Adaptive Bayesian SLOPE –
High-dimensional Model Selection with Missing Values

Wei Jiang\textsuperscript{1}  Malgorzata Bogdan\textsuperscript{2}  Julie Josse\textsuperscript{1}  Błażej Miasojedow\textsuperscript{3}
Veronika Ročková\textsuperscript{4}  TraumaBase\textsuperscript{©} Group\textsuperscript{5}

September 17, 2019

Abstract

The selection of variables with high-dimensional and missing data is a major challenge and very few methods are available to solve this problem. Here we propose a new method – adaptive Bayesian SLOPE – which is an extension of the SLOPE method of sorted $l_1$ regularization within a Bayesian framework and which allows to simultaneously estimate the parameters and select variables for large data despite missing values. The method follows the idea of the Spike and Slab LASSO, but replaces the Laplace mixture prior with the frequentist motivated ”SLOPE” prior, which targets control of the False Discovery Rate. The regression parameters and the noise variance are estimated using stochastic approximation EM algorithm, which allows to incorporate missing values as well as latent model parameters, like the signal magnitude and its sparsity. Extensive simulations highlight the good behavior in terms of power, FDR and estimation bias under a wide range of simulation scenarios. Finally, we consider an application of severely traumatized patients from Paris hospitals to predict the level of platelet, and demonstrate, beyond the advantage of selecting relevant variables, which is crucial for interpretation, excellent predictive capabilities. The methodology is implemented in the R package ABSLOPE, which incorporates C++ code to improve the efficiency of the proposed method.

Keywords: incomplete data, FDR control, penalized regression, spike and slab prior, health data

\textsuperscript{1}Inria XPOP and CMAP, École Polytechnique, France
\textsuperscript{2}University of Wroclaw, Poland and Lund University, Sweden
\textsuperscript{3}University of Warsaw, Poland
\textsuperscript{4}University of Chicago Booth School of Business, USA
\textsuperscript{5}Hôpital Beaujon, APHP, France
## Contents

1 Introduction 3  
   1.1 Contribution 4  
   1.2 Previous work on selecting variables with missing data 5  

2 Statistical model and assumptions 6  
   2.1 SLOPE 7  
   2.2 Adaptive Bayesian SLOPE 8  
   2.3 Assumptions for missing values 11  
   2.4 Overview of modeling 12  

3 Parameter estimation and model selection 12  
   3.1 Maximizing the observed penalized likelihood 13  
   3.2 Simulation step: sampling the latent variables 14  
   3.3 Stochastic approximation and maximization steps 15  
      3.3.1 Step-size $\eta_t = 1$ 16  
      3.3.2 General step-size 17  
   3.4 SLOB: Quick version of ABSLOPE 18  

4 Simulation study 19  
   4.1 Simulation setting 19  
   4.2 Convergence of SAEM 20  
   4.3 Behavior of ABSLOPE - SLOB 21  
      4.3.1 Scenario 1 22  
      4.3.2 Scenario 2 25  
   4.4 Comparison with competitors 27  
   4.5 Comparison of computation time 28  

5 Application to Traumabase dataset 30  
   5.1 Details on the dataset and preprocessing 30  
   5.2 Model selection results 32  
   5.3 Prediction performances 33  
   5.4 Results with interaction 36  

6 Discussion 37  

A Appendix 37  
   A.1 Deviation of prior (3) started from SLOPE prior 37  
   A.2 Missing mechanism 38  
   A.3 Standardization for MAR 39  
   A.4 Details of the simulation step: sampling the latent variables 39  
   A.5 Proof of conditional distribution of missing data 41  
   A.6 Summary of algorithms 43  
   A.7 Initialization of ABSLOPE 43
1 Introduction

The selection of variables with high-dimensional data is an important issue in many areas. For example, in molecular genetics, a large number of variables are available but only a few are relevant to explain the biological phenomena. The LASSO (Tibshirani, 1996) is a computationally efficient procedure to simultaneously estimate parameters and select variables which comes with some theoretical guarantees. However, it can lead to the inclusion of many false discoveries (Su et al., 2017) and identify the true model only if the strict “irrepresentability” condition is satisfied (Wainwright, 2009; Tardivel and Bogdan, 2018). Adaptive LASSO (Zou, 2006) which uses a weighted $\ell_1$ penalty (it adjusts the elements of the regularization parameters based on some initial estimate of regression coefficients) instead of a $\ell_1$ penalty significantly reduces the bias in estimating the regression parameters. This allows adaptive LASSO to recover the true model even when the irrepresentability condition is not satisfied (see e.g. Fan et al. (2014); Tardivel and Bogdan (2018); Rejchel and Bogdan (2019)). However, the properties of adaptive LASSO still heavily depend on the weight function and the selection of the tuning parameter $\lambda$, and the optimal choices of these parameters depend on the unknown parameters, like the signal magnitude or its sparsity. Recently, in Ročková and George (2018) the spike and slab LASSO (SSL) was introduced, which places the adaptive LASSO within the Bayesian context. In SSL the penalty functions arise from a fully Bayes spike-and-slab formulation and are self-adaptive to the process generating the data. Simulations and theoretical results reported in Ročková and George (2018); Ročková (2018) show that SSL allows to obtain a near minimax rate of convergence of regression estimates and performs very well even when the columns in the design matrix are strongly correlated.

In this article we extend the idea of SSL and present the adaptive Bayesian version of Sorted L-One Penalized Estimator (SLOPE) of Bogdan et al. (2015). The main motivation behind SLOPE was the control of the False Discovery Rate (FDR). Control of FDR is a focus of many methodological developments in multiple regression (see e.g. Barber et al. (2015); Candès et al. (2018)). Compared to methods aiming at the perfect signal recovery, methods controlling FDR are more liberal and allow for small number of false discoveries. This in turn leads to substantial increase of power and improvement of prediction properties.
when the signal is weak. As shown in Bogdan et al. (2015), SLOPE controls FDR when the design matrix is orthogonal. Moreover, in Su and Candès (2016); Bellec et al. (2018), it is proved that, contrary to LASSO, SLOPE allows to obtain an exact minimax rate of convergence of estimators of regression coefficients in sparse high dimensional multiple regression. However, similarly as in case of LASSO, for finite sample sizes it is practically impossible to combine good prediction and selection properties by SLOPE. Large values of the shrinkage parameters, needed to control FDR at a low level, result in the large bias of important regression coefficients and usually lead to poor estimation properties. One solution in practice suggested by Bogdan et al. (2015); Brzyski et al. (2019) is applying 2 steps: i) using SLOPE to detect significant predictors; ii) applying the standard least-squares methods within selected predictors for the estimation of coefficients. This two step approach allows to correct for bias of SLOPE estimates but it does not prevent the loss of FDR control, which typically occurs unless the true number of non-zero regression coefficients in $\beta$ is very small. The loss of FDR control by SLOPE is again related to the shrinkage of large regression coefficients, whose unexplained effect is often picked by even slightly correlated “false” explanatory variables (see Su et al. (2017) for the theoretical analysis of the similar phenomenon for LASSO).

1.1 Contribution

The adaptive Bayesian version of SLOPE we propose, tackles these issues in a similar way as the SSL solves the analogous problems of LASSO. The major difference between these two methods is the selection of the priors. The adaptive SLOPE prior is substantially more complicated and constructed in such a way that its “spike” component effectively leads to the regular SLOPE penalization of very small regression coefficients. This, together with reduction of the bias for large signals, allows for FDR control under a wide range of possible scenarios, as suggested by our extensive simulation study. The “slab” component of the SLOPE prior also preserves the properties of the SLOPE penalty, which leads to “averaging” similar estimates of regression coefficients (see Figueiredo and Nowak (2016) for discussion of the SLOPE averaging effect). This allows to obtain very good prediction properties when regressors are substantially correlated. The hyper-parameters of the
SLOPE prior are iteratively updated using the full Bayesian model in the spirit of stochastic approximation EM (Lavielle, 2014), which also has the great advantage to handling missing data.

Our contribution is to provide a complete methodology with a theoretical and efficient basis in computation time, to perform selection of variables with high dimensional data and missing values, and available in practice for users. The methodology presented in this article is implemented as an R (R Core Team, 2017) package ABSLOPE (Jiang et al., 2019b). The code to reproduce all the experiments is also provided in GitHub (Jiang, 2019).

1.2 Previous work on selecting variables with missing data

In the context of high-dimensional data, in addition to model selection issues, the management of missing data is crucial. Indeed, missing data are omnipresent. For example, genetic data obtained from microarray experiments often contain missing values for several reasons: insufficient resolution, image corruption, manufacturing errors, etc. The most common practice for dealing with missing data, listwise deletion, leads to estimation bias, unless the missing data are generated completely randomly, and often results in a loss of important information, especially for large data. The literature on the management of missing values is abundant, see Little and Rubin (2002) and the platform R-miss-tastic\(^1\) (Mayer et al., 2019) for an overview of the state of the art. However, there are few methods for selecting a model when some values are missing. For example, in generalized linear models, Claeskens and Consentino (2008); Ibrahim et al. (2008); Jiang et al. (2018) adapted likelihood-based information criteria designed for complete data such as AIC. However, their methods cannot process large data where the dimension \(p\) is large and comparable to the sample size \(n\), or greater than \(n\). In linear models, Loh and Wainwright (2012) formulated LASSO with missing values by modifying the covariance matrix estimation in the case of missing values, and solved the resulting non-convex problem with an algorithm based on the projected gradient descent. However, this method assumes that \(l_1\) norm of the coefficients is bounded by a constant which depends on the sparsity, which is unknown.

\(^1\)https://rmisstastic.netlify.com
in practice. Zhao et al. (2017) suggested a method based on pseudo-likelihood with LASSO penalty, which can be used to select variables, but does not estimate the parameters. As for Liu et al. (2016), they combined penalized regression techniques with multiple imputation and stability selection.

This document is organized as follows: The section 2 presents the notations, assumptions and the ABSLOPE model. The section 3 describes the stochastic approximation EM algorithm for processing missing data. Here we also introduce a simplified version of the algorithm, which instead of drawing samples of parameters from their conditional distribution, updates them by the respective conditional expectations. The section 4 evaluates the methodology through comprehensive simulation studies, where we evaluate power, FDR and estimation bias. In the section 5, we apply our approach to medical data on trauma patients to establish a model to predict the rate of platelets in hospital using the (incomplete) information collected by the ambulance. Finally, the section 6 concludes our work and proposes a discussion.

2 Statistical model and assumptions

Let \( y = (y_i, 1 \leq i \leq n) \) be a vector of response of length \( n \); and \( X = (X_{ij}, 1 \leq i \leq n, 1 \leq j \leq p) \) a standardized design matrix of dimension \( n \times p \), i.e., each column has mean 0 and the unit \( l_2 \) norm, i.e., \( \sum_{i=1}^{n} X_{ij} = 0 \) and \( \sum_{i=1}^{n} X_{ij}^2 = 1 \), \( \forall j \ s.t. \ 1 \leq j \leq p \). The observed dataset \((X, y)\) contains \( n \) i.i.d. realizations of random variables. Consider estimating \( \beta \) in the linear regression model:

\[
y = X\beta + \varepsilon ,
\]

where \( \beta = (\beta_j, 1 \leq j \leq p) \) is the vector of regression coefficients of length \( p \), for which we assume a sparse structure; and \( \varepsilon \) is a vector of length \( n \) of Gaussian error with mean 0 and variance \( \sigma^2 \), i.e., \( \varepsilon \sim N(0, \sigma^2 I_n) \). Then we have that:

\[
y \mid X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n) . \tag{1}
\]
2.1 SLOPE

SLOPE (Bogdan et al., 2015) estimates coefficients by minimizing a regularized residual sum of squares using a sorted $l_1$ norm penalty which generalizes the $l_1$ regularization used in LASSO, by penalizing larger coefficients more stringently:

$$\hat{\beta}_{\text{SLOPE}} = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \| y - X \beta \|^2 + \sigma \sum_{j=1}^{p} \lambda_j |\beta|_{(j)},$$

(2)

where the penalty coefficients $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and the absolute values of elements in $\beta$ are sorted in decreasing order $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}$. The sorted $l_1$ penalty can also be written as:

$$\text{pen}(\lambda) = \sigma \sum_{j=1}^{p} \lambda_j |\beta|_{(j)} = \sigma \sum_{j=1}^{p} \lambda_{r(\beta,j)} |\beta_j|,$$

where $r(\beta, j) \in \{1, 2, \cdots, p\}$ is the rank of $\beta_j$ among elements in $\beta$ in descending order. To solve the convex but non-smooth optimization problem (2), a proximal gradient algorithm can be used as detailed in Bogdan et al. (2015). SLOPE formulation indicates that, the higher the rank, i.e., the stronger the signal, the larger the penalty. This behavior is quite similar to the Benjamini-Hochberg procedure (BH) (Benjamini and Hochberg, 1995), which compares more significant $p$-values with more stringent thresholds and SLOPE can be seen as building a bridge between LASSO and the False Discovery Rate (FDR) control for multiple testing. In the context of multiple regression we define FDR of an estimator $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ as

$$\text{FDR} = \mathbb{E} \left( \frac{V}{\max(1, R)} \right),$$

where

$$R = \# \{ j : \hat{\beta}_j \neq 0 \} \quad \text{and} \quad V = \# \{ j : \hat{\beta}_j \neq 0 \land \beta_j = 0 \}.$$

As shown in Bogdan et al. (2015), SLOPE with the sequence of parameters $\lambda_{\text{BH}} = (\lambda_{\text{BH},1}, \ldots, \lambda_{\text{BH},p})$, with

$$\lambda_{\text{BH},j} = \Phi^{-1} \left( 1 - j \times \frac{q}{2p} \right),$$

where $\Phi(\cdot)$ denotes the cdf of $\mathcal{N}(0, 1)$, controls the FDR at level $q$. 

7
2.2 Adaptive Bayesian SLOPE

As any penalized likelihood estimator, SLOPE can be seen as a posterior mode under the following prior (Sepehri, 2016):

\[
p(\beta | \sigma^2; \lambda) = C(\lambda, \sigma^2) \prod_{j=1}^{p} \exp\left( -\frac{1}{\sigma} \lambda r(\beta_j) | \beta_j | \right),
\]

where \( C(\lambda, \sigma^2) \) is a normalizing constant.

This prior depends on just one sequence of tuning parameters \( \lambda \), which regulates both the model selection and shrinkage. Simulation results reported in Bogdan et al. (2015) show that the selection of \( \lambda \) leading to FDR control leads also to over-excessive shrinkage and large estimation bias. To solve this problem we follow the idea of the Spike and Slab LASSO (SSL) (Ročková and George, 2018), where the prior is a scale mixture of two Laplace distributions. The spike component has a small value of the scale parameter and is used to model the noise, while the large value of the scale parameter of the slab component allows to model large signals. In this case the maximum a posteriori estimator (MAP) relies on performing the weighted LASSO, with the tuning parameters adjusted separately for each variable and dependent on the posterior probability that this variable is an important predictor. In SSL the approximation to the full Bayes approach is used for establishing the LASSO penalty. In adaptive Bayesian SLOPE (ABSLOPE), we propose a different hierarchical Bayesian model with the prior based on the sequence of SLOPE decaying parameters to provide FDR control and stabilize the estimators of large signals by additional shrinkage of regression parameters towards each other (see Brzyski et al. (2019) for some discussion of the SLOPE shrinkage). Compared to SSL, ABSLOPE is also closer to the full Bayes procedure and allows for the estimation of latent parameters like the signal sparsity or the average magnitude of the nonzero elements of \( \beta \). The procedure requires only three interpretable input parameters: FDR level \( q \) and the hyperparameters of the Beta prior for the sparsity level \( \theta \).

ABSLOPE prior for \( \beta \) is defined as:

\[
p(\beta | \gamma, c, \sigma^2; \lambda) \propto c^{\sum_{j=1}^{p} \mathbf{1}(\gamma_j = 1)} \prod_{j} \exp\left( -w_j | \beta_j | \frac{1}{\sigma} \lambda r(W_{\beta,j}) \right),
\]

which seems complicated but we explain how to form this prior as follows:
1. For $\beta_j \neq 0$, we interpret it as a signal. Otherwise it is considered to be noise.

2. $\gamma_j \in \{0, 1\}$ indicates whether $\beta_j$ is is substantially different from the noise level. $\gamma = (\gamma_1, \cdots, \gamma_p)^T$ indexes $2^p$ possible models, and we define its distribution as Bernoulli:

$$p(\gamma | \theta) = \prod_{j=1}^{p} \theta^{\gamma_j}(1 - \theta)^{1 - \gamma_j},$$

where $\theta = \mathbb{P}(\gamma_j = 1; \theta)$ is the expected fraction of large $\beta_j$, i.e., $\theta$ indicates the level of sparsity. We assume the Beta $\text{Beta}(a, b)$ for $\theta$, where the values of $a$ and $b$ can be selected by the user, according to an initial guess of the signal sparsity.

3. $c \in (0, 1]$ is the ratio of average signal magnitude between the null components and the non-null components. We assume a non-informative prior $c \sim \mathcal{U}[0, 1]$.

4. Denote a diagonal weighting matrix $W = \text{diag}(w_1, w_2, \cdots, w_p)$ and its diagonal element:

$$w_j = c\gamma_j + (1 - \gamma_j) = \begin{cases} c, & \gamma_j = 1 \\ 1, & \gamma_j = 0 \end{cases}.$$

5. In case when the noise variance $\sigma$ is unknown we estimate it using an uninformative prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$.

We give the following proposition to derive the prior (3).

**Proposition 1.** Assume that a random variable $z = (z_1, z_2, \cdots, z_p)$ has a SLOPE prior:

$$p(z | \sigma^2; \lambda) \propto \prod_{j=1}^{p} \exp\left\{-\frac{1}{\sigma} \lambda_r(z_j) | z_j \right\},$$

and then define $\beta = W^{-1}z = (\frac{z_1}{w_1}, \cdots, \frac{z_p}{w_p})$. Finally the prior of $\beta$ is given by (3). The detailed derivation is provided in Appendix A.1.

As a result, when $W$ is known and data are fully observed, the MAP for $\beta$, with prior (3) and likelihood (1), is given by solution of SLOPE (2) with a weighted design matrix $\tilde{X} = XW^{-1}$. Regarding the effect of introducing $W$, we observe that when $\gamma_j = 0$, $w_j = 1$, i.e., the null variables are treated with the regular SLOPE penalty; however, when $\gamma_j = 1$, $w_j = c < 1$, i.e., the variables classified as the true signals can be put in the end of sequence
of \((w_j|\beta_j|, 1 \leq j \leq p)\), which means that, with the same penalization coefficient \(\lambda\), we give
the true signal \(\beta_j\) a smaller penalty \(w_j\lambda r(W_{\beta,j})\) than the regular SLOPE one \(\lambda r(\beta,j)\). As a
result, compared to SLOPE, this adaptive version returns a more accurate estimation since
the shrinkage of \(l_1\) penalty on true signals are much smaller.

**Remark 1** (Full Bayesian). We can extend the model to a fully Bayesian version, by
specifying priors on \(\sigma^2\) and \(\theta\), for example, \(p(\sigma^2) \propto \frac{1}{\sigma^2}\) and \(\theta \sim\) Beta\((a,b)\). The value of \(a\)
and \(b\) can be given by users, according to an initial guess of the signal sparsity.

![Prior distribution of SLOPE and ABSLOPE, on \(\beta\) whose true value is non-null (a) or null (b).](image)

Figure 1 shows the difference between the SLOPE prior and ABSLOPE prior, on a
single coefficient \(\beta_j\) with the same penalty coefficient \(\lambda\). On the left is represented the
distribution of a non zero \(\beta_j\) which shows that ABSLOPE produces larger estimates: the
mass is greater in the tails compared to SLOPE; while for the zero \(\beta_j\) represented on the
right, ABSLOPE gives exactly the same results as SLOPE, with a double exponential peak
near 0 to threshold small effects.

The ABSLOPE prior can be seen as a spike and slab prior, so that the spike prior models
the regression coefficients close to the noise level and the slab component models large
regression coefficients. The algorithm described in Section 3.4 shows that the slab compo-
nent is estimated to debias the large regression coefficients while for the spike component
is selected to control the FDR.

## 2.3 Assumptions for missing values

We suppose that the missingness occurs only in the covariates $X$ but not in the response $y$. For each individual $i$, we denote $X_{i,\text{obs}}$ the observed elements of $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ and $X_{i,\text{mis}}$ the missing ones. We also decompose the matrix of covariates as $X = (X_{\text{obs}}, X_{\text{mis}})$, keeping in mind that the missing elements may differ from one individual to another. For each individual $i$, we define the missing data indicator vector $m_i = (m_{ij}, 1 \leq j \leq p)$, with $m_{ij} = 1$ if $X_{ij}$ is missing and $m_{ij} = 0$ otherwise. The matrix $m = (m_i, 1 \leq i \leq n)$ then defines the missing data pattern. The missing data mechanism is characterized by the conditional distribution of $m$ given $X$ and $y$, with parameter $\phi$, i.e., $p(m_i | X_i, y_i, \phi)$. In the literature on missing data (Little and Rubin, 2002), three mechanisms (Rubin, 1976) are available to describe the distribution of the missingness and code the different reasons for the missingness: 

i) Missing completely at random (MCAR): the absence is not related to any variable in the study; 

ii) Missing at random (MAR): the missing data depends only on the observed variables; 

iii) Missing not at random (MNAR): the absence depends on the value itself. Throughout this paper, we assume the MAR mechanism which implies that the missing values mechanism can therefore be ignored when maximizing the likelihood (Little and Rubin, 2002). A reminder of these concepts is given in the Appendix A.2.

We adopt a probabilistic framework by assuming that $X_i = (X_{i1}, \ldots, X_{ip})$ is normally distributed:

$$X_i \sim \mathcal{N}_p(\mu, \Sigma), \quad i = 1, \ldots, n.$$  

As assumed at the beginning of Section 2, the covariates should be standardized. Here we have to consider how to scale $X$ with existence of missing data. When the missing values are MCAR, the scaling can be performed as a pre-processing step before performing the analysis. Indeed, the observed values represent a random sample from the population, so that the standard deviations estimated using observed data only are unbiased estimates of the population standard deviation even if their variance is larger. When the missing data are MAR, standard deviations estimated using observed data can be severely biased. Indeed, consider a case where two variables are highly correlated and missing values occur
in one variable when the values of the other variable are larger than a constant, then the estimated standard deviation will be biased downward. Consequently, its estimation needs to be included in the analysis. We detail in the Appendix A.3, how we update mean and standard deviation at each iteration of the algorithm presented in Section 3.

2.4 Overview of modeling

Figure 2 shows ABSLOPE graphical model with the variables, parameters and their relation. We aims at estimating $\beta$ and $\sigma^2$ and there are nuisance parameters: $\mu$ and $\Sigma$.

Figure 2: ABSLOPE graphical model. Arrows indicate dependencies. White circles are for latent variables, gray ones for observed variables and squares for parameters.

3 Parameter estimation and model selection

First, we present the method ABSLOPE, which uses a stochastic approximation EM algorithm. This algorithm involves proper sampling steps, which is however time consuming. Therefore, we also present its heuristic simplified version called SLOB, where the stochastic step is replaced by deterministic approximations of parameter expected values. This version allows to consider models of larger dimensions and, according to our simulation study, performs very similarly to the stochastic version in lower dimensions.
3.1 Maximizing the observed penalized likelihood

According to the model defined in Section 2 and presented in Figure 2, the penalized complete-data log-likelihood can be written as:

\[
\ell_{\text{comp}} = \log p(y, X, \gamma, c; \beta, \theta, \sigma^2) + \text{pen}(\beta)
\]

\[
= \log \left\{ p(X | \mu, \Sigma) p(y | X; \beta, \sigma^2) p(\gamma | \theta) p(c) \right\} + \text{pen}(\beta)
\]

\[
= -\frac{1}{2} \log(2\pi|\Sigma|) - \frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) - n \log(\sigma) - \frac{1}{2\sigma^2} \|y - X\beta\|^2
\]

\[
+ \sum_{j=1}^p \mathbb{1}(\gamma_j = 1) \log \theta + \sum_{j=1}^p \mathbb{1}(\gamma_j = 0) \log(1 - \theta) - \frac{1}{\sigma} \sum_{j=1}^p w_j |\beta_j| \lambda_r(W_{\beta,j}).
\]

(4)

Note that we focus on getting the MAP for point estimates and do not adopt a fully Bayesian inference which aims at calculating the posterior distribution. Due the presence of latent variables \(X_{\text{mis}}, \gamma\) and \(c\), we estimate \(\beta\) by maximizing the observed log-likelihood which integrates over the latent variables: \(\ell_{\text{obs}} = \iint \ell_{\text{comp}} dX_{\text{mis}} dc d\gamma\). We use an EM algorithm (Dempster et al., 1977) to estimate \(\beta\), and in the meantime, obtain the simulated \(\gamma\) to distinguish the true signals from the noise, i.e., to select variables. Given the initialization, iteration \(t\) updates \(\beta^t\) to \(\beta^{t+1}\) with the following two steps:

- **E step:** The expectation of the complete-data log likelihood with respect to the conditional distribution of latent variables is computed, i.e.,

  \[
  Q^t = \mathbb{E}(\ell_{\text{comp}}) \quad \text{wrt} \quad p(X_{\text{mis}}, \gamma, c, \theta | y, X_{\text{obs}}, \beta^t, \sigma^t, \mu^t, \Sigma^t)
  \]

Since it is not tractable, we derive a stochastic approximation EM (SAEM) algorithm (Lavielle, 2014) by replacing the E step by a simulation step and a stochastic approximation step.

- **Simulation:** draw one sample \((X_{\text{mis}}^t, \gamma^t, c^t, \theta^t)\) from

  \[
  p(X_{\text{mis}}, \gamma, c, \theta | y, X_{\text{obs}}, \beta^{t-1}, \sigma^{t-1}, \mu^{t-1}, \Sigma^{t-1})
  \]

(5)

- **Stochastic approximation:** update function \(Q\) with

  \[
  Q^t = Q^{t-1} + \eta_t \left( \ell_{\text{comp}}|_{X_{\text{mis}}^t, \gamma^t, c^t, \theta^t} - Q^{t-1} \right),
  \]

(6)

where \(\eta_t\) is the step-size.

13
The step-size \( \eta_t \) is chosen as a decreasing sequence as described in Delyon et al. (1999) which ensures the almost sure convergence of SAEM to a maximum of the observed likelihood in their continuously differentiable case.

- **M step:** \( \beta^{t+1}, \sigma^{t+1}, \mu^{t+1}, \Sigma^{t+1} = \arg \max Q_t^{t+1} \).
  
  Note that \( \Sigma^{t+1} \) is estimated as above only when \( p << n \). Otherwise we consider a shrinkage estimation as discussed in Remark 2. Indeed, we regard \( (\mu, \Sigma) \) as auxiliary parameters, which are needed only to update the missing values.

Despite the apparent complexity of the algorithm, it turns out that the likelihood (4) can be decomposed into several terms: one term for the linear regression part, one term for the covariates distribution and terms for the latent variables \( \gamma \) and \( c \), as illustrated in Figure 2. Consequently one iteration can be divided into sub-problems, as detailed in the following subsections.

### 3.2 Simulation step: sampling the latent variables

To perform the simulation step (5), we use a Gibbs sampler. To simplify the notations, we hide the superscript, but note that all the conditional distributions are computed given the quantities from the previous iteration. We perform the following sampling procedure:

\[
\begin{align*}
\gamma & \sim \text{Bin} \left( \frac{\theta c \exp(-c \frac{2}{\Sigma} |\beta_j|^2 \lambda_r(W_{\beta,j}))}{(1-\theta) \exp(-\frac{1}{2} |\beta_j|^2 \lambda_r(W_{\beta,j})) + \theta c \exp(-c \frac{2}{\Sigma} |\beta_j|^2 \lambda_r(W_{\beta,j}))} \right) ; \\
\theta & \sim \text{Beta} \left( a + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1), b + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 0) \right) , \text{ with } \text{Beta}(a,b) \text{ a prior for } \theta ; \\
c & \sim \text{Gamma} \left( 1 + \frac{1}{2} \sum_{j=1}^{p} |\beta_j|^2 \lambda_r(W_{\beta,j}) \mathbb{1}(\gamma_j = 1) \right) \text{ truncated to } [0,1].
\end{align*}
\]

The detailed calculation and the interpretation can be found in Appendix A.4. In addition, to simulate the missing values \( X_{\text{mis}} \), we perform a decomposition:

\[
X_{\text{mis}} \sim p(X_{\text{mis}} | \gamma, c, y, X_{\text{obs}}, \beta, \sigma, \theta, \mu, \Sigma) = p(X_{\text{mis}} | y, X_{\text{obs}}, \beta, \sigma, \mu, \Sigma) \propto p(y | X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma) p(X_{\text{mis}} | X_{\text{obs}}, \mu, \Sigma).
\]

Here we observe that the target distribution (8) is normal distribution since the two terms after the factorization are both normal. In the following proposition, we give the explicit form of the target distribution as the solution of a system of linear equations.
Proposition 2. For a single observation \( x = (x_{\text{mis}}, x_{\text{obs}}) \) where \( x_{\text{obs}} \) and \( x_{\text{mis}} \) denotes observed and missing covariates respectively. Let \( \mathcal{M} \) be the set containing indexes for missing covariates and \( \mathcal{O} \) for the observed ones. Assume that \( p(x_{\text{obs}}, x_{\text{mis}}; \Sigma, \mu) \sim \mathcal{N}(\mu, \Sigma) \) and let \( y = x\beta + \varepsilon \) where \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \). For all the indexes of the missing covariates \( i \in \mathcal{M} \), we denote:

\[
\begin{align*}
m_i &= \sum_{q=1}^{p} \mu_j s_{iq}, \quad u_i = \sum_{k \in \mathcal{O}} x_{\text{obs}}^k s_{ik}, \quad r = y - x_{\text{obs}}\beta_{\text{obs}}, \quad \tau_i = \sqrt{s_{ii} + \beta_i^2 / \sigma^2}, \\
m_i - u_i &= \sum_{j \in \mathcal{M}, j \neq i} \frac{\beta_i \beta_j \sigma^2 + s_{ij} \tau_i \tau_j}{\tau_i \tau_j} \tilde{\mu}_j = \varphi_i, \quad \text{for all } i \in \mathcal{M},
\end{align*}
\]

with \( s_{ij} \) elements of \( \Sigma^{-1} \) and \( \beta_{\text{obs}} \) the observed elements of \( \beta \).

Let \( \bar{\mu} = (\bar{\mu}_i)_{i \in \mathcal{M}} \) be the solution of the following system of linear equations:

\[
\begin{align*}
\frac{r_i \beta_i \sigma^2 + m_i - u_i}{\tau_i} - \sum_{j \in \mathcal{M}, j \neq i} \frac{\beta_i \beta_j \sigma^2 + s_{ij} \tau_i \tau_j}{\tau_i \tau_j} \bar{\mu}_j = \bar{\mu}_i, \quad \text{for all } i \in \mathcal{M},
\end{align*}
\]

and let \( B \) be a matrix with elements:

\[
B_{ij} = \begin{cases} 
\frac{\beta_i \beta_j \sigma^2 + s_{ij}}{\tau_i \tau_j}, & \text{if } i \neq j \\
1, & \text{if } i = j
\end{cases}
\]

then for \( z = (z_i)_{i \in \mathcal{M}} \) where \( z_i = \tau_i x_{\text{mis}}^i \) we have:

\[
z \mid x_{\text{obs}}, y; \Sigma, \mu, \beta, \sigma^2 \sim \mathcal{N}(\bar{\mu}, B^{-1}).
\]

As a result, we can simulate missing covariates from:

\[
x_{\text{mis}} \mid x_{\text{obs}}, y; \Sigma, \mu, \beta, \sigma^2 \sim \mathcal{N}(\bar{\mu} \odot \tau, B^{-1} \odot (\tau \tau^T)),
\]

where \( \tau = (\tau_i)_{i \in \mathcal{M}} \odot \) is used for Hadamard division. The proof is provided in Appendix A.5.

3.3 Stochastic approximation and maximization steps

After the simulation step, we obtain one sample for each latent variable: \( X_{\text{mis}}^t, \gamma^t, c^t \), and thus \( W^t \) with diagonal element \( w_j^t = 1 - (1 - c^t) \gamma_j^t \). Now we have several parameters to estimate, but each parameter only concerns a part of terms in complete-data likelihood. This also helps us to simplify the calculation. But still, the maximization step is quiet difficult because the complete model doesn’t belong to a regular exponential family (if so we could update the sufficient statistics and more easily maximize).
Implementing SAEM seems quite challenging in the general step-size case. We first begin with the case when step-size $\eta_t = 1$. This algorithm is known as stochastic EM (SEM) (Celeux et al., 1995). Note that it causes larger variance compared to the step-size as decreasing sequence (Delyon et al., 1999). In addition, there is no guarantee for the convergence to the MLE, but only to its neighborhood.

### 3.3.1 Step-size $\eta_t = 1$

When step-size $\eta_t = 1$, the estimation exactly boils down to maximizing the complete-data likelihood completed by sampling the latent variables from their conditional distribution given the observed values.

1. Update $\beta$.

$$
\beta^t = \arg \max_{\beta} Q_1^t(\beta) := -\frac{1}{2(\sigma^{t-1})^2} \|y - X^t \beta\|^2 - \frac{1}{\sigma^{t-1}} \sum_{j=1}^p w_j^t |\beta_j^t| \lambda_{r(W^t \beta, j)},
$$

where $X^t = (X_{\text{obs}}, X_{\text{mis}})$. This estimate corresponds to the solution of SLOPE, given the value of $W$, $X_{\text{mis}}$ and $\sigma$ and can be solved using a proximal gradient algorithm (Bogdan et al., 2015).

2. Update $\sigma$.

$$
\sigma^t = \arg \max_{\sigma} Q_2^t(\sigma) := -n \log(\sigma) - \frac{1}{2\sigma^2} \|y - X^t \beta^t\|^2 - \frac{1}{\sigma} \sum_{j=1}^p w_j^t |\beta_j^t| \lambda_{r(W^t \beta, j)}.
$$

Given by the derivative, the solution to estimate $\sigma$ is:

$$
\sigma^t = \frac{1}{2n} \left[ \sum_{j=1}^p \lambda_{r(W^t \beta, j)} w_j^t |\beta_j^t| + \sqrt{\left( \sum_{j=1}^p \lambda_{r(W^t \beta, j)} w_j^t |\beta_j^t| \right)^2 + 4n \text{RSS}} \right], \quad (10)
$$

where the RSS (residual sum of squares) is $\|y - X^t \beta^t\|^2$.

If we omit the penalization term, (10) amounts to $\sigma^t = \sqrt{\frac{\text{RSS}}{n}}$, which is the classical formula for MLE of $\sigma$ when $\beta$ is also estimated by MLE. In this case this estimator would be biased downward. Interestingly, our posterior mode estimator of $\sqrt{n}\sigma$ is larger than the corresponding RSS, which, according to the simulation results in Subsection 4.2, often leads to a roughly less biased estimator if most of the true effects are detected by ABSLOPE.
3. Update $\mu, \Sigma$:

$$
\mu^t, \Sigma^t = \arg \max_{\mu, \Sigma} -\frac{1}{2} \log(2\pi|\Sigma|) - \frac{1}{2}(X^t - \mu)^T \Sigma^{-1}(X^t - \mu) .
$$

When $p \ll n$, the solution is given by the empirical means and empirical covariance matrix:

$$
\mu^t = \bar{X}^t = \frac{1}{n} \sum_{i=1}^{n} X^t_i \quad \text{and} \quad \Sigma^t = \frac{1}{n} \sum_{i=1}^{n} (X^t_i - \bar{X}^t)(X^t_i - \bar{X}^t)^T .
$$

In high dimensional setting, the estimation of $\Sigma^t$ by the empirical covariance matrix is replaced by a shrinkage estimation, as discussed in Remark 2.

**Remark 2.** To tackle the problem of estimation and inversion of the covariance matrix in high dimensions, one can resort to a shrinkage estimation as detailed in Ledoit and Wolf (2004). With the assumption that the ratio $\frac{n}{p}$ is bounded, they propose an optimal linear shrinkage estimator as a linear combination of identity matrix $I_p$ and the empirical covariance matrix $S$, i.e.:

$$
\hat{\Sigma} = \rho_1 I_p + \rho_2 S, \quad \text{where} \quad \rho_1, \rho_2 = \arg \min_{\rho_1, \rho_2} \mathbb{E} \| \hat{\Sigma} - \Sigma \|^2 .
$$

The method boils down to shrink the empirical eigenvalues towards their mean. The parameters $\rho_1$ and $\rho_2$ are chosen with asymptotically (as $n$ and $p$ go to infinity) uniformly minimum quadratic risk in its class.

### 3.3.2 General step-size

For the case with a general step-size, say $\eta_t = \frac{1}{t}$, here we propose: for any parameter $\psi$,

$$
\psi^{t+1} = \psi^t + \eta_t \left[ \hat{\psi}_{MLE}^t - \psi^t \right] ,
$$

where $\hat{\psi}_{MLE}^t$ is the MLE for complete-data likelihood completed by drawing the latent variables from their conditional distribution given the observed information, which corresponds exactly the estimate when $\eta_t = 1$ in Subsection 3.3.1. That means, we directly apply stochastic approximation formula on the parameters, instead of operating on likelihood as (6). When the likelihood (4) is a linear function of the parameters, the stochastic approximation step in equation (6) corresponds exactly to our proposal (11). In other situations it gives good results from empirical point of view.
3.4 SLOB: Quick version of ABSLOPE

The implementation of the steps in SAEM, described in the subsections 3.2 and 3.3, can still be costly in terms of computation time, even if the terms of the likelihood decompose well and we use an approximation (11). We therefore propose here a simplified version of the algorithm, called SLOB, which instead of drawing samples \((X_{\text{mis}}^t, \gamma^t, c^t, \theta^t)\) from their conditional distribution (5) in the simulation step, we approximate them by their conditional expectation, i.e.,

\[
(X_{\text{mis}}^t, \gamma^t, c^t, \theta^t) \sim E(X_{\text{mis}}^t, \gamma, c \mid y, X_{\text{obs}}^t, \beta^{t-1}, \sigma^{t-1}, \mu^{t-1}, \Sigma^{t-1}) ;
\]

To simplify the notations, we hide the superscript, but note that all the conditional expectations are computed given the quantities from the previous iteration.

1. Approximate \(\gamma_j\) by:

\[
\pi := E(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) = p(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) \tag{7}
\]

\[
\frac{\theta c \exp\left(-\frac{1}{\sigma} |\beta_j| \lambda_r(W_{\beta,j})\right)}{(1 - \theta) \exp\left(-\frac{1}{\sigma} |\beta_j| \lambda_r(W_{\beta,j})\right) + \theta c \exp\left(-\frac{1}{\sigma} |\beta_j| \lambda_r(W_{\beta,j})\right)} .
\]

2. Approximate \(\theta\) by:

\[
E(\theta \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, c, \mu, \Sigma, W) = E(\theta \mid \gamma, \beta, \sigma, W) \tag{7} \frac{a + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)}{a + b + p} , \tag{13}
\]

where \(a\) and \(b\) are fixed parameters in the prior of \(\theta\).

3. Approximate \(c\) by:

\[
E(c \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma, W) \tag{19} = \frac{\int_0^1 x^{a'} \exp(-b'x)dx}{\int_0^1 x^{a'-1} \exp(-b'x)dx} , \tag{14}
\]

where \(a' = 1 + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)\), \(b' = \frac{1}{\sigma} \sum_{j=1}^{p} |\beta_j| \lambda_r(W_{\beta,j}) \mathbb{1}(\gamma_j = 1)\).

4. In the case with missing values, for the \(i^{th}\) observation \(X_i\), approximate \(X_{i,\text{mis}}\) by:

\[
E(X_{i,\text{mis}} \mid \gamma, c, y, X_{i,\text{obs}}, \beta, \sigma, \theta, \mu, \Sigma) = E(X_{i,\text{mis}} \mid y, X_{i,\text{obs}}, \beta, \sigma, \mu, \Sigma) ,
\]

which is provided by Proposition 2.

18
Then, in step M, we maximize the likelihood of the complete data, as in Subsection 3.3.1. The impact of replacing the simulation step with conditional expectation is that we ignore the variability of latent variable sampling, which in a high dimensional setting helps to reduce noise of the algorithm, and which also leads to an acceleration of the algorithm as indicated by the simulation study in Subsection 4.5. We provide a summary of ABSLOPE and SLOB methods in Appendix A.6.

4 Simulation study

4.1 Simulation setting

To illustrate the performance of our methodology, we perform simulations by first generating data sets as follows:

1. A design matrix $X_{n \times p}$ is generated from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. The matrix is standardized, s.t., the mean of each column is 0 and its $l2$-norm is 1.

2. The signal magnitude is $c_0 \sqrt{2 \log p}$ when $c_0$ is large the signal strength is stronger. Only $k$ on the $p$ predictors are non-zero and all equal to $c_0 \sqrt{2 \log p}$.

3. the response vector is generated $y = X\beta + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I_n)$ and $\sigma = 1$ to start.

4. Missing values are entered into the design matrix using a MCAR or MAR mechanism. For the former, we randomly generate 10% of missing cells; for the later, we follow the multivariate amputation procedure proposed by Schouten et al. (2018).

We set the initialization and the hyperparameters as follows.

**Initialization** Appendix A.7 provides the default values we have taken for the following simulation studies. The algorithm is not sensible to the choices of values $a$ and $b$ (12), but initial values for $\beta$ may have a stronger impact. In practice, we use the LASSO estimates based on preliminary mean imputation (missing values replaced by the average of the observed values for each variable) to initialize the coefficients.

---

2This signal strength is inspired by the penalty coefficient of the Bonferroni method to control the family wise error rate (FWER): $\lambda_{Bonf} = \sigma \phi^{-1}(1 - \frac{\alpha}{2p}) = \sqrt{2\log p}$, for $p$ large and $\alpha$ fixed, say $\alpha = 0.05$. 

19
**Step-size** We set $\eta_t = 1$ for the first $t_0 = 20$ iterations to approach the neighborhood of the MLE, then, choose a positive decreasing sequence $\eta_t = \frac{1}{t-t_0}$ to approximate the MLE, with the stochastic approach formula (11).

$\lambda$ sequence A sequence of penalty coefficients $\lambda$ must be chosen before implementing the algorithm. As introduced in the Subsection 2.1, we use a BH sequence inspired by orthogonal designs:

$$
\lambda_{BH}(j) = \phi^{-1}(1 - q_j), \quad q_j = \frac{jq}{2p}, \quad j = 1, 2, \ldots, p.
$$

### 4.2 Convergence of SAEM

We first illustrate the convergence of SAEM. We set the size of design matrix as $n = p = 100$ while the number of true predictors is $k = 10$, the signal strength $3\sqrt{2\log p}$ and the percentage of missingness 10%. The covariance $\Sigma$ is an identity matrix to start.

![Figure 3](image_url)

Figure 3: Convergence plots for three coefficients with ABSLOPE (colored solid curves). Black dash lines represent the true value for each $\beta$. Estimates obtained with three different sets of simulated data are represented by three different colors.

Figure 3 shows the convergence of some coefficients with SAEM for three simulated data sets. These graphs are representative of all the observed results. There are large fluctuations...
during the first $t_0 = 20$ iterations, then after introducing the stochastic approximation at the 20th iteration, convergence is achieved gradually. Due to the existence of a sorted $l_1$ penalty, the estimates are still slightly biased.

In addition, we also represent the convergence curves for $\sigma$ with ABSLOPE in supplementary materials (Jiang et al., 2019a) in order to compare the estimate of $\sigma$ by ABSLOPE to the biased MLE estimator without prior knowledge, i.e., $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{\text{RSS}}{n}}$. We can see that the estimates of $\sigma$ with both methods are biased downward, but since ABSLOPE has an additional correction term (10), it leads to a less biased estimator.

4.3 Behavior of ABSLOPE - SLOB

We then evaluate ABSLOPE and SLOB in different setting of parametrization to see how the signal strength, the sparsity and other parameters influence their performances.

**Criterion** We apply ABSLOPE or SLOB on a synthetic dataset and get estimates for $\hat{\beta}$ and the sampled $\hat{\gamma}$ indicating the model selection results. We compare the model selected to the true one. The total number of true discoveries is $TP = \#\{j : |\beta_j| > 0 \text{ and } |\hat{\beta}_j| > 0\}$ and the total number of false discoveries is $FN = \#\{j : |\beta_j| > 0 \text{ and } \hat{\beta}_j = 0\}$.

To evaluate the performance, we consider the following quantities:

- Power $= \frac{TP}{TP+FN}$;
- FDR $= \frac{FP}{FP+TP}$;
- MSE of $\beta$ (Relative $l_2$ norm error) $= \frac{||\hat{\beta} - \beta||^2}{||\beta||^2}$;
- Relative prediction error $= \frac{||X\hat{\beta} - X\beta||^2}{||X\beta||^2}$.

For each set of parameters, we repeat the procedure 200 times: 
1) data generation 
2) estimation and model selection with ABSLOPE/SLOB 
3) evaluation with the criteria presented above and we compute the means over the 200 simulations. The simulations were implemented with parallel computing.
4.3.1 Scenario 1

We first consider \( n = p = 100 \) and vary:

- sparsity: number of true signal \( k = 5, 10, 15, 20; \)
- signal strength \( \sqrt{2 \log p}, 2 \sqrt{2 \log p}, 3 \sqrt{2 \log p}, 4 \sqrt{2 \log p}; \)
- percentage of missingness 0.1, 0.2, 0.3, generated randomly, i.e., MCAR;
- correlation between covariates \( \Sigma = \text{toeplitz}(\rho)^3 \) where \( \rho = 0, 0.5, 0.9. \)

Then we applied the Algorithm 1 on each synthetic dataset.

**Results 1: no correlation, 10% missingness - vary signal strength**  According to Figure 4:

- We observe that FDR is always controlled at expected level 0.1.
- Power increases and estimation bias decreases with larger sparsity or stronger signal.
- When the signal is too weak (signal strength = \( \sqrt{2 \log p} \)), the power is near 0, which is due to the identifiability issue that ABSLOPE cannot distinguish the signal from the noise. Indeed, the value \( c = \frac{\lambda_1}{\sigma \sqrt{2 \log p}} \) is greater than one where \( \lambda_1 \) is the largest penalization coefficient. In addition, the bias is significant. This behaviour can be explained by the fact that we choose the penalty \( \lambda \) to reduce the noise \( \sigma \); but when the signal is as weak as \( \sigma \), this choice of \( \lambda \) also "kills" the real signal.

---

\[ \Sigma = \begin{pmatrix}
1 & \rho & \cdots & \rho^{p-2} & \rho^{p-1} \\
\rho & 1 & \cdots & \rho^{p-2} & \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\rho^{p-2} & \cdots & \ddots & \rho & \\
\rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1
\end{pmatrix}_{p \times p}, \text{ where } \rho \in [0,1] \text{ is a constant. For the Toeplitz structure, adjacent pairs of covariates are highly correlated and those further away are less correlated, as in microarray study, genes are correlated due to their distance in the regularity pathway.} \]
Figure 4: Mean of power (a), FDR (b), bias of the estimate for $\beta$ (c) and prediction error (d), as function of length of true signal, over the 200 simulations. Results for $n = p = 100$, percentage of missingness 10% and $\Sigma$ orthogonal (no correlation).
Results 2: with correlation, strong signal - vary percentage of missingness  

Now we add the correlation as $\Sigma = \text{toeplitz}(\rho)$ where $\rho = 0.5$, and also fix a strong signal strength as $3\sqrt{2 \log p}$. We then vary the sparsity and percentage of missingness. The results in Figure 5 show that:

- The power increases and the estimation bias decreases when the percentage of missing data decreases.
- In the presence of correlation, the FDR control is slightly lost when the number of non-zero coefficients is greater than 10, but still near the nominal level.

Figure 5: Mean of power (a), FDR (b), bias of the estimate for $\beta$ (c) and prediction error (d), as function of length of true signal over the 200 simulations. Results for $n = p = 100$, with correlation and strong signal.
4.3.2 Scenario 2

Now we consider a larger dataset $n = p = 500$ and vary the same parametrization as Subsection 4.3.1, except the sparsity, for which we take wider range of choices among $k = 10, 20, 30, \ldots, 60$. In this scenario of larger dimension, we have applied the simplified SLOB algorithm as described in Subsection 3.4 to avoid intensive computation.

Results 1: no correlation, 10% missingness - vary signal strength

According to Figure 6:

![Graphs showing results](image)

(a) Power  
(b) FDR  
(c) Bias of $\beta$  
(d) Prediction error

Figure 6: Mean of power (a), FDR (b), bias of the estimate for $\beta$ (c) and prediction error (d), as function of length of true signal, over the 200 simulations. Results for $n = p = 500$, percentage of missingness 10% and $\Sigma$ orthogonal (no correlation).

- FDR is always controlled at expected level 0.1.

- Similar to Figure 4, power increases and estimation error decreases with larger sparsity and stronger signal. However in this larger dimension case, we can handle with
larger number of relevant features until 30 or 40, at which we observe a phase transition due to the identifiability issue.

Results 2: with correlation, strong signal - vary percentage of missingness  
Now we add the correlation as $\Sigma = \text{toeplitz}(\rho)$ where $\rho = 0.5$, and also fix a strong signal strength as $3\sqrt{2\log p}$. We then vary the sparsity and percentage of missingness. The results in Figure 7 show that:

Figure 7: Mean of power (a), FDR (b), bias of the estimate for $\beta$ (c) and prediction error (d), as function of length of true signal over the 200 simulations. Results for $n = p = 500$, with correlation and strong signal.

- Similar to Figure 5, the power increases and the estimation error decreases when the percentage of missing data decreases.
- Due to the existence of correlation, the FDR control is slight lost, especially in the less sparse and more missing case.
• With 10% missing values, if the number of relevant features is below 40, then we can always achieve an efficient power and perfect FDR control. With larger percentage of missing values, the sparsity of this changing point will be more conservative.

In addition, we present the results varying the correlations in the supplementary materials (Jiang et al., 2019a).

### 4.4 Comparison with competitors

We use the same simulation scenario and criteria as those used in the subsection 4.3 to compare ABSLOPE and SLOB to other approaches that can be considered to select variables in the presence of missing data.

- ncLASSO: Non-convex LASSO (Loh and Wainwright, 2012)
- Methods based on preliminary mean imputation (MeanImp): missing values are replaced by the average of the observed values for each variable, then on the completed data set is applied:
  - SLOPE: Applying two steps
    1. SLOPE (Bogdan et al., 2015)
    2. OLS on the selected predictors to estimate the parameters;
  - LASSO: LASSO with $\lambda$ selected by cross validation;
  - adaLASSO: adaptive LASSO (Zou, 2006);

For SLOPE, ABSLOPE and SLOB, we set the penalization coefficient $\lambda$ as BH sequence which controls the FDR at level 0.1. The values of the tuning parameters for the different methods can be found in the available code on GitHub (Jiang, 2019). We try to make the comparisons as fair as possible and also favor the competitors: we give the true $\sigma$ to SLOPE whereas we estimate it with ABSLOPE. ncLASSO requires to specify a bound on the $l_1$ norm of the coefficients, i.e., $\beta < R = b_0 \#\{\beta_j : \beta_j \neq 0\}$, for which we take the real value of sparsity and signal strength.

Note that we don’t compare with the widely used multiple imputation (van Buuren and Groothuis-Oudshoorn, 2011), where several imputed values are made for each missing
value to reflect the uncertainty in the missingness. There are several reasons, including the inability to perform model selection with multiple imputation and the difficulty to aggregate the estimates from the imputed datasets.

We present the results for the case $n = p = 100$ in the supplementary materials (Jiang et al., 2019a) while Figure 8 summarizes the result for the case $n = p = 500$, 10% missingness and with correlation toeplitz(0.5). Lighter colors indicate smaller values.

- ABSLOPE and SLOB both have a strong power and an accurate prediction, and FDR is always controlled.

- The power and FDR controlled achieved by ABSLOPE and SLOB are better than the case $n = p = 100$. On one hand, the correlation helps the generation of missing values, on the other hand, the sparsity considered here is less complicated.

- Other methods pay a price in FDR control to achieve good power.

### 4.5 Comparison of computation time

Table 1 presents the execution time of the different methods considered in the simulation. In addition, we have an implementation of proposed algorithm in C and use Rcpp (Eddelbuettel and Balamuta, 2017) to integrate these functions in R, to compare with the other methods. In the case $n = p = 100$, We observe that the most time consuming method is ncLASSO, which spent on average 20 seconds for one simulation. While ABSLOPE also took on average 14 seconds for one run, its simplified version SLOB reduced this cost to 0.6 seconds, which is comparable to MeanImp + adaLASSO. While when $n = p = 500$, the convergence of ABSLOPE requires much more time but SLOB helps to simplify the complexity. In addition, the version of C for SLOB is more accelerated, saving half of the computation time, which makes SLOB capable to handle with larger datasets.
Figure 8: Comparison of power (a), FDR (b), bias of $\beta$ (c) and prediction error (d) with varying sparsity and signal strength, with 10% missingness over 200 simulations in the case with correlation.
Table 1: Comparison of average execution time (in seconds) for one simulation, in the case without correlation and with 10% MCAR, for $n = p = 100$ and $n = p = 500$ calculated over 200 simulations. (MacBook Pro, 2.5 GHz, processor Intel Core i7)

| Execution time (seconds) for one simulation | $n = p = 100$ | $n = p = 500$ |
|---------------------------------------------|---------------|---------------|
|                                            | min mean max  | min mean max  |
| ABSLOPE                                    | 12.83 14.33 20.98 | 646.53 696.09 975.73 |
| SLOB                                       | 0.53 0.60 0.98 | 35.82 39.18 57.66 |
| SLOB (with Rcpp)                           | 0.31 0.34 0.66 | 14.23 15.07 29.52 |
| MeanImp + SLOPE                            | 0.01 0.02 0.09 | 0.24 0.28 0.53 |
| ncLASSO                                    | 16.38 20.89 51.35 | 91.90 100.71 171.00 |
| MeanImp + LASSO                            | 0.10 0.14 0.32 | 1.75 1.85 3.06 |
| MeanImp + adaLASSO                         | 0.45 0.58 1.12 | 45.06 47.20 71.24 |

5 Application to Traumabase dataset

5.1 Details on the dataset and preprocessing

Our work is motivated by an ongoing collaboration with the TraumaBase group\textsuperscript{4} at APHP (Public Assistance - Hospitals of Paris), which is dedicated to the management of severely traumatized patients. Major trauma is defined as any injury that endangers the life or the functional integrity of a person. The global burden of disease working group of the WHO has recently shown that major trauma in its various forms, including traffic accidents, interpersonal violence, self-harm, and falls, remains a public health challenge and a major source of mortality and handicap around the world (Hay et al., 2017). Effective and timely management of trauma is critical to improving outcomes. Delay, or errors in treatment have a direct impact on survival, especially for the two main causes of death in major trauma: hemorrhage and traumatic brain injury. Using a patient’s records measured in prehospital stage or on arrival at hospital, we aim to establish models of prediction to prepare an appropriate response upon arrival at the trauma center, e.g., massive transfusion protocol and/or immediate haemostatic procedures. Such models intend to give support

\textsuperscript{4}http://www.traumabase.eu/
to clinicians and professionals. Due to the highly stressful and multi-player environments involved, evidence suggests that patient management – even in mature trauma systems – often exceeds acceptable time frames (Hamada et al., 2014). In addition, discrepancies may be observed between the diagnoses made by emergency doctors in the ambulance, and those made when the patient arrives at the trauma center (Hamada et al., 2015). These discrepancies can result in poor outcomes such as inadequate hemorrhage control or delayed transfusion.

To improve decision-making and patient care, six trauma centers within the Ile de France region (Paris area) in France have collaborated to collect detailed high-quality clinical data from the accident scene, to the hospital. These centers have joined TraumaBase progressively between January 2011 and June 2015. Since then, data collection is exhaustive and covers the whole administrative area around Paris. The structure of the database integrates algorithm for consistency and coherence, and the data monitoring is performed by a central administrator. Sociodemographic, clinical, biological and therapeutic data (from the prehospital phase to the discharge if hospital) are systematically recorded for all trauma patients, and all patients transported in the trauma rooms of the participating centers are included in the registry. The resulting database: TraumaBase, a multicenter prospective Trauma registry, now has data from 7495 trauma cases with more than 250 variables, collected from January 2011 to March 2016, with age ranged from 12 to 96, and is continually updated. The granularity of collected data makes this dataset unique in Europe. However, the data is highly heterogeneous, as it comes from multiple sources, and furthermore, is often missing, which makes modeling challenging.

In the following, we focus on one specific challenge: establishing a statistical model with missing covariates, to predict the level of platelet on arrival at hospital, which helps propose an innovative response to the public health challenge of major trauma. The platelet is the cellular agent responsible for clot formation. An optimized organization is essential to control blood loss as quickly as possible and to reduce early mortality in severely traumatized patients. It is difficult to obtain the level of platelet in real time on arrival at hospital and its value would change the way of patient caring.

We focus on patients with primary origin (i.e., after the accident, the patients sent
directly to the hospital, but not sent to Emergency Care Unit). After this selection, 6384 patients remained in the data set. Based on clinical experience, to predict the level of platelet on arrival at hospital, 15 influential quantitative measurements were included as pre-selected variables. Detailed descriptions of these measurements are shown in the supplementary materials (Jiang et al., 2019a). These variables were chosen because they were all available to the pre-hospital team, and therefore could be used in real situations. In addition, by visualizing the individual map from PCA, we also detected some outliers and corrected errors of records, as detailed in the supplementary materials (Jiang et al., 2019a).

Figure 9 shows the percentage of missingness per variable, varying from 0 to 60%. If we perform complete case analysis (i.e., ignoring all the observations with missing values) on the dataset, in addition to the bias that can be introduced, only less than one third of the observations (1648 patients) still remain, which demonstrates the importance of taking appropriate account of missing data.

![Figure 9: Percentage of missing values in each pre-selected variable from TraumaBase.](image)

5.2 Model selection results

As a standard procedure in supervised learning, we divide the dataset into training and test sets. The training set contains a random selection of 80% of observations, and the test set contains the remaining 20%. First in the training set, we select a model and estimate
the parameters. We apply ABSLOPE and compare it with the same methods than those
described in Section 4, namely MeanImp + SLOPE, MeanImp + LASSO, MeanImp +
adaLASSO, MeanImp + SSL except ncLASSO since we do not known the sparsity and the
$l_1$ bound of coefficients. Moreover, we also include:

- BIC: Mean imputation followed by stepwise method based on BIC;
- RF: Mean imputation followed by random forest (Liaw and Wiener, 2002). This
  approach is assessed only for its prediction properties as it does not explicitly select
  variables.

In the SLOPE type methods, we set the penalization coefficient $\lambda$ as BH sequence which
controls the FDR at level 0.1. Since we consider that our design matrix is centered and
does not contain an intercept, we also center the vector of responses and let SLOPE type
methods work with $\tilde{y} = y - \bar{y}$, where $\bar{y}$ is the mean of $y$. We repeat the procedure of data
splitting (into training and test sets) 10 times and Table 2 shows that, over 10 replications,
how many times each variable is selected. In addition, we present in Table 3 whether
the selected variables by ABSLOPE have on average a positive or negative effect on the
platelet.

The TraumaBase medical team indicated that the signs of the coefficients were partially
in agreement with their a priori ideas: Large values of shock Index, vascular filling, blood
transfusion and lactate give a sign of greatly bleeding for the patients. All the others things
being equal, the more bleeding the patient, the less the rate of platelets. However, the effect
of delta Hemocue and heart rate on the platelet are not entirely in agreement with their
opinion.

5.3 Prediction performances

In supervised learning, after fitting a model using a training set, a natural step is to evaluate
the prediction performance, which can be done with a test set. Assuming $X = (X_{\text{obs}}, X_{\text{mis}})$
an observation in the test set, we want to predict the binary response $y$. One important
point is that test set also contains missing values, since the training set and the test
set have the same distribution (i.e., the distribution of covariates and the distribution of
Table 2: Number of times that each variable is selected over 10 replications. Bold numbers indicate which variables are included in the model selected by ABSLOPE.

| Variable     | ABSLOPE | SLOPE | LASSO | adaLASSO | BIC |
|--------------|---------|-------|-------|----------|-----|
| Age          | 10      | 10    | 4     | 10       | 10  |
| SI           | 10      | 2     | 0     | 0        | 9   |
| MBP          | 1       | 10    | 1     | 10       | 1   |
| Delta.hemo   | 10      | 10    | 8     | 10       | 10  |
| Time.amb     | 2       | 6     | 0     | 4        | 0   |
| Lactate      | 10      | 10    | 10    | 10       | 10  |
| Temp         | 2       | 10    | 0     | 0        | 0   |
| HR           | 10      | 10    | 1     | 10       | 10  |
| VE           | 10      | 10    | 2     | 10       | 10  |
| RBC          | 10      | 10    | 10    | 10       | 10  |
| SI.amb       | 0       | 0     | 0     | 0        | 0   |
| MBP.amb      | 0       | 0     | 0     | 0        | 0   |
| HR.max       | 3       | 9     | 0     | 1        | 0   |
| SBP.min      | 5       | 10    | 10    | 10       | 8   |
| DBP.min      | 2       | 10    | 2     | 1        | 0   |

missingness). Therefore, we can’t directly apply the fitted model to predict \( y \) from an incomplete observation of the test \( X \).

Our framework offers a natural way to tackle this issue by marginalizing over the distribution of missing data given the observed ones. More precisely, with \( S \) Monte Carlo samples \((X_{\text{mis}}^{(s)}, 1 \leq s \leq S) \sim p(X_{\text{mis}}|X_{\text{obs}})\), we estimate directly the response by maximum a posteriori:

\[
\hat{y} = \arg\max_y p(y|X_{\text{obs}}) = \arg\max_y \int p(y|X)p(X_{\text{mis}}|X_{\text{obs}})dX_{\text{mis}}
\]

\[
= \arg\max_y \mathbb{E}_{p(X_{\text{mis}}|X_{\text{obs}})} p(y|X)
\]

\[
= \arg\max_y \sum_{s=1}^{S} p\left( y|X_{\text{obs}}, X_{\text{mis}}^{(s)} \right). 
\]

Table 3: The effect of the selected variables by AB-SLOPE on the platelet. “+” indicates positive effect while “-” negative; 0 indicates insignificant variables.

| Variable    | Effect |
|-------------|--------|
| Age         | -      |
| SI          | -      |
| MBP         | 0      |
| Delta.Hemo  | +      |
| Time.amb    | 0      |
| Lactate     | -      |
| Temp        | 0      |
| HR          | +      |
| VE          | -      |
| RBC         | -      |
| SI.amb      | 0      |
| MBP.amb     | 0      |
| HR.max      | 0      |
| SBP.min     | 0      |
| DBP.min     | 0      |
Note that in the literature there are not many solutions to deal with the missing values in the test set (Josse et al., 2019). For those imputation based method, we imputed the test set with mean imputation and predicted the platelet by $\hat{y} = X^{\text{imp}} \hat{\beta}$. Finally we evaluate the relative $l_2$ prediction error: $\text{err} = \frac{\|\hat{y} - y\|^2}{\|y\|^2}$. Prediction results obtained are presented in Figure 10.

![Figure 10: Empirical distribution of prediction errors of different methods over 10 replications for the TraumaBase data. Results for SLOPE are not presented due to its large gap compared to others, with a mean of prediction error equals to 0.27.](image)

ABSLOPE’s performance is comparable to that of Random Forest and adaptive LASSO, and slightly better than that of the traditional stepwise regression and LASSO. There is a significant gap between the results of ABSLOPE and those of SLOPE. One of the possible reasons is that the classic version of SLOPE may encounter difficulties in the presence of correlation, while ABSLOPE allows to improve the estimation of parameters even with correlations. Random forests have excellent predictive capabilities which is consistent with the results of Josse et al. (2019) which shows the consistency of supervised learning, in the case of a learning method with an important learning capacity like Random Forest, even in the case of a simple imputation like mean imputation. However, it is difficult to interpret its results in terms of selected variables, which is crucial for physicians.
Figure 10 and Table 2 show that ABSLOPE and adaLASSO methods which have the best predictive capabilities select almost the same variables with adaLASSO selecting MBP (mean blood pressure) and ABSLOPE selecting SI (shock index). These two variables are highly correlated since both are measurements based on systolic blood pressure.

5.4 Results with interaction

We also consider a more complete model by adding second order interactions between the covariates in the model, which increases the dimensionality at $p = 55$. We apply the same procedure as used in the case without interaction and the predictive results are presented in Figure 4.

Table 5 shows which variables are selected more than 5 times out of the 10 replications. Results for SSL and SLOPE are not presented due to their large gap compared to others, with a mean of prediction error equals to 0.35 and 0.40 respectively. Again, ABSLOPE provides good results in terms of prediction while being sparse. We observe that when interactions are added, age often appears in combination with other variables. LASSO methods tended to include a larger number of variables with potentially higher false discovery rates. In brief, other methods, apart from ABSLOPE, have a tendency to overfit when interactions are present.

Table 4: Empirical distribution of prediction errors of different methods over 10 replications for the TraumaBase data, with interactions between each pair of variables.

Table 5: The variables selected more than 5 times out of the 10 replications, by each method. “∗” indicates the interaction between two variables.
6 Discussion

ABSLOPE penalizes larger coefficients more stringently to control FDR, meanwhile it applies a weighting matrix to improve the estimation. In addition, modeling in a Bayesian framework gives detailed structure of predictors as sparsity and signal strength and the SAEM algorithm allows to handle missing values. According to the simulation study, ABSLOPE is competitive to other methods in terms of power, FDR and prediction error. For the future research, we will consider the problem of high-dimensional model selection with missing values for categorical or mixed data. Dealing with other missing mechanisms as MNAR is also worth the efforts. Finally, the extension of SLOPE for logistic regression and other generalized linear model is also a subject to explore.

A Appendix

A.1 Deviation of prior (3) started from SLOPE prior

Proof of Proposition 1 is provided as follows.

Proof. We assume a random variable \( z = (z_1, z_2, \ldots, z_p) \) has a SLOPE prior:
\[
p(z \mid \sigma^2; \lambda) \propto \prod_{j=1}^{p} \exp \left\{-\frac{1}{\sigma} \lambda_r(z_j) | z_j | \right\},
\]
and then define \( \beta = W^{-1} z = \left( \frac{z_1}{w_1}, \ldots, \frac{z_p}{w_p} \right) \), or equally, \( z_j = \beta_j w_j \) where the diagonal elements in the weight matrix are \( w_j = c \gamma_j + (1 - \gamma_j) = \begin{cases} c, & \gamma_j = 1 \\ 1, & \gamma_j = 0 \end{cases}, \quad j = 1, 2, \ldots, p \). Then according to the transformation of variables, we have the prior distribution for \( \beta \):
\[
p(\beta \mid W, \sigma^2; \lambda) \propto \left| \det \left( \frac{dz}{d\beta} \right) \right| p_z(W \beta \mid W, \sigma^2; \lambda)
= \prod_{j=1}^{p} w_j \prod_{j=1}^{p} \exp \left\{-\frac{1}{\sigma} \lambda_r(W_{\beta,j}) | w_j \beta_j | \right\}
= c^{\Sigma_{j=1}^{p} 1(\gamma_j=1)} \prod_{j=1}^{p} \exp \left\{-w_j | \beta_j | \frac{1}{\sigma} \lambda_r(W_{\beta,j}) \right\},
\]
which corresponds to our proposed prior (3). \( \square \)
A.2 Missing mechanism

Missing completely at random (MCAR) means that there is no relationship between the missingness of the data and any values, observed or missing. In other words, for a single observation $X_i$, we have:

$$p(r_i \mid y, X_i, \phi) = p(r_i \mid \phi)$$

Missing at Random (MAR), means that the probability to have missing values may depend on the observed data, but not on the missing data. We must carefully define what this means in our case by decomposing the data $X_i$ into a subset $X_i^{(\text{mis})}$ of data that “can be missing”, and a subset $X_i^{(\text{obs})}$ of data that “cannot be missing”, i.e. that are always observed. Then, the observed data $X_{i,\text{obs}}$ necessarily includes the data that can be observed $X_i^{(\text{obs})}$, while the data that can be missing $X_i^{(\text{mis})}$ includes the missing data $X_{i,\text{mis}}$. Thus, MAR assumption implies that, for all individual $i$,

$$p(r_i \mid y_i, X_i; \phi) = p(r_i \mid y_i, X_i^{(\text{obs})}; \phi)$$

$$= p(r_i \mid y_i, X_i^{(\text{obs})}; \phi)$$

(15)

MAR assumption implies that, the observed likelihood can be maximize and the distribution of $r$ can be ignored (Little and Rubin, 2002). Assume that $\theta$ is the parameter to estimate. Indeed:

$$\mathcal{L}(\theta, \phi; y, X_{\text{obs}}, r) = p(y, X_{\text{obs}}, r; \theta, \phi) = \prod_{i=1}^{n} p(y_i, X_{i,\text{obs}}, r_i; \theta, \phi)$$

$$= \prod_{i=1}^{n} \int p(y_i, X_i, r_i; \theta, \phi) dX_{i,\text{mis}}$$

$$= \prod_{i=1}^{n} \int p(y_i, X_i; \theta) p(r_i \mid y_i, X_i; \phi) dX_{i,\text{mis}},$$

then according to the assumption MAR (15), we have:

$$\mathcal{L}(\theta, \phi; y, X_{\text{obs}}, r) = \prod_{i=1}^{n} \int p(y_i, X_i; \theta) p(r_i \mid y_i, X_{i,\text{obs}}; \phi) dX_{i,\text{mis}}$$

$$= \prod_{i=1}^{n} p(r_i \mid y_i, X_{i,\text{obs}}; \phi) \times \prod_{i=1}^{n} \int p(y_i, X_i; \theta) dX_{i,\text{mis}}$$

$$= p(r \mid y, X_{\text{obs}}; \phi) \times p(y, X_{\text{obs}}; \theta)$$

Therefore, to estimate $\theta$, we aim at maximizing $\mathcal{L}(\theta; y, X_{\text{obs}}) = p(y, X_{\text{obs}}; \theta)$. So the missing mechanism can be ignored in the case of MAR.
A.3 Standardization for MAR

We update mean and standard deviation at each iteration of algorithm.

1. Initialization: In the initialization step, we first substitute missing values $X_{\text{mis}}$ with the mean of non-missing entries in each column, and obtain a imputed matrix $\tilde{X}^0 = (X_{\text{obs}}, X_{\text{mis}}^0)$, where $X_{\text{mis}}^0$ contains imputed values. We denote the mean and standard deviation of each column of $X^0$, by the vectors $m^0$ and $s^0$ respectively. Then we centered and scaled the imputed $X^0$, s.t., for each observation $i$:

$$\hat{X}^0_i = \left( X^0_i - m^0 \right) \odot \left( \sqrt{n}s^0 \right),$$

where the $\odot$ is used for Hadamard division.

2. During $t^{th}$ iteration of the algorithm, we obtain a new imputed dataset $X^t = (X_{\text{obs}}, X_{\text{mis}}^t)$, where $X_{\text{mis}}^t$ contains imputed values in $t^{th}$ iteration. Then we first reverse scaling using:

$$\tilde{X}^t = (\sqrt{nt^{-1}}) \circ X^t + m^{-1}t,$$

where the $\circ$ is used for Hadamard product. The vectors $m^t$ and $s^t$ are then updated as the means and standard deviations of $\tilde{X}^t$. Finally we perform scaling on $\tilde{X}^t$ to obtain a scaled matrix:

$$\hat{X}^t_i = (\tilde{X}^t - m^t) \odot (\sqrt{nt^t}).$$

A.4 Details of the simulation step: sampling the latent variables

To perform the simulation step (5), we use a Gibbs sampler. To simplify the notations, we hide the superscript, but note that all the conditional distributions are computed given the quantities from the previous iteration.

1. Simulate $\gamma$. According to the dependency between variables presented in Figure 2, simulating the element $\gamma_j$ boils down to:

$$\gamma_j \sim p(\gamma_j \mid \gamma_{-j}, c, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma) = p(\gamma_j \mid \gamma_{-j}, c, \beta, \sigma, \theta),$$
where \( \gamma_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p) \); i.e., sampling from a Binomial distribution with probability:

\[
\mathbb{P}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta) = \frac{\mathbb{P}(\gamma_j = 1 \mid \theta) \mathbb{P}(\beta \mid \gamma_j = 1, \gamma_{-j}, c, \sigma)}{\sum_{\gamma_j \in \{0,1\}} \mathbb{P}(\gamma_j \mid \theta) \mathbb{P}(\beta \mid \gamma_j, \gamma_{-j}, c, \sigma)}
= \left[ 1 + \frac{(1 - \theta) \exp \left( -\frac{1}{\sigma} \beta_j \lambda_r(w^0_{\beta,j}) \right) \times (c) \sum_{-j} 1 \exp \left( -w^0_{-j} \beta_j \frac{1}{\sigma} \lambda_r(w^0_{\beta,-j}) \right)}{\theta c \exp \left( -\frac{1}{\sigma} \beta_j \lambda_r(W^1_{\beta,j}) \right) \times \sum_{-j} \exp \left( -w^1_{-j} \beta_j \frac{1}{\sigma} \lambda_r(W^1_{\beta,-j}) \right)} \right]^{-1}
= \left[ 1 + \frac{(1 - \theta) \exp \left( -\frac{1}{\sigma} \beta_j \lambda_r(W^0_{\beta,j}) \right) \times \sum_{-j} \exp \left( -w^0_{-j} \beta_j \frac{1}{\sigma} \lambda_r(w^0_{\beta,-j}) \right)}{\sum_{-j} \exp \left( -w^1_{-j} \beta_j \frac{1}{\sigma} \lambda_r(W^1_{\beta,-j}) \right)} \right]^{-1},
\]

(16)

where the weighting matrix \( W^1 \) and \( W^0 \) have the same diagonal elements \( w^1_{-j} = w^0_{-j} = 1 - (1 - c)\gamma_{-j} \), except for the position \( j \): \( w^1_j = c \) while \( w^0_j = 1 \). Sampling from (16) requires to store in memory ordered list which needs to be updated for every index \( j \), such an approach could be computationally exhaustive. So we use an approximation, which does not perturb solution significantly, by replacing both \( W^1 \) and \( W^0 \) by the estimate of weighting matrix from previous iteration, noted by \( W \). With the approximation, we partially retrieve the information of \( \gamma_j \) from the last iteration, so the difference between the estimates from last and the current iteration will be reduced. Consequently, (16) is drawn from:

\[
\mathbb{P}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) = \left[ 1 + \frac{(1 - \theta) \exp \left( -\frac{1}{\sigma} \beta_j \lambda_r(W_{\beta,j}) \right)}{\theta c \exp \left( -\frac{1}{\sigma} \beta_j \lambda_r(W_{\beta,j}) \right)} \right]^{-1}
= \frac{\theta c \exp \left( -\frac{1}{\sigma} \beta_j \lambda_r(W_{\beta,j}) \right)}{(1 - \theta) \exp \left( -\frac{1}{\sigma} \beta_j \lambda_r(W_{\beta,j}) \right) + \theta c \exp \left( -\frac{1}{\sigma} \beta_j \lambda_r(W_{\beta,j}) \right)},
\]

(17)

which can be interpreted as the posterior probability of binary signal indicator for \( j^{th} \) variable, given the prior guess \( \mathbb{P}(\gamma_j = 1 \mid \theta) = \theta \) and the conditional likelihood of the vector \( \beta \) given \( \gamma_j = 1 \) and \( \gamma_j = 0 \), see (3).

2. Simulate \( \theta \). The update of \( \theta \) boils down to generate from:

\[
\theta \sim \mathbb{P}(\theta \mid \gamma, c, y, X_{obs}, X_{mis}, \beta, \sigma, \Sigma, W)
= \mathbb{P}(\theta \mid \gamma, \beta, \sigma, W) \propto \mathbb{P}(\theta) \mathbb{P}(\gamma \mid \theta),
\]

40
where \( p(\gamma | \theta) \) is a Bernoulli distribution. In addition, if we also assume a prior for \( \theta \) as a Beta distribution \( Beta(a, b) \) with \( a \) and \( b \) known, to offer additional initial information for the sparsity of signal, then the posterior is:

\[
Beta\left(a + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1), b + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 0)\right),
\]

from which we can generate the latent variable \( \theta \). The target distribution (18) also takes the prior knowledge of the sparsity into consideration, for example:

- If \( a = \frac{n}{100} \) and \( b = \frac{n}{10} \), the prior mean on sparsity is 0.091, which has the same effect as a single observation;
- If \( a = \frac{2}{p} \) and \( b = 1 - \frac{2}{p} \), the prior mean on sparsity is \( \frac{2}{p} \), which assumes a sparse structure a priori.

3. Simulate \( c \). We also consider the weighting matrix \( W \) from the previous iteration.

\[
c \sim p(c | \gamma, y, X_{obs}, X_{mis}, \beta, \sigma, \theta, \mu, \Sigma, W)
\]

\[
= p(c | \gamma, \beta, \sigma, W) \propto p(c) p(\beta | c, \gamma, \sigma, W)
\]

\[
= p(c) c^{\sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)} \exp \left( -\frac{c}{\sigma} \sum_{j=1}^{p} |\beta|_{r}(W_{\beta,j}) \mathbb{1}(\gamma_j = 1) \right),
\]

where \( p(c) \) is the prior distribution of \( c \). If the prior is chosen as \( c \sim U[0,1] \) then we just need to sample from a Gamma distribution truncated to [0,1]:

\[
Gamma\left(1 + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1), \frac{1}{\sigma} \sum_{j=1}^{p} |\beta|_{r}(W_{\beta,j}) \mathbb{1}(\gamma_j = 1)\right).
\]

If the signal is strong enough, \( i.e., |\beta_j| \) is relative large compared to level of noise \( \sigma \) when \( \gamma_j = 1 \), we will observe that the most typical values from the above Gamma distribution fall in the interval \([0, 1]\). As a result, the simulation will be closer to the original Gamma distribution without truncation. However, if the signal strength go down, then the distribution will be more truncated and skewed towards 1, where \( c \) exactly corresponds the inverse of average signal magnitude.

### A.5 Proof of conditional distribution of missing data

Proof of Proposition 2 is provided as follows.
Proof. For a single observation $x = (x_{\text{mis}}, x_{\text{obs}})$ where $x_{\text{obs}}$, and $x_{\text{mis}}$ denotes observed and missing covariates respectively. Assume that $p(x_{\text{obs}}, x_{\text{mis}}; \Sigma, \mu)$ follows $\mathcal{N}(\mu, \Sigma)$ and let $y = x\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$. Then we have the following conditional distribution of the missing covariate with index $i$:

$$
p(x_{\text{mis}}^i \mid x_{\text{obs}}, y, \sigma, \beta, \Sigma, \mu) \propto p(x_{\text{mis}}^i \mid \Sigma, \mu) p(y \mid x_{\text{mis}}^i, x_{\text{mis}}^i, \beta, \sigma),
$$

where $x_{\text{mis}}^i = (x_{\text{mis}}^j, j \neq i)$. Denote $\mathcal{M}$ the set containing indexes for the missing covariates and $\mathcal{O}$ for the observed ones. We then explicitly give the formula, with $s_{ij}$ elements of $\Sigma^{-1}$:

$$
p(x_{\text{mis}}^i \mid x_{\text{obs}}, y, \sigma, \beta, \Sigma, \mu) \propto \exp\left[-\frac{1}{2\sigma^2}(y - x\beta)^2 - \frac{1}{2}(x - \mu)^\top\Sigma^{-1}(x - \mu)\right]$$

$$
\propto \exp\left[-\frac{1}{2\sigma^2}\left(y - x_{\text{obs}}\beta_{\text{obs}} - x_{\text{mis}}^i\beta_i - \sum_{j \in \mathcal{M}, j \neq i} x_{\text{mis}}^j\beta_j\right)^2 - \frac{1}{2}\left(s_{ii}(x_{\text{mis}}^i - \mu_i)^2 + 2x_{\text{mis}}^i \sum_{j \in \mathcal{M}, j \neq i} (x_{\text{mis}}^j - \mu_j)s_{ij} + 2x_{\text{mis}}^i \sum_{k \in \mathcal{O}} (x_{\text{obs}}^k - \mu_k)s_{ik}\right)\right].
$$

After rearranging terms, with notations:

$$
m_i := \sum_{q=1}^p \mu_q s_{iq}, \quad u_i := \sum_{k \in \mathcal{O}} x_{\text{obs}}^k s_{ik}, \quad r := y - x_{\text{obs}}\beta_{\text{obs}}, \quad \tau_i := \sqrt{s_{ii} + \frac{\beta_i^2}{\sigma^2}},
$$

we get:

$$
p(x_{\text{mis}}^i \mid x_{\text{obs}}, y, \sigma, \beta, \Sigma, \mu, x_{\text{mis}}^i) \propto \exp\left[-\frac{1}{2}\left((x_{\text{mis}}^i)^2 s_{ii} + \frac{\beta_i^2}{\sigma^2} - 2x_{\text{mis}}^i (r\beta_i + m_i - u_i) + 2x_{\text{mis}}^i \sum_{j \in \mathcal{M}, j \neq i} \frac{\beta_j \beta_j^\top}{\sigma^2 + s_{ij}} x_{\text{mis}}^j\right)\right] \tag{20}
$$

$$
\propto \exp\left[-\frac{1}{2}\left(x_{\text{mis}}^i \tau_i - \frac{r\beta_i}{\tau_i} + m_i - u_i + \sum_{j \in \mathcal{M}, j \neq i} \frac{\beta_j \beta_j^\top}{\tau_i \tau_j} x_{\text{mis}}^j\right)\right].
$$

By the other hand, we can conclude from equations (4.9) (4.10) in Besag (1974), that, for $z = (z_i)_{i \in \mathcal{M}}$ where $z_i = \tau_i x_{\text{mis}}^i$ we have:

$$
p(z_i \mid x_{\text{obs}}, y, \sigma, \beta, \Sigma, \mu, x_{\text{mis}}^i) \propto \exp\left[-\frac{1}{2}\left(z_i - \bar{\mu}_i + \sum_{j \in \mathcal{M}, j \neq i} B_{ij} (z_j - \bar{\mu}_j)\right)^2\right], \tag{21}
$$

and

$$
z \mid x_{\text{obs}}, y; \Sigma, \mu, \beta, \sigma^2 \sim N(\bar{\mu}, B^{-1}).
$$
Combine equations (20) and (21), we obtain the solution:

\[
\frac{r\beta_j / \sigma^2 + m_i - u_i}{\tau_i} - \sum_{j \in M, j \neq i} \frac{\beta_i \beta_j / \sigma^2 + s_{ij}}{\tau_i \tau_j} \bar{\mu}_j = \bar{\mu}_i , \quad \text{for all } i \in M ,
\]

and

\[
B_{ij} = \begin{cases} 
\frac{\beta_i \beta_j / \sigma^2 + s_{ij}}{\tau_i \tau_j} , & \text{if } i \neq j \\
1 , & \text{if } i = j
\end{cases} , \quad \text{for all } i, j \in M .
\]

\[\square\]

A.6 Summary of algorithms

We propose the ABSLOPE model and solve the problem of the maximization of the penalized likelihood using the SAEM algorithm, described in the Algorithm 1. We also give the SLOB algorithm in 2 which is an approximated and accelerated version.

A.7 Initialization of ABSLOPE

Here we suggest the following starting values:

• \(\beta^0\) is obtained from elastic net LASSO (Simon et al., 2011), or Spike and Slab LASSO (Ročková and George, 2018);

• \(X^0_{mis}\) are imputed by PCA (imputePCA) (Josse and Husson, 2016), or imputed by the mean of column (imputeMean);

• \(\mu^0\) and \(\Sigma^0\) are estimated with the empirical estimators obtained from the imputed initial data;

• \(\sigma^0\) is given by the standard deviation: \(\frac{\| y - X^0_{mis} \beta^0 \|}{\sqrt{n-1}}\);

• \(c^0 = \min \left\{ \left( \frac{\sum_{j=1}^{p} \beta_j^0}{\# \{j : |\beta_j^0| > 0\} + 1} \right)^{-1} \sigma^0 \lambda_{1/2} \right\} \), where the sign \# means the cardinality of a set. \(c^0\) can be interpreted as the inverse of average magnitude for the true signal, i.e, \(\beta_j^0 \neq 0\);

• \(\theta^0 = \frac{\# \{j : |\beta_j^0| > 0\} + a}{p + b}\) where \(a\) and \(b\) are known parameters of the prior Beta distribution on \(\theta\). Here we choose \(i) a = \frac{2}{p}\) and \(b = 1 - \frac{2}{p}\), such that the prior mean on sparsity is
Algorithm 1 Solving ABSLOPE with SAEM.

**Input:** Initialization $\beta^0, \sigma^0, c^0, \theta^0, X^0_{\text{mis}}, \mu^0, \Sigma^0$;

for $t = 1, 2, \cdots, \text{Maxit}$ do

(Simulation step)

1. Generate $\gamma^t$ from (17);

2. Generate $\theta^t$ from Beta distribution (18);

3. Generate $c^t$ from truncated Gamma distribution (19);

4. Generate $X^t_{\text{mis}}$ from Gaussian distribution (9);

(Stochastic Approximation step)

1. Calculate $(\beta_{\text{MLE}}^t, \sigma_{\text{MLE}}^t, \mu_{\text{MLE}}^t, \Sigma_{\text{MLE}}^t)$, which are the MLE for complete-data likelihood integrating sampled missing values, as detailed in Subsection 3.3.1;

2. With step-size $\eta_t = \begin{cases} 1, & \text{if } t \leq 20, \\ \frac{1}{t-20}, & \text{if } t > 20 \end{cases}$, update

   $\beta^{t+1} \leftarrow \beta^t + \eta_t [\beta_{\text{MLE}}^t - \beta^t]$.

   Update $\sigma$, $\mu$ and $\Sigma$ similarly;

if $\|\beta^{t+1} - \beta^t\|^2 < \text{tol}$ then

Stop;

end if

end for

**Output:** Estimates $\hat{\beta} \leftarrow \beta^t$ and indexes for model selection $\hat{\gamma} \leftarrow \gamma^t$
Algorithm 2 SLOB: a quick version of ABSLOPE.

**Input:** Initialization $\beta^0, \sigma^0, c^0, \theta^0, X_{\text{mis}}^0, \mu^0, \Sigma^0$;

**for** $t = 1, 2, \ldots, \text{Maxit}$ **do**

(Imputation by expectation)

1. **for** $j = 1, 2, \ldots, p$ **do** $\gamma^t_j \leftarrow \mathbb{E}(\gamma_j = 1 | \gamma_{-j}, c, \beta, \sigma, \theta, W)$, according to (12);

2. $\theta^t \leftarrow \mathbb{E}(\theta | \gamma, \beta, \sigma, W)$, according to (13);

3. $c^t \leftarrow \mathbb{E}(c | \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma, W)$, according to (14);

4. **for** $i = 1, 2, \ldots, n$ **do** $X_{i,\text{mis}}^t \leftarrow \mathbb{E}(X_{i,\text{mis}} | y, X_{i,\text{obs}}, \beta, \sigma, \mu, \Sigma)$, according to Proposition 2;

(Maximization of integrated likelihood)

- $(\beta^{t+1}, \sigma^{t+1}, \mu^{t+1}, \Sigma^{t+1}) \leftarrow (\beta_{\text{MLE}}^t, \sigma_{\text{MLE}}^t, \mu_{\text{MLE}}^t, \Sigma_{\text{MLE}}^t)$, which are the MLE for complete-data likelihood integrating the imputed missing values by expectation.

if $\|\beta^{t+1} - \beta^t\|^2 < \text{tol}$ **then**

Stop;

end if

end for

**Output:** Estimates $\hat{\beta} \leftarrow \beta^t$ and indexes for model selection $\hat{\gamma} \leftarrow \gamma^t$
\frac{2}{p}; \ ii) \ a = 0.01n \ and \ b = 0.01n; \ iii) \ a = 1 \ and \ b = p. \ Our \ estimation \ results \ are \ not 
sensible \ to \ the \ choice \ of \ hyperparameters \ a \ and \ b.

**Supplementary Material**

**package:** R-package ABSLOPE containing the implementation of the proposed methodology, available in Jiang et al. (2019b).

**Codes:** Code to reproduce the experiments are provided in Jiang (2019).

**Additional supplementary materials:** Some supplementary simulation results are presented in Jiang et al. (2019a).

**Acknowledgment**

This work was supported by grants from Région Île-de-France: https://www.dim-mathinnov.fr. The authors are thankful for fruitful discussion with Marc Lavielle, Imke Mayer, Geneviève Robin and Aude Sportisse.

**References**

Barber, R. F., Candès, E. J., et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Bellec, P., Lecué, G., and Tsybakov, A. (2018). Slope meets Lasso: improved oracle bounds and optimality. *Ann.Statist.*, 46(6B):3603–3642.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.
Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140.

Brzyski, D., Gossmann, A., Su, W., and Bogdan, M. (2019). Group SLOPE – adaptive selection of groups of predictors. *Journal of the American Statistical Association*, 114(525):419–433.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Celeux, G., Chauveau, D., and Diebolt, J. (1995). *On stochastic versions of the EM algorithm*. PhD thesis, INRIA.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, 64:1062–9.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Eddelbuettel, D. and Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, 5:e3188v1.

Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *Annals of Statistics*, 42(1):324–351.

Figueiredo, M. A. T. and Nowak, R. D. (2016). Ordered weighted $l_1$ regularized regression with strongly correlated covariates: Theoretical aspects. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, JMLR:W&CP*, 51:930–938.

Guo, Y., Hastie, T., and Tibshirani, R. (2006). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100.
Hamada, S. R., Gauss, T., Duchateau, F.-X., Truchot, J., Harrois, A., Raux, M., Duranteau, J., Mantz, J., and Paugam-Burtz, C. (2014). Evaluation of the performance of french physician-staffed emergency medical service in the triage of major trauma patients. *Journal of Trauma and Acute Care Surgery*, 76(6):1476–1483.

Hamada, S. R., Gauss, T., Pann, J., Dünser, M. W., Léone, M., and Duranteau, J. (2015). European trauma guideline compliance assessment: the ETRAUSS study. *Critical care*, 19:423.

Hay, S. I. et al. (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100):1260 – 1344.

Ibrahim, J., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, 103(484):1648–1658.

Jiang, W. (2019). Codes and implementations for ABSLOPE. [https://github.com/wjiang94/ABSLOPE/tree/master/ABSLOPE](https://github.com/wjiang94/ABSLOPE/tree/master/ABSLOPE).

Jiang, W., Bogdan, M., Josse, J., Miasojedow, B., Ročková, V., and Group, T. (2019a). Additional supplementary materials for "Adaptive Bayesian SLOPE – high-dimensional model selection with missing values". [https://github.com/wjiang94/ABSLOPE/tree/master/ABSLOPE/OnlineSupp](https://github.com/wjiang94/ABSLOPE/tree/master/ABSLOPE/OnlineSupp).

Jiang, W., Josse, J., Lavielle, M., and TraumaBase Group (2018). Logistic regression with missing covariates – parameter estimation, model selection and prediction within a a joint-modeling framework. *arXiv e-prints*. arXiv:1805.04602.

Jiang, W., Miasojedow, B., and Majewski, S. (2019b). ABSLOPE: a package for high-dimensional model selection with missing values. [https://github.com/wjiang94/ABSLOPE](https://github.com/wjiang94/ABSLOPE).
Josse, J. and Husson, F. (2016). missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31.

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. (2019). On the consistency of supervised learning with missing values. *arXiv e-prints*. arXiv:1902.06931.

Lavielle, M. (2014). *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*. Chapman and Hall/CRC.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365 – 411.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3):18–22.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

Liu, Y., Wang, Y., Feng, Y., and Wall, M. M. (2016). Variable selection and prediction with incomplete high-dimensional data. *Ann. Appl. Stat.*, 10(1):418–450.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664.

Mayer, I., Josse, J., Tierney, N., and Vialaneix, N. (2019). R-miss-tastic: a unified platform for missing values methods and workflows. *arXiv e-prints*. arXiv:1902.06931.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rejchel, W. and Bogdan, M. (2019). Fast and robust model selection based on ranks. *arXiv preprint 1905.05876*.

Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Annals of Statistics*, (46):401–437.

Ročková, V. and George, E. I. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444.
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581.

Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930.

Sepehri, A. (2016). The Bayesian SLOPE. arXiv:1608.08968.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Su, W., Bogdan, M., Candès, E., et al. (2017). False discoveries occur early on the Lasso path. *The Annals of Statistics*, 45(5):2133–2150.

Su, W. and Candès, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068.

Tardivel, P. J. and Bogdan, M. (2018). On the sign recovery given by the thresholded LASSO and thresholded Basis Pursuit. *arXiv preprint arXiv:1812.05723*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Zhao, J., Yang, Y., and Ning, Y. (2017). Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data, 28.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.