
Imputation of Clinical Covariates in Time Series

Dimitris Bertsimas

Sloan School of Management and Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02142
dbertsim@mit.edu

Agni Orfanoudaki

Sloan School of Management and Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02142
agniorf@mit.edu

Colin Pawlowski

Sloan School of Management and Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02142
cpawlows@mit.edu

Abstract

Missing data is a common problem in real-world settings and particularly relevant in healthcare applications where researchers use Electronic Health Records (EHR) and results of observational studies to apply analytics methods. This issue becomes even more prominent for longitudinal data sets, where multiple instances of the same individual correspond to different observations in time. Standard imputation methods do not take into account patient specific information incorporated in multivariate panel data. We introduce the novel imputation algorithm `med.impute` that addresses this problem, extending the flexible framework of `opt.impute` suggested by Bertsimas et al. [1]. Our algorithm provides imputations for data sets with missing continuous and categorical features, and we present the formulation and implement scalable first-order methods for a K -NN model. We test the performance of our algorithm on longitudinal data from the Framingham Heart Study when data are missing completely at random (MCAR). We demonstrate that `med.impute` leads to significant improvements in both imputation accuracy and downstream model AUC compared to state-of-the-art methods.

1 Introduction

Despite the implementation of complex Electronic Healthcare Records (EHR) Systems, missing data are ubiquitous in clinical epidemiological research [2] posing considerable challenges in the analyses and interpretation of results and potentially weakening their validity [3]. Often, those data sets contain numerous visits of the same person corresponding to various patterns of missing data. This particularity challenges state-of-the-art missing data methods that do not consider the connection of multiple observations to the same individual [4].

A variety of machine learning approaches have been introduced in the literature to deal with missing data. The simplest approach is the mean imputation that uses the mean of the observed values

to replace the missing for the same covariate [5]. Another common method called `bpca` uses the singular value decomposition of the data matrix and information from a prior distribution on the model parameters to impute the missing values [6]. Recent studies, though, show that such methods can lead to seriously misleading results, advising to consider multiple imputation [7, 5]. The latter, implemented in the package `mice` [8], allows for uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them [9]. However, multiple imputation methods are slower and require pooling results, which may not be appropriate for certain applications. Bertsimas et al. [1] showed that a general optimization framework with a predictive model-based cost function can explicitly handle both continuous and categorical variables and can be used to generate single as well as multiple imputations. This optimization perspective leads to new scalable algorithms for more accurate data imputation.

The algorithms above are not tailored to multivariate time series data sets even though in time series prediction missing values and their missing patterns are often correlated with the target labels [10]. Nevertheless, preliminary work was done in demonstrating their performance in that setting [11]. Recurrent Neural Network approaches have also been employed [10, 4] when the missing values were imputed as part of a prediction task tailored to the particular data set.

Given multivariate time-series data, we develop a novel imputation method that utilizes traditional optimization and machine learning techniques leading to superior performance over state-of-the-art algorithms. We formulate the problem of missing data imputation with time series information as a family of optimization problems under the `med.impute` framework. We derive fast first-order solutions to these problems for K -NN which can be easily extended to SVM and tree models. We run experiments on the Framingham Heart Study, a large-scale longitudinal clinical study, focusing on downstream models which predict 10-year risk of stroke. We demonstrate that `med.impute` leads to significant improvements in imputation accuracy and downstream model accuracy.

2 Methods

2.1 Framework Definition

We consider the single imputation problem, for which our task is to fill in the missing values of data set $\mathbf{X} \in \mathbb{R}^{n \times p}$ with n observations (rows) and p features (columns). We assume that the first p_0 features are continuous, with missing and known indices $\mathcal{M}_0, \mathcal{N}_0$ respectively, and that the next $p_1 = p - p_0$ features are categorical, with missing and known indices $\mathcal{M}_1, \mathcal{N}_1$ respectively.

In addition, we assume that each observation i corresponds to an individual $y_i \in \{1, \dots, M\}$. For data sets with multiple observations of individuals over time, we have $M < n$. We define $t_i \in \mathbb{R}^+$ as the number of (days/months/years) after a reference date that observation i was recorded. It follows that $|t_i - t_j|$ is the time difference in (days/months/years) between observations i and j .

For each feature $d = 1, \dots, p$, we introduce the parameters α_d, λ_d . The first parameter $\alpha_d \in [0, 1]$ is the relative weight given to the time series component of the `med.impute` objective function for variable d . At the extremes, $\alpha_d = 0$ corresponds to imputing feature d under the original `opt.impute` objective, and $\alpha_d = 1$ corresponds to imputing feature d using each individual’s time series information independently.

The second parameter $\lambda_d \in (0, 1]$ is the exponential time decay parameter for variable d . We introduce this parameter so that observations from the same individual at nearby points in time will be weighted most heavily in the imputation. For each pair of observations i, j , we define

$$C_{ijd} = \begin{cases} \lambda_d^{|t_i - t_j|}, & \text{if } y_i = y_j, \\ 0, & \text{otherwise.} \end{cases}$$

These constants will be coefficients in the time series component of the objective function. We learn α_d and λ_d via cross-validation.

2.2 A K -NN Formulation of the Problem

Given the general optimization-based imputation model suggested in [1], we present an adjusted formulation that accounts for multiple instances of the same observation in time under the K -NN framework.

In order to weight instances of the same person in the imputation model, we will add a penalty term to the objective, with different weights α_d for each dimension $d = 1, \dots, p$. The key decision variables are the imputed continuous values $\{w_{id} \in \mathcal{M}_0\}$ and the imputed categorical values $\{v_{id} \in \mathcal{M}_1\}$. First, define the distance between observations i and j as

$$d_{ij} := \sum_{d=1}^{p_0} (w_{id} - w_{jd})^2 + \sum_{d=p_0+1}^{p_0+p_1} \mathbb{1}_{\{v_{id} \neq v_{jd}\}}. \quad (1)$$

Next, introduce the binary variables:

$$z_{ij} = \begin{cases} 1, & \text{if } j \text{ is among the } K\text{-nearest neighbors of } i \text{ with respect to distance metric (1),} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The `med.impute` formulation with the K -NN objective function is

$$\begin{aligned} \min \quad & \sum_{i \in \mathcal{I}} \sum_{j=1}^n z_{ij} \left(\sum_{d=1}^{p_0} (1 - \alpha_d) (w_{id} - w_{jd})^2 + \sum_{d=p_0+1}^{p_0+p_1} (1 - \alpha_d) \mathbb{1}_{\{v_{id} \neq v_{jd}\}} \right) \\ & + \sum_{i \in \mathcal{I}} \sum_{j=1}^n \left(\sum_{d=1}^{p_0} \alpha_d C_{ijd} (w_{id} - w_{jd})^2 + \sum_{d=p_0+1}^{p_0+p_1} \alpha_d C_{ijd} \mathbb{1}_{\{v_{id} \neq v_{jd}\}} \right) \\ \text{s.t.} \quad & w_{id} = x_{id} && (i, d) \in \mathcal{N}_0, \\ & v_{id} = x_{id} && (i, d) \in \mathcal{N}_1, \\ & z_{ii} = 0 && i \in \mathcal{I}, \\ & \sum_{j=1}^n z_{ij} = K && i \in \mathcal{I}, \\ & \mathbf{Z} \in \{0, 1\}^{|\mathcal{I}| \times n}, \end{aligned} \quad (3)$$

where $\mathcal{I} = \{i : \mathbf{x}_i \text{ has one or more missing values}\}$. At the optimal solution, the objective function is the sum of the distances from each point to its K -nearest neighbors with respect to distance metric (1), plus the sum of the distances from each point to other observations from the same individual. We use coordinate descent [12] with random restarts to find high quality solutions for this problem, alternatively updating the binary variables and the imputed values as in [1].

3 Real-world Experiments

3.1 Experimental Setup

To test the accuracy of our method, we run a series of computational experiments on data from the Framingham Heart Study (FHS), a large-scale longitudinal clinical study, focusing on downstream models which predict 10-year risk of stroke. We consider all individuals from the Original Cohort with 10 or more observations, which includes $M = 1,107$ unique patients. For each patient, we take the 10 most recent observations, so the data set has $n = 11,070$ observations total. We include $p = 13$ continuous (Body Mass Index, Systolic Blood Pressure, Hematocrit, etc.) and categorical (Gender, Smoking, etc.) covariates.

In our experiments, we generate patterns of missing data for various percentages ranging from 10% to 50% under the missing completely at random (MCAR) mechanism. We take the full data set to be the ground truth. We run some of the most commonly-used and state-of-the-art methods [1, 5, 6, 8] for imputation on these data sets to predict the missing values and compare against `med.impute`. We run further experiments to evaluate the impact of these imputations on the intended downstream machine learning task, which is to predict the 10-year risk of stroke given the most recent observation from

each patient. We compare the out-of-sample performance of an ℓ_1 -regularized logistic regression model fit using the imputed data sets for various levels of missing information. We run a second set of experiments with the missing percentage fixed at 50%, varying the number of observations per patient in the data set from 1 to 10. We report imputation accuracy and downstream task results for these experiments as well.

3.2 Results

In Figure 1, we show the results from the FHS experiments with varying levels of missing data. For all missing percentages considered, `med.impute` produces the imputations with the highest accuracy and the best performance on the downstream classification task. As the percentage of missing data increases, the relative improvement of `med.impute` over the reference methods increases. At 50% missing data, the mean absolute error (MAE) of `med.impute` is 0.343, compared to 0.509 for the next best method `opt.impute`. Further, at 50% missing data, the area under the curve (AUC) of the logistic model trained using the `med.impute` imputation is 0.848, compared to 0.768 for the next best methods `mean` and `bpca`. These results demonstrate that `med.impute` is able to leverage time series information to gain a substantial edge over methods which ignore this information.

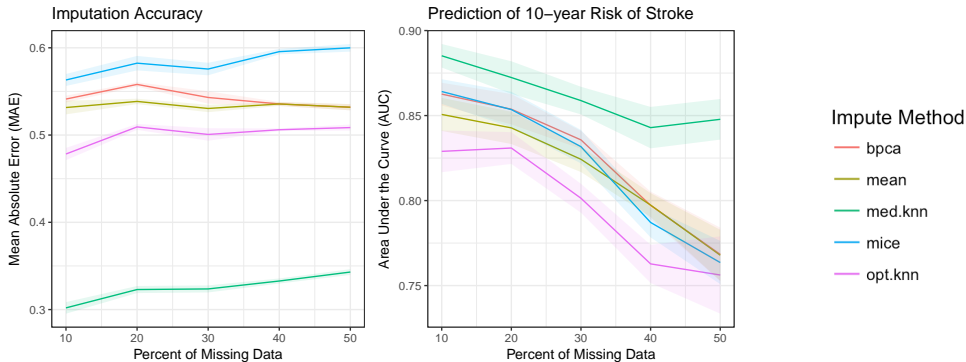


Figure 1: Results from experiments on the Framingham Heart Study data set with 10 observations per patient, varying the percentage of missing data from 10% to 50%.

In Figure 2, we show the results from the FHS experiments with varying numbers of observations per patient (OPP). For the experiment with $OPP = k$, we take the k most recent observations for each patient. When $OPP = 1$, `med.impute` is equivalent to the `opt.impute` method. We observe that the MAE of `med.impute` decreases significantly as OPP goes up to 4-5, and then increases slightly beyond this point. Because the observations in FHS data set occur every 2 years, this suggests that the past 6-8 years of data are most useful for imputing an individual’s clinical covariates. Similarly, the AUC from `med.impute` peaks when $OPP = 4$, and then levels off around 0.85. In contrast, the AUC of the reference methods declines slightly as the OPP increases, indicating that adding more observations to the data set does not help traditional imputation methods in this case.

4 Conclusions

We propose a new imputation algorithm for multivariate data in time series that yields high quality solutions using a K -NN framework combined with fast first-order methods. Through computational experiments with real-world data sets from the Framingham Heart Study, we show that `med.impute` yields statistically significant gains in imputation quality over state-of-the-art imputation methods, which leads to improved out-of-sample performance on downstream tasks. In future work, we can extend this algorithm to incorporate time series information in SVM and Trees models.

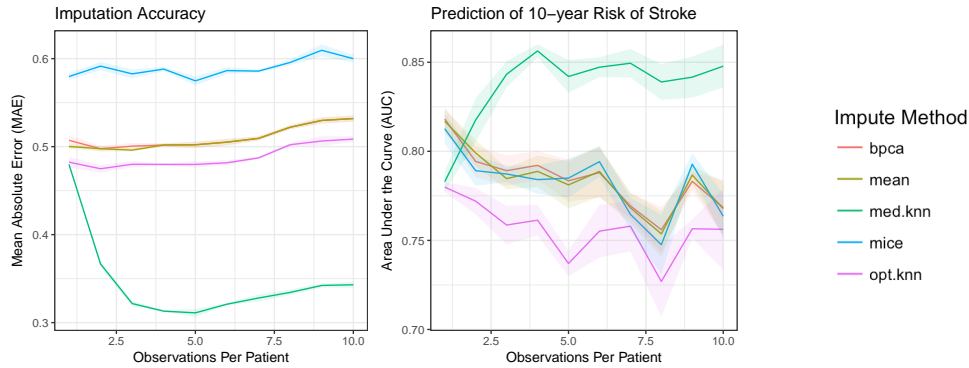


Figure 2: Results from experiments on the Framingham Heart Study data set with 50% missing data, varying the number of observations per patient from 1 to 10.

References

- [1] Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. “From Predictive Methods to Missing Data Imputation: An Optimization Approach”. In: *Journal of Machine Learning Research* 18.196 (2018), pp. 1–39. URL: <http://jmlr.org/papers/v18/17-073.html>.
- [2] Alma B Pedersen et al. “Missing data and multiple imputation in clinical epidemiological research”. In: *Clinical Epidemiology* 9 (2017). DOI: 10.2147/CLEP.S129785. URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5358992/>.
- [3] Angela M Wood, Ian R White, and Simon G Thompson. “Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals”. In: *Clinical Trials* 1.4 (2004). PMID: 16279275, pp. 368–376. DOI: 10.1191/1740774504cn032oa. eprint: <https://doi.org/10.1191/1740774504cn032oa>. URL: <https://doi.org/10.1191/1740774504cn032oa>.
- [4] Zhengping Che et al. “Recurrent Neural Networks for Multivariate Time Series with Missing Values”. In: *Scientific Reports* 8.1 (2018), p. 6085. ISSN: 2045-2322. DOI: 10.1038/s41598-018-24271-9. URL: <https://doi.org/10.1038/s41598-018-24271-9>.
- [5] Robert J. Mislevy. In: *Journal of Educational Statistics* 16.2 (1991), pp. 150–155. ISSN: 03629791. URL: <http://www.jstor.org/stable/1165119>.
- [6] Shigeyuki Oba et al. “A Bayesian missing value estimation method for gene expression profile data”. In: *Bioinformatics* 19.16 (2003), pp. 2088–2096. DOI: 10.1093/bioinformatics/btg287. eprint: [/oup/backfile/content_public/journal/bioinformatics/19/16/10.1093/bioinformatics/btg287/2/btg287.pdf](http://oup/backfile/content_public/journal/bioinformatics/19/16/10.1093/bioinformatics/btg287/2/btg287.pdf). URL: <http://dx.doi.org/10.1093/bioinformatics/btg287>.
- [7] Kristel J.M. Janssen et al. “Missing covariate data in medical research: To impute is better than to ignore”. In: *Journal of Clinical Epidemiology* 63.7 (2010), pp. 721–727. ISSN: 0895-4356. DOI: <https://doi.org/10.1016/j.jclinepi.2009.12.008>. URL: <http://www.sciencedirect.com/science/article/pii/S0895435610000193>.
- [8] Stef van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67. URL: <https://www.jstatsoft.org/v45/i03/>.
- [9] Joseph L. Schafer and Maren K. Olsen. “Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst’s Perspective”. In: *Multivariate Behavioral Research* 33.4 (1998). PMID: 26753828, pp. 545–571. DOI: 10.1207/s15327906mbr3304_5. eprint: https://doi.org/10.1207/s15327906mbr3304_5. URL: https://doi.org/10.1207/s15327906mbr3304_5.
- [10] Zachary C Lipton, David Kale, and Randall Wetzell. “Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series”. In: *Proceedings of the 1st Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 56. Proceedings of Machine Learning Research. Children’s Hospital LA, Los Angeles, CA, USA:

PMLR, 18–19 Aug 2016, pp. 253–270. URL: <http://proceedings.mlr.press/v56/Lipton16.html>.

- [11] Zhongheng Zhang. “Multiple imputation for time series data with Amelia package”. In: *Annals of Translational Medicine* 4.3 (2016). ISSN: 2305-5847. URL: <http://atm.amegroups.com/article/view/8846>.
- [12] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.