Comparative Assessment and Outlook on Methods for Imputing Proteomics Data

Minjie Shen
Virginia Polytechnic Institute and State University
Yi-Tan Chang
Virginia Polytechnic Institute and State University
Chiung-Ting Wu
Virginia Polytechnic Institute and State University
Sarah J Parker
Cedars Sinai Medical Center
Georgia Saylor
Wake Forest University
Yizhi Wang
Virginia Polytechnic Institute and State University
Guoqiang Yu
Virginia Polytechnic Institute and State University
Jennifer E. Van Eyk
Cedars Sinai Medical Center
Robert Clarke
University of Minnesota
David M. Herrington
Wake Forest University
Yue Wang (yuewang@vt.edu)
Virginia Polytechnic Institute and State University

Research Article

Keywords: quantitative proteomics data analysis, evaluation methodologies, low-rank matrix factorization framework

Posted Date: March 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-298864/v1

Comparative assessment and outlook on methods for imputing proteomics data

Minjie Shen[1,*], Yi-Tan Chang[1,*], Chiung-Ting Wu[1,*], Sarah J Parker[2], Georgia Saylor[3], Yizhi Wang[1], Guoqiang Yu[1], Jennifer E. Van Eyk[2], Robert Clarke[4], David M. Herrington[3], and Yue Wang[1,†]

[*]Equal contribution; [†]Author for correspondence

[1]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA; [2]Advanced Clinical Biosystems Research Institute, Cedars Sinai Medical Center, Los Angeles, CA 90048, USA; [3]Department of Internal Medicine, Wake Forest University, Winston-Salem, NC 27157, USA; [4]The Hormel Institute, University of Minnesota, Austin, MN 55912, USA

Abstract

Background: Missing values are a major issue in quantitative proteomics data analysis. While many methods have been developed for imputing missing values in high-throughput proteomics data, comparative assessment on the accuracy of existing methods remains inconclusive, mainly because the true missing mechanisms are complex and the existing evaluation methodologies are imperfect. Moreover, few studies have provided an outlook of current and future development.

Results: We first report an assessment of eight representative methods collectively targeting three typical missing mechanisms. The selected methods are compared on both realistic simulation and real proteomics datasets, and the performance is evaluated using three quantitative measures. We then discuss fused regularization matrix factorization, a popular low-rank matrix factorization framework with similarity

and/or biological regularization, which is extendable to integrating multi-omics data such as gene expressions or clinical variables. We further explore the potential application of convex analysis of mixtures, a biologically-inspired latent variable modeling strategy, to missing value imputation. The preliminary results on proteomics data are provided together with an outlook into future development directions.

Conclusion: While a few winners emerged from our comparative assessment, data-driven evaluation of imputation methods is imperfect because performance is evaluated indirectly on artificial missing or masked values not authentic missing values. Imputation accuracy may vary with signal intensity. Fused regularization matrix factorization provides a possibility of incorporating external information. Convex analysis of mixtures presents a biologically plausible new approach.

Background

Liquid chromatography coupled to mass spectrometry (LC-MS) is a popular method for high-throughput identification and quantification of thousands of proteins in a single analysis [1, 2]. The LC-MS signals can be displayed in a three-dimensional space with the mass-to-charge ratios, retention times and intensities for the observed peptides. However, this approach suffers from many missing values at the peptide or protein level, which significantly reduces the amount of quantifiable proteins with an average of 44% missing values [3-5].

While there are multiple causes for this missingness, three typical missing mechanisms are widely acknowledged. Low abundant proteins may be missing because their concentration is below the lower limit of detection (LLD); while poorly ionizing peptides may cause proteins to be Missing Not at Random (MNAR) [6]. However, missingness may also extend to mid- and even high-range intensities, statistically

categorized into Missing at Random (MAR) and Missing Completely at Random (MCAR) [7]. MAR is actually missing conditionally at random given the observed, known covariates, or even unknown covariates. MCAR depends neither on observed nor on the missing data, thus the incomplete data are representative for the entire data. Practically, MAR and MNAR cannot be distinguished because by definition missing values are unknown [8]. More importantly, missing values in reality can originate from a mix of both known and unknown missing mechanisms [7, 9].

A common solution for missingness is to impute the missing values based on assumed missing mechanisms. But, this comes at the expense of potentially introducing profound change in the distribution of protein-level intensities, because most of existing methods are designed specifically for a single missing mechanism. This can have unpredictable effects on downstream differential analyses. Moreover, while many imputation methods have been adopted for imputing missing values in proteomics data, comparative evaluation on their relative performance remains largely inconclusive, and few studies provide an outlook addressing unresolved problems or future development directions [4, 9, 10].

To gain first-hand insight into the strengths and limitations of both imputation methods and assessment designs, we conduct a collective assessment of eight representative methods involving three typical missing mechanisms in conjunction with authentic missing values. Compared on a set of realistic and preserving simulations derived from real proteomics data sets, the performance of the selected methods is measured by three criteria, root-mean-square error (RMSE), normalized root-mean-square error (NRMSE), and Sum of Ranks (SOR). There are several important observations from this comparison study. First, while imputation methods perform differentially under various missing mechanisms, algorithmic parameter settings, and

preprocessing procedures, there are a few methods that consistently outperformed peer methods across a range of realistic simulation studies. Second, the quality of performance assessment depends on the efficacy of simulation designs and a more realistic simulation design should include authentic missing values and preserve original overall data distribution. Third, existing assessment methodology is imperfect in that performance is indirectly assessed on imputing either artificial or masked, but not authentic missing values (see Discussion section).

To explore a more integrative strategy for improving imputation performance, we discuss a low-rank matrix factorization framework with fused regularization on sparsity and similarity – Fused Regularization Matrix factorization (FRMF) [11-13], which can naturally integrate other-omics data such as gene expression or clinical variables. We also introduce a biologically-inspired latent variable modeling strategy - Convex analysis of Mixtures (CAM) [13, 14], which performs data imputation using the original intensity data (before log-transformation). The preliminary results on real proteomics data are provided together with an outlook into future development directions.

Results

Experimental design and protocol

We selected eight representative methods for comparative assessment, based on their intended missing mechanism(s) and imputation principles, summarized in **Figure 1**. One method (Min/2) is devoted to MNAR (LLD) [7], two methods (swKNN and pwKNN) are tailored to MAR (local-similarity) [15], and five methods (Mean, PPCA, NIPALS, SVD, and SVT) are designed for MCAR/MAR (global-structure or low-rank matrix factorization) [7, 9, 16-18]. We then explored and tested several variants of

100 FRMF and CAM, where local similarity information is obtained from baseline or other
101 data acquired from the same samples.

102 We conducted the comparative assessments in two complementary simulation
103 settings. First, the realistic simulation data were generated from the observed data
104 portion (no authentic missing value) of a real proteomics dataset, where artificial
105 missing values were introduced involving two typical missing mechanisms and used
106 for performance assessment. Second, the realistic simulation data were generated from
107 the complete data matrix (including authentic missing values) of a real proteomics
108 dataset, where a small percentage of data points were randomly set-aside (masked
109 values) and used solely for performance assessment. The preprocessing eliminates
110 those proteins whose missing rates are higher than 80% and then performs log2
111 transformation [19]. The parameters were optimized for each imputation method by
112 parameter sweeping over a wide range of settings at each missing rate. The overall
113 experimental workflow is given in **Figure 2**.

114 **Real proteomics data**

115 The real LC-MS proteomics data form the base from which the simulation data sets
116 were produced [6]. The data were acquired using data-independent acquisition (DIA)
117 protocol, and protein level output was generated by mapDIA [20]. The dataset contains
118 200 samples associated with 2,682 proteins measured in human left anterior descending
119 (LAD) coronary arteries collected as part of a study of coronary and aortic
120 atherosclerosis [21]. The data were produced in three separate batches, indexed by A,
121 B, and C, and all have passed quality control and preprocessing procedures,
122 summarized in **Table 1** (Supplementary Information).

123 **Table 1.** Summary of real proteomics datasets used in this work.

	Sample size	Protein size	Total Missing Rate #MV/(#Sample*#Protein)	Setting #1 protein size (non-missing proteins)	Setting #2 protein size (proteins with <= 80% missing rate)
Batch A	98	2107	24.67%	751 (35.64%)	1935 (91.84%)
Batch B	55	2604	29.63%	819 (31.45%)	2324 (89.25%)
Batch C	47	2590	25.52%	976 (37.68%)	2325 (89.77%)

Simulation data generated from the observed portion of data matrix

Based on the observed data portion (no authentic missing values), we adopted a hybrid missing data model and used the R package imputeLCMD to introduce artificial missing values while preserving the original observed data patterns [22]. Specifically, MCAR missing values were introduced by randomly replacing some data points with 'NA' (not available) according to the designed missing rates (approximately from 1% to 50%); MNAR missing values were introduced by quantile cut-off for the full data set [7, 9, 19]; and mixed MCAR and MNAR missing values were introduced by assigning $(1-\beta)$ portion of MCAR and β portion of MNAR; corresponding to missing rate α and $\beta = 0, 0.1, 1$ (Supplementary Information).

Simulation data with set-aside masked values from the full data matrix

In this simulation setting, we used the full data matrix (including both observed and authentic missing values) from the human coronary proteomics data set. To preserve the original patterns of both observed and authentic missing data, for each protein, a small percentage of data points in the complete data matrix were randomly set-aside as 'NA' (masked values) with the masking rate(s) proportional to the authentic missing rate(s). This procedure was repeated for all proteins and the masked values were

142 considered as a mix of MNAR and MAR conditioned on the observed missing rates
143 and data patterns (**Figure 3**, Supplementary Information).

144 **Performance assessment focused on MNAR**

145 As shown in **Figure 4** (see additional results in Supplementary Information), under
146 MNAR missing mechanism assumption, SVT and Min/2 yielded the best performance
147 in both simulation settings, the relative performance of SVT and Min/2 depends on the
148 missing rates and criterion used for evaluation. The MNAR-devoted method, Min/2,
149 performs much better than the others as expected while the baseline method, Mean,
150 performance is the worst among all methods (see additional results in Supplementary
151 Information). Note that SOR increases expectedly when the missing rate increases
152 because SOR is positively associated with the number of missing proteins, therefore the
153 value of SOR may not imply the absolute performance of a method.

154 **Performance assessment focused on MCAR**

155 The imputation performance of the eight methods on MCAR mechanism is shown in
156 **Figure 5** (see additional results in Supplementary Information). The experimental
157 results show that NIPALS outperforms all other methods in both simulation settings
158 and for almost all three evaluation criteria; while SVT is the best when RMSE is used;
159 Min/2 performs the worst among all other methods in all cases that may be expected
160 due to its design for MNAR mechanism; and Mean performs flatly over different
161 missing rates with an expected baseline performance except for Min/2. In addition,
162 while all methods perform worse when total missing rate increases, the ranking of their
163 relative performances remains unchanged (see additional results in Supplementary
164 Information).

Performance assessment focused on authentic missing values

As shown in **Figure 6** (see additional results in Supplementary Information), NIPALS, SVT, and protein-wise or sample-wise KNN achieve the best performance, where authentic missing values are dominant and imputation accuracy is evaluated on the masked values. MNAR-devoted method Min/2 performs the worst as expected. Similar to the case of MCAR, low-rank methods and local-similarity methods perform worse when total missing rate increases, and among these methods, SVD and PPCA perform even worse than the baseline method Mean when total missing rate is large. Note that because low abundant proteins often have higher authentic missing rates and accordingly higher masking rates, more low abundant proteins (possibly the minimum values) are masked over highly expressed proteins, and the counterintuitive decrease in NRMSE by Min/2 is expected when the authentic missing rate increases.

Evaluation of the FRMF method focused on authentic missing values

We evaluated three variants of the FRMF method. RMF serves as a baseline regularized matrix factorization algorithm; FRMF_self introduces a fused-regularization utilizing the similarity among samples embedded within data matrix; and FRMF_cross_patho exploits external pathological scores again via fused-regularization strategy where the pathological scores are the qualitative percentages of the intimal surface involvement of various atherosclerotic changes graded by pathologists [6].

The experimental results are shown in **Figure 7**. While RMF performs comparably and expectedly to SVD, both FRMF_self and FRMF_cross_patho significantly outperform RMF. This preliminary result indicates a potential benefit of combining global low-rank and local-similarly regularizations, as well as leveraging external information via fused regularization.

Evaluation of the CAM method focused on authentic missing values

Based on biologically-inspired latent variable modeling of complex tissues - CAM [13, 14], we proposed and evaluated three variants of the CAM based imputation method. CAM_complete performs CAM based imputation using the non-missing portion of full data matrix; CAM_SVT and CAM_NIPALS perform CAM based imputation using full data matrix while initialized by SVT and NIPALS, respectively.

The experimental results are shown in **Figure 8**. Expectedly, CAM_complete performs much better than the baseline method Mean. More importantly, both CAM_NIPALS and CAM_SVT consistently outperform NIPALS and SVT - the two top performers indicated in our earlier comparative assessment. This preliminary result suggests that biologically-plausible latent variable modeling may potentially improve imputation accuracy within the framework of low-rank optimization.

Method

Brief introduction to the eight existing methods

- **Min/2 (half minimum):** Assuming the MNAR missing mechanism, for each protein, replacing missing values with half the minimum value of observed intensities in that protein across samples [6, 10].
- **Mean:** Assuming the MAR/MCAR missing mechanism, for each protein, replacing missing values with mean value of observed intensities in that protein across samples [6, 10].
- **swKNN (sample-wise k-nearest neighbors):** Assuming the MAR missing mechanism and leveraging local similarity among samples, for each protein, replacing missing values with weighted average of observed intensities in that protein proportional to the proximities of k-nearest neighboring samples [10].

- **pwKNN (protein-wise k-nearest neighbors):** Assuming the MAR missing mechanism and leveraging local similarity among proteins, for each sample, replacing the missing values with weighted average of observed intensities in that sample proportional to the proximities of k-nearest neighboring proteins (with protein-wise normalization) [10].
- **PPCA (probabilistic PCA):** Assuming the MCAR/MAR missing mechanism, a low-rank probabilistic PCA matrix factorization is estimated by the expectation maximization (EM) algorithm and subsequently used to impute missing values [23].
- **NIPALS (non-linear estimation by iterative partial least squares):** Assuming the MCAR/MAR missing mechanism, a low-rank missing-data-tolerant PCA matrix factorization is estimated by iterative regression and subsequently used to impute missing values [24, 25].
- **SVD (SVDImpute):** Assuming the MCAR/MAR missing mechanism, a low-rank SVD matrix factorization is estimated by the EM algorithm and subsequently used to impute missing values [24, 26].
- **SVT (singular value thresholding):** Assuming MCAR/MAR missing mechanism, a low-rank SVT matrix factorization is estimated by iteratively solving a nuclear norm minimization problem and subsequently used to impute missing values [18].

Performance measures

Three quantitative measures are used to evaluate imputation accuracy, namely Root Mean Square Error (RMSE), Normalized Root Mean Square Error (NRMSE), and Sum of Ranks (SOR). Specifically, RMSE and NRMSE are given by [27, 28]

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\Omega}(\hat{X}_{\Omega} - X_{\Omega})^2}{|\Omega|}}, \quad \mathrm{NRMSE} = \sqrt{\frac{\sum_{\Omega}(\hat{X}_{\Omega} - X_{\Omega})^2}{|\Omega|\sigma^2_{X_{\Omega}}}},$$

respectively, where Ω is the index set of missing values in complete data matrix X, $|\Omega|$ is the total number of missing values, $\hat{X}$ is the imputed complete data matrix, and $\sigma^2_{X_{\Omega}}$ is the variance of missing values. To address the bias of NRMSE under MNAR missing mechanism, SOR has been proposed as [19]

$$\mathrm{SOR} = \sum_{i=1}^{P} rank(\mathrm{NRMSE}_i),$$

where P is the number of proteins containing at least one missing value, i is the protein index in this protein subset, and $rank(\mathrm{NRMSE}_i)$ is the ranks of protein-wise NRMSE across different imputation methods.

Introduction to FRMF method

As aforementioned, low-rank matrix factorization has been a popular and effective approach for missing data imputation [12]. For imputing proteomics data, the assumption is that there is only a small number of biological processes determining the expression profiles. Consider an $m \times n$ complete data matrix X describing m samples and n proteins. A low-rank matrix factorization approach seeks to approximate X containing missing values by a linear latent variable model,

$$X_{m\times n} = A_{m\times l} \times S_{l\times n}, \quad (1)$$

where $A_{m\times l}$ and $S_{l\times n}$ are the low-rank factor matrices, and $l \ll \min(m, n)$. In order to prevent overfitting, the solution is often formulated as a regularized sparse SVD minimization problem on the observed values

$$\min \sum_{i=1}^{m}\sum_{j=1}^{n} I\left(X_{ij} \neq \mathrm{NA}\right)\left(X_{ij} - A_i S_j\right)^2 + \lambda_{\mathrm{A}} \|A\|_F^2 + \lambda_{\mathrm{S}} \|S\|_F^2,$$

where $\|.\|_F^2$ denotes the Frobenius norm, $I(.)$ is the indicator function, and $\lambda_A, \lambda_S > 0$ are the regularization parameters. When local similarity information is available, FRMF can be formulated by adding a fused regularization term

$$\min \sum_{i=1}^{m}\sum_{j=1}^{n} I(X_{ij} \neq \mathrm{NA})(X_{ij} - A_i S_j)^2 + \lambda_A \|A\|_F^2 + \lambda_S \|S\|_F^2$$

$$+ \alpha \sum_{i=1}^{m} \sum_{k \in \mathcal{F}(i)} \|A_i - A_k\|_F^2,$$

where α is the fused regularization parameter, and $\mathcal{F}(i)$ denotes the neighborhood sample subset of sample i and can be determined using baseline data or other relevant measurements e.g. gene expression or pathological score. In our study, $\mathcal{F}(i)$ is determined by the between-sample cosine similarity $\cos(X_i, X_k)$ based on data matrix in FRMF_self, or $\cos(P_i, P_k)$ based on pathological scores in FRMF_cross_patho.

Introduction to CAM method

CAM is a latent variable modeling and deconvolution technique previously used for identifying biologically-interpretable cell subtypes $S_{l\times n}$ and their composition $A_{m\times l}$ in complex tissues [6, 13, 14, 21]. We adopt the CAM framework into (1) and demonstrate that hybrid CAM_SVT and CAM_NIPALS can handle missing values naturally and this combination leads to a novel and biologically-plausible imputation strategy. The workflow of CAM based method with three variants is given in **Figure 9**.

Discussion

The quality of simulating assumed missing mechanisms (MNAR and MCAR) depends on the efficacy of simulation tools. However, because the simulation uses only the observed portion of data matrix, the introduced artificial missing values cannot fully resemble authentic missing mechanisms and/or patterns in relation to original overall

data distribution. More critically, performance is actually assessed on imputing artificial not authentic missing values, where the overall data distribution may be distorted.

To address the aforementioned issues in the presence of authentic missing values, a small percentage of set-aside values are introduced into complete data matrix and used solely for assessment purpose. Because masked values are randomly assigned onto both observed and authentic missing values, the simulation maximally preserves original overall data distribution. Note that masked values may represent a mix of MNAR (high missing rate associated with low protein abundance) and MAR (joint distribution of both observed and authentic missing values). However, performance is assessed indirectly on imputing masked not authentic missing values.

Imputation accuracy would be arguably affected by data preprocessing and algorithmic parameter setting. In this study, sample-wise normalization and protein-wise standardization are performed based on the requirements of each method. While these preprocessing notably affects the scale of NRMSE, relative performances across methods remain consistent. The experimental results show that imputation performance varies with parameter setting, while there appears no theoretical guideline for optimizing parameter setting.

While FRMF is a promising and novel imputation approach, its effectiveness for improving classic low-rank methods would depend on diversity among samples, discriminatory power of similarity measure, and complementary nature of additional and relevant measurements. Newly proposed CAM method represents an interesting direction for further development. More importantly, CAM performs missing value imputation using original intensity rather than log-transformed data, and this is

mathematically more rigorous because log-transformation violates the linear nature of low-rank matrix factorization [29].

Declarations

- Ethics approval and consent to participate

  Not applicable

- Consent for publication

  Not applicable

- Availability of data and materials

  The scripts used in the paper is available in R script ProImput. Code for all experiments can be found in the vignette at https://github.com/MinjieSh/ProImput. The operation system can be any system supporting R language.

- Competing interests

  The authors declare that they have no competing interests.

- Funding

  This work has been supported by the National Institutes of Health under Grants HL111362-05A1, HL133932, NS115658-01, and the Department of Defense under Grant W81XWH-18-1-0723 (BC171885P1).

- Authors' Contributions

  MJS, YTC, CTW and YW initiated this research and developed the methods. MJS and YW drafted the manuscript. MJS and CTW developed the assessment procedure and software. SJP

and GS contributed with biomedical case study. YZW, GQY, JEVE and DMH contributed with biomedical significance and discussion of the work. All authors have read, commented on and accepted the final manuscript.

- Acknowledgements

Not applicable

References

1. Canterbury JD, Merrihew GE, MacCoss MJ, Goodlett DR, Shaffer SA: **Comparison of data acquisition strategies on quadrupole ion trap instrumentation for shotgun proteomics**. *Journal of The American Society for Mass Spectrometry* 2014, **25**(12):2048-2059.
2. Doerr A: **DIA mass spectrometry**. *Nature methods* 2014, **12**(1):35.
3. Goeminne LJE, Sticker A, Martens L, Gevaert K, Clement L: **MSqRob Takes the Missing Hurdle: Uniting Intensity- and Count-Based Proteomics**. *Anal Chem* 2020, **92**(9):6278-6287.
4. Ma W, al. e, Wang P: **DreamAI: algorithm for the imputation of proteomics data**. *bioRxiv* 2020.
5. Dabke K, Kreimer S, Jones MR, Parker SJ: **A Simple Optimization Workflow to Enable Precise and Accurate Imputation of Missing Values in Proteomic Datasets**. *bioRxiv* 2020.
6. Herrington DM, Mao C, Parker SJ, Fu Z, Yu G, Chen L, Venkatraman V, Fu Y, Wang Y, Howard TD *et al*: **Proteomic Architecture of Human Coronary and Aortic Atherosclerosis**. *Circulation* 2018, **137**(25):2741-2756.
7. Lazar C, Gatto L, Ferro M, Bruley C, Burger T: **Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies**. *Journal of proteome research* 2016, **15**(4):1116-1125.
8. Jakobsen JC, Gluud C, Wetterslev J, Winkel P: **When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts**. *BMC Med Res Methodol* 2017, **17**(1):162.
9. Webb-Robertson B-JM, Wiberg HK, Matzke MM, Brown JN, Wang J, McDermott JE, Smith RD, Rodland KD, Metz TO, Pounds JG: **Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics**. *Journal of proteome research* 2015, **14**(5):1993-2001.
10. Liu M, Dongre A: **Proper imputation of missing values in proteomics datasets for differential expression analysis**. *Brief Bioinform* 2020.
11. Lin X, Boutros PC: **Optimization and expansion of non-negative matrix factorization**. *BMC bioinformatics* 2020, **21**(1):7.

12. Ma H, Zhou D, Liu C, Lyu MR, King I: **Recommender systems with social regularization**. In: *The fourth ACM international conference on Web search and data mining: 2011; Hong Kong*. ACM Press: 287-296.
13. Wang N, Hoffman EP, Chen L, Chen L, Zhang Z, Liu C, Yu G, Herrington DM, Clarke R, Wang Y: **Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues**. *Scientific Reports* 2016, **6**:18909.
14. Chen L, Wu CT, Wang N, Herrington DM, Clarke R, Wang Y: **debCAM: a bioconductor R package for fully unsupervised deconvolution of complex tissues**. *Bioinformatics (Oxford, England)* 2020, **36**(12):3927-3929.
15. Rahman SA, Huang Y, Claassen J, Heintzman N, Kleinberg S: **Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data**. *Journal of biomedical informatics* 2015, **58**:198-207.
16. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I: **Missing data and multiple imputation in clinical epidemiological research**. *Clinical epidemiology* 2017, **9**:157.
17. John C, Ekpenyong EJ, Nworu CC: **Imputation of missing values in economic and financial time series data using five principal component analysis approaches**. *CBN Journal of Applied Statistics* 2019, **10**(1):51-73.
18. Cai J-F, Candès EJ, Shen Z: **A singular value thresholding algorithm for matrix completion**. *SIAM Journal on optimization* 2010, **20**(4):1956-1982.
19. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y: **Missing value imputation approach for mass spectrometry-based metabolomics data**. *Scientific reports* 2018, **8**(1):1-10.
20. Teo G, Kim S, Tsou C-C, Collins B, Gingras A-C, Nesvizhskii AI, Choi H: **mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry**. *Journal of proteomics* 2015, **129**:108-120.
21. Parker SJ, Chen L, Spivia W, Saylor G, Mao C, Venkatraman V, Holewinski RJ, Mastali M, Pandey R, Athas G *et al*: **Identification of Putative Early Atherosclerosis Biomarkers by Unsupervised Deconvolution of Heterogeneous Vascular Proteomes**. *J Proteome Res* 2020, **19**(7):2794-2806.
22. **imputeLCMD: A collection of methods for left-censored missing data imputation** [https://cran.r-project.org/package=imputeLCMD]
23. Tipping ME, Bishop CM: **Probabilistic principal component analysis**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1999, **61**(3):611-622.
24. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J: **pcaMethods—a bioconductor package providing PCA methods for incomplete data**. *Bioinformatics* 2007, **23**(9):1164-1167.
25. Ochoa-Muñoz AF, González-Rojas VM, Pardo CE: **Missing data in multiple correspondence analysis under the available data principle of the NIPALS algorithm**. *DYNA* 2019, **86**(211):249-257.
26. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays**. *Bioinformatics* 2001, **17**(6):520-525.
27. Stekhoven DJ, Bühlmann P: **MissForest—non-parametric missing value imputation for mixed-type data**. *Bioinformatics* 2012, **28**(1):112-118.

28. Oba S, Sato M, Takemasa I, Monden M, Matsubara K, Ishii S: **A Bayesian missing value estimation method for gene expression profile data**. *Bioinformatics* 2003, **19**(16):2088-2096.
29. Zhong Y, Liu Z: **Gene expression deconvolution in linear space**. *Nat Methods* 2011, **9**(1):8-9; author reply 9.

Figures

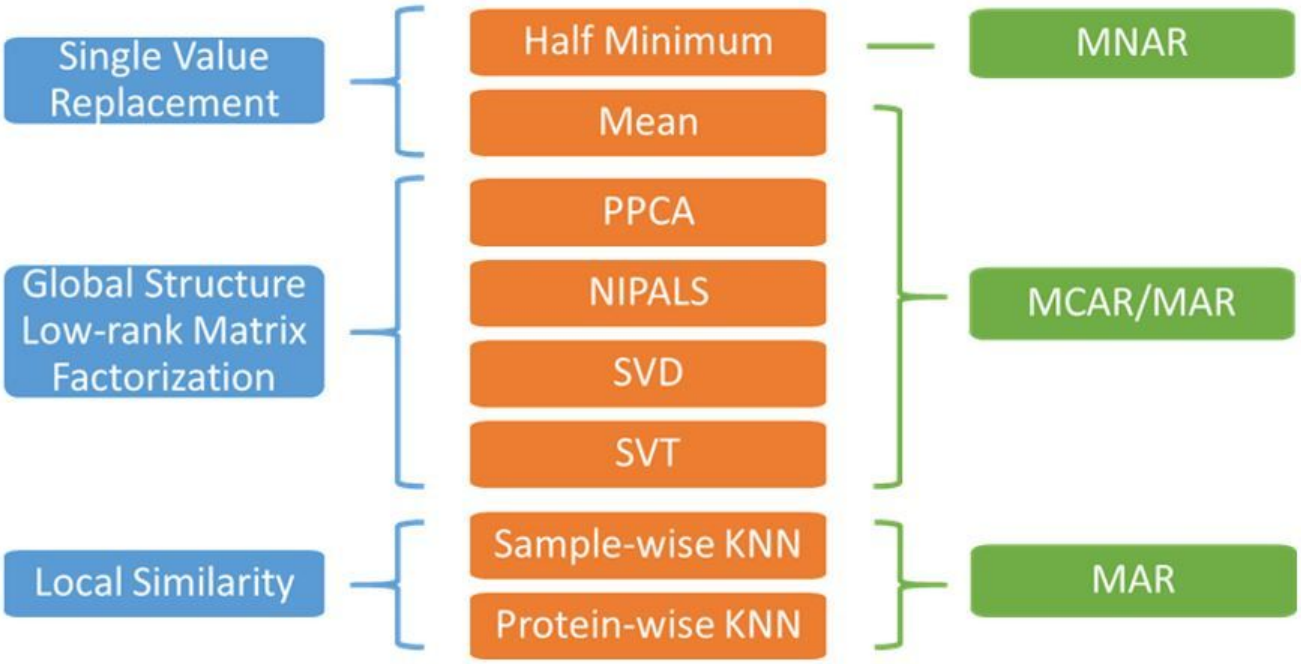

Figure 1

Comparative assessment of eight representative missing value imputation methods, divided into three categories.

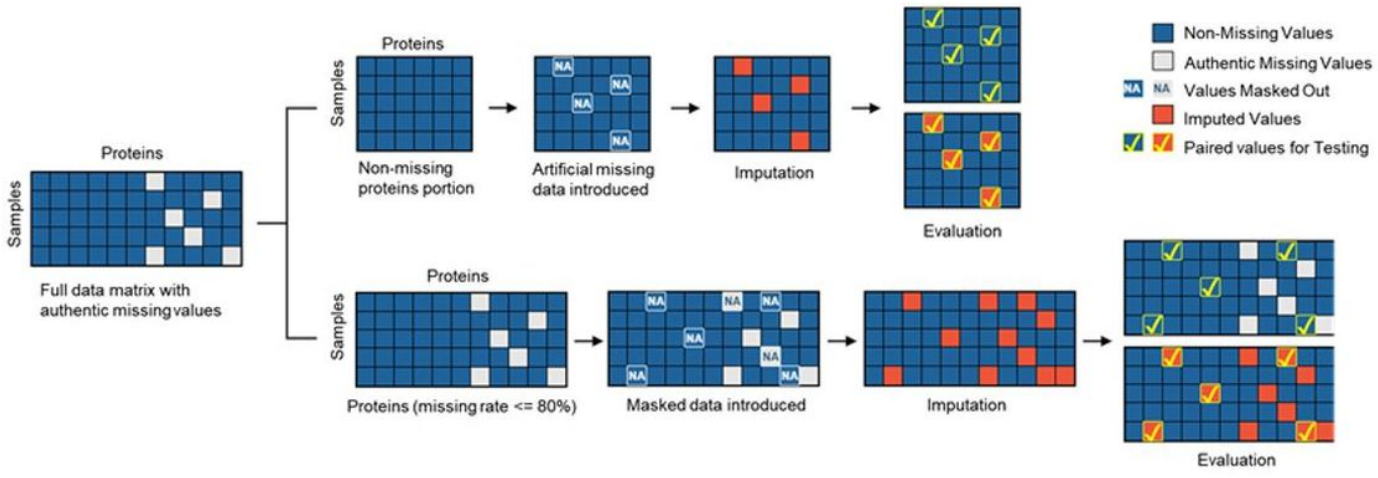

Figure 2

Two-phased workflow of realistic simulation-based assessment on missing value imputation methods.

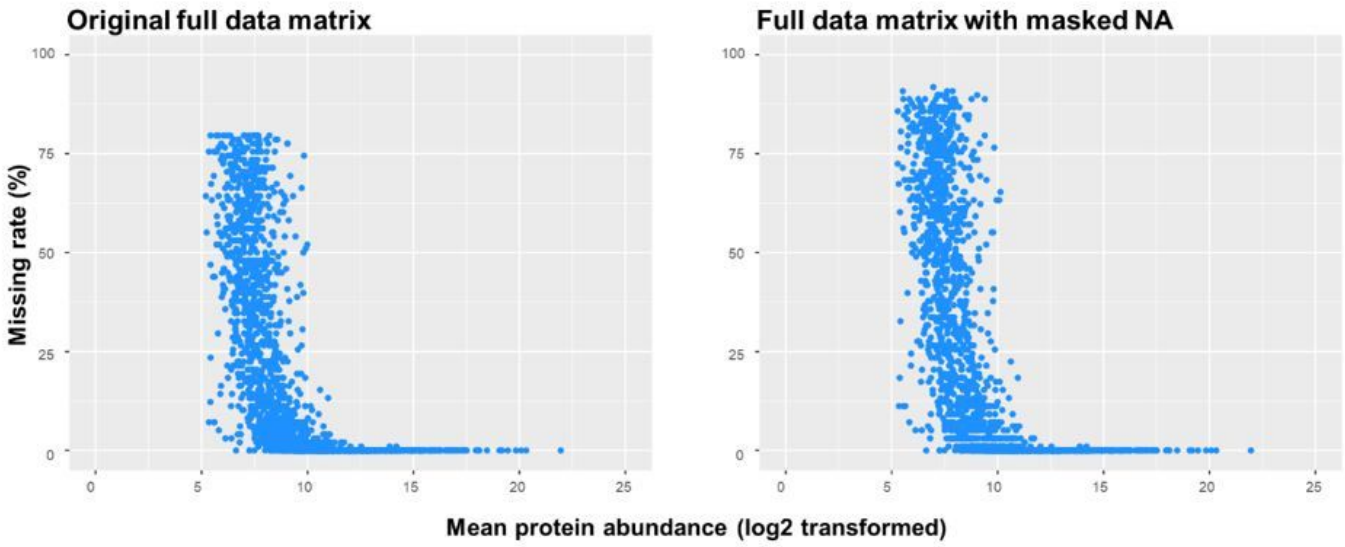

Figure 3

The overall pattern of missing values illustrated by the relationship between protein missing rate and protein mean intensity, before (left panel) and after (right panel) introducing masked NA.

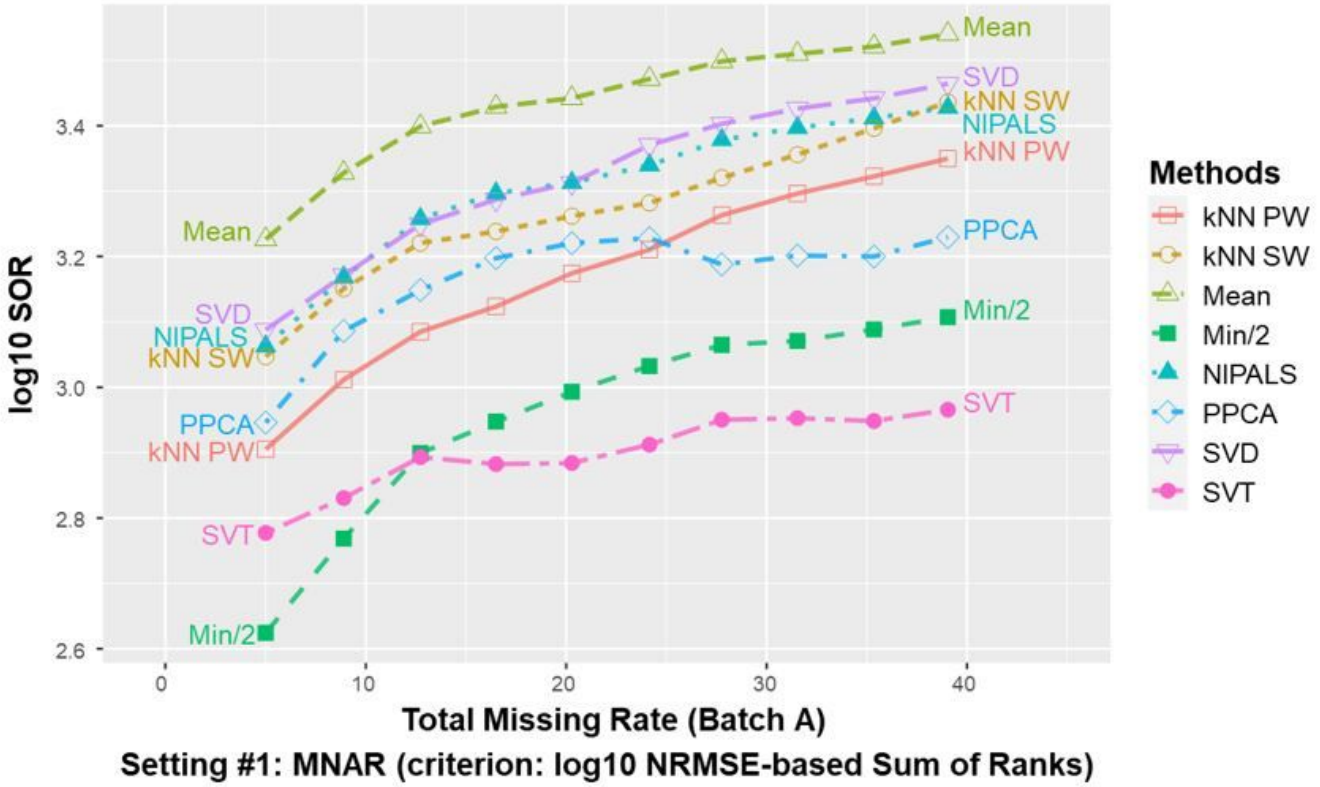

Figure 4

Imputation performance of the eight methods on the simulation data of setting #1, with assumed MNAR missing mechanism and varying total missing rates.

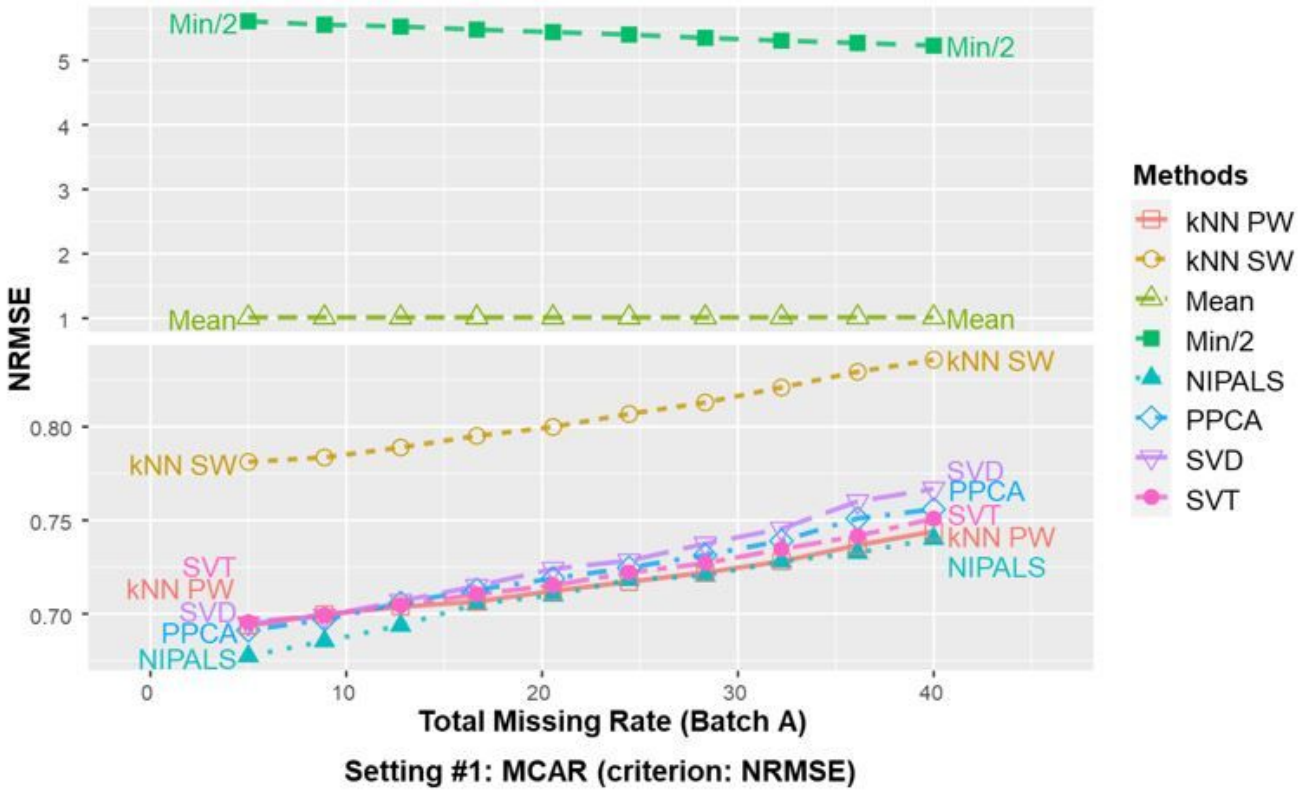

Figure 5

Imputation performance of the eight methods on the simulation data of setting #1, with assumed MCAR missing mechanism and varying total missing rates.

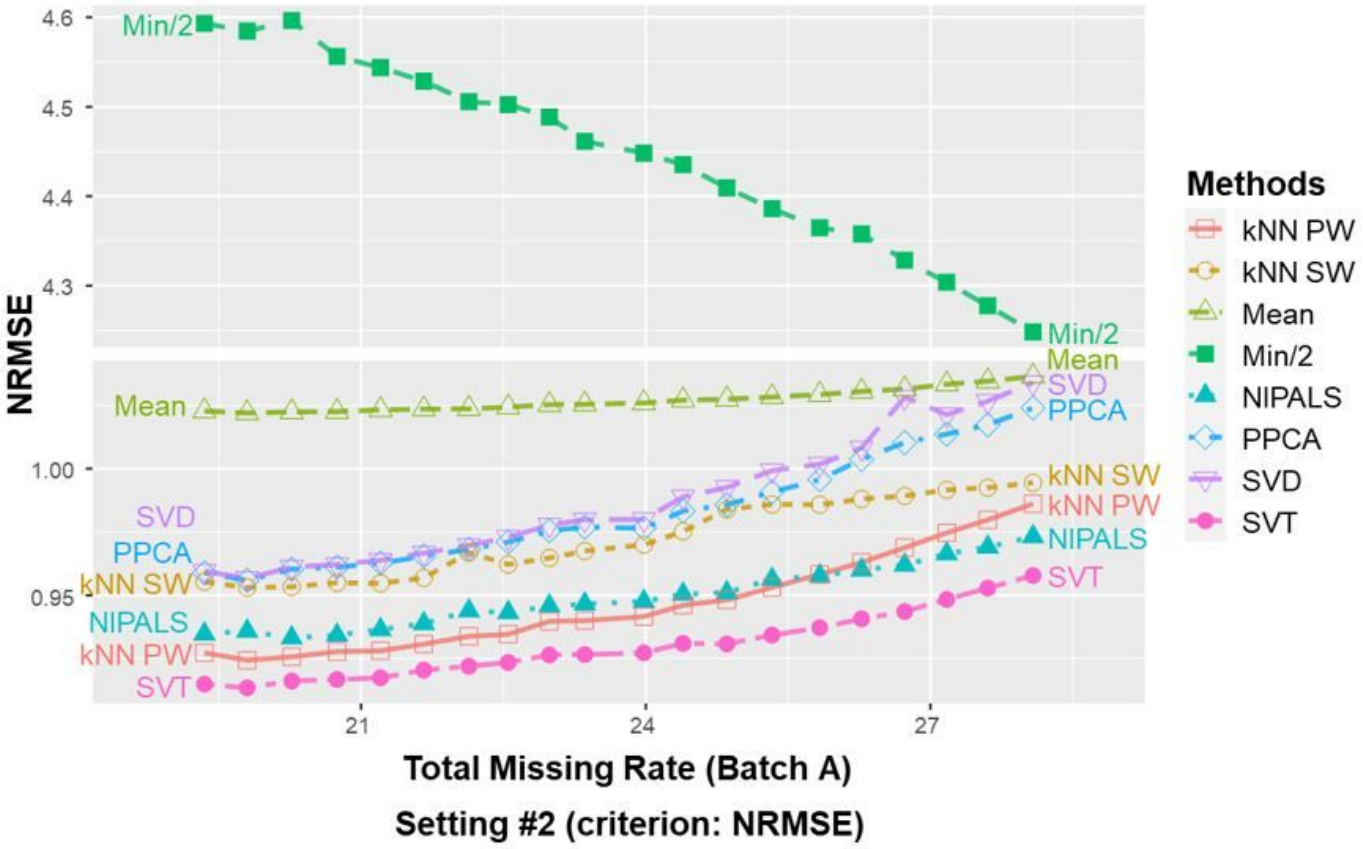

Figure 6

Imputation performance of the eight methods on the simulation data of setting #2, focusing on authentic missing mechanism and varying masked rates.

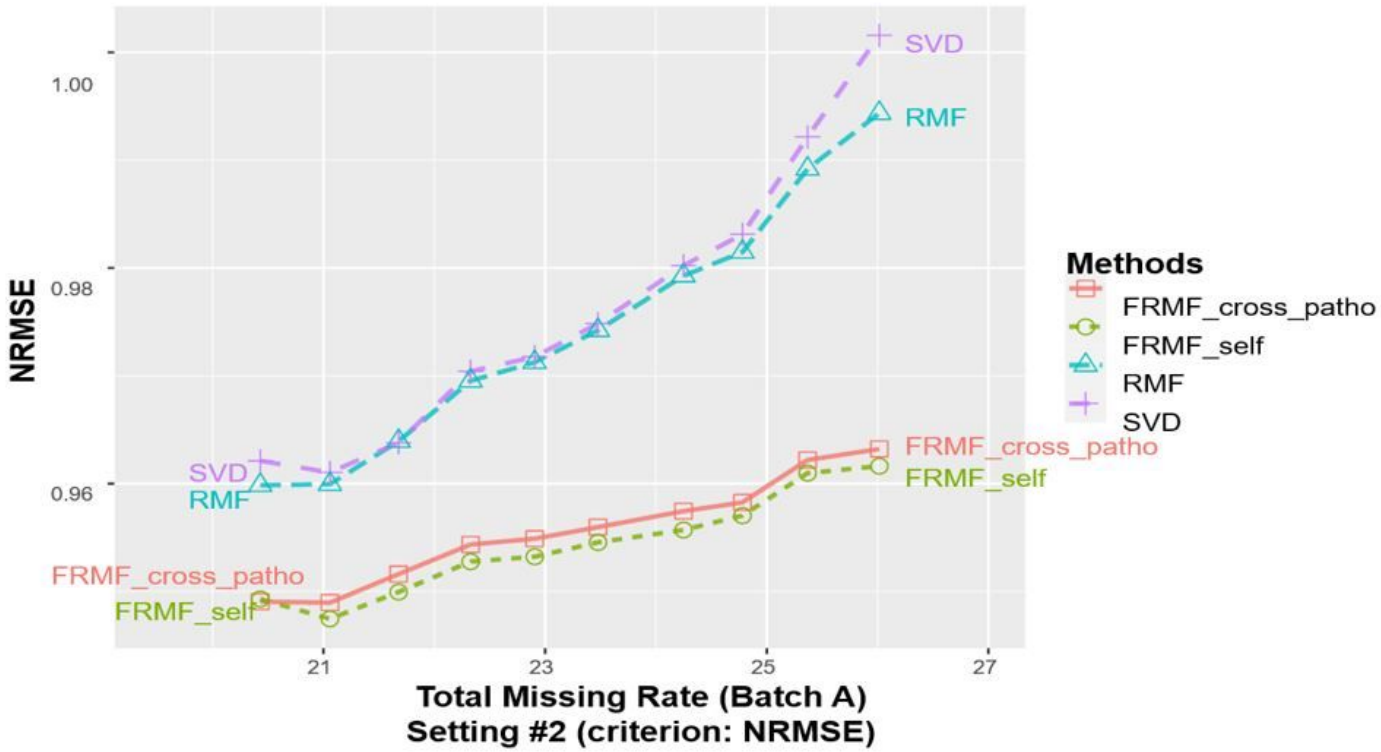

Figure 7

Imputation performance of the FRMF variants on the simulation data of setting #2, with varying masked rates.

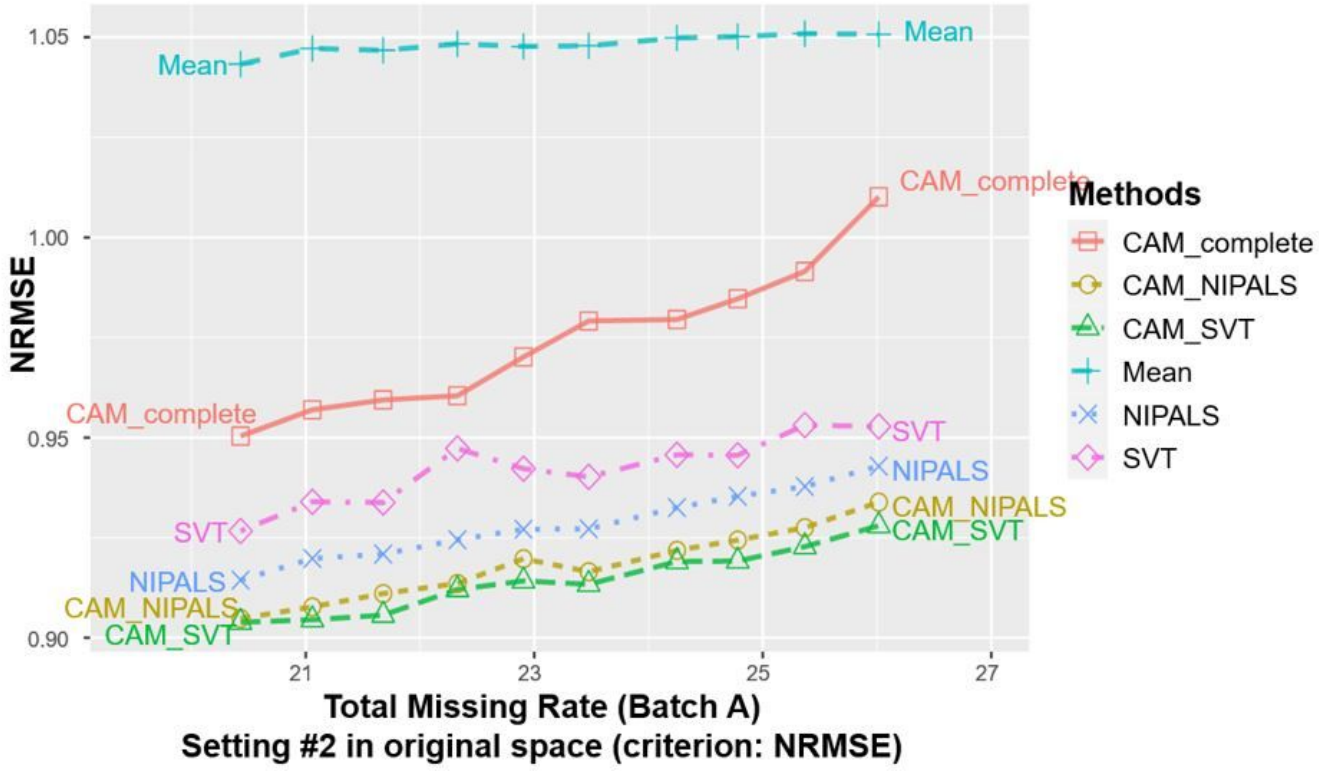

Figure 8

Imputation performance of CAM variants on the simulation data of setting #2, with varying masked rates, in comparison to that of Mean, SVT, and NIPALS. The imputation accuracy is evaluated in the original intensity space (before log-transformation).

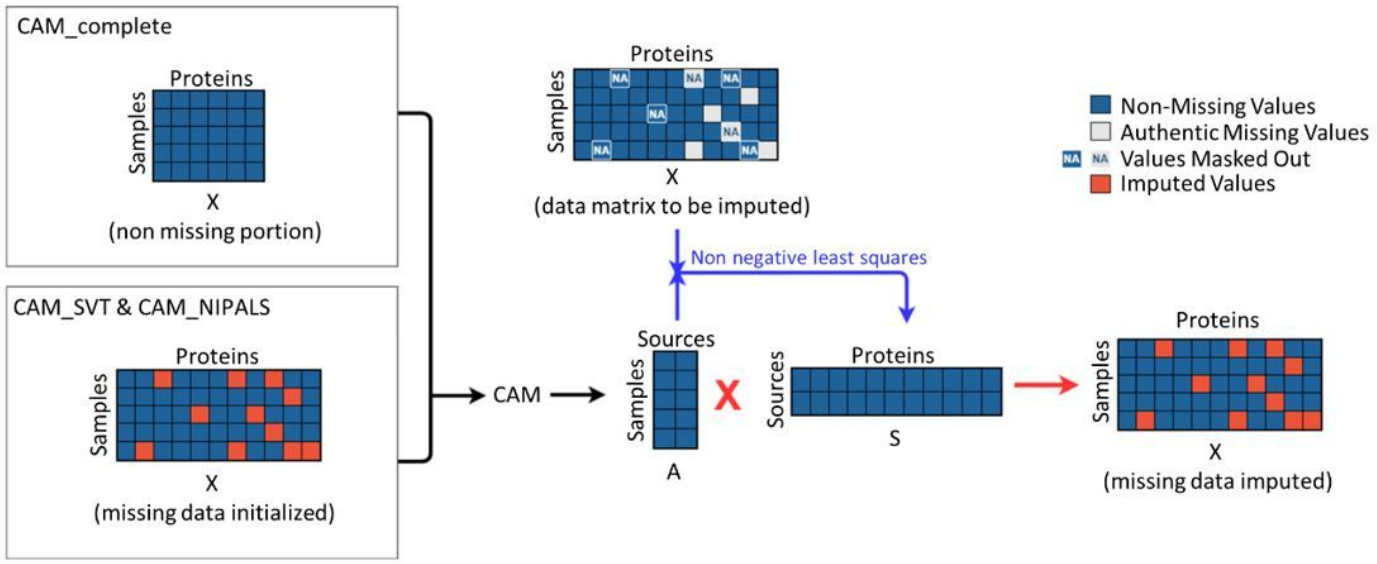

Figure 9

Workflow of the CAM based imputation method with two variant algorithms.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx