Bioinformatics, 34(9), 2018, 1555–1561
doi: 10.1093/bioinformatics/btx816
Advance Access Publication Date: 20 December 2017
Original Paper

OXFORD

Data and text mining

Missing value imputation for LC-MS metabolomics data by incorporating metabolic network and adduct ion relations

Zhuxuan Jin[1], Jian Kang[2,*] and Tianwei Yu[1,*]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA and [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 5, 2017; revised on September 29, 2017; editorial decision on December 17, 2017; accepted on December 19, 2017

Abstract

Motivation: Metabolomics data generated from liquid chromatography-mass spectrometry platforms often contain missing values. Existing imputation methods do not consider underlying feature relations and the metabolic network information. As a result, the imputation results may not be optimal.

Results: We proposed an imputation algorithm that incorporates the existing metabolic network, adduct ion relations even for unknown compounds, as well as linear and nonlinear associations between feature intensities to build a feature-level network. The algorithm uses support vector regression for missing value imputation based on features in the neighborhood on the network. We compared our proposed method with methods being widely used. As judged by the normalized root mean squared error in real data-based simulations, our proposed methods can achieve better accuracy.

Availability and implementation: The R package is available at http://web1.sph.emory.edu/users/tyu8/MINMA.

Contact: jiankang@umich.edu or tianwei.yu@emory.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metabolomics aims to comprehensively identify and quantify all metabolites in a system and to study their changes in relation to diet, environment, disease status, genetic effects, pharmaceutical interventions, etc. (Lindon *et al.*, 2007). By profiling and analyzing metabolite abundance, it can be helpful for unveiling the etiology of diseases and providing a functional readout of the physiological state of the human body. Liquid chromatography-mass spectrometry (LC-MS) is a commonly used metabolomics platform due to its feasibility to measure complex samples, such as human plasma and urine (Jones *et al.*, 2012).

The quality of the LC-MS data influences the downstream analysis, including metabolite quantitation, functional interpretation, pathway analysis for disease mechanisms. The datasets normally contain large portions of metabolites with missing observations in some samples. The underlying missingness mechanism is complex. As discussed by Gromski *et al.* (2014), the missingness can be the result of one or any combination of the following factors: (i) the failure in computational detection, (ii) measurement error, (iii) signals are of low intensity which cannot be distinguished from background noise, (iv) imperfection of the detection algorithms used and (v) deconvolution that may result in false negatives. They also argued that imputation techniques should be favored over other methods of handling missingness in LC-MS metabolomics studies.

Various imputation techniques have been developed and applied in metabolomics studies (Armitage *et al.*, 2015; Gromski *et al.*, 2014; Hrydziuszko and Viant, 2012; Taylor *et al.*, 2016), many of which were carried over from the field of microarray gene

expression. They do not utilize two pieces of valuable information that are unique to metabolomics data. The first piece of information is the known metabolic network, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000). There are plenty of literature supporting the idea of utilizing network information in data analysis procedures to improve variable selection and functional interpretation (Aggio *et al.*, 2010; Barupal *et al.*, 2012; Cai *et al.*, 2017; Kessler *et al.*, 2013; Li *et al.*, 2013; Ravasz *et al.*, 2002; Stelling *et al.*, 2002; Xia and Wishart, 2010). Given their close co-regulation, features matched to neighboring metabolites on the network could help predict each other's abundance in the sample. This may only be true for a subset of the metabolites, and the relation could be non-linear, creating a challenge in utilizing such information. However advanced machine learning techniques such as support vector regression (SVR) can utilize non-linear relations, as well as resist the impact of nuisance variables, i.e. those included in the model but have no true predictive power. With the help of such techniques, network information could contribute to missing value imputation.

The second piece of information that we try to utilize is the relationship between features that are likely derived from the same metabolite. Grouping and annotating features based on their mass-to-charge ratio (m/z) and retention time (RT) characteristics have been utilized in feature identification (Kuhl *et al.*, 2012; Silva *et al.*, 2014; Uppal *et al.*, 2017). Potentially features derived from the same metabolite, even if the identity of the metabolite is unknown, can help the imputation of each other. For example, if the monoisotopic weight of a hypothetical molecule M is 100.000, then in data from positive ion mode with ESI ionization, the theoretical m/z values of two of its likely adduct ions are: $[M+H]^+, 101.007276$ and $[M+Na]^+, 122.989218$. Here 'M' represents the metabolite, the element after the plus sign represents the adduct, and the '+' outside the bracket represents the charge state. The difference between the two m/z values does not change with the molecular weight of M. That is, even if a chemical is not in the database, its adduct ions still follow the same pattern in terms of the difference between their m/z values. For example, if we observe two m/z values in the data, and $|m/z_1 - m/z_2|$ is different from $22.989218 - 1.007276$ by no more than $m/z_2 \times 10^{-5}$, and the two features have close RT values, then we consider they are highly likely to be derived from the same metabolite. We note that this relation is *likely* but not *definitive*. We will again rely on the SVR's capability to resist nuisance variables when a false relation is included in the imputation.

Combining the afore-mentioned information and traditional approaches, we propose a missing value imputation algorithm for LC-MS metabolomics data by applying the support vector regression (SVR) algorithm to a predictor network newly constructed among the features. To be specific, the predictor network is built by incorporating the metabolic network and adduct ion relations, together with linear and nonlinear associations between feature abundance levels calculated directly from the data (Fig. 1a). And then we impute each feature with missing values by fitting an SVR model on

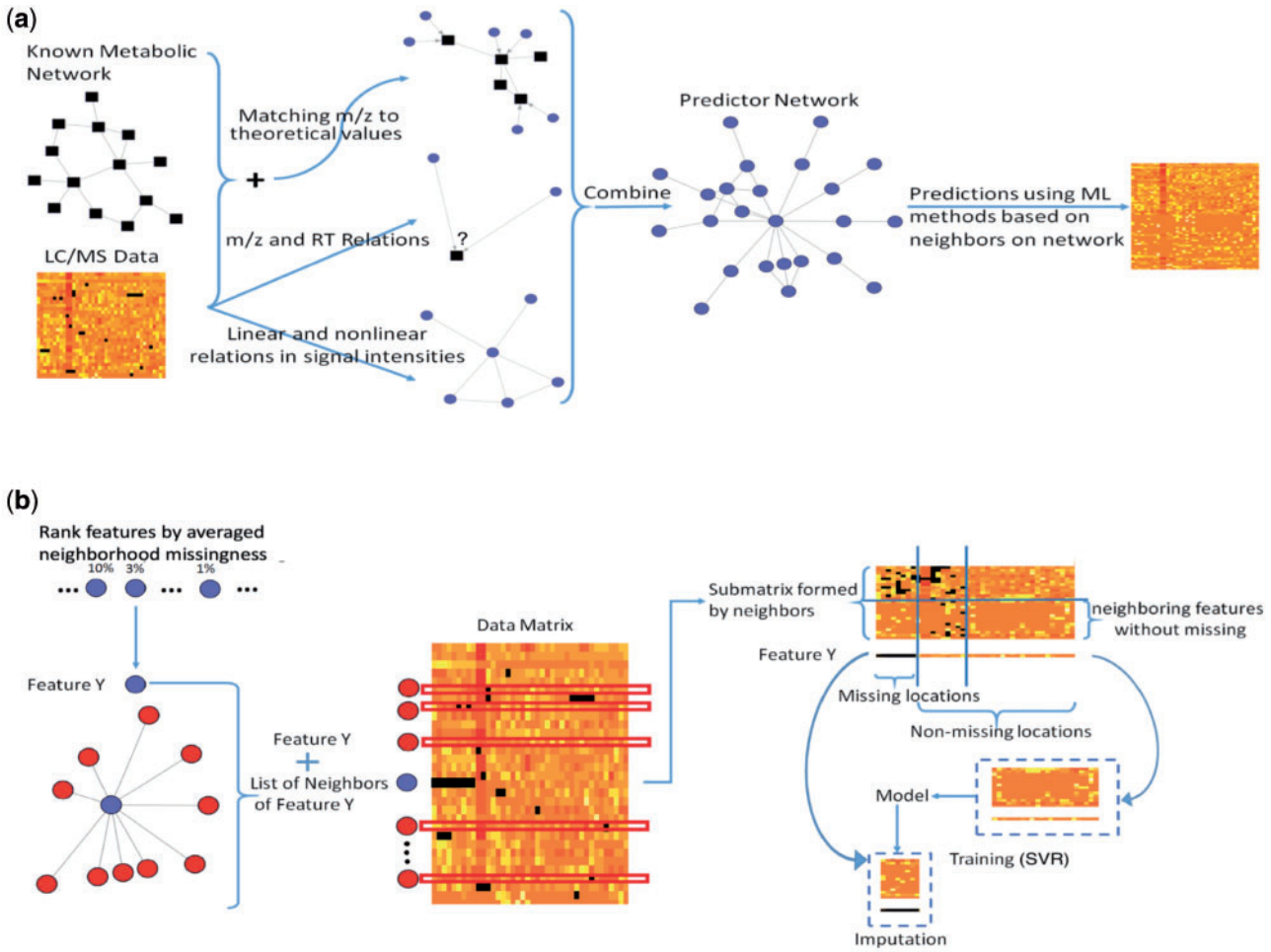

Fig. 1. The workflow of the proposed method. (**a**) Building the predictor network for imputation; (**b**) the imputation procedure given the predictor network

the dataset where the neighboring features on the predictor network are utilized (Fig. 1b). An R package called MINMA (Missing data Imputation incorporating Network and adduct ion information in Metabolomics Analysis) has been developed to implement the algorithm.

2 Materials and methods

2.1 Building the predictor network

The predictor network was constructed on the feature level. The purpose of this network was to represent the feature relations. Essentially, every node on this network was a feature. If two features were considered as 'potentially helpful in imputing each other's missing values', they were connected by an edge between them in the network. To define the 'potentially helpful' features, we mainly considered the feature relations from three sources (Fig. 1a):

- Metabolic Network
  The metabolic network we used in this paper was extracted from the KEGG database (Kanehisa and Goto, 2000). If two metabolites are involved in the same reaction, then they are linked in the metabolic network. Features matched to these two metabolites were considered connected. The matching of features to metabolites was based on matching the theoretical *m/z* of some common adduct ions of the metabolites to the observed *m/z* values of the features at a certain tolerance level (10 ppm in this study). In this proof-of-concept study, as the data were generated from positive ion mode with electrospray ionization (ESI), we considered five adduct ions that are common in this type of data: $[M+H]^+$, $[M+NH4]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+2Na-H]^+$. The specification of ion types can easily be done by user choice of the package.
- *m/z* value differences of common adduct ions
  First we determined what adduct ion forms were included. Then the *m/z* differences between the adduct ions of the same charge were calculated. Pairwise *m/z* differences were calculated for all features in the data. When the *m/z* difference between two features match closely with the theoretical difference between two adduct ions (10 ppm in this study), and their RT difference was less than a pre-defined threshold (100 es in this study), the two features were considered likely to be derived from the same metabolite. They were connected in the feature level network. The same set of five common adduct ions as mentioned above were used in this study.
- Correlation Inferred from Data Matrix
  We consider two features 'neighbors' if they were highly correlated base on the following correlation measures:

  1. Linear correlation: consider 'neighbors' based on n_1 largest pairwise Pearson correlations ($n_1 = 10$ in this study).
  2. DCOL correlation: consider 'neighbors' based on n_2 largest pairwise nonlinear correlations defined by Distance based on Conditional Ordered List (DCOL) (Yu and Peng, 2013). ($n_2 = 10$ in this study)
  3. dCov dependency: consider 'neighbors' based on n_3 largest pairwise general dependencies defined by Brownian distance covariance (Kosorok, 2009). ($n_3 = 10$ in this study)

  These three criteria might generate overlapping feature pairs.

By building the predictor network from multiple sources, it was guaranteed that each feature had at least k connections in the network.

2.2 The imputation procedure

The imputation was based on the predictor network. In the following discussions when network neighborhood is mentioned, we refer to the predictor network. In the imputation of every feature, only its connected features on the predictor network were used as predictors. We firstly introduce some mathematical notations here: $e_{(i,j)}$ represents the value at location (i, j) in data matrix $E = \{e_{(i,j)}, i = 1, \ldots, m, j = 1, \ldots, n\}$, i represents the ith feature (row) and j represents the jth sample (column). If the ith row, feature $e_i = \{e_{(i,j)}, j = 1, \ldots, n\}$ has missing locations, we denote: $e_{i,mis} = \{e_{(i,j)}, j = 1, \ldots, n, \text{where}, e_{(i,j)} = \text{NA}\}$, similarly, we denote $e_{i,obs} = \{e_{(i,j)}, j = 1, \ldots, n, \text{where}, e_{(i,j)} \neq \text{NA}\}$ as the observed locations in feature e_i, here the *mis* and *obs* only indicate locations instead of the values. All the neighboring features of feature i are indexed as $nbr(i)$.

For feature i, we selected the non-missing locations of feature i and used $e_{i,obs}$ as response vector and those neighboring features where they were fully observed in these observed locations denoted as $e_{nbr(i),obs}$ were formed as the predictor matrix. Then we trained the SVR model using $e_{i,obs} \sim e_{nbr(i),obs}$ and extracted the predicted value $\hat{e}_{i,mis}$ when $e_{nbr(i),mis}$ was used as the testing data for imputation.

Before imputation, the sequence for imputing the features with missing locations needs to be decided first or to be updated along the way. In this paper, we utilized a pre-fixed imputation sequence scheme for computation consideration. Specifically speaking, features were firstly ranked by a measure called averaged neighborhood missingness. The averaged neighborhood missingness of one feature was defined as the average number of missing locations of its neighboring features. Then the features with smaller averaged neighborhood missingness were imputed first. After imputing the feature, the imputed values were filled in the original missing locations and were treated as non-missing locations in the following iterations (Fig. 1b). However, the imputation sequence still stayed the same.

2.3 Performance comparision

We compared the proposed imputation algorithm (denoted as Net_SVR) with other commonly used imputation algorithms in metabolomics studies, including the K-Nearest Neighbors (KNN) (Troyanskaya *et al.*, 2001), the Bayesian Principal Component Analysis (BPCA) (Oba *et al.*, 2003), the imputation based on Simple Linear Regression (SLR), the imputation based on Singular Value Decomposition (SVD), the imputation by inserting Single Values (SVI: Min/2, Mean, Median). We briefly describe those methods:

- The K-Nearest Neighbors (KNN) (Troyanskaya *et al.*, 2001) finds the k nearest neighboring features $\{e_j, j = l_1, \ldots, l_k\}$ by a Euclidean metric calculated among those whose feature columns are not missing at location *mis*, and then takes the average values of non-missing locations $e_{j,mis}$ calculated as $\frac{1}{k}\sum_{j=l_1}^{l_k} e_{j,mis}$ for imputation.
- The Bayesian Principal Component Analysis (BPCA) (Oba *et al.*, 2003) simultaneously estimates a probabilistic model for the data matrix and estimates some latent parameter sets within the framework of Bayesian inference, and then impute the missing values in the data matrix by the expectation with respect to the estimated posterior distribution.
- The imputation based on Simple Linear Regression (SLR) is conducted by first fitting a series of univariate simple linear regression models and collecting the predicted value from each SLR model, and then imputing $e_{i,mis}$ by a weighted

summation of all these predicted values, where the weights are decided by their pairwise Pearson correlation only using observed data.

- The imputation based on Singular Value Decomposition (SVD) (Troyanskaya *et al.*, 2001) firstly initializes all missing values by their row means. Each time, given a complete observed matrix, it conducts a SVD procedure that obtains a set of mutually orthogonal expression patterns (eigen-features). And then it imputes the missing values by regressing the features with missing values against the *nPC* eigen-features (*nPCs* need to be pre-specified). This imputation is repeated until the total change of two successive imputations is less than the tolerance value.
- The imputation by inserting a Single Value (SVI: Min/2, Mean, Median). These methods replace all missing values by a pre-calculated value. Common choices are: half of the minimum (Min/2), the mean (Mean) and the median (Median) calculated from all the observed values in the data matrix.

In order to evaluate the performance of each method, we calculated the normalized root mean squared error (NRMSE) of the imputed values. The NRMSE was calculated for all the simulated missing locations that were non-zero in the original data matrix.

Suppose the total number of locations we use in calculation is K, the imputed values are $\widehat{e} = \{\widehat{e}_k, k = 1, \ldots, K\}$ and the ground-truth from the original observed data matrix are $e = \{e_k, k = 1, \ldots, K\}$. The NRMSE is defined as follows:

$$\mathrm{NRMSE}(\widehat{e}, e) = \sqrt{\frac{\sum_{k=1}^{K} (\widehat{e}_k - e_k)^2 / K}{Var(e)}}.$$

The smaller NRMSE is, the lower the prediction errors and the better the imputation method. For better illustration, we further used a metric called 'NRMSE Ratio' for algorithm comparison, such that the plot is on similar scale for all missing rates. For every missing imputation method (MI), it is defined as the ratio of its NRMSE taken over the NRMSE of KNN. Due to the popularity of KNN in this field, we chose to use KNN in the denominator for calculation.

$$\mathrm{NRMSE\,Ratio(MI)} = \mathrm{NRMSE}(\widehat{e}_{\mathrm{MI}}, e)/\mathrm{NRMSE}(\widehat{e}_{\mathrm{KNN}}, e).$$

Based on the definition, if we compare two methods, the smaller the NRMSE Ratio is, the better imputation performance.

In summary, the pseudocodes for the proposed algorithm are listed in the following:

Algorithm 1 BUILD_NET

Input data matrix $E_{m\times n}$; metabolite network G; adduct info-matrix $A = (A_{mz}, A_{rt})$; reference ions names I; tolerance level $tol.mz$, $tol.rt$; number of neighbors: n_1, n_2, n_3;
procedure Create the feature-level predictor network
  neighbors $N = list()$
  for feature i in $1 : m$ **do**
    nbrs.net=$\{j : i \sim j$ in $G\}$
    nbrs.ion=$\{j : \exists p, q \in I, s.t$
    $\frac{||A_{mz}[i]-A_{mz}[j]|-|A_{mz}[p]-A_{mz}[q]||}{|A_{mz}[p]-A_{mz}[q]|} \leq tol.mz$
    and $||A_{rt}[i] - A_{rt}[j]| - |A_{rt}[p] - A_{rt}[q]|| \leq tol.rt\}$
    nbrs.corr = c()
      nbrs.corr1=$\{n_1$ largest linear-correlated features with $i\}$
      nbrs.corr2=$\{n_2$ largest DCOL-correlated features with $i\}$
      nbrs.corr3=$\{n_3$ largest dCov-correlated features with $i\}$
    nbrs.corr = nbrs.corr1 ∪ nbrs.corr2 ∪ nbrs.corr3
    $N[[i]]$=nbrs.net ∪ nbrs.ion ∪ nbrs.corr
  Return N

Algorithm 2 IMP_SEQ

Input data matrix E; predictor network denoted as neighbors list N
procedure Rank features by averaged neighborhood missingness
  impseq = c()
  avemiss = c()
  E.nmiss = apply(E, 1, function(e){sum(is.na(e))})
  for feature i in $1 : m$ **do**
    nbrs.i = N[[i]]
    avemiss[i]=mean(E.nmiss[nbrs.i])
  impseq = rank $(1 : m)$ by avemiss
Return impseq

Algorithm 3 NET_SVR

Input data matrix E; metabolite network G; adduct info-matrix $A = (A_{mz}, A_{rt})$; reference ions names I; tolerance level $tol.mz$, $tol.rt$; number of neighbors: n_1, n_2, n_3;
procedure Build predictor network
  N = BUILD_NET$(E, G, A, I, tol.mz, tol.rt, n1, n2, n3)$
procedure Creat an imputation sequence
  $impseq$ = IMP_SEQ(E, N)
procedure Imputation
  Initialize $\widehat{E} = E$
  for feature i in $impseq$ **do**
    create $e_i = E[i,]$, $e_{i,obs}, e_{i,mis}$
    extract neighbor locations from $N[[i]]$ as $nbr(i)$
    train a SVR model $e_{i,obs} \sim e_{nbr(i),obs}$
    predict $e_{i,mis}$ as $\widehat{e}_{i,mis}$ using $e_{nbr(i),mis}$
    set $\widehat{E}[i, mis] = \widehat{e}_{i,mis}$
  Return $\widehat{E}$

3 Results

3.1 Datasets and simulation setup

In this study, we used two metabolomics datasets denoted as CAD and CHD to assess the performance of different methods. The CAD dataset is from the Emory Cardiovascular Biobank, which consists of patients who have undergone coronary angiography to document the presence/absence of coronary artery disease. Demographic characteristics, medical histories, behavioral factors and fasting blood samples have been documented and details about risk factor definitions and coronary angiographic phenotyping have been described previously (Patel *et al.*, 2012). Each sample was analyzed in triplicate with high-resolution liquid chromatography-mass spectrometry (LC-MS), using anion exchange column combined with the Thermo-

Orbitrap-Velos (Thermo Fisher, San Diego, CA) mass spectrometer in positive ion mode, with a *m/z* range of 85 to 850.

The CHD dataset is a dataset from the Emory-Georgia Tech Predictive Health Initiative Cohort of the Center for Health Discovery and Well Being. This is a cohort of generally healthy university employees aged 18 and older (http://predictivehealth.emory.edu) (Brigham, 2010). The data was generated by C18 column combined with the Thermo-Orbitrap-Velos mass spectrometer in positive ion mode, with a *m/z* range of 85 to 850.

Both datasets were pre-processed using xMSAnalyzer (Uppal *et al.*, 2013) in combination with apLCMS (Yu *et al.*, 2009, 2013). Each sample was run in triplicates in the datasets. For each feature, there were three readings per subject. An average feature intensity value was calculated from the non-zero readings of the three. For filtering the data matrix, rows with more than 20% of zeros were removed. Finally, the data matrix was log-transformed by the function $y = \log(1 + x)$. The CAD dataset contains 18 434 features and 489 samples with 41.34% of the locations being zero. We removed rows with over 20% zeros, resulting in a data matrix of 7033 rows with an overall missing rate of 2%. The CHD dataset contains 8942 features and 415 samples with 43.54% zeros. We removed rows with over 20% zeros, resulting in a data matrix of 3187 rows with an overall missing rate of 7%. In the following simulation procedure, the non-zero values in these matrices served as ground truth. They were knocked out and then imputed, and the imputation accuracy was assessed by NRMSE over these non-zero ground truth values.

As described in the Materials and methods section, we built the predictor network using: (1) linear correlation, (2) DCOL correlation, (3) dCov dependency, (4) difference in *m/z* (relative difference is less than 10 ppm) and RT values (difference less than 100) between any pair of features, indicating high likelihood of them being derived from the same metabolite, (5) *m/z* matching to neighboring metabolites on the KEGG metabolic network. For all KEGG metabolites, we first computed the theoretical *m/z* values of common adduct ions, and then computed the difference between these *m/z* values and the feature *m/z* values. A relative difference less than 10 ppm suggests a potential match. Two features matched to connected metabolites on the KEGG network are connected in the predictor network. For (4) and (5), five adduct ions were considered in this study: $[M + H]^+$, $[M + NH4]^+$, $[M + Na]^+$, $[M + K]^+$, $[M + 2Na - H]^+$. The MINMA package provides the option of using other adduct ions.

3.2 Computation

As indicated by Hrydziuszko and Viant (2012), the missingness may not occurr randomly in metabolomics data based on the analysis of DI FT-ICR MS metabolomics datasets. As a result, assuming a complete random missing mechanism may not be appropriate for imputation. Inspired by their work, we created the simulated datasets by knocking out a portion of locations from the ground-truth matrix by mimicking real missing patterns, and then evaluated each of the algorithms. To be specific, when we simulated a missing rate of *r*, each time we randomly selected one feature *a* from this ground-truth matrix, and one feature *b* from the original input matrix (before removing rows with $> 20\%$ missing). We knocked out the locations (encoded as NA) in feature *a* where there were observed zero values in the corresponding location in feature *b*, until the simulated dataset hit the missing rate of *r*. Similar approach has been taken in microarray missing value imputation study (Yu *et al.*, 2011). In this way, without any assumptions of missing mechanism, imputation algorithms were all evaluated based on the real data missing pattern.

We simulated the datasets with various missing rates: 1, 5, 10, 15, 20, 25, 30, 35 and 40%. For each of them, we generated 50 datasets and used the averaged NRMSE Ratio for evaluation. For each missing percentage, we tested various parameter settings for each method, i.e. $k = 5$, 10, 15 for KNN and $n_1, n_2, n_3 = 5, 10, 15$ for Net_SVR and $nPCs = 5$, 10, 15 for BPCA and SVD, using 5 simulations, and then used the best parameter setting in the full simulation of 50 datasets.

All computations were run under R version 3.3.1. KNN was implemented using the function 'impute.knn' from the package 'impute'; BPCA was performed using the function 'bpca' from the package 'pcaMethods' (Stacklies *et al.*, 2007); SVD was applied using function 'impute.svd' in the package 'bcv'. For our mehtod Net_SVR, the SVR model was fitted using the function 'svm' from the package 'e1071' (Meyer *et al.*, 2017). The packages 'impute' and 'pcaMethods' are Bioconductor packages.

3.3 Simulation results

The simulation results are presented in Figure 2, where we applied all the candidate algorithms for imputation to two real datasets: CAD and CHD. For the simulation results of CAD dataset (Fig. 2a), at each missing rate ranging from 1 to 40%, BPCA, Median Imputation and Net_SVR were below the dash line of 1, which means all three methods outperformed KNN (recall that NRMSE Ratio of KNN is always 1). The averaged NRMSE Ratio for them were 0.893, 0.890 and 0.727, respectively. SLR, SVD and Mean Imputation outperformed KNN only when missing rate was 1% and performed worse than KNN when missing rate was increased. Among all top three methods: BPCA, Median and Net_SVR, when missing rate was as low as 1%, all three of them performed significantly better than KNN, as the missing rate increased, the gap compared to KNN shrank. Across all missing rates, our proposed algorithm Net_SVR performed the best as it obtained the smallest NRMSE Ratio compared to others with a minimum of 0.579 and maximum of 0.762.

The Net_SVR method also outperformed others when we applied all algorithms to the CHD dataset (Fig. 1b). It was the only algorithm that achieved an NRMSE Ratio below 1 across all missing rates. The averaged NRMSE Ratio of Net_SVR was 0.726 with a minimum of 0.518 and a maximum of 0.806. BPCA performed slightly better than KNN in most of the cases, but still yielded larger NRMSE Ratio at missing rate 10% (1.013), and was very close to KNN at missing rates 15% (NRMSE Ratio 1.000) and 20% (NRMSE Ratio 0.996). Median performed worse than KNN for the CHD dataset while it performed better in the CAD dataset, but the overall NRMSE Ratio of Median is around 1.

Additionally, of all the algorithms evaluated, imputing the missing locations by half of the minimum value yielded the largest NRMSE values, even though the data was already log-transformed. It is because the Min/2 approach takes a different assumption than all the other methods. It assumes the unobserved values are missed only when the signal is below a detection threshold, which largely doesn't hold true in metabolomics data, thus, it is the worst among all the imputation algorithms. Our results are generally consistent with previous studies. The studies were somewhat diverse in terms of the data used, as well as the objectives used in judging the performance. Overall they showed a mixed performance between KNN, BPCA and SVD, while simple imputation methods such as

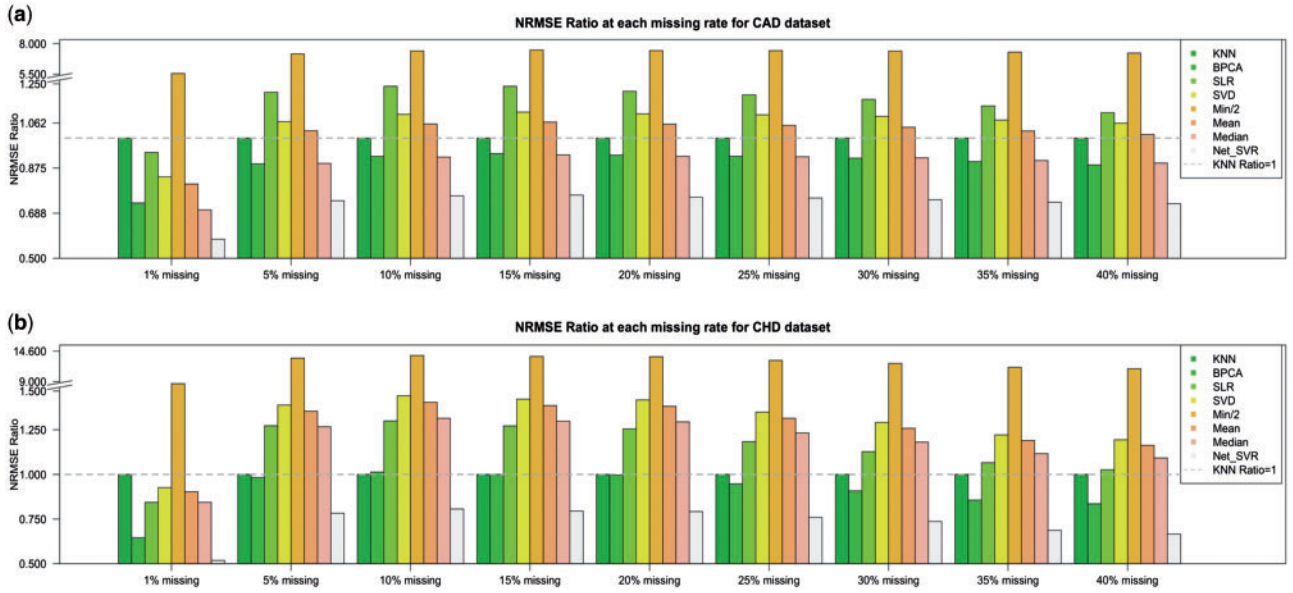

Fig. 2. Simulation results. (**a**) CAD (AE) data; (**b**) CHD (C18) data

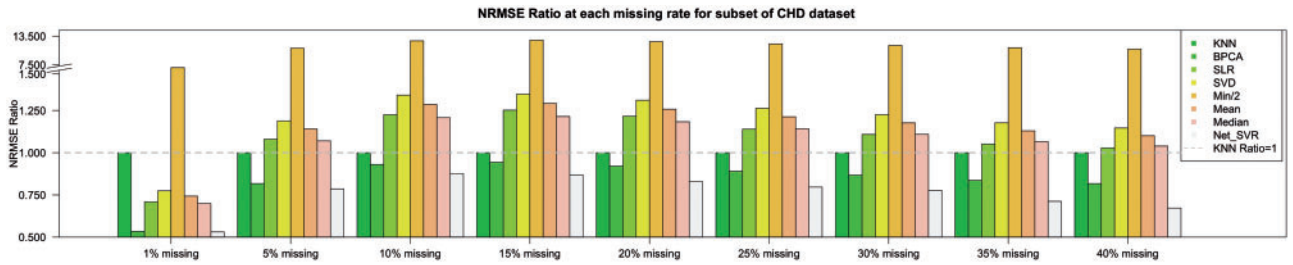

Fig. 3. Simulation results from a subset of the CHD data with 100 columns

Min/2 are in general unfavorable (Armitage *et al.*, 2015; Gromski *et al.*, 2014; Hrydziuszko and Viant, 2012; Taylor *et al.*, 2016). Given the methods' performance may depend on the data type, sample size and missing mechanisms, it is most likely that no method is universally better. On the other hand, the knockout-impute simulation approach can be helpful. Given a specific dataset, a simulation similar to the current manuscript or those previously reported may be helpful in determining which imputation method best suites the data.

In metabolomics data, the underlying missing data pattern is unclear, and the assumptions needed for modeling missing mechanism is hard to justify. Thus in these two simulation studies, the missing locations were generated in a way to mimic the real data missing pattern, which is not in favor of any of the algorithms tested. The results indicated that Net_SVR may be a safer choice given it utilizes diverse information.

The two datasets used both contained over 400 samples. However, some datasets in real-world applications may contain fewer samples. In order to evaluate how the methods perform under the situation of smaller sample sizes, we randomly subsampled the columns of the CHD dataset. The simulation result when we subsampled 100 columns is presented in Figure 3. All the algorithms performed similarly to the results in Figure 2b. With the sample size reduction, BPCA had better relative performance compared to itself but still worse than Net_SVR. Our proposed method still outperformed the others at most missing rates. Further reduction of sample size produced similar results (see Supplementary materials for details).

4 Discussion and conclusion

Imputation techniques are widely used for handling missing data in metabolomics studies. In this paper, we proposed a missing data imputation algorithm where a feature-level predictor network is constructed and then utilized for imputation. We incorporated different information for constructing the predictor network: the existing metabolic network structure, adduct ion relations among features and various linear/nonlinear pairwise correlations calculated from feature abundance levels. They are believed to be potentially helpful in depicting related features which may help in imputing each other's missing values. As this predictor network may include some false edges, hence noise in the imputation model, we applied the SVR model for reducing the influence of possible nuisance variables in the imputation process.

In real-world metabolomics studies, missing mechanism is hard to ascertain and the assumptions needed for modeling real data missing pattern is sometimes hard to justify. In order to better compare some of the widely-used algorithms in this field, we randomly sampled missing patterns from real features to mimic the real data missing pattern in the simulation studies. Simulation results showed that in high-resolution LC-MS data, the proposed algorithm Net_SVR outperforms the others at most missing rate settings.

In the application of the Net_SVR method, correctly specifying the types of adduct ions is important. Using too few adduct ion types causes the loss of valuable links that could contribute to imputation, while using adduct ions that are uncommon in the specific

experimental platform may add many false edges in the predictor network. MINMA provides a function to match feature *m/z* values to 32 positive adduct ions, or 13 negative adduct ions. Alternatively, xMSannotator provides matching to more adduct ions (Uppal *et al.*, 2017). Although *m/z* matching can always yield some false positives, nonetheless the frequency of adduct ion in the match can indicate which types of adduct ions are more common in the data, which can serve as the basis for selecting adduct ions to use.

To summarize, by constructing a feature-level predictor network and then imputing missing values using a SVR model that uses neighborhood predictors on the network, the Net_SVR is an effective imputation method. The method can be extended in several directions: 1. other machine learning methods that are better resistant to nuisance variables can be used in place of the SVR; 2. when constructing the predictor network, different sources of information could be weighted differently based on the user's prior knowledge; 3. other feature relations can be incorporated; 4. if computationally feasible, the imputation sequence can be constantly updated along the way for better utilizing the network information.

Acknowledgements

The authors thank Dr Shuzhao Li and Mr Qingpo Cai for helpful discussions.

Funding

This study was partially funded by National Institutes of Health [grant numbers R01MH105561, R01GM124061] and MSM/Emory Cardiovascular (MECA) Center for Health Equity.

Conflict of Interest: none declared.

References

Aggio,R.B.M. *et al.* (2010) Pathway activity profiling (papi): from the metabolite profile to the metabolic pathway activity. *Bioinformatics*, **26**, 2969–2976.

Armitage,E.G. *et al.* (2015) Missing value imputation strategies for metabolomics data. *Electrophoresis*, **36**, 3050–3060.

Barupal,D.K. *et al.* (2012) Metamapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics*, **13**, 99.

Brigham,K.L. (2010) Predictive health: the imminent revolution in health care. *J. Am. Geriatr. Soc.*, **58**, S298–S302.

Cai,Q. *et al.* (2017) Network marker selection for untargeted lc-ms metabolomics data. *J. Proteome Res.*, **16**, 1261–1269.

Dimitriadou,E. *et al.* (2009) *Package 'e1071'. R Software Package*.

Gromski,P.S. *et al.* (2014) Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*, **4**, 433–452.

Hrydziuszko,O. and Viant,M.R. (2012) Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, **8**, 161–174.

Jones,D.P. *et al.* (2012) Nutritional metabolomics: progress in addressing complexity in diet and health. *Annu. Rev. Nutr.*, **32**, 183.

Kanehisa,M. and Goto,S. (2000) Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kessler,N. *et al.* (2013) Meltdb 2.0-advances of the metabolomics software system. *Bioinformatics*, **29**, 2452–2459.

Kosorok,M.R. (2009) On brownian distance covariance and high dimensional data. *Ann. Appl. Stat.*, **3**, 1266–1269.

Kuhl,C. *et al.* (2012) Camera: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.*, **84**, 283–289.

Li,S. *et al.* (2013) Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.*, **9**, e1003123.

Lindon,J.C. *et al.* (2007) Metabonomics in pharmaceutical R&D. *FEBS J.*, **274**, 1140–1151.

Meyer,D. *et al.* (2017) e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. Available at https://CRAN.R-project.org/package=e1071.

Oba,S. *et al.* (2003) A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Patel,R.S. *et al.* (2012) Association of a genetic risk score with prevalent and incident myocardial infarction in subjects undergoing coronary angiography. *Circ. Cardiovasc. Genet.*, **5**, 441–449.

Ravasz,E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Silva,R.R. *et al.* (2014) Probmetab: an r package for bayesian probabilistic annotation of lc-ms-based metabolomics. *Bioinformatics*, **30**, 1336–1337.

Stacklies,W. *et al.* (2007) pcamethods—a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, **23**, 1164–1167.

Stelling,J. *et al.* (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.

Taylor,S.L. *et al.* (2016) Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices. *Brief. Bioinf.*, bbw010.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for dna microarrays. *Bioinformatics*, **17**, 520–525.

Uppal,K. *et al.* (2013) xmsanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics*, **14**, 15.

Uppal,K. *et al.* (2017) xmsannotator: an r package for network-based annotation of high-resolution metabolomics data. *Anal. Chem.*, **89**, 1063–1067.

Xia,J. and Wishart,D.S. (2010) Metpa: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, **26**, 2342–2344.

Yu,T. and Peng,H. (2013) Hierarchical clustering of high-throughput expression data based on general dependences. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **10**, 1080–1085.

Yu,T. *et al.* (2009) aplcms–adaptive processing of high-resolution lc/ms data. *Bioinformatics*, **25**, 1930–1936.

Yu,T. *et al.* (2011) Incorporating nonlinear relationships in microarray missing value imputation. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 723–731.

Yu,T. *et al.* (2013) Hybrid feature detection and information accumulation using high-resolution lc-ms metabolomics data. *J. Proteome Res.*, **12**, 1419–1427.