


Variable selection for high-dimensional incomplete data using horseshoe estimation with data augmentation

Yunxi Zhang & Soeun Kim



To cite this article: Yunxi Zhang & Soeun Kim (2023): Variable selection for high-dimensional incomplete data using horseshoe estimation with data augmentation, Communications in Statistics - Theory and Methods, DOI: [10.1080/03610926.2023.2177107](https://doi.org/10.1080/03610926.2023.2177107)

To link to this article: <https://doi.org/10.1080/03610926.2023.2177107>

 View supplementary material 

 Published online: 23 Feb 2023.

 Submit your article to this journal 

 View related articles 

 View Crossmark data 



Variable selection for high-dimensional incomplete data using horseshoe estimation with data augmentation

Yunxi Zhang^a and Soeun Kim^b

^aDepartment of Data Science, University of Mississippi Medical Center, Jackson, USA; ^bDepartment of Mathematics, Physics, Statistics, Azusa Pacific University, Azusa, USA

ABSTRACT

Bayesian shrinkage methods have been widely employed to perform variable selection with high-dimensional data. However, the presence of missing data hinders the implementation of these methods. Since complete case analyses can lead to biased estimates, applicable and efficient methods of variable selection with imputation are needed to obtain valid results. In order to address this issue, we propose an algorithm that employs the horseshoe shrinkage prior for shrinkage and multiple imputation for missing data in high-dimensional settings with a practical suggestion on model selection decision strategy. Simulation studies and real data analyses are presented and compared with those of other possible approaches. The simulation results suggest that the proposed algorithm can be considered as a general strategy for model selection of incomplete continuous data.

ARTICLE HISTORY

Received 19 January 2022
Accepted 31 January 2023

KEYWORDS



High-dimensional data; missing data; multiple imputation; Bayesian shrinkage; variable selection

1. Introduction

Selecting a subset of variables for linear regression models from high-dimensional data in which the number of variables p exceeds the number of observations n has garnered widespread research attention. Although many variable selection methods have been developed, most assume that all variables in the specified model are completely observed. However, this is often not the case in practice, where data sets inevitably include missing data due to incomplete cases. The default option for variable selection methods implemented in statistical software for linear regression is to simply drop or delete incomplete cases; however, deleting all incomplete cases can give unreliable inference for missing data because the remaining complete cases are less likely to represent the target population. Given that high-dimensional data may lack any complete cases, a variable selection method that appropriately handles missing data without sacrificing useful information is needed for this scenario in particular to avoid incorrect selection and parameter estimation in presence of high-dimensional data.

Consider data with n observations of an outcome variable y_i and predictor variables $X_i = (X_{i1}, \dots, X_{ip})$ with dimension p , where $n \ll p$. This relationship can be expressed with the following linear regression model

$$y_i = \beta_0 + \sum_{j=1}^p X_j \beta_j + \varepsilon_i, i = 1, \dots, n, \quad (1)$$

CONTACT Soeun Kim  soeun.s.kim@gmail.com  Department of Mathematics, Physics Statistics Azusa Pacific University Azusa USA.

Supplemental data for this article can be accessed online at <https://doi.org/10.1080/03610926.2023.2177107>.

© 2023 Taylor & Francis Group, LLC

where β_0 is the intercept term, $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^p$ is the set of regression coefficients corresponding to each predictor variable $\mathbf{X}_j, j = 1, \dots, p$, and $\boldsymbol{\varepsilon} = \{\varepsilon_i\}_{i=1}^n$ are the residuals following normal distribution $N(0, \sigma_\varepsilon^2)$.

One commonly employed algorithm for handling high-dimensional data is the lasso (Tibshirani 1996). It was originally solved using the least angle regression (LARS) algorithm under the frequentist framework. With the l_1 -regularization, lasso shrinks the coefficients of noise variables to exactly zero. Cross-validation is commonly conducted with the frequentist lasso to give the optimized estimates of $\boldsymbol{\beta}$.

In Bayesian estimation, the horseshoe prior, a scaled mixture of a global shrinkage parameter and a set of local shrinkage parameters with priors following the half-Cauchy distribution, was developed to perform shrinkage on high-dimensional data (Carvalho, Polson, and Scott 2009, 2010). The heavy tail in the half-Cauchy distribution is robust to sparse variables, and this prior has been widely discussed and applied in regression models. For example, a fast sampling strategy in high-dimensional regression was based on Gaussian sampling with the horseshoe prior to avoid the Cholesky factorization, which is often computationally inefficient (Bhattacharya, Chakraborty, and Mallick 2016). The Gibbs sampling approach for zero-intercept linear regression model on standardized data by decomposing the half-Cauchy distribution following its association with inverse-Gamma distribution in mean field variational Bayes (Makalic and Schmidt 2016; Wand et al. 2011).

In practice, multiple imputation (Donald B. Rubin 1976) has become one of the most widely used methods for handling missing data. By reflecting both between- and within-imputation variability, multiple imputation represents both uncertainty due to values being missing and estimation uncertainty that would arise without missing data. Multiple imputation can be implemented using data augmentation (Tanner and Wong 1987; Schafer 1997; van Dyk and Meng 2001), a Markov Chain Monte Carlo (MCMC) iterative simulation algorithm where unobserved values are filled in by drawing samples from conditional predictive distributions.

Variable selection with incomplete data on a modest number of predictors of interest (with $n > p$) has been the focus of several studies, proposing selection procedures. Garcia, Ibrahim, and Zhu (2010) proposed a variable selection procedure within a penalized-likelihood framework. Moreover, Yang, Belin, and Boscardin (2005) employed data augmentation for imputation in combination with stochastic search variable selection (George and McCulloch 1993), suggesting that it is preferable for MCMC iterations to incorporate both imputation and variable selection rather than first performing imputation and then selecting variables. In a case study of variable selection for multistate models with time-to-event outcomes, Beesley and Taylor (2021) provided a practical strategy of employing fully conditional specification (Bartlett et al. 2015) for imputation with several high-dimensional variable selection techniques including the horseshoe priors. Though the fully conditional specification is known for its compatibility, being outside of the Bayesian framework, it cannot be blended with the Gibbs samplers of horseshoe priors in the MCMC procedure.

Despite these advances in statistical practice, variable selection with missing data imputation for high-dimensional continuous variables remains a challenge. With a relatively large number of predictors p , it is difficult to use methods that incorporate joint-modeling assumptions for a sizable number of covariates, as the implied covariance matrix is apt to be singular for high-dimensional data. One strategy to avoid singularity is to resample the

complete cases. Long and Johnson (2015) proposed an approach for selection and imputation by bootstrap. A limitation of resampling-based imputation is that it requires at least one complete case in the data set, and this requirement might not be met in high-dimensional data sets. Moreover, Liu et al. (2016) employed multiple imputation to handle missing values and lasso procedure with cross-validation to perform variable selection. Carvalho, Polson, and Scott (2009) discussed imputation for Gaussian graphical model in its illustration of the horseshoe prior without an extension to the linear regression model. Liang et al. (2018) proposed an imputation-conditional consistency algorithm giving several applications and extensions, but numerical studies were limited to scenarios with missing completely at random (MCAR) mechanism, where the probability of missing is independent. It is necessary to consider data with more realistic assumptions, like missing at random (MAR) where the probability of missing depends on observed data (Donald B. Rubin 1976).

Given that access to high-dimensional data has expanded enormously, this article aims to develop an algorithm for variable selection with imputation using high-dimensional continuous data with incomplete predictor variables. To accomplish this, we mix data augmentation and Bayesian shrinkage estimation with horseshoe priors in an MCMC procedure and propose the data augmentation with horseshoe estimation (DA-HS) to simultaneously perform variable selection with multiple imputation for data with incomplete predictor variables. The DA-HS provides a practical solution to high-dimensional variable selection with expected accurate coefficient estimations and variable selection for data with or without complete cases. We conduct extensive simulation studies to compare DA-HS with most popular approaches in dealing with variable selection of incomplete high-dimensional data. The feasibility of the methods is illustrated through applications to observational data from World Bank Group (2021) and genetic data from Riboflavin study (Lee et al. 2001; Carvalho, Polson, and Scott 2010).

2. Methods

2.1. Variable selection using horseshoe estimator

Building on the linear regression model in Equation (1), we apply “horseshoe” prior distributions to the coefficients β as in Makalic and Schmidt (2016) while assuming a Jeffreys prior distribution for the residual variance. The corresponding model specification can be written in Bayesian hierarchical form:

$$\begin{aligned} y_i | \beta_0, \beta, \sigma_\varepsilon &\sim N \left(\beta_0 + \sum_{j=1}^p \mathbf{X}_j \beta_j, \sigma_\varepsilon^2 \right), \\ \beta_j | \lambda_j^2, \tau^2, \sigma_\varepsilon^2 &\sim N \left(0, \lambda_j^2 \tau^2 \sigma_\varepsilon^2 \right), \\ \sigma_\varepsilon^2 &\sim \sigma_\varepsilon^{-2} d\sigma_\varepsilon^2, \\ \lambda_j &\sim C^+(0, 1), \\ \tau &\sim C^+(0, 1). \end{aligned} \quad (2)$$

Here, λ_j and τ are the local and global shrinkage parameters, respectively, and $C^+(0, 1)$ is the standard half-Cauchy distribution with probability density function $f(t) = \frac{2}{\pi(1+t^2)}$ $I\{t > 0\}$. Makalic and Schmidt (2016) introduced auxiliary variable for each shrinkage

parameter following half-Cauchy distribution so that $\lambda_j^2 | v_j \sim IG(1/2, v_j^{-1})$, $\tau^2 | \xi \sim IG(1/2, \xi^{-1})$ and $v_j, \xi \sim IG(1/2, 1)$, where $IG(a, b)$ is the inverse-Gamma distribution, and a and b are shape and scale parameters respectively.

The hierarchical model described in Equation (2) is designed only for data with standardized predictor variables. Because we will use a Bayesian regression model for imputation, we keep the variance of each variable without scaling before model fitting. We adjust scales in the model for each local shrinkage parameter $\lambda_j, j = 1, \dots, p$, as suggested by Piironen and Vehtari (2017) as $\lambda_j \sim C^+(0, \sigma_j^{-2})$, where σ_j^2 is the variance of \mathbf{X}_j .

Then auxiliary variables for each shrinkage parameter are used to decompose and simplify sampling from half-Cauchy distributions. We then can express the Bayesian regression model as

$$\begin{aligned} y_i | \boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma_\varepsilon &\sim N\left(\boldsymbol{\beta}_0 + \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\beta}_j, \sigma_\varepsilon^2\right), \\ \boldsymbol{\beta}_j | \lambda_j^2, \tau^2, \sigma_\varepsilon^2 &\sim N\left(0, \lambda_j^2 \tau^2 \sigma_\varepsilon^2\right), \\ \sigma_\varepsilon^2 &\sim \sigma_\varepsilon^{-2} d\sigma_\varepsilon^2, \\ \lambda_j^2 | v_j &\sim IG(1/2, v_j^{-1}), v_j \sim IG(1/2, \sigma_j^4), \\ \tau^2 | \xi &\sim IG(1/2, \xi^{-1}), \xi \sim IG(1/2, 1). \end{aligned} \quad (3)$$

Here, $v_j, j = 1, \dots, p$, and ξ are the auxiliary variables. We then derive the full conditional posterior distributions of the hierarchical regression parameters as:

$$\begin{aligned} \boldsymbol{\beta} | \mathbf{y}, \mathbf{X}, \sigma_\varepsilon^2, \boldsymbol{\Lambda}, \tau^2 &\sim N\left(\left[\mathbf{X}^T \mathbf{X} + \tilde{\tau}^{-2} \tilde{\boldsymbol{\Lambda}}^{-1}\right]^{-1} \mathbf{X}^T \mathbf{y}, \tilde{\sigma}_\varepsilon^2 \left[\mathbf{X}^T \mathbf{X} + \tilde{\tau}^{-2} \tilde{\boldsymbol{\Lambda}}^{-1}\right]^{-1}\right), \\ \sigma_\varepsilon^2 | \mathbf{y}, \mathbf{X}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \tau^2 &\sim IG\left(\frac{n+p}{2}, \frac{1}{2}(\mathbf{y} - \tilde{\boldsymbol{\beta}}_0 - \mathbf{X} \tilde{\boldsymbol{\beta}})^T (\mathbf{y} - \tilde{\boldsymbol{\beta}}_0 - \mathbf{X} \tilde{\boldsymbol{\beta}}) + \frac{1}{2\tilde{\tau}^2} \tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{\Lambda}}^{-1} \tilde{\boldsymbol{\beta}}\right), \\ \lambda_j^2 | v_j, \boldsymbol{\beta}_j, \tau^2, \sigma_\varepsilon^2 &\sim IG\left(1, \frac{1}{v_j} + \frac{\tilde{\boldsymbol{\beta}}_j^2}{2\tilde{\tau}^2 \tilde{\sigma}_\varepsilon^2}\right), \\ \tau^2 | \xi, \sigma_\varepsilon^2, \boldsymbol{\beta}_j, \lambda_j^2 &\sim IG\left(\frac{p+1}{2}, \frac{1}{\xi} + \frac{1}{2\tilde{\sigma}_\varepsilon^2} \sum_{j=1}^p \frac{\tilde{\boldsymbol{\beta}}_j^2}{\tilde{\lambda}_j^2}\right), \\ (v_j | \lambda_j^2 &\sim IG(1, \tilde{\sigma}_j^4 + 1/\tilde{\lambda}_j^2), \\ \xi | \tau^2 &\sim IG(1, 1 + 1/\tilde{\tau}^2), \end{aligned} \quad (4)$$

where $\boldsymbol{\Lambda} = (\lambda_1^2, \dots, \lambda_p^2)$ and symbols with tilde indicate random samples. The intercept term of the regression model (1) is sampled through normal distribution

$$\boldsymbol{\beta}_0 \sim N\left(\bar{\mathbf{Y}} - \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}} \tilde{\sigma}_\varepsilon^2 / n\right), \quad (5)$$

where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}_\varepsilon^2$ are estimations calculated from (4).

The maximum likelihood estimate of $\tilde{\boldsymbol{\beta}}$ in traditional nonhierarchical linear regression is $\tilde{\boldsymbol{\beta}}^{\text{ML}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. For our model (3), expressing the Bayesian posterior mean estimator of

β as $E[\beta|y, \mathbf{X}] = (1 - \kappa) \hat{\beta}^{\text{ML}}$, we have $\kappa = (\kappa_1, \dots, \kappa_p)$ and $\kappa_j = 1/(1 + \tau^2 \lambda_j^2 \mathbf{X}_j^T \mathbf{X}_j)$ is the shrinkage factor that indicates the degree of shrinkage for each coefficient and the importance of the corresponding variables. On the one hand, if $\kappa_j \rightarrow 0$, no shrinkage is performed on β_j such that $E[\beta_j|y, \mathbf{X}] \rightarrow \hat{\beta}_j^{\text{ML}}$, which means \mathbf{X}_j should be selected as an important predictor variable. On the other hand, if $\kappa_j \rightarrow 1$, $E[\beta_j|y, \mathbf{X}] \rightarrow 0$, suggesting that no signal emerges from \mathbf{X}_j so that \mathbf{X}_j might therefore be described as a “noise variable.” Thus, variables with a comparatively smaller κ_j should be selected in the final model. However, the threshold of κ_j for selection is not clear and remains as a problem.

The prior distribution of κ_j can be derived from λ_j such that

$$p(\kappa_j | \sigma_j^2, \tau) \propto \frac{(\tau^2 \mathbf{X}_j^T \mathbf{X}_j)^{\frac{1}{2}}}{\pi [\kappa_j (\tau^2 \sigma_j^2 \mathbf{X}_j^T \mathbf{X}_j - 1) + 1]} \frac{1}{\sqrt{\kappa_j (1 - \kappa_j)}}.$$

The distribution of κ_j keeps the shape of the horseshoe of *Beta* (1/2, 1/2) only when $\tau^2 \sigma_j^2 \mathbf{X}_j^T \mathbf{X}_j = 1$. Thus, when $\tau^2 \sigma_j^2 \mathbf{X}_j^T \mathbf{X}_j \neq 1$, the suggested threshold value at 0.5 (Carvalho, Polson, and Scott 2010) cannot be applied. However, the range of κ_j remains between (0, 1). Relying on the robust shrinkage behavior and the scaled estimation of local shrinkage parameter λ_j , we select variables at the same shrinkage level as most important variables. Specifically, we select variables with a small shrinkage factor κ_j differing from κ_{\min} , the smallest estimation in κ , by less than one order of magnitude.

2.2. Variable selection and missing data imputation

Here we extend the high-dimensional variable selection problem to incomplete data. We assume that the predictor variables follow an independent multivariate normal distribution. The covariance matrix of high-dimensional data is usually sparse, and its likelihood estimate is non-positive definite, and it is difficult to obtain an unbiased estimate of the covariance matrix. The shrinkage method is commonly employed to estimate the covariance matrix for high-dimensional data. Li, Craig, and Bhadra (2019) compared several methods and found that all of them more or less have different levels of bias. Therefore, we assume that each predictor variable follows normal distribution, where $\mathbf{X}_j \sim N(\mu_j, \sigma_j^2)$, $j = 1, \dots, p$. For the imputation model we take into account the sparsity of covariance matrix and use parameters derived from the regression model for estimation. Denote X_k as a predictor variable with missing values, then $\mathbf{X}_{-k} = \{\mathbf{X}_1, \dots, \mathbf{X}_{k-1}, \mathbf{X}_{k+1}, \dots, \mathbf{X}_p\}$ is the set of predictor variables excluding \mathbf{X}_k , and $\boldsymbol{\beta}_{-k} = (\beta_1, \dots, \beta_{k-1}, \beta_{k+1}, \dots, \beta_p)$ is the set of corresponding regression coefficients. Using Bayes theorem, the conditional distribution $p(\mathbf{X}_k | y, \mathbf{X}_{-k}, \mu_k, \sigma_k^2, \beta_0, \boldsymbol{\beta}) \propto P(y | \mathbf{X}, \beta_0, \boldsymbol{\beta}) P(\mathbf{X}_k | \mathbf{X}_{-k}, \mu_k, \sigma_k^2)$ (Kim, Sugar, and Belin 2015). As a result, we have

$$\begin{aligned} \mathbf{X}_k | y, \mathbf{X}_{-k}, \mu_k, \sigma_k^2, \beta_0, \boldsymbol{\beta} &\sim N(m_k, \Sigma_k), \\ \text{where } m_k &= \frac{\beta_k \sigma_k^2 (y - \beta_0 - \mathbf{X}_{-k} \boldsymbol{\beta}_{-k}) + \mu_k \sigma_\varepsilon^2}{\beta_k^2 \sigma_k^2 + \sigma_\varepsilon^2} \\ \text{and } \Sigma_k &= \frac{\sigma_\varepsilon^2 \sigma_k^2}{\beta_k^2 \sigma_k^2 + \sigma_\varepsilon^2} \end{aligned} \quad (6)$$

For the mean μ_k and variance σ_k^2 of variable \mathbf{X}_k , we use standard non-informative prior $p(\mu_k, \sigma_k^2) \propto |\sigma_k^2|^{-(p+1)/2}$ such that the full conditionals of posterior distribution are

$$\begin{aligned}\mu_k | \mathbf{X}_k, \sigma_k^2 &\sim N(\bar{\mathbf{X}}_k, \sigma_k^2/n), \\ \sigma_k^2 | \mathbf{X}_k, \mu_k &\sim W^{-1}((\mathbf{X}_k - \mu_k)^T(\mathbf{X}_k - \mu_k), n-1).\end{aligned}\quad (7)$$

Here, the variance of incomplete variables is sampled from the inverse-Wishart distribution. The resulting data-augmentation with horseshoe estimation (DA-HS) algorithm can be written:

- (1) *Setting Initial Values.* Set $\boldsymbol{\beta}^{(0)} = (1, \dots, 1)$, $\lambda_j^{2(0)} \sim U(0, 1)$, $\tau^{2(0)} = 1$, $v_j^{(0)} = 1$, $\xi^{(0)} = 1$, $j = 1, \dots, p$, where $U(0, 1)$ is the standard uniform distribution. Use mean imputation for missing values as the starting values and get an imputed data $\mathbf{X}_{\text{aug}}^{(0)}$.
- (2) *Generating Imputation Parameters.* Draw imputation parameters $(\mu_k^{(t+1)}, \sigma_k^{2(t+1)})$ from $P(\mu_k^{(t+1)}, \sigma_k^{2(t+1)} | \mathbf{X}_{\text{aug}}^{(t)})$, $k = 1, \dots, p$ through Equation (7), where $\mathbf{X}_{\text{aug}}^{(t)}$ is the t -th imputed data of \mathbf{X} .
- (3) *Generating Regression Parameters.* Draw regression parameters $(\sigma_\varepsilon^{(t+1)}, \boldsymbol{\Lambda}^{(t+1)}, \tau^{2(t+1)}, v_1^{(t+1)}, \dots, v_p^{(t+1)}, \xi^{(t+1)}, \boldsymbol{\beta}^{(t+1)})$ from $P(\sigma_\varepsilon^{(t+1)}, \boldsymbol{\Lambda}^{(t+1)}, \tau^{2(t+1)}, v_1^{(t+1)}, \dots, v_p^{(t+1)}, \xi^{(t+1)}, \boldsymbol{\beta}^{(t+1)} | \mathbf{y}, \mathbf{X}_{\text{aug}}^{(t)})$ through the full conditional posterior distributions expressed in Equation (4). Sample $\beta_0^{(t+1)}$ through Equation (5).
- (4) *Calculating the Shrinkage Factor.* Calculate the shrinkage factor $\kappa_j^{(t+1)} = 1/(n\sigma_j^{2(t+1)}\tau^{2(t+1)}\lambda_j^{2(t+1)})$.
- (5) *Data Imputation.* For incomplete variable \mathbf{X}_j , draw imputations for missing values from normal distribution $N(m_j^{(t+1)}, \Sigma_j^{(t+1)})$ through Equation (6) and get $\mathbf{X}_{\text{aug}}^{(t+1)}$.
- (6) *Iteration.* Iterate Steps (2) – (5) to produce a Markov chain.

As a multiple imputation algorithm, we pool m imputed datasets after the burn-in period, and, for each of them, we select the subset of variables by including variables with $\kappa_j \leq 10\kappa_{\min}$ of each imputation. Variables that are selected for all m imputed datasets are included in the final model. The posterior mean estimations of $\boldsymbol{\beta}$ are calculated and combined following Rubin's Rule (Donald B Rubin 2004). After selection, we shrink the β_j estimates for the noise variables to 0.

3. Simulation study

We conduct two simulation studies to evaluate the performance of DA-HS. First, we generate data with a small number of incomplete variables to compare DA-HS with complete-case analysis in which variable selection is conducted only on cases without missing values. Then we compare DA-HS with an alternative algorithm in a more general high-dimensional data scenario where cases may contain missing values. We evaluate the performance of DA-HS and alternative algorithms in terms of the coefficient estimation and variable selection.

3.1. Synthetic data with a small proportion of missing

We simulate data sets where (n, p) is either $(200, 500)$ or $(300, 500)$. Each data set includes a fully observed outcome variable y and p predictor variables $\mathbf{X}_1, \dots, \mathbf{X}_p$ following a multivariate normal distribution with a compound-symmetry correlation structure where off-diagonal elements are either 0.3 or 0.7. We consider $\mathbf{X}_1, \dots, \mathbf{X}_8$ as signal variables and the remaining \mathbf{X} 's as noise variables. Using linear regression, we generate the outcome variable $y = \sum_{j=1}^8 \mathbf{X}_j \beta_j + \epsilon$, where the residual ϵ follows a standard normal distribution. The true regression coefficients $(\beta_1, \dots, \beta_8) = (1, 1, 1, 1, 1, 1, 1, 1)$ or $(5, 5, 5, 5, 5, 5, 5, 5)$, and $\beta_j = 0$, for $9 \leq j \leq p$.

In addition, we evaluate the impact of missing data between two missing data mechanisms and two missing data rates. We generate missing values considering both missing completely at random (MCAR) and missing at random (MAR) mechanisms. In each missingness mechanism, we generate missing values with a 10% or 20% rate for each of $\mathbf{X}_6, \dots, \mathbf{X}_9$. The remaining variables are considered as completely observed. In the MCAR mechanism, data are dropped randomly. In the MAR mechanism, missing values are generated for \mathbf{X}_k depending on \mathbf{X}_{k+10} using logistic regression models with the form $\text{logit}(\Pr(\mathbf{X}_k \text{ is missing})) = w_k + \mathbf{X}_{k+10}$, where w_k is the parameter controlling the missing data rate. Thus, at least 60% of the observations are complete cases for data with a 10% missingness rate and at least 20% of the observations are complete with a 20% missingness rate. We have 32 scenarios in this simulation, and we replicate each scenario 500 times. We apply both DA-HS and the most common strategy for variable selection of incomplete high-dimensional data, complete-case analysis (CC-LS), on each generated data set. The CC-LS removes incomplete cases and uses the frequentist lasso with 10-fold cross-validation to select variables and ordinary least squares (OLS) estimator to estimate coefficients of linear regression. In implementing DA-HS, we sample β following the strategy described in (Makalic and Schmidt 2016); that is, when high-dimensional data set has $p > 200$, the β is sampled using a fast Gaussian sampling algorithm (Bhattacharya, Chakraborty, and Mallick 2016); otherwise, it is sampled using the sampling algorithm based on Cholesky factorization (Rue 2001).

For analysis, we conduct DA-HS with 1,000 burn-in period, $m = 10$ imputations, and 1,000 iterations for each imputation. Figure 1 displays the MCMC trace plot of posterior estimates for $\beta_5, \dots, \beta_{10}$ among which missing values existed in the corresponding predictors of $\beta_6, \beta_7, \beta_8$ and β_9 during the 8,500 iterations. The true values of $\beta_5, \beta_6, \beta_7$, and β_8 are equal to 5 and those of β_9 and β_{10} are 0. The trace plot shows that MCMC mixes fast. In addition, the traces of the nonzero coefficients have obviously larger variance than those of zero coefficients. The posterior estimates of $\beta_5, \beta_6, \beta_7$, and β_8 shift around 1, and those of β_9 and β_{10} shift around 0. The trace plot of β_5 looks similar to that of β_6, β_7 , and β_8 ; that is the posterior estimates of coefficients corresponding to incomplete predictors and complete predictors are similar.

We then investigate and compare the regression coefficient estimation of the simulation results that are summarized across all predictor variables based on the bias (the difference between the estimated coefficients and the true coefficients) and mean squared error (MSE; the average of the error squares) over 500 simulated data sets. Both bias and MSE are calculated for all predictor variables in data sets.

Moreover, we assess the variable selection efficiency of the simulation results through the sensitivity, specificity, and Matthews correlation coefficient (MCC) (Matthews 1975). We

Table 1. Simulation with a small proportion of missing values under the MCAR mechanism using data augmentation with horseshoe estimation (DA-HS) and complete-case analysis with lasso selection and OLS estimation (CC-LS)^a.

n	p	Method	Bias	MSE	Sensitivity	Specificity	MCC
Missing rate: 10%							
200	500	DA-HS	−0.001	0.000	0.99 (0.01)	1.00 (0.00)	1.00 (0.01)
		CC-LS	0.000	0.002	1.00 (0.00)	0.95 (0.01)	0.51 (0.05)
300	500	DA-HS	0.000	0.000	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		CC-LS	0.000	0.001	1.00 (0.00)	0.96 (0.01)	0.52 (0.06)
Missing rate: 20%							
200	500	DA-HS	−0.001	0.001	0.97 (0.03)	1.00 (0.00)	0.99 (0.02)
		CC-LS	0.000	0.004	0.99 (0.02)	0.95 (0.01)	0.50 (0.06)
300	500	DA-HS	−0.001	0.000	0.99 (0.01)	1.00 (0.00)	1.00 (0.01)
		CC-LS	0.000	0.002	1.00 (0.00)	0.95 (0.01)	0.51 (0.06)

^aBias, MSE, sensitivity, specificity, and MCC are calculated over 4 scenarios under the MCAR mechanism. The mean of bias and MSE are calculated for all variables.

Table 2. Simulation with a small proportion of missing values under the MAR mechanism using data augmentation with horseshoe estimation (DA-HS) and complete-case analysis with lasso selection and OLS estimation (CC-LS)^a.

n	p	Method	Bias	MSE	Sensitivity	Specificity	MCC
Missing rate: 10%							
200	500	DA-HS	−0.001	0.000	0.99 (0.02)	1.00 (0.00)	0.99 (0.01)
		CC-LS	0.000	0.002	1.00 (0.00)	0.95 (0.01)	0.51 (0.05)
300	500	DA-HS	0.000	0.000	1.00 (0.01)	1.00 (0.00)	1.00 (0.00)
		CC-LS	0.000	0.001	1.00 (0.00)	0.96 (0.01)	0.52 (0.06)
Missing rate: 20%							
200	500	DA-HS	−0.001	0.001	0.96 (0.03)	1.00 (0.00)	0.98 (0.02)
		CC-LS	0.000	0.004	1.00 (0.02)	0.95 (0.01)	0.50 (0.06)
300	500	DA-HS	−0.001	0.001	0.99 (0.02)	1.00 (0.00)	0.99 (0.01)
		CC-LS	0.000	0.002	1.00 (0.00)	0.95 (0.01)	0.51 (0.05)

^aBias, MSE, sensitivity, specificity, and MCC are calculated over 4 scenarios under the MAR mechanism. The mean of bias and MSE are calculated for all variables.

denote the number of true positives, true negatives, false positives, and false negatives as TP, TN, FP, and FN, respectively. Then sensitivity, specificity, and MCC are defined as

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP},$$

$$\text{and MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}.$$

Sensitivity, also called as the true positive rate, measures the proportion of signal variables that are correctly selected in the final model, while the specificity, also called the true negative rate, refers to the proportion of noise variables that the final model correctly eliminated. As a correlation coefficient, an MCC with a value closer to 1 indicates a better performance of variable selection, with high values requiring good performance on both positive and negative classifications.

Tables 1 and 2 present the estimation and selection results for DA-HS and CC-LS under the MCAR and MAR mechanisms, respectively, with 10% and 20% missing rates. As shown in all scenarios, DA-HS has outstanding performance in variable estimation and much higher

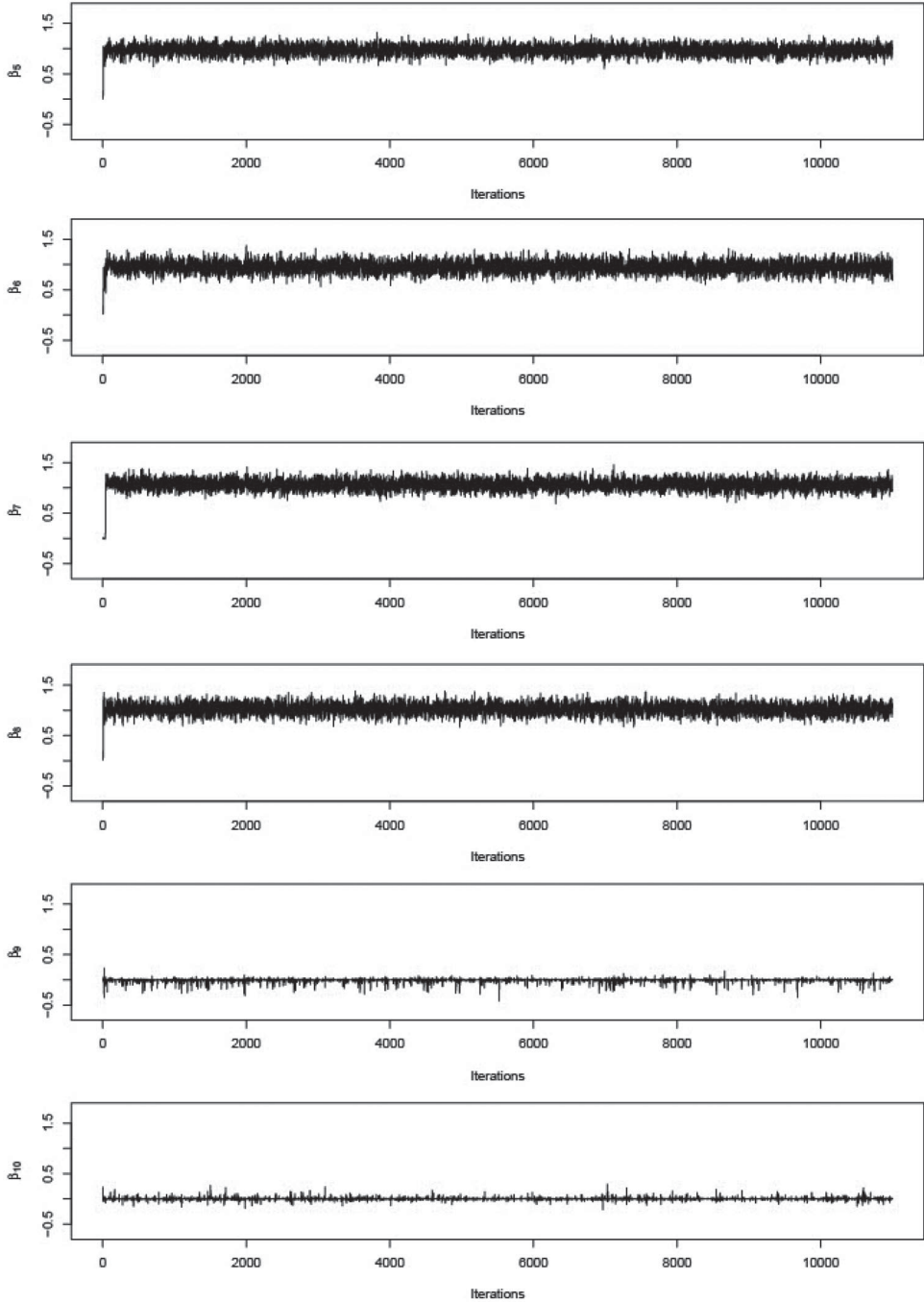


Figure 1. MCMC trace plot for $\beta_5 - \beta_{10}$

efficiency in variable selection than CC-LS. The bias and MSE of DA-HS estimations are negligible. Moreover, though the sensitivity and specificity are not low for CC-LS, the MCC is around 0.5 because of the lower TN and higher FP values as shown in Tables 1 and 2 of

the supplemental material. The remarkably smaller MSE and much higher MCC of DA-HS suggest that DA-HS offers a good imputation with estimation and selection. The result of CC-LS implies that complete-case analysis should not be used for high-dimensional incomplete data even if the proportion of missing values is small. Additionally, the standard deviations of MCC were lower for DA-HS than CC-LS, which suggests that imputation is necessary to avoid losing too much useful information as happens in complete-case analysis.

3.2. Synthetic data with a large proportion of missing

We conduct a series of simulations with larger proportion of missing values in a larger number of predictor variables, which may arise in practical situations. We investigate the performance of DA-HS and compare it with that of an alternative algorithm, mean imputation followed by lasso and 10-folded cross-validation for selection and OLS for estimation. Because missing values may exist in all observations, we apply mean-LS given that it is likely to be used in practice to address the issue of variable selection for high-dimensional incomplete data.

We simulate data sets for 6 different settings for (n, p) : $\{(200, 300), (200, 500), (200, 1000), (300, 300), (300, 500), (300, 1000)\}$. Like the synthetic data generated in the [Section 3.1](#), each data set includes n observations of a fully observed outcome variable y and p predictor variables $\mathbf{X}_1, \dots, \mathbf{X}_p$ following the multivariate normal distribution with a compound symmetry correlation structure with off-diagonal elements either 0.3 or 0.7. In the context of $\mathbf{y} = \sum_{j=1}^8 \mathbf{X}_j \beta_j + \boldsymbol{\varepsilon}$, we consider $\mathbf{X}_1, \dots, \mathbf{X}_8$ as signal variables and the remaining \mathbf{X} 's as noise variables. The true regression coefficients $(\beta_1, \dots, \beta_8) = (1, 1, 1, 1, 1, 1, 1, 1)$, $(5, 5, 5, 5, 5, 5, 5, 5)$ or $(3, 5, 3, 5, 3, 5, 3, 5)$, and, for $9 \leq j \leq p$, $\beta_j = 0$. The residual $\boldsymbol{\varepsilon}$ follows a normal distribution with variance taken to be either 1 or 4.

Using the same MCAR and MAR mechanisms as described in [Section 3.1](#), we generate missing values for more variables. Specifically, we generate missing values with a 10% or 20% missing rate for each of $\mathbf{X}_5, \dots, \mathbf{X}_{14}$. The remaining variables are considered as completely observed. The overall proportion of complete cases is approximately 15% or 40%, but it is possible that there are no complete cases in the data sets due to randomness in data generation. Therefore, we have 288 scenarios in this simulation, and each are replicated 500 times.

[Tables 3](#) and [4](#) present the variable selection results for DA-HS and mean-LS under the MCAR and MAR mechanisms, respectively, with 10% and 20% missing rates. As shown, DA-HS remarkably outperforms the mean-LS in all scenarios. Compared with mean-LS, DA-HS has a higher specificity very close to 1, which means that it is less likely to exclude true signal variables from the final model. Moreover, the sensitivity of both mean-LS and DA-HS is close to 1, which indicates that both algorithms can detect true signal variables in the final model. However, not like DA-HS, mean-LS is unable to exclude noise variables and likely to overfit the final model. [Tables 3](#) and [4](#) of the supplemental material show that mean-LS gives lower TN and higher FP of selection than DA-HS does. In addition, the MCC of DA-HS are higher than 0.90 across all scenarios, while that of mean-LS gets lower and round 0.35 with the increasing of dimensions. The mean of MCC is lower for mean-LS than for DA-HS (0.40 vs 0.96, respectively). MCC only approaches 1 when all the TN, TP, FN, and FP values are close to the true values. Thus, considering the close standard deviation of MCC by these two algorithms, DA-HS gives much more reliable selection than mean-LS.

[Table 5](#) displays the estimation results for DA-HS and mean-LS under the MCAR and MAR mechanisms with 10% and 20% missing rates. Similar to CC-LS, mean-LS estimates

Table 3. Selection results of simulation with a large proportion of missing values under MCAR mechanism using data augmentation with horseshoe estimation (DA-HS) and mean imputation with lasso selection and OLS estimation (mean-LS) across scenarios within the same data dimensions^a.

<i>n</i>	<i>p</i>	DA-HS			mean-LS		
		Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC
Missing rate: 10%							
200	200	0.96 (0.03)	1.00 (0.02)	0.97 (0.02)	1.00 (0.01)	0.87 (0.03)	0.47 (0.04)
	500	0.95 (0.04)	0.99 (0.02)	0.96 (0.03)	0.99 (0.01)	0.93 (0.01)	0.43 (0.04)
	1000	0.94 (0.04)	0.98 (0.04)	0.94 (0.03)	0.99 (0.01)	0.96 (0.01)	0.39 (0.04)
300	300	0.98 (0.02)	1.00 (0.00)	0.98 (0.01)	1.00 (0.00)	0.89 (0.02)	0.44 (0.04)
	500	0.97 (0.02)	1.00 (0.00)	0.98 (0.01)	1.00 (0.00)	0.93 (0.01)	0.42 (0.04)
	1000	0.96 (0.02)	1.00 (0.00)	0.98 (0.01)	1.00 (0.01)	0.96 (0.01)	0.39 (0.03)
Missing rate: 20%							
200	200	0.94 (0.05)	0.99 (0.02)	0.95 (0.04)	0.99 (0.01)	0.85 (0.03)	0.43 (0.04)
	500	0.92 (0.06)	0.99 (0.03)	0.94 (0.04)	0.99 (0.02)	0.92 (0.01)	0.39 (0.03)
	1000	0.91 (0.06)	0.98 (0.03)	0.92 (0.04)	0.98 (0.03)	0.95 (0.01)	0.36 (0.03)
300	300	0.96 (0.03)	1.00 (0.01)	0.98 (0.02)	1.00 (0.01)	0.87 (0.02)	0.40 (0.03)
	500	0.96 (0.03)	1.00 (0.01)	0.97 (0.02)	1.00 (0.01)	0.91 (0.01)	0.38 (0.03)
	1000	0.95 (0.03)	1.00 (0.01)	0.97 (0.02)	0.99 (0.01)	0.95 (0.01)	0.36 (0.03)

^aThe mean and standard deviation of sensitivity, specificity, and MCC are calculated over 12 scenarios under the MCAR mechanism.

Table 4. Selection results of simulation with a large proportion of missing values under MAR mechanism using data augmentation with horseshoe estimation (DA-HS) and mean imputation with lasso selection and OLS estimation (mean-LS) across scenarios within the same data dimensions^a.

<i>n</i>	<i>p</i>	DA-HS			mean-LS		
		Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC
Missing rate: 10%							
200	200	0.96 (0.03)	1.00 (0.02)	0.97 (0.03)	0.99 (0.01)	0.86 (0.03)	0.45 (0.04)
	500	0.94 (0.04)	0.99 (0.03)	0.95 (0.03)	0.99 (0.01)	0.93 (0.01)	0.41 (0.04)
	1000	0.94 (0.05)	0.99 (0.03)	0.94 (0.03)	0.98 (0.02)	0.96 (0.01)	0.38 (0.03)
300	300	0.97 (0.02)	1.00 (0.00)	0.98 (0.01)	1.00 (0.00)	0.88 (0.02)	0.42 (0.04)
	500	0.97 (0.02)	1.00 (0.00)	0.98 (0.01)	1.00 (0.01)	0.92 (0.01)	0.40 (0.03)
	1000	0.96 (0.02)	1.00 (0.00)	0.97 (0.01)	1.00 (0.01)	0.95 (0.01)	0.37 (0.03)
Missing rate: 20%							
200	200	0.93 (0.05)	0.99 (0.02)	0.94 (0.04)	0.99 (0.02)	0.84 (0.03)	0.41 (0.04)
	500	0.91 (0.06)	0.98 (0.03)	0.93 (0.04)	0.98 (0.03)	0.91 (0.01)	0.38 (0.04)
	1000	0.90 (0.07)	0.98 (0.04)	0.91 (0.05)	0.97 (0.04)	0.95 (0.01)	0.35 (0.03)
300	300	0.95 (0.03)	1.00 (0.00)	0.97 (0.02)	0.99 (0.01)	0.86 (0.02)	0.38 (0.03)
	500	0.94 (0.04)	1.00 (0.01)	0.96 (0.02)	0.99 (0.02)	0.90 (0.01)	0.37 (0.03)
	1000	0.93 (0.04)	1.00 (0.01)	0.96 (0.03)	0.99 (0.02)	0.94 (0.01)	0.35 (0.03)

^aThe mean and standard deviation of sensitivity, specificity, and MCC are calculated over 12 scenarios under the MAR mechanism.

coefficients using OLS estimator, which is UMVUE. However, the MSE and bias of DA-HS are lower in most scenarios. Considering the undesired selection performance, we suggest that mean imputation with lasso approach, such as the mean-LS approach in this study, should not be used for variable selection for high-dimensional incomplete data.

Table 5. Estimation results of simulation with a large proportion of missing values under MCAR and MAR mechanisms using data augmentation with horseshoe estimation (DA-HS) and mean imputation with lasso selection and OLS estimation (mean-LS) across scenarios within the same data dimensions^a.

		DA-HS				mean-LS			
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
<i>n</i>	<i>p</i>	MCAR		MAR		MCAR		MAR	
Missing rate: 10%									
200	200	−0.003	0.004	−0.003	0.004	0.007	0.026	0.008	0.038
	500	−0.001	0.002	−0.002	0.002	0.003	0.013	0.003	0.018
	1000	−0.001	0.001	−0.001	0.001	0.001	0.008	0.002	0.011
300	300	−0.002	0.002	−0.002	0.002	0.004	0.014	0.005	0.022
	500	−0.001	0.001	−0.001	0.001	0.003	0.010	0.003	0.014
	1000	−0.001	0.001	−0.001	0.001	0.001	0.005	0.002	0.008
Missing rate: 20%									
200	200	−0.007	0.01	−0.007	0.011	0.011	0.049	0.012	0.067
	500	−0.003	0.004	−0.003	0.005	0.005	0.024	0.005	0.032
	1000	−0.002	0.002	−0.002	0.003	0.002	0.014	0.003	0.049
300	300	−0.004	0.004	−0.004	0.005	0.008	0.049	0.008	0.041
	500	−0.003	0.003	−0.003	0.003	0.005	0.049	0.005	0.026
	1000	−0.001	0.001	−0.001	0.002	0.002	0.049	0.003	0.014

^aResults are summarized over 12 scenarios under the MCAR and MAR mechanisms. The mean of bias and MSE are calculated for all variables.

4. World Bank data application: life expectancy study

4.1. Data description

As an international financial institution that provides loans to developing countries, the World Bank provides a web resource, Data Bank, that includes simple and quick access to collections of data on various topics from the World Bank Group (2021). These data are a vital source of information on financial and technical assistance to developing countries around the world. Specifically, it contains 20 topics, 1,823 annual indicators, 247 countries and areas, and 59 years of data from 1950–2018. We use the World Bank data and perform DA-HS as an application to health data with a large number of continuous variables.

A major challenge for the analysis of the World Bank arises due to a high proportion of missing values. Here, we illustrate DA-HS on World Bank data to identify health factors associated with life expectancy at birth of countries located in Europe, Asia, and Pacific in 2010. The life expectancy at birth provided by the World Bank uses the mortality patterns at the time of birth to predict the average lifespan of newborns (World Bank Group 2021).

4.2. Data analysis

We focus on 61 countries with reported life expectancy at birth. Figure 2 shows the life expectancy at birth of these countries. In this heat map, red color indicates higher life expectancy, and light yellow color indicates lower life expectancy. Countries with life expectancy at birth missing are shown in white color. As shown, the life expectancy at birth is highest in countries in western Europe (~ 80) and lowest in countries in eastern Europe and central Asia (~ 60). To identify the factors associated with life expectancy at birth, we include 154 factors in the Health topic from the World Bank with at least 5 observed values.

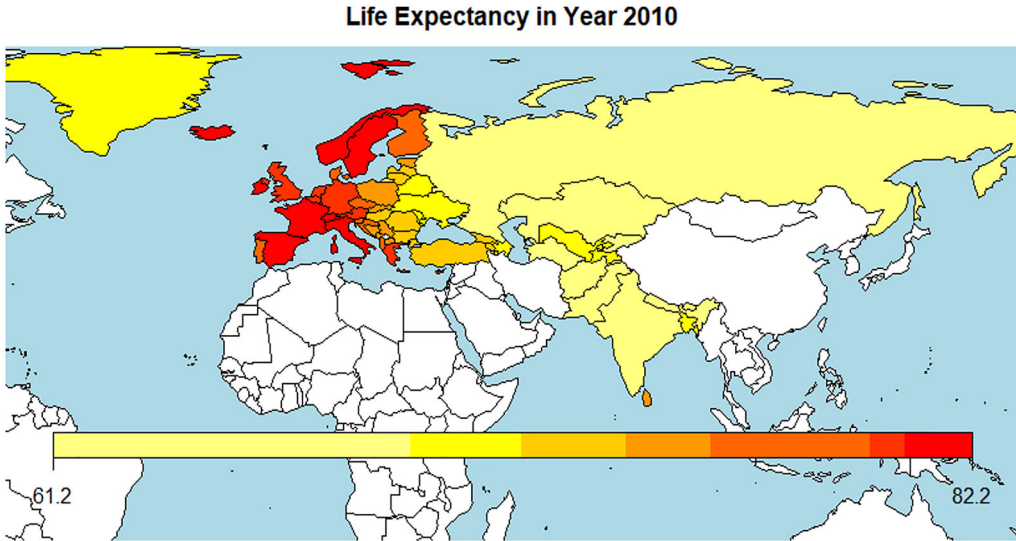


Figure 2. Heat map of life expectancy at birth in 2010.

The list of variables and corresponding number of missing values are displayed in the [Table 5](#) of the [supplemental material](#).

We first investigate the missing data pattern of the data set. [Figure 3](#) shows the pattern of missing data in this data set, showing that there is no complete case in this data. In addition to the outcome variable, we have 7 completely observed predictor variables for all countries: birth rate, death rate, male life expectancy at birth, female life expectancy at birth, fertility rate, total population, and population growth. The country with the highest number of missing items had missing values in 140 variables.

After this investigation, we impute data and perform variable selection using DA-HS. Male and female life expectancy at birth are listed in the data set, and since these are directly related to the outcome variable, we apply DA-HS on the data excluding these two variables to identify potentially related factors. Using DA-HS, we apply horseshoe prior on the regression parameters with data augmentation. After 8,500 iterations, including 1,000 burn-in iterations, we pool the result by multiple imputation with shrinkage factors to decide on the final selection subset.

4.3. Results

[Table 6](#) shows variable selection results from all 154 predictor variables and their coefficient estimation. Two variables are selected in the final model. Both male and female life expectancy at birth are selected with shrinkage factors nearly 0. As life expectancy are calculated using male and female, it is obvious that male and female life expectancy are highly correlated with the outcome variable. To further detect variables related to the life expectancy outcome, we conduct the analysis excluding male and female life expectancy. [Table 7](#) displays the results that percentage of female population ages 80 and above and percentage of male survive to age 65 are selected. The selection results take both male and female into consideration. Also, these two variables are of the oldest age among all variables about percentages of population

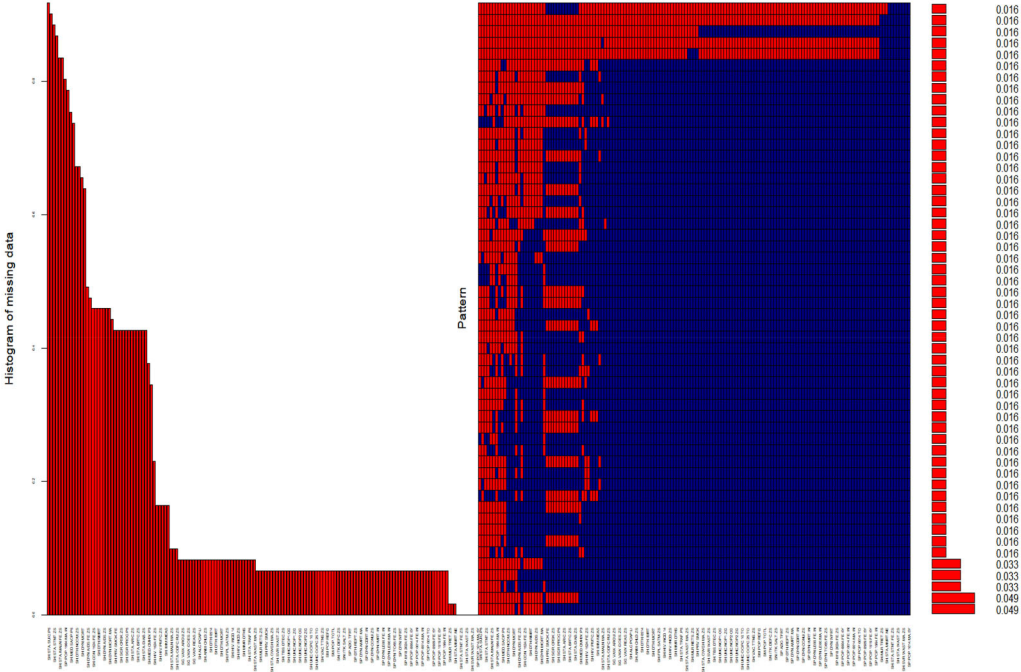


Figure 3. Missing values of World Bank Data. The left histogram displays the probability of missingness for each variable; the right pattern displays all missing patterns in the data. Missing values are represented in red; observed values are represented in blue.

Table 6. Variable selection for life expectancy Study 1.

	Coefficient estimates	Shrinkage factor estimates
Life expectancy at birth, female (years)	0.486	0.000
Life expectancy at birth, male (years)	0.515	0.000

Table 7. Variable selection for life expectancy Study 2.

	Coefficient estimates	Shrinkage factor estimates
Population ages 80 and above, female (% of female population)	0.252	0.000
Survival to age 65, male (% of cohort)	0.720	0.015

age and survival age. As it is widely recognized that females live longer than males, having percentages of male survival to age 65 counts more male populations.

5. Gene expression microarray data application: riboflavin study

5.1. Data description

Here, the DA-HS algorithm is applied to a genetic study with high-dimensional microarray data set riboflavin (Maathuis, Kalisch, and Bühlmann 2009) which includes 71 fully observed samples, 4,088 predictor variables, and an outcome variable, log-transformed riboflavin production rate. This data set is originally described in the gene study of *Bacillus subtilis* (Lee et al. 2001) and is available in the R package “hdi” (Dezeure et al. 2015).

Table 8. Riboflavin data selection result by DA-HS.

Missing mechanism	Incomplete case proportion	Selected genes (coefficient estimates)
Missing rate: 10%		
MCAR	070%	YOAB_at (−1.289)
	100%	YOAB_at (−1.196)
MAR	070%	YOAB_at (−1.393)
	100%	YOAB_at (−1.392)
Missing rate: 20%		
MCAR	070%	YXLE_at (−0.443)
	100%	YOAB_at (−1.410)
MAR	070%	YXLE_at (−0.417)
	100%	YOAB_at (−1.423)

5.2. Data analysis

To examine the performance of DA-HS, we generate 10% or 20% missing values in each of 20% random selected predictor variables. Missing values are generated under MCAR and MAR mechanism. To further examine DA-HS using data with a limited number of complete cases, we set the overall proportion of samples with missing entries as either 70% or 100% by generating 10% or 20% missing values under MCAR or MAR in the randomly selected 70% data entries or the whole data set. For each of the 8 scenarios, we simulate 50 data sets.

We conduct the DA-HS on each simulated data sets with 6,500 iterations, among which 1,500 iterations are considered as the burn-in period. The selection and regression result is obtained for each of the generated data.

5.3. Results

Within each scenario, we take the mean number of selected variables as the final selection model size and list the genes with highest selection frequencies as well as the coefficient estimates for each selected genes using the mean of estimates in Table 8. Though the riboflavin data consist of a large number of predictor variables, the final model of each scenario selected 1 gene only. The selection and estimation results are very similar among all scenarios. The gene YOAB_at is selected in 6 out of 8 scenarios with all negative estimated coefficients approximately −1.4. The gene YXLE_at is selected in 2 scenarios with coefficient estimates approximately −0.4.

We investigate the results over all simulated missing data analyses, and two genes, YOAB_at and YXLE_at, are selected in the final model. They are also presented as most important predictors in a selection analysis on the completely observed riboflavin data using cross-validated Lasso and TREX, a fast approach to shrinkage avoiding tuning parameter (Lederer Lederer and Müller 2015), where the coefficient estimates presented are much closer to 0. As we discussed before, the coefficients estimated by DA-HS are calculated through horseshoe priors. With the local shrinkage parameters, it is less likely to over-shrink the coefficients of important predictor variables.

6. Discussion

In this study, we provide a practical high-dimensional regression regularization algorithm DA-HS for incomplete data. Specifically, we use horseshoe prior in linear regression and mix

the Bayesian shrinkage with multiple imputation in MCMC as suggested by Yang, Belin, and Boscardin (2005). The Bayesian shrinkage methods are unable to shrink the coefficients of noise variables to exact zero, so our selection relies on the magnitude of shrinkage from the shrinkage factor for each coefficient.

The simulation studies suggest the necessity of multiple imputation on high-dimensional incomplete data. Compared with the complete case analysis, DA-HS uses more information from incomplete cases and will lead to better selection efficiency. Compared with mean imputation followed by frequentist lasso and OLS estimation, DA-HS provide more accurate estimates and remarkably more efficient selection on synthetic data with a large proportion of missing values. The estimates of DA-HS are close to the true values even with shrinkage estimator. The selection results of DA-HS are impressive, especially when data dimension is higher. Additionally, though for the sake of computational efficiency, the imputation step of DA-HS takes data covariance into account through the regression parameters, the outstanding performance of DA-HS in simulation studies with high covariance of 0.7 demonstrated the feasibility of applying DA-HS to data with highly correlated covariates. Moreover, by checking the standard deviation in the selection efficiency measurements, we infer that the performance of DA-HS is stable, for which the selection is given with the multiple imputation. Because the horseshoe prior has been proven to be robust to the signal variables (Carvalho, Polson, and Scott 2010), the performance of DA-HS is also grounded in a strong theoretical foundation.

We illustrate DA-HS using the World Bank data as an application, and the DA-HS provides appropriate selection with imputation of high proportion of missing values. The coefficient estimates by DA-HS also give reliable interpretation of the variable effects on the outcome of interest. The illustration of these methods for World Bank data suggests that the DA-HS can be used for health science data sets in a similar way.

We further apply DA-HS on microarray data, with the simulated high proportion of incomplete cases. Our algorithm show advance in less over-shrinkage in estimation by comparing with analysis using fully observed data.

This study assumes normally distributed predictor variables, however, in real data analysis, there may be other types of variables involved in the data. When predictor variables are continuous and follow skewed distributions, we suggest conducting variable transformation before using DA-HS. However, when predictor variables are nominal or ordinal, extension of this algorithm to a more general case with other types of variables will remain of interest for future study.

References

- Bartlett, J. W., S. R. Seaman, I. R. White, and J. R. Carpenter. 2015. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research* 24 (4):462–87. doi:[10.1177/0962280214521348](https://doi.org/10.1177/0962280214521348).
- Beesley, L. J., and J. M. Taylor. 2021. Bayesian variable selection and shrinkage strategies in a complicated modelling setting with missing data: A case study using multistate models. *Statistical Modelling* 21 (1–2):11–29. doi:[10.1177/1471082x20920972](https://doi.org/10.1177/1471082x20920972). <https://journals.sagepub.com/doi/abs/10.1177/1471082X20920972>.
- Bhattacharya, A., A. Chakraborty, and B. K. Mallick. 2016. Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *Biometrika* 103 (4):985–91.
- Carvalho, C. M., N. G. Polson, and J. G. Scott. 2009. Handling Sparsity via the Horseshoe. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*. <https://proceedings.mlr.press/v5/carvalho09a.html>.

- Carvalho, C. M., N. G. Polson, and J. G. Scott. 2010. The horseshoe estimator for sparse signals. *Biometrika* 97 (2):465–80.
- Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen. 2015. High-dimensional Inference: Confidence intervals, p-values and R-software hdi. *Statistical Science* 30 (4):533–58, 26.
- Garcia, R. I., J. G. Ibrahim, and H. Zhu. 2010. Variable selection for regression models with missing data. *Statistica Sinica* 20 (1):149–65.
- George, E. I., and R. E. McCulloch. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88 (423):881–9.
- Kim, S., C. A. Sugar, and T. R. Belin. 2015. Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in Medicine* 34 (11):1876–88.
- Lederer, J., and C. L. Müller. 2015. Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX. *Proceedings of the AAAI Conference on Artificial Intelligence* 29 (1):2729–35. <https://ojs.aaai.org/index.php/AAAI/article/view/9550>.
- Lee, J. M., S. Zhang, S. Saha, S. Santa Anna, C. Jiang, and J. Perkins. 2001. RNA expression analysis using an antisense *Bacillus subtilis* genome array. *Journal of Bacteriology* 183 (24):7371–80.
- Li, Y., B. A. Craig, and A. Bhadra. 2019. The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics* 28 (3):747–57.
- Liang, F., B. Jia, J. Xue, Q. Li, and Y. Luo. 2018. An imputation-regularized optimization algorithm for high dimensional missing data problems and beyond. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 80 (5):899–926.
- Liu, Y., Y. Wang, Y. Feng, and M. M. Wall. 2016. Variable selection and prediction with incomplete high-dimensional data. *The Annals of Applied Statistics* 10 (1):418–50.
- Long, Q., and B. A. Johnson. 2015. Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics (Oxford, England)* 16 (3):596–610.
- Maathuis, M. H., M. Kalisch, and P. Bühlmann. 2009. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics* 37 (6A):3133–64, 32.
- Makalic, E., and D. F. Schmidt. 2016. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters* 23 (1):179–82.
- Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta* 405 (2):442–51.
- Piironen, J., and A. Vehtari. 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* 11 (2):5018–51, 34.
- Rubin, D. B. 2004. Vol. 81 of *Multiple imputation for nonresponse in surveys*. United States: John Wiley & Sons.
- Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63 (3):581–92.
- Rue, H. 2001. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2):325–38.
- Schafer, J. L. 1997. *Analysis of incomplete multivariate data*. New York, NY: CRC Press.
- Tanner, M. A., and W. H. Wong. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82 (398):528–40.
- Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):267–88.
- van Dyk, D. A., and X.-L. Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics* 10 (1):1–50.
- Wand, M. P., J. T. Ormerod, S. A. Padoan, and R. Frühwirth. 2011. Mean field variational bayes for elaborate distributions. *Bayesian Analysis* 6 (4):847–900.
- World Bank Group. 2021. The world bank data. Accessed May 8, 2021. <http://data.worldbank.org>.
- Yang, X., T. R. Belin, and W. J. Boscardin. 2005. Imputation and variable selection in linear regression models with missing covariates. *Biometrics* 61 (2):498–506.