Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

When to Impute? Imputation before and during cross-validation

Byron C. Jaeger*[1] | Nicholas J. Tierney[2] | Noah R. Simon[3]

[1]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama
[2]Very Important Stuff Committee, Institute B, City B, Country B
[3]Department of Biostatistics, University of Washington, Seattle, Washington

Correspondence
*Byron C. Jaeger. Email: bcjaeger@uab.edu

Present Address
327M Ryals Public Health Building 1665 University Blvd Birmingham, Alabama 35294-0022

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut elit odio. Donec fermentum tellus neque, vitae fringilla orci pretium vitae. Fusce maximus finibus facilisis. Donec ut ullamcorper turpis. Donec ut porta ipsum. Nullam cursus mauris a sapien ornare pulvinar. Aenean malesuada molestie erat quis mattis. Praesent scelerisque posuere faucibus. Praesent nunc nulla, ullamcorper ut ullamcorper sed, molestie ut est. Donec consequat libero nisi, non semper velit vulputate et. Quisque eleifend tincidunt ligula, bibendum finibus massa cursus eget. Curabitur aliquet vehicula quam non pulvinar. Aliquam facilisis tortor nec purus finibus, sit amet elementum eros sodales. Ut porta porttitor vestibulum. Integer molestie, leo ut maximus aliquam, velit dui iaculis nibh, eget hendrerit purus risus sit amet dolor. Sed sed tincidunt ex. Curabitur imperdiet egestas tellus in iaculis. Maecenas ante neque, pretium vel nisl at, lobortis lacinia neque. In gravida elit vel volutpat imperdiet. Sed ut nulla arcu. Proin blandit interdum ex sit amet laoreet. Phasellus efficitur, sem hendrerit mattis dapibus, nunc tellus ornare nisi, nec eleifend enim nibh ac ipsum. Aenean tincidunt nisl sit amet facilisis faucibus. Donec odio erat, bibendum eu imperdiet sed, gravida luctus turpis.

KEYWORDS:
Class file; LaTeX; Statist. Med.; Rmarkdown;

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
```

1 | INTRODUCTION

In evaluating the performance of complex predictive models, it is now well understood that so-called training error (the predictive error measured on observations used to fit the model) is a poor proxy for generalization error (the performance of the model on future, as-yet-unseen, observations). The training error of a model will often be overly optimistic for the generalization error. For more complex modeling procedures (or on high dimensional data) this over-optimism can be quite substantial. As such, it

is now standard to use sample-splitting methods to estimate generalization error. These methods train and evaluate models on separate data. Cross-validation (CV) is a common sample-splitting technique that repeatedly splits a single dataset into training and validation subsets to efficiently estimate generalization error.

For multi-step modeling procedures, it is recommended that the entire sequence of steps be carried out during each replicate of CV to mimic the application of the entire procedure to an independent test set. This is particularly important for supervised steps (those steps that use the outcome variable). It is thought that most unsupervised steps (*i.e.,* steps that do not access information about outcome variables) can be completed before conducting CV with only minimal bias. This has been empirically validated for PCA and variance-based dimension reduction, ??? (we need to fill in some stuff here)

Missing data (MD) occur frequently in applied settings, and several machine learning algorithms (e.g., regression) are incompatible with MD. Imputation is a technique that replaces MD with estimated values. Imputation of MD is often one of the most computationally expensive steps in modeling procedures. For example, the `missForest` imputation algorithm may fit one random forest model for each column that contains MD. Computational expense of applying `missForest` or other imputation strategies with similar complexity during each replicate of CV scales poorly to larger datasets and to higher numbers of replications. High computational cost may lead analysts to prefer more convenient but less effective strategies to handle MD. A more computationally efficient approach would be to implement 'unsupervised imputation' (*i.e.,* imputing MD without accessing outcome information) *before* conducting CV. However, there is a lack of evidence substantiating the claim that unsupervised imputation conducted before CV will provide reliable estimates of a multi-step modeling procedure's generalization error. More importantly, there is a lack of evidence outlining scenarios when unsupervised imputation before CV can lead to biased estimates of model error.

In this manuscript, we identify scenarios when reliable model error estimates can (or cannot) be computed using unsupervised imputation *before* CV (a strategy we will refer to as $\texttt{I} \rightarrow \texttt{CV}$). We compare estimated model error according to $\texttt{I} \rightarrow \texttt{CV}$ with estimated model error when imputation is applied *during each replicate* of CV (a strategy hereafter referred to as $\texttt{CV} \circlearrowright \texttt{I}$). To objectively compare the accuracy of model error estimates using $\texttt{I} \rightarrow \texttt{CV}$ and $\texttt{CV} \circlearrowright \texttt{I}$, we compute 'true' model error using an external validation set with MD. Both simulated and real data are leveraged to draw these comparisons. Our analysis also introduces and applies the `ipa` R package (**i**mputation for **p**redictive **a**nalytics), which provides functions to create single or multiple imputed training and testing sets for prediction modeling.

The rest of this manuscript is organized as follows. In Section 5, we summarize results from experiments using real data from public repositories. Results are summarized alongside our recommendations for applied practice in Section 6

2 | MISSING DATA AND CROSS-VALIDATION

MD mechanisms were first formalized by Rubin, who developed a framework to analyze MD that supposes each data point has some probability of being missing. If the probability of missingness is unrelated to the data (*i.e.,* all data are equally likely to be missing), then the data are missing completely at random (MCAR). When the probability of missingness is related to observed variables in the data (*i.e.,* all data within observed groups are equally likely to be missing), the data are missing at random (MAR). If the probability of missingness is determined by reasons that are unknown or unobserved, the data are missing not at random (MNAR). To illustrate, if a doctor did not run labs for a patient because the clinic was too crowded at the time, the patient's data are MCAR. If instead the doctor chose not to measure the patient's labs because the patient was too young, the patient's data are MAR. If the patient missed the appointment because the patient was too sick, the patient's data are MNAR. In the context of statistical learning, previous findings have shown that when data are MNAR, imputation alone is often less efficient than incorporating features that characterize missing patterns (*e.g.,* missingness incorporated as an attribute)[1,2,3]. Since the primary aim of the current study is to assess the differences between two implementations of imputation (*i.e.,* $\texttt{I} \rightarrow \texttt{CV}$ and $\texttt{CV} \circlearrowright \texttt{I}$), we focus analyses on cases where data are MAR or MCAR.

The mechanisms of MD and methods to engage with MD have been studied thoroughly in the context of statistical inference. The primary focus in this setting is to obtain valid test statistics for statistical hypotheses in the presence of MD. Imputation to the mean and, more broadly, MD strategies that create a single imputed value, have been shown to increase type I errors for inferential statistics by artificially reducing the variance of observed data and ignoring the uncertainty attributed to MD, respectively. Multiple imputation, a widely recommended strategy to handle MD for statistical inference, is capable of producing valid test statistics when data are MCAR or MAR because it can simultaneously address the two shortcomings listed in the previous sentence. It is notable that the 'accuracy' of imputed values is not critical for the success of multiple imputation, given sufficient estimates of conditional distributions[4]. Instead, it is the consistency of the estimated covariance matrix for regression coefficients that makes this strategy ideal for statistical inference.

In the context of supervised statistical learning, data often include a training set and an external validation set. commonly ordered steps for model development are (1) data pre-processing, (2) tuning model hyper-parameters, (3) training the model(s) on the full training set, and (4) validating the model(s) using the validation set. Handling MD is a pre-processing task conducted in the training and validation sets, separately. In contrast to statistical inference, single imputation is often used in supervised learning workflows and strategies with greater imputation 'accuracy' are preferred. When k-fold CV is applied, the training data are split into k non-overlapping subsets (*i.e.,* folds). Each fold is then used as a testing set for the model(s) developed on the $k - 1$ other folds. Aggregating model errors over all k steps provides an estimate of the modeling procedure's generalization error, making k-fold CV an ideal resampling technique for tuning model hyper-parameters or comparing different modeling

procedures. To mimic the application of the entire modeling procedure to an independent validation set, it is recommended that all data pre-processing steps be executed in each replicate of CV. Following CV, the generalization error of the modeling procedure(s) is measured by applying the 'tuned' modeling procedure to the entire training set and then evaluating the resulting prediction function's accuracy in the external validation set.

% additional text that I left of the previous paragraph: Although multiple imputation has been shown to increase the prediction accuracy of downstream models, the computational overhead of multiple imputation is not ideal. It is generally assumed that the accuracy of downstream models trained on the imputed data will be increased by making the imputed data more closely resemble the unobserved complete data.

There are at least two ways to evaluate the effectiveness of an imputation strategy. One approach is to evaluate how accurately the strategy imputes unobserved data. However, imputation accuracy can only be measured when the values of unobserved data are known, which limits its practical use. Another approach is to evaluate the generalization error of models fitted to the imputed data. This approach is applicable in any setting involving supervised learning. Evaluating imputation strategies in this manner also allows imputation parameters to be 'tuned' in the same manner as parameters in the prediction model. As the focus of this manuscript is to investigate whether imputation can occur before CV without biasing estimates for generalization error, we do not highlight the accuracy of imputed values and focus exclusively on generalization error.

3 | ORDER OF OPERATIONS

When employing supervised learning methods with split-sample-validation or cross-validation, it is critical that training data is separated from validation data before any "learning" is done: *The entire supervised pipeline must be run using only training data*. This applies both to the fitting procedure (eg. applying linear regression, or random forests, etc…), and to any supervised preprocessing steps. There are a number of examples in the literature where wildly optimistic estimates of validation error have been obtained because supervised variable selection was performed on the entire dataset, rather than just the training sample (CITE). In scenarios with a larger number of features, even simple methodologies, eg. selecting those features with high individual correlation with outcome, can induce substantial bias (CITE).

It is not unusual to separate training and validation data before performing even unsupervised preprocessing steps. For example, in many supervised learning frameworks, it is common to center and scale features, such that they are mean 0, variance 1 before engaging with the outcome. As this does not involve the outcome, it is entirely unsupervised. Nonetheless to most accurately replicate the "fitting process" in each replicate of CV, it is common to learn the centering and scaling parameters for each feature in each training subset, and apply those parameters to center/scale features in the corresponding validation subset (CITE).

3.1 | Testing data

4 | SIMULATED EXPERIMENTS

Due in part to the plurality of methods used for prediction modeling and resampling, it is difficult to obtain meaningful theoretical results related to $\texttt{I}\rightarrow\texttt{CV}$ and $\texttt{CV}\circlearrowleft\texttt{I}$. The goal of this section is to assess empirical findings comparing these two strategies. Our primary goal is to measure and compare how well each strategy approximates a model's generalization error. A secondary goal is to monitor the computation time of each approach and conduct relative comparisons. For these empirical experiments, we vary multiple parameters that may impact results, including the number of potential predictor variables, the number of training observations, and the missing mechanism (MCAR or MAR). As the difference between $\texttt{I}\rightarrow\texttt{CV}$ and $\texttt{CV}\circlearrowleft\texttt{I}$ grows negligible when there are few missing data, we focus our comparisons to settings where 90% of observations contain at least 1 missing value and 30-50% of the data are missing. We apply nearest-neighbor imputation to handle MD and least absolute shrinkage and selection operator (LASSO) regression to develop prediction functions throughout the simulated experiments since these methods are widely recommended and used in applied settings. To generate data that matched the structure of our modeling procedure, we generated outcomes using linear effects without interaction. To be comprehensive, a separate appendix investigates the same simulation study where random forests

In addition to varying parameters in the data-generation process, we also consider three general data-generation scenarios. In scenario 1, the observed data are independent and identically distributed (iid). In scenario 2, the data are iid conditional on an observed grouping variable. A total of 11 groups are formed, one in the validation set and the remaining 10 in the training set. Each group is characterized by a unique mean value for predictor variables. During CV, the observed groups are separated into ten folds to mimic the prediction of outcomes in a population with different characteristics. Scenario 3 is identical to scenario 2 in all regards except that the grouping variable is latent. Consequently, CV does not break the observed groups into separate folds for scenario 3. Scenario 1 is an idealistic setting where $\texttt{I}\rightarrow\texttt{CV}$ and $\texttt{CV}\circlearrowleft\texttt{I}$ should provide almost identical estimates of generalization error. Scenarios 2 and 3 are meant to test whether $\texttt{I}\rightarrow\texttt{CV}$ produces biased estimates of generalization error by omitting imputation from the CV process. In these scenarios, it is theoretically clear that $\texttt{CV}\circlearrowleft\texttt{I}$ is more appropriate, but the degree to which $\texttt{I}\rightarrow\texttt{CV}$ is inappropriate has not yet been quantified and may be negligible in larger samples.

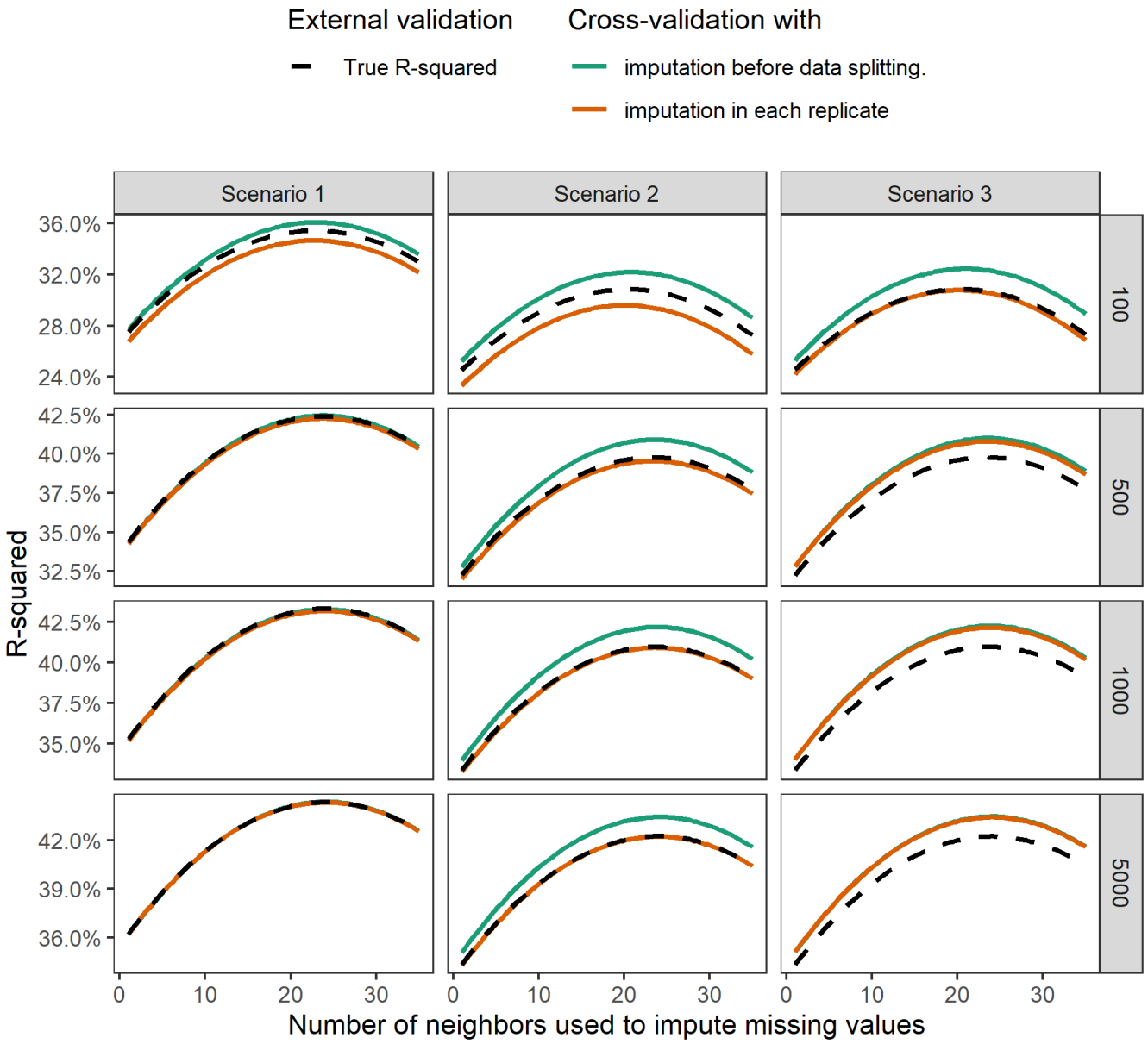

FIGURE 1 External generalization error and internal estimates of generalization error using I→CV and CV↻I

4.1 | Results

5 | REAL DATA EXPERIMENTS

6 | DISCUSSION AND RECOMMENDATIONS

References

1. Twala B, Jones M, Hand DJ. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters* 2008; 29(7): 950–956.

2. Twala B. An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence* 2009; 23(5): 373–405.

3. Tang F, Ishwaran H. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 2017; 10(6): 363–377.

4. Van Buuren S. *Flexible imputation of missing data*. CRC press . 2018.