Research Article

Statistics in Medicine

Received 28 February 2012, Accepted 22 February 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5783

Variable selection for multiply-imputed data with application to dioxin exposure study

Qixuan Chen[a*†‡] **and Sijian Wang**[b‡]

Multiple imputation (MI) is a commonly used technique for handling missing data in large-scale medical and public health studies. However, variable selection on multiply-imputed data remains an important and long-standing statistical problem. If a variable selection method is applied to each imputed dataset separately, it may select different variables for different imputed datasets, which makes it difficult to interpret the final model or draw scientific conclusions. In this paper, we propose a novel multiple imputation-least absolute shrinkage and selection operator (MI-LASSO) variable selection method as an extension of the least absolute shrinkage and selection operator (LASSO) method to multiply-imputed data. The MI-LASSO method treats the estimated regression coefficients of the same variable across all imputed datasets as a group and applies the group LASSO penalty to yield a consistent variable selection across multiple-imputed datasets. We use a simulation study to demonstrate the advantage of the MI-LASSO method compared with the alternatives. We also apply the MI-LASSO method to the University of Michigan Dioxin Exposure Study to identify important circumstances and exposure factors that are associated with human serum dioxin concentration in Midland, Michigan. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: group LASSO penalty; multiple imputation; regularization; Rubin's rules; variable selection

1. Introduction

Dioxins are a class of chemical contaminants that are highly toxic and persist in the environment. Studies of highly exposed populations show that dioxins can cause excessive cases of cancer, diabetes, immune system suppression, and skin problems [1–5]. Because dioxins are very difficult to resolve once they enter the human body, it is important to investigate the site-specific exposure factors and pathways by which people accumulate body burdens of dioxins. This information can aid the development of regulations to reduce the risk of exposure.

The work of this article is motivated by analyzing the University of Michigan Dioxin Exposure Study (UMDES) data. The UMDES is the first population-based dioxin exposure study of residents living in the Midland, Michigan, area, one of the largest and best characterized dioxin contamination sites in North America [6–8]. In the UMDES, adults older than 18 years who had lived in their current residence for at least 5 years were eligible to participate. Eligible subjects were selected from the populations of five counties in Michigan, and were invited to complete an interview, donate an 80-mL whole blood sample, and have their household dust and soil sample collected. The study generated an unprecedented dataset with an extensive collection of information on demographics, health, residence, activities, work history, lifetime food consumption, current diet, and measurements of dioxins in participants' household dust, soil, and serum.

As often encountered in large epidemiological studies, the statistical analysis of the UMDES was hindered by missing data, which were caused by item nonresponse in survey questionnaire, and missing dust

[a] *Department of Biostatistics, Columbia University Mailman School of Public Health, New York, NY, U.S.A.*
[b] *Department of Biostatistics & Medical Informatics, Department of Statistics, University of Wisconsin–Madison, Madison, WI, U.S.A.*
**Correspondence to: Qixuan Chen, Department of Biostatistics, Columbia University Mailman School of Public Health, New York, NY, U.S.A.*
[†] *E-mail: qc2138@columbia.edu*
[‡] *Q. Chen and S. Wang contributed equally to this work.*

and soil values from subjects who refused to provide samples. Although the nonresponse rate is small for each individual variable, the missing values are dispersed throughout the data in a haphazard pattern. Thus, ignoring missing data by deleting incomplete cases is wasteful of the costly collected data and can lead to biased statistical inferences. Alternatively, the multiple imputation (MI) framework suggested by Rubin [9] can be used to address the problem of missing data. MI refers to a procedure of replacing each missing value with a plausible value multiple times to generate multiple complete datasets. MI has a practical advantage of allowing standard complete-data methods of analysis to be used. Thus, the complete-data inferences from each imputed dataset can be combined to form one inference that properly reflects the uncertainty of imputation due to the missing data. Compared with maximum-likelihood estimates calculated directly from incomplete data, MI is more attractive for data with general missing patterns and for multiple-purpose analysis with many estimands like the UMDES data. Unless the rate of missing information is high, in most situations only 3 to 10 imputations are needed to yield an excellent efficiency [9]. For the UMDES data, Olson *et al*. [10] used a sequential regression imputation procedure [11] with five imputations to generate values for missing items. Our analysis is based on these five imputed datasets. Our goal is to identify important exposure factors that are associated with human population's serum dioxin concentration in Midland, which is a variable selection problem.

Variable selection has been extensively studied in statistical literature. Some classical methods including stepwise selection or best subset selection search for the best model based either on significance tests or on a certain information-based criterion, such as Akaike information criterion [12] and Bayesian information criterion (BIC) [13]. Penalized likelihood-based methods have drawn a lot of attention in recent literature, including the least absolute shrinkage and selection operator (LASSO) [14], smoothly clipped absolute deviation [15], smooth integration of counting and absolute deviation [16], minimax concave penalty [17], truncated L1 penalty [18], group LASSO [19], composite absolute penalties [20], and grouping pursuit [21], among others. Bayesian variable selection strategies [22–24] have also become popular in many applications. However, these variable selection methods may have inadequate performance if they are directly applied to multiply-imputed data, because the selection may not be consistent across the multiple datasets generated by imputation. Specifically, if a variable selection method is applied to each imputed dataset separately, it will probably identify different important variables in each imputed dataset. This makes it difficult to produce the overall parameter estimates across all imputed datasets and hence makes it difficult to interpret the model or draw scientific conclusions.

When MI is used to handle missing data, some naïve variable selection strategies, such as conducting variable selection with complete cases only, that is, discarding cases with any missing values, or with a single imputed dataset, are often used. These naïve strategies are easy to implement, but they ignore either possible differences between the complete cases and incomplete cases or uncertainty of imputation caused by missing information. To incorporate the missing-data uncertainty, we can conduct variable selection in each imputed dataset separately and claim a variable to be important if it is selected with a frequency greater than some subjective threshold (say 60%), which is one popular approach [25] in epidemiological literature. Obviously, this approach may identify different important variables with different thresholds, but there is no clear guide on how to choose a proper threshold. Two studies in statistical literature have addressed the variable selection problem for multiply-imputed data. Yang, Belin, and Boscardin [26] proposed a Bayesian variable selection method for multiply-imputed data, by applying Rubin's rules (RR) [9,27] to synthesize different sets of Markov chain Monte Carlo estimates into a final summary. Wood, White, and Royston [28] described a stepwise variable selection method by repeated use of RR. They also proposed a 'stacked' method that combines the multiply-imputed datasets and uses a weighting scheme to account for the fraction of missing data in each covariate. They then concluded that the RR approach was recommended for performing stepwise selection in multiply-imputed data as it is the only approach to preserve the type I error.

In this paper, we extend the LASSO method to multiply-imputed data by treating the estimated regression coefficients associated with the same variable across different imputed datasets as a group and by selecting or removing the whole group together. We call this method the multiple imputation-least absolute shrinkage and selection operator (MI-LASSO). The MI-LASSO method incorporates imputation uncertainty in the variable selection procedure and yields a consistent selection across multiple-imputed datasets. The rest of the paper is organized as follows. In Section 2, we review imputation strategies and RR for MI inference. We introduce the MI-LASSO method in Section 3. In Section 4, we show the performance of the MI-LASSO method in multiply-imputed data by using simulations. In Section 5, we apply the MI-LASSO method to the UMDES data to identify important factors that are associated with human serum dioxin concentration in Midland, Michigan. We conclude the paper with Section 6.

2. Multiple imputation

2.1. Imputation

Multiple imputation can be summarized into three steps: *imputation*, *analysis*, and *combination* [9, 29–33]. In the first step, missing data are imputed D times to generate D imputed complete datasets. Imputations are often created under Bayesian arguments by specifying a model for the variables with missing data given observed variables, prior distributions for parameters, and a model for the missing data mechanism if missing data are non-ignorable. Imputated values are then the draws from the posterior predictive distributions of the missing data conditional on the observed data. The imputation model should be chosen as general as possible to accommodate potential different data analysis purposes [34]. Although MI does not require the missing at random (MAR) assumption in the sense defined by Rubin [27, 35], most of current MI techniques assume that the missing data are MAR.

Two general multivariate techniques for multiply imputing missing data with general missing pattern are joint modeling and sequential regression imputation strategies. Schafer [36] used Markov chain Monte Carlo methods to jointly model multivariate continuous data using the multivariate normal distribution, multivariate categorical data using loglinear models, and mixed continuous and categorical data using the general location model. These joint modeling approaches were implemented in the S-PLUS/R packages 'norm', 'cat', and 'mix', respectively. The use of the joint modeling strategy can be challenging for large datasets with hundreds of variables of varying types [37]. In contrast, sequential regression imputation specifies separate conditional models for each variable given other variables. It is flexible in allowing variables in the data to have varying types, including but not limited to continuous, dichotomous, polytomous, and counts. It is capable of handling imputations with restrictions, such as survey skip pattern and logical or consistency bounds in questionnaires [11]. The sequential regression MI is also known as 'multivariate imputation by chained equations' [38], and was implemented in the R(R Development Core Team, Vienna, Austria) packages 'mice' [38] and 'mi' [39], and in the SAS(SAS Institute Inc., Cary, NC) callable software IVEware [11].

At each step of the sequential regression imputation, the conditional distribution of each variable is modeled using an appropriate regression model given the other variables. The imputed values are then the draws from the posterior predictive distribution of the missing data given the observed data. The imputed variables are then used in the imputation of the next variables until all variables with missing values are imputed. This process repeats until it converges. Often only a small number of iterations are needed to reach the convergence for problems with moderate fraction of missing data [38]. Thus, instead of modeling the multivariate data by a joint distribution, the sequential regression imputation models conditional distributions, which are much easier to accommodate. Specifically, linear regression models can be used for continuous variables, logistic regression models for dichotomous variables, generalized logit regression models for polytomous variables, and Poisson regression models for count variables. Despite the lack of theoretical justification, the sequential regression imputation algorithm works well in real application settings and has been widely used to handle missing data in large survey data [40, 41]. The sequential regression imputation was used by Olson *et al.* [10] to handle the missing data in the UMDES data and also was used in our simulation study in Section 4.

2.2. Rubin's rules for multiple imputation inference

The analysis of a multiply-imputed data is straightforward [9, 27]. Each dataset completed by imputation is first analyzed by a same complete-data method. RR are then used to obtain a combined estimate from the D estimates calculated from the D imputations. Specifically, let $\hat{\theta}_d$ and V_d, $d = 1, \ldots, D$, and be D estimates and their associated variances for a scalar parameter θ, calculated from the dth imputation. The combined estimate of θ is the average $\bar{\theta} = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$. The variance of $\bar{\theta}$ has two components: the average within-imputation variance $\bar{V} = \frac{1}{D}\sum_{d=1}^{D} V_d$, and the between-imputation variance $B = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \bar{\theta}\right)^2$. The combined variance associated with $\bar{\theta}$ is $T = \bar{V} + \frac{D+1}{D}B$. When sample size is large,

$$\left(\bar{\theta} - \theta\right) T^{-1/2} \sim t_\nu,$$

where the degrees of freedom $\nu = (D-1)\left(1 + \frac{1}{D+1}\frac{\bar{V}}{B}\right)^2$ [9, 42].

Notice that RR cannot be applied to obtain combined coefficient estimates if covariates of regression models in each imputation are different. Thus, it is desired to have a variable selection method that yields a consistent selection across all imputed datasets. After a model is selected, this same regression model can be fitted on each of D imputations, and RR can be applied to obtain combined coefficient estimates and variances.

3. Multiple imputation-least absolute shrinkage and selection operator method

Suppose that we have n subjects and p candidate covariates, $X_1, \ldots, X_p$. Let Y_i be the outcome variable and X_{ij} be the jth candidate covariate ($j = 1, \ldots, p$) for the ith subject. We assume that missing data can occur in either Y or X_j. The observation indicators for Y_i and X_{ij} are denoted by R_{iy} and R_{ij}, respectively. If $R_{iy} = 1$, Y_i is observed and, if $R_{iy} = 0$, Y_i is missing. Similarly, if $R_{ij} = 1$, X_{ij} is observed and, if $R_{ij} = 0$, X_{ij} is missing. MI is used to handle missing data in Y and X_j. We denote each of the D imputed datasets as $(y_{d,i}; x_{d,i1}, \ldots, x_{d,ip})_{i=1}^{n}$, $d = 1, \ldots, D$, where $y_{d,i}$ and $x_{d,ij}$ are the values of Y and X_j for the ith subject in the dth imputed dataset, respectively. If $R_{ij} = 1$, we have $x_{1,ij} = \cdots = x_{D,ij} = x_{ij}$, and if $R_{ij} = 0$, $x_{d,ij}$ can take different values in each imputation. We assume that the complete data are generated from the following linear regression model:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where β_j's are regression coefficients, and terms $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are error terms. We are interested in selecting important covariates X_j that predict Y.

Variable selection based on penalized likelihood estimation has become popular in recent statistical research. In particular, the LASSO method [14] has gained much attentions. It penalizes the L_1 norm of the regression coefficients to achieve a sparse model:

$$\min_{\beta_j} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2}$$

where λ is a non-negative tuning parameter that controls the amount of shrinkage. Because of the singularity at $\beta_j = 0$ of the L_1-norm penalty, some estimated $\hat{\beta}_j$ will be exactly zero, which realizes the variable selection. The theoretical properties of the LASSO method have been thoroughly studied in literature, say in [15, 43–45] and references therein.

When multiply-imputed data are present, if LASSO is applied to each imputed dataset separately, it is possible that it selects different variables in different imputed datasets. This inconsistency in the selection is due to fitting the model on each dataset *separately*, which motivates us to consider fitting models on all imputed datasets *jointly* to yield a consistent variable selection across all imputed datasets. Denote $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}$ be the D estimated coefficients for covariate X_j on the D imputed datasets. If X_j is unimportant, $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}$ should *all* be zero and if X_j is important, $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}$ should all be nonzero. Motivated by this desired consistency, we treat $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}$ as a group and apply the group LASSO penalty [19]. To be specific, we consider the following optimization approach:

$$\min_{\beta_{d,j}} \sum_{d=1}^{D} \sum_{i=1}^{n} \left(y_{d,i} - \beta_{d,0} - \sum_{j=1}^{p} \beta_{d,j} x_{d,ij} \right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{d=1}^{D} \beta_{d,j}^2}, \tag{3}$$

where $\sum_{j=1}^{p} \sqrt{\sum_{d=1}^{D} \beta_{d,j}^2}$ is called the group LASSO penalty. It was originally proposed by Yuan *et al*. [19] for linear regression when the covariates can be separated into several groups. Its theoretical property was studied in [46] and references therein. We will talk about how to tune λ in Section 4.

We call the resulting variable selection procedure the MI-LASSO. We can see that, because of the group LASSO penalty, $\hat{\beta}_{d,j}$, the estimated regression coefficient for X_j on the dth imputed dataset depends on all imputed datasets instead of the dth imputed dataset only. In other words, the proposed method fits D models on all imputed datasets jointly instead of fitting models *separately*. Furthermore,

because of the nature of the group LASSO penalty [19], the estimated coefficients $\left(\hat{\beta}_{1,j},\ldots,\hat{\beta}_{D,j}\right)$ for each covariate X_j will either be all exactly zero or be all nonzero. This property guarantees the desired consistency of variable selection across all imputed datasets.

It is not trivial to get the solution to the optimization problem (3), because the group LASSO penalty function is singular at the origin point. To overcome this optimization difficulty, we apply the local quadratic-approximation method as proposed by Fan and Li [15]. To be specific, we iteratively solve the optimization problem (3). Suppose we already have the estimates $\hat{\beta}_{d,j}^{(t)}$, $d=1,\ldots,D$, at the tth iteration. As long as $\sqrt{\sum_{d=1}^{D}\left(\hat{\beta}_{d,j}^{(t)}\right)^2}>0$, we have the following approximation:

$$\sqrt{\sum_{d=1}^{D}\beta_{d,j}^2}\approx\frac{\sum_{d=1}^{D}\beta_{d,j}^2}{\sqrt{\sum_{d=1}^{D}\left(\hat{\beta}_{d,j}^{(t)}\right)^2}}.$$

Correspondingly, the optimization problem (3) can be approximated by

$$\min_{\beta_{d,j}}\sum_{d=1}^{D}\left\{\sum_{i=1}^{n}\left(y_{d,i}-\beta_{d,0}-\sum_{j=1}^{p}\beta_{d,j}x_{d,ij}\right)^2+\lambda\sum_{j=1}^{p}c_j\beta_{d,j}^2\right\}, \tag{4}$$

where $c_j=1/\sqrt{\sum_{d=1}^{D}\left(\hat{\beta}_{d,j}^{(t)}\right)^2}$. We can see that in the optimization problem (4), the estimated regression coefficients $\hat{\beta}_{d,j}^{(t+1)}$'s can be obtained by solving D separate ridge regressions, which is easy to implement in practice. The iterations continue until the convergence is claimed. One possible limitation for this approximation is that once a group of coefficients are shrunken to zero, they will stay at zero. To avoid this inflexibility, we propose to fix $\hat{\beta}_{1,j}^{(t)}=\cdots=\hat{\beta}_{D,j}^{(t)}=\delta$ when $\sum_{d=1}^{D}\left(\hat{\beta}_{d,j}^{(t)}\right)^2\leqslant D\delta^2$. In our algorithm, we choose $\delta=10^{-10}$, and our numerical studies show that the algorithm performs well.

4. Simulation study

4.1. Design

In this section, we perform a simulation study to evaluate the finite sample performance of the MI-LASSO method and compare the results with four existing methods: (i) LASSO, the LASSO method applied to the full data before generating any missing data; (ii) stepwise, the stepwise selection method applied to the full data before generating any missing data; (iii) CC-LASSO, the LASSO method applied to the complete cases by discarding observations with any missing data; and (iv) RR-stepwise, the stepwise selection method by repeated use of RR applied to the multiply-imputed data proposed in [28]. To make our paper self contained, we present the details of RR-stepwise selection procedure in Appendix.

We considered three examples. In the first two examples, all covariates are continuous. In the third example, all covariates are dichotomous. In each example, the outcome variable Y in the complete data are generated from the following linear regression model:

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)^T$ and $\epsilon\overset{i.i.d.}{\sim}N(0,\sigma^2)$. We chose the σ value so that $\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}/\sigma^2=1$, where Σ is the population covariance matrix of $\boldsymbol{X}$.

In each example, we let Y be fully observed and generated missing data in $\boldsymbol{X}$. We considered ignorable missing data mechanisms: missing completely at random (MCAR) and MAR. The missing data were then multiply imputed using the 'mice' package in R [38], where the default predictive mean matching method was used to impute continuous covariates and the logistic regression imputation method was used for dichotomous covariates. In the imputation model for each incomplete candidate covariate, all the other candidate covariates and Y were used as predictors. Five imputations were completed for each incomplete data.

For stepwise and RR-stepwise, denote α_1 and α_2 to be the P-value thresholds for including and removing a variable, respectively. We considered two sets of thresholds in our simulations:

($\alpha_1 = 0.05, \alpha_2 = 0.06$) and ($\alpha_1 = 0.20, \alpha_2 = 0.21$). For the LASSO, CC-LASSO and MI-LASSO methods, the tuning parameter λ was selected by minimizing BIC.

For the LASSO method, BIC has the following formula:

$$\text{BIC} = \log\left(\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 / n\right) + df_1 * \log(n)/n,$$

where df_1 is the degrees of freedom of the fitted model. Zou and Hastie [47] proved that df_1 can be estimated by the number of nonzero fitted coefficients. For the MI-LASSO method, we define the corresponding BIC as

$$\text{BIC} = \log\left(\sum_{d=1}^{D}\sum_{i=1}^{n}\left(y_{d,i} - \hat{\beta}_{d,0} - \sum_{j=1}^{p}\hat{\beta}_{d,j} x_{d,ij}\right)^2 / (Dn)\right) + df_2 * \log(Dn)/(Dn). \quad (5)$$

Following Yuan and Lin [19], it can be shown that the degrees of freedom of our fitted model, df_2, can be estimated by

$$df_2 = \sum_{j=1}^{p} \boldsymbol{I}\left(\sqrt{\textstyle\sum_{d=1}^{D}\hat{\beta}_{d,j}^2} > 0\right) + \sum_{j=1}^{p}\frac{\sqrt{\sum_{d=1}^{D}\hat{\beta}_{d,j}^2}}{\sqrt{\sum_{d=1}^{D}\tilde{\beta}_{d,j}^2}}(D-1),$$

where $\hat{\beta}_{d,j}$ and $\tilde{\beta}_{d,j}$ are the MI-LASSO estimate and the ordinary least square estimate for the jth covariate on the dth dataset, respectively. The first term of df_2 is equal to the number of nonzero fitted coefficients. The details of derivations are provided in Appendix.

To compare the performance of variable selection among different methods, we considered three criteria: sensitivity of selection (SEN)

$$\text{SEN} = \frac{\text{\# of selected important variables}}{\text{\# of true important variables}},$$

specificity of selection (SPE)

$$\text{SPE} = \frac{\text{\# of removed unimportant variables}}{\text{\# of true unimportant variables}},$$

and mean-squared error (MSE) [14]

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

A higher value of sensitivity and specificity and a lower value of MSE are desirable.

4.2. *Simulation one: Continuous candidate covariates with compound symmetry covariance structure*

In this example, we have the sample size $n = 100$ and the number of candidate covariates $p = 20$. For the true regression coefficient β_j's, we let $\beta_j = 1$ for $j = 1, 2, 5, 11, 12, 15$ and $\beta_j = 0$ otherwise. The covariates $\boldsymbol{X}$ were generated from a multivariate normal distribution with zero mean, unit variance, and a compound symmetry correlation structure, that is, $cov(X_i, X_j) = \rho$, $\forall i \neq j$. The correlation ρ takes a value of 0.1 for low correlation and a value of 0.5 for high correlation among covariates. We generated missing data in X_{11}-X_{20}. For MCAR, we independently dropped 5% of data in $X_{11} - X_{20}$ to yield datasets with about 60% complete cases. For MAR, the missing data indicators R_{ij} were generated using the following logistic regression model:

$$\text{logit}\{\Pr(R_{ij} = 0 | X_{i(j-10)}, Y_i)\} = \alpha_0 + 0.5X_{i(j-10)} + 0.5Y_i,$$

where α_0 was chosen to yield around 60% complete cases.

Table I shows that the MI-LASSO method has similar or slightly higher sensitivity but lower specificity of selection and similar or slightly larger MSE compared with the LASSO method applied to

Statistics in Medicine

Table I. Mean sensitivity (SEN) and specificity (SPE) of selection and median mean-squared error (MSE) in simulation one: continuous candidate covariates with compound symmetry covariance structure ($\rho = 0.1$ or 0.5).

	$\rho = 0.1$			$\rho = 0.5$		
	SEN	SPE	MSE	SEN	SPE	MSE
Full data						
LASSO	95.2	82.9	1.6	75.5	78.5	2.7
stepwise ($\alpha_1 = 0.05$)	87.0	93.8	1.8	44.6	92.7	4.5
stepwise ($\alpha_1 = 0.20$)	95.3	78.4	2.0	61.7	78.7	4.9
MCAR						
MI-LASSO	96.0	82.1	1.5	79.9	74.0	2.6
CC-LASSO	76.5	84.4	3.6	59.9	81.0	4.6
RR-stepwise ($\alpha_1 = 0.05$)	83.9	94.6	1.9	42.8	93.3	4.7
RR-stepwise ($\alpha_1 = 0.20$)	94.3	79.3	2.1	59.2	78.6	5.2
MAR						
MI-LASSO	94.7	80.0	1.8	76.9	72.6	3.2
CC-LASSO	46.5	91.7	7.4	34.6	89.3	15.0
RR-stepwise ($\alpha_1 = 0.05$)	79.7	94.4	2.3	38.2	92.3	5.3
RR-stepwise ($\alpha_1 = 0.20$)	91.8	79.1	2.4	54.7	78.6	5.9

MCAR, missing completely at random; MAR, missing at random.

the full data before generating any missing data. This indicates that the use of MI-LASSO on multiply-imputed data can identify important covariates in a linear regression model as well as the LASSO method would have achieved if the data were complete. By deleting incomplete observations, CC-LASSO has lower sensitivity of selection and much larger MSE than MI-LASSO, especially when missing data are MAR. Compared with MI-LASSO, RR-stepwise has lower sensitivity but higher specificity when using 0.05 as the P-value to enter, and lower sensitivity and lower specificity when using 0.2 as the P-value to enter. Both RR-stepwise procedures yield models with larger MSE than MI-LASSO. Furthermore, as ρ increases from 0.1 to 0.5, the mean sensitivity drops to a half, and the MSE doubles in RR-stepwise, but the deterioration is less severe in the use of MI-LASSO. This poor performance of the stepwise strategies when multicollinearity is present is also evident in complete-data analysis using the full data.

4.3. Simulation two: Continuous candidate covariates with first-order autoregressive (AR(1)) covariance structure

In this example, we considered three n and p combinations: ($n = 100, p = 20$), ($n = 200, p = 20$), and ($n = 100, p = 40$). We generated $\boldsymbol{X}$ from a multivariate normal distribution with zero mean, unit variance, and an AR(1) correlation with $\rho = 0.5$. For the true regression coefficients β_j's, when $p = 20$, we let $\beta_j = 1$ for $j = 1, 2, 5, 11, 12, 15$ and $\beta_j = 0$ otherwise. When $p = 40$, we let $\beta_j = 1$ for $j = 1, 2, 5, 11, 12, 15, 21, 22, 25, 31, 32, 35$ and $\beta_j = 0$ otherwise. For missing data generation, when $p = 20$, we generated missing data in $X_{11} - X_{20}$. When $p = 40$, we generated missing data in $X_{11} - X_{20}$ and $X_{31} - X_{40}$. The procedure of missing data generation was the same as presented in simulation one. For the case with ($n = 100, p = 40$) and the case with ($n = 200, p = 20$), the yielded datasets have around 60% complete cases. For the case with ($n = 100, p = 20$), we considered both 60% complete cases and 35% complete cases.

Table II presents the simulation results. Overall MI-LASSO has similar sensitivity and specificity of selection and similar MSE compared with LASSO applied to the full data, and performs better than CC-LASSO and RR-stepwise. As p increases from 20 to 40, the sensitivity of selection decreases, and the MSE increases for all approaches. Increasing n from 100 to 200, both the sensitivity of selection and the MSE are largely improved, especially for the stepwise strategies. In the scenario of $p = 20$ and $n = 200$, the sensitivity of selection is close to perfect for all approaches, except CC-LASSO, and MI-LASSO and RR-stepwise ($\alpha_1 = 0.05$) yield the same MSE. This suggests that MI-LASSO has a greater advantage than RR-stepwise when number of candidate covariates relative to sample size is large. Finally, the sensitivity of selection decreases, and the MSE increases as the percentage of complete cases drops from

Table II. Mean sensitivity (SEN) and specificity (SPE) of selection and median mean-squared error (MSE) in simulation two: continuous candidate covariates with AR(1) covariance structure under varying sample sizes (n) and numbers of candidate covariates (p). Incomplete data contain 35% complete cases in the high-missing proportion (HM) scenario and 60% complete cases in the other scenarios.

	$p = 20, n = 100$			$p = 40, n = 100$			$p = 20, n = 200$			$p = 20, n = 100$ (HM)		
	SEN	SPE	MSE	SEN	SPE	MSE	SEN	SPE	MSE	SEN	SPE	MSE
Full data												
LASSO	92.6	82.3	1.5	58.9	88.6	8.4	99.8	84.0	0.6	92.6	82.3	1.5
stepwise ($\alpha_1 = 0.05$)	77.6	92.8	2.0	50.2	92.6	8.9	98.1	94.9	0.6	77.6	92.8	2.0
stepwise ($\alpha_1 = 0.20$)	89.5	79.8	2.0	66.3	78.9	9.5	99.2	80.9	0.8	89.5	79.8	2.0
MCAR												
MI-LASSO	93.6	82.2	1.5	64.8	86.9	7.7	99.9	82.1	0.6	93.2	80.0	1.6
CC-LASSO	75.4	84.1	3.2	36.5	89.2	14.4	96.3	81.9	1.2	58.6	76.8	8.2
RR-stepwise ($\alpha_1 = 0.05$)	74.6	93.5	2.1	48.1	92.9	9.2	97.1	95.5	0.6	72.4	93.2	2.3
RR-stepwise ($\alpha_1 = 0.20$)	88.5	80.8	2.0	65.3	78.8	9.8	99.2	81.2	0.8	86.2	80.6	2.2
MAR												
MI-LASSO	91.8	79.6	1.7	65.1	83.2	8.3	99.7	80.1	0.7	90.7	69.0	2.6
CC-LASSO	47.3	91.4	6.8	12.6	96.7	23.0	86.2	85.5	3.3	25.3	87.5	37.0
RR-stepwise ($\alpha_1 = 0.05$)	70.5	93.1	2.4	43.4	93.0	10.3	95.2	94.8	0.7	60.1	92.6	3.2
RR-stepwise ($\alpha_1 = 0.20$)	84.9	79.1	2.3	60.2	78.6	11.1	98.4	81.0	0.9	76.7	78.5	3.2

MCAR, missing completely at random; MAR, missing at random.

60% to 35% for the case with (n=100, p=20). Not surprising, the performance of CC-LASSO is most affected by the increase in the proportion of missing data when missing data are MAR, where the MSE increases from 6.8 to 37.0.

4.4. Simulation three: Dichotomous candidate covariates

In this example, we have $n = 100$ and $p = 20$. To generate dichotomous covariates, we first simulated $\boldsymbol{X}$ from a multivariate normal distribution with zero mean, unit variance, and an AR(1) correlation with $\rho = 0.5$. We then dichotomized the simulated continuous variables into 0/1 variables using 0 as the cut-off value. For the true regression coefficient β_j, we let $\beta_j = 1$ for $j = 1, 2, 5, 11, 12, 15$ and $\beta_j = 0$ otherwise. We generated missing data in $X_{11} - X_{20}$ to have around 60% complete cases. The missing data generation procedures are the same as described in simulation one.

Simulation results are shown in Table III, and the findings are similar to what we observed in the scenarios with continuous covariates.

5. Real data application

The data in this example come from the UMDES. Detailed description of the UMDES study design and a report of study findings using data of all study participants can be found elsewhere [6, 7]. In this application, we used the data of the 448 residents who resided in or near the flood plain of the Tittabawassee between the Dow Chemical Company plant in Midland and the confluence of the Tittabawassee and Shiawassee rivers in Saginaw. Our goal was to identify important circumstances and exposure pathways that predict serum concentration of the most toxic dioxin compound, 2,3,7,8-tetrachlorodibenzop-dioxin (TCDD). Candidate covariates we considered here include health and demographics (age, sex, body mass index (BMI), BMI change, pregnancy and breast-feeding history among women, smoking, education, income, and race/ethnicity), region of current residence, soil and household dust TCDD concentrations, recreational activities in the contaminated areas (fishing, hunting, and water activities), occupations (focused on those with potential dioxin exposure), diet (especially sport-caught fish, game meat, eggs, poultry, dairy, and vegetables raised in contaminated soil), and property-use factors (e.g., backyard burning). We converted categorical/nominal variables into separate dichotomous variables. The final data for analysis contain 63 continuous and 68 dichotomous candidate covariates, most of which are mildly correlated with a mean absolute correlation coefficient of 0.06 (95% CI: [0.00,0.20]). However, a few

Table III. Mean sensitivity (SEN) and specificity (SPE) of selection and median mean-squared error (MSE) in simulation three: dichotomous candidate covariates.

	SEN	SPE	MSE
Full data			
LASSO	95.0	82.1	0.4
stepwise ($\alpha_1 = 0.05$)	87.3	93.1	0.4
stepwise ($\alpha_1 = 0.20$)	95.4	78.4	0.5
MCAR			
MI-LASSO	96.2	82.0	0.4
CC-LASSO	74.6	83.6	1.1
RR-stepwise ($\alpha_1 = 0.05$)	85.3	93.9	0.5
RR-stepwise ($\alpha_1 = 0.20$)	95.2	79.1	0.5
MAR			
MI-LASSO	95.7	81.4	0.4
CC-LASSO	60.5	87.2	1.1
RR-stepwise ($\alpha_1 = 0.05$)	83.6	93.6	0.5
RR-stepwise ($\alpha_1 = 0.20$)	94.0	78.9	0.6

MCAR, missing completely at random; MAR, missing at random.

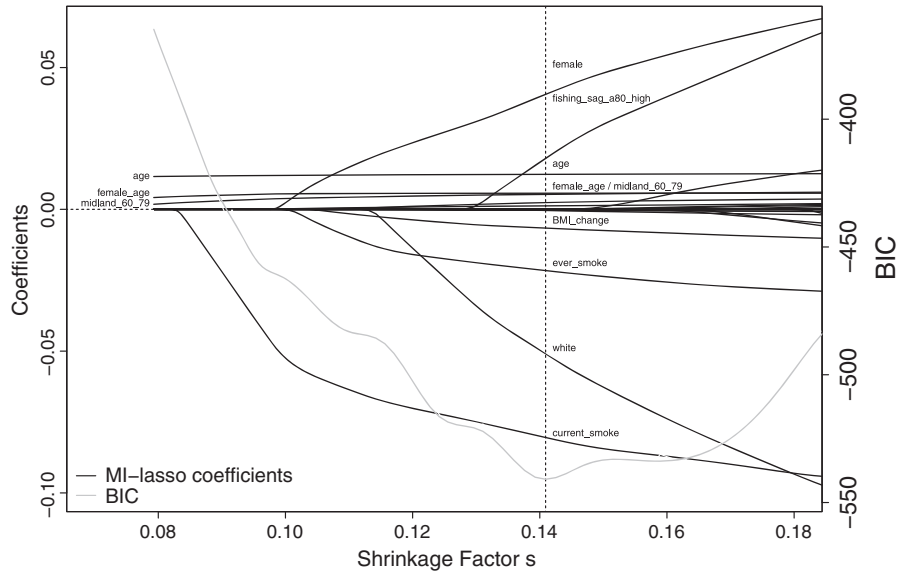

Figure 1. Profiles of MI-LASSO coefficients and BIC value, as tuning parameter λ changes. Average regression coefficients across five imputations are plotted versus shrinkage factor s using black curves. The vertical dashed line represents the selected model with $s = 0.141$ chosen by the smallest BIC, where the profile of BIC is plotted using a gray curve. Five selected covariates in Table IV (months of breast-feeding, lifetime pack-years smoking, years of using wood-burning stoves in 1960-1979, years of living in Midland in 1940-1959, and years of working in paper industry after 1980) are not labeled because of their small absolute values of coefficients. BMI, body mass index.

covariates are highly correlated with the correlation as high as 0.85. We also included the age and gender interaction, because previous study showed its importance in predicting serum TCDD concentration [7, 48]. The outcome serum TCDD concentration was log-10 transformed.

Our analysis was based on the five imputations completed by Olson *et al*. [10] using sequential regression MI strategy implemented in IVEware [11]. We specified a series of tuning parameter λ values and showed the profiles of MI-LASSO coefficients and BIC values as λ changes in Figure 1. The y-axis on the left is the average regression coefficient estimate, $\frac{1}{5}\sum_{d=1}^{5}\hat{\beta}_{d,j}$, for $j = 1, \ldots, 131$, calculated from the MI-LASSO optimization equation (3). The y-axis on the right is the BIC value defined in equation

(5). The coefficient estimates and BIC are displayed using black and gray curves, respectively. The x-axis is the standardized shrinkage factor s, defined as

$$s = \frac{\sum_{j=1}^{131} \sqrt{\sum_{d=1}^{5} \hat{\beta}_{d,j}^2}}{\sum_{j=1}^{131} \sqrt{\sum_{d=1}^{5} \tilde{\beta}_{d,j}^2}},$$

where $\tilde{\beta}_{d,j}$ is the ordinary least square estimate for the jth covariate on the dth dataset. When $s = 0$, the null model is selected, and when $s = 1$, no shrinkage is applied, and $\hat{\beta}_{d,j} = \tilde{\beta}_{d,j}$. Here, we only present

Table IV. Comparison of RR-stepwise and MI-LASSO in selecting important factors that predict $\log_{10}$ serum 2,3,7,8-tetrachlorodibenzop-dioxin concentration in the University of Michigan Dioxin Exposure Study flood plain and near-flood plain sample

	RR-stepwise			MI-LASSO		
	Covariate	Estimate	95% CI	Covariate	Estimate	95% CI
(0)	Intercept	0.426	(0.290, 0.562)	Intercept	0.374	(0.180, 0.570)
(1)	Age	0.015	(0.011, 0.018)	Age	0.013	(0.010, 0.017)
(2)	Female	0.161	(0.087, 0.235)	Female	0.147	(0.078, 0.216)
(3)	Age × female	0.006	(0.002, 0.009)	Age × female	0.006	(0.002, 0.010)
(4)	Months of breast-feeding	-0.006	(-0.009, -0.003)	Months of breast-feeding	-0.006	(-0.008, -0.003)
(5)	Lifetime pack-years smoking	-0.002	(-0.004, -0.001)	Lifetime pack-years smoking	-0.002	(-0.004, 0.0004)
(6)	Current smoker	-0.155	(-0.264, -0.046)	Current smoker	-0.139	(-0.252, -0.025)
(7)	BMI change	-0.022	(-0.031, -0.013)	BMI change	-0.018	(-0.027, -0.009)
(8)	Years of living in Midland: 1960-79	0.011	(0.006, 0.016)	Years of living in Midland: 1960-79	0.008	(0.003, 0.013)
(9)	Fishing in Saginaw river and bay after 1980 (≥ 1/moth)	0.165	(0.044, 0.286)	Fishing in Saginaw river and bay after 1980 (≥ 1/month)	0.195	(0.075, 0.316)
(10)	White	-0.190	(-0.351, -0.030)	White	-0.213	(-0.415, -0.011)
(11)	Years of using wood burning stoves in 1960-1979	0.008	(0.002, 0.013)	Years of using wood burning stoves in 1960-1979	0.009	(0.003, 0.014)
(12)	Years of working in paper industry after 1980	0.010	(0.003, 0.017)	Years of working in paper industry after 1980	0.008	(0.003, 0.014)
(13)	BMI	0.006	(0.002,0.011)	Ever smoke	-0.051	(-0.129, 0.027)
(14)	Years of living in Midland: after 1980	-0.006	(-0.012, -0.0002)	Years of living in Midland: 1940-59	0.003	(-0.002, 0.008)
(15)	Years of living in a property with crops livestock or poultry after 1980	0.008	(0.001,0.015)			
(16)	Years of spraying chemicals to kill plants: 1940-59	-0.034	(-0.059, -0.009)			
(17)	Years of spraying chemicals to kill plants: 1960-79	0.024	(0.011, 0.038)			
(18)	Years of working in foundry after 1980	-0.008	(-0.016, -0.0005)			
(19)	Eating meat raised in Saginaw river and bay (> 0, <1/month)	0.115	(0.020, 0.210)			

BMI, body mass index.

the profiles when s lies between 0.08 and 0.18. We used a vertical dashed line in Figure 1 to represent the selected model at $s = 0.141$, where the BIC achieves the smallest value. The covariates with nonzero coefficients are age, female, age and female interaction, BMI change, months of breastfeeding, whether a current smoker or ever smoke, lifetime pack-years smoking, white, years of living in Midland county in 1940-1959 and 1960-1979, fishing in Saginaw river and bay after 1980 more than once per month, using wood burning stoves in 1960-1979, and years of working in paper industry after 1980.

We refitted the selected model by using least square estimation and used RR to obtain combined regression coefficient estimates and 95% CIs (Table IV). This is similar to the approach suggested in Efron *et al*. [49]. For comparison, we also included the coefficients and 95% CIs of the model selected, by using the RR-stepwise method, with $\alpha_1 = 0.05$ and $\alpha_2 = 0.06$. The top part of Table IV lists the 12 covariates that were selected by both methods, and the bottom part shows the covariates that were selected by one method but not the other. The 12 covariates selected by both methods have similar coefficients and are all statistically significant at the significance level of 0.05. The analysis suggests that the effect of age on serum TCDD concentration differs by gender – the serum TCDD concentration cumulates more among females than males as they age. In addition, serum TCDD concentration is negatively associated with white ethnicity, months of breast-feeding among women, smoking, and weight gain in the past year. These findings are consistent with previous reports [50, 51]. Furthermore, we also identified some important residence, local recreational activity, occupation, and property-use factors, which can be important to local residents, government agencies, and policy makers. Specifically, the serum TCDD concentration is higher among people living more years in Midland county in 1960-1979, fishing frequently in Saginaw river and bay after 1980, using wood burning stoves in 1960-1979, or working in paper industry after 1980. In addition to the 12 common selected covariates, another two covariates (whether ever smoke or years of living in Midland county in 1940-1959) were selected by the use of MI-LASSO. Effects of these two covariates are in the same direction as we expected, although they are not statistically significant. The RR-stepwise method also identified another seven covariates, but it is hard to interpret the coefficients of some of these variables. For example, it cannot be true that spraying chemicals to kill plants in 1940-1959 is negatively associated with serum TCDD concentration. Moreover, we observed the instability in the selection of models obtained using the RR-stepwise method. Analyses not shown here found that some of the seven covariates selected by RR-stepwise but not MI became insignificant when we dropped or added a few unimportant candidate covariates. Thus, the importance of these seven covariates requires further investigation.

6. Discussion

Missing data arise in almost all research problems and can cause difficulties in data analysis. MI has become one of the most popular methods to deal with missing data because of its generality, ease of implementation, and availability of relevant software. In this paper, we propose an MI-LASSO method for variable selection with multiply-imputed datasets. The MI-LASSO method jointly fits models across multiple-imputed datasets by applying a group LASSO penalty on the regression coefficients of the same covariate in different imputed datasets. It has an important feature of consistency in the selection of variables across MIs. That is, the coefficients of the same covariate estimated by MI-LASSO are either all zero or all nonzero in different imputations. Our simulation study shows that MI-LASSO can identify important covariates in a linear regression model similar to LASSO would have achieved if the data were complete, and consequently, the selected models have similar MSEs. By discarding incomplete observations, the use of LASSO on complete-cases results in models with poor sensitivity of selection and much larger MSE than MI-LASSO, especially when proportion of complete-cases is low or missing data are MAR. Furthermore, as proportion of missing information increases, the sensitivity, specificity, and mean-squared error are much worse by the use of LASSO on complete cases but do not differ much by the use of MI-LASSO. This suggests against conducting variable selection on complete cases, which may fail to identify important covariates, but recommends the use of MI-LASSO, especially when proportion of missing information is moderate to high.

The MI-LASSO method is superior to the stepwise variable selection method by repeated use of RR (RR-stepwise). The MI-LASSO method results in models with better sensitivity and MSE than the RR-stepwise method, and the advantage is especially pronounced for small sample size data with a large number of covariates and for data with the presence of multicollinearity. The MI-LASSO method also yields more stable selection of models than the RR-stepwise method as evident in the UMDES data application. Moreover, the MI-LASSO method is easy to implement, and the computation is fast.

By applying the local quadratic-approximation method, the coefficients of MI-LASSO can be easily estimated by iteratively solving ridge regressions. Both SAS macro and R code for implementing the MI-LASSO method are available upon request. This makes MI-LASSO more attractive to practitioners than alternative computationally intensive methods, such as the Bayesian variable selection method for multiply-imputed data [26].

In practice, we often assume that missing data are ignorable when MI is used to handle missing data. In the simulation study, we assumed that missing data are MCAR or MAR and studied the performance of MI-LASSO under varying sample sizes, varying numbers of candidate covariates, varying correlations between candidate covariates, varying types of data, and varying proportions of missing data. All scenarios of simulation show the good performance of MI-LASSO when missing data are ignorable. MI in non-ignorable missing data is challenging. There is no general MI method or software program available that can be used to handle non-ignorable missing data. However, it is important to note that MI paradigm and the use of RR do not require missing data mechanism to be ignorable, nor does MI-LASSO. MI-LASSO will perform well in non-ignorable missing data if imputations are well conducted and missing data mechanism is correctly modeled. On the other hand, with a large proportion of missing data, a misspecified imputation model may distort the observed relationship between the outcome and covariates, so that MI-LASSO may lead to a poor selection of models based on bad imputations using naïve imputation methods.

In this paper, we mainly focus on the methodology development and computation aspect of the MI-LASSO method. The corresponding theoretical properties of our proposed method are under exploration. We expect to show the sign consistency of our method for variable selection by assuming some conditions similar to irrepresentative condition for LASSO [44]. The challenging part of the development of the theoretical properties is to assess the effect of uncertainty in imputation on the variable selection. Our proposed method can be also straightforwardly extended to generalized linear models.

Appendix

A.1. Stepwise selection method for multiply-imputed data using Rubin's rules

Wood, White, and Royston [28] described a stepwise variable selection method for multiply-imputed data by repeated use of RR and referred to it as the RR approach. We present in the following the detailed RR-stepwise selection procedure:

Step 0: Choose α_1, the P-value to enter, and α_2, the P-value to remove. Specify the initial model M_0, the model with no covariates. Set $t = 0$.

Step 1: Let $t = t + 1$. For each covariate X that is not included in M_{t-1}, fit D regressions with the model $\{M_{t-1}, X\}$ on D imputed datasets. Calculate the combined P-value for each newly added covariate X by using RR. Let X_a be the covariate with the smallest combined P-value p_a. If $p_a \leqslant \alpha_1$, update the model M_t to be $\{M_{t-1}, X_a\}$; otherwise, $M_t = M_{t-1}$, and the procedure terminates.

Step 2: Fit D regressions with the model M_t on D imputed datasets. Calculate the combined P-values for covariates in the model. Let X_b be the covariate with the largest combined P-value p_b. If $p_b > \alpha_2$, update the model M_t to be $\{M_t, -X_b\}$, where the minus sign means removing X_b from M_t.

Step 3: Repeat step 2 until the largest combined P-value p_b is smaller than or equal to α_2.

Step 4: Go back to step 1, and iterate until the procedure terminates.

When the RR-stepwise iteration terminates, the combined P-values for all the covariates in the model are not bigger than α_2, and none of the covariates not in the model has combined P-value smaller than or equal to α_1 if it is added into the model. To avoid endless iteration, we need the condition of $\alpha_1 \leqslant \alpha_2$.

A.2. Degrees of freedom estimation for multiple imputation-least absolute shrinkage and selection operator

We re-represent our MI-LASSO method as a group LASSO regression problem and derive its degrees of freedom for the fitted model. For the ith subject, let $\boldsymbol{w}_{ij}$ be a square diagonal matrix with the diagonal entries be $\boldsymbol{x_{ij}} = (x_{1,ij}, \ldots, x_{D,ij})^T$. Denote $z_i = (y_{1,i}, \ldots, y_{D,i})^T$, $\boldsymbol{\beta}_0 = (\beta_{1,0}, \ldots, \beta_{D,0})^T$, and

$\boldsymbol{\beta}_j = (\beta_{1,j}, \ldots, \beta_{D,j})^T$ for $j = 1, \ldots, p$. Given λ, the optimization problem (3) is equivalent to the following group LASSO regression estimation:

$$\min_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} \|z_i - \boldsymbol{\beta}_0 - \sum_{j=1}^{p} \boldsymbol{w}_{ij}^T \boldsymbol{\beta}_j\|_2^2 + \lambda \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_2,$$

where $\|\boldsymbol{\beta}_j\|_2 = \sqrt{\sum_{d=1}^{D} \beta_{d,j}^2}$. This equivalence can be verified using matrix calculation straightforwardly, and the details are omitted here.

Following the degrees of freedom for group LASSO regression, which is derived in Yuan and Lin (Equation 6.3) [19], we estimate the degrees of freedom for the fitted model by the MI-LASSO method as

$$df_2 = \sum_{j=1}^{p} \boldsymbol{I}\left(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0\right) + \sum_{j=1}^{p} \frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\tilde{\boldsymbol{\beta}}_j\|_2}(D-1),$$

or

$$df_2 = \sum_{j=1}^{p} \boldsymbol{I}\left(\sqrt{\sum_{d=1}^{D} \hat{\beta}_{d,j}^2} > 0\right) + \sum_{j=1}^{p} \frac{\sqrt{\sum_{d=1}^{D} \hat{\beta}_{d,j}^2}}{\sqrt{\sum_{d=1}^{D} \tilde{\beta}_{d,j}^2}}(D-1),$$

where $\hat{\boldsymbol{\beta}}_j = \left(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}\right)^T$ and $\tilde{\boldsymbol{\beta}}_j = \left(\tilde{\beta}_{1,j}, \ldots, \tilde{\beta}_{D,j}\right)^T$ with $\hat{\beta}_{d,j}$ and $\tilde{\beta}_{d,j}$ as the MI-LASSO estimate and the ordinary least square estimate for the jth variable on the dth dataset, respectively.

Acknowledgements

The authors thank Dr. David Garabrant for making the UMDES data available and Dr. Jun Shao for a very helpful discussion. We also thank an associate editor and referees for their helpful comments on the original version of this paper. This research was supported in part by the Calderone Junior Faculty Research Prize from Columbia University Mailman School of Public Health.

References

1. Arisawa K, Takeda H, Mikasa H. Background exposure to PCDDs/PCDFs/PCBs and its potential health effects: A review of epidemiologic studies. *The Journal of Medical Investigation* 2001; **52**:10–21.
2. Baccarelli A, Mocarelli P, Patterson Jr. DG, Bonzini M, Pesatori AC, Caporaso N, Landi MT. Immunologic effects of dioxin: New results from Seveso and comparison with other studies. *Environmental Health Perspectives* 2002; **110**:1169–1173.
3. Dalton T, Kerzee J, Wang B, Miller M, Dieter M, Lorenz M, Shertzer H, Nebert D, Puga A. Dioxin exposure is an environmental risk factor for ischemic heart disease. *Cardiovascular Toxicology* 2001; **1**:285–298.
4. Passi S, Nazzaro-Porro M, Boniforti L, Gianottij F. Analysis of lipids and dioxin in chloracne due to tetrachloro-2,5,7,8-p-dibenzodioxin. *British Journal of Dermatology* 1981; **105**:137–143.
5. Zambon P, Ricci P, Bovo E, Casula A, Gattolin M, Fiore AR, Chiosi F, Guzzinati S. Sarcoma risk and dioxin emissions from incinerators and industrial plants: A population-based case-control study (Italy). *Environmental Health: A Global Access Science Source* 2007; **6**:19.
6. Garabrant DH, Fransblau A, Lepkowski J, Gillespie BW, Adriaens P, Demond A, Ward B, Ladronka K, Hedgeman E, Knutson K, Zwica L, Olson K, Towey T, Chen Q, Hong B. The University of Michigan Dioxin Exposure Study: Methods for an environmental exposure study of polychlorinated dioxins, furans, and biphenyls. *Environmental Health Perspectives* 2009; **117**:803–810.
7. Garabrant DH, Fransblau A, Lepkowski J, Gillespie BW, Adriaens P, Demond A, Hedgeman E, Knutson K, Zwica L, Olson K, Towey T, Chen Q, Hong B, Chang CW, Lee SY, Ward B, Ladronka K, Luksemburg W, Maier M. The University of Michigan Dioxin Exposure Study: Predictors of human serum dioxin concentrations in Midland and Saginaw, Michigan. *Environmental Health Perspectives* 2009; **117**:818–824.
8. Hedgeman E, Chen Q, Hong B, Chang CW, Olson K, Ladronka K, Ward B, Adriaens P, Demond A, Gillespie BW, Lepkowski J, Franzblau A, Garabrant DH. The University of Michigan Dioxin Exposure Study: Population survey results and serum concentrations for Polychlorinated Dioxins, Furans, and Biphenyls. *Environmental Health Perspectives* 2009; **117**:811–817.
9. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
10. Olson K, Sinibaldi J, Lepkowski JM, Ward B, Ladronka K, Towey T, Wright D, Gillespie BW. Missing data in an environmental exposure study: Imputation to improve survey estimation. *Organohalogen Compounds* 2006; **68**:1346–1349.

11. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A Multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; **27**:85–95.
12. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**:716–723.
13. Schwarz GE. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**:461–464.
14. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 1996; **58**:267–288.
15. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96**:1348–1360.
16. Lv J, Fan Y. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* 2009; **37**:3498–3528.
17. Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 2010; **38**:894–942.
18. Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association* 2012; **107**(497):223–232.
19. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 2006; **68**:49–67.
20. Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* 2009; **37**:3468–3497.
21. Shen X, Huang H-C. Grouping pursuit through a regularization solution surface. *Journal of American Statistical Association* 2010; **105**:727–739.
22. George E, Foster D. Calibration and empirical Bayes variable selection. *Biometrika* 2000; **87**:731–747.
23. George E, McCulloch R. Variable selection via Gibbs sampling. *Journal of American Statistical Association* 1993; **88**:881–889.
24. Park T, Casella G. The Bayesian Lasso. *Journal of the American Statistical Association* 2008; **103**:681–687.
25. Heymans MW, Van Buuren S, Knol DL, Van Mechelenen W, de Vet HCW. Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology* 2007; **7**:33–42.
26. Yang X, Belin TR, Boscardin WJ. Imputation and variable selection in linear regression models with missing covariates. *Biometrics* 2005; **61**:498–506.
27. Little JA, Rubin DB. *Statistical Analysis with Missing Data*, (2nd edn). Wiley: Hoboken, 2002.
28. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Statistics in Medicine* 2008; **27**:3227–3246.
29. Rubin DB. Multiple imputation after 18+ years (with discussion). *Journal of American Statistical Association* 1996; **91**:473–489.
30. Harel O, Zhou XH. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine* 2007; **26**(16):3057–3077.
31. Barnard J, Meng X. Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods and Medical Research* 1999; **8**:17–36.
32. Schafer JL. Multiple imputation: A primer. *Statistical Methods in Medical Research* 1999; **8**:3–15.
33. Zhou XH, Eckert G, Tierney WM. Multiple imputation in public health research. *Statistics in Medicine* 2001; **20**:1541–1549.
34. Meng XL. Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* 1995; **10**:538–573.
35. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.
36. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman and Hall: London, 1997.
37. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research* 2011; **20**:40–49.
38. Van Buuren S, Groothuis-Oudshoorn K. mice:Multivariate imputation by chained equations in R. *Journal of Statistical Software* 2011; **45**:1–67.
39. Su Y-S, Gelman A, Hill J, Yajima M. Multiple imputation with diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software* 2011; **45**(2):1–31. http://www.jstatsoft.org/v45/i02/.
40. van Buuren S, Brand JPL, Groothuis-Oudshoorn K, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006; **76**:1049–1064.
41. Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychological Methods* 2002; **7**:147–177.
42. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of American Statistical Association* 1986; **81**:366–374.
43. Knigh TK, Fu W. A symptotics for Lasso-type estimators. *The Annals of Statistics* 2000; **28**:1356–1378.
44. Zhao P, Yu B. On model selection consistency of Lasso. *Journal of Machine Learning Research* 2006; **7**:2541–2563.
45. Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statististics* 2009; **37**:1705–1732.
46. Huang J, Zhang T. The benefit of group sparsity. *Annals of Statististics* 2010; **38**:1978–2004.
47. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 2005; **67**:301–320.
48. Chen Q, Garabrant DH, Hedgeman E, Little RJA, Elliott MR, Gillespie B, Lee S-Y, Lepkowski JM, Franzblau A, Adriaens P, Demond AH, Patterson Jr. DG. Estimation of background serum 2,3,7,8-TCDD concentrations by using quantile regression in the UMDES and NHANES populations. *Epidemiology* 2010; **21**:S51–S57.
49. Efron B, Hastie T, Johnstone I, Tibshirani T. Least angle regression. *Annals of Statistics* 2004; **32**:407–499.
50. Patterson Jr. DG, Patterson D, Canady R, Wong L-Y, Lee R, Turner W, Caudil S, Needham L, Henderson A. Age specific dioxin TEQ reference range. *Organohalogen Compounds* 2004; **66**:2844–2849.
51. Wittsiepe J, Schrey P, Ewers U, Selenka F, Wilhelm M. Decrease of PCDD/F levels in human blood from Germany over the past ten years (1989-1998). *Chemosphere* 2000; **40**:1103–1109.