \end{aligned}$$

where the inversion of the Taylor expansion and the expectation in the last line is valid if  $f$  admits uniformly bounded derivatives and  $\mathbb{E}[\varepsilon^n]$  decay sufficiently fast with  $n$ , e.g.,  $\varepsilon$  is sub-Gaussian (Wainwright 2019). We make the following observations:

- In general,  $\mathbb{E}[Y | \mathbf{X}_{2:d} = \mathbf{x}_{2:d}] \neq f(\mu(\mathbf{x}_{2:d}), \mathbf{x}_{2:d})$  meaning that impute-then-regress leads to sub-optimal predictions.
- As a first order approximation,  $\mathbb{E}[Y | \mathbf{X}_{2:d} = \mathbf{x}_{2:d}] - f(\mu(\mathbf{x}_{2:d}), \mathbf{x}_{2:d})$  scales like the average imputation bias  $\mathbb{E}[\varepsilon] = \mathbb{E}[X_1 - \mu(\mathbf{X}_{2:d})]$ . Unless MAR is assumed, this quantity cannot be estimated from data, leading to systematic prediction errors. Under MAR, however, taking  $\mu(\mathbf{x}_{2:d}) = \mathbb{E}[X_1 | \mathbf{X}_{2:d}, M_1 = 0] = \mathbb{E}[X_1 | \mathbf{X}_{2:d}]$  leads to  $\mathbb{E}[\varepsilon] = 0$  and this first-order error vanishes.
- The second-order term scales like the square mean imputation error  $\mathbb{E}[\varepsilon^2]$ . Hence, the more accurate the imputation method, the lower the prediction error. However, unless  $X_1$  is a deterministic function of the other variables, this second term does not vanish. Using linear models (for which  $f^{(2)} = 0$ ) or adding a second-order correction to the impute-then-regress prediction appear to be more reasonable predictive strategies.

### 3.5. Implications for research and practice

In this section, we illustrated the extent to which the data imputation methodology developed with inference in mind (Section 2) are inadequate for predictive tasks. For instance, they recommend multiple imputation of the test set, which is neither trivial nor computationally viable, while simple imputation rules can be optimal from a predictive point of view (Section 3.1-3.2). The key assumption for valid imputation is the MAR assumption, which, from a prediction point of view, actually less favorable than NMAR (Section 3.3). The analysis we conducted suggests a few practical guidelines and remarks:

1. Consistent predictions are obtained whenever the prediction model can de-impute with certainty.
2. Missing data on the test set should be imputed using the same method as on the training set. This simple observation favors simple imputation rules such as mean-imputation over black-box methods.
3. For categorical or discrete variables, missingness should be encoded as a separate category instead of imputing the mode. For continuous variables, mean imputation can lead to consistent predictions. However, we should emphasize that this theoretical conclusion is only valid: if the down-stream prediction task uses a universally consistent learner such as  $k$ -nearest neighbors or random forests; if the data is MAR; asymptotically as the number of observations  $n \rightarrow \infty$ .

4. Dealing with so-called “distributional shifts”, i.e., different missing data patterns between training and testing, remains an open question for research. Practitioners should always investigate the underlying mechanisms leading to a variable being missing.

#### 4. A Framework for Regression with Missing Values

Section 3 presents several counter-intuitive insights about prediction with missing data and shows, in particular, that simple impute-then-regress methods can asymptotically converge to the Bayes-optimal predictor  $\mathbb{E}[Y|\mathbf{M} = \mathbf{m}, \mathbf{o}(\mathbf{X}, \mathbf{M}) = \mathbf{o}(\mathbf{x}, \mathbf{m})]$ . Yet, a natural question to ask is whether one could learn this optimal predictor directly, i.e., build a predictive model that (a) leverages information from the missingness pattern  $\mathbf{M}$  and (b) applies to the observed vectors  $\mathbf{o}(\mathbf{X}, \mathbf{M})$  directly.

Formally, we want to solve an empirical risk minimization problem

$$\min_{f \in \mathcal{F}} \sum_{i=1}^n \ell(y_i, f(\mathbf{o}(\mathbf{x}_i, \mathbf{m}_i), \mathbf{m}_i)) + R(f),$$

where the learners in the class  $\mathcal{F}$  take  $\mathbf{o}(\mathbf{x}, \mathbf{m})$  and  $\mathbf{m}$  as arguments.  $R(\cdot)$  is a regularization term introduced to reduce over-fitting and improve out-of-sample performance (Bousquet and Elisseeff 2002). The main complication is that many learners, including linear models, cannot cope with the variable dimension of the vector  $\mathbf{o}(\mathbf{x}, \mathbf{m})$  by design. We propose a novel framework for regression with missing features in Section 4.1, leading to a hierarchy of linear models presented in Section 4.2. In Section 4.3, we show that some stages of this hierarchy can be seen as an impute-then-regress strategy where the imputation and the regression models are performed simultaneously.

##### 4.1. Linear regression with missing values

We restrict our attention to linear models of the form  $f(\mathbf{o}(\mathbf{x}, \mathbf{m}), \mathbf{m}) = \sum_{j:m_j=0} w_j(\mathbf{m})x_j$ . In other words, we allow the parameters  $\mathbf{w}(\mathbf{m})$  of the linear model to depend on the set of observed covariates and the prediction is then obtained by taking the inner product of  $\mathbf{x}$  with  $\mathbf{w}(\mathbf{m})$ . Since not all coordinates of  $\mathbf{x}$  are observed, we consider a restricted inner product on the available covariates only,  $\mathbf{o}(\mathbf{x}, \mathbf{m})$ . For ease of notation we define

$$\langle \mathbf{w}(\mathbf{m}), \mathbf{x} \rangle_{\mathbf{m}} := \sum_{\substack{j=1 \\ m_j=0}}^d w_j(\mathbf{m})x_j = \mathbf{o}(\mathbf{w}(\mathbf{m}), \mathbf{m})^\top \mathbf{o}(\mathbf{x}, \mathbf{m}) = \mathbf{w}(\mathbf{m})^\top \text{diag}(\mathbf{1} - \mathbf{m})\mathbf{x}.$$

Observe that  $\langle \mathbf{w}(\mathbf{m}), \mathbf{x} \rangle_{\mathbf{m}}$  is the usual inner product of  $\mathbf{w}(\mathbf{m}) \in \mathbb{R}^d$  with  $\text{diag}(\mathbf{1} - \mathbf{m})\mathbf{x} \in \mathbb{R}^d$ , which is obtained by replacing the missing entries of  $\mathbf{x}$  by 0. With this notation,  $f(\mathbf{o}(\mathbf{x}, \mathbf{m}), \mathbf{m}) = \langle \mathbf{w}(\mathbf{m}), \mathbf{x} \rangle_{\mathbf{m}}$ . To find the linear model  $\mathbf{w} : \{0, 1\}^d \mapsto \mathbb{R}^d$ , we then solve

$$\min_{\mathbf{w}(\cdot)} \sum_{i=1}^n \ell(y_i, \langle \mathbf{w}(\mathbf{m}_i), \mathbf{x}_i \rangle_{\mathbf{m}_i}) + R(\mathbf{w}(\cdot)). \quad (6)$$

#### 4.2. A hierarchy of adaptive regression models

We now derive a hierarchy of linear models, according to how  $\mathbf{w}(\mathbf{m})$  depends on the missingness pattern, borrowing ideas from the multi-stage adaptive optimization literature.

*Fully adaptive regression.* Formulation (6) considers a different linear model for all potential missingness patterns. For a given pattern  $\mathbf{m} \in \{0, 1\}^d$ , the model  $\mathbf{w}(\mathbf{m})$  is obtained by solving

$$\min_{\mathbf{w}} \sum_{\substack{i=1 \\ \mathbf{m}_i = \mathbf{m}}}^n \ell(y_i, \langle \mathbf{w}, \mathbf{x}_i \rangle_{\mathbf{m}}) + R(\mathbf{w}).$$

This approach has two major drawbacks: first, it treats each missingness pattern separately, hence substantially reducing the number of observations available to fit each model. Second, there can be as many as  $2^d$  potential missingness patterns, i.e.,  $2^d$  models to be trained, rendering the fully adaptive approach potentially intractable. Yet, in practice, only a fraction of the  $d$  covariates might be subject to missingness and the actual number of patterns to consider can be substantially smaller.

*Static regression.* On the opposite side of the spectrum, we can consider a static model  $\mathbf{w}(\mathbf{m}) = \mathbf{w}$  which does not depend on  $\mathbf{m}$  and solve

$$\min_{\mathbf{w}} \sum_{i=1}^n \ell(y_i, \langle \mathbf{w}, \mathbf{x}_i \rangle_{\mathbf{m}_i}) + R(\mathbf{w}).$$

This is equivalent to fitting a linear model on a full data set where missing values are replaced by 0. Note that this is not equivalent to applying mean-impute on the data before training a linear model. By imputing missing values with a 0, a missing feature effectively does not contribute to the output of the model, while any non-zero value would have affected the final prediction.

*Affinely adaptive regression.* Affine policies are a successful and often-used tool in adaptive optimization, as they are typically more powerful than a static policy but more tractable than a fully adaptive one. An affinely adaptive linear model takes the form  $\mathbf{w}^{\text{affine}}(\mathbf{m}) = \mathbf{w}_0 + \mathbf{W}\mathbf{m}$ , where  $\mathbf{w}_0$  is a static policy, and  $\mathbf{W}_{ij}$  represents a correction to apply to the coefficient  $w_{0i}$  whenever feature  $j$  is missing. For a given point  $(\mathbf{o}(\mathbf{x}, \mathbf{m}), \mathbf{m})$ , a prediction is obtained by computing  $\langle \mathbf{w}_0 + \mathbf{W}\mathbf{m}, \mathbf{x} \rangle_{\mathbf{m}} = \sum_j w_{0j}(1 - m_j)x_j + \sum_{j,k} W_{jk}m_k(1 - m_j)x_j$ . By again replacing missing values with a zero, this model is equivalent to a linear regression model on  $\mathcal{O}(d^2)$  variables and can be computed by solving:

$$\min_{\mathbf{w}_0, \mathbf{W}} \sum_{i=1}^n \ell(y_i, \langle \mathbf{w}_0 + \mathbf{W}\mathbf{m}_i, \mathbf{x}_i \rangle_{\mathbf{m}_i}) + R(\mathbf{w}_0, \mathbf{W}).$$

Note, that if we were to consider adaptability on the intercept term only, i.e., a feature  $x_j$  that is constant equal to 1 and never missing ( $m_j = 0$ ), it would be equivalent to a linear regression model on the variables  $(\text{diag}(\mathbf{1} - \mathbf{m})\mathbf{x}, \mathbf{m})$ .

*Finitely adaptive regression.* Another way to balance the tradeoff between expressiveness of the adaptive model and tractability is finite adaptability. Formally, let  $\mathcal{M} = \{\mathbf{m} \in \{0, 1\}^d : \exists i, \mathbf{m} = \mathbf{m}_i\} \subseteq \{0, 1\}^d$  designate the set of unique missingness patterns in the data. We can partition  $\mathcal{M}$  into  $Q$  disjoint subsets  $\{\mathcal{M}_q\}_{q=1}^Q$  such that  $\mathcal{M} = \cup_{q=1}^Q \mathcal{M}_q$ , then train a distinct  $\mathbf{w}_q$  for each subset  $\mathcal{M}_q$ , by solving  $Q$  optimization problems:

$$\min_{\mathbf{w}_q} \sum_{i \in \mathcal{I}_q} \ell(y_i, \langle \mathbf{w}_q, \mathbf{x}_i \rangle_{\mathbf{m}_i}) + R(\mathbf{w}_q),$$

where  $\mathcal{I}_q = \{1 \leq i \leq n : \mathbf{m}_i \in \mathcal{M}_q\}$  indexes data points with a missingness pattern in  $\mathcal{M}_q$ .

The main difficulty of a finitely adaptive approach is in choosing the partition  $\{\mathcal{M}_q\}_{q=1}^Q$ . It may arise naturally from the data: for example, if the number of distinct missingness patterns  $|\mathcal{M}|$  is small enough (e.g., less than 20), we can set  $Q = |\mathcal{M}|$  and cover  $\mathcal{M}$  with singletons. In this case, the finitely adaptive strategy is equivalent to the fully adaptive one. More generally, we propose an iterative method to simultaneously learn the partition of missingness patterns and the appropriate regression models (Algorithm 1).

We begin with the set of all potential patterns  $\mathcal{M}$ , and an initial regression model  $\mathbf{w}$ , then update the subsets and adaptive regression models as follows. Consider one particular subset, denoted  $\mathcal{M}_q$  without loss of generality. Any feature index  $j$  can be used to further divide  $\mathcal{M}_q$  into  $\mathcal{M}_q^{j0} = \{\mathbf{m} \in \mathcal{M}_q : m_j = 0\}$  and  $\mathcal{M}_q^{j1} = \{\mathbf{m} \in \mathcal{M}_q : m_j = 1\}$ . Correspondingly



$j$  splits the data points  $\mathcal{I}_q$  into subsets  $\mathcal{I}_q^{j0}$  and  $\mathcal{I}_q^{j1}$ . We assume that neither  $\mathcal{I}_q^{j0}$  nor  $\mathcal{I}_q^{j1}$  are empty. Hence, we can train regression models  $\mathbf{w}_q^{j0}$  and  $\mathbf{w}_q^{j1}$  on each subset. Finally, we decide to split on the feature  $j$  which minimizes in-sample prediction error

$$\sum_{i \in \mathcal{I}_q^{j0}} \ell(y_i, \langle \mathbf{w}_q^{j0}, \mathbf{x}_i \rangle_{\mathbf{m}_i}) + \sum_{i \in \mathcal{I}_q^{j1}} \ell(y_i, \langle \mathbf{w}_q^{j1}, \mathbf{x}_i \rangle_{\mathbf{m}_i}),$$

To prevent overfitting, we can add different stopping criteria that make further splitting inadmissible. For instance, we can forbid splits that partition the data into subsets  $\mathcal{I}_q^{j0}$  and  $\mathcal{I}_q^{j1}$  that are too small (minimum bucket size), bound the total depth of the resulting partitioning tree, or split according to a feature  $j$  only if the improvement in cost/error is substantial enough compared to using  $\mathbf{w}_q$  on the entire data (cost improvement tolerance).

---

**Algorithm 1:** Iterative procedure for finitely adaptive regression

---

**Result:** Partition  $\mathcal{P} = \{\mathcal{M}_q, q \in [Q]\}$  and models  $\{\mathbf{w}^q, q \in [Q]\}$ .

initialization:  $\mathcal{P} = \{\mathcal{M}_1 = \mathcal{M}\}$ ;

**for**  $\mathcal{M}_q \in \mathcal{P}$  **do**

$j^* \leftarrow \arg \min_{(j,t)} \sum_{i \in \mathcal{I}_q^{j0}} \ell(y_i, \langle \mathbf{w}_q^{j0}, \mathbf{x}_i \rangle_{\mathbf{m}_i}) + \sum_{i \in \mathcal{I}_q^{j1}} \ell(y_i, \langle \mathbf{w}_q^{j1}, \mathbf{x}_i \rangle_{\mathbf{m}_i}), ;$

**if** *stopping criterion is not met* **then**

split  $\mathcal{M}_q$  along  $j^*$ :  $\mathcal{P} \leftarrow (\mathcal{P} \setminus \{\mathcal{M}_q\}) \cup \{\mathcal{M}_q^{j*0}, \mathcal{M}_q^{j*1}\}$

**end**

**end**

---

### 4.3. Connection between adaptive regression and optimal impute-then-regress

In this section, we demonstrate that some adaptive regression models can be viewed as impute-then-regress strategies where the imputation and the regression models are designed simultaneously.

We consider a special case of affinely adaptive regression where all the regression coefficients are static except for the intercept, which depends on the missingness pattern, i.e.,  $f(\mathbf{o}(\mathbf{x}, \mathbf{m}), \mathbf{m}) = b(\mathbf{m}) + \langle \mathbf{w}, \mathbf{x} \rangle_{\mathbf{m}}$ , where  $b(\mathbf{m})$  is an adaptive intercept term. When the intercept is affinely adaptive, i.e.,  $b(\mathbf{m}) = b_0 + \sum_j b_j m_j$ , the prediction function is

$$f(\mathbf{o}(\mathbf{x}, \mathbf{m}), \mathbf{m}) = b_0 + \sum_{j=1}^d (w_j(1 - m_j)x_j + b_j m_j) = b_0 + \sum_{j=1}^d w_j \left( (1 - m_j)x_j + m_j \frac{b_j}{w_j} \right).$$

In other words, a static regression model with affinely adaptive intercept can be viewed as imputing  $\mu_j := b_j/w_j$  for feature  $j$  whenever it is missing, and then applying a linear model  $\mathbf{w}$ . The key difference with standard impute-then-regress strategies, however, is that the vector of imputed values  $\boldsymbol{\mu}$  and the linear model  $\mathbf{w}$  are computed simultaneously, instead of sequentially, hence leading to greater predictive power<sup>1</sup>. In the simple case where there is only one feature  $X_1$ , this family of models would learn the rule  $w_1(X_1(1 - M_1) + \mu_1 M_1)$  with

$$\mu_1 = \frac{\mathbb{E}[Y|M_1 = 1]}{\text{cor}(Y, X_1|M_1 = 0)}.$$

Compared to classical imputation methods, we observe that the imputed value does not only depend on the distribution of  $X_1$  on the samples where it is observed ( $M_1 = 0$ ). Rather, (a) it depends on the target variable  $Y$ , and (b) it involves observations for which  $X_1$  is missing ( $M_1 = 1$ ). In particular, if  $Y$  satisfies a linear relationship  $Y = w_1^* X_1$ , then  $\mu_1 = \mathbb{E}[X_1|M_1 = 1]$ . In contrast, standard mean-imputation would select  $\mu_1 = \mathbb{E}[X_1|M_1 = 0]$ . In the same vein, we can view more sophisticated imputation rules as more complex adaptive rules for the intercept.

Finally, we observe that static regression models with an affinely adaptive intercept are very easy to compute in practice. They correspond to a simple linear regression model over  $2d$  variables, the  $d$  coordinates of  $\mathbf{x}$  (with missing values imputed as 0) along with the  $d$  coordinates of  $\mathbf{m}$ , as well as an intercept term  $b_0$ . Accordingly, in the following numerical experiments, we consider these models as our “static” benchmarks.

## 5. Numerical experiments

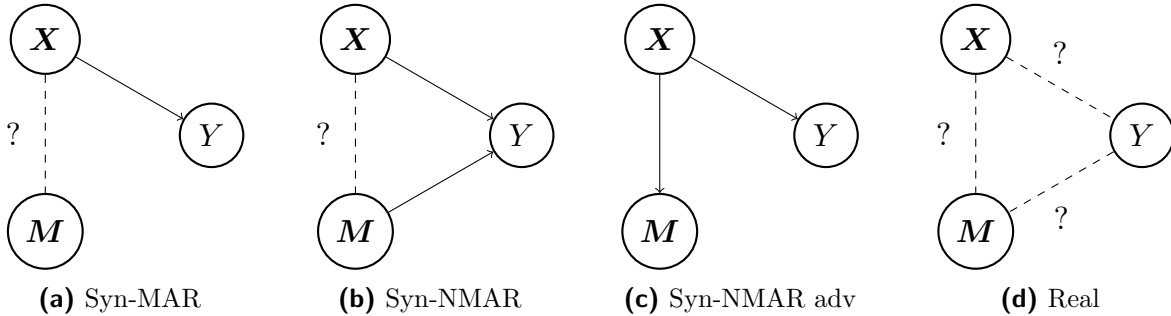
In this section, we assemble a corpus of 71 publicly available data sets with missing data, from the UCI Machine Learning Repository and the RDatasets Repository<sup>2</sup> with two objectives in mind: first, confirm the practical implications of Theorem 1 regarding the benefit of mean and mode imputation; second, evaluate the empirical performance of our adaptive regression framework.

<sup>1</sup> Note that our algebraic manipulation, and the resulting interpretation, is valid only if  $w_j \neq 0$ . If  $w_j = 0$  and  $b_j \neq 0$ , it means that feature  $j$  is not a strong predictor of the outcome variable  $y$  but the fact that it is missing,  $m_j$ , is.

<sup>2</sup> <https://archive.ics.uci.edu> and <https://github.com/vincentarelbundock/Rdatasets>

### 5.1. Methodology

We consider four experimental settings, with both synthetic and real signals  $Y$ . They differ in the relationships between the missingness pattern  $M$ , the design matrix  $X$  and the signal  $Y$  as depicted on Figure 1.



**Figure 1** Graphical representation of the 4 experimental designs implemented in our benchmark simulations. Solid (resp. dashed) lines correspond to correlations explicitly (resp. not explicitly) controlled in our experiments.

*Synthetic signal.* Given a data set with  $n$  observations  $(\mathbf{o}(\mathbf{x}^{(i)}, \mathbf{m}^{(i)}), \mathbf{m}^{(i)})$ , we first generate a fully observed version of the data by performing missing data imputation using the R package `missForest` (Stekhoven and Bühlmann 2012), obtaining a new data set  $\{(\mathbf{x}_{\text{full}}^{(i)}, \mathbf{m}^{(i)})\}_{i \in [n]}$ .

In the first setting, referred to as “Syn-MAR”, we synthesize  $Y$  according to the rule  $y^{(i)} = \mathbf{w}^\top \mathbf{x}_{\text{full}}^{(i)} + \epsilon^{(i)}$ . Here, we take a random vector  $\mathbf{w}$  with only  $k = 10$  non-zero entries distributed uniformly over  $[-1, 1]$ . We control  $k_{\text{missing}}$ , the number of features in the support of  $\mathbf{w}$  that are sometimes missing in the data. We independently sample noise  $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$  where  $\sigma^2$  controls the signal-to-noise ratio (we take  $\sigma^2$  so that  $SNR = 2$  in our experiments). Hence, the resulting response  $Y$  depends directly on the covariates  $X$  but not on the missingness pattern  $M$ . However, we do not control the correlation between  $X$  and  $M$  for two reasons: First, they both come from a real-world dataset which might not satisfy the MAR assumption. Second, as previously observed (Remark 2), imputation does induce some correlation between the imputed data set  $X_{\text{full}}$  and  $M$ .

The second setting resembles the first one except that  $Y$  is now a function of both  $X$  and  $M$ :  $y^{(i)} = \mathbf{w}_x^\top \mathbf{x}_{\text{full}}^{(i)} + \mathbf{w}_m^\top \mathbf{m}^{(i)} + \epsilon^{(i)}$ , where  $\mathbf{w}_x$  and  $\mathbf{w}_m$  are  $k$ - and  $k_{\text{missing}}$ -sparse respectively. Consequently, we refer to this setting as “Syn-NMAR”.

For the third synthetic setting, “Syn-NMAR adv”, we follow the same methodology as for “Syn-MAR” and obtain observations  $(\mathbf{x}_{\text{full}}^{(i)}, \mathbf{m}^{(i)}, y^{(i)})$ ,  $i \in [n]$ . Then, we reallocate the missingness patterns across observations so as to ensure the data is NMAR. Formally, we consider the observations  $(\mathbf{o}(\mathbf{x}_{\text{full}}^{(i)}, \mathbf{m}^{(\sigma_i)}), \mathbf{m}^{(\sigma_i)}, y^{(\sigma_i)})$ ,  $i \in [n]$ , where  $\sigma$  is the permutation maximizing the total sum of missing values  $\sum_{i=1}^n \mathbf{x}_{\text{full}}^{(i)\top} \mathbf{m}^{(\sigma_i)}$ . Hence,  $Y$  and  $\mathbf{M}$  explicitly depend on  $\mathbf{X}$ .

*Real signal.* Finally, we also consider the “real” signals  $y^{(i)}$  when indicated in the data description (i.e., for 52 of the 71 datasets). It can be binary, in which case we adapt the loss function accordingly.

## 5.2. Theorem 1 in practice

*Numerical performance of mode imputation.* For missing categorical variables, Theorem 1 recommends encoding missingness as an attribute over mode imputation. Accordingly, we numerically compare the out-of-sample performance of a regularized linear regression model (Zou and Hastie 2005) with these two approaches.

Our analysis comprises 44 datasets with at least one missing categorical feature (missing numerical features are mean-imputed). In synthetic-signal experiments, we generate several targets  $Y$ , setting the number of features in the signal  $k$  to 10 (or all available features if fewer) and varying the number of missing features in the signal  $k_{\text{missing}}$  from 0 to  $k$ . For each experiment, we use 70% of the observations to train a linear regression model, cross-validating regularization parameters, and evaluate the out-of-sample  $R^2$  on the remaining 30%. We repeat this procedure 10 times, on 10 different train/test splits obtained via stratified sampling on the missing data patterns. We compare the out-of-sample performance of the models obtained by encoding missingness as a new category with the ones obtained after mode imputation.

To quantitatively assess the effect of mode imputation, we regress out-of-sample accuracy onto a binary variable encoding whether the model was trained after mode imputation or not. Control variables include a dataset fixed effect and  $k_{\text{missing}}$  (for synthetic  $Y$ ’s). Output from this regression analysis is reported in Table 1. As we can see, mode imputation has a negative (detrimental) effect on predictive power. It reduces out-of-sample  $R^2$  by roughly 0.05 (in absolute terms) on the synthetic examples ( $p$ -value  $< 0.01$ ). We do not observe a significant effect on the real  $Y$  examples, suggesting other factors beyond missing data might impact the validity of Theorem 1 in practice.

**Table 1** Regression output for predicting the out-of-accuracy ( $R^2$  or AUC) based on whether mode imputation was used. We include dataset and  $k_{missing}$  fixed effects. We report regression coefficient values (and clustered standard errors).

Setting	Mode Imputation coefficient	$p$ -value	Adjusted $R^2$
Syn-MAR	-0.0470 (0.013)	0.0005	0.3871
Syn-NMAR	-0.0475 (0.0136)	0.0005	0.5282
Syn-NMAR adv	-0.0460 (0.0128)	0.0003	0.3692
Real	0.0011 (0.0032)	0.7345	0.9141

Controls: Dataset,  $k_{missing}$  (if available)

**Table 2** Difference in means/median and one-sided  $p$ -values from a  $t$  and Wilcoxon test applied to assess the negative impact of mode imputation on downstream accuracy.

Setting	$\Delta$ mean ( $p$ -value)	$\Delta$ median ( $p$ -value)
Syn-MAR	-0.0470 ( $< 2 \cdot 10^{-16}$ )	-0.0157 ( $< 2 \cdot 10^{-16}$ )
Syn-NMAR	-0.0475 ( $< 2 \cdot 10^{-16}$ )	-0.0174 ( $< 2 \cdot 10^{-16}$ )
Syn-NMAR adv	-0.0460 ( $< 2 \cdot 10^{-16}$ )	-0.0121 ( $< 2 \cdot 10^{-16}$ )
Real	0.0011 (0.5872)	0.0000 (0.4636)

For each dataset and each random training/test split, we can also directly compare the out-of-sample  $R^2$  of the models obtained with missingness as a category and mode imputation respectively and rigorously test whether mode imputation achieves lower out-of-sample  $R^2$ . Results from a paired  $t$ -test (difference in means) and paired Wilcoxon test (difference in medians) are reported in Table 2 and corroborate the findings from the regression analysis.

*Comparing impute-then-regress methods.* Theoretically, mean imputation can lead to Bayes-consistent predictors. In addition, how to implement impute-then-regress methods in practice, especially out-of-sample imputation, remains an open question. Here, we consider and compare four variants of impute-then-regress:

- (V1) In the first variant, we simultaneously impute the train and test set using R package `mice` (van Buuren and Groothuis-Oudshoorn 2010). This implementation has the advantage of jointly imputing the train and test sets. However, it cannot be applied in practice since the test set, i.e., future observations, are unavailable during calibration.

- (V2) Secondly, we use `mice` to impute the train set alone, and then impute the test set with the *imputed* training data. Since the test set is imputed with knowledge of the previously imputed train set, we can hope that the imputation on test data will mimic the strategy used in training, despite the black-box nature of the imputation method.
- (V3) As a third option, we impute the train set alone with `mice`, and then the test set with the *original* train set. We intuit that (V3) will be less powerful than (V2) since the rules learned for imputing the test set might differ from the ones used for the training set.
- (V4) Finally, we consider mean-imputation, where the mean is estimated on the observed training data only. This implementation is light in storage requirements, fast, and theoretically optimal (see Theorem 1).

Note that for (V1-3) the imputation method, `mice` is one of the state-of-the-art methods for missing data imputation, while being less powerful than the method we used to generate the ground truth, `missForest`.

For this analysis, we select datasets in our corpus with at least one missing numerical feature (59 datasets in total) and conduct the same analysis as in the previous paragraph with two objectives in mind: First, confirm empirically that mean imputation competes with the other imputation methods as far as the downstream prediction task is concerned. Second, demonstrate that (V3) is less powerful than (V2).

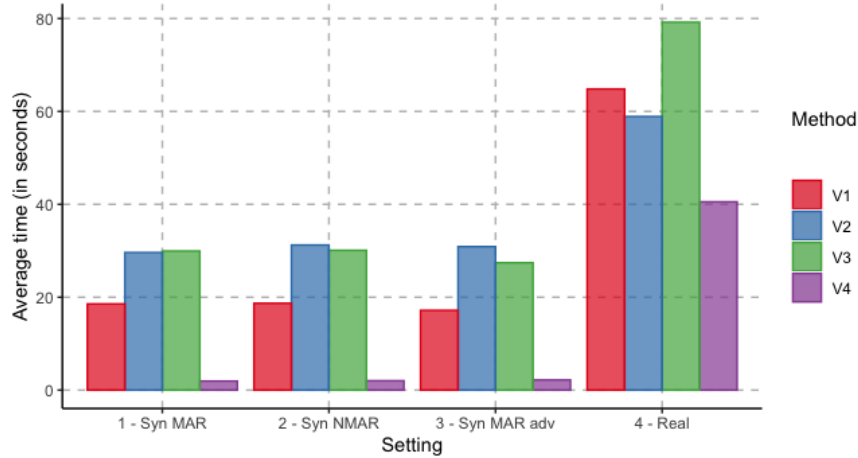
We conduct a regression analysis (Table 3), to assess the relative benefit of using (V1-3) over mean imputation (V4). We observe that no other impute-then-regress method significantly outperforms mean-impute. On the contrary, mean imputation is very competitive; It improves upon all methods in the “Syn-MAR” and “Syn-NMAR” settings, and displays no significant difference in accuracy in other settings, despite its simplicity and ease of implementation. As depicted on Figure 2, it requires significantly less computational effort than all other methods. We confirm these findings by directly comparing (V1-3) with (V4) in paired *t*- and Wilcoxon tests (see Table 5 in Appendix B).

Finally, to confirm our theoretical insight that imputing the test set with the imputed train set (V2) should have an edge over imputing the test set with the original train (V3), we directly compare these two variants and report the results from a paired *t*- and Wilcoxon test in Table 4, which corroborate the validity of these findings.

**Table 3** Regression output for predicting the out-of-accuracy ( $R^2$  or AUC) based on the impute-then-regress method. We include dataset and  $k_{missing}$  fixed effects. We report regression coefficient values (and clustered standard errors).

Setting	V1 vs. V4 coeff.	V2 vs. V4 coeff.	V3 vs. V4 coeff.	Adjusted $R^2$
Syn-MAR	-0.0183 (0.0047)***	-0.0231 (0.0050)***	-0.0253 (0.0057)***	0.2137
Syn-NMAR	-0.0189 (0.0050)***	-0.0161 (0.0053)**	-0.0248 (0.0066)***	0.2991
Syn-NMAR adv	0.0042 (0.0053)	0.0051 (0.0050)	0.0003 (0.0065)	0.4309
Real	0.0001 (0.0018)	-0.0018 (0.0019)	-0.0023 (0.0022)	0.9457

Controls: Dataset,  $k_{missing}$  (if available);  $p$ -value: \*: < 0.1, \*\*: < 0.01, \*\*\*: < 0.001



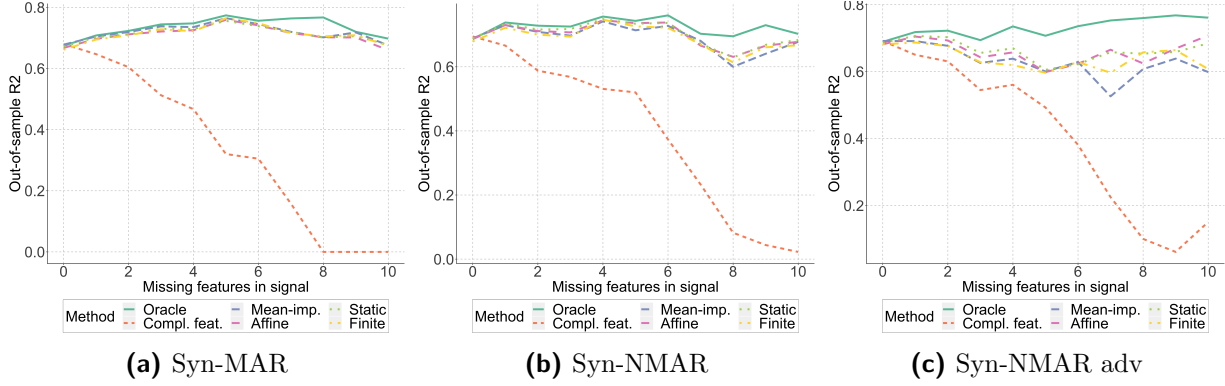
**Figure 2** Average computational time of each impute-then-regress variant.

**Table 4** Difference in means/median and one-sided  $p$ -values from a  $t$  and Wilcoxon test applied to assess the positive impact on downstream accuracy of (V2) over (V3).

Setting	$\Delta$ mean ( $p$ -value)	$\Delta$ median ( $p$ -value)
Syn-MAR	0.0022 (0.2959)	0.0006 (0.0003)
Syn-NMAR	0.0087 ( $< 10^{-11}$ )	0.0010 ( $< 10^{-7}$ )
Syn-NMAR adv	0.0048 ( $< 10^{-5}$ )	0.0013 ( $< 10^{-12}$ )
Real	0.0005 (0.8459)	0.0004 (0.2685)

### 5.3. Performance of adaptive linear regression

Finally, we compare the adaptive regression models proposed in Section 4 (static with affine intercept, affine and finite), with an impute-then-regress model with mean imputation. As benchmarks, we also compare to an “oracle” model, which has access to the fully observed



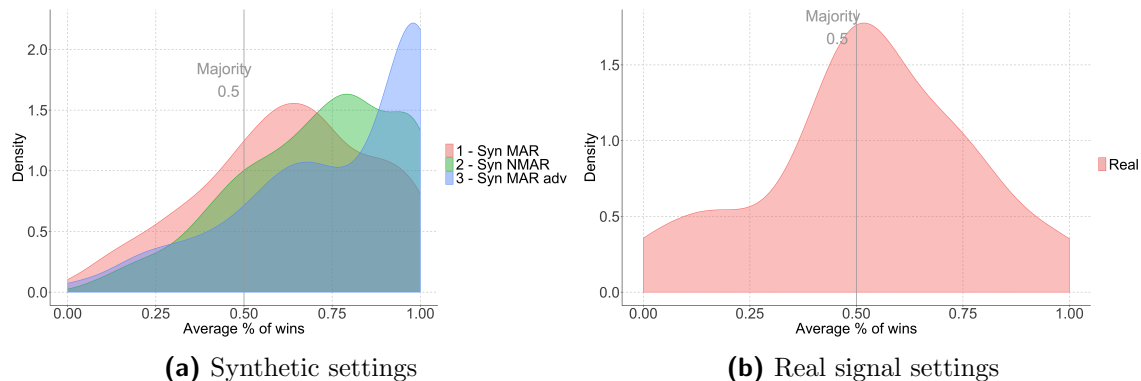
**Figure 3** Comparison of adaptive regression methods vs. impute-then-regress on synthetic data, as  $k_{missing}$  increases. We also compare to two extremes: the “Oracle” which has access to the fully observed data and the “Complete Feature” regression which regresses of features that are never missing only.

data set  $\mathbf{x}_{full}$  and “complete feature” regression, i.e., using only features that are never missing. In line with our previous findings, we consider the same datasets as in the previous section, with missing categorical variables encoded as a new category.

Figure 3 compares the out-of-sample  $R^2$  of the five approaches (the two benchmarks, mean-impute-then-regress and the three adaptive methods) for the three synthetic settings as  $k_{missing}$  increases. First, we observe that “complete feature” regression is dominated by all approaches and its performance quickly degrades as  $k_{missing}$  increases, suggesting that excluding a feature based on the fact that it can be missing strongly hinders predictive power. Note that this conclusion holds despite the fact that features are correlated. Second, under MAR, all other methods, namely mean-impute-then-regress and adaptive regression models, perform almost as well as the oracle. The gap between their performance and the oracle’s widens as the number of missing features contributing to the signal  $k_{missing}$  increases and as the missingness mechanism departs from the MAR assumption. In this regard, while adaptive regression and mean-imputation are comparable in the MAR setting, the edge of adaptive models grows in settings where data is not missing at random, since these models are able to leverage information in the missingness patterns directly.

To verify this finding, for each experiment, i.e., for each setting, dataset and  $k_{missing}$  value, we count how often one adaptive model outperforms (out-of-sample) an impute-then-regress method (we implement both (V2) and (V4) variants in this experiment). Figure 4a represents the density of this percentage of “wins” across all data sets and for the three synthetic settings. The trend is clear: as the missingness mechanisms departs further





**Figure 4** Frequency of “wins” from adaptive models over impute-then-regress ones.

away from the MAR assumption, adaptive models improve more often over impute-then-regress. On real-world signals (Figure 4b), we observe a slimer edge of adaptive models over impute-then-regress strategies, a difference which could be explained by the fact that we restricted our attentions to linear models.

## 6. Conclusion

Causal inference and prediction are two distinct but intertwined functions of statistical models (Shmueli et al. 2010). In inference settings, data imputation, in particular multiple data imputation, is the gold standard for dealing with missing information. In this paper, we questioned its relevance for predictive tasks. From a theoretical perspective, we prove that impute-then-regress strategies are consistent if the data can surely be de-imputed (Corollary 1), the intuition being that missingness carries potentially predictive information which is concealed by complex imputation rules. On the contrary, practical predictive models should be trained and applicable directly on partially observed data and leverage the fact that some features are missing. To this end, we develop adaptive linear regression models, and demonstrate their relevance on real-world numerical examples. More broadly, we believe that making machine learning models adaptive to the set of features available constitutes an exciting direction for future work.

## References

- D. Bertsimas, C. Pawlowski, and Y. D. Zhuo. From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 18(196):1–39, 2018. URL <http://jmlr.org/papers/v18/17-073.html>.
- T. H. Bø, B. Dysvik, and I. Jonassen. Lsimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic acids research*, 32(3):e34–e34, 2004.

- O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.
- L. P. Brás and J. C. Menezes. Improving cluster-based missing value estimation of dna microarray data. *Biomolecular engineering*, 24(2):273–282, 2007.
- L. F. Burgette and J. P. Reiter. Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, 172(9):1070–1076, 2010.
- Z. Cai, M. Heydari, and G. Lin. Iterated local least squares microarray missing value imputation. *Journal of bioinformatics and computational biology*, 4(05):935–957, 2006.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
- F. Husson, J. Josse, B. Narasimhan, and G. Robin. Imputation of mixed data with multilevel singular value decomposition. *Journal of Computational and Graphical Statistics*, 28(3):552–566, 2019.
- J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.
- W. Jiang, J. Josse, M. Lavielle, and T. Group. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145: 106907, 2020.
- J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values. feb 2019.
- H. Kim, G. H. Golub, and H. Park. Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198, 2005.
- R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. Wiley & Sons, 2019.
- T. A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233, 1982.
- X.-L. Meng and D. B. Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991.
- S. Mohamed, Z. Ghahramani, and K. A. Heller. Bayesian exponential family pca. In *Advances in neural information processing systems*, pages 1089–1096, 2009.
- T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, P. Solenberger, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.
- D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. Wiley & Sons, 1987.
- S. Seaman, J. Galati, D. Jackson, and J. Carlin. What is meant by” missing at random”? *Statistical Science*, pages 257–268, 2013.
- G. Shmueli et al. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.
- D. J. Stekhoven and P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

- O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.
- B. Twala, M. Jones, and D. J. Hand. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.
- S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.
- M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- X. Wang, A. Li, Z. Jiang, and H. Feng. Missing value estimation for dna microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC bioinformatics*, 7(1):32, 2006.
- J. Yoon, J. Jordon, and M. Van Der Schaar. Gain: Missing data imputation using generative adversarial nets. *arXiv preprint arXiv:1806.02920*, 2018.
- X. Zhang, X. Song, H. Wang, and H. Zhang. Sequential local least squares imputation estimating missing value of microarray data. *Computers in biology and medicine*, 38(10):1112–1120, 2008.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

## Appendix A: Proof of Theorem 3.1

*Proof* Following the approach laid out in Josse et al. (2019), We elicit the value of the predictor  $f_{\mu\text{-impute}}(\mathbf{x})$  by partitioning the training data set ( $\mu$ -imputed). We consider some imputed vector  $\mathbf{x}$  such that  $\mu$  is continuous at  $\mathbf{x}_{2:d}$ . For concision, we denote  $\mu = \mu(\mathbf{x}_{2:d})$ .

**Case 1:** If  $x_1 \neq \mu$ , then, by continuity of  $\mu$ , we can construct a ball  $B(\mathbf{x}, h)$  centered around  $\mathbf{x}$  and of radius  $h$  that does not contain  $\mu$ . As a result,  $\mathbb{E}[Y|\mathbf{X} \in B(\mathbf{x}, h)] = \mathbb{E}[Y|\mathbf{X} \in B(\mathbf{x}, h), M_1 = 0]$ , and we take the limit as  $h \rightarrow 0$ .

**Case 2:** If  $x_1 = \mu$  and  $\eta(\mathbf{x}) = \mathbb{P}(M_1 = 1|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}) = 0$ , then  $\{\mathbf{X} = \mathbf{x}\} = \{\mathbf{X} = \mathbf{x}, M_1 = 0\}$  with probability 1.

**Case 3:** If  $x_1 = \mu$  and  $\eta(\mathbf{x}) = \mathbb{P}(M_1 = 1|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}) = \eta > 0$ , we separate the discrete and continuous cases.

*Discrete case.* For any ball  $B(\mathbf{x}, h)$  small enough, conditioning on the event  $\{\mathbf{X} \in B(\mathbf{x}, h)\}$  is equivalent to conditioning on  $\{X_1 = \mu, \mathbf{X}_{2:d} \in B(\mathbf{x}_{2:d}, h)\}$  since  $X_1$  only takes discrete values. Conditioning on the value of  $\mathbf{X}_{2:d}$ , we have

$$\begin{aligned} \mathbb{E}[Y|X_1 = \mu, \mathbf{X}_{2:d}] &= \frac{\mathbb{E}[Y\mathbb{I}(X_1 = \mu, M_1 = 1)|\mathbf{X}_{2:d}] + \mathbb{E}[Y\mathbb{I}(X_1 = \mu, M_1 = 0)|\mathbf{X}_{2:d}]}{\mathbb{P}(M_1 = 1, X_1 = \mu|\mathbf{X}_{2:d}) + \mathbb{P}(M_1 = 0, X_1 = \mu|\mathbf{X}_{2:d})} \\ &= \frac{\mathbb{E}[Y\mathbb{I}(M_1 = 1)|\mathbf{X}_{2:d}] + \mathbb{E}[Y\mathbb{I}(X_1 = \mu)\mathbb{I}(M_1 = 0)|\mathbf{X}_{2:d}]}{\mathbb{P}(M_1 = 1|\mathbf{X}_{2:d}) + \mathbb{P}(M_1 = 0|\mathbf{X}_{2:d})\mathbb{P}(X_1 = \mu|\mathbf{X}_{2:d})}, \end{aligned}$$

where the equality follows from the conditional independence assumption,  $X_1 \perp\!\!\!\perp M_1|\mathbf{X}_{2:d}$ . Let us denote  $p_\mu = \mathbb{P}(X_1 = \mu|\mathbf{X}_{2:d} = \mathbf{x}_{2:d})$ . By taking the expectation over  $\mathbf{X}_{2:d} \in B(\mathbf{x}_{2:d}, h)$  and the limit as  $h \rightarrow 0$ , the order of which can be switched by dominated convergence, we obtain the desired result.

*Continuous case.* With a similar line of argument, for any ball  $B(\mathbf{x}, h)$ ,

$$\mathbb{E}[Y|\mathbf{X} \in B(\mathbf{x}, h)] = \mathbb{E}[Y|\mathbf{X}_{2:d} \in B(\mathbf{x}_{2:d}, h), M_1 = 1] T_1(h) + T_2(h),$$

where we denote

$$T_1(h) = \frac{\mathbb{P}(\mathbf{X}_{2:d} \in B(\mathbf{x}_{2:d}, h), M_1 = 1)}{\mathbb{P}(\mathbf{X}_{2:d} \in B(\mathbf{x}_{2:d}, h), M_1 = 1) + \mathbb{P}(\mathbf{X} \in B(\mathbf{x}, h), M_1 = 0)},$$

$$T_2(h) = \frac{\mathbb{E}[Y\mathbb{I}(\mathbf{X} \in B(\mathbf{x}, h), M_1 = 0)]}{\mathbb{P}(\mathbf{X}_{2:d} \in B(\mathbf{x}_{2:d}, h), M_1 = 1) + \mathbb{P}(\mathbf{X} \in B(\mathbf{x}, h), M_1 = 0)}.$$

Let us denote  $\text{vol}(h, d)$  the volume of the ball  $B(\mathbf{x}, h)$ . Observe that  $\text{vol}(h, d)$  does not depend on  $\mathbf{x}$  and that  $\text{vol}(h, d)$  is proportional to  $h^d$ . Since  $\mathbf{X} \in \mathbb{R}^d$  admits a continuous density, we can show that

$$\mathbb{P}(\mathbf{X} \in B(\mathbf{x}, h), M_1 = 0) \lesssim \text{vol}(h, d),$$

$$\mathbb{E}[Y\mathbb{I}(\mathbf{X} \in B(\mathbf{x}, h), M_1 = 0)] \lesssim \text{vol}(h, d),$$

$$\mathbb{P}(\mathbf{X}_{2:d} \in B(\mathbf{x}_{2:d}, h), M_1 = 1) \gtrsim \text{vol}(h, d - 1),$$

where the inequalities are given up to a multiplicative constant. This proves that  $T_1(h) \rightarrow 1$  and  $T_2(h) \rightarrow 0$  when  $h \rightarrow 0$ , so that

$$\mathbb{E}[Y|\mathbf{X} \in B(\mathbf{x}, h)] \rightarrow \mathbb{E}[Y|\mathbf{X}_{2:d} = \mathbf{x}_{2:d}, M_1 = 1].$$

## Appendix B: Additional numerical evidence

**Table 5** Difference in means/median and two-sided  $p$ -values from a  $t$  and Wilcoxon test applied to assess the difference in downstream accuracy between (V4) over all others impute-then-regress strategies.

Comparison	Setting	$\Delta$ mean ( $p$ -value)	$\Delta$ median ( $p$ -value)
(V1) - (V4)	Syn-MAR	-0.0183 ( $< 10^{-10}$ )	-0.0087 ( $< 10^{-16}$ )
	Syn-NMAR	-0.0189 ( $< 10^{-15}$ )	-0.0073 ( $< 10^{-16}$ )
	Syn-NMAR adv	0.0042 (0.0128)	-0.0009 ( $< 10^{-4}$ )
	Real	0.0001 ( $< 0.9843$ )	-0.0003 (0.5904)
(V2) - (V4)	Syn-MAR	-0.0231 ( $< 10^{-8}$ )	-0.0100 ( $< 10^{-16}$ )
	Syn-NMAR	-0.0161 ( $< 10^{-10}$ )	-0.0097 ( $< 10^{-16}$ )
	Syn-NMAR adv	0.0051 (0.0052)	-0.0006 (0.0021)
	Real	-0.0018 (0.4605)	0.0003 (0.6032)
(V3) - (V4)	Syn-MAR	-0.0253 ( $< 10^{-16}$ )	-0.0104 ( $< 10^{-16}$ )
	Syn-NMAR	-0.0248 ( $< 10^{-16}$ )	-0.0106 ( $< 10^{-16}$ )
	Syn-NMAR adv	0.0003 (0.8855)	-0.0016 ( $< 10^{-9}$ )
	Real	-0.0023 (0.4120)	-0.0009 (0.0577)