Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Data and text mining

Imputation for Lipidomics and Metabolomics (ImpLiMet): Online optimization and method selection for missing data imputation

Huiting Ou[1-2+], Anuradha Surendra[3+], Graeme S.V. McDowell[3], Emily Hashimoto-Roth[4,6,7], Jianguo Xia[1,5], Steffany A.L. Bennett[4,6,8*], Miroslava Čuperlović-Culf[3,6*]

[1]Department of Human Genetics, McGill University, Montreal, Quebec, Canada
[2]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[3]National Research Council of Canada, Digital Technologies Research Centre, Ottawa, Ontario, Canada
[4]Neurolipidomcs Laboratory and India Taylor Lipidomic Research Platform, Ottawa, Ontario, Canada
[5]Institute of Parasitology, McGill University, Montreal, Quebec, Canada
[6]Department of Biochemistry, Microbiology, and Immunology and Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada
[7]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[8]Department of Cellular and Molecular Medicine, University of Ottawa Brain and Mind Research Institute, & Department of Chemistry and Biomolecular Sciences, Centre for Catalysis Research and Innovation, University of Ottawa, Ottawa, Ontario, Canada

+ Equal first authors *To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Missing values are often unavoidable in modern high-throughput measurements due to various experimental or analytical reasons. Imputation, the process of replacing missing values in a dataset with estimated values, plays an important role in multivariate and machine learning analyses. Three missingness patterns have been conceptualized: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Each describes unique dependencies between the missing and observed data. As a result, the optimal imputation method for each dataset depends on the type of data, the cause of the missing data, and the nature of relationships between the missing and observed data. The challenge is to identify the optimal imputation solution for a given dataset.

Results: Imputation for Lipidomics and Metabolomics (ImpLiMet) is a user-friendly UI-platform that enables users to impute missing data using eight different methods. Using the user's dataset, ImpLiMet can then suggest the optimal imputation solution through a grid search-based investigation of the error rate for imputation across three missingness data simulations. The effect of imputation can be visually assessed by principal component analysis (PCA) comparing the impact of removing all features and samples with missing data with the chosen imputation method.

Availability and implementation: ImpLiMet is freely available at https://complimet.ca/shiny/implimet/ with software accessible at https://github.com/complimet/ImpLiMet

Contact: steffanyann.bennett@uottawa.ca or miroslava.cuperlovic-culf@nrc-cnrc.gca.

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Missing data are a major problem for multivariate, machine learning (ML) and network analyses. For example, in large lipidomic or metabolic datasets, measurements for some analytes may not be available in every sample due to routine technical variability, low abundance, ion suppression from co-eluting analytes, inaccurate feature assignment in annotation pipelines, or because analytes are simply not present in a subset of samples. This "missingness" confounds ML approaches, limits the number of methodologies that can be utilized, and reduces the statistical power of models that exclude samples with missing values. Sample exclusion further alters cohort representation, notably when "missingness" is an indicator of a particular subgroup, biasing results

towards the groups in which all analytes are observed, and potentially leading to inaccurate interpretations (Jäger et al. 2021; Stoyanovich et al. 2020; Sterne et al. 2009)

Imputing missing values is a valid solution to the problems of "missingness" in multivariate analyses and can reduce data bias resulting from sample exclusion. Three types of missingness have been conceptualized that can be addressed by imputation (Mack et al. 2018; Scheffer 2002):

1) Missing completely at random (MCAR) refers to values whose absence is completely independent of any other data feature or covariate. In this type of missingness, each sample has the same probability of presenting an MCAR value because there is no underlying difference between the samples with or without missing data (Mack et al. 2018; Rubin 1976). A real-world example of MCAR is transient (aka random) technological failure over the course of data collection such that there is no relationship between the samples with missing or observed values.

2) Missing at random (MAR) refers to missing values whose absence are related to the values of other measured features but not to the measured values of the same feature (Schafer 1997). Here, missing values do not depend on the variable in question but on the values of the other analytes present in each sample. An example of MAR would be when the value for one analyte is missing because its measurement is obscured by the abundance of another analyte in the same sample (i.e., ion suppression of co-eluting analytes in the case of lipidomic or metabolomic datasets).

3) Missing not at random (MNAR) refers to missing values that are absent because of a feature, condition, or covariate that is responsible for the absence in that sample. Here, the probability of missingness depends on the sample itself. A biological example of this group would be analytes that are not synthesized, thus not present, in every condition. A technological example would be when analytes are present in a given sample but are below the limit of quantification of the technology used to measure the data.

Multiple imputation methods can approximate missing values (Bertsimas et al. 2018; Schafer 1997). Recently, both Jäger et al., and Chilimoniuk et al., have compared and evaluated different imputation approaches with respect to the quality of the imputed data and their downstream impact on ML pipelines (Chilimoniuk et al. 2024; Jäger et al. 2021). Jager et al. (Jäger et al. 2021) present method for testing imputation quality based on the error rate and downstream use of data and in their work show that in almost all shown examples Random-Forest (RF) provides the most optimal result. However, both Jager et al. and Chilimoniuk et al. show that the optimal imputation solution is dataset dependent, influenced by the type of missing data, the cause of the missing data, and the nature of relationships between the missing and observed data (Chilimoniuk et al. 2024; Jäger et al. 2021). With this knowledge, the bioinformatic challenge becomes to identify the optimal imputation solution for a given dataset.

To address this challenge, we present **Imp**utation for **Li**pidomics and **Met**abolomics – ImpLiMet. ImpLiMet is an R package available at https://github.com/complimet/ImpLiMet and online with UI at Computational Lipidomics and Metabolomics: CompLiMet: https://complimet.ca/shiny/implimet/. ImpLiMet enables users to impute missing data using 8 different methods. The effect of each method can be visualized by principal component analysis (PCA) comparing the impact of removing all features and samples with missing data to the chosen imputation method. To identify the optimal imputation solution, ImpLiMet further offers an optimization option wherein the error of each imputation method is evaluated, and the user is informed of the method with the lowest mean absolute percentage error (MAPE) across three "missingness" simulations for their dataset.

2 Implementation

ImpLiMet is written in R and deployed with a RShiny UI. It is compatible with all modern web browsers. Figure 1 presents the ImpLiMet workflow and pseudocode for the optimization procedure. To run ImpLiMet, the user uploads their dataset with missing values as a .CSV file. If the dataset includes features measured in different units by different platforms (multiple feature measurement groups), the user has the option to format their data such that the imputation methods consider features measured using different analyzers or in different units separately. An example of different measurement groups could be the combination of lipidomic and metabolomic data measured using different platforms. The user can then specify the number of features or samples with the selected percentage(s) % of missing values to be removed prior to choosing an imputation measure or optimizing across measures. Eight imputation methods are available: 1) replacing with the minimum by feature, 2) replacing with the minimum divided by 5 by feature, 3) replacing with the maximum by feature, 4) replacing with the median by feature, 5) replacing with the mean by feature, 6) using K-Nearest Neighbors (kNN) (Hastie, et al. 1999; Troyanskaya, et al. 2001), 7) using Random Forest (RF) (Pantanowitz, Marwala, 2009) or 8) using Multivariate Imputation by Chained Equations (MICE) (van Buuren, et al. 1999; van Buuren et al. 2006). For kNN, RF, and MICE, the user is queried to provide the number of neighbours for kNN, the number of trees for RF, and the number of iterations for MICE. kNN is implemented using *impute.KNN* function; RF imputation utilizes *missRanger.RF* function (Stekhoven, Buehlman, 2011; Wright et al. 2017) and MICE function mice (van Buuren et al. 2011).

If users have at least 3 features and 6 samples with no missing values in the dataset, ImpLiMet further offers an optimization option wherein the error of each imputation method is evaluated by simulating the three different sources of missingness in the user's dataset once all missing data is removed and testing all available imputation methods. Optimization then suggests as the best imputation method the one with the lowest mean absolute percentage error (MAPE) across the three "missingness" data simulations. The selected approach is used to impute the original dataset and this result is provided as a download.

In the optimization step, samples without any missing values are selected to create a complete set. If the cleaned dataset obtained by removing all samples (rows) with missing values has no remaining values, optimization instead selects features (columns) without missing values. If there are less than at least 6 samples and 3 features remaining, optimization of imputation cannot be done. From the dataset devoid of missing features, ImpLiMet removes data values at the sample threshold percentage initially provided by the user for filtering. If threshold percentage is not provided, i.e., user opts not to remove any features or samples from their dataset prior to imputation, ImpLiMet uses 30% as the threshold percentage in the optimization process. Threshold percentage is used to simulate the optimal imputation method for a given dataset at the level of the user's specified tolerance for imputation. For extremely small dataset sizes (e.g., a 6x3 matrix), only a 10% threshold for full optimization will enable simulation as all other thresholds will result in an insufficient sample size for imputation method testing and error calculation. The known values removed for simulation are kept as the hold-out set and are used to evaluate error of imputation as follows:

Given dataset: $X = \{x_{ij}\}, i = 1 \dots, N_s; j = 1 \dots, N_f$ where N_s is the number of samples and N_f is the number of features; with missing elements $x_{km}, (k, m) \in M$ the goal of imputation is to determine values for the missing elements that resemble the complete data. As the first step in optimization, any row or column with missing elements are removed leading to the subset $X' = \{x_{ij}'\}, i = 1 \dots, N_s', \backslash x_{km}, (k, m); j = 1 \dots, N_f$. From this subset data, removal is performed separately to simulate MAR, MCAR, and MNAR mechanisms. Pseudocode for each missingness mechanism is provided in Figure 1B.

For **MCAR**, a filtering matrix of dimension $N'_s x N'_f$ is created by

random sampling from a uniform distribution (minimum = 0 and maximum = 1) generated from the function *runif* in R. Random values in the matrix are ranked and values below the imputation threshold are set to NA for missing and above are set to one for remaining. The element-wise product between this filtering matrix and full data matrix provides the MCAR example set for further testing.

For **MNAR**, the missing value assignment is performed individually for each feature as follows: 1) A list of values is generated by sampling from a logistic distribution (location = 0, scale = 1), denoted $L_1 = \{l_i^{(1)}\}, i = 1 \dots, N'_s$. 2) A second list is generated by sampling from the uniform distribution (minimum = 0 and maximum = 1), denoted $L_2 = \{l_i^{(2)}\}, i = 1 \dots, N'_s$. 3) A third list is generated from the product of $L_3 = L_1 \cdot L_2$, $L_3 = \{l_i^{(3)} = l_i^{(1)} l_i^{(2)}\}$. 4) The ranks for the values in L_1, L_3, as well as the feature measurements, are computed. 5) The highest and lowest ranks from L_3, with the number of missing values dependent on the assigned threshold, are determined and the corresponding (feature-wise) ranks in L_1 are assigned. Equivalent ranks in the data set are removed as missing.

in L_3, are retrieved, and the corresponding order in the filtering matrix column for the co-dependent feature are assigned as NA.

After generating the three types of missing datasets, each dataset is imputed using each of the eight available methods. For multivariate methods, users are prompted to select a simple or full version of parameter optimization. Simple parameter optimization uses the following default parameters: K-value=10, Tree Value=500, and Mice Iteration=2. If full parameter search is selected, the accuracy of the imputed values is tested over a range of hyperparameters for kNN, MICE and RF. Specifically, for kNN, the K-values tested range from 10 to 100 incremented by 20. For the optimal K-value in in this range, a refined search is conducted from k-4 to k+4 in single value increments to identify the K-value with the lowest error rate. For RF, the number of trees in the sequence of 5, 10, 20, 50, 100, 150, 200, 500 are examined to determine the optimal tree size. For MICE, 1-3 iterations are tested. Error rates are calculated by mean absolute error rate (MAPE) defined as:

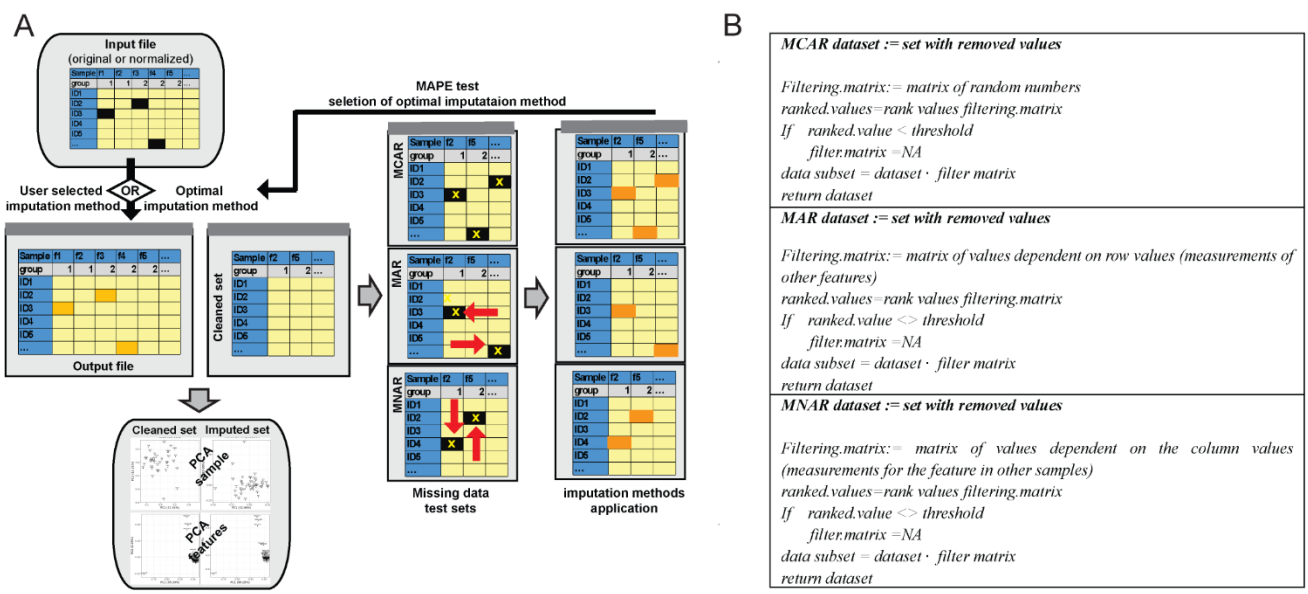

Figure 1. A. Schematic workflow of ImpLiMet. In the case of automated optimization, ImpLiMet first removes all columns in the dataset with missing values then simulates missing elements following three types of missingness: MCAR, MAR and MNAR. Missing values are imputed with all methods and the error of imputation is determined using MAPE. Imputation is then performed on the original dataset using the method with the lowest MAPE value. The dataset with imputed values is returned to the user and effect of imputation on the dataset is visualized by PCA. B. Schematic pseudo code of the process of data removal for the three different missingness types during optimization. Matrix multiplication indicates element-wise product. More detailed pseudo code is provided in the Supplementary Materials. A detailed flowchart is presented at https://complimet.ca/shiny/implimet.

For **MAR,** a co-dependance group is created by summating all feature values in a sample except the values in the current cell. If the input file contains information about the feature groups, based on biological or analytical characteristics, the summation calculation is performed within each feature group for each sample for the co-dependence matrix. The MAR process follows MNAR steps 1 through 3. In step 4, the ranks for the values in L_1, L_3, and the sample values in the filtering matrix are computed. Missing indices are assigned to the highest and lowest ranks from L_3, with the number of missing values dependent on the sample threshold. The order of the values in L_1, that produces the missing indices

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \frac{|x_i - y_i|}{x_i}$$

where N is the number of missing values, x_i is the actual value and y_i is the prediction. The MAPE results for each of the eight imputation methods assessed for each missingness mechanism are displayed and the method with the lowest MAPE value across the missingness mechanisms is highlighted and used for imputation.

The effect of imputation on the dataset is visualized by PCA comparing the original dataset following removal of all samples and features with missing data to that of the imputed dataset. PCA for both

samples, calculating PCs using features as variables, and features, using their values across samples as variables are shown. The user provided sample and feature names are shown in the plots for reference. An example of the optimization utilization as well as comparison of errors in imputation using recommended and other imputation methods is presented in the Supplementary Materials. Briefly, from the subset of metabolomics data published by Li et al. (Li et al. 2019) with complete data for 50 samples and 50 features we have removed values from 120 cells and tested the error rate for the imputed values using different methods. Result shows that the recommended method, in this case RF, provides imputation with the lowest error when comparing the imputed and the original data and the best agreement in PCA with the original dataset (full information is provided in Supplementary Materials).

3 Conclusion

ImpLiMet is a versatile and open-access Web-based application designed to assist users in identifying the optimal imputation solution for their given dataset. Presented imputation recommendation shows the best method based on the lowest error rate overall, while at the same time presenting error rates of imputation for different types of missingness for all methods. ImpLiMet currently includes eight previously presented imputation methods. Future work will include the addition of other published and newly developed imputation methods as well analysis of the type of missingness that is present in the dataset.

Funding

This work was supported in part by RGPIN-2019-06796 to SALB from the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as operating grant AI-4D-102-3 to SALB and MCC from the NRC AI4D Program. HO received an NSERC CREATE Matrix Metabolomics Scholarship. *Conflict of Interest:* none declared.

References

Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. (2018), 'From Predictive Methods to Missing Data Imputation: An Optimization Approach', *Journal of Machine Learning Research*, 18.

Chilimoniuk, J., et al. (2024), 'imputomics: web server and R package for missing values imputation in metabolomics data', *Bioinformatics*, 40 (3).

Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol.1(2): research0003.

Jäger, S., Allhorn, A., and Biessmann, F. (2021), 'A Benchmark for Data Imputation Methods', *Frontiers in Big Data*, 4.

Mack, C., Su, Z., and Westreich, D. (2018), 'AHRQ Methods for Effective Health Care', *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition* (Rockville (MD): Agency for Healthcare Research and Quality (US)).

Pantanowitz, A., Marwala, T. (2009). Missing Data Imputation Through the Use of the Random Forest Algorithm. In: Yu, W., Sanchez, E.N. (eds) Advances in Computational Intelligence. Advances in Intelligent and Soft Computing, vol 116. Springer, Berlin, Heidelberg.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika*, 63 (3), 581-92.

Schafer, J.L (1997), 'Analysis of Incomplete Multivariate Data (1st ed.)', *Chapman and Hall/CRC*.

Scheffer, Judi (2002), 'Dealing with missing data', *Research Letters in the Information and Mathematical Sciences*, 3, 7.

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

Stekhoven, Daniel J., and Peter Buehlmann. (2011) MissForest-non-parametric missing value imputation for mixed-type data." Bioinformatics 28 (1): 112–18.

Stoyanovich, J., Howe, B., and Jagadish, H. V. (2020), 'Responsible Data Management', *Proceedings of the Vldb Endowment*, 13 (12), 3474-88.

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

van Buuren, S., Boshuizen, H.C., Knook, D.L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18, 681–694.

van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn C.G.M., Rubin, D.B. (2006) Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 76, 12, 1049–1064.

van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1–67.

Wright, M., Ziegler. A. (2017) "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in c++ and r." Journal of Statistical Software, 77 (1): 1–17.