HHS Public Access

Author manuscript

J Stat Comput Simul. Author manuscript; available in PMC 2015 September 25.

Published in final edited form as:

J Stat Comput Simul. 2015 ; 85(9): 1902–1916. doi:10.1080/00949655.2014.907801.

Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect

Y. Wan[a], **S. Datta**[a], **D.J. Conklin**[b], and **M. Kong**[a,*]

[a]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA

[b]Division of Cardiovascular Medicine, Department of Medicine, University of Louisville, Louisville, KY, USA

Abstract

The statistical methods for variable selection and prediction could be challenging when missing covariates exist. Although multiple imputation (MI) is a universally accepted technique for solving missing data problem, how to combine the MI results for variable selection is not quite clear, because different imputations may result in different selections. The widely applied variable selection methods include the sparse partial least-squares (SPLS) method and the penalized least-squares method, e.g. the elastic net (ENet) method. In this paper, we propose an MI-based weighted elastic net (MI-WENet) method that is based on stacked MI data and a weighting scheme for each observation in the stacked data set. In the MI-WENet method, MI accounts for sampling and imputation uncertainty for missing values, and the weight accounts for the observed information. Extensive numerical simulations are carried out to compare the proposed MI-WENet method with the other competing alternatives, such as the SPLS and ENet. In addition, we applied the MIWENet method to examine the predictor variables for the endothelial function that can be characterized by median effective dose (ED50) and maximum effect (Emax) in an ex-vivo phenylephrine-induced extension and acetylcholine-induced relaxation experiment.

Keywords

variable selection; multiple imputation; elastic net; penalized least squares

1. Introduction

Missing data are a common problem in various settings including clinical trials, animal studies, and survey sampling.[1–3] When analysing data with missing values, a straightforward strategy is to conduct a complete case analysis, where the observations with any missing values are ignored. This approach is simple yet ignores the possible differences between the complete cases and incomplete cases that may result in a substantial bias when the subjects with complete observations are not a random sub-sample of all subjects.[1] The complete case analysis also may lose information, and thus, results in incorrect inferences.

*Corresponding author. maiying.kong@louisville.edu, m0kong02@exchange.louisville.edu.

[1,4] Because experiments in medical research are usually expensive, the need for adequate handling of missing data is a constantly recognized source of concern.[5] Instead of the complete case analysis, a more sophisticated approach called single imputation is used to impute the missing values with plausible values, and then statistical analyses are carried out on the imputed data set. However, the single imputation method ignores the uncertainty of imputation on the missing values that may lead to the underestimation of variances and the distortion of the correlation structure of the data. Therefore, simple single imputation is usually not recommended.[1–4] Multiple imputation (MI) has gradually become a more wellaccepted imputation-based statistical technique for handling missing data since the publication of Rubin's pioneering work for nonresponses in survey.[1] MI procedure involves imputing each missing value with $M\ (>1)$ independent plausible values, and then applying the standard analysis to each imputed data set. The final estimates of the parameters and their variances are obtained from the M sets of estimates using Rubin's rules, with accounting for the uncertainty among MIs.[3,4] The objective of MI method is not to predict missing values as close as possible to the true values but to handle missing data so that valid statistical inferences can be made.[3,4] Rubin's rules have become the gold standard when data are missing at random (MAR).[6–8] By the definition of Little and Rubin,[1] the three general types of missing mechanism are: (1) missing complete at random (MCAR); (2) MAR; and, (3) not missing at random (NMAR).[1–3] Standard implementation of MI relies on an assumption that missing data are either MCAR or MAR, while the MI procedure may also be extended to the cases where missing data are NMAR. [7,9,10]

Variable selection is increasingly important in modern data analysis. Many techniques, such as the least absolute shrinkage and selection operator (LASSO),[11] the elastic net (ENet), [12] and the sparse partial least squares (SPLS),[13] have been developed to select important variables that are associated with outcome variables. LASSO minimizes the restricted least squares with the constraint on the absolute values of the parameters (i.e. L_1 norm), and ENet minimizes the constrained least squares with the constraint on the combination of the absolute and the squared values of parameters.[11,12,14] SPLS maximizes the correlation between outcome variables and the linear combinations of predictor variables (covariates) with constraints on the L_1 norm of the parameters.[13] The constraint for LASSO can be considered as a special case of the ENet, and several studies have shown that ENet performs better than LASSO.[12] These methods have assumed that the observations in the data set are complete. How to apply these variable selection methods to the situation when there are missing values is an important yet unresolved problem.

Several approaches to combine the variable selection methods with MI techniques have been proposed recently.[9,15,16] Wood et al. [9] proposed a 'stacked' approach by combining the multiply imputed data sets into one and using a weighting scheme to account for the fraction of missing data in each predictor variable. However, the variable selection method used by them was the classical backward stepwise selection approach. Heymans et al. developed and tested a methodology combining MI with bootstrapping techniques for studying prognostic variable selection.[15] Chen and Wang proposed an MI-LASSO variable selection method as an extension of the LASSO method to MI-based data, which is, to the best of our

knowledge, the only work combining the penalized least-squares method with MI-based data.[16] In the work,[16] the observations with missing values and those without missing values are treated with equal importance. In this article, we propose an MI-based weighted ENet (MI-WENet) method as an extension of the ENet to the stacked multiple imputed data, with a weight accounting for the proportion of the observed information for each observation. The cyclical coordinate descent methods [17] are applied to minimize the weighted penalized least squares associated with the MI-WENet variable selection method.

To describe our new approach, in Section 2, we first review the two most popular variable selection methods: SPLS and ENet, and then we propose the MI-based SPLS (MI-SPLS) and the MI-WENet for analysing data with missing values. Then in Section 3, we carry out extensive numerical simulations to evaluate the performance of the proposed methods, and compare the performance of the proposed methods with the other competing methods. For Section 4, we apply the proposed MI-WENet method to examine the predictor variables for the maximum effect and the median effective dose in an ex-vivo phenylephrine-induced extension and acetylcholineinduced relaxation experiment study. Finally, we provide a discussion of the pros and cons of our current approach in Section 5.

2. Methods

Let Y_i denote the outcome variable and X_{ij} be the jth predictor variable $(j = 1, \ldots, p)$ for the ith subject $(i = 1, \ldots, n)$. Without loss of generality, we assume that Y_i and X_{ij} are standardized to have zero mean and unit standard deviation. For simplicity, we consider the following linear regression model:

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i, \quad i=1,\ldots,n, \quad (1)$$

where the regression coefficients $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ are unknown parameters to be estimated, and the error term ε_i are independently identically distributed as $N(0, \sigma^2)$.

2.1. Review of SPLS and ENet

The SPLS [13] is an extension of partial least-squares regression (PLS) [18] to achieve simultaneous dimension reduction and variable selection. The PLS begins with calculating the first latent direction vector t_1 as $X\hat{\beta}^{(1)}$, where $\hat{\beta}^{(1)}$ is obtained by maximizing the correlation between the response variable Y and the linear combination of covariates, $X\beta$, i.e.

$$\hat{\beta}^{(1)} = \arg\max_{\beta} \left\{\beta^T X^T Y Y^T X \beta\right\}, \quad subject \quad to \quad \beta^T\beta = 1. \quad (2)$$

Suppose the kth ($k \geq 1$) direction vector, $t_k = X\hat{\beta}(k)$, has been obtained. Denote $T = (t_1, t_2, \ldots, t_k)$ and $M_T = I - T(T^{\mathrm{T}}T)^{-1}T^{\mathrm{T}}$, the $(k + 1)$th direction vector can be obtained by solving Equation (2), with Y replaced by its orthogonal projection onto the complementary of the column space of the known direction vectors T, i.e. replacing Y by $M_T Y$. This process is repeated to obtain a small number of direction vectors. Regressing the original Y on those direction vectors result in a relationship between Y and X due to each direction

vector is a linear combination of the covariates *X*. PLS has become a very popular tool in the field of chemometrics and bioinformatics.[19,20] The SPLS achieves the sparsity of the coefficients on *X* by adding the L_1 constraints on β.[13] For example, $\hat{\beta}^{(1)}(1)$ is updated as

$$arg \max_{\beta} \left\{ \beta^T X^T Y Y^T X \beta \right\}, \quad subject \quad to \quad \beta^T \beta = 1 \quad and \quad \| \beta \|_{L_1} \leq \lambda, \quad (3)$$

where $\| \beta \|_{L1} = \sum_{j=1}^{P} |\beta_j|$. The L_1 constraint is added to obtain each direction vector.[13] SPLS obtains good performance in prediction and variable selection by producing sparse linear combinations of the original predictors, and is especially applicable when *p* is much greater than *n*.[13]

The ENet [12] is a widely applied regulation and the variable selection method. The ENet estimator is obtained by undoing the shrinkage for the naïve ENet estimator that is obtained by minimizing the penalized least squares

$$L(\lambda, \alpha, \beta_0, \beta) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i^T \beta \right)^2 + \lambda P_\alpha(\beta), \quad (4)$$

where

$$P_\alpha(\beta) = \alpha \| \beta \|_{L_1} + \frac{1}{2}(1-\alpha) \| \beta \|_{L_2} = \sum_{j=1}^{p} \left\{ \alpha |\beta_j| + \frac{1}{2}(1-\alpha) \beta_j^2 \right\}. \quad (5)$$

Here, P_α is the ENet penalty that is a compromise between the ridge regression penalty (α = 0) [21] and the LASSO penalty (α = 1).[11] Ridge regression is known to shrink the coefficients of correlated predictor variables, allowing them to borrow strength from each other.[14,21] The ENet penalty with $\alpha = 1 - \varepsilon$, for some small $\varepsilon > 0$, performs much like the LASSO but removes any degeneracies and wild behaviour caused by extreme correlations. [17] For a given λ, as α increases from 0 to 1, the sparsity of the solution to Equation (4), i.e. the number of coefficients being zero, increases monotonically from 0 to the sparsity of the LASSO solution. The naïve ENet estimator obtained from Equations (4) and (5) does not perform satisfactorily,[12] while the ENet estimator that undoes the shrinkage for the naïve ENet, performs much better even compared with LASSO and ridge regression. The ENet estimator is obtained as

$$\hat{\beta}(ENet) = (1 + \lambda(1-\alpha)) \hat{\beta}(naive \quad ENet). \quad (6)$$

The ENet penalty is particularly useful in the cases that *p* is greater than *n* and there are many correlated predictors,[12] which has also been shown in our simulation studies.

2.2. MI-SPLS and MI-WENet

Both the SPLS and ENet methods assume that all covariates and outcome variables are fully observed. In the cases that there are missing values, Rubin's rules provide a general framework to handle missing problems provided missing data are MAR or MCAR.[1–4] However, Rubin's rules can not be directly applied to SPLS or ENet, because the variables

selected for one imputed data set may be quite different from those based on another imputed data set. To the best of our knowledge, there is no standard rule to combine the selected variables resulted from different imputed data sets.[8,9,16,22]

To overcome the shortcoming in combining the multiple results from MI data, we propose to select variables based on the stacked MI data. To be specific, let us assume that the outcome variable is fully observed, but the predictor variables may have some missing values. The missing values in the variables are imputed *M* times independently to generate *M* imputed data sets. We denote the *m*th imputed data set as $\left(y_i;\quad x_{i1}^{(m)},\ldots,x_{ip}^{(m)}\right)_{i=1}^{n}$, for *m* = 1, … , *M*, where $x_{ij}^{(m)}$ is the value of the *j*th predictor variable for the *i*th subject in the *m*th imputed data set. If X_{ij} is observed, then we have $x_{ij}^{(1)}=\cdots=x_{ij}^{(M)}=x_{ij}$ and if X_{ij} is missing, then x_{ij} may take different values in each imputation. Popular softwares for implementing MI procedure include the R-packages *mice* [23] and *mi*,[24] the SAS software *IVEware*,[25] and a module named *MULTIPLE IMPUTATION* in SPSS. In the simulation studies, we applied the R-package *mice* that is based on the sequential regression MI, i.e. the multivariate imputation by chained equations, to impute missing data.[23,25] In applying the R-package *mice*, users are allowed to specify the conditional distribution of each variable on the other variables in the data. The imputation was carried out based on the specified conditional distribution for the missing variables.[23]

Once *M* imputed data sets are obtained, one may stack the *M* imputed data sets as a large complete data set having $M \times n$ observations. SPLS and ENet can be directly applied to this single stacked data set. These approaches are called MI-SPLS and MI-based ENet (MI-ENet), respectively. In general, the estimates based on the stacked MI data are unbiased if the estimates based on a single data set are unbiased, while the standard errors based on the stacked MI data will be under-estimated if they can be estimated.[8] For the MI-ENet method, a simple way to correct the under-estimated errors is to apply a weight to each observation. Denote this weight by w_i for subject *i*. For the stacked *M* imputed data sets, one could assign $w_i = 1/M$ thus the overall weight for a subject is 1. This weighting scheme puts the same weight for each subject and ignores the degree of missing information. A more legitimate way is to assign weights according to the quality of the observed information. If a subject has more missing predictor variables, the weight assigned to the subject should be smaller. We propose to assign the weight $w_i = f_i/M$, where f_i is the fraction of observed values for subject *i*, i.e. the ratio of number of observed variables for the subject *i* to the total number of predictor variables *p*. This approach is named as the MI-WENet method.

The MI-WENet minimizes the following penalized weighted least squares:

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i\left(y_i-\beta_0-x_i^{(m)T}\beta\right)^2 \quad + \quad \lambda P_\alpha(\beta), \quad (7)$$

where $\beta = (\beta_1, \ldots, \beta_p)$. The penalty here is the same as the ENet penalty in Equation (4). We propose to standardize each predictor variable first based on the available data, then carry out the MI to get M imputed data sets. In the stacked data, the values for each variable may

not have mean zero and variance 1 and the intercept may not be the mean of the observed responses anymore. Thus β_0 needs to be estimated in the same manner as the other regression parameters β. By avoiding any re-standardization in the stacked data, $\sum_{m=1}^{M} w_i\left(y_i-\beta_0-x_i^{(m)^T}\beta\right)^2$ will reduce to $\left(y_i-\beta_0-x_i^{\mathrm{T}}\beta\right)^2$, if there is no missing predictor variable for subject i. Thus, the objective function is reduced exactly to the standard ENet, when there is no missing value at all in the original data.

Denote the objective function (7) as $R(\beta_0, \beta)$. To solve for (β_0, β), a coordinate descent method can be applied.[17] Assuming the current estimated $\hat{\beta}_0$ and $\hat{\beta}$ are known, we wish to update $\hat{\beta}_j$ as $\hat{\beta}_j+\Delta\beta_j$ by partially optimizing $R(\beta_0, \beta)$ with respect to β_j ($j = 0, 1, \ldots, p$). Note that the gradient for $\Delta\beta_j$ at $\beta_j = \beta_j \neq 0$, which only exists if $\beta_j = 0$, is

$$\begin{aligned}\frac{\partial R(\beta_0,\beta)}{\partial\Delta\beta_j} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i\left(y_i-\beta_0-x_i^{(m)T}\beta-x_{ij}^{(m)}\Delta\beta_j\right)\left(-x_{ij}^{(m)}\right)\\ &\quad +\lambda(1-\alpha)(\beta_j+\Delta\beta_j)+sign(\beta_j)\lambda\alpha,\end{aligned} \tag{8}$$

where $(\beta_0,\beta)=\left(\hat{\beta}_0,\hat{\beta}\right)$. $Set\quad \partial R(\beta_0,\beta)/\partial\Delta\beta_j=0$, one can get

$$\Delta\beta_j=\frac{(1/n)\,\Sigma_{i=1}^{n}\Sigma_{m=1}^{M} w_i\left(y_i-\beta_0-x_i^{(m)T}\beta\right)x_{ij}^{(m)}-sign(\beta_j)\lambda\alpha-\lambda(1-\alpha)\beta_j}{(1/n)\,\Sigma_{i=1}^{n}\Sigma_{m=1}^{M} w_i x_{ij}^{(m)2}+\lambda(1-\alpha)}, \tag{9}$$

where $(\beta_0,\beta)=\left(\hat{\beta}_0,\hat{\beta}\right)$. $Set\quad \partial R(\beta_0,\beta)/\partial\Delta\beta_j=0$. Then β_j is updated as follows:

$$\begin{aligned}\tilde{\beta}_j^{(new)} &= \tilde{\beta}_j+\Delta\beta_j\\ &= \frac{(1/n)\Sigma_{i=1}^{n}\Sigma_{m=1}^{M} w_i\left(y_i-\beta_0-\Sigma_{l=l,l\neq j}^{p}\right)x_{il}^{(m)}\beta_l\Big)x_{il}^{(m)}-\lambda\alpha sign(\beta_j)}{(1/n)\Sigma_{i=1}^{n}\Sigma_{m=1}^{M} w_i x_{ij}^{(m)2}+\lambda(1-\alpha)}\\ &= \frac{S(1/n)\Sigma_{i=1}^{n}\Sigma_{m=1}^{M} w_i\left(y_i-\beta_0-\Sigma_{l=l,l\neq j}^{p}\right)x_{il}^{(m)}\beta_l\Big)x_{il}^{(m)},\lambda\alpha}{(1/n)\Sigma_{i=1}^{n}\Sigma_{m=1}^{M} w_i x_{ij}^{(m)2}+\lambda(1-\alpha)},\end{aligned} \tag{10}$$

where $(\beta_0,\beta)=\left(\hat{\beta}_0,\hat{\beta}\right)$. $Set\quad \partial R(\beta_0,\beta)/\partial\Delta\beta_j=0$, and $S(z, \gamma)$ is the soft-thresholding operator with value

$$sign(z)(|z|-\gamma)_+=\begin{cases} z-y & if\, z>0 \quad and\, \gamma<|z|,\\ z+y & if\, z<0 \quad and\, \gamma<|z|,\\ 0 & if\, \gamma\geq|z|.\end{cases}$$

To reduce imputation burden, for a given multiple imputed stacked data set and a given weight, one may first calculate and store the following quantities:

$$XY_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i y_i x_{ij}^{(m)} \quad for \quad j=0,1,\ldots,p.$$

$$XX_{jj'} = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i y_i x_{ij}^{(m)} x_{ij'}^{(m)} \quad for \quad 0 \le j \le j' \le p.$$

Here $x_{ij}^{(m)}$ is set to 1 for $j = 0$. Suppose that $\hat{\beta}_j^{(old)}$ $(j=0, 1, \ldots, p)$ are the available values at the previous iteration, one may update β_j $(j = 0, 1, \ldots, p)$ by

$$\tilde{\beta}_j^{(new)} = \frac{S\left(XY_j - \Sigma_{l<j} XX_{jl}\tilde{\beta}_l^{(new)} - \Sigma_{l>j} XX_{jl}\tilde{\beta}_l^{(old)}, \lambda\alpha\right)}{XX_{jj}+\lambda(1-\alpha)}. \quad (11)$$

The procedure is repeated until convergence to get the estimates for β_j $(j = 0, 1, \ldots, p)$. These estimates are similar to the naïve ENet estimates,[12] which can be obtained by a truncation at $\lambda\alpha$ and a shrinkage with a factor $XX_{jj} + \lambda(1 - \alpha)$ for β_j . A better estimate that undoes the shrinkage is obtained by

$$\hat{\beta}_j\,(weighted \quad ENet) = (XX_{jj}+\lambda(1-\alpha))\,\hat{\beta}_j\,(weighted \quad naive \quad ENet). \quad (12)$$

The weighted ENet estimates in Equation (12) are used in the simulations in Section 3 and the case study in Section 4, and performs well in both variable selection and prediction.

In the present work, we applied 10-fold cross validation method to select the tuning parameters α and λ. Here $\alpha \in (0, 1)$, and $\lambda > 0$. Because (α, λ) determines the soft-threshold boundary, we start with a sequence grid value for α. For each fixed α, we compute the solution for a decreasing sequence of values for λ starting at the largest value λ_{max} for which the entire vector $\hat{\beta}$, i.e.

$$\alpha\lambda_{max} = \max_{0\le j\le p} |XY_j|,$$

and set $\lambda_{min} = E\lambda_{max}$ with $E = 0.001$. We construct a sequence of λ values decreasing from λ_{max} to λ_{min} on the log-scale. The pair of (α, λ) is chosen such that the cross validation error is minimized.

3. Simulation

In this section, we design different simulation schemes to examine the performance of the proposed MI-WENet method and compare it with the other methods, such as MI-SPLS and MI-ENet. The different simulation scenarios are reported in Section 3.1; the corresponding simulation results are reported in Section 3.2.

3.1. Simulation settings

In the simulation studies, we assume that the underlying model is known and has the form of $Y_i = X_i\beta + \varepsilon_i$, for $i = 1, \ldots, n$, where $X_i = (X_{i,1}, \ldots, X_{i,p})$, $\beta = (\beta_1, \ldots, \beta_p)^T$, and $\varepsilon_i \sim N(0, \sigma^2)$. The predictor variables for each subject were generated from a multivariate normal

distribution with mean zero and a covariance matrix I:. σ was set as the value such that the signal to noise ratio is 2, i.e. $\sqrt{\beta^{\mathrm{T}}\Sigma\beta}/\sigma=2$

Simulation scenarios were designed based on various assumptions of sample size n, number of predictor variables p, missing mechanism, missing pattern and correlation structure of the predictor variables. Correlation structure for the predictor variables of the ith subject (i = 1, … , n) was tested under three specifications: (1) compound symmetry with low correlation, i.e. corr($X_{i,j}$, $X_i{}^{j1}$) = 0.1; (2) compound symmetry with medium correlation, i.e. corr($X_{i,j}$, $X_i{}^{j1}$) = 0.5; and (3) first-order autoregressive (AR(1)), i.e. corr($X_{i,j}$, $X_i{}^{j1}$) = $0.8^{|j-j^1|}$, for j, j^1 = 1, … , p and $j = j^1$, respectively. We set the homogenous variances as 1 for all $X_{i,j}$, so the covariance matrix I: was same as the correlation matrix. Under each specification, we induced missing values under the MCAR and MAR mechanisms, respectively; and for each missing mechanism, missing values were generated with independent and monotone missing patterns, respectively. In total, 17 scenarios were tested in our simulations, which we believe have covered most situations in practical application. The independent missing pattern means that the missing observations for different variables are independent, and the monotone missing pattern is that a missing observation in x_{ij} (where i is the subject index, and j is the variable index) implies that all observations $x_{ij}1$ for $j \le j^1 \le p$ are missing.

For each scenario with fixed n, p, I:, missing mechanism and missing pattern, the following steps are carried out:

(1) Generate fully observed predictor variables for X_i (i = 1, … , n).

(2) Generate the outcome variable for Y_i from the underlying model $Y_i = X_i\boldsymbol{\beta} + \varepsilon_i$ (i = 1, … , n), where $\varepsilon_i \sim N(0, \sigma^2)$.

(3) Independently generate test data set (x_t , y_t) (t = 1, … , n_t) by repeating steps 1 and 2, where the sample size n_t is larger than n (n_t = 1000 in our simulations).

(4) Fit the full data set that has a sample size n and has been generated in steps 1 and 2 by using SPLS and ENet, respectively (see the rows named as *Full-SPLS* and *Full-ENet* in Tables 1–3).

(5) Induce missing values for the predictor variables according to each pre-specified missing mechanism and missing pattern.

(6) Fit the data set including complete cases only by using SPLS and ENet, respectively (see the rows named as *CC-SPLS* and *CC-ENet* in Tables 1–3).

(7) Impute missing values M times (M = 5), and stack the M imputed data sets into an enlarged one.

(8) Perform SPLS, ENet and WENet based on the first single imputed data set (see the rows named *SI-SPLS*, *SI-ENet* and *SI-WENet* in Tables 1–3), and based on the stacked data set (see the rows named *MI-SPLS*, *MI-ENet* and *MI-WENet* in Tables 1–3).

(9) Repeat Steps 1-8 100 times, and summarize the averaged key performance measures of each method.

The key performance measures for each method under each simulation scenario are predicted mean-squared error (PMSE), mean-squared error (MSE), sensitivity (SENS in Tables 1–4) and specificity (SPEC in Tables 1–4). The PMSE is defined as

$$PMSE\left(\hat{\beta}\right) = \frac{1}{n_t}\sum_{t=1}^{n_t}\left(y_t - x_t\hat{\beta}\right)^2,$$

where x_t and y_t are fully observed independent test data generated in Step 3, and $\hat{\beta}$ is the estimate of the underlying regression parameter β for each model. PMSE is obtained by averaging the predicted errors on a large number of observations, where we have set n_t as 1000. The MSE is defined as

$$MSE = \left(\hat{\beta} - \beta\right)^T \Sigma \left(\hat{\beta} - \beta\right),$$

where $\hat{\beta}$ and β are the same as for PMSE, lower values of PMSE and MSE are desirable. The sensitivity is defined as the fraction of variables selected among those whose coefficients are not zero in the underlying model, and the specificity is defined as the fraction of variables not selected among those whose coefficients are zeros in the underlying model. Larger sensitivity and specificity indicate a better performance.

To examine the performance of different methods, we first fixed p = 12, n = 50 and β = (3, 1.5, 0, 0, 2, 0, 3, 1.5, 0, 0, 2, 0)^T, and we considered the combinations of different missing mechanism (MCAR and MAR), different missing pattern (independent and monotone) and different correlation structure for the predictor variables. Under the MCAR scheme, the independent missing pattern was generated by independently removing 16% of the observations from each of the first six predictor variables, which resulted in around 50% observations containing missing values; the monotone missing pattern was generated by first inducing missing values to the 8% of randomly sampled observations from the first to sixth predictors, and then repeatedly adding missing values to another 8% randomly sampled observations from the second to sixth, third to sixth, fourth to sixth, fifth to sixth, and the sixth only predictor variables, which eventually resulted in 48% subjects containing missing values. The simulation results for MCAR, with different missing patterns and different correlation structures for the predictor variables, are reported in Table 1. For MAR, missing values were induced by the following logistic regression model:

$$logit\left\{Pr\left(X_{ij(m)} \quad is \quad missing \quad |X_{ij(c)}, Y_i\right)\right\} = X_{ij(c)}+Y_i, \quad for \quad i=1,\ldots,n. \quad (13)$$

Here, $j^{(m)}=1,\ldots,p_1$ are indices for the predictor variables in which missing values are to be induced, and $j^{(c)}=p_1+j^{(m)}$ are indexes for the completely observed predictor variables. When p = 12, p_1 is set as 6. For independent missing pattern, the procedure to generate

missing values was the same as in the MCAR cases, except that the 16% removed observations for each of the six missing predictor variables were selected by the highest probabilities calculated from the logistic model (13). For monotone missing pattern, we applied the logistic model (13) to the whole data set first, and removed 8% observations from the first to sixth predictor variables according to the missing probabilities for the first predictor variable. We then applied the logistic model (13) to the remaining data set with complete cases only, and removed 8% additional observations from the second to sixth predictor variables according to the missing probabilities of the second predictor variable. Repeating above procedure until 8% additional observations were removed for the sixth predictor variable only, resulted in 48% subjects containing missing values in total. The corresponding simulation results for MAR are reported in Table 2.

We also conducted simulations with different combinations of p and n, under the specification of monotone MAR and AR(1) correlation structure, so that the performance of different methods with large p and small n (say $p = 24$, 48, and 60, with n fixed at 50) and with small p and large n (say $n = 50$, 100, and 200, with p fixed at 12) can be examined. Here, when $p > 12$, the β in the underlying models were set as the repetitions of $(3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0)$. The procedures for generating monotone MAR missing values were similar as when $p = 12$ and $n = 50$. In the cases when $p = 24$ and $p = 48$, p_1 in Equation (13) were set as $p/2$, and the percentages of missing values in each iteration were controlled at 4% and 2%, respectively. When $p = 60$, p_1 was set as 24 and the percentage of missing value in each iteration was controlled at 2%. The total missing percentage was fixed at 48% under each scenario. The corresponding simulation results are reported in Table 3.

The number of simulation runs is 100 in Tables 1–3. To examine whether a large number of simulation runs impacts the simulation results, we carried out the simulations with 500 runs for each scenario given in Table 3. The corresponding results are reported in Table 4.

3.2. Simulation results

The results for MCAR with different missing patterns and different correlations for X are summarized in Table 1, and results for MAR are summarized in Table 2. The results for MCAR (Table 1) and MAR (Table 2) explain consistent improvement in the estimation and prediction errors using the MI-WENet procedure compared with others. From Tables 1 and 2, we see that: (1) Full-ENet is consistently having lower PMSE and MSE than those from Full-SPLS. When correlations of X are low, Full-ENet has both higher sensitivity and specificity compared with Full-SPLS; when correlations of X are medium to high, Full-ENet has similar or a little lower sensitivity (within 12%), while the specificity is around 30% higher than those from Full-SPLS, indicating that the ENet method has better performances than the SPLS method for the variable selection and prediction in our simulations. (2) Based on complete cases analysis, both SPLS and ENet (CC-SPLS and CC-ENet) methods have much higher PMSE and MSE than all other imputation based methods; the sensitivity for CC-ENet dropped 30–50% compared with the Full-ENet, and the specificity for CC-SPLS are generally low. All these measurements indicate that CC-SPLS and CC-ENet are not recommended. (3) MI-SPLS has a high sensitivity but the specificity is at least 30% lower than MI-WENet, indicating that MI-SPLS would select more variables of those should not

be selected. (4) In all the tested simulation scenarios, the MI-WENet method generally obtains the lowest PMSE and MSE among all competing imputation methods considered here with an exception in Table 1. That is, for the independent MCAR case when the correlations are following an AR(1) process, the PMSE and MSE for MI-WEnet are slightly larger than the other imputation-based Enet method. The sensitivity and specificity of the MI-WENet is always close to the full-ENet model. Opposed to that, other imputed ENet models gain sensitivity with a significant loss in specificity compared with the full-ENet model. MI-WENet also maintains a reasonable sensitivity and specificity across all the simulation scenarios. This demonstrates that the MI-WENet method outperforms all the other methods.

Table 3 displays the results based on different combinations of p and n, under the specification of monotone MAR and AR(1) correlation structure. The first column in Table 3 shows the performance of different methods for fixed $p = 12$, when n increases, say $n = 50$, 100, and 200. The results demonstrate that: as n increases, (1) the PMSE and MSE for each method decreases, which means that the prediction becomes more accurate as n goes larger; (2) the sensitivity increases, indicating that as n increases, the percentage of correctly selected variables increases; (3) the specificity stays almost the same, indicating that the sample size does not impact the percentage of correctly rejected variables effectively; (4) among all the imputation methods, the MI-WENet method has the best performance in terms of smallest PMSE and MSE, and relatively high sensitivity and specificity compared with all the ENet-based imputation methods. However, we observe a higher sensitivity with a significant loss of specificity in the SPLS-based imputation methods. Although we see the reduced sensitivity in MI-ENet imputations for highly correlated data compared with many SPLS-based imputations. The lower sensitivity is not as severe compared with the loss of specificity in the SPLS-based imputations. The second column in Table 3 illustrates the performance of different methods when the number of predictor variables increases from 24 to 60 with fixed sample size n at 50, from which we conclude that: (1) as p increases (say $p = 24$, 48, 60), the PMSE and MSE for each method increase apparently; (2) as p gets larger, the sensitivity decreases, and the specificity slightly decreases as well for SPLS methods, while increase slightly for ENet methods; (3) in general, the performance of MI-WENet is as good as the Full-ENet.

To examine whether a large number of simulation runs impacts the simulation results, we carried out the simulations of the same scenarios as presented in Table 3 but with 500 simulation runs. The results are presented in Table 4, from which we can see the results with 500 runs are very similar to those with 100 simulation runs (see Table 3).

Based on all simulation results, we conclude that the MI-WENet method obtains more or less the lowest PMSE and MSE among all the imputation-based methods. The sensitivity and specificity of the MI-ENet method is better than all other ENet-based imputation methods. In some cases although it looses in terms of sensitivity to some of the SPLS-based imputation methods its loss in sensitivity is not as severe as the loss of specificity in some of the SPLS-based imputations. Moreover, in most of our simulation scenarios, the PMSE, MSE, sensitivity and specificity from MI-WENet are closest to those from ENet on fully

observed data. MI-WENet is therefore recommended for variable selection and prediction when missing data exist.

In the following section, we applied the MI-WENet method to examine which variables were associated with the median effective dose and maximum effect in an ex-vivo phenylephrineinduced extension and acetylcholine-induced relaxation experiment.

4. Case study

The high-fat diet and normal chow fed mouse model has been used to examine the mechanisms by which high-fat diet impacts cardiovascular function. Early on, high-fat diet feeding induces endothelium infiammation, insulin resistance and endothelium dysfunction, which precedes the onset of diabetes.[26] Thus, endothelium dysfunction, characterized by decreased nitric oxide (NO) production or bioavailability, is used as a robust and early indicator of cardiovascular injury.[27] In the mouse model, mice were randomly assigned to high-fat diet and normal chow groups. The mice were fed for 12 weeks. Their body weight (BW), organ weight, blood variables and an array of plasma compositions and the ex-vivo endothelial functional outcomes were measured. Organ weights included heart, liver, kidney and spleen weight. The blood variables included percentage of red blood counts (%RBC, i.e. hematocrit) and percentage of white blood counts (%buffy). The plasma parameters included the counts of cholesterol , triglyceride, albumin, total protein (TP), high density lipoprotein (HDL), low density lipoprotein (LDL), alanine aminotransferase, aspartate aminotransferase, creatine kinase, alkaline phosphatase, creatinine, haemoglobin A1c (HbA1c), insulin, and nitrogen oxide species (NO_x , i.e. the sum of nitrite (NO_2) and nitrate (NO_3)), the ratio of HDL to LDL, and the percentage of albumin to total protein (Alb/TP). Isolated aorta were contracted with phenylephrine and relaxed with acetylcholine as previously published.[28] Percentage relaxation based on maximal contraction was calculated for each aorta. The percentage of maximal relaxation is called the Emax, and the acetylcholine concentration needed to achieve 50% relaxation is called the effective concentration producing 50% response, i.e. EC50. Emax and EC50 are two important parameters used to quantify endothelial function. In this section, we examined whether the two measurements of endothelial function, Emax and EC50, were related to any of the blood variables, plasma parameters, organ and body weights of the mice.

The final data set included 22 mice and 28 measured predictor variables. Some values in the predictor variables were missing due to inadequate volume of plasma. In total, eight mice had missing observations. In order to include the eight mice in the analysis, we applied the MIWENet method to examine what variables were closely associated with the measurements of endothelial function. To apply the MI-WENet method, we imputed five realizations for each missing value, and stacked the five imputed data sets into one large data set. Each variable was scaled to have unit variance before MI, and there was no additional standardization carried out after imputation. Thus, the subjects without missing values remained the same in the stacked data set. The log-transformation for EC50 was applied to ensure the normality of residuals. We applied the MI-WENet method to the stacked multiple imputed data set to obtain the coefficient estimates and select the important predictor variables. In addition, we applied leave-one-out cross validated samples to

construct 95% confidence intervals (95% CIs) for the estimated coefficients. The predictor variables whose 95% CIs did not contain zero were selected as the important variables for predicting the measurements of endothelial function. The estimates for the selected important predictors and their 95% CIs are given in Table 5.

The selected important predictors for Emax were NO_x and the ratio of kidney to body weight. The selected important predictors for the log-transformed EC50 were NO_x , kidney weight, the ratio of kidney to body weight, spleen weight, the ratio of spleen to body weight, the ratio of heart to body weight, TP, Alb/TP, HDL, and LDL. Endothelium dysfunction is commonly associated with decreased nitric oxide production and/or bioavailability.[29–31] The current results show that the decreased NO_x is associated with decreased Emax and increased EC50, which is consistent with previous findings. The other findings, such as association between endothelium dysfunction and kidney/BW, are also interesting and may be investigated further. The selected important predictor NO_x for Emax and EC50 demonstrates the selection precision of our proposed model, and thus, re-emphasizes the importance of using these endpoints to highlight the fundamental role of the endothelium in diet-induced cardiovascular injury.

5. Discussion and conclusions

Missing data are common in animal experiments and clinical studies. In this project we concentrated on the cases with missing covariate values. One of the frequently used methods in practice is the complete case analysis which ignores the covariates with missing observations. This method is easy to carry out, while it is inefficient and sometimes incorrect because the missing observations may not be a random subset of the whole sample. In this paper, we proposed an MI-WENet method (MI-WEnet) for variable selection and prediction. Our simulation studies demonstrated that the proposed MI-WENet method was able to identify important predictor variables with similar precisions as the SPLS and ENet methods would have achieved if the data were completely observed. Sensitivity and specificity obtained by the MI-WENet method were close to the results from the ENet method based on the full data in all the tested simulation scenarios. In addition, the MI-WENet method had the lowest MSE and PMSE among almost all methods we have evaluated. Our simulations also showed that the use of SPLS and ENet on complete cases only resulted in models with poor sensitivity and much larger PMSE and MSE than MI-WENet, especially when proportion of missing data is high and the missing patterns are MAR. This again indicates that the use of MI-WENet is especially recommended when proportion of missing values of the covariates is moderate to high.

MI-WENet maintained a balanced sensitivity and specificity in all the simulation scenarios and all the imputation schemes. The MI-WENet is also easy to implement. By applying the cyclical coordinate descent algorithm,[17] the coefficients of MI-WENet can be easily estimated by iteratively minimizing the weighted penalized least squares. The computational cost is mainly affected by the number of predictor variables not the sample size. R code for implementing the MIWENet method can be obtained upon request. At last, it should be pointed out that the weights we proposed account for the available information in an

observation; how to account for the available information more accurately is challenging and is beyond the scope of the current work.

Acknowledgments

Funding

This work was partially supported by NIH/NHLBI [grant number U24HL094373] for YW and MK, the funding of diabetes and obesity center provided by NIGMS [grant number GM103492] for DJC and MK, NIH [grant number HL89380] for DJC, and NIH/NCI [grant number CA170091-01A1] for SD.

References

[1]. Rubin, DB. Multiple imputation for nonresponse in surveys. Wiley; New York: 1987.

[2]. Little, RJ.; Rubin, DB. Statistical analysis with missing data. Vol. 539. Wiley; New York: 1987.

[3]. Little, RJ.; Rubin, DB. Statistical analysis with missing data. 2nd. Wiley; New York: 2002.

[4]. Van Buuren, S. Flexible imputation of missing data. Chapman and Hall/CRC Press; 2012. Print ISBN: 978-1-4398-6824-9; eBook ISBN: 978-1-4398-6825-6

[5]. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handledfi A review of published randomized controlled trials in major medical journals. Clin Trials. 2004; 1:368–376. [PubMed: 16279275]

[6]. Wood AM, White IR, Hillsdon M, Carpenter J. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. Int J Epidemiol. 2005; 34:89–99. [PubMed: 15333619]

[7]. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999; 18:681–694. [PubMed: 10204197]

[8]. Cohen, J.; Cohen, P.; Stephen, G.; Leona, S. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd. Lawrence Erlbaum Associates; Mahwah, NJ: 2003.

[9]. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed datafi. Stat Med. 2008; 27:3227–3246. [PubMed: 18203127]

[10]. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res. 2007; 16:259–275. [PubMed: 17621471]

[11]. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc. Ser B (Methodological). 1996; 58:267–288.

[12]. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc: Ser (Stat Method). 2005; 67:301–320.

[13]. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc: Ser B (Stat Method). 2010; 72:3–25.

[14]. Hastie, T.; Tibshirani, R.; Friedman, JH. 2nd. Springer; New York: 2009. The elements of statistical learning: data mining, inference, and prediction.

[15]. Heymans MW, Buurenvan S, Knol DL, Mechelenvan W, Vetde HC. Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Method. 2007; 7:33–42.

[16]. Chen Q, Wang S. Variable selection for multiply-imputed data with application to dioxin exposure study. Stat Med. 2013

[17]. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Soft. 2010; 33:1–22.

[18]. Wold, H. Partial least squares. In: Kots, S.; Johnson, NL., editors. Encyclopedia of statistical sciences. Vol. 6. Wiley; New York: 1985. p. 581-591.

[19]. Datta S. Exploring relationships in gene expressions: a partial least squares approach. Gene Expression. 2001; 9:249–255. [PubMed: 11763996]

[20]. Pihur V, Datta S, Datta S. Reconstruction of genetic association networks from microarray data: a partial least squares approach. Bioinformatics. 2008; 24:561–568. [PubMed: 18204062]

[21]. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12:55–67.

[22]. Schomaker M, Heumann C. Model selection and model averaging after multiple imputation. Comput Stat Data Anal. 2013

[23]. Groothuis-Oudshoorn K. Van Buuren S. mice: multivariate imputation by chained equations in R. J Stat Softw. 2011; 45:1–67.

[24]. Su YS, Yajima M, Gelman AE, Hill J. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. J Stat Softw. 2011; 45:1–31.

[25]. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Method. 2001; 27:85–96.

[26]. Kim F, Pham M, Maloney E, Rizzo NO, Morton GJ, Wisse BE, Kirk EA, Chait A, Schwartz MW. Vascular inflammation, insulin resistance, and reduced nitric oxide production precede the onset of peripheral insulin resistance. Arteriosclerosis, Thrombosis Vascular Biol. 2008; 28:1982–1988.

[27]. Rizzo NO, Maloney E, Pham M, Luttrell I, Wessells H, Tateya S, Daum G, Handa P, Schwartz MW, Kim F. Reduced NO-cGMP signaling contributes to vascular infiammation and insulin resistance induced by high-fat feeding. Arteriosclerosis, Thrombosis, Vascular Biol. 2010; 30:758–765.

[28]. Conklin D, Haberzettl P, Prough R, Bhatnagar A. Glutathione-S-transferase P protects against endothelial dysfunction induced by exposure to tobacco smoke. Am J Physiol Heart Circulatory Physiol. 2009; 296:1586–1597.

[29]. Hadi HA, Carr CS, Al Suwaidi J. Endothelial dysfunction: cardiovascular risk factors, therapy, and outcome. Vascular Health Risk Manag. 2005; 1:183–198.

[30]. Davignon J, Ganz P. Role of endothelial dysfunction in atherosclerosis. Circulation. 2004; 109:III–27–III-32.

[31]. Versari D, Daghini E, Virdis A, Ghiadoni L, Taddei S. Endothelial dysfunction as a target for prevention of cardiovascular disease. Diabetes Care. 2009; 32:S314–S321. [PubMed: 19875572]

Table 1

Simulation results for MCAR scenarios with different missing patterns (independent and monotone) and different correlation structures (compound symmetry with low correlation, compound symmetry with medium correlation and first-order autoregressive (AR(1))) for $p = 12$ and $n = 50$.

	Independent MCAR				Monotone MCAR			
	PMSE	**MSE**	**SENS**	**SPEC**	**PMSE**	**MSE**	**SENS**	**SPEC**
	Low correlation: corr($X_{i,j}$, $X_{i,j'}$) = 0.1							
Full-SPLS	14.72	3.52	97.2	67.0	15.05	3.96	95.7	70.2
CC-SPLS	28.84	15.34	80.7	47.7	22.08	10.23	84.5	52.0
SI-SPLS	17.72	6.20	91.7	60.3	18.64	7.21	86.8	70.0
MI-SPLS	16.10	4.68	98.3	36.5	16.59	5.24	98.7	42.5
Full-ENet	13.98	2.83	97.7	74.0	14.28	3.07	97.2	74.7
CC-ENet	30.27	18.51	52.5	87.5	21.99	10.47	68.8	89.3
SI-ENet	15.85	4.61	94.5	71.5	17.18	5.96	89.0	75.3
MI-ENet	15.43	4.22	96.3	71.7	16.38	5.16	94.8	72.2
SI-WENet	15.70	4.45	94.5	74.0	16.56	5.29	91.8	75.7
MI-WENet	15.34	4.14	96.0	74.8	15.90	4.66	94.0	72.8
	Medium correlation: corr($X_{i,j}$, $X_{i,j'}$) = 0.5							
Full-SPLS	32.86	7.60	90.8	41.0	33.30	7.87	86.5	44.3
CC-SPLS	44.56	17.55	81.5	34.5	37.40	11.01	86.8	31.2
SI-SPLS	34.06	8.76	88.5	36.5	34.76	9.25	86.8	40.0
MI-SPLS	33.98	8.28	93.8	23.2	34.28	8.73	91.5	32.2
Full-ENet	30.63	5.67	87.3	71.7	30.99	5.84	86.0	71.5
CC-ENet	58.10	33.02	33.8	84.0	50.03	24.86	42.7	86.0
SI-ENet	32.45	7.42	82.5	66.2	33.18	8.10	77.0	69.0
MI-ENet	31.72	6.67	84.8	68.2	31.99	6.93	81.8	69.2
SI-WENet	32.00	6.95	83.5	69.7	32.73	7.60	79.3	70.8
MI-WENet	31.35	6.35	86.7	68.2	32.01	6.87	81.8	70.3
	AR(1) correlation: corr($X_{i,j}$,$X_{i,j'}$) = $0.8^{\lvert j-j' \rvert}$							
Full-SPLS	27.77	5.91	87.7	28.3	27.93	6.00	88.3	30.2
CC-SPLS	38.42	15.22	85.8	23.7	31.88	9.41	89.3	22.3
SI-SPLS	28.01	5.95	91.0	23.5	29.40	6.94	85.8	35.2
MI-SPLS	28.26	6.04	94.3	17.3	29.71	7.56	89.7	25.3
Full-ENet	27.14	5.39	78.5	65.8	26.94	4.99	80.0	66.7
CC-ENet	48.79	26.79	36.8	83.0	43.19	21.26	42.7	83.3
SI-ENet	27.33	5.61	76.8	65.7	28.54	6.64	71.3	65.3
MI-ENet	27.07	5.35	78.3	63.7	27.76	5.80	74.8	66.2
SI-WENet	27.20	5.50	77.7	63.8	28.03	6.07	74.8	64.2
MI-WENet	27.40	5.65	77.3	62.8	27.53	5.56	76.8	64.2

Table 2

Simulation results for MAR scenarios with different missing patterns (independent and monotone) and different correlation structures (compound symmetry with low correlation, compound symmetry with medium correlation and first-order autoregressive (AR(1))) for $p = 12$ and $n = 50$.

	Independent MAR				Monotone MAR			
	PMSE	**MSE**	**SENS**	**SPEC**	**PMSE**	**MSE**	**SENS**	**SPEC**
	Low correlation: corr($X_{i,j}$,$X_{i,j'}$) = 0.1							
Full-SPLS	14.45	3.33	94.8	75.7	14.66	3.54	95.3	70.5
CC-SPLS	27.27	12.69	78.7	59.2	33.38	14.95	72.5	56.3
SI-SPLS	18.03	6.61	89.3	68.0	19.66	8.19	84.0	66.0
MI-SPLS	16.52	4.93	98.0	38.7	17.18	5.72	95.0	44.3
Full-ENet	13.83	2.71	96.8	78.8	13.93	2.88	97.2	75.2
CC-ENet	26.83	12.82	64.5	92.0	31.14	15.20	58.8	91.5
SI-ENet	16.90	5.73	90.7	77.2	18.25	7.07	85.2	72.8
MI-ENet	16.38	5.07	92.8	74.7	16.88	5.70	90.0	75.5
SI-WENet	16.57	5.35	91.3	78.3	17.20	5.96	88.7	73.5
MI-WENet	15.96	4.74	93.5	76.5	16.41	5.23	89.8	73.7
	Medium correlation: corr($X_{i,j}$,$X_{i,j'}$) = 0.5							
Full-SPLS	32.28	7.24	90.5	34.5	32.77	7.67	89.0	38.0
CC-SPLS	40.92	16.02	81.3	40.2	49.39	19.42	79.8	38.2
SI-SPLS	35.15	9.57	80.7	45.0	35.49	9.47	81.3	41.5
MI-SPLS	34.69	8.83	91.5	30.8	34.88	8.69	90.7	34.3
Full-ENet	31.27	6.27	83.3	72.8	30.76	6.01	84.2	73.2
CC-ENet	53.89	27.42	47.5	91.3	58.18	28.58	43.8	88.8
SI-ENet	33.21	8.31	77.3	70.5	33.31	8.17	76.2	69.3
MI-ENet	32.72	7.79	78.2	69.0	32.22	7.20	77.5	68.2
SI-WENet	33.47	8.57	76.8	71.3	33.00	8.06	75.8	71.7
MI-WENet	32.63	7.66	78.8	69.0	31.94	6.99	78.2	67.8
	AR(1) correlation: corr($X_{i,j}$,$X_{i,j'}$) = $0.8^{\lvert j-j' \rvert}$							
Full-SPLS	27.36	5.66	89.7	30.2	27.45	5.82	87.5	31.0
CC-SPLS	33.11	11.68	81.0	36.3	43.08	15.44	75.8	35.8
SI-SPLS	30.46	8.28	81.3	40.3	30.33	7.31	81.3	42.0
MI-SPLS	30.96	8.46	91.0	23.5	30.02	7.15	89.2	27.3
Full-ENet	26.91	5.19	78.5	66.8	26.59	5.04	78.5	70.5
CC-ENet	45.50	22.77	41.8	84.5	53.26	28.26	38.0	87.2
SI-ENet	30.21	8.30	67.8	70.5	28.28	6.68	69.3	72.2
MI-ENet	28.76	6.93	72.8	69.2	27.76	6.05	72.7	70.5
SI-WENet	28.81	7.01	70.3	67.5	27.95	6.37	69.0	72.0
MI-WENet	27.97	6.34	73.8	68.8	27.22	5.63	73.5	69.8

Table 3

Simulation results for scenarios with different combinations of p and n under monotone MAR and AR(1) correlation structure specifications based on 100 simulation runs.

	PMSE	MSE	SENS	SPEC	PMSE	MSE	SENS	SPEC
	$p = 12, n = 50$				$p = 24, n = 50$			
Full-SPLS	27.45	5.82	87.5	31.0	24.25	7.33	87.7	54.4
CC-SPLS	43.08	15.44	75.8	35.8	43.97	18.29	78.2	47.2
SI-SPLS	30.33	7.31	81.3	42.0	26.40	9.15	80.8	59.6
MI-SPLS	30.02	7.15	89.2	27.3	29.33	11.66	91.0	26.8
Full-ENet	26.59	5.04	78.5	70.5	22.81	5.89	82.5	82.7
CC-ENet	53.26	28.26	38.0	87.2	47.90	27.29	41.7	93.4
SI-ENet	28.28	6.68	69.3	72.2	24.17	7.24	75.0	83.2
MI-ENet	27.76	6.05	72.7	70.5	23.15	6.19	78.3	82.7
SI-WENet	27.95	6.37	69.0	72.0	23.78	6.85	78.3	82.7
MI-WENet	27.22	5.63	73.5	69.8	23.03	6.08	79.3	84.3
	$p = 12, n = 100$				$p = 48, n = 50$			
Full-SPLS	24.89	2.96	90.7	35.0	64.14	26.81	84.6	38.7
CC-SPLS	40.66	13.77	80.0	40.5	125.58	63.08	68.1	39.9
SI-SPLS	27.12	4.67	82.8	47.7	71.94	35.72	75.8	44.2
MI-SPLS	26.46	4.13	91.8	33.3	122.30	81.38	87.8	24.7
Full-ENet	24.35	2.44	89.5	68.2	62.51	26.75	55.6	87.1
CC-ENet	46.01	19.01	53.0	86.8	139.63	95.08	16.2	95.6
SI-ENet	26.13	4.06	79.5	70.7	79.43	43.50	45.1	86.8
MI-ENET	25.98	3.85	81.2	70.0	72.33	36.19	52.2	83.0
SI-WENet	25.61	3.71	80.5	69.8	77.47	41.56	47.2	86.9
MI-WENet	25.34	3.41	83.0	72.2	69.98	34.01	53.5	84.4
	$p = 12, n = 200$				$p = 60, n = 50$			
Full-SPLS	23.26	1.53	97.2	35.3	86.32	39.61	85.1	32.0
CC-SPLS	33.61	9.76	87.2	34.8	172.15	89.69	65.5	42.1
SI-SPLS	24.92	3.07	90.0	54.5	95.03	48.63	79.5	38.5
MI-SPLS	24.69	2.83	95.8	39.2	276.42	218.87	91.8	11.6
Full-ENet	22.95	1.20	97.3	71.2	91.66	46.41	48.9	86.7
CC-ENet	38.12	11.94	68.0	83.8	183.00	126.53	14.5	95.4
SI-ENet	24.63	2.76	89.5	75.2	105.60	60.05	40.9	87.2
MI-ENet	24.36	2.48	92.2	71.2	97.83	52.21	46.9	86.8
SI-WENet	23.84	2.06	92.0	72.5	103.51	58.41	41.7	87.9
MI-WENet	23.72	1.97	93.0	69.7	94.69	49.27	47.9	86.6

Table 4

Simulation results for scenarios with different combinations of p and n under monotone MAR and AR(1) correlation structure specifications based on 500 simulation runs.

	PMSE	MSE	SENS	SPEC	PMSE	MSE	SENS	SPEC
	p = 12, n = 50				p = 24, n = 50			
Full-SPLS	27.90	6.11	88.1	30.4	24.36	7.16	88.0	52.3
CC-SPLS	44.22	16.62	77.0	38.0	44.36	18.91	77.0	47.1
SI-SPLS	30.82	8.16	79.2	45.6	26.67	8.97	81.2	60.8
MI-SPLS	30.67	8.10	87.4	29.8	29.30	11.45	92.5	27.5
Full-ENet	27.11	5.34	78.2	68.6	22.95	5.93	80.9	82.7
CC-ENet	53.80	28.20	38.1	85.6	47.73	27.26	40.1	93.0
SI-ENet	29.49	7.63	68.4	73.5	24.56	7.53	74.2	82.2
MI-ENet	28.62	6.70	70.7	73.7	23.36	6.34	77.9	82.8
SI-WENet	28.85	7.04	69.8	72.5	24.09	7.03	76.6	82.3
MI-WENet	27.97	6.16	73.0	72.4	23.25	6.20	78.1	83.1
	p =12, n = 100				p = 48, n = 50			
Full-SPLS	24.80	3.16	92.8	32.7	64.41	27.52	84.9	40.0
CC-SPLS	37.82	12.56	82.2	38.2	117.09	56.28	74.1	36.0
SI-SPLS	26.90	4.79	83.8	48.6	75.06	38.56	75.6	45.2
MI-SPLS	26.34	4.35	91.8	33.0	116.28	76.58	85.4	25.8
Full-ENet	24.18	2.52	90.3	67.2	64.41	28.48	57.2	85.2
CC-ENet	44.37	18.25	54.4	85.2	137.51	93.63	17.0	95.4
SI-ENet	26.15	4.38	79.4	70.2	78.09	42.30	45.5	84.7
MI-ENET	25.63	3.79	82.7	70.5	72.12	36.19	52.5	83.5
SI-WENet	25.21	3.53	83.3	68.9	76.92	40.97	47.1	85.1
MI-WENet	24.88	3.20	85.6	69.1	69.80	33.94	54.7	83.6
	p = 12, n = 200				p = 60, n = 50			
Full-SPLS	23.27	1.57	96.2	36.7	84.43	37.53	83.6	34.8
CC-SPLS	33.71	9.82	87.8	33.9	164.11	82.84	69.4	39.1
SI-SPLS	24.86	2.98	88.7	55.2	96.66	49.80	76.1	40.4
MI-SPLS	24.55	2.70	96.0	36.9	285.05	225.15	91.0	15.6
Full-ENet	23.02	1.27	97.2	70.0	88.60	43.01	50.6	86.3
CC-ENet	36.95	10.69	70.1	84.8	183.30	128.18	13.8	95.7
SI-ENet	24.49	2.64	88.2	71.1	104.67	59.05	40.6	87.2
MI-ENet	24.26	2.37	92.6	70.4	99.70	54.00	45.3	86.1
SI-WENet	23.76	2.00	91.8	71.5	103.13	57.48	41.4	87.2
MI-WENet	23.60	1.85	94.2	70.2	97.24	51.57	46.3	86.4

Table 5

The estimated coefficients and their 95% CIs based on leave-one-out samples for Emax and EC50 using the MI-WENet method.

Covariate	Estimate	95% CI-low	X95% CI-up
	Emax		
NO_x	0.1894	0.0792	0.2997
Kidney/BW	0.2664	0.0128	0.5200
	EC50		
NO_x	−1.0803	−1.6983	−0.4623
Kidney	−1.2246	−1.9112	−0.5379
Kidney/BW	−1.7004	−2.6188	−0.7821
Spleen	−1.5503	−2.3522	−0.7484
Spleen/BW	−1.9629	−2.5398	−1.3860
Heart/BW	−0.9037	−1.5152	−0.2923
TP	−1.0097	−1.4712	−0.5481
Alb/TP	−0.9950	−1.4623	−0.5276
HDL	−1.1533	−1.6846	−0.6220
LDL	−1.0116	−1.4736	−0.5497