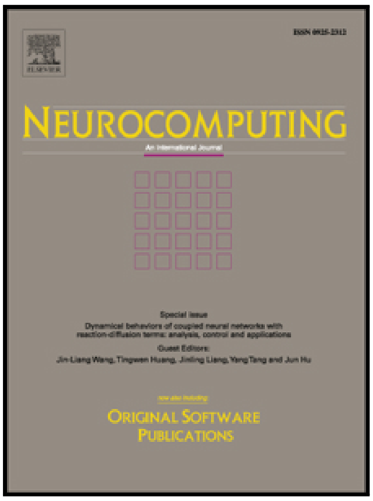

ISSN 0925-2312
ELSEVIER
NEUROCOMPUTING
An International Journal
Special issue
Dynamical behaviors of coupled neural networks with reaction-diffusion terms: analysis, control and applications
Guest Editors:
Jin-Liang Wang, Tingwen Huang, Jinling Liang, Yang Tang and Jun Hu
now also including
ORIGINAL SOFTWARE PUBLICATIONS

Biases in feature selection with missing data

Borja Seijo-Pardo, Amparo Alonso-Betanzos, Kristin P. Bennett,
Verónica Bolón-Canedo, Julie Josse, Mehreen Saeed,
Isabelle Guyon

PII: S0925-2312(19)30149-3
DOI: https://doi.org/10.1016/j.neucom.2018.10.085
Reference: NEUCOM 20420

To appear in: *Neurocomputing*

Received date: 16 July 2018
Revised date: 4 October 2018
Accepted date: 10 October 2018

Please cite this article as: Borja Seijo-Pardo, Amparo Alonso-Betanzos, Kristin P. Bennett, Verónica Bolón-Canedo, Julie Josse, Mehreen Saeed, Isabelle Guyon, Biases in feature selection with missing data, *Neurocomputing* (2019), doi: https://doi.org/10.1016/j.neucom.2018.10.085

Biases in feature selection with missing data

Borja Seijo-Pardo[1], Amparo Alonso-Betanzos[1], Kristin P. Bennett[2],
Verónica Bolón-Canedo[1], Julie Josse[3], Mehreen Saeed[4], Isabelle Guyon [5]

1 . CITIC, University of A Coruña, A Coruña 15006, Spain.
2. Rensselaer Polytechnic Institute, Troy 12180-3590, New York, USA.
3. CMAP, Ecole Polytechnique, Palaiseau 91128, France.
4. FAST, National University of Computer and Emerging Sciences, Punjab 54000, Lahore, Pakistan.
5. LRI, University Paris-Saclay, Orsay 91405, France.

Abstract

Feature selection is of great importance for two possible scenarios: (1) prediction, i.e. improving (or minimally degrading) the predictions of a target variable while discarding redundant or uninformative features and (2) discovery, i.e. identifying features that are truly dependent on the target and may be genuine causes to be determined in experimental verifications (for example for the task of drug target discovery in genomics). In both cases, if variables have a large number of missing values, imputing them may lead to false positives; features that are not associated with the target become dependent as a result of imputation. In the first scenario, this may not harm prediction, but in the second one, it will erroneously select irrelevant features.

In this paper, we study the risk/benefit trade-off of missing value imputation in the context of feature selection, using causal graphs to characterize when structural bias arises. Our aim is also to investigate situations in which imputing missing values may be beneficial to reduce false negatives, a situation that might arise when there is a dependency between feature and target, but the dependency is below the significance level when only complete cases are considered. However, the benefits of reducing false negatives must be balanced against the increased number of false positives.

In the case of binary target variable and continuous features, the t-test is often used for univariate feature selection. In this paper, we also introduce a de-biased version of the t-test allowing us to reap the benefits of imputation, while not incurring the penalty of increasing the number of false positives.

Keywords: `feature selection, missing data, de-biased t-test`

1. Introduction

Missing data is nowadays commonly encountered in many real-world datasets. In early studies, missing data has been treated as a secondary problem, and was dealt with standard simple procedures, such as replacing missing values by the variable median, eliminating them, or adding indicator variables of missingness [1, 2]. However, the problem of missing data is at the very core of machine learning, even when no data are missing. Indeed, target values in supervised learning problems are missing in test data, and if they are not missing at random (that is, missingness mechanism depends solely on variables with complete information)[3], this introduces a form of sample bias; collaborative filtering recommendation systems also present cases of missing data, occurring for at least two reasons: users generally do not evaluate all items, and new questions/items are often introduced in the course of the study.

The analysis and resolution of missing data problems have been the subject of extensive studies [2, 4, 5]. In these works, the authors have indicated basic and general notions for dealing with certain missing data situations and also on the principles used for the recovering of such missing data. Traditionally, discarding incomplete samples, or filling in missing values using median or average values of the features being missed, have been used to deal with the missing data problem. Methodological approaches [4] appeared in the 70's, providing Maximum Likelihood Estimation routines, Bayesian Estimation or Multiple Imputation. In particular, Judea Pearl et al. [6, 7] studied the recoverability and testability of missing data problems, using the causal theory to deal with the different possible missing mechanisms and simplifying the seminal theoretical work of Rubin et al. [4], in which most practices are based. Moreover, the authors deepened on this problem in another study [8], where they use “missingness graphs” with the aim of determining the recoverability of the problem at hand.

Missing data might occur in several parts of a study such as data sources, study designs, registration time, registration frequency, etc., and for different reasons, as failure to complete questionnaires, inaccurate data transfer, sensor failure, loss to follow-up, etc. Thus, it has been a common problem in almost any discipline, mainly in social, behavioral and medical sciences. Among those, however, clinical research is one of the main areas where the treatment of missing data takes a significant relevance. Therefore, several recent studies focus on that area. Enders [9] describes a number of practical issues that can be given by applying imputation on clinical problems, including mixtures of categorical and continuous variables, item-level missing data in questionnaires, significance testing, interaction effects and multilevel missing data. Pedersen et al. [10] study the effect of missing data in clinical epidemiological research, providing insights on the type of missing data, and recommending multiple imputation as the best way to obtain unbiased and valid estimates of associations based on information from the available data in this area. Missing data is problematic due to the risk of bias, which depends on the type of missingness, the relative size of the data that are missing, and the way of dealing with these missing values [11]. Several authors focus on carrying out different works to interpret and reduce the risk of bias in their respective fields. As examples, Nguyen et al. [12] studied the impact of missing data strategies in the parental employment and health areas, while Tomita et al. [13] proposed a new method for multiple imputation to obtain a consistent final estimate, tested on a real medical dataset.

In general, the difficulty of such problems varies depending on the nature of the “missingness mechanism” [1, 14, 15, 6]. In the simplest case values are Missing Completely At Random (MCAR), which occurs when the “missingness mechanism” is completely unrelated to any study variable. In large datasets plagued with MCAR missing data, samples with missing values can be discarded without biasing the distribution of the remaining data. A slightly more general and frequent case concerns data Missing At Random (MAR) for which the missingness mechanism depends solely on variables with complete information (*i.e.* with no missing values) [3]. For example, in a study of the relationship between age and salary for which the age of all respondents is known, if older participants tend to skip more often the question "What is your salary?", an analysis not taking into account this MAR pattern may falsely fail to find a positive association between age and salary. However, since the missingness mechanism depends on a fully observed variable (age) there are means of recovering the missing information (to some extent). Data that are neither MCAR nor MAR are referred to as Missing Not At Random (MNAR). In MNAR data, the missingness mechanism depends on the variable of interest itself or other variables not completely known. For example, if in a salary study participants with higher salaries skip the question "What is your salary?", the missingness mechanism is MNAR.

In this paper, we consider the case where values are not MNAR and sporadically missing. As stated above, there are several research works that have proposed algorithms addressing missing data with imputation (replacement of missing values), partial imputation, partial deletion, full analysis, and interpolation [2, 16, 5]. Practitioners often favor imputing missing values before applying a generic algorithm for classification, regression, and/or feature selection. Possible biases introduced by such procedures can be alleviated with multiple imputation [5, 14, 17, 18]. The simple way of dealing with missing values by ignoring missing samples that we mentioned above, asymptotically leads to unbiased results for the MCAR situation [1, 19, 15]. This is due to the fact that the subsample of cases with complete data is equivalent to a simple random sample– that does not cause bias for MCAR assumption– from the original dataset. In the context of feature selection, however, discarding samples with missing values results in a loss of statistical power (risk of false negatives). Thus imputation (*e.g.* by regression) is tempting.

In this work, our aim is to study the potential bias introduced in feature selection when handling missing data. Should we resist the temptation to impute missing values and then utilize standard feature selection methods? Particularly when large amounts of data are missing (e.g. 80%) imputation seems almost unavoidable, as obviously full records are almost impossible to find, which renders multivariate selection methods very difficult to apply. However, we will show that, particularly when large amounts of data are missing, imputation may introduce bias in data, even in the presumably "nicest" case of MCAR data. Thus it will be more advisable to devise feature selection methods that are "robust" to the presence of missing data rather than imputing them [20, 21]. To the best of our knowledge, the modified t-statistic we propose in this paper as well as the use of causal graph to evidence the notion of structural bias are both new[1].

The rest of the paper is organized as follows: In Section 2, we exhibit a case of severe bias introduced by imputation on an example derived from the Gisette benchmark (from the NIPS feature selection challenge) for illustrative purposes. In Section 3, we explain the application domain, which motivates this research and formalizes our problem setting. In Section 4 we derive our proposal for a modified t-statistics. In Section 5, we study a first category of bias introduced by improper imputation: statistical bias, and use the modified t-statistic, which takes care of bias introduced with single imputation by regression. In Section 6, we study a second category of bias, which we call "structural" bias, and which arises when the imputation model reverses causal arrows compared to the original data generating model, and also analyze the results of our proposed modified t-statistic. Section 7 presents our findings on a real life dataset containing information about patients with diabetes.

Finally, we would like to emphasize that this paper is not about uncovering causal relationships, but its aim is to address the problem of establishing dependencies between variables/features and a target variable (the classical feature selection problem). However, as we shall see later on, we make use of functional causal models to describe the underlying data generating process and explain the notion of structural bias introduced by imputation.

[1] Source code available in `https://github.com/chalearn/missing-causal-relation.git`

2. An illustrative example of imputation bias

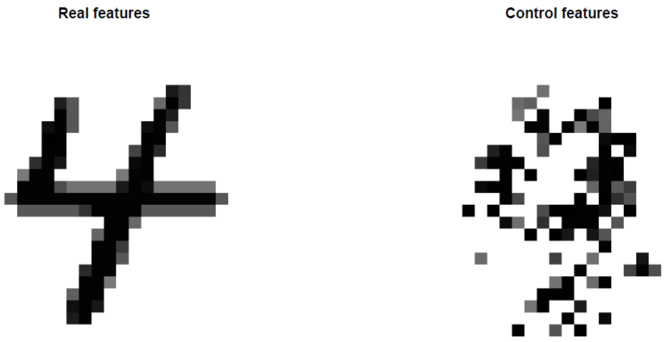
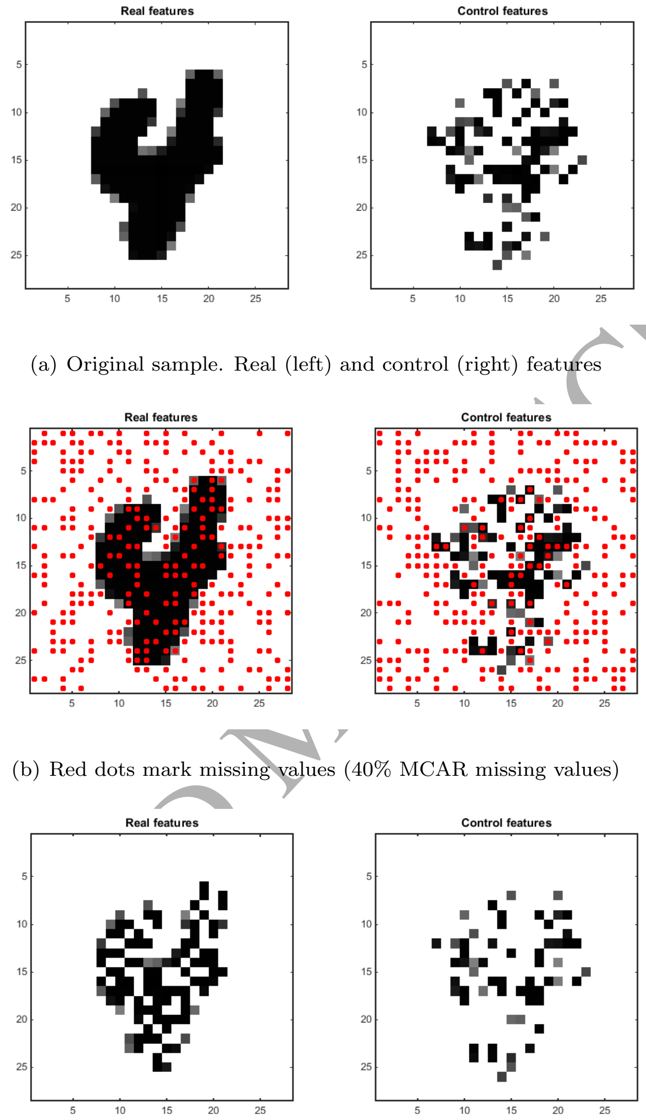


Figure 1: **An example of real features (left image) and control features or distractors (right image) for one sample of the class "4" digit.**

In this section, we will illustrate the imputation bias problem that we face using a reduced version of the standard MNIST dataset benchmark [22], in which only two confusable classes are used: digits 4 and 9. We thus have 6000 examples per class available. In the original representation, each digit image has 28x28 gray level pixels (784 features). We modified the feature set for our purpose by complementing the real features with matched random distractors (control features obtained by random permutations of values of the corresponding real feature, also called "probes")[23, 24, 25]. This led to a double image containing 784 real pixels and 784 control pixels for a total of 1568 features. In Figure 1 we show an example of real and control sub-images for one of the patterns (originally a 4). To the left is the real image and to the right the control image. We notice that the control image has a distribution of dark pixels at locations where dark pixels of a "4" or a "9" are found. Indeed, the marginal distributions of the pixels in control images are the same as those in real images. Control features, however, are statistically independent of real features and of one another.

We used this example to study the impact of missing data on feature selection, where the missing data is MCAR. Since the MCAR missingness mecha-



(a) Original sample. Real (left) and control (right) features

(b) Red dots mark missing values (40% MCAR missing values)

(c) Rendering w. missing values replaced by white pixels

Figure 2: **MCAR missing sample with 40% missing values**

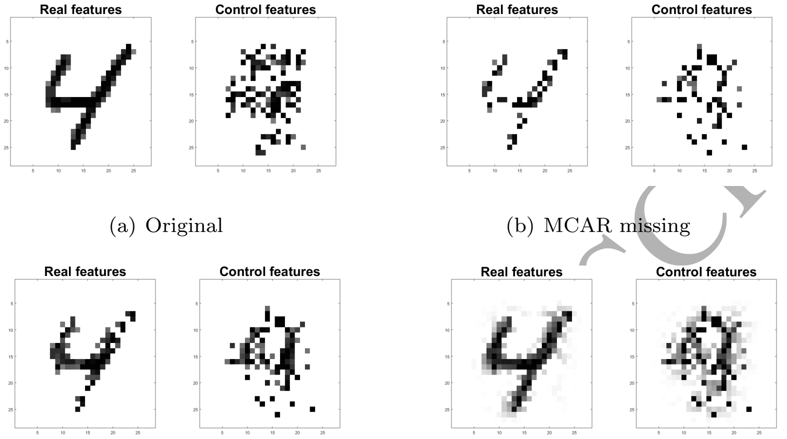
nism is external and unrelated to the data, the probability that a feature value is missing is the same for both classes and for all pixels. In Figure 2, we show an example in which we deleted 40% of the pixels completely at random. Figure 2(a) shows the real and the control features; Figure 2(b) indicates with red dots the 40% missing values; and finally Figure 2(c) shows the resulting sample after replacing all missing values by white pixels (for rendering purpose only). For the above cases, we can see that it will be challenging for a classifier to obtain good results with these type of samples. Thus in most cases, researchers perform an imputation method to substitute values for the missing data using some model.

Let us now assume that we will impute the missing values by two of the most popular imputation methods:

- Median, the missing value in a particular sample f_{miss} is replaced by the median value of the affected feature f obtained from the data samples with no missing data.

- Singular Value Decomposition method (SVD). The SVD imputation method performs a matrix factorization of the data matrix X into $X = USV$, where U and V are orthogonal matrices of eigenvectors and S is a diagonal matrix of singular values. Only the k largest singular values are retained (with their corresponding eigenvectors) and X is reconstructed after restricting U, S, and V to the relevant lines and columns. This fills the missing values with a linear combination of the components of the retained eigenvectors. In practice, we fixed $k = 20$ rather arbitrarily, and iterated the process several times, initializing the missing values to the feature medians, then repeatedly performing SVD, reconstructing X, then substituting the imputed missing values in the original matrix [26].

Obviously, the first method above (median) only considers information from the given feature itself to input its missing values. The second method has been praised by many authors as a simple and efficient imputation method [27, 28, 29]. The missing values f_{miss} of a particular feature are imputed by correlations of lines and columns, and thus in this case it considers information from the whole dataset for missing value imputation. As it is shown in Figure 3, both methods allow for a reconstruction of the original sample (3(a)) when using a sample with 40% of missing values (3(b)), although the result using the SVD imputation method (3(d)) apparently is much better than using the median (3(c)).



(a) Original

(b) MCAR missing

(c) Median Imputation

(d) SVD Imputation

Figure 3: **An example of the use of SVD and median imputation methods over a sample with 40% missing data following a MCAR mechanism.**

We conducted systematic experiments on the "4" and "9" classes of the MNIST dataset [30]. From the original data (28x28 pixel maps) we constructed features including products of pairs of pixels. An equal number of probes was added to real features to obtain 2500 "real" features and 2500 "probes". We used a data split into 6000 training examples and 1000 test examples. This corresponds to the Gisette dataset of the NIPS 2003 feature selection challenge [2]. We added missing data completely at random and varied the proportion of missing data as: 30%, 60% and 80%. As predictor, we used the ridge regression algorithm with default parameters from the CLOP package[3]. We used the S2N filter (analogous to the t-statistic) for ranking the features [31]: $s2n = |\mu_1 - \mu_2|/(s_1 + s_2)$ where μ_1 and μ_2 are the means and s_1 and s_2 the standard

[2] `http://clopinet.com/isabelle/Projects/NIPS2003/`

[3] `http://clopinet.com/CLOP/`

deviations for the two classes. We thus obtained the learning curves (Area under ROC curve as a function of number of features) shown in Figure 4. The learning curves show the superiority of SVD over median imputation (better prediction power). However, the performance of SVD is suspiciously “too good” at low number of selected features: better results are obtained with more missing data! At very high proportion of missing values (80%, for example), using just one selected feature, after imputation leads to better prediction accuracy than using 13 selected features with no missing values. In general, fewer selected features lead to better performance at high levels of missing data in the first part of the learning curves in the Figure 4. This can be explained by the fact that a feature with many missing values, acts as “place holders” being replaced by imputation with linear combinations of all other features. Hence, such a feature, which may originally be completely meaningless (a distractor or “probe”), may become a powerful predictor after imputation. In reality they are “constructed features”.

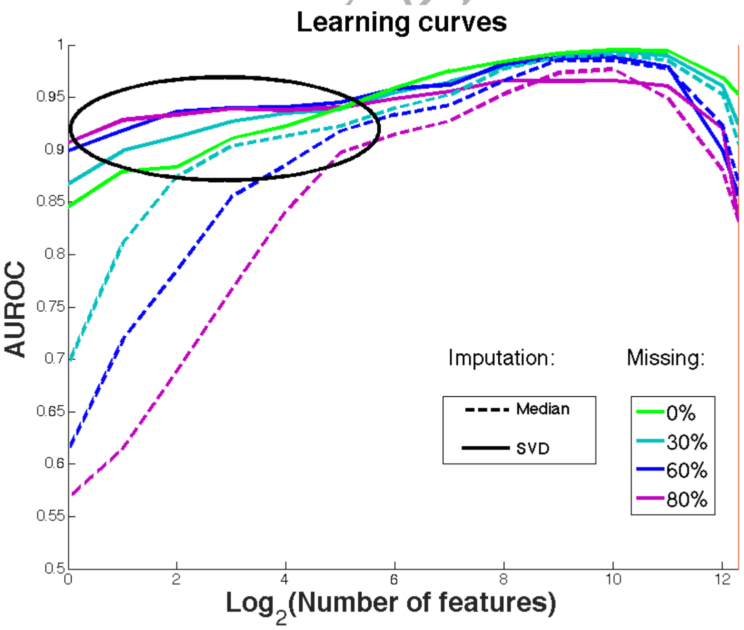


Figure 4: **Predictive power:** Learning curves showing the Area under the ROC curve (AUROC) versus the number of features selected.

It turns out that imputing with SVD before feature selection also results in more false positives (meaningless features falsely called significant). This can be seen in the precision-recall curves shown in Figure 5. For large fractions of missing values, the figure shows that the SVD curves drop more quickly 200 than those of median imputation. As before, our interpretation is that SVD constructs falsely relevant features, which may seem more relevant than real features. This is consistent with the results shown in Figure 4.

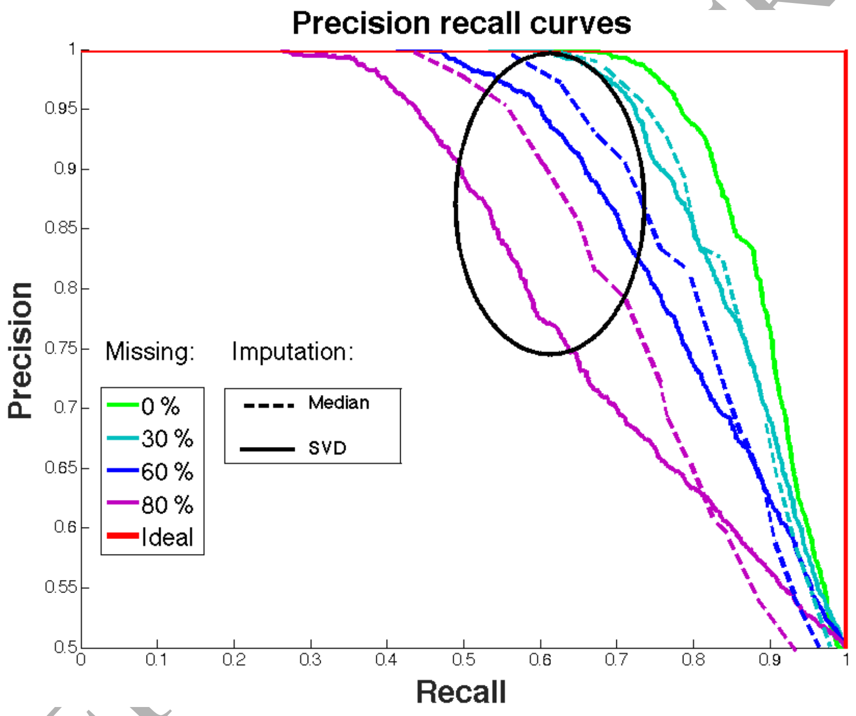


Figure 5: **Discovery power.** Precision obtained versus Recall for the same percentages of missing values as in Figure 4. (Precision = fraction of "true features" retrieved of all features. Recall = fraction of "true features" retrieved of "all true features", a.k.a. True Positive Rate.

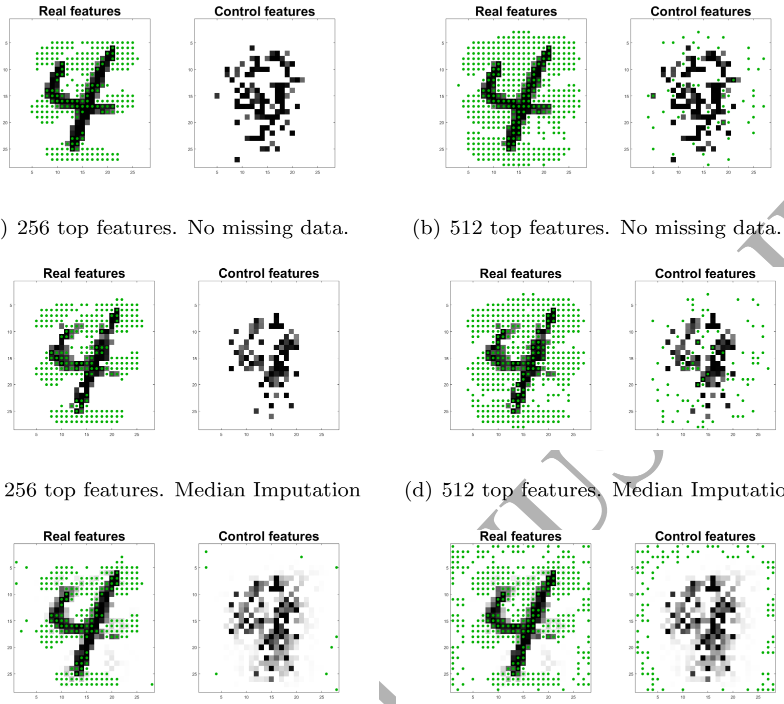
Let us visualize the situation with raw MNIST images. In Figure 6 we show the results of selecting the top 256 and 512 features, shown as green dots in 205 the sample images. As can be seen in Figure 6(a), when there is no missing data (first row), the 256 top ranking features are all selected from real features (not from probes). If we increase the number of top ranking features to 512, a few features are also selected from the control pool (probes). In the case of median imputation (second row) for 40% MCAR missing values, the results for the 256 top ranking features (Figure 6(c)) are similar to those with non missing data (Figure 6(a)). If we extend the selection to the top 512 then a few more features are selected from the control image (Figure 6(d). Finally, for SVD imputation, the 256 top ranking features (Figure 6(e)) already include a few features from the control image, and it is clearly much worse for the top 512 features (Figure 6(f)). This illustrates that missing data imputation with SVD leads to the selection of more false positives (selection of probes in the control image).

3. General problem setting and motivation

Feature selection (FS) is a preprocessing step that has become essential during the last few years, due to the proliferation of large datasets [32, 33]. The goals of FS can be diverse [34], such as reducing the training time needed for subsequent machine learning algorithms. Among others, the purpose of FS is twofold:

- Regression/Classification, which is the problem of predicting a target variable from a set of given variables/features. In the context of FS, the subset of features that increases or minimally degrades prediction performance must be selected, eliminating those features that are irrelevant or noisy.
- Discovery, which is the problem of finding features/variables which may influence the target, with the intention of later testing the causal relationships in a controlled experiment. As the features selected by the FS algorithm are a subset of the original set of features, the underlying phenomena governing a given problem can be explained by identifying those features associated with a given target variable.

While prediction and discovery are sometimes addressed jointly, they differ in their emphasis on Type I and Type II errors [35]:



(a) 256 top features. No missing data.

(b) 512 top features. No missing data.

(c) 256 top features. Median Imputation

(d) 512 top features. Median Imputation

(e) 256 top features. SVD Imputation

(f) 512 top features. SVD Imputation

Figure 6: **Visualization of the results of applying S2N feature ranking.** The first row corresponds to selecting the first 256 (left sub-figure) and 512 (right sub-figure) features for the case of no missing data. The second row is analogous but for median imputation, and the last row correspond to the figures of SVD imputation. All examples except first line correspond with 40% missing data following a MCAR mechanism.

- A Type I error involves rejecting a true null hypothesis (false positives). Incorrectly deducing that a given drug cures a certain disease is an example of a Type I error. Such errors affect the discovery tasks more, as wrong associations or false positives would lead to inferring wrong causal relationships, which may lead to harmful decisions (inefficient or detrimental new policies or treatments).

- A Type II error involves failing to reject a false null hypothesis (false negatives). An example of a Type II error is failing to show that an effective drug cures a certain disease. In this case, it is the prediction task that is more affected due to the absence of a good predictor. Some studies [34] have already shown that predictors are very tolerant to large amounts of irrelevant variables, but their performance deteriorates considerably when key predictive features are omitted.

Although the analysis that will be carried out in this study extends to multivariate feature selection, in order to present our findings more clearly, we will illustrate our reasoning using univariate feature ranking methods. Thus, we will investigate the bias that can be introduced by imputing missing values by regression of continuous helper variable H (auxiliary variable with complete data) onto a continuous "source variable" S (variable of interest having missing values), for which the relevance to a binary "target variable" $T \in \{1, 2\}$ (a 2-class problem) is to be tested. We explicitly forbid imputing values of S using T, since this will obviously corrupt our estimation of the dependency between S and T. We further simplify the problem, assuming data missing completely at random (MCAR) (that is, the missingness mechanism is independent of S, H and T), to separate the problems of bias introduced by the "missingness mechanism" and bias introduced by imputation.

Testing univariate dependencies between several source variables or features and a target can be performed as a preprocessing step in many multivariate feature selection methods in order to select relevant features without overfitting [36]. In cases where the number of samples is very small (in the order of hundreds), compared to the number of features to be tested, it is often advisable to perform only univariate feature selection as preprocessing. Thresholds on p-values (or false discovery rate) are often set to discard features which are below certain significance level. Thus, for our simple scenario and as a criterion of relevance, we use the simple t-statistic, which computes the ratio of: (1) The difference in the means of S for the 2 classes $\bar{S_1} - \bar{S_2}$, and (2) the standard deviation of the difference of the means, which assuming equal class variance σ^2, is approximately $\sqrt{2/n}\sigma$, where n is the number of samples.

For example, the following t-statistic is commonly used for balanced binary classification problems with continuous features:

$$t_{orig} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sigma_p \sqrt{2/n}} \qquad \text{with} \quad \sigma_p^2 = \frac{\sigma_1^2 + \sigma_2^2}{2} \quad \text{and} \quad \sigma_i^2 = \sum_{k=1}^{n_i} \frac{(x_k - \hat{\mu}_i)}{n_i} \ , \qquad (1)$$

where $\hat{\mu}_i$ are the sample means, σ_i the sample standard deviations, σ_p the pooled within class standard deviation, and $n_1 = n_2 = n/2$ the number of samples per class.

Standard methods of evaluation will be used. Thus, the fraction of Type I errors (false positive rate) will be assessed by the p-value of a statistical test (e.g. the t-test for the t-statistic). Corrections for multiple testing such as the Bonferroni correction, may be applied on top of our analysis. For non-tabulated test statistics, we will use “distractors” to emulate a null distribution (features obtained by randomly permuting the values of real features). The p-value is then estimated as the fraction of distractors whose test statistic exceeds the value obtained for the feature being tested. As Type II errors are more difficult to assess in real data, for which no ground truth of the relevant features is known, the prediction performances of a given predictor will be used in order to quantify them indirectly.

3.1. Imputation mechanisms

The simplest way to deal with missing values is to compute the t-statistic on the basis of only the n_{obs} observed samples, while ignoring all the n_{mis} missing samples. This has been shown to lead asymptotically to unbiased results for MCAR situations [4]. But discarding samples with missing values will unavoidably result in a loss of statistical power (risk of false negatives), making imputation (*e.g.* by regression) very tempting. In the context of feature selection, however, there are at least three types of biases that can be introduced with imputation by regression:

- Optimistic sample count: a simple imputation with the mean of S (which could be thought of as regression with a constant model) allows us to use all n samples instead of only the n_{obs}. However, this leads to a false increase in the number of samples, which artificially inflates the t-statistic producing too small estimates of the p-value.

- Optimistic variance estimate by ignoring the regression residual: imputing with the mean of S or with an estimate of the expected value of S from H (regression: $\hat{S} = aH + b$) leads to an under-estimate of the variance of S because imputed values have less variance than real values. To avoid that, imputed values should be drawn from $P(S|H)$ to take into account the intrinsic noise (captured e.g. by the residual in a least-square regression), as performed in multiple imputations.

- Optimistic variance estimate by ignoring the uncertainty on the regression parameter(s): when imputing by regression, we must estimate the parameters of the regression model (*e.g.* the slope a of linear regression) using the n_{obs} available values. Our estimator of the variance of S using values imputed by regression must also take into account the uncertainty on our estimate of coefficient a.

In what follows, we derive adjusted t-statistics accounting for such biases. Our purpose is mostly didactic: we want to arrive at a simple formula, which includes terms accounting for systematic errors introduced by "intrinsic noise" (regression residual) and "finite training sample" (regression parameter error bar). We demonstrate that even for simple cases where the best case data generation process is known with certainty, the usual t-test is optimistic as it underestimates the increased variance arising from imputation.

Thus, we illustrate a simple case in which the t-statistic models the data generating process effectively using the following assumptions:

- T is a Bernouilli process, $T \in \{1, 2\}$ with given apriori probabilities $\mathrm{P}(T = 1)$ and $\mathrm{P}(T = 2)$.

- S is drawn given T with: $\mathrm{P}(S|T) \sim N(\mu_T, \sigma_T)$.

H is a variable correlated with S. We use linear regression to perform imputation:

$$\hat{S} = aH + b$$

Variables T, S, and H are jointly observed.

Other authors have addressed the related problem of variance estimation in the presence of imputation of missing data in a more detailed and general way (e.g. [37, 38, 39]), not directly relating it to the problem of feature selection. These could be useful follow up readings for generalizing our findings to other univariate feature selection statistics. Our purpose in this paper is to arrive at a formula, which exhibits terms showing schematically the influence of the various types of biases and to alert the machine learning community of such problems.

4. Our proposal for a modified t-statistic

In this section, we will describe our proposal for a modified t-statistics that can take into account the uncertainty of linear regression imputation. To start with, we will describe first our assumptions and the notation that will be employed.

- $T \in \{1, 2\}$ is the binary target variable, with each class having n_T points.
- The samples of interest S_{Ti}, $i = 1 \dots n_T$ are the i.i.d samples in class T with mean μ_T and variance σ_T.
- The set of n_{obsT} observed samples in Class T are obs_T. The set of n_{misT} missing samples in Class T are mis_T.
- The helper samples H_{Ti}, $i = 1 \dots n_T$ are the i.i.d samples in class T. H is a fully observed variable, although we can still index it using the above index scheme as necessary.
- We produce estimates of $S_{Ti}, i \in mis_T$ using linear regression on S. This process creates the variable $\hat{S}_{Ti} = aH_{Ti} + b + \epsilon_i$ (and as for now we will ignore the mechanism of how the coefficients of $\hat{S}$ are estimated).

We assume ϵ are i.i.d. with mean 0 and standard deviation σ_ϵ, and also assume a and σ_e are known.

Using the imputed data, μ_T is estimated by:

$$\bar{S}_T = \frac{\sum_{i \in obs_T} S_{Ti} + \sum_{j \in mis_T} \hat{S}_{Ti}}{n_T} \tag{2}$$

Our aim is to test the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ against the null hypotheses $H_0 : \mu_1 = \mu_2$.

Using only the observed data and the assumption that the standard deviations are equal for the two classes this can be accomplished using the two sample t-test with common standard deviation:

$$t = \frac{\bar{S}_1 - \bar{S}_2}{\sigma_p \sqrt{\frac{1}{n_{obs1}} + \frac{1}{n_{obs2}}}} \tag{3}$$

where the pooled standard deviation is:

$$\sigma_p = \sqrt{\frac{(n_{obs1} - 1)\sigma^2_{obs1} + (n_{obs2} - 1)\sigma^2_{obs2}}{n_{obs1} + n_{obs2} - 2}}. \tag{4}$$

Our next step is to see how this test changes when we incorporate the imputed data. In order to do this, we begin by deriving the mean and variance of our estimate of $\bar{S}_1 - \bar{S}_2$ (after imputation). For this, we need to compute the expected value of the difference of the adjusted means, $E(\bar{S}_1 - \bar{S}_2)$. For any $j \in m_T$.

$$\begin{aligned} E(\hat{S}_{Tj}) &= E(\hat{a}H_{Tj} + \hat{b} + \epsilon_j) \\ &= E(\hat{a}H_{Tj}) + E(\hat{b}) + E(\epsilon_j) \\ &= E(\hat{a}H_{Tj}) + b + 0 \\ &= cov(\hat{a}, h_{Tj}) + E(\hat{a})E(h_{Tj}) + b \\ &= a\mu_{HT} + b \text{ if a is fixed} \end{aligned} \tag{5}$$

The later comes from the fact that Ordinary Least Squares creates unbiased estimates of a and b and for fixed a, $cov(\hat{a}, h_{Tj}) = 0$. If a is calculated then

$cov(\hat{a}, h_{Tj}) = cov(g(S_{obs1}, H_{obs1}), H_{Tj})$ may not be zero, where g is a fixed function of the observed data.

Using this result we can find the expected difference of the estimates (for the sake of clarity in the formula, abbreviating *obs* with o and *mis* with m):

$$\begin{aligned}
&E(\bar{S}_1 - \bar{S}_2) = \\
&\qquad E\left(\frac{\sum_{i\in o_1} S_{1i} + \sum_{j\in m_1} \hat{S}_{1j}}{n_1} - \frac{\sum_{i\in o_2} S_{2i} + \sum_{j\in m_2} \hat{S}_{2j}}{n_2}\right) \\
&\qquad = \frac{n_{o1}}{n_1}\mu_1 + \frac{n_{m1}}{n_1}E(\hat{S}_1) - \frac{n_{o2}}{n_2}\mu_2 - \frac{n_{m2}}{n_2}E(\hat{S}_2) \\
&\qquad = \frac{n_{o1}}{n_1}\mu_1 + \frac{n_{m1}}{n_1}(cov(\hat{a}, h_1) + E(\hat{a})E(h_1) + b)) \\
&\qquad - \frac{n_{o2}}{n_2}\mu_2 - \frac{n_{m2}}{n_2}(cov(\hat{a}, h_2) + E(\hat{a})E(h_2) + b))
\end{aligned} \tag{6}$$

If we assume that a is fixed or that $\frac{n_{m1}}{n_1}(cov(\hat{a}, h_1) = \frac{n_{m2}}{n_2}(cov(\hat{a}, h_2)$, then the expression can be simplified as:

$$E(\bar{S}_1 - \bar{S}_2) = \left(\frac{n_{o1}}{n_1}\mu_1 + \frac{n_{m1}}{n_1}a\mu_{H1}\right) - \left(\frac{n_{o2}}{n_2}\mu_2 + \frac{n_{m1}}{n_2}a\mu_{H2}\right) \tag{7}$$

And if we assume that the sample sizes for missing and observed data are the same for both classes, the expression simplifies to:

$$E(\bar{S}_1 - \bar{S}_2) = (f_o\mu_1 + f_m a\mu_{H1}) - (f_o\mu_2 + f_m a\mu_{H2}) \tag{8}$$

where f_o and f_m are the fractions of observed and missing data, respectively.

Now, let us derive the variance. We define:

$$Q_o = \frac{\sum_{i\in o_1} S_{1i}}{n_1} - \frac{\sum_{i\in o_2} S_{2i}}{n_2} \tag{9}$$

$$Q_m = \frac{\sum_{j\in m_1} \hat{S}_{1j}}{n_1} - \frac{\sum_{j\in m_2} \hat{S}_{2j}}{n_2} \tag{10}$$

Also, the result exploits the facts that $S_i \perp S_j$ and $S_i \perp \hat{S}_j$ for $i \neq j$, and

$S_i \perp H_j$ and $\hat{a} \perp H_j$for $i \in obsT$, $H_j \in misT$.

$$\begin{aligned} var(\bar{S}_1 - \bar{S}_2) &= var(Q_o + Q_m) \\ &= var(Q_o) + var(Q_m) + 2cov(Q_o, Q_m) \\ &= \frac{n_{o1}}{n_1^2} var(S_1) + \frac{n_{o2}}{n_2^2} var(S_2) \\ &+ \frac{n_{m1}}{n_1^2} var(\hat{S}_1) + \frac{n_{m2}}{n_2^2} var(\hat{S}_2) + 2cov(Q_o, Q_m) \\ &= \frac{n_{o1}}{n_1^2} var(S_1) + \frac{n_{o2}}{n_2^2} var(S_2) + \frac{n_{m1}}{n_1^2} var(aH_1 + \epsilon) \\ &+ \frac{n_{m2}}{n_2^2} var(aH_2 + \epsilon) + 2cov(Q_o, Q_m) \end{aligned} \tag{11}$$

As we are in the case in which a is fixed, we also can exploit $cov(Q_o, Q_m) = 0$ (This covariance may not be 0 for estimated $\hat{a}$). Thus results can be further simplified to:

$$\begin{aligned} var(\bar{S}_1 - \bar{S}_2) &= \frac{n_{o1}}{n_1^2} \sigma_1^2 + \frac{n_{o2}}{n_2^2} \sigma_2^2 \\ &+ \frac{n_{m1}}{n_1^2} (a^2 \sigma_{H1}^2 + \sigma_\epsilon^2) + \frac{n_{m2}}{n_2^2} (a^2 \sigma_{H2}^2 + \sigma_\epsilon^2) \end{aligned} \tag{12}$$

And again, as our assumptions are a fixed, the classes have the same variances, and equal sample sizes for both classes, the expression finally reduces to:

$$var(\bar{S}_1 - \bar{S}_2) = \frac{2}{n}(f_o \sigma^2 + f_m(a^2 \sigma_H^2 + \sigma_\epsilon^2)) \tag{13}$$

The distribution of this statistic would depend on the assumptions of the problem.

4.1. Creating the test for fixed a

For the case of a fixed, we use as our statistic:

$$t = \frac{\bar{S}_1 - \bar{S}_2}{\sqrt{\frac{2}{n}(f_o \sigma^2 + f_m(a^2 \sigma_H^2 + \sigma_\epsilon^2)}} \tag{14}$$

with the variables replaced by their corresponding sample estimates.

If we further assume S and H have same variance and $a = 1$, then

$$t = \frac{\bar{S}_1 - \bar{S}_2}{\sqrt{\frac{2}{n}(\sigma_{all}^2 + f_m \sigma_\epsilon^2)}} \tag{15}$$

where σ_{all}^2 is the sample covariance of observed S_T and imputed $\hat{S}_T$ combined.

4.2. Creating the test for estimated a

The coefficients of the regression function have a closed form, so we can determine their distributions.

Let us define $\bar{S_o}$ as the sample mean of $S_i, i \in obs$, $\bar{H_o}$ and $var(H_o)$ as the sample mean and variance of $H_i, i \in obs$, and $cov(H_o, S_o)$ as the corresponding sample covariance. Then

$$\hat{a} = \frac{cov(H_o, S_o)}{var(H_o)} \tag{16}$$

$$\hat{b} = \bar{S_o} - \hat{a}\bar{H_o} \tag{17}$$

$$E(\hat{a}) = a \tag{18}$$

$$var(\hat{a}) = \frac{\sigma_\epsilon^2}{n_o \sigma_H^2} \tag{19}$$

380 where $\sigma_R^2 = \frac{\sum_{i \in o}(S_i - \hat{s_i})}{n_o - 2}$. Note that we can remove b in $\hat{S} = aH + b$ by normalizing observed data, so let's assume we do that and neglect b from further calculations.

We need a key result here on the variance of the estimated S for both classes:

$$\begin{aligned} var(\hat{S}_{Ti}) &= var(\hat{a}H_{Ti} + \epsilon) \\ &= var(\hat{a}H_{Ti}) + var(\epsilon) \\ &= var(\hat{a})var(H_{Ti}) + var(\hat{a})E(H_{Ti})^2 \\ &+ var(H_{Ti})E(\hat{a})^2 + var(\epsilon) \\ &= \frac{\sigma_\epsilon^2}{n_o \sigma_{H_o}^2}\sigma_{H_T}^2 + \frac{\sigma_\epsilon^2}{n_o \sigma_{H_o}^2}\mu_{H_T}^2 + \sigma_{H_T}^2 a^2 + \sigma_\epsilon^2 \\ &= \sigma_{H_o}^2 a^2 + \left[1 + \frac{1}{n_o} + \frac{\mu_{H_T}^2}{n_o \sigma_{H_o}^2}\right]\sigma_\epsilon^2 \end{aligned} \tag{20}$$

Note that this depends on

$$var(XY) = var(X)var(Y) + var(X)E(Y)^2 + var(Y)E(X)^2.$$

This last step assumes $\sigma^2_{H_o} = \sigma^2_{H_T}$, i.e. that the variance of H is the same for all T. Estimation of a increases the impact of the residual on the variance, but this additional impact goes to 0 as $n_0 \to \infty$.

Note

$$var(Q_o) = \frac{n_{o1}}{n_1^2} var(S_1) + \frac{n_{o2}}{n_2^2} var(S_2) = \frac{n_{o1}}{n_1^2}\sigma_1^2 + \frac{n_{o2}}{n_2^2}\sigma_2^2 \tag{21}$$

since the $cov(S_1, S_2) = 0$, and

$$var(Q_m) = \frac{n_{m1}}{n_1^2} var(\hat{S}_1) + \frac{n_{m2}}{n_2^2} var(\hat{S}_2) - \frac{2}{n_1 n_2} cov(\sum_{i \in m_1} \hat{S}_{1i}, \sum_{j \in m_2} \hat{S}_{2j}) \tag{22}$$

Here we assume that the covariance term is negligible since its limit is 0 as the sample size grows and the only dependence between $\hat{S}_1$ and $\hat{S}_2$ is through $\hat{a}$.

Under simplifying assumptions that S and H have the same variance and that a=1, this becomes

$$\begin{aligned} var(Q_o) &= \frac{2f_o}{n}\sigma^2 \\ var(Q_m) &= \frac{2f_m}{n}\left[\sigma^2 + \left[1 + \frac{(2\sigma^2 + \mu^2_{H_1} + \mu^2_{H_2})}{2n_o\sigma^2}\right]\sigma^2_\epsilon\right] \end{aligned} \tag{23}$$

Thus under the assumptions that the variance of S and H are equal and that the sample size is sufficiently large,

$$\begin{aligned} var(\bar{S}_1 - \bar{S}_2) &= var(Q_o + Q_m) \\ &= var(Q_o) + var(Q_m) + 2cov(Q_o, Q_m) \\ &= \frac{2}{n}\left[\sigma^2 + f_m\left[1 + \frac{(2\sigma^2 + \mu^2_{H_1} + \mu^2_{H_2})}{2n_o\sigma^2}\right]\sigma^2_\epsilon\right] + 2cov(Q_o, Q_m) \\ &\approx \frac{2}{n}\left[\sigma^2 + f_m\left[1 + \frac{(2\sigma^2 + \mu^2_{H_1} + \mu^2_{H_2})}{2n_o\sigma^2}\right]\sigma^2_\epsilon\right] \end{aligned} \tag{24}$$

Neglecting the covariance will result in an underestimate of the variance of $\bar{S}_1 - \bar{S}_2$. Recall that Q_0 is the difference in the estimates of the means of the classes on the observed data and that Q_m is the difference in the means of the classes on the imputed data. Thus, Q_0 and Q_m will be positively correlated. As the number of observed samples goes to infinity for a fixed f_m, then Q_o and a

rapidly converge to constants. Consequently, the problem converges to the case of a known which has the $cov(Q_0, Q_m) = 0$ so the covariance term would have little impact for problems with large numbers of observed data.

This suggests a statistic of the form

$$t = \frac{\bar{S}_1 - \bar{S}_2}{\sqrt{\frac{2}{n}(\sigma_{all}^2 + f_m \left[1 + \frac{\alpha}{n_o}\right] \sigma_\epsilon^2)}} \tag{25}$$

with $\alpha > 0$ chosen appropriately. The best statistic and the distribution of that statistic depend on the assumptions about the joint distribution of S and H, and thus this remains an open question. But clearly when used with imputed data, the typical t-test may induce false positive errors due to the underestimation of variance.

5. Statistical bias

To start with the simplest problem, it is well known that the use of imputation causes statistical biases [40, 17]. The application of Eq. 1 to imputed data may yield false positive discoveries by under-estimating the variance in data, with the first source of bias being an over-estimation of the number of samples n_i. The imputed values violate the independence and identically distributed assumptions of the formula. As imputed values do not bear novel information, one simple manner of correcting the problem is to divide the class variances by the number of observed values n_{oi} (instead of dividing by the total number of values). This simple arrangement can be done for the case of the median imputation, however the SVD imputation's bias analysis is not that simple. Going back to our continuous helper variable H with complete data, our continuous source variable S and our binary target variable T mentioned in Section 3, let us assume for simplification that we perform (single) imputation of the missing values of a feature of interest S by linear regression of the fully observed helper variable H (correlated both to S and the target T). If we use the imputed values in the calculation of σ_1 and σ_2 the variance might be underestimated due to two reasons:

- If the linear model $S = aH + noise$ is assumed as correct, when single imputation is carried out, the missing values are replaced by their expected value $\hat{S} = aH$. Thus, the noise term, which can be estimated by the RMSE residual of the fit σ_r, is completely ignored.
- The regression coefficient is evaluated only from a small and finite quantity of observed data, and thus some uncertainty appears.

In the simple and approximate formula below, both type of biases are corrected:

$$t_{modified} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{2/n}\sqrt{\sigma_p^2 + \underbrace{f_m \sigma_r^2}_{\text{regression residual}} \underbrace{(1 + \alpha/n_o)}_{\text{regression coeff. uncertainty}}}} \tag{26}$$

Compared to Equation 1, σ_p is the pooled within class standard deviation (estimated on all samples after imputation), f_m is the fraction of missing values, σ_r is the RMSE residual of the fit, n_o the number of observed (non missing) values, and α a positive coefficient.

Equation 26 corrects two types of biases, which create false positive results: (1) We add at the denominator a term proportional to s_r, taking into account the intrinsic regression error. This term vanishes when the fraction of missing values go to zero. (2) We add another corrective term to take into account the uncertainty on the regression coefficient a, which goes to zero as the number of observed examples n_o increases. The coefficient of proportionality is computed with $\alpha =< \mu_H^2 > /\sigma_H^2$, where $< \mu_H^2 >$ is the average of the square class-wise means of H and σ_H is the variance of H.

To arrive at this formula we made several simplifying assumptions, however, the formula clearly reflects qualitatively the influence of the two types of biases and the necessity to correct them.

6. Structural bias

Regarding other problems that appear when using imputation with incorrect assumptions about the data generating process, we will make use of simple structural causal models to analyze the imputation mechanisms. These models have been used successfully used before by other authors [6] for analyzing missingness mechanisms. However, in our case we will be utilizing them with a different goal in mind, that is, the analysis of imputation mechanisms assuming an MCAR context. Let us remember that we have defined three different variables in the previous sections: The feature of interest S, for which some values might be missing; fully observed variables H and T, where H is a feature correlated to both S and T; and finally Σ, which is a feature S after the imputation of missing values using H.

And we ask ourselves the following question:

Does $\Sigma \perp T \Rightarrow S \perp T$ and $\Sigma \angle T \Rightarrow S \angle T$ (where $\perp \doteq$ independent and $\angle \doteq$ dependent)?

First we carry out a listwise deletion (LWD) independence test between S and T, in which the records containing missing values are ignored. Subsequently, we challenge the LWD test result by re-testing after imputation by regression with H. The t-test has been employed for revealing dependencies between S (source) and T (target) (p-value ≤ 0.01). H (helper) is an "auxiliary" variable. We will exemplify two cases using histograms and scatter plots of S and H shown in Figures 7 and 8. The binary variable $T = \pm 1$ is color coded (red/blue) and the missing values are represented in yellow. Depending on the outcome of the LWD test, we have two different cases:

1. Case 1 (model shown in subfigure 7(a)), in which imputation yields false negatives, that is, Listwise Deletion (LWD) reveals that $S \angle T$. To challenge this result with imputation, we must use a variable H such that $H \perp T$. This results in a set of dependencies for the null model and alternative model shown in Table 1. Imputation does not reverse any causal arrow and can be considered "legitimate". However, as exemplified in Figure 7, it can result in false negatives. Thus, if we find a dependency

Table 1: **Challenging the results of Listwise Deletion (LWD test).** We represent schematically all possible attempts to obtain better results with imputation instead of LWD. The first column indicates the conclusion drawn after performing the LWD test. The second column indicates the dependencies between variables for the corresponding null model and alternative model (interestingly, the dependencies between S and T impose a single meaningful choice of dependencies between S and H and between H and T, see text). The model graph summarizes all causal models consistent with such dependencies. Directed arrows mean a causal relationship and bidirected arrows the presence of a latent common cause. Undirected edges mean any dependency (causal direction irrelevant). Stars are "wild cards" coding for "arrow or not arrow" e.g. $A \leftarrow *B$ means $A \leftrightarrow B$ or $A \leftarrow B$. The imputation graph represents the potential change in causal direction incurred by imputation of missing data. Double arrows mean imputation.

LWD	Dependencies	Model graph	Imput. graph
$S\angle T$	Null model ($S\perp T$,$S\angle H$,$H\perp T$)	$T \quad S * - * H$	$T \quad \Sigma \Leftarrow H$
	Alt. model ($S\angle T$,$S\angle H$,$H\perp T$)	$T* \rightarrow S \leftarrow *H$	$T* \rightarrow \Sigma \Leftarrow H$
$S\perp T$	Null model ($S\perp T$,$S\angle H$,$H\angle T$)	$S* \rightarrow H \leftarrow *T$	$\Sigma \Leftarrow H \leftarrow *T$
	Alt. model ($S\angle T$,$S\angle H$,$H\angle T$)	$S - H$ $\diagdown \; \diagup$ T	$\Sigma \Leftarrow H$ $\diagdown \; \diagup$ T

that is significant with LWD, we rather satisfy ourselves with this result than impute and risk to get a false negative result.

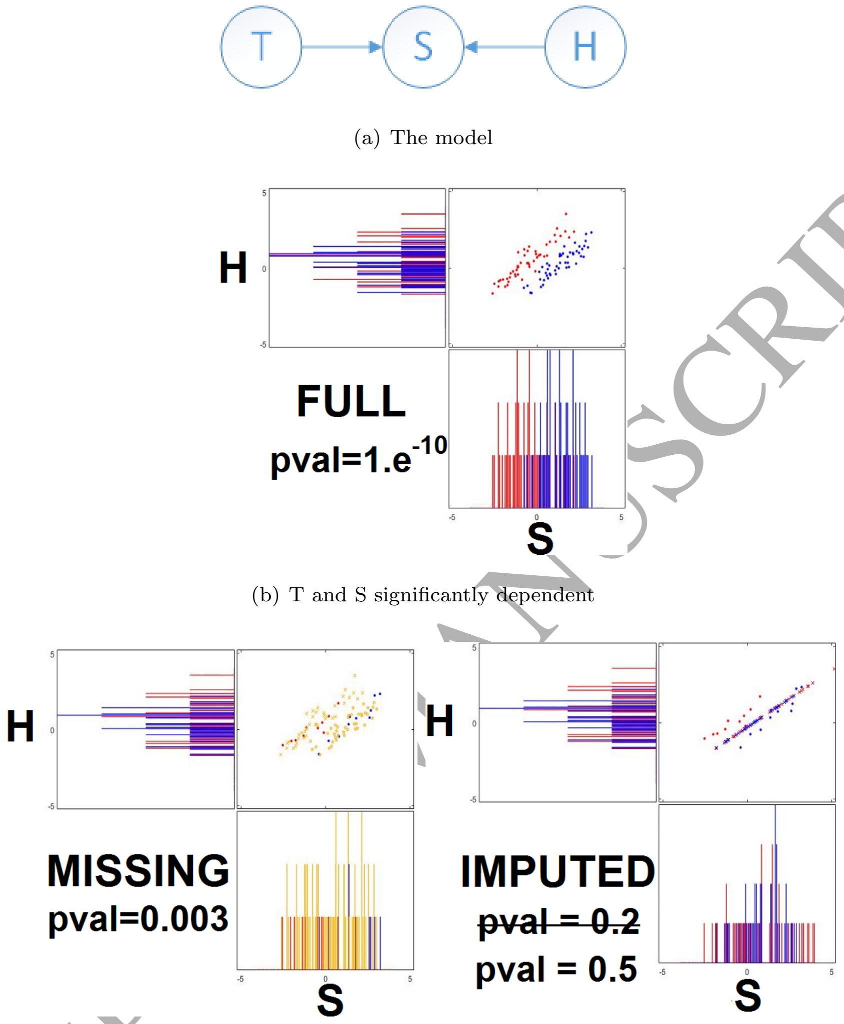
We will exemplify this case in Figure 7. We draw 100 points following the data generation model $T \sim \text{Bern}\,(p = 1/2)\,; S \sim \mathcal{N}(0, 1), ; H = T + S + \textit{noise}$. As can be seen in subfigure 7(b), T and S are significantly dependent. Then we make S have 80% of missing values simulating an MCAR mechanism. As H and T are fully known, the p-value indicates that T and S remain significantly dependent based on the 20% of remaining complete data (subfigure 7(c)). On subfigure 7(d) we show the result of carrying out an imputation of values by regressing H on S. It can be seen that the imputation results in a loss of separation of blue/red in the histogram. Checking the p-value (strike-through is the original p-value, non-strike through is our modified t-statistic), the dependence between S and T can no longer be detected. Thus, imputation using H, that contains no information about T, contributes to noise with respect to detecting the dependence between S and T.

2. Case 2 (model shown in subfigure 8(a), in which imputation yields false positives, that is, LWD reveals that $S \perp T$. To challenge this result with imputation, we must use a variable H such that $H \angle T$. As shown in Table 1, this time imputation does reverse a causal arrow, which may lead to a false positive discovery.

We will exemplify this case in Figure 8. As in the case above, we draw 100 points with the model $T \sim \text{Bern}\,(p = 1/2)\,; S \sim \mathcal{N}(0, 1); H = T + S + noise$. As can be seen in subfigure 8(b) the p-value of the test confirms that there is not a significant dependency between S and T. Again, S is simulated to have 80% of missing values following a MCAR mechanism, H and T are fully known and it can be seen that there is no change in the p-value of the test. Finally, analogous to Case 1, we impute missing values by regressing S on H. The consequence is that the imputation model reverses the causal relation between S and H, thus S and T become dependent, as strike-trough p-value shows, and as can be seen in subfigure 8(d), a separation of red/blue appears in the histogram of S that was not in the original data. Our proposed correction to the t-statistic brings the p-value above the chosen significance level (0.01).

7. Experiments on a real life dataset

We carried out further experiments on a real life dataset, of patients with diabetes, collected over a period of ten years [41]. The dataset has several features including detailed survey items that have important information about the patients. As all questions in the questionnaire were not answered by patients, there are many missing values in the raw collected data. To illustrate our findings, we took a subset of patients who filled out the questionnaire and we selected those questionnaire items which have more than 90% complete information (so less than 10% missing values). This reduced our data to a size



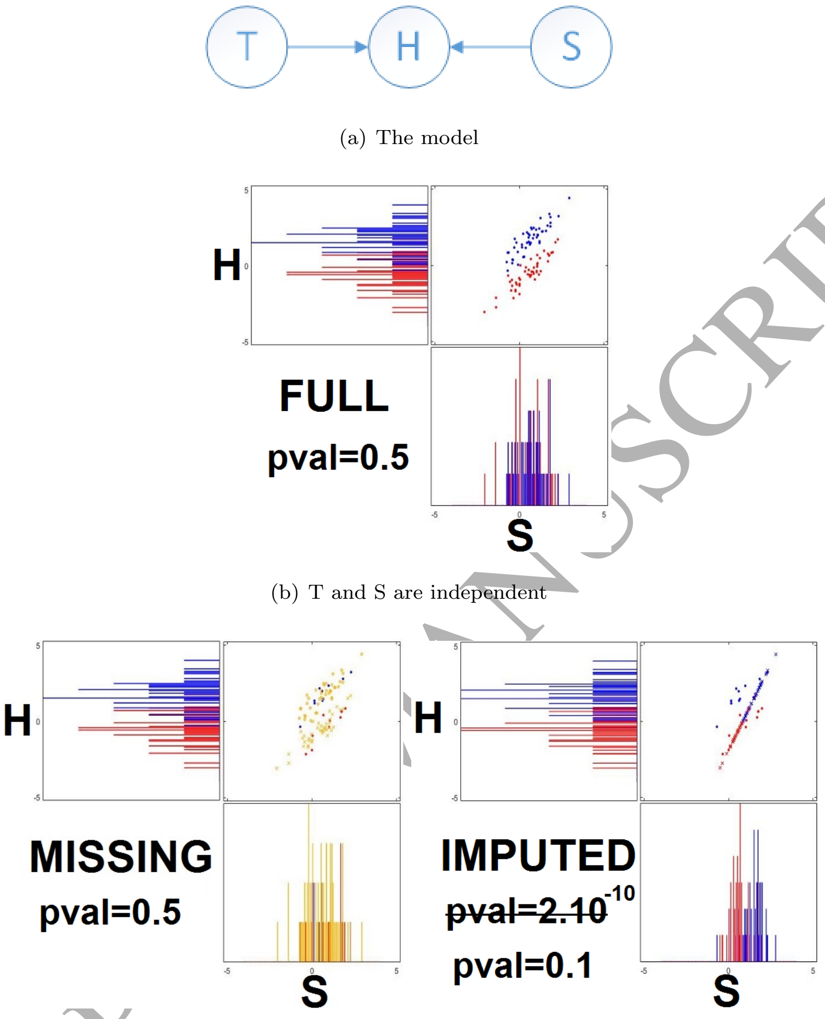
(a) The model

(b) T and S significantly dependent

(c) T and S remain significantly dependent

(d) S/T dependency not detectable

Figure 7: **An example in which it can be seen that imputation yields FALSE POSITIVES.** The t-test is employed to reveal dependencies between S (source) and T (target) (p-value ≤ 0.01). H (helper) is an "auxiliary" variable. In histograms and scatter plots of S and H, binary variable $T = \pm 1$ is color coded (red/blue) and missing values are represented in yellow. Subfigure (b) displays 100 points randomly drawn following the data generating model $T \sim \text{Bern}\,(p = 1/2)\,; H \sim \mathcal{N}(0,1); S = T + H + noise$. The p-value reveals that T **and** S **are significantly DEPENDENT**. In Subfigure (c), S has 80% of values Missing Completely At Random (MCAR). H and T are fully known. The p-value indicates that T **and** S **remain significantly DEPENDENT** based on the remaining 20% of **complete data**. For subfigure (d), we impute missing values by regressing S on H. Imputation results in a loss of blue/red separation in the histogram of S. According to the p-value, **the dependency between** S **and** T **is no longer detectable**. Imputation using H, carrying no information about T, contributed *noise* w.r.t. detecting the dependency between S and T.



(a) The model

(b) T and S are independent

(c) T and S remain independent

(d) S and T become dependent

Figure 8: **An example in which it can be seen that imputation yields false NEGATIVES.** Same problem setting as in the previous figure, but with a different data generating model for which now S **and** T **are INDEPENDENT**. In subfigure (b) we draw 100 points with the model $T \sim \text{Bern}\,(p = 1/2)\,; S \sim \mathcal{N}(0, 1); H = T + S + noise$. No significant dependency between S and T is found according to the p-value of the T-test. Thus, T and S are INDEPENDENT. In Subfigure (c) S has 80% of values Missing Completely At Random (MCAR). H and T are fully known. The situation does not change, and thus T and S remain independent. For subfigure (d) we impute missing values by regressing S on H. The imputation model $H \Rightarrow S$ reverses the causal arrow $S \rightarrow H$. After imputation there is a blue/red separation in the histogram of S, which did not exist in the original data, that is S **and** T **become DEPENDENT** (strikethrough p-value). Our proposed correction to the t statistic (non strikethrough text) brings the p-value below the chosen significance level (0.01).

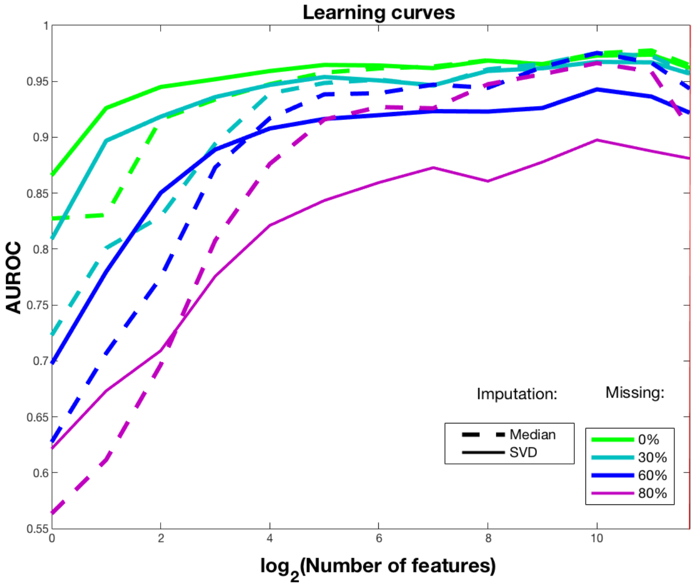


Figure 9: **Predictive power on diabetes dataset:** Learning curves showing the Area under the ROC curve (AUROC) versus the number of features selected for the ISIS diabetes dataset.

of 1110x555 matrix. The target variable is the age at diagnosis of a patient, binarized to whether the age at diagnosis was 4 years or more (positive class) or less than 4 years (negative class). The data has both binary and continuous features representing various aspects of a patient's life like diet, lifestyle, family information, etc.

In order to keep the experiments consistent with the Gisette dataset, described in the earlier sections, we added 555 features, which were the product of two features selected at random, increasing the number of relevant features to 1665. To the 1665 features we added 1665 probes (irrelevant), which were random permutation of existing features. We repeated the experiments for different proportions of missing data, i.e., 30%, 60% and 80%, generated via the MCAR mechanism. This is similar in spirit to the simulations performed on the Gisette dataset described earlier.

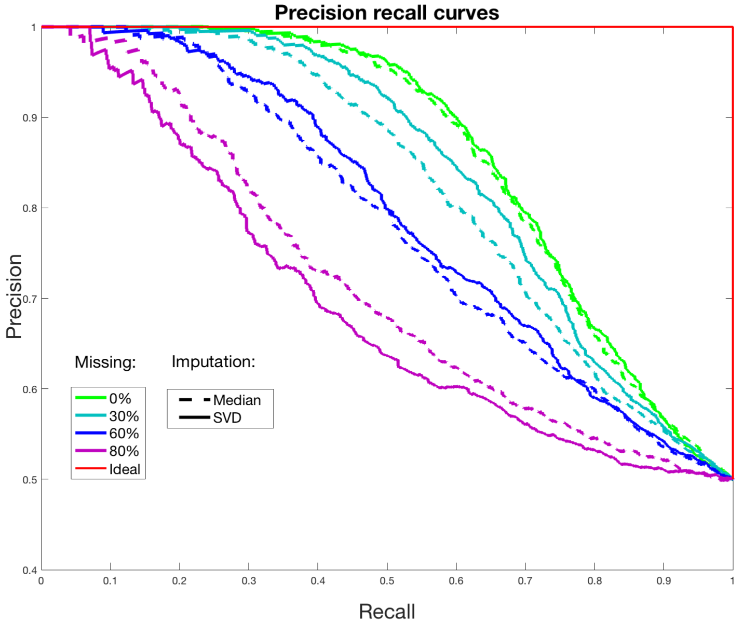


Figure 10: **PR curves for the diabetes dataset:** Precision recall curves for the ISIS diabetes dataset for different percentage of missing values. Missing mechanism is MCAR.

530 When we have a mix of binary and continuous variables, it is recommended in the machine learning literature to always add a pre-processing step that standardizes or rescales the data, to bring each variable to the same scale. However, the presence of missing values in our original data poses a problem for the initial processing of data. For example, when standardizing a feature, 535 the mean and standard deviation of the feature are required. Computing the mean and standard deviation would then require an imputation method to fill in missing values. Alternatively, case wise deletion can be used for computing the mean and the standard deviation. As this is a research problem in its own right and deserves further investigation, we leave the pre-processing of 540 data in the presence of missing values as a subject of our future research. For the experiments mentioned in this section we applied a method of imputation (median or SVD) to the initial raw data and then applied S2N (as explained on

page 9) to rank each feature according to its relevance to the target variable. Once imputation was applied to the data, we applied standardization and then the logistic regression classifier.

Figure 9 illustrates the learning curves for different percentages of missing data, when 5 fold cross validation accuracy (AUC of the ROC curve) was recorded from a logistic regression classifier. Comparing with the learning curve of Figure 4 on page 10 for the Gisette dataset, we can see similar trends for both SVD and median imputation. For a smaller set of features the performance of SVD is significantly better when the percentage of missing data is very high. However, with more features, median imputation appears to be a superior method.

As noted for the Gisette dataset, for a larger fraction of MCAR missing values, the SVD imputation method, applied before feature selection, has poorer feature discovery power as presented by the precision recall curve of Figure 10. Note in this figure that the green curves representing 0% missing values imply 0 added missing values via MCAR mechanism. The raw data originally has missing values, which were imputed either via median or SVD imputation and hence the difference in the two curves. We also experimented with case wise deletion before applying the S2N filter. The performance of case wise deletion was very similar to that of median imputation and hence those curves are not included in the figure.

The results on the diabetes dataset strengthen our claims that SVD leads to the generation of more false positives when the percentage of missing values in data is high. Hence, using SVD for imputing missing values would lead to a higher number of irrelevant features, which can be falsely detected as relevant.

8. Conclusions

Missing data arises in almost all analyses nowadays, and thus has become an undeniably ubiquitous problem, that can not be handled with the simple exclusion of those cases containing missing values of variables, a strategy that is known as listwise deletion (LWD). LWD is widely used because of its distribution-preserving properties, but has the drawback of excluding potentially a large fraction of the data. Thus, several methods for imputing missing values have been developed, whose adequacy and validity depend on various assumptions, that are easily violated. Imputing missing values before any data processing and particularly before performing feature selection when there is a large fraction of missing values (above 80%), is tempting. Yet this is precisely when it is important to be cautious. The newly imputed values may introduce bias in data, with adverse effects on both type I errors (false positives) and type II errors (false negatives). This situation occurs even for the "nicest" type of missingness mechanisms: Missing Completely at Random (MCAR), in which the assumption is that the probability of missing data on a certain variable is unrelated to the value of the variable itself or to the values of any other variables in the dataset.

The types of bias that can be introduced by imputation are of two different natures: statistical and structural. Statistical bias results in an improper estimation of variance and/or co-variance between variables and can be corrected either analytically of by multiple imputation. Our proposal is a modified t-statistic which takes into account the uncertainty of linear regression imputation analytically, for the case of univariate feature selection of continuous features and a binary target variable. The new t-statistic captures both the uncertainty due to the limited accuracy of the regression coefficients (estimated from a small data sample) and the residual of the fit. The other type of bias, structural bias, is more insidious. It stems from the reversal of causal arrows by the imputation mechanism and can result in an increasing rate of false positives. For problems of prediction, this may not be an issue. But for problems of discovery, when a large fraction of variable values are missing, it is not advisable to use imputation methods such as regression or SVD, if one wants to avoid increasing the false discovery rate. Future work includes devising novel feature selection methods robust to missing data, without requiring imputation of missing values. Also, the problem of pre-processing data using standardization or normalization, in the presence of missing values, also deserves further attention and investigation.

9. Acknowledgements

We thank our colleagues of the Center for Data Science and in particular Balázs Kégl for supporting the launching event of the Paris-Saclay workgroup on missing data and for supporting the visit of Kristin Bennett. We are particularly grateful to the Isis-Diab collaborative group for sharing their data with us and for help and numerous discussions with Alain-Jacques Valleron, Pierre Bougnères, Sophie Le Fur, and Sophie Valtat. This research has also been financially supported in part by the Spanish Ministerio de Economía y Competitividad (research project TIN2015-65069-C2-1-R), the Consellería de Industria of the Xunta de Galicia (projects GRC2014/035 and ED431G/01), and the European Union (European Regional Development Funds - ERDF).

10. References

References

[1] P. D. Allison, Missing data, Sage Publishers, 2002.

[2] C. K. Enders, Applied missing data analysis, Guilford Press, 2010.

[3] S. Seaman, J. Galati, D. Jackson, J. Carlin, What is meant by missing at random?, Statistical Science 28 (2) (2013) 257–268. `doi:10.1214/13-STS415`.

[4] R. J. Little, D. B. Rubin, Statistical analysis with missing data, John Wiley & Sons, 2014.

[5] N. Longford, Missing data and small-area estimation: Modern analytical equipment for the survey statistician, Springer Science & Business Media, 2006.

[6] J. Pearl, K. Mohan, Recoverability and testability of missing data: Introduction and summary of results, Available at SSRN 2343873.

[7] I. Shpitser, K. Mohan, J. Pearl, Missing data as a causal and probabilistic problem, in: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, Amsterdam, Netherlands, 2015, pp. 802–811.

[8] K. Mohan, J. Pearl, T. Jin, Missing data as a causal inference problem, in: Proceedings of the Neural Information Processing Systems Conference (NIPS), 2013.
URL https://ssrn.com/abstract=2343794

[9] C. K. Enders, Multiple imputation as a flexible tool for missing data handling in clinical research, Behaviour research and therapy 98 (2017) 4–18.

[10] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, I. Petersen, Missing data and multiple imputation in clinical epidemiological research, Clinical epidemiology 9 (2017) 157.

[11] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, J. R. Carpenter, Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, Bmj 338 (2009) b2393.

[12] C. D. Nguyen, L. Strazdins, J. M. Nicholson, A. R. Cooklin, Impact of missing data strategies in studies of parental employment and health: Missing items, missing waves, and missing mothers, Social Science & Medicine 209 (2018) 160–168.

[13] H. Tomita, H. Fujisawa, M. Henmi, A bias-corrected estimator in multiple imputation for missing data, Statistics in Medicine.

[14] D. B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[15] J. Schafer, J. W. Graham, Missing data: our view of the state of the art., Psychological methods 7 (2) (2002) 147.

[16] S. García, J. Luengo, F. Herrera, Data preprocessing in data mining, Springer, 2015.

[17] S. Van Buuren, Flexible imputation of missing data, CRC press, 2012.

[18] P. Royston, et al., Multiple imputation of missing values: update, Stata Journal 5 (2) (2005) 188.

[19] P. McKnight, K. McKnight, S. Sidani, A. Figueredo, Missing data: A gentle introduction, The Guildford Press, 2007.

[20] G. Doquire, M. Verleysen, Feature selection with missing data using mutual information estimators, Neurocomputing 90 (2012) 3 – 11.

[21] G. Doquire, M. Verleysen, Feature selection with missing data using mutual information estimators, Neurocomputing 90 (2012) 3–11.

[22] I. Guyon, S. Gunn, A. Ben-Hur, G. Dror, Result analysis of the NIPS 2003 feature selection challenge, in: NIPS, 2004, pp. 545–552.

[23] A. F. Hayes, J. Matthes, Computational procedures for probing interactions in ols and logistic regression: SPSS and SAS implementations, Behavior Research Methods 41 (2009) 924–936.

[24] J. Sessa, D. Syed, Techniques to deal with missing data, in: 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016, pp. 1–4. `doi:10.1109/ICEDSA.2016.7818486`.

[25] Y. Dong, C.-Y. Peng, Principled missing data methods for researchers, SpringerPlus 2 (2013) 222.

[26] J. Josse, F. Husson, missMDA: A package for handling missing values in multivariate data analysis, Journal of Statistical Software, Articles 70 (1) (2016) 1–31. `doi:10.18637/jss.v070.i01`.
URL `https://www.jstatsoft.org/v070/i01`

[27] K. M., A. Benczùr, K. Csalogàny, Methods for large scale SVD with missing values, in: Proceedings of KDD Cup and workshop, Vol. 12, 2007, pp. 31–38.

[28] M. K., M. Mohamad, S. bin Denis, A review on missing value imputation algorithms for microarray gene expression data, Current Bioinformatics 9 (1) (2014) 18–22.

[29] T. O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (6) (2001) 520–525.

[30] Y. LeCun, C. Cortes, C. J.C. Burges, Mnist database of handwritten digits, `http://yann.lecun.com/exdb/mnist/`, accessed: 2016-11-22.

[31] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, science 286 (5439) (1999) 531–537.

[32] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of machine learning research 3 (Mar) (2003) 1157–1182.

[33] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Feature selection for high-dimensional data, Springer-Verlag, 2015.

[34] I. Guyon, A. Elisseeff, An introduction to feature extraction, Feature extraction (2006) 1–25.

[35] A. Banerjee, U. B. Chitnis, S. L. Jadhav, J. S. Bhawalkar, S. Chaudhury, Hypothesis testing, type i and type ii errors, Industrial Psychiatry Journal 18 (2) (2009) 127–131.

[36] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19).

[37] J. C. Deville, C. E. Sändal, Variance estimation for the regression imputed Horvitz-Thompson estimator 10 (1994) 381–394.

[38] J. N. K. Rao, Variance estimation in the presence of imputation for missing data, in: AMSTAT, 2000.

[39] J.-K. Kim, Variance estimation after imputation 27 (2001) 75–83.

[40] A. Donders, G. van der Heijdenc, T. Stijnend, K. Moons, A gentle introduction to imputation of missing values, Journal of Clinical Epidemiology 59 (10) (2006) 1087–1091.

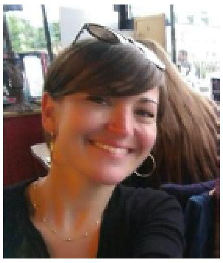
[41] F. Balazard, S. Le Fur, S. Valtat, A. J. Valleron, P. I. c. G. Bougnères, Association of environmental markers with childhood type 1 diabetes mellitus revealed by a long questionnaire on early life exposures and lifestyle in a case-control study, BMC Public Health 16.

Borja Seijo- Pardo, obtained his B.S. degree in Computer Science from University of A Coruña (Spain) in 2015. He is currently a Ph.D. student in the Department of Computer Science at University of A Coruña. His research interests include machine learning, ensemble approaches, feature selection and missing data.

Amparo Alonso-Betanzos received the Ph.D. degree for her work in the area of medical expert systems in 1988 at the University of Santiago de Compostela (Spain). Later, she was a postdoctoral fellow in the Medical College of Georgia, Augusta (USA). She is currently a Full Professor in the Department of Computer Science, University of A Coruña (Spain). Her main current areas are intelligent systems, scalable machine learning and feature selection.

Kristin Bennett, is a Professor at the Department of Mathematical Sciences, Department of Computer Sciences of the Lally School of Management at Rensselaer Polytechnic Institute, Troy, New York (USA): Her research interests are Combining operations research and artificial intelligence problem solving methods. Mathematical programming approaches to problems in data science, data mining, artificial intelligence and machine learning such as machine learning, support vector machines, neural networks, pattern recognition, and planning. Application of these techniques to medical, financial and scientific problems. Applications of machine learning to systems biology, cheminformatics, bioinformatics, tissue engineering, molecular epidemiology, and population biology.

Verónica Bolón-Canedo received her B.S. (2009), M.S. (2010) and Ph.D. (2014) degrees in Computer Science from the University of A Coruña (Spain). After a Postdoctoral fellowship in the University of Manchester, UK (2015) she is currently teaching in the Department of Computer Science of the University of A Coruña. She has extensively published in the area of machine learning and feature selection. On these topics she has co-authored one book, one book chapter, and more than 50 research papers in international conferences and journals.



Julie Josse, is a professor of Statistics at the Applied Math department at Ecole Polytechnique (Saclay) and member of the data-science initiative and XPOP INRIA team. Her main research fields are missing values, visualization with dimensionality reduction (PCA, correspondence analysis), multi-blocks data, low rank matrix estimation, questionnaire analyses. She has specialized in missing data, visualization and the nonparametric analyses of complex data structures. Her work was rewarded by a European Union grant in 2013 to increase her research potential and to spend a year at Stanford University. She spent a year as a researcher in INRIA before joining Polytechnique in 2016.

Mehreen Saeed is an associate professor at the Department of Computer Science and has been teaching at NUCES, Lahore Campus since December 2006. She has an MSc. in Computer Science from Quaid-e-Azam University, Islamabad. Her doctorate degree was completed in 1999 from the department of Engineering Mathematics, University of Bristol, UK. Her research interests include AI, Computer Vision and Machine Learning. She is also a member of the board of directors of Chalearn.

Isabelle Guyon, is professor of data science at the Université Paris-Saclay (UPSud/INRIA, Orsay), specialized in statistical data analysis, pattern recognition and machine learning. Her areas of expertise include computer vision, bioinformatics, and power systems. Her recent interest is in applications of machine learning to the discovery of causal relationships. Prior to joining Paris-Saclay she worked as an independent consultant and was a researcher at AT&T Bell Laboratories, where she pioneered applications of neural networks to pen computer interfaces (with collaborators including Yann LeCun and Yoshua Bengio) and co-invented with Bernhard Boser and Vladimir Vapnik Support Vector Machines (SVM), which became a textbook machine learning method. She is also the primary inventor of SVM-RFE, a variable selection technique based on SVM. The SVM-RFE paper has thousands of citations and is often used as a reference method against which new feature selection methods are benchmarked. She also authored a seminal paper on feature selection that received thousands of citations. She organized many challenges in Machine Learning since 2003 supported by the EU network Pascal2, NSF, and DARPA, with prizes sponsored by Microsoft, Google, Facebook, Amazon, Disney Research, and Texas Instrument. Isabelle Guyon holds a Ph.D. degree in Physical Sciences of the University Pierre and Marie Curie, Paris, France. She is president of Chalearn, a non-profit dedicated to organizing challenges, vice-president of the Unipen foundation, adjunct professor at New-York University, action editor of the Journal of Machine Learning Research, editor of the Springer series of Challenges in Machine Learning, program co-chair of NIPS 2016 and general co-chair of NIPS 2017.