Original Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280213511027
smm.sagepub.com
SAGE

Multiple imputation in the presence of high-dimensional data

Yize Zhao and Qi Long

Abstract

Missing data are frequently encountered in biomedical, epidemiologic and social research. It is well known that a naive analysis without adequate handling of missing data may lead to bias and/or loss of efficiency. Partly due to its ease of use, multiple imputation has become increasingly popular in practice for handling missing data. However, it is unclear what is the best strategy to conduct multiple imputation in the presence of high-dimensional data. To answer this question, we investigate several approaches of using regularized regression and Bayesian lasso regression to impute missing values in the presence of high-dimensional data. We compare the performance of these methods through numerical studies, in which we also evaluate the impact of the dimension of the data, the size of the true active set for imputation, and the strength of correlation. Our numerical studies show that in the presence of high-dimensional data the standard multiple imputation approach performs poorly and the imputation approach using Bayesian lasso regression achieves, in most cases, better performance than the other imputation methods including the standard imputation approach using the correctly specified imputation model. Our results suggest that Bayesian lasso regression and its extensions are better suited for multiple imputation in the presence of high-dimensional data than the other regression methods.

Keywords

Bayesian lasso regression, high-dimensional data, missing data, multiple imputation, regularized regression

1 Introduction

Missing data are frequently encountered in biomedical, epidemiologic, and social research. It is well known that a naive analysis without adequate handling of missing data may lead to bias and/or loss of efficiency. Among statistical methods that have been developed for handle missing data, multiple imputation (MI)[1,2] can be readily conducted using existing software package[3,4] in a wide range of situations and it allows data analysts to apply standard complete-data analysis directly to imputed data sets. As a result, MI has become increasingly popular in practice. The key idea of MI is to replace each missing value with a set of "plausible" values drawn from their predictive distributions

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA

Corresponding author:
Qi Long, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA.
Email: qlong@emory.edu

conditional on the observed data and generate multiple imputed data sets to account for uncertainty of imputing missing values. Subsequently, each imputed data set is analyzed separately using standard complete-data methods and the results are combined across imputed data sets using Rubin's rule.[1,2] MI has been investigated extensively in many settings.[5–12] Harel and Zhou[13] provide a nice review of theory, implementation, and software for MI.

While MI has proven to be very flexible, its validity in practice is predicated on several important conditions. First, most MI methods assume that data are missing at random (MAR)[2]; in particular, we are not aware of any available MI software package which can deal with the case that data are not missing at random (NMAR). While sensitivity analysis has been proposed in combination with MI in the presence of NMAR, this approach has not been widely used in practice. Most practitioners prefer direct application of MI, often implicitly or explicitly relying on the assumption of MAR. Second, imputation models need to be correctly specified or close to the true models and it has been advocated[14,15] that imputation models should be as general as the data allow them to be so as to accommodate a wide range of statistical models that will be used on the imputed data sets. One reason is that it is often unclear to an imputer what analysis will eventually be conducted by data analysts; furthermore, more general imputation models make it more likely that the assumption of MAR will hold and imputation is proper. To build a sensible, general imputation model, a major challenge is to avoid leaving out important predictors, since leaving out important variables may lead to imputation models that are less general than analysis models and then biased results. In particular, in the case of high-dimensional data where the number of variables (p) is large or even greater than the sample size (n), it is often not feasible to include all possibly relevant predictors and their interactions in imputation models. When conducting imputation in such cases, model trimming or regularization becomes imperative but classical model trimming techniques such as stepwise selection-based Akaike information criterion[16] or other criteria are known to perform poorly. To the best of our knowledge, principled MI approaches that can handle high-dimensional data ($p > n$ or $p \gg n$) have not been investigated in the literature and this issue is not addressed in existing MI software packages, which likely perform poorly or fail as shown in our numerical studies.

Regularized regression has been proposed to conduct simultaneous parameter estimation and model selection, which seems to offer a natural solution for the issue of constructing imputation models in the presence of high-dimensional data. We briefly review here the key concepts of regularized regression and Bayesian lasso (BLasso) regression and explore how they may fit in with imputation. Consider a linear regression model with n observations and p predictors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ is the $n \times p$ design matrix, and ϵ is white noise. In the case of $p < n$, one could estimate the coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ by minimizing the residual sum of squares

$$L(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

which leads to the ordinary least squares estimator. However, in the cases of $p > n$ and $p \approx n$, the aforementioned approach fails and instead we can use the regularized least squares estimator

$$\hat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + p_\lambda(\boldsymbol{\beta})$$

where $p_\lambda(\boldsymbol{\beta})$ is a regularization function, shrinking some parameter estimates toward/to zero. Several forms of $p_\lambda(\boldsymbol{\beta})$ have been proposed, leading to different regularized estimators; some popular choices include ridge penalty,[17] $p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2$, lasso penalty,[18] $p_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$, elastic net (EN) penalty,[19] $p_\lambda(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$, and adaptive lasso (ALasso),[20] $p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|$, where $\hat{\boldsymbol{\omega}} = 1/|\hat{\boldsymbol{\beta}}|^\gamma$ and $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$ consistent estimator. Alternatively, we can use BLasso regression and its generalizations[21–23] to fit the aforementioned model in the presence of high-dimensional data. While there are some connections between BLasso regression and regularized regression, there are also some key differences; in particular, prediction, not model selection, is of primary interest in BLasso regression, making it a more natural fit for MI. However, the performance of these methods has not been evaluated in the context of MI.

In this article, we investigate approaches for multiply imputing missing values in the presence of high-dimensional data and compare their finite sample performance through numerical studies. The remainder of this article is organized as follows. In Section 2, we describe three MI approaches based on regularized regression and BLasso regression. In Section 3, we present numerical results for evaluating the performance of the proposed and existing approaches in the presence of high-dimensional data. We conclude this article with some discussion remarks in Section 4.

2 MI methods

Suppose that we have a sample of n independent observations from a target population; each observation consists of p variables $\mathbf{z}_i = \left(z_{i,1}, z_{i,2}, \ldots, z_{i,p}\right)^T$ $(i = 1, 2, \ldots, n)$, representing a random draw from a multivariate distribution with a set of unknown parameters $\boldsymbol{\theta}$. Some of the p variables have missing values. Define the missing data indicator matrix $\Delta = (\delta_{i,j})$ such that $\delta_{i,j} = 1$ if $z_{i,j}$ is missing and $\delta_{i,j} = 0$ if $z_{i,j}$ is observed. Let $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)^T$ denote the complete data, $\mathbf{Z}_{\text{obs}}$ denote the observed components of $\mathbf{Z}$, and $\mathbf{Z}_{\text{mis}}$ denote the missing components.

Throughout, we assume that data are MAR, that is, $f(\Delta|\mathbf{Z}, \phi) = f(\Delta|\mathbf{Z}_{\text{obs}}, \phi)$, where ϕ denotes the set of parameters associated with the missing data mechanism; we also assume that ϕ and θ are distinct, resulting in ignorable missingness. It then follows from the assumption of MAR that

$$f(\mathbf{Z}_{\text{mis}}|\mathbf{Z}_{\text{obs}}) = \int f(\mathbf{Z}_{\text{mis}}|\mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{Z}_{\text{obs}}) \mathrm{d}\boldsymbol{\theta} \tag{1}$$

$$f(\boldsymbol{\theta}|\mathbf{Z}_{\text{obs}}) \propto f(\boldsymbol{\theta}) \int f(\mathbf{Z}_{\text{obs}}, \mathbf{Z}_{\text{mis}}|\boldsymbol{\theta}) \mathrm{d}\mathbf{Z}_{\text{mis}} \tag{2}$$

where $f(\boldsymbol{\theta})$ represents a prior distribution of θ. $f(\mathbf{Z}_{\text{mis}}|\mathbf{Z}_{\text{obs}})$ and $f(\mathbf{Z}_{\text{mis}}|\mathbf{Z}_{\text{obs}}, \boldsymbol{\theta})$ in equation (1) are the conditional predictive distributions of $\mathbf{Z}_{\text{mis}}$ given $\mathbf{Z}_{\text{obs}}$ and given $\mathbf{Z}_{\text{obs}}$ and $\boldsymbol{\theta}$, respectively. $f(\boldsymbol{\theta}|\mathbf{Z}_{\text{obs}})$ in equations (1) and (2) is the posterior distribution of $\boldsymbol{\theta}$ conditional on the observed data. Equations (1) and (2) motivate the standard MI procedure, which typically consists of two steps: in the m-th imputation $(m = 1, \ldots, M)$, one first generates a random draw for $\boldsymbol{\theta}$ from its posterior distribution, denoted by $\hat{\boldsymbol{\theta}}^{(m)}$, and then impute each component of $\mathbf{Z}_{\text{mis}}$ by a random draw from the conditional predictive distribution $f(\mathbf{Z}_{\text{mis}}|\mathbf{Z}_{\text{obs}}, \hat{\boldsymbol{\theta}}^{(m)})$. Subsequently, statistical methods for complete data can be directly applied to each of the M imputed data sets and Rubin's rule can be used to combine the results across the M imputed data sets. Specifically, assuming that the parameter of interest in the data analysis is β and its estimate and associated variance estimate in the m-th imputed data set are $\hat{\beta}^{(m)}$ and $U^{(m)}$, respectively, the combined estimate for the M imputed data sets is $\bar{\beta} = M^{-1} \sum_{m=1}^{M} \hat{\beta}^{(m)}$

and its variance is estimated by $\widehat{\mathrm{Var}}(\bar{\beta}) = \bar{U} + (1 + M^{-1})B$, where $B = (M-1)^{-1}\sum_{m=1}^{M}(\bar{\beta} - \hat{\beta}^{(m)})^2$ and $\bar{U} = M^{-1}\sum_{m=1}^{M} U^{(m)}$.

In order to conduct MI, it is essential to postulate a statistical model for $f(\mathbf{Z}_{\mathrm{mis}}|\mathbf{Z}_{\mathrm{obs}}, \boldsymbol{\theta})$ or more broadly for $f(\mathbf{Z}|\boldsymbol{\theta})$ so as to obtain the posterior distribution of $\boldsymbol{\theta}$ in equation (2). For example, for continuous $\mathbf{Z}$, one could assume that $\mathbf{Z}$ follows a multivariate normal distribution with unknown mean and variance–covariance. When the dimension of the data (p) is much smaller than the sample size of the observed data, one could fit a fairly complex model without any restrictions, say, on the correlation structure of a multivariate normal $\mathbf{Z}$, and obtain the posterior distribution of $\boldsymbol{\theta}$ using the observed data. If, however, p is large relative to the sample size, model trimming or regularization are needed. Since classical model trimming techniques such as stepwise selection do not work well in the presence of high-dimensional data, the key idea underlying the proposed approaches is to use regularized regression or BLasso regression to estimate the posterior distribution of $\boldsymbol{\theta}$ in equation (2).

2.1 A simplified data setup

For the ease of exposition, we consider a simplified data setup where only one variable has missing values with the others fully observed. Without loss of generality, we assume that $\mathbf{z}_1$ is continuous and contains missing values with the first r components observed, $\mathbf{z}_{\mathrm{obs},1} = (z_{1,1}, \ldots, z_{r,1})^T$, and the remaining $n - r$ components missing, $\mathbf{z}_{\mathrm{mis},1} = (z_{r+1,1}, \ldots, z_{n,1})^T$. Define the complement data set for $\mathbf{z}_1$ as $\mathbf{Z}_{-1} = (\mathbf{z}_{1,-1}, \mathbf{z}_{2,-1}, \ldots, \mathbf{z}_{n,-1})^T = (\mathbf{Z}_{\mathrm{obs},-1}, \mathbf{Z}_{\mathrm{mis},-1})^T$ with $\mathbf{z}_{i,-1} = (z_{i,2}, z_{i,3}, \ldots, z_{i,p})^T$, and define the complement data sets for $\mathbf{z}_{\mathrm{obs},1}$ and $\mathbf{z}_{\mathrm{mis},1}$ as $\mathbf{Z}_{\mathrm{obs},-1} = (\mathbf{z}_{1,-1}, \ldots, \mathbf{z}_{r,-1})^T$ and $\mathbf{Z}_{\mathrm{mis},-1} = (\mathbf{z}_{r+1,-1}, \ldots, \mathbf{z}_{n,-1})^T$, respectively. Then the observed data are $\mathbf{Z}_{\mathrm{obs}} = (\mathbf{z}_{\mathrm{obs},1}, \mathbf{Z}_{\mathrm{obs},-1}, \mathbf{Z}_{\mathrm{mis},-1})$ and the missing data are $\mathbf{Z}_{\mathrm{mis}} = (\mathbf{z}_{\mathrm{mis},1})$; there are r complete cases and n–r incomplete cases with $\mathbf{z}_1$ missing. It follows that the imputation model (1) reduces to

$$f(\mathbf{z}_{\mathrm{mis},1}|\mathbf{z}_{\mathrm{obs},1}, \mathbf{Z}_{-1}) = \int f(\mathbf{z}_{\mathrm{mis},1}|\mathbf{Z}_{\mathrm{mis},-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{z}_{\mathrm{obs},1}, \mathbf{Z}_{\mathrm{obs},-1}) \mathrm{d}\boldsymbol{\theta} \tag{3}$$

To obtain the posterior distribution of $\boldsymbol{\theta}$, $f(\boldsymbol{\theta}|\mathbf{z}_{\mathrm{obs},1}, \mathbf{Z}_{\mathrm{obs},-1})$, we can posit and fit a regression model with $\mathbf{z}_{\mathrm{obs},1}$ as the outcome variable and $\mathbf{Z}_{\mathrm{obs},-1}$ as the set of predictors. For the purpose of illustration, we consider here a linear regression model

$$\mathbf{z}_{\mathrm{obs},1} = \alpha_0 + \mathbf{Z}_{\mathrm{obs},-1}\boldsymbol{\alpha} + \epsilon \tag{4}$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_r)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p-1})^T$. Under Model (4), $\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\alpha}, \sigma^2)^T$. When $p \ll r$, MI software packages that implement existing MI procedures such as R packages `mi` and `mice` can be directly used to fit Model (4), obtain $f(\boldsymbol{\theta}|\mathbf{z}_{\mathrm{obs},1}, \mathbf{Z}_{\mathrm{obs},-1})$, and conduct imputation; however, when $p > r$ or $p \approx r$, the existing procedures and software packages are not directly applicable or do not perform well.

In Sections 2.2 to 2.4, we discuss three approaches to obtain $f(\boldsymbol{\theta}|\mathbf{z}_{\mathrm{obs},1}, \mathbf{Z}_{\mathrm{obs},-1})$ in the cases of $p > r$ or $p \approx r$. In Section 2.5, we briefly discuss extensions to more general missing patterns. Of note, the proposed methods can be readily extended to the cases where interaction terms between individual columns of $\mathbf{Z}_{\mathrm{obs},-1}$ are included in Model (4) or the variables with missing values such as $\mathbf{z}_1$ follow other distributions.

2.2 MI through direct use of regularized regression

We consider the data setup as introduced in Section 2.1 and our goal is to conduct MI for $\mathbf{z}_{\text{mis},1}$. Specifically, we need to fit the imputation model (4) using r complete cases. We denote by $\mathcal{S}$ the set of variables in $\mathbf{Z}_{\text{obs},-1}$ that are associated with $\mathbf{z}_{\text{obs},1}$, also known as the true active set, and denote by $|\mathcal{S}| = q$ its cardinality, that is, the number of important variables for imputing $\mathbf{z}_1$, and by $\mathbf{Z}_{\text{obs},\mathcal{S}}$ the corresponding design matrix. As stated previously, it is imperative to conduct model trimming or regularization when fitting Model (4); we define the subset of predictors that are selected to impute $\mathbf{z}_1$ as the active set, denoted by $\hat{\mathcal{S}}$, and denote the corresponding design matrix as $\mathbf{Z}_{\text{obs},\hat{\mathcal{S}}}$.

To achieve model trimming when fitting the imputation model (4), we propose to use regularization methods such as lasso or ALasso. However, it is not trivial to obtain the distribution $f(\theta|\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},-1})$ when regularized regression is used for (4). We first consider an approach where a regularization method is used to conduct both model trimming and parameter estimation and a bootstrap step is incorporated to simulate random draws from $f(\boldsymbol{\theta}|\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},\hat{\mathcal{S}}})$, ensuring that imputation is proper.[2] Similar bootstrap steps have been used in MI in other settings[24,25] for the same purpose. This approach is referred to as the direct use of regularized regression (DURR). In the DURR approach, the algorithm for the m-th imputation can be described as follows:

(1) Generate a bootstrap data set $\mathbf{Z}^{(m)}$ of size n by randomly drawing n observations from $\mathbf{Z}$ with replacement.
(2) Use a regularized regression method to fit Model (4) based on the complete cases in $\mathbf{Z}^{(m)}$, that is, $(\mathbf{z}^{(m)}_{\text{obs},1}, \mathbf{Z}^{(m)}_{\text{obs},-1})$, and obtain parameter estimate $\hat{\boldsymbol{\theta}}^{(m)}$, noting that $\hat{\boldsymbol{\theta}}^{(m)}$ can be considered a random draw from $f(\boldsymbol{\theta}|\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},-1})$.
(3) Impute $\mathbf{z}_{\text{mis},1}$ with $\mathbf{z}^{(m)}_{\text{mis},1}$ by drawing randomly from the predictive distribution $f(\mathbf{z}_{\text{mis},1}|\mathbf{Z}_{\text{mis},-1}, \hat{\boldsymbol{\theta}}^{(m)})$, noting that imputation is conducted on the original data set, not the bootstrap data set.

Repeating the above procedure for M times results in M imputed data sets. Subsequently, standard complete-data analysis can be applied to each one of the M imputed data sets.

2.3 MI through indirect use of regularized regression

We also investigate an alternative approach to DURR: a regularization method is used for model trimming only and is followed by a standard MI procedure using the estimated active set ($\hat{\mathcal{S}}$), say, through a maximum likelihood inference procedure. This approach is referred to as the indirect use of regularized regression (IURR). The algorithm of the IURR approach is described as follows:

(1) Use a regularized regression method to fit Model (4) based on the r complete cases in the observed data, that is, $(\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},-1})$, and identify the active set, $\hat{\mathcal{S}}$.
(2) Approximate the distribution, $f(\boldsymbol{\theta}|\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},\hat{\mathcal{S}}})$, using a standard inference procedure such as maximum likelihood.
(3) Conduct MI for $\mathbf{z}_{\text{mis},1}$: in the m-th imputation, randomly draw $\hat{\boldsymbol{\theta}}^{(m)}$ from $f(\boldsymbol{\theta}|\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},\hat{\mathcal{S}}})$, and subsequently impute $\mathbf{z}_{\text{mis},1}$ with $\mathbf{z}^{(m)}_{\text{mis},1}$ by drawing randomly from the predictive distribution $f(\mathbf{z}_{\text{mis},1}|\mathbf{Z}_{\text{mis},-1}, \hat{\boldsymbol{\theta}}^{(m)})$.

In this approach, step 3 is repeated for M times to obtain M imputed data sets. The most notable difference between DURR and IURR is that IURR only uses regularized regression to conduct model trimming for the imputation model whereas DURR uses regularized regression to conduct both model trimming and parameter estimation for the imputation model. In cases where p is large (e.g. $p = 200$ or 1000) and q is small (e.g. $q = 4$ or 20), we expect IURR to outperform DURR, since DURR is likely to shrink the regression coefficient estimates in $\mathcal{S}$ toward 0 in order to filter out noisy variables in $Z_{\text{obs},-1}$.

2.4 MI through Bayesian lasso regression

For various regularization methods, their Bayesian counterparts through hierarchical Bayesian formulations have been investigated in the literature.[21,26] The basic idea underlying the BLasso[21] is to set a conditional double-exponential prior $f(\boldsymbol{\alpha}|\sigma^2) = \prod_{j=1}^{p-1} \frac{\lambda}{2\sqrt{\sigma^2}} \exp \frac{-\lambda|\alpha_j|}{\sqrt{\sigma^2}}$ for regression parameters and noninformative scale-invariant marginal prior $f(\sigma^2) = 1/\sigma^2$ for σ^2, where the BLasso parameter λ is selected by marginal maximum likelihood. It follows that lasso estimates can be interpreted as the mode of the posterior distribution under a fully Bayesian formulation. In addition, to account for the uncertainty about regression model specification, Hans[23] proposed a mixture prior for α in the high-dimensional case

$$f(\alpha|\sigma^2, \lambda, \rho) = \prod_{j=1}^{p-1} \left\{ (1-\rho)\delta_0(\alpha_j) + \rho \frac{\lambda}{2\sqrt{\sigma^2}} \exp \frac{-\lambda|\alpha_j|}{\sqrt{\sigma^2}} \right\} \tag{5}$$

where $\delta(\cdot)$ is a point mass at zero. Priors for parameters σ^2, λ and ρ are

$$\sigma^2 \sim \text{Inverse} - \text{Gamma}(a, b), \quad \lambda \sim \text{Gamma}(r, s), \quad \rho \sim \text{Beta}(g, h) \tag{6}$$

with prespecified hyperparameters (a, b), (r, s) and (g, h).

There are several distinct advantages to use BLasso regression for imputation. First, it produces a valid posterior distribution of θ and a valid posterior predictive distribution of $\mathbf{z}_{\text{mis},1}$ via Markov chain Monte Carlo (MCMC) in a principled framework, avoiding the difficulty on generating predictive distributions for missing data that are associated with the use of frequentist approaches as described in Sections 2.2 and 2.3. Second, when conducting imputation, we are more interested in accurate prediction (imputation) than in variable selection and it has been shown that BLasso performs similarly or better in prediction when compared with frequentist lasso in finite samples.[21,26] Third, even though the prior probability of the event $\alpha_j = 0$ is nonzero in (5), Bayesian lasso tends to induce a weaker shrinkage effect compared with the frequentist regularization in the case of high-dimensional data; as a result, BLasso likely leads to more general imputation models in (4) and hence is better suited for imputation.

The imputation algorithm that incorporates BLasso regression is described as follows:

(1) Formulate a hierarchical BLasso model for the data based on Model (4), that is, assigning prior (5) for α and hyperpriors (6) for σ^2, λ, and ρ to account for model uncertainty and conduct Bayesian model trimming.
(2) Simulate $f(\boldsymbol{\theta}|\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},-1})$ via MCMC.

(3) Conduct MI for $\mathbf{z}_{\text{mis},1}$: in the m-th imputation, randomly draw $\hat{\boldsymbol{\theta}}^{(m)}$ from the posterior distribution, $f(\boldsymbol{\theta}|\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},-1})$, and subsequently impute $\mathbf{z}_{\text{mis},1}$ with $\mathbf{z}_{\text{mis},1}^{(m)}$ by drawing randomly from the posterior predictive distribution $f(\mathbf{z}_{\text{mis},1}|\mathbf{Z}_{\text{mis},-1}, \hat{\boldsymbol{\theta}}^{(m)})$.

2.5 MI for general missing pattern

In Sections 2.2 to 2.4, we present three MI approaches for the simplified data setup as defined in Section 2.1. Here, we consider a general missing pattern. Following the previous notation, for the data set $\mathbf{Z}$, we denote the observed components and missing components for variable j by $\mathbf{z}_{\text{obs},j}$ and $\mathbf{z}_{\text{mis},j}$ with the corresponding complement data set for the remaining variables denoted by $\mathbf{Z}_{\text{obs},-j}$ and $\mathbf{Z}_{\text{mis},-j}$ ($j = 1, 2, \ldots, p$), respectively. Note that unlike the simplified setup in Section 2.1, $\mathbf{Z}_{\text{obs},-j}$ and $\mathbf{Z}_{\text{mis},-j}$ may themselves contain missing values since the variables in the complement set of j may also contain missing values. In addition, we assume that the conditional distribution of $\mathbf{z}_{\text{obs},j}$ given $\mathbf{Z}_{\text{obs},-j}$ is parameterized by $\boldsymbol{\theta}_j$.

For the general missing pattern, we extend the methods described in Sections 2.2 to 2.4 based on the technique of *chained equations*; the basic idea is to impute the missing values of each variable using the remaining variables and conduct imputation iteratively. This approach has been adopted in multivariate imputation by chained equations (MICE)[27] and MICE has been implemented in several statistical software packages including `SPSS` and `R` and is applicable in cases where $p \ll n$. While MICE lacks rigorous theoretical justification for some general missing patterns, it has been shown in practice to achieve good performance in a wide range of settings.[3,27,28]

In the setup of our interest, we start the iterative procedure with some initial values. In the m-th step, we generate a random draw, $\hat{\boldsymbol{\theta}}_j^{(m)}$, from $f(\boldsymbol{\theta}_j|\mathbf{z}_{\text{obs},j}, \mathbf{Z}_{\text{obs},-j})$ through a regularized regression or BLasso regression approach as described in Sections 2.2 to 2.4; we then impute $\mathbf{z}_{\text{mis},j}^{(m)}$ based on the distribution $f(\mathbf{z}_{\text{mis},j}|\mathbf{Z}_{\text{mis},-j}, \hat{\boldsymbol{\theta}}_j^{(m)})$ for each $j = 1, 2, \ldots, p$. We repeat this step iteratively until convergence. The complete algorithm can be summarized as follows:

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_1^{(m)} &\sim f\left(\boldsymbol{\theta}_1|\mathbf{z}_{\text{obs},1}, \mathbf{Z}_{\text{obs},-1}^{(m-1)}\right) \\
\mathbf{z}_{\text{mis},1}^{(m)} &\sim f(\mathbf{z}_{\text{mis},1}|\mathbf{Z}_{\text{mis},-1}^{(m-1)}, \hat{\boldsymbol{\theta}}_1^{(m)}) \\
&\vdots \\
\hat{\boldsymbol{\theta}}_p^{(m)} &\sim f(\boldsymbol{\theta}_p|\mathbf{z}_{\text{obs},p}, \mathbf{Z}_{\text{obs},-p}^{(m-1)}) \\
\mathbf{z}_{\text{mis},p}^{(m)} &\sim f(\mathbf{z}_{\text{mis},p}|\mathbf{Z}_{\text{mis},-p}^{(m-1)}, \hat{\boldsymbol{\theta}}_p^{(m)}),
\end{aligned}
$$

where $\hat{\boldsymbol{\theta}}_1^{(m)}$ through $\hat{\boldsymbol{\theta}}_p^{(m)}$ are obtained using regularized regression or BLasso regression. Note that the superscript $(m-1)$ in $\mathbf{Z}_{\text{obs},-j}^{(m-1)}$ and $\mathbf{Z}_{\text{mis},-j}^{(m-1)}$ implies that the missing values in $\mathbf{Z}_{\text{obs},-j}$ and $\mathbf{Z}_{\text{mis},-j}$ are filled in using their previous updates; in other words, while the observed data $\mathbf{Z}_{\text{obs}}$ do not change in the iterative updating procedure, the missing data $\mathbf{Z}_{\text{mis}}$ do change from one iteration to another. After convergence, the last M imputed data sets after appropriate thinning are chosen for subsequent standard complete-data analysis.

The MI approach incorporating BLasso regression (Section 2.4) directly simulates the posterior distribution of the unknown parameters and posterior predictive distribution of missing data. Thus, it is straightforward to incorporate this method to simulate $f(\boldsymbol{\theta}_j|\mathbf{z}_{\text{obs},j}, \mathbf{Z}_{\text{obs},-j})$ in the above iterative procedure. We simply need to posit a hierarchical Bayesian model for each conditional distribution.

On the other hand, it is considerably more involved to incorporate the DURR and IURR methods in this iterative procedure.

3 Numerical studies

We conduct numerical studies to evaluate the finite sample performance of the three proposed methods including DURR, IURR, and BLasso in the presence of high-dimensional data. For DURR and IURR, we consider three widely used regularization methods, namely, lasso, EN, and ALasso, with a 10-fold cross-validation used to select tuning parameters. For BLasso, we use the approach proposed by Hans[23] with hyperparameters $(a, b) = (0.1, 0.1)$, $(r, s) = (0.01, 0.01)$, and $(g, h) = (1, 1)$ in (6). We also compare these methods with a standard parametric imputation approach that either is based on the true imputation model or uses all variables in the data set without model trimming. For all MI methods, 30 imputed data sets are generated for subsequent complete-data analysis.

In our numerical studies, we focus on settings where the primary goal is to conduct regression analysis and estimate the regression coefficients, denoted by $\boldsymbol{\beta}$, in the presence of missing data. Specifically, we first conduct imputation using each approach and then estimate $\boldsymbol{\beta}$ using imputed data sets; the regression coefficient estimate $\hat{\boldsymbol{\beta}}$ is then used to compare the performance of different methods. To benchmark bias and loss of efficiency for estimating $\boldsymbol{\beta}$, two additional methods that do not involve imputation are also used in numerical studies: a gold standard (GS) method that estimates $\boldsymbol{\beta}$ using the underlying complete data before missing data are generated and a complete-case (CC) analysis method that estimates $\boldsymbol{\beta}$ using only the complete cases.

3.1 Simulations

In all simulations, the sample size is fixed at $n = 100$. Each simulated data set includes $\mathbf{y}$, the fully observed outcome variable, and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_p)$, the set of predictors and auxiliary variables where $\mathbf{z}_1$ contains missing values and $\mathbf{Z}_{-1}$ is associated with $\mathbf{y}$ and/or missingness of the data. $(\mathbf{z}_2, \ldots, \mathbf{z}_p)$ is first generated from a multivariate normal distribution with mean $(0, \ldots, 0)_{p-1}$ and a first-order autoregressive covariance matrix with autocorrelation ρ varying as 0, 0.5, and 0.9. We consider settings with $p = 200$ and $p = 1000$. For each combination of p and ρ, $\mathbf{z}_1$ is generated from a normal distribution with variance $\sigma^2_{z_1} = 1$ and mean $\mu_{z_1} = \alpha_0 + \mathbf{Z}_{\mathcal{S}}\boldsymbol{\alpha}$, where $\mathcal{S}$ represents the true active set with a cardinality of q. We considered cases where $q = 4$, 20, or 50, and the corresponding design matrices $\mathbf{Z}_{\mathcal{S}} = \{\mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_{50}, \mathbf{z}_{51}\}$, $\{\mathbf{z}_2, \ldots, \mathbf{z}_{11}, \mathbf{z}_{50}, \ldots, \mathbf{z}_{59}\}$ and $\{\mathbf{z}_2, \ldots, \mathbf{z}_{11}, \mathbf{z}_{50}, \ldots, \mathbf{z}_{59}, \mathbf{z}_{70}, \ldots, \mathbf{z}_{79}, \mathbf{z}_{90}, \ldots, \mathbf{z}_{99}, \mathbf{z}_{110}, \ldots, \mathbf{z}_{119}\}$, respectively. To fix the signal-to-noise ratio when generating $\mathbf{z}_1$, α is set to $1' \times 1$, $1' \times \sqrt{0.2}$ and $1' \times \sqrt{0.08}$ for $q = 4, 20,$ or 50, respectively. Given $\mathbf{Z}$, $\mathbf{y}$ is generated from a normal distribution with mean $\mu_y = \beta_0 + \beta_1\mathbf{z}_1 + \beta_2\mathbf{z}_2 + \beta_3\mathbf{z}_3$ $(\beta_j = 1)$ and variance $\sigma^2_y = 3$. Missing values are generated from only one variable $\mathbf{z}_1$ and the missing data indicator δ_1 for $\mathbf{z}_1$ is generated from a logistic model $\text{logit}[Pr(\delta_1 = 1|\mathbf{Z}_{-1}, \mathbf{y})] = -1 - 0.1\mathbf{z}_2 + 2\mathbf{z}_3 - 2\mathbf{y}$, resulting in approximately 30% of $\mathbf{z}_1$ missing.

For each simulated data set, MI is conducted using each method of interest; then the linear regression model for $\mathbf{y}$ is fit using the imputed data sets for $(\mathbf{y}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$ and Rubin's rule is used to obtain $\hat{\boldsymbol{\beta}}$ and their standard errors. When conducting imputation using DURR, IURR, and BLasso, the entire data set $(\mathbf{y}, \mathbf{Z})$ is used. However, since the standard parametric MI procedure cannot be directly used in the cases of $p > n$, we consider two ways to use the standard parametric MI: (1) the true active set $\mathcal{S}$ plus $\mathbf{y}$ are used to impute $\mathbf{z}_1$, denoted by MI-true and (2) 50 (or 80) variables including the true active set $\mathcal{S}$ plus $\mathbf{y}$ are used to impute $\mathbf{z}_1$ in the cases of $q = 4, 20$ (or

$q = 50$), denoted by MI-50 (or MI-80). Note that MI-true is not applicable in practice since the true active set is unknown; MI-50 (or MI-80) is analogous to an approach used in practice that includes an initial screening step to reduce the number of predictors to less than n or r. For MI-true, MI-50, and MI-80, we use the R packages `mice` with its default method (i.e. Bayesian linear regression with a ridge prior) and `mi` with its default method (i.e. Bayesian regression for Gaussian data) and we obtain similar results; thus, we only report the results using `mice`. When using the R package `mice`, we also specify different ridge parameter values (namely 10^{-5}, 10^{-8}, and 10^{-10}) and observe little change in performance; as a result, we present only the results based on a ridge parameter value of 10^{-5}.

The simulation results are summarized for $\hat{\beta}_1$ based on the following measures that are calculated over 500 Monte Carlo data sets: mean bias of $\hat{\beta}_1$, mean standard error of $\hat{\beta}_1$ (SE), Monte Carlo standard deviation of $\hat{\beta}_1$ (SD), mean square error of $\hat{\beta}_1$ (MSE), and coverage rate of the 95% confidence interval of $\hat{\beta}_1$ (CR). In addition to compare different methods, we also investigate the effect of the dimension of the data (p), the correlation among the data (ρ), and the size of the true active set for imputation (q). Note that when $\rho = 0$ (i.e. the predictors are independent of each other), the IURR method is expected to perform well since the regularization methods are known to achieve model selection consistency under such independence condition.

The simulation results are presented in Tables 1 to 3 for $q = 4$, 20, and 50, respectively; within each table different methods are compared and the effects of correlation ρ and dimension p are evaluated with q fixed. In all scenarios, the naive approach (CC) and the standard MI method that does not use the true active set (i.e. MI-50 or MI-80) lead to considerable to substantial bias and larger MSE, underperforming IURR and BLasso; in addition, both methods underperform DURR in most cases except when $p = 1000$, $\rho = 0.9$ and $q = 20$ or 50.

When the size of the true active set is small, that is, $q = 4$ (Table 1), BLasso exhibits negligible bias and its CR is close to the nominal level; its performance is comparable or better than that of the DURR and IURR methods in all cases. In addition, the IURR methods exhibit smaller bias and MSE compared with their DURR counterparts for all values of ρ and p. Within IURR or DURR, ALasso tends to achieve better performance than lasso and EN in terms of bias and MSE, which is more pronounced when $\rho = 0$ or 0.5. When correlation is high, all IURR methods perform reasonably well with EN and ALasso exhibiting negligible bias and small MSE. IURR and BLasso achieve clearly better performance in the case of $\rho = 0.9$ compared with the cases of $\rho = 0$ or 0.5, whereas DURR shows somewhat mixed results. This suggests that when variables are strongly correlated the variables that are selected to impute missing values, although may not be identical to the true active set, still provide sufficient information for imputation, which leads to improved performance of IURR and BLasso. As the dimension of data increases from $p = 200$ to 1000, the performance of each method tends to deteriorate with BLasso showing the least amount of deterioration.

Compared with Table 1 ($q = 4$), Tables 2 ($q = 20$) and 3 ($q = 50$) show similar patterns on comparisons among the imputation methods. BLasso is shown to achieve overall better performance compared to DURR and IURR and its performance is comparable to MI-true (the standard MI using the true active set) when $q = 4$ or $q = 20$ and better than MI-true when $q = 50$. Several new trends also emerge as the size of the active set (q) increases. When $q = 20$ or 50, the impact of ρ on the performance of IURR becomes somewhat mixed; in particular under $p = 1000$, IURR tends to perform better in the case of $\rho = 0$ compared with $\rho = 0.5$ or 0.9. As q increases, the performance of each imputation method deteriorates with larger bias and MSE and such deterioration is considerably more pronounced with IURR and DURR than with BLasso. While MI-true achieves satisfactory performance when q is small to moderate, it exhibits substantial bias when $q = 50$ whereas the performance of BLasso remains satisfactory, indicating that even in the

Table 1. Simulation results for estimating $\beta_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets, where $n = 100$ and $q = 4$.

		$\rho = 0.0$					$\rho = 0.5$					$\rho = 0.9$				
		Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
	GS	0.001	0.101	0.108	0.012	0.932	−0.007	0.089	0.091	0.008	0.938	−0.003	0.081	0.077	0.006	0.954
	CC	−0.206	0.123	0.123	0.058	0.630	−0.176	0.112	0.114	0.044	0.648	−0.173	0.105	0.102	0.040	0.598
	MI-true	−0.008	0.115	0.116	0.014	0.938	−0.005	0.100	0.097	0.009	0.954	−0.008	0.090	0.087	0.008	0.960
	MI-50	−0.356	0.154	0.164	0.154	0.378	−0.333	0.145	0.153	0.134	0.394	−0.300	0.138	0.141	0.110	0.406
							DURR									
	Lasso	0.074	0.129	0.131	0.023	0.894	0.073	0.105	0.107	0.017	0.896	0.051	0.096	0.101	0.013	0.924
$p = 200$	EN	0.066	0.137	0.135	0.022	0.922	0.075	0.110	0.110	0.018	0.892	0.056	0.098	0.104	0.014	0.916
	ALasso	0.056	0.120	0.118	0.017	0.918	0.053	0.100	0.106	0.014	0.906	0.025	0.091	0.098	0.010	0.922
	Lasso	0.066	0.146	0.145	0.025	0.904	0.093	0.117	0.128	0.025	0.838	0.079	0.097	0.105	0.017	0.862
$p = 1000$	EN	0.028	0.158	0.152	0.024	0.952	0.075	0.130	0.134	0.024	0.896	0.079	0.103	0.110	0.018	0.866
	ALasso	0.050	0.134	0.129	0.019	0.924	0.069	0.107	0.115	0.018	0.872	0.050	0.093	0.103	0.013	0.900
							IURR									
	Lasso	0.017	0.111	0.116	0.014	0.930	0.013	0.097	0.108	0.012	0.912	−0.006	0.091	0.100	0.010	0.924
$p = 200$	EN	0.002	0.118	0.123	0.015	0.930	0.010	0.102	0.105	0.011	0.930	−0.017	0.093	0.097	0.010	0.930
	ALasso	0.008	0.107	0.110	0.012	0.942	0.004	0.094	0.104	0.011	0.926	−0.003	0.088	0.092	0.008	0.940
	Lasso	0.030	0.113	0.127	0.017	0.918	0.044	0.098	0.112	0.015	0.894	0.019	0.090	0.112	0.013	0.906
$p = 1000$	EN	0.020	0.126	0.134	0.018	0.926	0.037	0.107	0.113	0.014	0.904	0.003	0.094	0.110	0.012	0.934
	ALasso	0.006	0.108	0.112	0.013	0.940	0.002	0.094	0.096	0.009	0.952	0.001	0.086	0.086	0.007	0.944
$p = 200$	BLasso	−0.023	0.115	0.120	0.015	0.946	−0.005	0.098	0.098	0.010	0.950	−0.001	0.090	0.092	0.009	0.952
$p = 1000$	BLasso	−0.030	0.118	0.123	0.016	0.934	−0.009	0.101	0.102	0.010	0.950	−0.005	0.092	0.094	0.009	0.948

Bias: mean bias of $\hat{\beta}_1$; SE: mean standard error of $\hat{\beta}_1$; SD: Monte Carlo standard deviation of $\hat{\beta}_1$; MSE: mean square error of $\hat{\beta}_1$; CR: coverage rate of 95% confidence interval for $\hat{\beta}_1$; GS: gold standard; CC: complete-case; EN: elastic net; DURR: direct use of regularized regression; IURR: indirect use of regularized regression; ALasso: adaptive lasso; BLasso: Bayesian lasso.

case that the true active set is known it is still advantageous to incorporate regularization in fitting imputation models, especially when the size of the true active set (q) is large.

We conduct additional simulations in settings that are similar to those in Tables 1 to 3 but have three variables ($\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_3$) with missing values. The findings on comparisons of different imputation methods are consistent with those reported in Tables 1 to 3. However, the computational cost for our proposed approaches, in particular, the BLasso imputation, is considerably higher in this set of simulations.

3.2 Data from a cancer study

The proposed methods are further compared using gene expression data collected from a cancer study. The data set includes $n = 200$ subjects, for which expression data for 1036 gene probe sets are obtained through microarray experiments, and we refer to them as gene biomarkers. We randomly pick one gene biomarker as the outcome variable of interest, $\mathbf{y}$, and four biomarkers as predictors of interest, $(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4)$. The main goal of the analysis is to fit the regression model: $E(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_4) = \beta_0 + \beta_1\mathbf{z}_1 + \cdots + \beta_4\mathbf{z}_4$. The remaining biomarkers are denoted by $(\mathbf{z}_5, \mathbf{z}_6, \ldots, \mathbf{z}_{1035})$, noting that the order in which the biomarkers are arranged is randomly chosen.

Table 2. Simulation results for estimating $\beta_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets, where $n = 100$ and $q = 20$.

		$\rho = 0.0$					$\rho = 0.5$					$\rho = 0.9$				
		Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
	GS	−0.005	0.082	0.083	0.007	0.958	−0.002	0.056	0.054	0.003	0.948	−0.001	0.039	0.038	0.001	0.960
	CC	−0.224	0.104	0.103	0.061	0.412	−0.163	0.081	0.082	0.033	0.484	−0.093	0.063	0.064	0.013	0.700
	MI-true	−0.046	0.097	0.093	0.011	0.940	−0.025	0.070	0.073	0.006	0.916	−0.018	0.052	0.049	0.003	0.938
	MI-50	−0.264	0.135	0.131	0.087	0.530	−0.201	0.113	0.133	0.058	0.606	−0.156	0.101	0.105	0.035	0.718
							DURR									
	Lasso	0.053	0.129	0.136	0.021	0.886	0.117	0.084	0.085	0.021	0.718	0.050	0.054	0.056	0.006	0.872
$p = 200$	EN	0.055	0.132	0.138	0.022	0.896	0.134	0.085	0.088	0.026	0.662	0.059	0.054	0.058	0.007	0.824
	ALasso	0.018	0.129	0.133	0.018	0.932	0.072	0.083	0.081	0.012	0.904	0.038	0.054	0.054	0.004	0.910
	Lasso	0.032	0.152	0.154	0.025	0.914	0.149	0.098	0.101	0.032	0.672	0.110	0.060	0.065	0.016	0.572
$p = 1000$	EN	0.019	0.156	0.152	0.023	0.936	0.157	0.106	0.106	0.036	0.688	0.131	0.062	0.070	0.022	0.462
	ALasso	0.006	0.153	0.142	0.020	0.978	0.102	0.097	0.088	0.018	0.834	0.062	0.060	0.061	0.008	0.868
							IURR									
	Lasso	−0.012	0.108	0.116	0.014	0.918	0.035	0.070	0.079	0.007	0.900	0.002	0.047	0.052	0.003	0.922
$p = 200$	EN	−0.024	0.112	0.130	0.017	0.932	0.026	0.073	0.079	0.007	0.922	−0.001	0.050	0.051	0.003	0.936
	ALasso	−0.019	0.106	0.110	0.012	0.934	0.022	0.071	0.080	0.007	0.908	0.006	0.047	0.054	0.003	0.914
	Lasso	−0.003	0.117	0.125	0.016	0.934	0.079	0.078	0.104	0.017	0.790	0.038	0.047	0.059	0.005	0.824
$p = 1000$	EN	−0.023	0.126	0.137	0.019	0.946	0.081	0.085	0.106	0.018	0.802	0.046	0.051	0.060	0.006	0.838
	ALasso	−0.024	0.108	0.119	0.015	0.924	0.033	0.078	0.085	0.008	0.920	0.017	0.047	0.056	0.003	0.902
$p = 200$	BLasso	−0.037	0.114	0.112	0.014	0.942	−0.003	0.078	0.075	0.006	0.954	0.003	0.051	0.055	0.003	0.936
$p = 1000$	BLasso	−0.056	0.126	0.116	0.017	0.956	0.004	0.088	0.082	0.007	0.962	0.006	0.056	0.059	0.004	0.932

Bias: mean bias of $\hat{\beta}_1$; SE: mean standard error of $\hat{\beta}_1$; SD: Monte Carlo standard deviation of $\hat{\beta}_1$; MSE: mean square error of $\hat{\beta}_1$; CR: coverage rate of 95% confidence interval for $\hat{\beta}_1$; GS: gold standard; CC: complete-case; EN: elastic net; DURR: direct use of regularized regression; IURR: indirect use of regularized regression; ALasso: adaptive lasso; BLasso: Bayesian lasso.

There are no missing data in the original data set, so we adopt the following scheme to compare different imputation methods. About 500 pseudo data sets with a sample size of $n = 200$ are generated from the original data set through randomly sampling $n = 200$ observations with replacement. For each pseudo data set, we generate missing values for $\mathbf{z}_1$ with the missing indicator δ following a logistic model: $\text{logit}[Pr(\delta = 1)] = 1 + y + z_4 - 2z_5$. The proposed imputation methods are applied to each pseudo data set with missing data and subsequently a linear regression analysis is conducted to estimate $\boldsymbol{\beta}$ based on the imputed data sets. For comparison, we also use the naive approach based on the CC in each pseudo data and a GS approach based on the analysis of the original complete data. The estimates of $\boldsymbol{\beta}$ averaged over 500 pseudo data sets are used to evaluate the performance of different imputation methods. It is noteworthy that unlike the simulation studies the true imputation model for the cancer data is unknown; however, our analysis scheme still allows us to access the underlying complete data and hence compare the imputation methods with the GS approach to benchmark their performance.

Similar to the simulation studies, Table 4 presents only the results for estimating β_1. In addition, the computational cost (3.4 GHz CPU, 8GB Memory, Windows System) for the MI methods is also provided. Consistent with the simulation results, BLasso again achieves best performance with its estimate being closest to the GS estimate, and the CC analysis exhibits substantial bias compared with

Table 3. Simulation results for estimating $\beta_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets, where $n = 100$ and $q = 50$.

		$\rho = 0.0$					$\rho = 0.5$					$\rho = 0.9$				
		Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
	GS	−0.003	0.080	0.078	0.006	0.954	−0.003	0.053	0.055	0.003	0.940	0.001	0.031	0.031	0.001	0.958
	CC	−0.217	0.102	0.102	0.058	0.408	−0.166	0.078	0.084	0.035	0.470	−0.087	0.054	0.057	0.011	0.628
	MI-true	−0.273	0.136	0.136	0.093	0.502	−0.221	0.118	0.136	0.067	0.568	−0.123	0.097	0.102	0.025	0.846
	MI-80	−0.222	0.102	0.110	0.061	0.406	−0.172	0.080	0.088	0.037	0.406	−0.087	0.057	0.083	0.014	0.686
								DURR								
	Lasso	0.032	0.134	0.129	0.018	0.946	0.113	0.090	0.093	0.022	0.760	0.069	0.053	0.052	0.007	0.788
$p = 200$	EN	0.036	0.135	0.130	0.018	0.948	0.130	0.092	0.096	0.026	0.718	0.084	0.053	0.053	0.010	0.664
	ALasso	0.014	0.136	0.126	0.016	0.952	0.082	0.090	0.086	0.014	0.860	0.053	0.055	0.052	0.006	0.874
	Lasso	0.011	0.162	0.154	0.024	0.942	0.130	0.111	0.108	0.029	0.758	0.155	0.067	0.077	0.030	0.354
$p = 1000$	EN	0.003	0.163	0.151	0.023	0.960	0.135	0.119	0.111	0.031	0.782	0.195	0.070	0.084	0.045	0.170
	ALasso	0.007	0.163	0.151	0.023	0.952	0.089	0.110	0.098	0.017	0.870	0.110	0.067	0.064	0.016	0.678
								IURR								
	Lasso	−0.012	0.111	0.111	0.012	0.950	0.030	0.075	0.094	0.010	0.926	0.010	0.041	0.050	0.003	0.896
$p = 200$	EN	−0.030	0.118	0.126	0.017	0.938	0.026	0.080	0.101	0.011	0.916	0.002	0.046	0.053	0.003	0.922
	ALasso	−0.031	0.108	0.116	0.014	0.930	0.023	0.077	0.081	0.007	0.924	0.020	0.040	0.052	0.003	0.876
	Lasso	−0.024	0.118	0.131	0.018	0.930	0.071	0.087	0.094	0.014	0.838	0.065	0.044	0.065	0.008	0.650
$p = 1000$	EN	−0.036	0.124	0.128	0.018	0.942	0.066	0.097	0.103	0.015	0.874	0.066	0.052	0.083	0.011	0.712
	ALasso	−0.040	0.110	0.120	0.016	0.936	0.021	0.083	0.091	0.009	0.910	0.021	0.083	0.091	0.009	0.910
$p = 200$	BLasso	−0.055	0.116	0.109	0.015	0.944	0.000	0.081	0.076	0.006	0.962	0.005	0.048	0.048	0.002	0.940
$p = 1000$	BLasso	−0.079	0.131	0.121	0.021	0.936	−0.002	0.090	0.084	0.007	0.962	0.016	0.057	0.060	0.004	0.936

Bias: mean bias of $\hat{\beta}_1$; SE: mean standard error of $\hat{\beta}_1$; SD: Monte Carlo standard deviation of $\hat{\beta}_1$; MSE: mean square error of $\hat{\beta}_1$; CR: coverage rate of 95% confidence interval for $\hat{\beta}_1$; GS: gold standard; CC: complete-case; EN: elastic net; DURR: direct use of regularized regression; IURR: indirect use of regularized regression; ALasso: adaptive lasso; BLasso: Bayesian lasso.

GS. For each regularization method, DURR and IURR perform similarly. Furthermore, the results show that ALasso underperforms Lasso and EN within DURR and this trend is not as pronounced within IURR. As for computational time, BLasso is more computationally intensive than DURR and IURR, though its computational time is still acceptable. DURR requires longer computational time than IURR, which is expected since DURR requires fitting regularized regression in each of the M imputed data sets, whereas IURR only requires fitting regularized regression once.

4 Discussion

Since MI is inherently Bayesian, the BLasso regression is a natural fit for multiply imputing missing values in the presence of high-dimensional data. Our numerical results show that the BLasso imputation achieves, in most cases, better performance than the other imputation methods including several existing imputation methods. In addition, the DURR and IURR methods have two other disadvantages. First, it is not straightforward to extend the DURR and IURR approaches to the case of general missing pattern, where the use of bootstrap in DURR further complicates matters. Second, it has been shown that the standard residual bootstrap method, while works for ALasso, may not work for lasso.[29,30] Even though the DURR approach uses the bootstrap that

Table 4. Estimation of β_1 using the MI methods, gold standard approach and CC analysis for the cancer data example.

Method	Estimate	SE	SD	TIME (min)
GS	0.394	0.292		
CC	0.622	0.392	0.431	
		DURR		
Lasso	0.402	0.327	0.349	0.262
EN	0.412	0.331	0.357	0.264
ALasso	0.331	0.306	0.267	6.322
		IURR		
Lasso	0.408	0.303	0.353	0.108
EN	0.388	0.300	0.335	0.120
ALasso	0.350	0.305	0.257	0.240
BLasso	0.390	0.299	0.323	4.804

For the MI methods and CC, SE: mean standard error of $\hat{\beta}_1$; SD: Monte Carlo standard deviation of $\hat{\beta}_1$; TIME: mean computational time for one data set; GS: gold standard; CC: complete-case; EN: elastic net; DURR: direct use of regularized regression; IURR: indirect use of regularized regression; ALasso: adaptive lasso; BLasso: Bayesian lasso.

resample the cases rather than residuals, it is also likely plagued by the problem inherent with the standard residual bootstrap method, likely contributing to its poor performance in the numerical studies. In summary, our results suggest that the BLasso regression and its extensions are better suited for MI in the presence of high-dimensional data than the other regression methods.

Many researchers[31–35] have investigated optimal strategies for MI in the presence of a large number of variables and in particular, the question that how many variables should be included in imputation models in such cases. Van Buuren et al.[31] suggested to select no more than 15–25 variables for imputation purposes. More recently, Hardt et al.[35] investigated the same question in small sample research; they used the existing MI methods that are implemented in the R package mice[3] and arrived at a rule of thumb that the ratio of variables used for imputation to the number of complete cases should not be more than 1:3. It is worth noting that both advices were derived based on the classical regression techniques for fitting imputation model that were available at the time. Consistent with recommendations by others,[14,15] van Buuren et al.[31] also stated that the generally accepted principle for imputation "implies that the number of predictors should be as large as possible." However, the performance of the classical regression techniques is known to deteriorate as p increases; as a result, if p is large, it is generally not feasible to include all available variables in imputation models when the classical regression techniques are used for imputation. Consequently, while their advices likely work well in the low-dimensional setting ($n > p$) when the classical regression techniques are used, their applicability to the high-dimensional setting ($p > n$ or $p \gg n$) as investigated in the current work may be questionable. As mentioned previously, state-of-the-art modern regression techniques have been developed in recent years including those investigated in our paper, allowing one to include all available predictors in imputation models and achieve simultaneous predictor selection and parameter estimation when fitting imputation models. These new regression techniques open the door for us to conduct imputation in the high-dimensional setting ($p > n$ or $p \gg n$). Our research fills this gap and provides some initial insights on applying modern regression techniques to imputation. Compared

with the existing MI approaches based on the classical regression techniques, the main advantages of the proposed methodology are as follows: (1) it is directly applicable to both low-dimensional and high-dimensional data and (2) it is a principled approach achieving simultaneous predictor selection and parameter estimation in imputation models and does not rely on heuristic rules of thumb for predictor selection.

One limitation of the proposed approaches, in particular, the BLasso approach, is their computational cost. While their computational cost is acceptable for the settings considered in this work, it can become computationally very expensive or even infeasible to apply them in the settings where there are a large number of variables and many of these variables have missing values. Additional research is needed for development of efficient algorithms for the proposed approaches and for extensive evaluations of their performance in situations with more general missing patterns. For example, one potential extension is to embed the iterative procedure in Section 2.5 as part of the BLasso imputation.

Our numerical studies focus on continuous data that are normally distributed, a potentially restrictive, unrealistic assumption in practice. The sampling (or posterior) distribution of $\hat{\boldsymbol{\theta}}$ in the imputation model is obtained from $f(\boldsymbol{\theta}|\mathbf{Z}_{\text{obs}})$ in Model (1), which is likely insensitive to the assumption of multivariate normality if there are enough data and in particular, the regularization methods such as ALasso are used. However, imputing $\mathbf{Z}_{\text{mis}}$ is based on the conditional distribution $f(\mathbf{Z}_{\text{mis}}|\mathbf{Z}_{\text{obs}}, \boldsymbol{\theta})$ in Model (1), which is likely sensitive to the normality assumption. To the best of our knowledge, this issue has not been investigated in the context of imputation. Future work is needed to assess the robustness of the proposed imputation methods to violations of the normality assumption and extend the proposed methodology to other types of data such as count or discrete data.

In parallel with existing robust imputation approaches, nonparametric regression approaches such as regularized additive models[36] or boosting[37] may also be adapted to multiply impute missing values in the presence of high-dimensional data, which can relax some modeling assumptions. On a related note, a direct application of classical nonparametric regression techniques such as K-nearest neighbors[38] is unlikely to be satisfactory for imputation in the presence of high-dimensional data due to several well-known issues; one major issue is the curse of dimensionality, that is, the convergence rate of the resulting estimator in nonparametric regression decreases as the dimension of the predictors (p) increases, which is particularly acute in the presence of high-dimensional data.

Acknowledgments

The authors thank the editor and two referees for their suggestions that helped greatly improve the article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the PCORI.

Funding

This work was supported in part by a PCORI contract (ME-1303-5840).

Conflict of interest

None declared.

References

1. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley, 1987.
2. Little RJA and Rubin DB. *Statistical analysis with missing data*. New York, NY: Wiley, 2002.
3. van Buuren S and Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011; **45**(3): 1–67.
4. Su Y, Gelman A, Hill J, et al. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J Stat Softw* 2011; **45**(2): 1–31.
5. Harel O and Zhou XH. Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies. *Stat Med* 2007; **26**: 2370–2388.
6. He Y, Yucel R and Raghunathan TE. A functional multiple imputation approach to incomplete longitudinal data. *Stat Med* 2011; **30**: 1137–1156.
7. Little R and An H. Robust likelihood-based analysis of multivariate data with missing values. *Stat Sin* 2004; **14**: 949–968.
8. Long Q, Zhang X and Johnson BA. Robust estimation of area under ROC curve using auxiliary variables in the presence of missing biomarker values. *Biometrics* 2010; **67**: 559–567.
9. Long Q, Hsu CH and Li Y. Doubly Robust Nonparametric Multiple Imputation for Ignorable Missing Data. *Stat Sin* 2012; **22**: 149–172.
10. Qi L, Wang YF and He Y. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Stat Med* 2010; **29**: 2592–2604.
11. Stuart EA, Azur M, Frangakis C, et al. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *Am J Epidemiol* 2009; **169**: 1133–1139.
12. Zhang G and Little R. Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics* 2008; **65**: 911–918.
13. Harel O and Zhou X. Multiple imputation: review of theory, implementation and software. *Stat Med* 2007; **26**: 3055–3057.
14. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci* 1994; **9**: 538–558.
15. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; **91**: 473–489.
16. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Control* 1974; (19): 716–723.
17. Hoerl AE and Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**: 55–67.
18. Tibshirani RJ. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* 1996; **58**: 267–288.
19. Zou H and Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc SerB* 2005; **67**: 301–320.
20. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.
21. Park T and Casella G. The Bayesian lasso. *J Am Stat Assoc* 2008; **103**: 681–686.
22. Hans C. Bayesian lasso regression. *Biometrika* 2009; **96**: 835–845.
23. Hans C. Model uncertainty and variable selection in Bayesian lasso regression. *Stat Comput* 2010; **20**: 221–229.
24. Heitjan DF and Little RJ. Multiple imputation for the fatal accident reporting system. *Appl Stat* 1991; **40**: 13–29.
25. Rubin DB and Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Stat Med* 2006; **10**: 585–598.
26. Kyung M, Gill J, Ghosh M, et al. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal* 2010; **5**: 369–412.
27. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007; **16**: 219–242.
28. van Buuren S, Brand JPL, Groothuis-Oudshoorn C, et al. Fully conditional specification in multivariate imputation. *J Stat Comput Simul* 2006; **76**: 1049–1064.
29. Chatterjee A and Lahiri SN. Bootstrapping lasso estimators. *J Am Stat Assoc* 2011; **106**: 608–625.
30. Chatterjee A and Lahiri S. Asymptotic properties of the residual bootstrap for Lasso estimators. *Proc Am Math Soc* 2010; **138**: 4497–4509.
31. van Buuren S, Boshuizen HC, Knook DL, et al. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; **18**: 681–694.
32. Collins L, Schafer J and Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001; **6**: 330–351.
33. Hoo J. The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. *Educ Psychol Meas* 2009; **69**: 929–947.
34. White I, Royston P and Wood A. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; **30**: 377–399.
35. Hardt J, Herke M and Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Med Res Methodol* 2012; **12**: 184http://www.biomedcentral.com/1471-2288/12/184).
36. Huang J, Horowitz JL and Wei F. Variable selection in nonparametric additive models. *Annal Stat* 2010; **38**: 2282–2313.
37. Bühlmann P and Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci* 2007; **22**: 477–505.
38. Devroye LP and Wagner T. The strong uniform consistency of nearest neighbor density estimates. *Annal Stat* 1977; **5**: 536–540.
39. Stone CJ. Optimal rates of convergence for nonparametric estimators. *Annal Stat* 1980; **8**: 1348–1360.