# Adaptive Bayesian SLOPE: Model Selection with Incomplete Data

Wei Jiang[1] Małgorzata Bogdan[2] Julie Josse[1] Szymon Majewski[3]
Błażej Miasojedow[4] Veronika Ročková[5] TraumaBase® Group[6]

October 2020

**Abstract**

We consider the problem of variable selection in high-dimensional settings with missing observations among the covariates. To address this relatively understudied problem, we propose a new synergistic procedure—adaptive Bayesian SLOPE with missing values—which effectively combines SLOPE (sorted $l_1$ regularization) with the Spike-and-Slab LASSO (SSL) and is accompanied by an efficient Stochastic Approximation of Expected Maximization (SAEM) algorithm to handle missing data. Similarly as in SSL, the regression coefficients are regarded as arising from a hierarchical model consisting of two groups: the spike for the inactive and the slab for the active. However, instead of assigning independent spike and slab Laplace priors for each covariate, here we deploy a joint SLOPE "spike and slab" prior which takes into account the ordering of coefficient magnitudes in order to control for false discoveries. We position our approach within a Bayesian framework which allows for simultaneous variable selection and parameter estimation while handling missing data. Through extensive simulations, we demonstrate satisfactory performance in terms of power, false discovery rate (FDR) and estimation bias under a wide range of scenarios including complete data and existence of missingness. Finally, we analyze a real dataset consisting of patients from Paris hospitals who underwent severe trauma, where we show competitive performance in predicting platelet levels. Our methodology has been implemented in C++ and wrapped into open source R programs for public use.

*Keywords:* incomplete data, FDR control, penalized regression, spike and slab prior, stochastic approximation EM, health data

[1]Inria XPOP and CMAP, École Polytechnique, France
[2]University of Wroclaw, Poland and Lund University, Sweden
[3]CMAP, École Polytechnique, France
[4]University of Warsaw, Poland
[5]University of Chicago Booth School of Business, USA
[6]Hôpital Beaujon, APHP, France

# Contents


**1 Introduction** **3**
1.1 Our Contributions . . . 6

**2 Model and Assumptions** **7**
2.1 SLOPE . . . 8
2.2 Adaptive Bayesian SLOPE . . . 9
2.3 Assumptions of missing values . . . 12
2.4 Overview of Modeling . . . 13

**3 Parameter Estimation and Model Selection** **14**
3.1 Maximizing the Observed Penalized Likelihood . . . 14
3.2 Simulation Step: Sampling the Latent Variables . . . 16
3.3 Stochastic Approximation and Maximization Steps . . . 17
3.3.1 Step-size $\eta_t = 1$ . . . 17
3.3.2 General Step-size . . . 19
3.4 SLOBE: Quick Version of ABSLOPE . . . 19
3.5 Details of implementation of ABSLOPE and SLOBE . . . 20

**4 Simulation study** **21**
4.1 Simulation setting . . . 21
4.2 Convergence of SAEM and comparison of ABSLOPE and SLOBE algorithms 22
4.3 Behavior of ABSLOPE with missing values . . . 23
4.4 Empirical comparison of SLOBE with other regularization methods for $n = p = 500$. . . . 26
4.4.1 Complete data . . . 28
4.4.2 With 10 % of Missing Data . . . 31

**5 Application to Traumabase dataset** **32**
5.1 Details on the dataset and preprocessing . . . 32
5.2 Model selection results . . . 37
5.3 Prediction performance . . . 40

**6 Discussion** **41**

**A Appendix** **47**
A.1 Deviation of prior (5) started from SLOPE prior . . . 47
A.2 Missing mechanism . . . 47
A.3 Standardization for MAR . . . 48
A.4 Details of the simulation step: sampling the latent variables . . . 49
A.5 Proof of conditional distribution of missing data . . . 51
A.6 Summary of algorithms . . . 54
A.7 Initialization of ABSLOPE . . . 54

# 1 Introduction

The problem of variable selection is ubiquitous in contemporary data applications. In molecular genetics, for instance, a vast number of predictors is available but only a few are deemed relevant for explaining biological phenomena. In high-dimensional data, variable selection can be plagued by the presence of missing values. For example, genetic data obtained from microarray experiments often contain missing values due to several reasons: insufficient resolution, image corruption, manufacturing errors, etc. This work develops a unified framework that tackles variable selection in such challenging scenarios. Our contributions are twofold: (1) proposal of a new penalized likelihood procedure (called adaptive Bayesian SLOPE) and (2) development of a variant of an EM algorithm for subset selection which can simultaneously handle missing covariate values. Below, we position our methodological contributions within the context of existing literature on high-dimensional variable selection and missing values.

The LASSO (Tibshirani, 1996), now a default penalized likelihood method, proved itself successful at simultaneously estimating parameters and covariate sets. Its theoretical guarantees include model selection consistency which, however, requires very stringent "irrepresentability" conditions on sparsity and the correlation structure between explanatory variables (see e.g., Zhao and Yu (2006); van de Geer and Bühlmann (2009); Wainwright (2009); Tardivel and Bogdan (2018)). To obtain good model selection properties, the penalty parameter $\lambda$ needs to be sufficiently large to discard irrelevant predictors. However, large $\lambda$ leads to underestimation of important regression coefficients and interference of their effects with even slightly correlated variables. As a result, false discoveries occur early along the LASSO path (see Su et al. (2017)) and often prevent this method from identifying the true model. To remedy these problems, adaptive LASSO (Zou, 2006) uses a weighted $\ell_1$ penalty with weights depending on some initial estimates of regression coefficients, leading to a smaller shrinkage of large effects. In this way, adaptive LASSO reduces estimation bias and can be consistent for variable selection even when irrepresentability is not satisfied (see *e.g.* Fan et al. (2014); Tardivel and Bogdan (2018); Rejchel and Bogdan (2019)). However, performance properties of adaptive LASSO still rely heavily on the weight function and tuning parameters, whose optimal choices depend on unknown aspects

of the estimation problem such as signal magnitude or sparsity. More recently, Ročková and George (2018) developed the Spike-and-Slab LASSO (SSL) procedure which bridges the default penalized likelihood approach (the LASSO) and the default Bayesian variable selection approach (spike-and-slab). In SSL, the penalty function arises from a fully Bayes spike-and-slab formulation and, as such, exerts self-adaptation properties with less hyper-parameter tuning required. In addition, SSL alleviates over-shrinkage of important signals by providing enough prior support for large effects. Theoretical results and simulations reported in Ročková and George (2018) and Ročková (2018) show that SSL attains near rate-minimax convergence (for the posterior mode *as well as* the entire posterior) and performs very well even when the columns in the design matrix are strongly correlated. In this article, we build on the Spike-and-Slab LASSO framework by incorporating aspects of the Sorted L-One Penalized Estimator (SLOPE) method of Bogdan et al. (2015). The main motivation behind SLOPE was False Discovery Rate (FDR) control, one of the central goals of methodological developments in multiple regression (see *e.g.* Barber et al. (2015); Candès et al. (2018)). Compared to methods aiming at perfect signal recovery, controlling for FDR is more liberal as it allows for some small number of mistakes. As a result, this leads to substantial gains in power and in prediction improvements when the signal is weak. As shown in Bogdan et al. (2015), SLOPE controls for FDR when the design matrix is orthogonal. Moreover, Su and Candès (2016) and Bellec et al. (2018) showed that, contrary to the LASSO, SLOPE allows one to achieve the exact minimax convergence rate for regression coefficients in sparse high dimensional regression. However, similarly as with the LASSO, it is challenging to attain good prediction and, at the same time, good variable selection with SLOPE in finite samples. Large amounts of shrinkage, needed to keep FDR small, result in large estimation bias of important regression coefficients and thereby poor estimation. One practical remedy, suggested by Bogdan et al. (2015); Brzyski et al. (2019), is proceeding in two steps: *i)* using SLOPE to detect relevant predictors; *ii)* applying standard least-squares with selected predictors for estimation. This two-step approach allows one to diminish the bias of SLOPE. However, there still remains the problem of the loss of FDR control, which typically occurs when the columns of the design matrix are correlated. This loss of FDR control results from over-shrinkage of large regression coefficients, whose

unexplained effect is often compensated by even slightly correlated "false" explanatory variables (see Su et al. (2017) for the theoretical analysis of the similar phenomenon for the LASSO). Combining SLOPE with SSL, we hope to address these issues be designing the Adaptive Bayesian version of SLOPE.

The second objective of our work has been to develop a variable selection framework for missing data among covariates. The most common practice of dealing with missing data is the list-wise deletion (complete case analysis), which confines the analysis to the observations with no missing attributes. This approach leads to estimation bias, unless the missing data are generated completely randomly, and to a huge information loss. Moreover, this approach is no longer feasible in a large-scale context, where even a small proportion of missing values for each explanatory variable could lead to elimination of the majority of observations. As Zhu et al. (2019) says: "One of the ironies of working with Big Data is that missing data play an ever more significant role, and often present serious difficulties for analysis." There is no shortage of literature on missing values management, e.g. see Little and Rubin (2019) and the platform `R-miss-tastic`[1] (Mayer et al., 2019) for an overview of the state of the art. However, there are only a few methods for selecting a model where the data contain missing values. For example, in generalized linear models, Claeskens and Consentino (2008); Ibrahim et al. (2008); Jiang et al. (2019) adapted likelihood-based information criteria designed for complete data such as AIC. However, their methods cannot process large data where the dimension $p$ is larger than (or comparable to) the sample size $n$. To handle high-dimensional incomplete data in linear models, Loh and Wainwright (2012) formulated a LASSO variant by modifying the covariance matrix estimation for 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 the $l_1$ norm of the vector of true regression coefficients is bounded by a constant which depends on the sparsity level rarely known in practice. In another related work, Zhao et al. (2017) suggested a pseudo-likelihood method with a LASSO penalty, which can be used to select variables, but does not estimate the parameters. Other extensions based on LASSO include a convex conditioned LASSO of Datta et al. (2017), with asymptotic sign-consistency, but

[1] `https://rmisstastic.netlify.com`

capable of handling only data missing completely at random. More recently, Descloux et al. (2020) focused on sign recovery by reframing missingness as a sparse corruption problem and then solving it with a LASSO-Zero method robust to missing not at random (MNAR) assumption (Little and Rubin, 2019). A simple alternative to perform variable selection with missing values could be to (1) first impute missing data and then (2) proceed with selection. To mitigate underestimation of variance stemming from single imputation (see e.g., Little and Rubin (2019)), results can be aggregated from multiple imputations (MI). However, different imputed datasets can return different models (different sets of variables) and the Rubin's rules (see Rubin (2009)) only serve for aggregating estimators of the same regression coefficients. An interesting solution is proposed in Liu et al. (2016) where penalized regression is combined with MI to give a probability of selection for each variable, followed by cross validation to find a cutoff for final selection. However, aggregating different models from the resulting multiple imputed data sets becomes complex when the number of variables in the data set is very large.

Despite these recent advances, model selection with missing values remains largely underdeveloped. Interesting theoretical guarantees are often obtained only under restrictive assumptions. Methodology for specific purposes, such as FDR control considered here, has not been explored yet with missing values.

## 1.1 Our Contributions

Our contributions bridge two seemingly unrelated areas: (1) penalized likelihood regression methodology and (2) missing data treatments. Our first contribution is the proposal of an adaptive Bayesian version of SLOPE (ABSLOPE) which builds on the Spike-and-Slab LASSO framework by incorporating aspects of the Sorted L-One Penalized Estimator (SLOPE). By embedding SLOPE within a Bayesian spike-and-slab framework, our prior is constructed so that the "spike" component effectively reduces to regular SLOPE for very small regression coefficients. Together with a bias-reducing slab for large signals, this allows for FDR control under a wide range of possible scenarios, as will be seen from our extensive simulation study. In addition, the "slab" component of our mixture prior preserves the averaging property of SLOPE for similar regression coefficients (see Figueiredo and Nowak (2016) for discussion of the SLOPE averaging effect). This leads to very good

prediction properties when regressors are substantially correlated. The computation with our mixture SLOPE prior is based on an algorithm with an Expectation Maximization (EM) spirit (Lavielle, 2014). While ABSLOPE is a standalone contribution, the nature of the computation (i.e. the EM algorithm) suggests a compelling possibility that it could be elegantly extended to missing data problems.

To address the missing covariates problem, we propose a stochastic approximation EM algorithm (Lavielle, 2014) to estimate the parameters of our model and, at the same time, deal with missing values. Missing data and hyper-parameters of the mixture SLOPE prior are iteratively updated inside the algorithm, automating the tuning of the prior. Additionally, we propose a computationally efficient approximation algorithm, *SLOBE*, where instead of generating from the respective conditional distributions, parameters are updated based on the approximation to their conditional expectation.

Our aim is to develop a complete and efficient methodology for selection of variables with high dimensional data and with incomplete data. The methodology has been implemented in R (R Core Team, 2018) programs ABSLOPE and SLOBE. The codes that reproduce all our experiments are available from GitHub (Jiang et al., 2021a).

This manuscript is organized as follows: Section 2 introduces notation and assumptions about our ABSLOPE model. Section 3 describes the stochastic approximation EM algorithm (and its simplified variant) for processing missing data. Section 4 evaluates the methodology with a comprehensive simulation study focusing on power, FDR and estimation bias while distinguishing the complete case and the incomplete case. In Section 5, we apply our approach to a medical dataset of trauma patients to develop a model that predicts the rate of platelets using (incomplete) medical information collected by the ambulance. Finally, Section 6 concludes our work with a discussion.

# 2 Model and Assumptions

Let $y = (y_i, 1 \leq i \leq n)$ be a vector of $n$ responses and $X = (X_{ij}, 1 \leq i \leq n, 1 \leq j \leq p)$ a design matrix of dimension $n \times p$ standardized so that each column has mean 0 and a unit $l_2$ norm, *i.e.* $\sum_{i=1}^{n} X_{ij} = 0$ and $\sum_{i=1}^{n} X_{ij}^2 = 1$ for $1 \leq j \leq p$. We consider the problem of estimating $\beta$

based on realizations $y$ from the linear regression model:

$$y = X\beta + \varepsilon, \tag{1}$$

where $\beta = (\beta_j, 1 \le j \le p)$ 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 independent Gaussian errors with mean 0 and variance $\sigma^2$, *i.e.* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

## 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 LASSO by penalizing larger coefficients more stringently:

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

where the penalty coefficients $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$ and the absolute values of elements in $\beta$ are sorted in a decreasing order $|\beta|_{(1)} \ge |\beta|_{(2)} \ge \cdots \ge |\beta|_{(p)}$. 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 a 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). Unlike in SSL, the SLOPE formulation operates under the following premise: 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. In this way, SLOPE can be seen as building a bridge between the 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

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

where

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

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

$$\lambda_{\text{BH},j} = \Phi^{-1}\left(1 - j \times \frac{q}{2p}\right), \tag{3}$$

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

## 2.2 Adaptive Bayesian SLOPE

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

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

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

This prior depends on just one sequence of tuning parameters $\lambda$, which regulates both model selection and shrinkage. Simulation results reported in Bogdan et al. (2015) show that the selection of $\lambda$ leading to FDR control also leads 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). SSL avoids over-shrinkage of large effects with a two-point Laplace mixture prior, where large coefficients can escape shrinkage by migrating towards the slab portion of the prior. The mixture prior formally writes as

$$\mathrm{p}(\beta \mid \gamma) = \prod_{j=1}^{p} \left[\gamma_j \phi_1(\beta_j) + (1 - \gamma_j)\phi_0(\beta_j)\right], \tag{4}$$

where $\phi_1(\beta_j) = 0.5\lambda_1 e^{-\lambda_1|\beta_j|}$ serves as a slab distribution for modeling large effects, $\phi_0(\beta_j) = 0.5\lambda_0 e^{-\lambda_0|\beta_j|}$ with $\lambda_0 >> \lambda_1$ is a spike distribution for modeling negligibly small effects, and $\gamma_j \in \{0,1\}$ is the indicator of the true signal. The spike component is assigned a large penalty $\lambda_0$ (small variance) to weed out noise, while the slab component has a small penalty $\lambda_1$ (large variance) to provide enough support for large signals. The Spike-and-Slab LASSO procedure is based on maximum a posteriori estimation (MAP) which relies on fast weighted LASSO calculations with weights automatically adjusted throughout the algorithm. Namely, separately for each variable, we have a penalty which depends on the (conditional) posterior probability that this variable is an important predictor. The SSL prior also automatically learns the level of sparsity through an empirical-Bayes plug-in inside the algorithm. The optimal choice of the spike penalty $\lambda_0$ relates to the prior mixing

weight $\theta$ and should reflect the inherent sparsity of the signal (Ročková, 2018). The SSL procedure does not rely on a single value $\lambda_0$ but, similarly as the LASSO, creates a solution path indexed by increasing values of $\lambda_0$. Typically, the path stabilizes when $\lambda_0$ increases and one can report a solution from this stable region. Since the SLOPE procedure was shown to be adaptive to the level of sparsity, we will replace the spike portion of the SSL prior with the Bayesian SLOPE prior to achieve more automatic sparsity adaptation (hoping for a reasonable FDR control).

In our adaptive Bayesian SLOPE (ABSLOPE), we thereby consider a different hierarchical Bayesian model with the spike prior based on the sequence of SLOPE decaying parameters to provide FDR control and to stabilize estimation of large signals by additional shrinkage of regression parameters towards one another (see Brzyski et al. (2019) for some discussion of the SLOPE shrinkage). ABSLOPE borrows strength across covariates (by tying them together through the spike distribution) and, similarly as SSL, allows for estimation of latent inclusion parameters and the level of sparsity (i.e. number of nonzero $\beta$ coefficients). The procedure requires only three interpretable input parameters: FDR level $q$ and the hyperparameters $a$ and $b$ of the Beta prior for the sparsity level $\theta \sim Beta(a, b)$.

The ABSLOPE prior on the regression vector $\beta$ is formally defined as:

$$\mathrm{p}(\beta \mid \gamma, c, \sigma^2; \lambda) \propto c^{\sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)} \prod_{j=1}^{p} \exp\left\{-w_j |\beta_j| \frac{1}{\sigma} \lambda_{r(W\beta, j)}\right\}. \tag{5}$$

This formulation may seem a bit complicated at first sight and so we carefully explain its components below:

1. Each $\beta_j \neq 0$ is regarded as signal and noise otherwise.

2. As is customary with spike-and-slab priors, each covariate $x_j$ is equipped with a binary inclusion indicator $\gamma_j \in \{0, 1\}$ which indicates whether $\beta_j$ is is substantially different from the noise level. The vector $\gamma = (\gamma_1, \cdots, \gamma_p)$ then indexes $2^p$ possible model configurations. Conditionally on a mixing (prior inclusion) weight $\theta \in (0, 1)$, we define the model distribution as an independent Bernoulli product:

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

where $\theta = \mathbb{P}(\gamma_j = 1; \theta)$ is formally defined as the expected fraction of large $\beta_j$, *i.e.*, $\theta$ indicates the level of sparsity. We assume that $\theta$ is either fixed or arose from a

beta distribution $Beta(a, b)$, where the values of $a$ and $b$ can be selected by the user, according to an initial guess of the signal sparsity.

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

4. We define a diagonal weighting matrix $W = \mathrm{diag}(w_1, w_2, \cdots, w_p)$ consisting of elements

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

5. For the case when the noise variance $\sigma$ is unknown, we can assume an uninformative prior $\mathrm{p}(\sigma^2) \propto \frac{1}{\sigma^2}$.

To motivate the prior (5), it is useful to note (detailed derivations is provided in Appendix A.1) that simple rescaling of coefficients drawn from (2.2) yields the desired spike-and-slab variant.

**Proposition 1.** *Assume that entries in a random vector $z = (z_1, z_2, \cdots, z_p)'$ have a SLOPE prior, i.e.*

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

*Then $\beta = W^{-1}z = (\frac{z_1}{w_1}, \cdots, \frac{z_p}{w_p})$ follows a prior given by (5).*

As a result, when $W$ is known (i.e. we know the signal and noise variables from $\gamma_j \in \{0, 1\}$) and when the data are fully observed, the MAP for $\beta$ under the ABSLOPE prior (5) can be obtained as a solution to SLOPE (2) with a weighted design matrix $\tilde{X} = XW^{-1}$. Let us now clarify the value of introducing the weighting matrix $W$. It turns out that when $\gamma_j = 0$ we have $w_j = 1$, *i.e.*, noise variables are treated with the regular SLOPE penalty. This penalty shrinks larger estimates more than the smaller ones, according to their expected values in the ordered sequence of estimates of noise regression coefficients. Since these expected values increase with $p$, our spike prior leads to the multiple testing adjustment, similar as in case of the Benjamini and Hochberg (1995) correction for multiple testing.

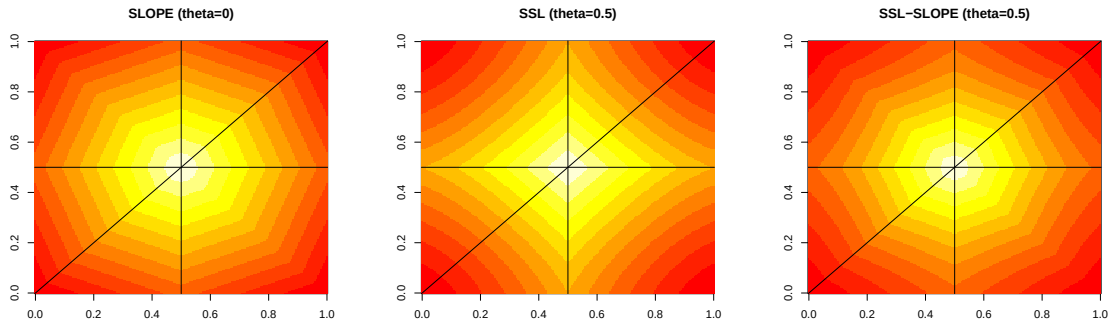


Figure 1: Heat-maps of the SLOPE prior (obtained with $\theta = 0$), SSL prior and ABSLOPE prior in two dimensions.

On the other hand, when $\gamma_j = 1$ we have $w_j = c < 1$ and the variables are treated as true signals and thereby not shrunk as much. This is achieved by multiplying the respective elements of the vector of tuning parameters by $c$ and, additionally, by moving these variables towards the end of sequence. This implies that, under ABSLOPE, the large effects $\beta_j$ will be assigned a penalty $c\lambda_{r(W\beta,j)}$ that is substantially smaller than $\lambda_{r(\beta,j)}$ obtained under the regular SLOPE. As a result, this adaptive version is poised to yield more accurate estimation due to the smaller shrinkage of large regression coefficients.

Figure 1 depicts heatmaps of the SLOPE, SSL and ABSLOPE priors for two coefficients. We can see how SSL puts a bit more mass on coordinate axes, supporting more the larger values (less shrinkage). The ABSLOPE prior is seen to retain the clustering property of SLOPE (putting a prior mass on the diagonal lines) but is more star shaped, again supporting larger values.

## 2.3 Assumptions of 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 $X_i = (X_{i1}, X_{i2}, \cdots, X_{ip})$ 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 $r_i = (r_{ij}, 1 \leq j \leq p)$, with $r_{ij} = 1$ if $X_{ij}$ is missing and $r_{ij} = 0$ otherwise. The matrix $r = (r_i, 1 \leq i \leq n)$ then defines the missing data pattern. The missing data mechanism is characterized by the conditional distribution

of $r$ given $X$ and $y$, with a parameter $\phi$, *i.e.*, $\mathtt{p}(r_i|X_i, y_i, \phi)$. In the literature on missing data (Little and Rubin, 2019), 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, 2019). 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:

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

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 deviations. However, they are more variable. 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 need to be included in the analysis. We detail in the Appendix A.3, how we update mean and standard deviation at each iteration of algorithm presented in Section 3.

## 2.4 Overview of Modeling

Figure 2 shows our ABSLOPE graphical model with variables, parameters and their relations. We aim at estimating $\beta$ and $\sigma^2$, treating parameters $\mu$ and $\Sigma$ as nuisance.

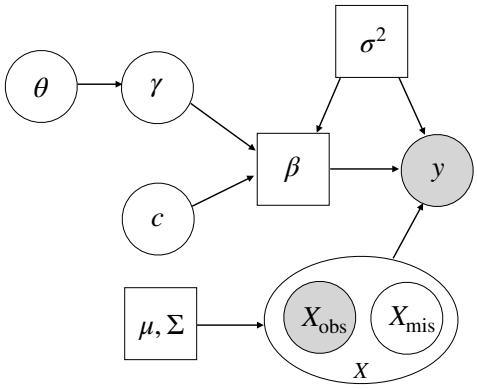


Figure 2: ABSLOPE graphical model. Arrows indicate dependencies. White circles are for latent variables, gray ones for observed variables and squares for parameters.

# 3 Parameter Estimation and Model Selection

In this section, we develop an ABSLOPE method based on the stochastic approximation EM algorithm. As this algorithm entails proper sampling which can be quite time consuming, we also provide a simplified heuristic version called SLOBE, where the stochastic step is replaced with deterministic approximations of parameter expected values. This faster variant allows us to consider models of larger dimensions and, according to our simulation study, performs very similarly to the stochastic version.

## 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:

$$\begin{aligned}
\ell_{\text{comp}} &= \log \mathtt{p}(y, X, \gamma, c;\, \beta, \theta, \sigma^2) + pen(\beta) \\
&= \log \left\{ \mathtt{p}(X \mid \mu, \Sigma)\, \mathtt{p}(y \mid X;\, \beta, \sigma^2)\, \mathtt{p}(\gamma \mid \theta)\, \mathtt{p}(c) \right\} + pen(\beta) \\
&= -\frac{1}{2}\log(2\pi|\Sigma|) - \frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) - n\log(\sigma) - \frac{1}{2\sigma^2}\|y - X\beta\|^2 \\
&\quad + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)\log\theta + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 0)\log(1-\theta) - \frac{1}{\sigma}\sum_{j=1}^{p} w_j|\beta_j|\lambda_{r(W\beta,j)}.
\end{aligned} \tag{6}$$

Similarly as the EMVS variable selection procedure of Ročková and George (2014), we focus on obtaining the MAP point estimates and do not aspire at fully Bayesian inference which would entail calculating the entire posterior distribution. Due to 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: $\ell_{\text{obs}} = \iiint \ell_{\text{comp}}\, dX_{\text{mis}}\, dc\, d\gamma$. We use the EM algorithm (Dempster et al., 1977) to estimate $\beta$, and in the meantime, obtain simulated $\gamma$ to distinguish the true signals from the noise, *i.e.* to select variables. Given the initialization, each 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.*,

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

  Since this 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 $(X^t_{\text{mis}}, \gamma^t, c^t, \theta^t)$ from

$$\mathtt{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^{t-1}, \sigma^{t-1}, \mu^{t-1}, \Sigma^{t-1})\ ; \tag{7}$$

  - *Stochastic approximation:* update function Q with

$$Q^t = Q^{t-1} + \eta_t \left( \ell_{\text{comp}}\Big|_{X^t_{\text{mis}}, \gamma^t, c^t, \theta^t} - Q^{t-1} \right)\ , \tag{8}$$

    where $\eta_t$ is the step-size.

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

- *M step:* $(\beta^{t+1}, \sigma^{t+1}, \mu^{t+1}, \Sigma^{t+1}) = \arg\max Q^{t+1}$.
  Note that $\Sigma^{t+1}$ is estimated as above only when $p << n$. Otherwise we consider a shrinkage estimation as discussed in Remark 1. 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 (6) 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 tractable sub-problems, as detailed in the following subsections.

### 3.2 Simulation Step: Sampling the Latent Variables

To perform the simulation step (7), we use the Gibbs sampler. To simplify notation, we hide the superscript and note that all conditional distributions are computed given the quantities from the previous iteration. We perform the following sampling procedure:

$$\begin{cases} \gamma \sim Bin\left(\frac{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)+\theta c\exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}\right); \\ \theta \sim Beta\left(a+\sum_{j=1}^{p}\mathbb{1}(\gamma_j=1), b+\sum_{j=1}^{p}\mathbb{1}(\gamma_j=0)\right), \text{ with } Beta(a,b) \text{ a prior for } \theta; \\ c \sim Gamma\left(1+\sum_{j=1}^{p}\mathbb{1}(\gamma_j=1), \quad \frac{1}{\sigma}\sum_{j=1}^{p}|\beta_j|\lambda_{r(W\beta,j)}\mathbb{1}(\gamma_j=1)\right) \text{ truncated to } [0,1]. \end{cases} \tag{9}$$

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

$$\begin{aligned} X_{\text{mis}} &\sim \mathtt{p}(X_{\text{mis}} \mid \gamma, c, y, X_{\text{obs}}, \beta, \sigma, \theta, \mu, \Sigma) \\ &= \mathtt{p}(X_{\text{mis}} \mid y, X_{\text{obs}}, \beta, \sigma, \mu, \Sigma) \\ &\propto \mathtt{p}(y \mid X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma)\, \mathtt{p}(X_{\text{mis}} \mid X_{\text{obs}}, \mu, \Sigma)\,. \end{aligned} \tag{10}$$

Here, we observe that the target distribution (10) is a normal distribution since the two terms after factorization are both normal. In the following proposition, we give the explicit form of the target distribution as a solution to a system of linear equations.

**Proposition 2.** *For a single observation $x = (x_{\text{mis}}, x_{\text{obs}})$ we denote with $x_{\text{obs}}$ and $x_{\text{mis}}$ 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 $\varepsilon \sim N(0, \sigma^2)$. For all the indexes of the missing covariates $i \in \mathcal{M}$, we denote:*

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

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

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

$$\frac{r\beta_i/\sigma^2 + m_i - u_i}{\tau_i} - \sum_{j \in \mathcal{M}, j \neq i} \frac{\beta_i\beta_j/\sigma^2 + s_{ij}}{\tau_i\tau_j}\tilde{\mu}_j = \tilde{\mu}_i\,, \quad \text{for all } i \in \mathcal{M}\,, \tag{11}$$

*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 $z_i = \tau_i x^i_{\text{mis}}$ we have:*

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

*As a result, we can simulate missing covariates from:*

$$x_{\text{mis}} \mid x_{\text{obs}}, y; \Sigma, \mu, \beta, \sigma^2 \sim N(\tilde{\mu} \oslash \tau, B^{-1} \oslash (\tau \tau^T)) ,$$

*where $\tau = (\tau_i)_{i \in \mathcal{M}}$ $\oslash$ 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: $X^t_{\text{mis}}, \gamma^t$, $c^t$, and thus $W^t$ with diagonal elements $w^t_j = 1 - (1 - c^t)\gamma^t_j$. Now we have several parameters to estimate, but each parameter only concerns some of the terms in the complete-data likelihood. This helps us simplify calculations. The maximization step is nevertheless quite difficult because the complete model does not belong to a regular exponential family (if so we could update the sufficient statistics and maximize more easily).

As the implementation of SAEM is quite challenging in the general step-size case, we start with the simpler case of fixed step-size $\eta_t = 1$. It is important to note that this causes larger variance compared to setting the step-size as a decreasing sequence Delyon et al. (1999) and there is no guarantee of convergence to the actual mode, only to its neighborhood.

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

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

1. Update $\beta$.

$$\beta^t = \arg\max_{\beta} Q^t_1(\beta) := -\frac{1}{2(\sigma^{t-1})^2} \|y - X^t \beta\|^2 - \frac{1}{\sigma^{t-1}} \sum_{j=1}^{p} w^t_j |\beta_j| \lambda_{r(W^t \beta, j)} ,$$

where $X^t = (X_{\text{obs}}, X^t_{\text{mis}})$. This estimate corresponds to the solution of SLOPE, given the value of $W$, $X_{\text{mis}}$ and $\sigma$. In our implementation of ABSLOPE we solve the SLOPE optimization problem using the Alternative Direction Method of Multipliers of Boyd et al. (2011), which turns out to be much quicker than the proximal gradient algorithm of Bogdan et al. (2015) when the regressors are strongly correlated or when they are on different scales, as in our reweighting scheme.

2. When $\sigma$ is unknown it may be updated according to the formula

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

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

$$\sigma^t = \frac{1}{2n}\left[\sum_{j=1}^{p}\lambda_{r(W^t\beta^t,j)}w_j^t|\beta_j^t| + \sqrt{\left(\sum_{j=1}^{p}\lambda_{r(W^t\beta^t,j)}w_j^t|\beta_j^t|\right)^2 + 4n\text{RSS}}\right] , \tag{12}$$

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

If we omit the penalization term, (12) amounts to $\sigma^t = \sqrt{\frac{RSS}{n}}$, which is the classical formula for MLE of $\sigma$ when $\beta$ is also estimated by MLE. In this case this estimator would be biased downwards. Interestingly, our posterior mode estimator of $\sqrt{n}\sigma$ is larger than the corresponding RSS, which, according to our simulation results often leads to a less biased estimator when most of the true effects are detected by ABSLOPE.

3. Update $\mu, \Sigma$:

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

When $p << n$, the solution is given by the empirical mean and the empirical covariance matrix:

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

In high dimensional setting, estimation of $\Sigma^t$ by the empirical covariance matrix is replaced by shrinkage estimation, as discussed in Remark 1.

**Remark 1.** *To tackle the problem of estimation and inversion of the covariance matrix in high dimensions, one can resort to 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.:*

$$\hat{\Sigma} = \rho_1 I_p + \rho_2 S, \qquad \textit{where } \rho_1, \rho_2 = \underset{\rho_1,\rho_2}{\arg\min}\, \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \ .$$

*The method boils down to shrinking 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

With a general step-size (say $\eta_t = \frac{1}{t}$), for a model parameter $\psi$ we set

$$\psi^{t+1} = \psi^t + \eta_t \left[\hat{\psi}^t_{MLE} - \psi^t\right] \ , \tag{13}$$

where $\hat{\psi}^t_{MLE}$ is the MLE estimator of the complete-data likelihood completed by drawing the latent variables from their conditional distributions given the observed information. This exactly corresponds to the estimate in Subsection 3.3.1 when $\eta_t = 1$. In other words, we apply stochastic approximations on the model parameters, instead of directly operating on the likelihood in (8). When the likelihood (6) is a linear function of the parameters, the stochastic approximation step in equation (8) corresponds exactly to our proposal (13). In other situations, it gives good results from an empirical point of view.

## 3.4 SLOBE: Quick Version of ABSLOPE

The implementation of SAEM, as described in Subsection 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 the approximation (13). We therefore propose a simplified version of the algorithm, called SLOBE, which instead of drawing samples $(X^t_{\text{mis}}, \gamma^t, c^t, \theta^t)$ from their conditional distribution (7) in the simulation step, approximates them by their conditional expectation, i.e.,

$$(X^t_{\text{mis}}, \gamma^t, c^t, \theta^t) \leftarrow \mathbb{E}(X_{\text{mis}}, \gamma, c \,|\, y, X_{\text{obs}}, \beta^{t-1}, \sigma^{t-1}, \mu^{t-1}, \Sigma^{t-1}) \ ;$$

To simplify notation, 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:

$$\begin{aligned}\pi &:= \mathbb{E}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) = p(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) \\ &\overset{(9)}{=} \frac{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right) + \theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)} .\end{aligned} \tag{14}$$

2. Approximate $\theta$ by:

$$\mathbb{E}(\theta \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, c, \mu, \Sigma, W) = \mathbb{E}(\theta \mid \gamma, \beta, \sigma, W) \overset{(9)}{=} \frac{a + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)}{a + b + p} , \tag{15}$$

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

3. Approximate $c$ by:

$$\mathbb{E}(c \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma, W) \overset{(22)}{=} \frac{\int_0^1 x^{a'} \exp(-b'x)dx}{\int_0^1 x^{a'-1} \exp(-b'x)dx} , \tag{16}$$

   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 $i^{\text{th}}$ observation $X_i$, approximate $X_{i,\text{mis}}$ by:

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

   which is provided by Proposition 2.

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 a conditional expectation is that we ignore the variability of latent variable sampling, which in high dimensional settings helps reduce noise of the algorithm, and which also leads to accelerations as shown in our simulation study in the supplementary materials (Jiang et al., 2021b). We provide a summary of ABSLOPE and SLOBE methods in Appendix A.6.

## 3.5 Details of implementation of ABSLOPE and SLOBE

**Standardization** In our simulation studies and the real life application of ABSLOPE we decided to not penalize the intercept term, which is estimated by the average value of the response variable in the training sample. The remaining parameters of our regression model are then estimated by running ABSLOPE using the centered values of the response variable and the centered and standardized design matrix $X$, as assumed in Section 2.

**Initialization** Appendix A.7 provides the default initialization and prior parameters we have taken for the following simulation studies. The algorithm is not sensitive to the choice of values $a$ and $b$ (14), but initial values for $\beta$ may have a stronger impact, particularly when $\beta$ has many small non-zero elements. Based on extensive simulation studies we recommend to start from the cross-validated LASSO estimates, based on preliminary imputation by PCA implemented in *missMDA* package(Josse and Husson, 2016), or by the Multivariate Imputation by Chained Equations from the *mice* package (van Buuren and Groothuis-Oudshoorn, 2011).

**Step-size** For ABSLOPE we set $\eta_t = 1$ for the first $t_0 = 20$ iterations to approach the neighborhood of the MAP estimator, then, choose a positive decreasing sequence $\eta_t = \frac{1}{t-t_0}$ to approximate the MAP, with the stochastic approach formula (13).

# 4 Simulation study

## 4.1 Simulation setting

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

• A design matrix $X_{n\times p}$ is generated from a multivariate normal distribution $\mathcal{N}(0,\Sigma)$ , with all diagonal elements of the covariance matrix equal to $1/n$.

• The response variable is generated from the model

$$Y = X\beta + \epsilon,$$

where $\epsilon \sim N(0, I_n)$ and the nonzero elements of $\beta$ are of the form $c\sqrt{2\log p}$. This signal strength is inspired by an observation that for the orthogonal design matrix $X'X = I$ the maximal least square estimator over the false predictors is close to $\sqrt{2\log p}$. Thus, when $c > 3$ then signal is strong when compared to the background noise, while values of $c$ close to 1 correspond to signals which are barely distinguishable from the noise.

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

## 4.2 Convergence of SAEM and comparison of ABSLOPE and SLOBE algorithms

We first illustrate the convergence of SAEM and compare the performance of ABSLOPE and SLOBE algorithms. We set the size of design matrix as $n = p = 100$, with 10% of missing values. We simulated $k = 10$ true predictors, with values $\beta_{15} = \beta_{25} = \beta_{35} = \beta_{65} = \beta_{75} = \beta_{85} = \beta_{95} = 3\sqrt{2\log p}$ and $\beta_{45} = 1.5\sqrt{2\log p}$, $\beta_{55} = 5\sqrt{2\log p}$.

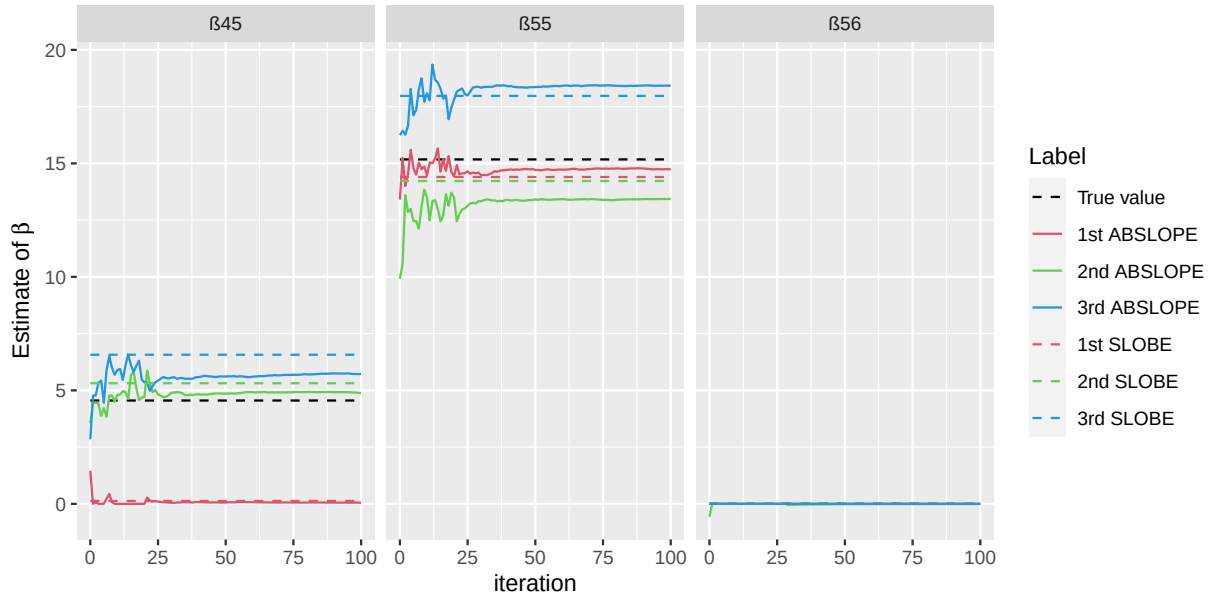


Figure 3: Convergence plots for three coefficients with ABSLOPE and SLOBE (colored solid curves). Black dash lines represent the true value for each $\beta$. Estimates obtained with three different sets of simulated data are represented by three different colors.

The solid lines in Figure 3 represent the path of the SAEM solutions 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. We can also observe that a weak effect $\beta_{45}$ was missed in one of our simulation runs.

The dashed lines in Figure 3 represent the final results of the SLOBE algorithm, which converged after 33, 35 and 68 iterations, respectively. We can see that ABSLOPE and SLOBE yield very similar results. The most significant differences occur for $\hat{\beta}_{45}$ in sim-

ulation 2, where ABSLOPE is slightly more accurate, and for $\hat{\beta}_{55}$ in simulation 3, where SLOBE is slightly more accurate.

In addition, we also represent the convergence curves for $\sigma$ with ABSLOPE in supplementary materials Jiang et al. (2021b) in order to compare the estimate of $\sigma$ by ABSLOPE to the biased MLE estimator without prior knowledge, *i.e.*, $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{RSS}{n}}$. These results show that that the estimates of $\sigma$ with both methods are biased downward, but since ABSLOPE has an additional correction term (12), it leads to a less biased estimator.

## 4.3 Behavior of ABSLOPE with missing values

We then evaluate ABSLOPE in different parametrization settings to see how the signal strength, sparsity and other parameters influence its performance.

**Criterion** We apply ABSLOPE on a synthetic data set and obtain a sequence of estimated indicator vectors $\hat{\gamma}$. A variable $i$ is identified as an important predictor if the average of sampled $\hat{\gamma}_i$ in the last 20 iterations exceeds 0.5. For the final estimation of $\beta$, we keep only the selected terms in the estimated $\beta$ from the last iteration. We compare the selected model 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 = $\frac{FP}{max(FP+TP,1)}$ ;
- MSE of $\beta$ (Relative mean squared error) = $\frac{\|\hat{\beta}-\beta\|^2}{\|\beta\|^2}$;
- MSP: Relative squared prediction error = $\frac{\|X\hat{\beta}-X\beta\|^2}{\|X\beta\|^2}$.

For each set of parameters, we repeat the procedure 200 times: *i)* data generation *ii)* estimation and model selection with ABSLOPE *iii)* evaluation with the criteria presented above and we compute the means over the 200 simulations. The simulations were implemented with parallel computing. We consider $n = p = 100$ and vary:

- sparsity: number of true signal $k = 3, 6, 10, 12, 15$,
- signal strength: weak $1.3\sqrt{2\log p}$ or strong $3\sqrt{2\log p}$;
- percentage of missingness 0.05, 0.1, 0.15, generated randomly, *i.e.,* MCAR;
- correlation between each pair of covariates $\rho = 0, 0.5$

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 in our simulation example FDR is controlled below the expected

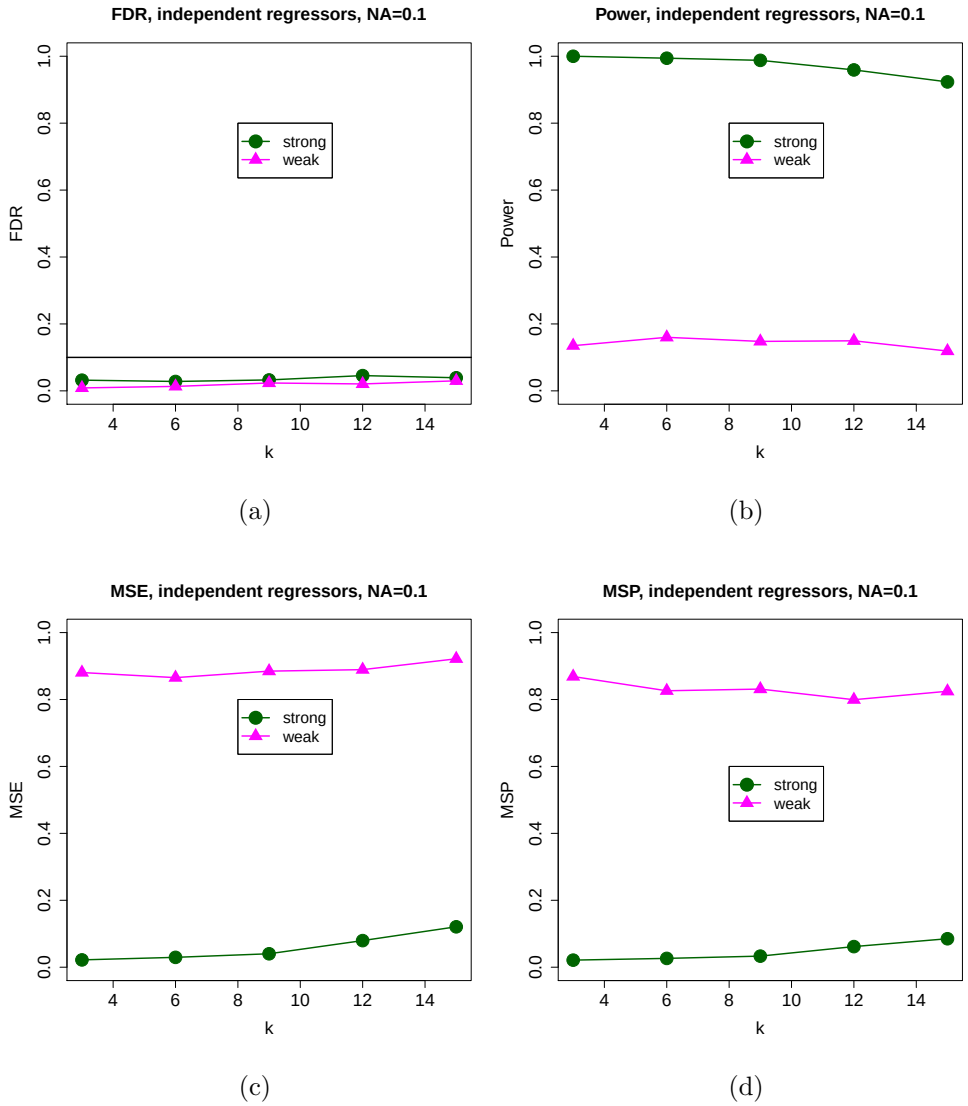


Figure 4: Estimated FDR (a), power (b), relative mean squared estimation error (c) and relative mean squared prediction error (d), as functions of the number of nonzero regression coefficients. Results for $n = p = 100$, percentage of missingness 10% and $\Sigma$ orthogonal (no correlation).

level 0.1. The power slightly decreases and the estimation error slightly increases with increased number of important predictors. When the signal is very weak (signal strength

$=1.3\sqrt{2\log p}$), the power is below 0.2. This is partially related to the confounding between different model parameters, difficult to resolve when the signal is very weak.

**Results 2: with correlation ($\rho = 0.5$, strong signal, varied percentage of missing values).** The results in Figure 5 show that: The power decreases and the FDR and the

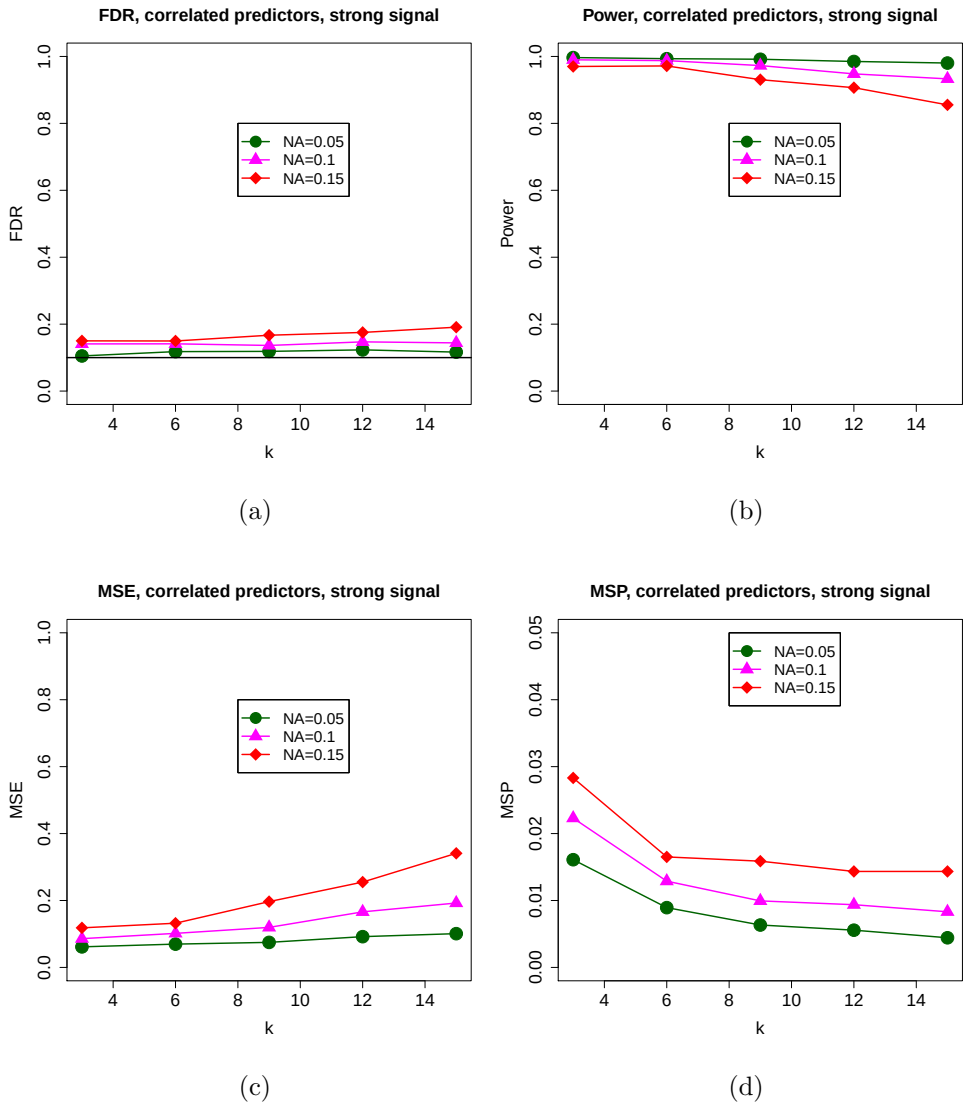


Figure 5: Estimated FDR(a), power(b), relative mean squared error (c) and relative mean squared prediction error (d), as functions of the number of nonzero regression coefficients. Results for $n = p = 100$, with correlation $\rho = 0.5$ and strong signal $3\sqrt{2\log p}$.

estimation error increase when the percentage of missing data increases. However, we

can observe that FDR is only slightly above the nominal level. Interestingly, in case of correlated predictors the prediction error is much smaller than the estimation error and decreases when the number of important predictors increases.

In addition, we present the results varying the correlations and testing the robustness to the normal assumption for covariates in the supplementary materials (Jiang et al., 2021b). The power, FDR and estimation and prediction errors obtained with the proposed AB-SLOPE based on the Gaussian assumptions are robust to moderate deviations from the assumed probabilistic model.

## 4.4 Empirical comparison of SLOBE with other regularization methods for $n = p = 500$.

In this section we present the results of the simulation study comparing different model selection methods in a high-dimensional setup $n = p = 500$.

**Estimation Procedures.** We consider the following estimation procedures

- Efficient SLOBE implementation of Adaptive Bayesian SLOPE based on the BH sequence $\lambda_{\text{BH}}$ (3) with the nominal FDR level $q = 0.1$.

- Adaptive Bayesian LASSO (ABLAS) - a new version of spike and slab Lasso, with the spike prior fixed using the multiple testing principles. ABLAS is obatined by running SLOBE algorithm with the constant Bonferroni sequence $\lambda_1 = \ldots = \lambda_p = \lambda_{\text{BH},1}$ with $q = 0.1$. Similarly as in SLOBE, ABLAS slab prior is estimated based on the data.

- Cross-validated LASSO (LCV), as implemented in *glmnet* R package (Friedman et al., 2010), with $\lambda$ selected by minimizing the cross-validated prediction error (option *s='lambda.min'* in *cv.glmnet*).

- SLOPE two stage procedure. In the first stage SLOPE is used for the model selection and in the second stage the regression coefficients are estimated using the least squares method within the selected model. We use *SLOPE* R package (Larsson et al., 2020) and the heuristic *gaussian* sequence of the tuning parameters recommended in Bogdan et al. (2015) for FDR control when the predictors are independent. We use the nominal FDR level $q = 0.1$.

- SSL: Spike and Slab Lasso as implemented in *SSLASSO* R package (Ročková and George, 2018).

- Adaptive Lasso (ALAS) with weights determined by the cross-validated LASSO, $w_j = \frac{\sigma}{|\beta_j^{LCV}|}$ and with the Bonferoni adjusted $\lambda = \sigma\lambda_{\mathrm{BH},1}$.

SLOPE, SLOBE, ABLAS and ALAS procedures use the tuning parameters adjusted to control the number of false discoveries, which depend on the standard deviation of the error term $\sigma$. In this section we use a consistent approach to deal with the unknown $\sigma$ case, which relies on replacing $\sigma$ with

$$\hat{\sigma} = \sqrt{\frac{||Y - X\beta^{LCV}||^2}{n - ||\beta^{LCV}||_0}}\ , \tag{17}$$

where $||\beta^{LCV}||_0$ is the size of the support of $\beta^{LCV}$. In our simulations this estimator turned out to be very accurate and the respective versions of ABSLOPE and SSL usually had better properties than the versions using build-in routines for the estimation of $\sigma$ (see the supplementary materials, Jiang et al. (2021b)).

When performing these large scale simulations we observed the dependency of SLOBE and ABLAS convergence statistics on the signal strength and sparsity. In case of strong and sparse signals the convergence is usually obtained in less than 10 iterations, while in case of dense and weak signals the algorithm needs more iterations to converge and sometimes has a tendency to oscillate between different modes of the multivariate posterior distribution. If such a situation occurs we stop the SLOBE or ABLAS algorithm after 100 iterations.

**Simulation setting.** In all the simulations in this section $n = p = 500$. The rows of the design matrix are generated as independent random vectors from a multivariate normal distribution $N\left(0, \frac{1}{n}\Sigma\right)$. We consider two scenarios, one with *independent* regressors, where the correlation matrix $\Sigma = I$, and the one with *correlated* regressors, where $\Sigma$ is the compound symmetry matrix with $\Sigma_{i,j} = 0.5$ for $i \neq j$. We generate the response variable using the multiple regression model (1) with $\epsilon \sim N(0, I)$ (i.e. $\sigma = 1$). The number $k$ of nonzero regression coefficients in the vector $\beta$ takes values from the set $k \in \{5, 10, 20, 40, 60\}$. We consider weak signals with $\beta_1 = \ldots = \beta_k = 1.3\sqrt{2\log p}$, and moderately strong signals with $\beta_1 = \ldots = \beta_k = 2\sqrt{2\log p}$.

**Summary Statistics.** Similarly as in Section 4.3 we report FDR, Power, MSE and MSP, which are obtained from averaging the results of 200 simulation runs.

### 4.4.1 Complete data

In this section we report the results of the analysis of the complete data, i.e., of the data without missing values. In this situation the estimator of $\sigma$ (17) turned out to be very precise and we have not observed significant deviations in performance of the methods using this estimator and the true $\sigma$ value. Therefore we report only the results for the unknown $\sigma$ case, which are illustrated in Figures 6 and 7.

Summarizing results reported in Figures 6 and 7 we can observe that SLOPE based on the heuristic sequence of tuning parameters controls FDR when the predictors are independent and $k < 40$ and then its FDR slightly increases above the nominal level. When the signals are weak this FDR control comes at the price of the loss of power for larger values of $k$. When predictors are strongly correlated SLOPE does not longer control FDR, which varies between 0.8 and 0.6 for sparse and denser models.

Instead, SLOBE based on the BH sequence of tuning parameters allows for FDR control at the nominal level when the signal is strong or dense enough, so that its composition can be learned from the data. When the signals are large, SLOBE controls FDR both for the independent and correlated setups, has a high power and superior estimation and prediction properties. When the signals are weak and sparse, so there is very little information to learn the signal composition, FDR of SLOBE exceeds the nominal level. This effect is stronger when predictors are strongly correlated. When the number of nonzero elements in $\beta$ increases such that the estimation of the signal sparsity and its magnitude becomes more feasible then FDR of SLOBE seems to converge to the nominal level even for weak signals and correlated predictors. These observations suggest possible theoretical developments concerning the asymptotic FDR control by SLOBE.

Comparing Adaptive Bayes LASSO (ABLAS) to SLOBE we can observe a somewhat unexpected phenomenon. For independent predictors FDR of ABLAS systematically increases with $k$ and for $k > 60$ it actually exceeds FDR of SLOBE. For correlated predictors ABLAS has FDR which is systematically larger than FDR of SLOBE. This is a bit surprising, since the SLOBE decaying sequence of tuning parameters is "smaller" than the

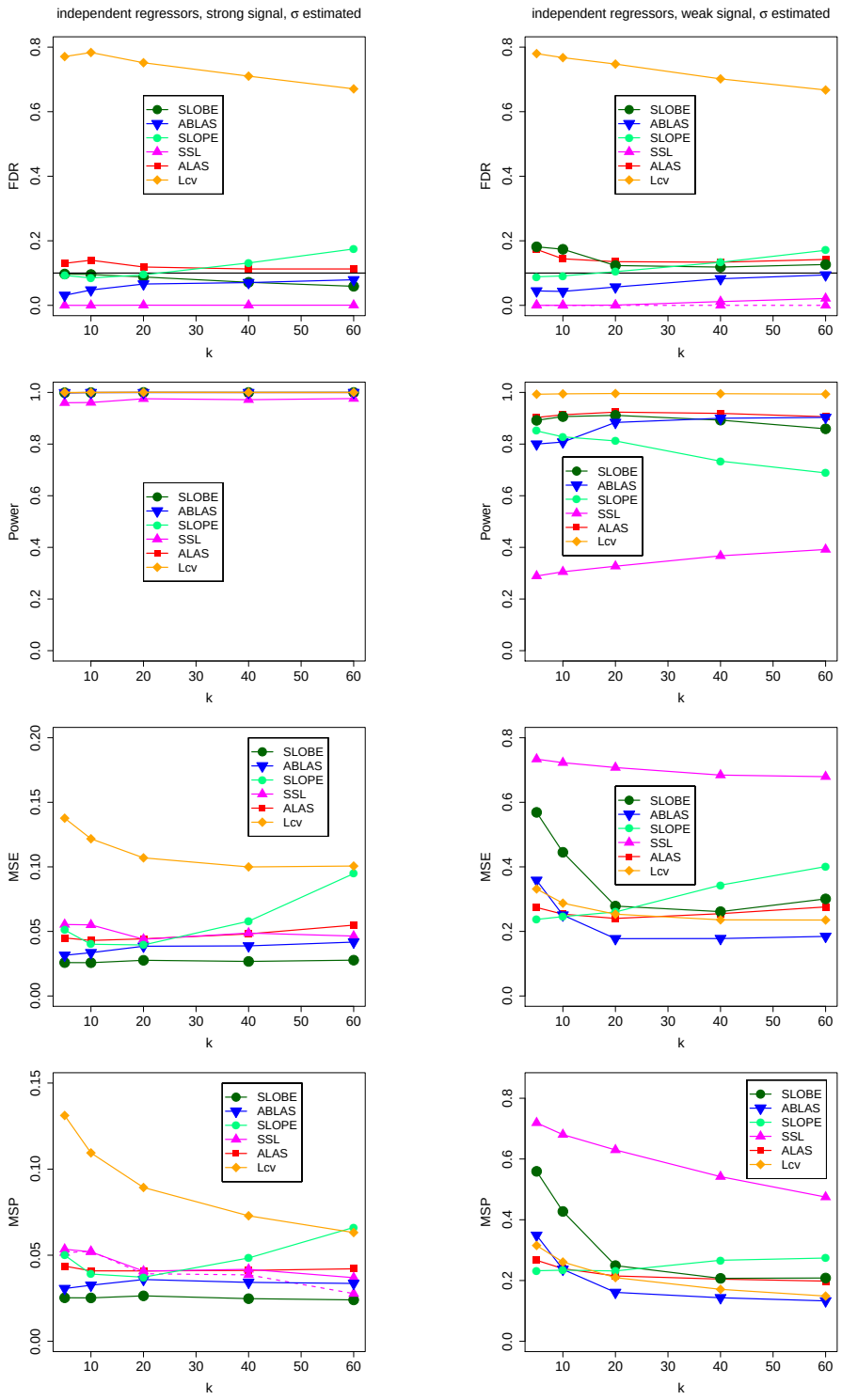


Figure 6: Different performance measures as the function of the number of true nonzero regression coefficients. Complete data with independent predictors and strong (left panel) and weak (right panel) signals.

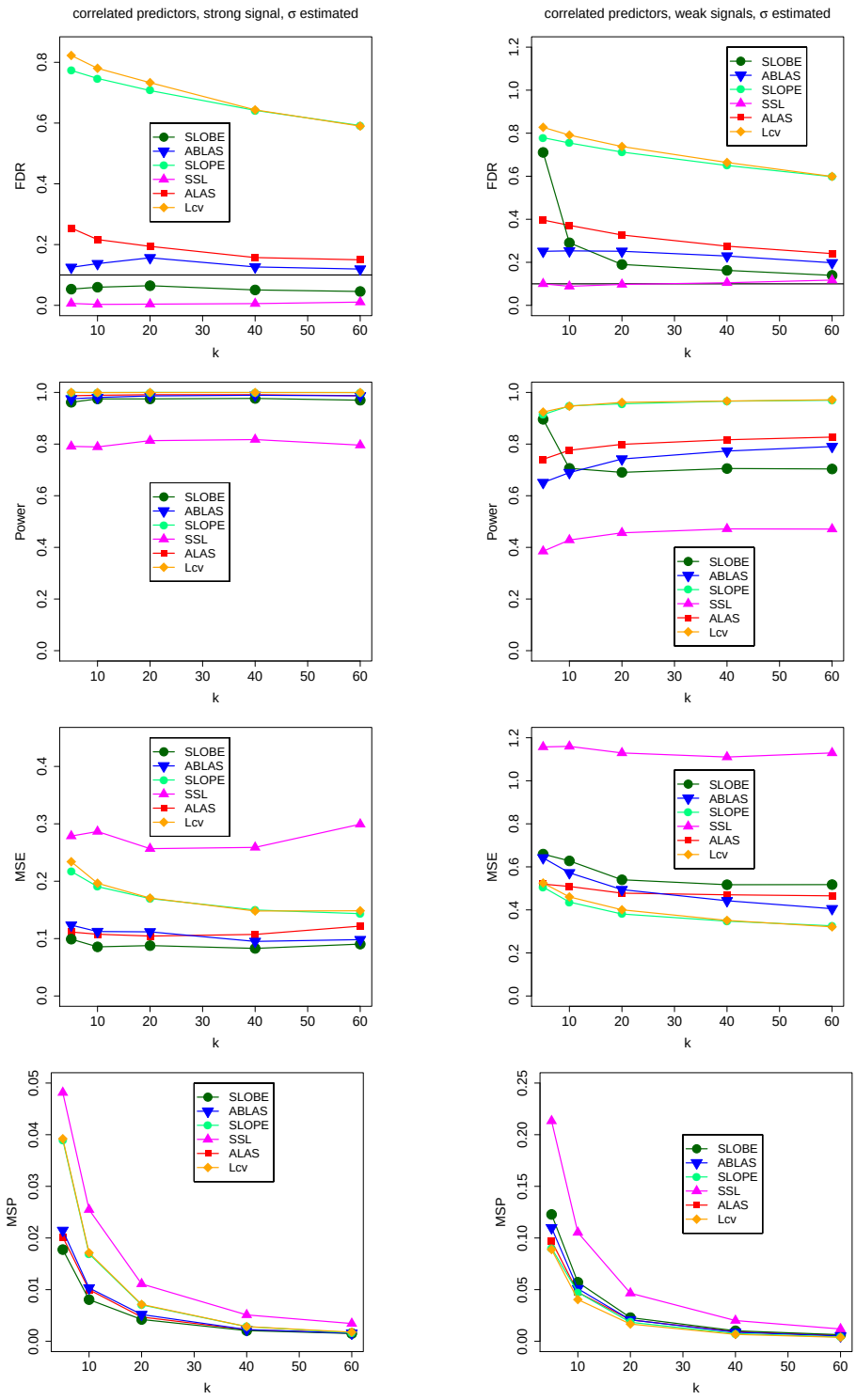


Figure 7: Different performance measures as the function of the number of true nonzero regression coefficients. Complete data with correlated predictors and strong (left panel) and weak (right panel) signals.

ABLAS constant sequence. We believe that this is due to a fact that some of the estimates of the zero elements of the $\beta$ vector are larger than it is suggested by the more concentrated ABLAS spike prior and are more easily classified as signals. Interestingly, the related overestimation of the sparsity parameter $\theta$ leads to the smaller bias of nonzero regression coefficients and to the superior estimation and prediction properties of ABLAS when the signals are weak.

Comparing different versions of adaptive LASSO we can observe that SSL systematically has the smallest FDR and is competitive for large signals. Instead, it looses a lot of power and estimation and prediction accuracy when signals are weak or moderately large. We believe that this is due to the default selection of a small $\lambda_1$ value, which suggests that the true signals are large. Thus, moderately large or weak signals are attracted by the spike component of the prior. Comparing ABLAS to ALAS we can see that ABLAS systematically has a smaller FDR. When the signals are strong or weak and dense ABLAS is also better with respect to prediction and estimation properties.

Least squares estimators within a SLOPE model are competitive when predictors are independent and the signal is sparse ($k < 20$) but loose accuracy when $k$ increases. Interestingly, for correlated predictors these estimators perform very similarly to the cross-validated LASSO estimators, being worse than other methods for the strong signals but having superior properties when the signal is weak. This however comes at the price of large FDR ($> 0.6$).

### 4.4.2 With 10 % of Missing Data

In this section we report the results of the analysis of the data with 10% of observations missing completely at random. We impute the missing data using the Principal Components Analysis model from the *missMDA* R package (Josse and Husson, 2016). This imputed data set is used for the estimation by Lasso CV, SLOPE, SSL and ALAS procedures and is the starting point of SLOBE and ABLAS algorithms.

We observed that in case of missing data and independent predictors all methods are sensitive to the inaccuracy of $\sigma$ estimation. Therefore in Figures 8 and 9 we compare the results for the cases when $\sigma$ is known and estimated. In case of strongly correlated data the difference between the "known" and "estimated" $\sigma$ cases was hardly visible, so we report

the results only for the "estimated" case in Figure 10.

Figures 8-10 illustrate that the presence of missing values had a relatively small influence on the performance of the compared methods. In case of SLOBE we can observe that FDR is controlled roughly at the same level as when the data are complete, i.e., it is below the nominal level when the signal is strong and converges to the nominal level when the number of weak signals increases. Here the deterioration of properties is mainly represented by some slight loss of power and the respective slight loss of estimation and prediction properties as compared to the complete data case. In case of ABLAS and ALAS we see a different pattern. Both these procedures have larger FDR than in the complete data case, which in most cases substantially exceeds FDR of SLOBE and the nominal level. Instead, the loss of power is smaller than in case of SLOBE and both ABLAS and ALAS do not loose much in terms of estimation and prediction properties when compared to the complete data case. Here we can observe that ABLAS has a systematically smaller FDR than ALAS and better estimation and prediction properties when predictors are independent, while ALAS has better estimation and prediction properties for correlated regressors.

Interestingly, all methods seem to be most sensitive to the inacuracy of $\sigma$ estimation when the predictors are independent and the signal is strong. This is specifically true about SSL which looses a lot of power and estimation and prediction accuracy when $\sigma$ is estimated.

# 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[2] 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 life or functional integrity of a person. 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

[2] http://www.traumabase.eu/

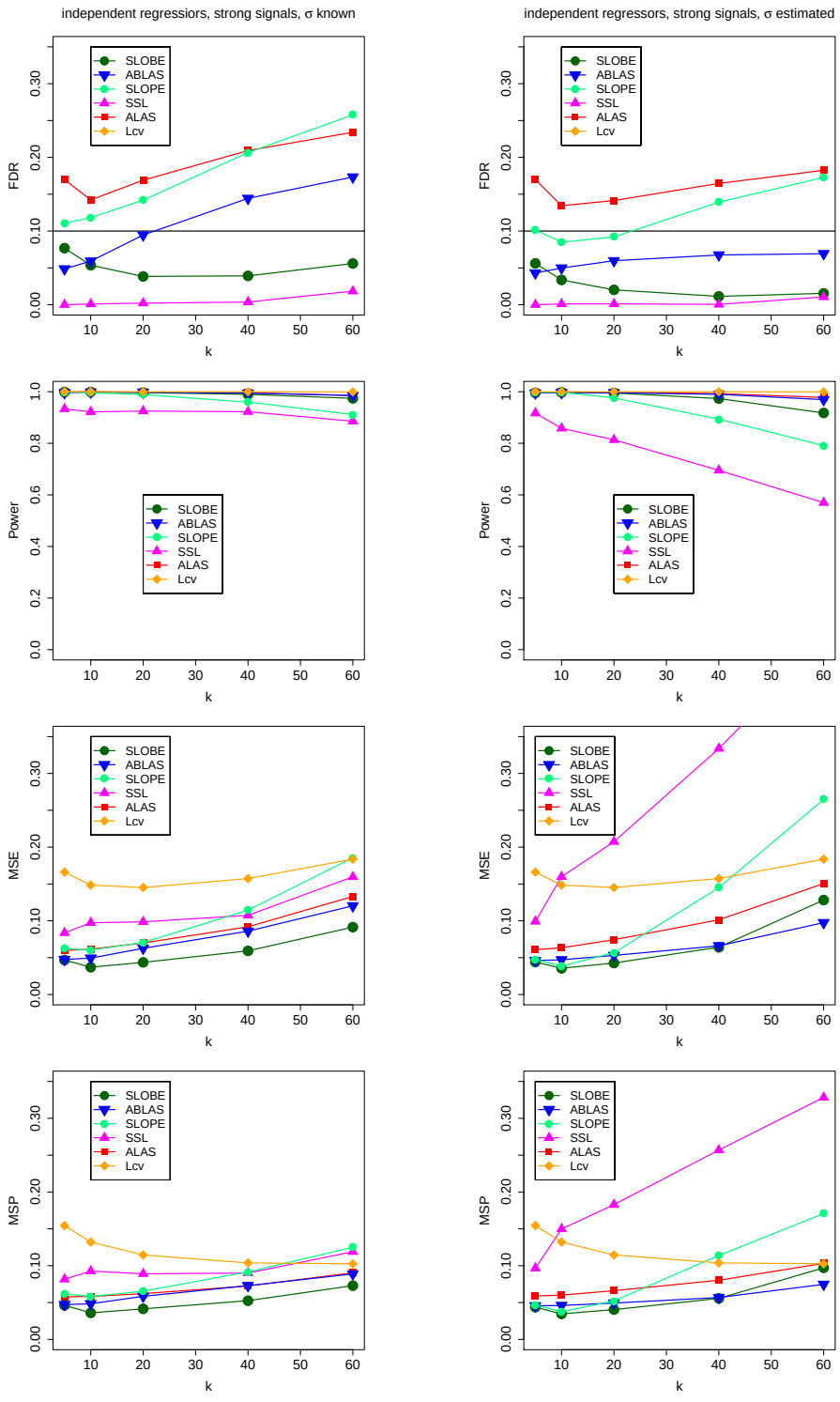


Figure 8: Different performance measures as the function of the number of true nonzero regression coefficients for 10% of data missing completely at random, independent regressors and strong signals. Left and right panel provide the results for $\sigma$ known and estimated, correspondingly.

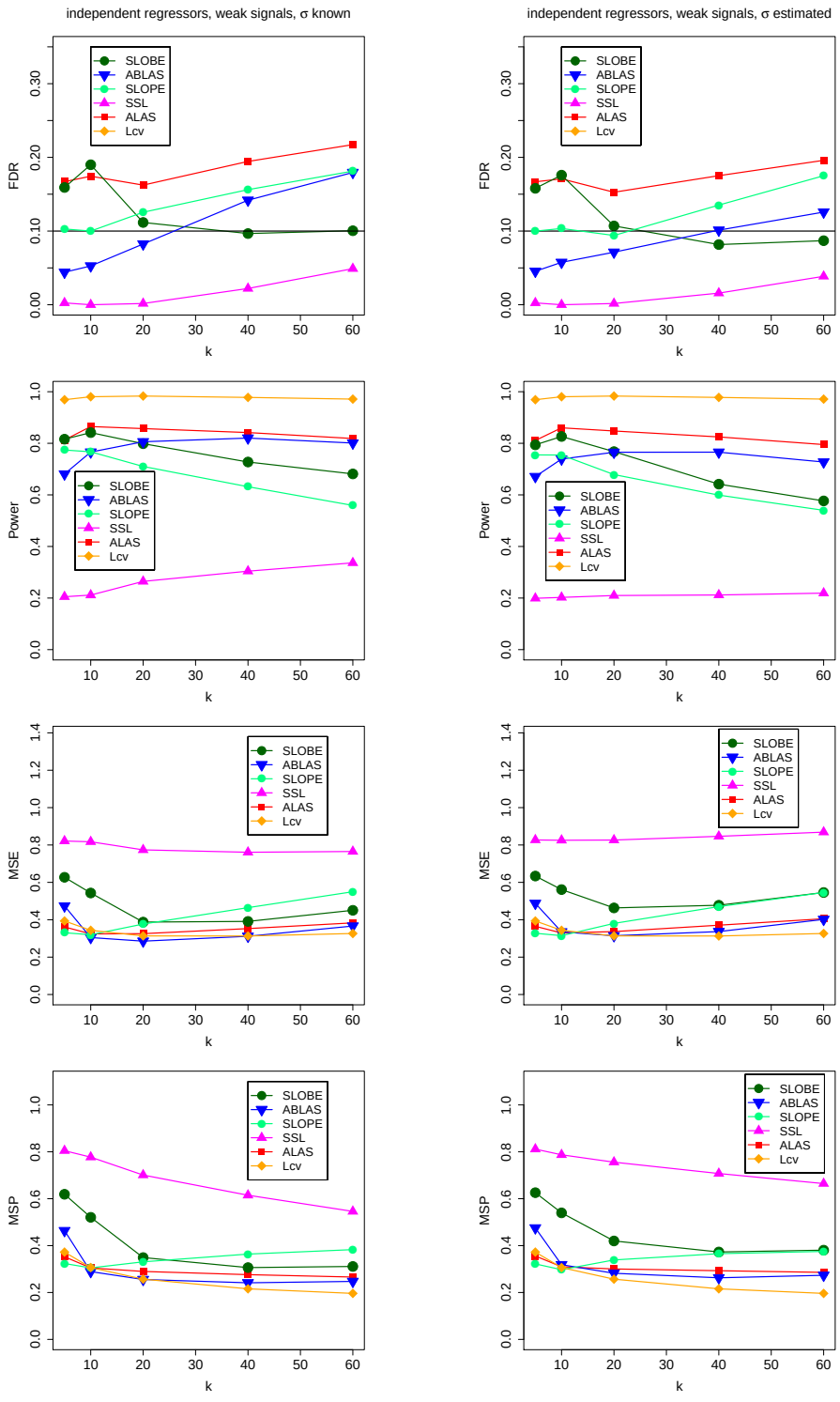


Figure 9: Different performance measures as the function of the number of true nonzero regression coefficients for 10% of data missing completely at random, independent regressors and weak signals. Left and right panel provide the results for $\sigma$ known and estimated, correspondingly.

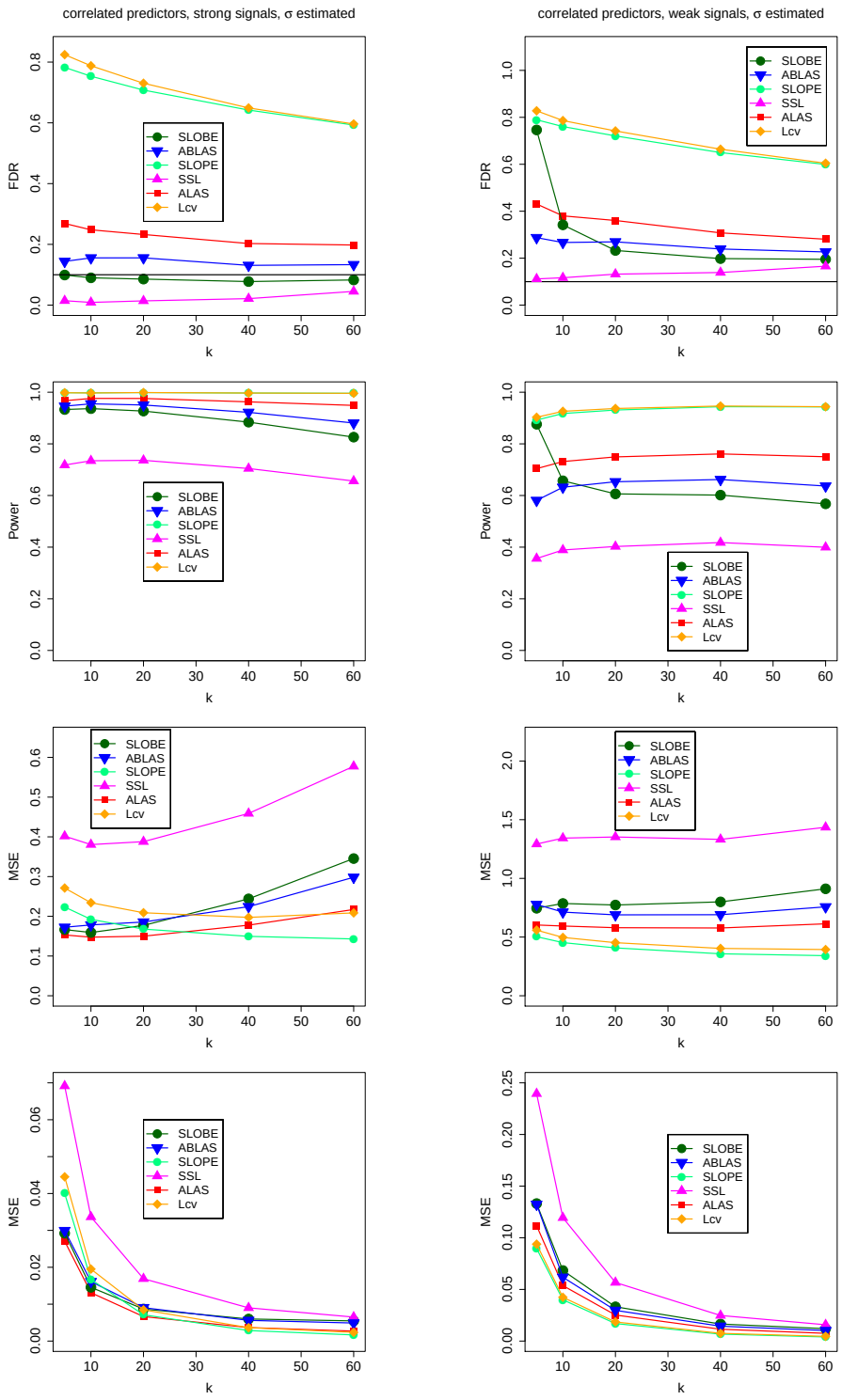


Figure 10: Different performance measures as the function of the number of true nonzero regression coefficients for 10% of data missing completely at random and correlated regressors, with $\sigma$ estimated. Left and right panel provide the results for strong and weak signals, correspondingly.

outcomes. Delays and/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 patients' records measured in the prehospital stage or on arrival to the hospital, we aim to establish prediction models in order 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 to clinicians and professionals. Due to the highly stressful multi-player environment, 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 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 accident scenes to the hospital. These centers have joined TraumaBase progressively between January 2011 and June 2015. The database integrates algorithms for consistency and coherence and data monitoring is performed by a central administrator. Sociodemographic, clinical, biological and therapeutic data (from the prehospital phase to the discharge) 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 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 plagued with missing values, which makes modeling challenging.

In our analysis, we have focused on one specific challenge: developing a statistical model with missing covariates in order to predict the level of platelet upon arrival at the hospital. This model can aid creating an innovative response to the public health challenge of major trauma. The platelet is a cellular agent responsible for clot formation. It is essential to control its levels to prevent blood loss as quickly as possible in order 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, if available, its levels would determine how the patients are treated. Accurate prediction of this metric is thereby crucial for making important treatment decisions in real time.

We focus on patients after an accident who were sent directly to the hospital (not sent to Emergency Care Unit). After this pre-selection, 6384 patients remained in the data set. Based on clinical experience, in order to predict the level of platelet on arrival at the 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., 2021b). These variables were included here because they were all available to the pre-hospital team, and therefore could be used in real situations.

Figure 11 shows the percentage of missingness per variable, varying from 0 to 60%. If we were to perform the complete case analysis (*i.e.*, ignoring all the observations with missing values) only less than one third of the observations (1648 patients) would still remain in the dataset. This loss of data demonstrates the importance of appropriately handling the missing values.

## 5.2 Model selection results

As is customary in supervised learning, we divide the dataset into training and test sets. To obtain high quality test sets we at first used the whole available information to create the *imputed* data set with the Multivariate Imputation by Chained Equations (MICE, van Buuren and Groothuis-Oudshoorn (2011)). Then we randomly selected 70% of observations with missing values for the training set, while the test set contains the remaining 30% of observations from the *imputed* data set. Since the current implementation of ABSLOPE/SLOBE imputations can handle only quantitative explanatory variables we replaced the missing values for the binary explanatory variable RBC with the values from the *imputed* data set. We apply SLOBE and ABLAS compare them with the following methods:

- MICE imputation followed by SLOPE;
- MICE imputation followed by cross-validated Lasso;
- MICE imputation followed by ALAS;
- MICE imputation followed by SSL;

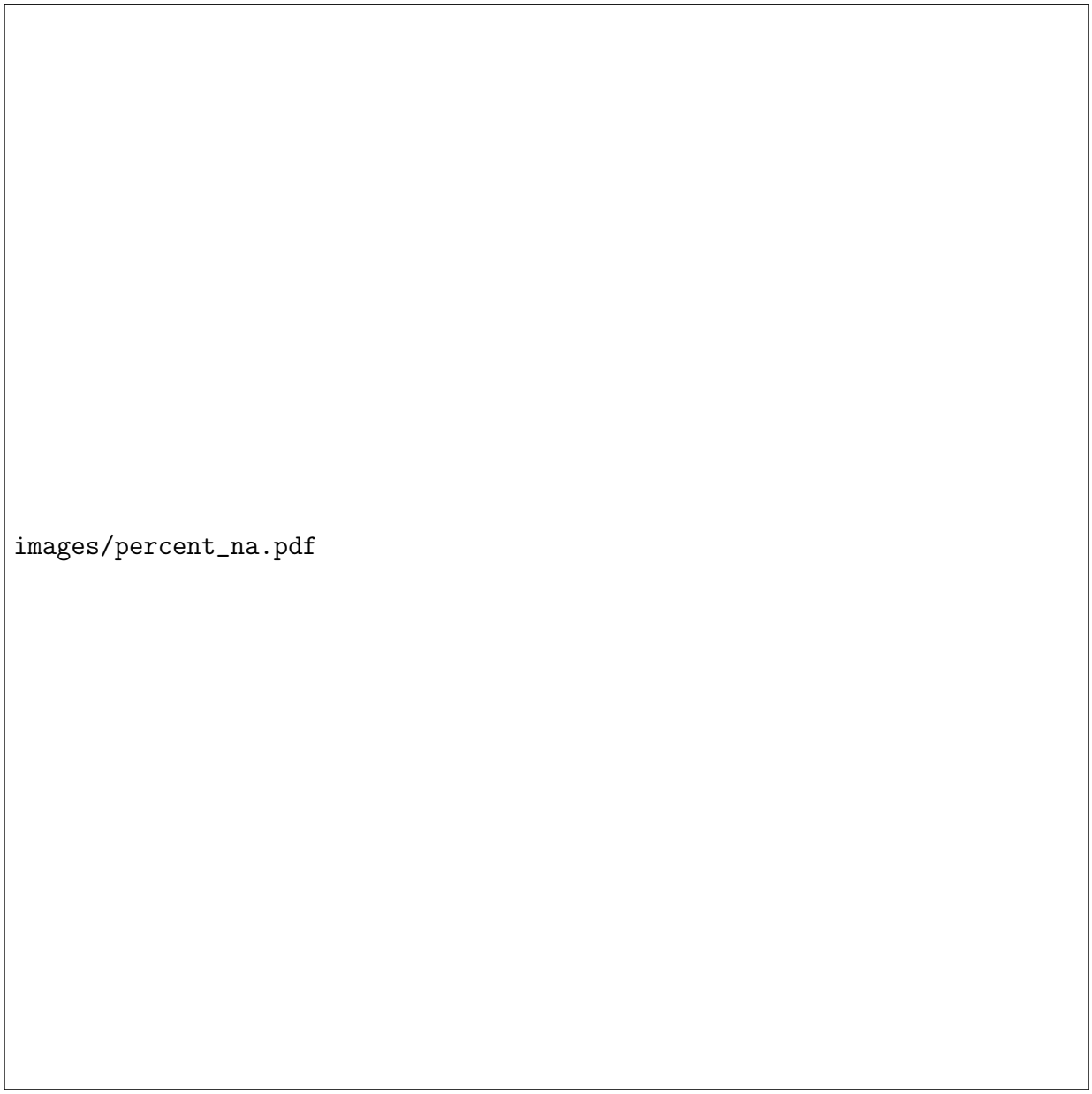


Figure 11: Percentage of missing values in each pre-selected variable from TraumaBase.

Table 1: Number of times that each variable is selected over 10 replications. Bold numbers indicate which variables are included in the models selected by SLOBE and ABLAS.

| Variable | SLOBE | ABLAS | SLOPE | LASSO | ALAS | SSL |
|---|---|---|---|---|---|---|
| Age | **10** | **10** | 10 | 10 | 10 | 10 |
| SI | **10** | **10** | 5 | 10 | 10 | 10 |
| MBP | 0 | 0 | 10 | 5 | 1 | 10 |
| Delta.hemo | **10** | **10** | 10 | 10 | 10 | 10 |
| Time.amb | 1 | 0 | 4 | 10 | 3 | 8 |
| Lactate | **10** | **10** | 10 | 10 | 10 | 10 |
| Temp | 1 | 0 | 10 | 8 | 3 | 9 |
| HR | **10** | **10** | 10 | 10 | 10 | 10 |
| VE | **10** | **10** | 10 | 10 | 10 | 10 |
| RBC | **10** | **10** | 10 | 10 | 10 | 10 |
| SI.amb | 0 | 0 | 4 | 8 | 2 | 5 |
| MBP.amb | 0 | 0 | 1 | 6 | 1 | 1 |
| HR.max | 4 | 1 | 10 | 10 | 4 | 10 |
| SBP.min | 3 | 2 | 10 | 9 | 7 | 10 |
| DBP.min | 2 | 0 | 9 | 6 | 1 | 4 |

• MICE imputation followed by a random forest (RF) (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 our design matrix being centered and without an intercept, we also center the vector of responses and apply the procedure on $\tilde{y} = y - \bar{y}$, where $\bar{y}$ is the mean of $y$. We repeat the procedure of data splitting (into training and test sets) 10 times and Table 1 shows that, over 10 replications, how many times each variable is selected.

Here we can observe that SLOBE and ABLAS consistently select 7 explanatory vari-

ables. The signs of the corresponding regression coefficients are negative for age, shock index, vascular filling, blood transfusion and lactate are negative, which is in agreement with the expectations of the TraumaBase medical team. However, the estimated positive effects of delta Hemocue and the heart rate on the platelet were not entirely in agreement with their opinion.

## 5.3 Prediction performance

We compare the prediction properties of different methods by calculating the relative mean square prediction error: err = $\frac{\|\hat{y}-y\|^2}{\|y\|^2}$, where the explanatory variables in the test set were imputed with *mice* using the whole available information.

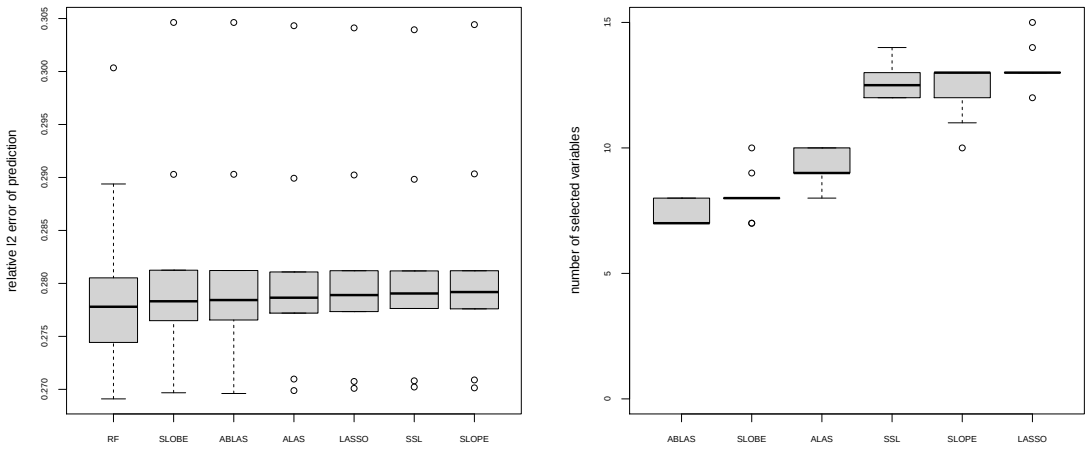


Figure 12: Empirical distribution of prediction errors of different methods over 10 replications for the TraumaBase data and of the number of variables selected by different methods.

As we can observe, the predictive properties of different methods do not differ much. As expected, Random Forest has the highest prediction accuracy but the gain is rather small as compared to the complexity of the RF predictive model. RF predictions require all 15 measurements used to fit the model, which might be costly to obtain in the stressful post-trauma ambulance situation. Moreover, RF predictions require a specialized software with an implementation of the Random Forest fitted model.

Among other methods SLOBE and ABLAS yield the smallest median of the prediction

error and the smallest median number of selected variables. The main advantage of these two new approaches is that they provide relatively small models which remain stable over different random partitions of the data into the training and tests.

Finally, we also performed SLOBE on the whole standardized data set and obtained the model

$100Platelets = -8.71\text{Age} - 10.52\text{SI} + 9.16\text{Delta.hemo} - 14.7\text{Lactate} + 14.2\text{HR} - 6.54\text{VE} - 11\text{RBC} + 0.076\text{HR.max} + 0.076\text{SBP.min} + 0.006\text{DBP.min}.$

The standardized coefficients by the three last variables are very small, which confirms the cross-validation results pointing at the first seven variables as the most important predictors.

# 6 Discussion

ABSLOPE penalizes noise coefficients more stringently to control for FDR while leaving larger effects relatively unbiased through an adaptive weighting matrix. In addition, casting our method within a Bayesian framework allows one to assign a probabilistic structure over models and estimate the pattern of sparsity. We develop an SAEM algorithm which handles missing values and which treats model indicators as missing data. According to the simulation study, ABSLOPE is competitive with other methods in terms of power, FDR and prediction error. For future research, we will consider the problem of high-dimensional model selection with missing values for categorical or mixed data and other missing mechanisms such as MNAR.

# Supplementary Materials

R programs *ABSLOPE* and *SLOBE* containing the implementation of the proposed methodology, codes to reproduce the experiments and some supplementary simulation results are provided in Jiang et al. (2021a).

# Acknowledgment

Wei Jiang was supported by grants from Région Île-de-France: https://www.dim-mathinnov.fr. The research of Małgorzata Bogdan and Szymon Majewski was supported by the grant of the Polish National Center of Science Nr 2016/23/B/ST1/00454. Blazej Miasojedow was

supported by the Polish National Science Center grant: NCN UMO-2018/31/B/ST1/00253. Veronika Rockova gratefully acknowledges support from the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business. The authors are thankful for fruitful discussion with Edward I. George, 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 Tysbakov, 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.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

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.

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

Datta, A., Zou, H., et al. (2017). Cocolasso for high-dimensional error-in-variables regression. *The Annals of statistics*, 45(6):2400–2426.

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.

Descloux, P., Boyer, C., Josse, J., Sportisse, A., and Sardy, S. (2020). Robust lasso-zero for sparse corruption and model selection with missing covariates.

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.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

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., Bogdan, M., Josse, J., Majewski, S., Miasojedow, B., Ročková, V., and Group, T. (2021a). Codes and implementations for ABSLOPE. https://github.com/mbogdan-work/ABSLOPE.

Jiang, W., Bogdan, M., Josse, J., Miasojedow, B., Ročková, V., and Group, T. (2021b). Additional supplementary materials for "Adaptive Bayesian SLOPE: Model selection with incomplete data". https://github.com/mbogdan-work/ABSLOPE.

Jiang, W., Josse, J., and Lavielle, M. (2019). Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, page 106907.

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.

Larsson, J., Wallin, J., Bogdan, M., van den Berg, E., Sabatti, C., Candès, E., Patterson, E., and Su, W. (2020). *SLOPE: Sorted L1 Penalized Estimation*. R package version 0.3.2.

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. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

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 (2018). *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. (2014). EMVS: The Bayesian approach to Bayesian variable selection. *Journal of the American Statistical Association*, (109):828–836.

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.

Rubin, D. B. (2009). *Multiple Imputation for Nonresponse in Surveys*, volume 307. John Wiley & Sons.

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.

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.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

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.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zhu, Z., Wang, T., and Samworth, R. J. (2019). High-dimensional principal component analysis with heterogeneous missingness.

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

# A Appendix

## A.1 Deviation of prior (5) started from SLOPE prior

We assume a random variable $z = (z_1, z_2, \cdots, z_p)$ has a SLOPE prior:

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

and then define $\beta = W^{-1}z = (\frac{z_1}{w_1}, \cdots, \frac{z_p}{w_p})$, or equally, $z_j = \beta_j w_j$ where the diagonal elements in the weight matrix are $w_j = c\gamma_j + (1-\gamma_j) = \begin{cases} c, & \gamma_j = 1 \\ 1, & \gamma_j = 0 \end{cases}$, $j = 1, 2, \cdots, p$. Then according to the transformation of variables, we have the prior distribution for $\beta$:

$$\begin{aligned}
\mathtt{p}(\beta \mid W, \sigma^2; \lambda) &\propto \left|\det\left(\frac{dz}{d\beta}\right)\right| \mathtt{p}_z(W\beta \mid W, \sigma^2; \lambda) \\
&= \prod_{j=1}^{p} w_j \prod_{j=1}^{p} \exp\left\{-\frac{1}{\sigma}\lambda_{r(W\beta,j)}|w_j\beta_j|\right\} \\
&= c^{\Sigma_{j=1}^{p}\mathbb{1}(\gamma_j=1)} \prod_{j=1}^{p} \exp\left\{-w_j|\beta_j|\frac{1}{\sigma}\lambda_{r(W\beta,j)}\right\},
\end{aligned}$$

which corresponds to our proposed prior (5).

## 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:

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

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 $i$,

$$\begin{aligned} \mathtt{p}(r_i \mid y_i, X_i; \phi) &= \mathtt{p}(r_i \mid y_i, X_i^{(\text{obs})}; \phi) \\ &= \mathtt{p}(r_i \mid y_i, X_{i,\text{obs}}; \phi) \end{aligned} \tag{18}$$

MAR assumption implies that, the observed likelihood can be maximize and the distribution of $r$ can be ignored Little and Rubin (2019). Assume that $\theta$ is the parameter to estimate. Indeed:

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

then according to the assumption MAR (18), we have:

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

Therefore, to estimate $\theta$, we aim at maximizing $\mathcal{L}(\theta; y, X_{\text{obs}}) = \mathtt{p}(y, X_{\text{obs}}; \theta)$. 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 $i$:

$$\hat{X}_i^0 = (X_i^0 - m^0) \oslash (\sqrt{n} s^0),$$

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

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

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

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:

$$\hat{X}^t_i = (\tilde{X}^t - m^t) \oslash (\sqrt{n}s^t).$$

The final estimates of regression coefficients are then rescaled to match the original values of the response and explanatory variables.

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

To perform the simulation step (7), we use a Gibbs sampler. To simplify notation, we hide the superscript, and note that all 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:

$$\begin{aligned}\gamma_j &\sim \mathtt{p}(\gamma_j \mid \gamma_{-j}, c, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma) \\ &= \mathtt{p}(\gamma_j \mid \gamma_{-j}, c, \beta, \sigma, \theta)\ ,\end{aligned}$$

where $\gamma_{-j} = (\gamma_1, \cdots, \gamma_{j-1}, \gamma_{j+1}, \cdots, \gamma_p)$; *i.e.*, sampling from a Binomial distribution with probability:

$$\begin{aligned}\mathbb{P}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta) &= \frac{\mathbb{P}(\gamma_j = 1 \mid \theta)\mathtt{p}(\beta \mid \gamma_j = 1, \gamma_{-j}, c, \sigma)}{\sum_{\gamma_j \in \{0,1\}} \mathbb{P}(\gamma_j \mid \theta)\mathtt{p}(\beta \mid \gamma_j, \gamma_{-j}, c, \sigma)} \\ &= \left[1 + \frac{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W^0\beta,j)}\right) \times (c)^{\sum_{-j}\mathbb{1}(\gamma_{-j}=1)} \prod_{-j}\exp\left(-w^0_{-j}|\beta_{-j}|\frac{1}{\sigma}\lambda_{r(W^0\beta,-j)}\right)}{\theta c\exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W^1\beta,j)}\right) \times (c)^{\sum_{-j}\mathbb{1}(\gamma_{-j}=1)} \prod_{-j}\exp\left(-w^1_{-j}|\beta_{-j}|\frac{1}{\sigma^t}\lambda_{r(W^1\beta,-j)}\right)}\right]^{-1} \\ &= \left[1 + \frac{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W^0\beta,j)}\right)}{\theta c\exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W^1\beta,j)}\right)} \times \frac{\prod_{-j}\exp\left(-w^0_{-j}|\beta_{-j}|\frac{1}{\sigma}\lambda_{r(W^0\beta,-j)}\right)}{\prod_{-j}\exp\left(-w^1_{-j}|\beta_{-j}|\frac{1}{\sigma}\lambda_{r(W^1\beta,-j)}\right)}\right]^{-1},\end{aligned} \tag{19}$$

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 (19) 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, (19) is drawn from:

$$\begin{aligned}\mathbb{P}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) &= \left[1 + \frac{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}\right]^{-1} \\ &= \frac{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right) + \theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}\,,\end{aligned} \tag{20}$$

which can be interpreted as the posterior probability of binary signal indicator for $j^{\text{th}}$ 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 (5).

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

$$\begin{aligned}\theta &\sim \mathtt{p}(\theta \mid \gamma, c, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \mu, \Sigma, W) \\ &= \mathtt{p}(\theta \mid \gamma, \beta, \sigma, W) \propto \mathtt{p}(\theta)\,\mathtt{p}(\gamma \mid \theta)\,,\end{aligned}$$

where $\mathtt{p}(\gamma \mid \theta)$ 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:

$$Beta\left(a + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1), b + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 0)\right)\,, \tag{21}$$

from which we can generate the latent variable $\theta$. The target distribution (21) 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.

$$\begin{aligned} c &\sim \mathtt{p}(c \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma, W) \\ &= \mathtt{p}(c \mid \gamma, \beta, \sigma, W) \propto \mathtt{p}(c)\, \mathtt{p}(\beta \mid c, \gamma, \sigma, W) \\ &= p(c)\, c^{\sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)} \exp\left( -\frac{c}{\sigma} \sum_{j=1}^{p} |\beta_j| \lambda_{r(W\beta, j)} \mathbb{1}(\gamma_j = 1) \right) , \end{aligned}$$

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

$$Gamma\left( 1 + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1), \quad \frac{1}{\sigma} \sum_{j=1}^{p} |\beta_j| \lambda_{r(W\beta, j)} \mathbb{1}(\gamma_j = 1) \right) . \tag{22}$$

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 $p(x_{\text{obs}}, x_{\text{mis}}; \Sigma, \mu) \sim \mathcal{N}(\mu, \Sigma)$ 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 $i$:

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

where $x^{-i}_{\text{mis}} = \left( x^j_{\text{mis}},\, j \neq i \right)$. 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 $\Sigma^{-1}$:

$$\begin{aligned}
&\mathtt{p}(x^i_{\text{mis}} \mid x_{\text{obs}}, y, \sigma, \beta, \Sigma, \mu, x^{-i}_{\text{mis}}) \propto \exp\left[-\frac{1}{2\sigma^2}(y - x\beta)^2 - \frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right] \\
&\propto \exp\Bigg[-\frac{1}{2\sigma^2}\left(y - x_{\text{obs}}\beta_{\text{obs}} - x^i_{\text{mis}}\beta_i - \sum_{j\in\mathcal{M}, j\neq i} x^j_{\text{mis}}\beta_j\right)^2 \\
&\qquad - \frac{1}{2}\left(s_{ii}(x^i_{\text{mis}} - \mu_i)^2 + 2x^i_{\text{mis}} \sum_{j\in\mathcal{M}, j\neq i}(x^j_{\text{mis}} - \mu_j)s_{ij} + 2x^i_{\text{mis}} \sum_{k\in\mathcal{O}}(x^k_{\text{obs}} - \mu_k)s_{ik}\right)\Bigg] .
\end{aligned}$$

After rearranging terms, with notations:

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

we get:

$$\begin{aligned}
&\mathtt{p}(x^i_{\text{mis}} \mid x_{\text{obs}}, y, \sigma, \beta, \Sigma, \mu, x^{-i}_{\text{mis}}) \\
&\propto \exp\left\{-\frac{1}{2}\left[(x^i_{\text{mis}})^2\left(s_{ii} + \frac{\beta_i^2}{\sigma^2}\right) - 2x^i_{\text{mis}}\left(\frac{r\beta_i}{\sigma^2} + m_i - u_i\right) + 2x^i_{\text{mis}} \sum_{j\in\mathcal{M}, j\neq i}\left(\frac{\beta_i\beta_j}{\sigma^2} + s_{ij}\right)x^j_{\text{mis}}\right]\right\} \qquad (23) \\
&\propto \exp\left\{-\frac{1}{2}\left[x^i_{\text{mis}}\tau_i - \frac{r\beta_i/\sigma^2 + m_i - u_i}{\tau_i} + \sum_{j\in\mathcal{M}, j\neq i} \frac{\beta_i\beta_j/\sigma^2 + s_{ij}}{\tau_i\tau_j} x^j_{\text{mis}}\tau_j\right]^2\right\} .
\end{aligned}$$

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 $z_i = \tau_i x^i_{\text{mis}}$ we have:

$$\mathtt{p}(z_i \mid x_{\text{obs}}, y, \sigma, \beta, \Sigma, \mu, x^{-i}_{\text{mis}}) \propto \exp\left[-\frac{1}{2}\left(z_i - \tilde{\mu}_i + \sum_{j\in\mathcal{M}, j\neq i} B_{ij}\,(z_j - \tilde{\mu}_j)\right)^2\right] , \qquad (24)$$

and

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

Combine equations (23) and (24), we obtain the solution:

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

and

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

□

**Algorithm 1** Solving ABSLOPE with SAEM.

**Input:** Initialization $\beta^0$, $\sigma^0$, $c^0, \theta^0$, $X^0_{\text{mis}}$, $\mu^0$, $\Sigma^0$; Choose maximum iteration number Maxit > 20;

**for** $t = 1, 2, \cdots,$ Maxit **do**

*(Simulation step)*

1. Generate $\gamma^t$ from (20);
2. Generate $\theta^t$ from Beta distribution (21);
3. Generate $c^t$ from truncated Gamma distribution (22);
4. Generate $X^t_{\text{mis}}$ from Gaussian distribution (11);

*(Stochastic Approximation step)*

1. Calculate $(\beta^t_{\text{MLE}}, \sigma^t_{\text{MLE}}, \mu^t_{\text{MLE}}, \Sigma^t_{\text{MLE}})$, which are the MLE for complete-data likelihood integrating sampled missing values, as detailed in Subsection 3.3.1;
2. With step-size $\eta_t = \begin{cases} 1, & \text{if } t \le 20 \\ \frac{1}{t-20}, & \text{if } t > 20 \end{cases}$, update

$$\beta^{t+1} \leftarrow \beta^t + \eta_t \left[\beta^t_{MLE} - \beta^t\right].$$

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

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

**Stop**;

**end if**

**end for**

**Output:** Probability of selecting variables $\hat{\gamma} \leftarrow \frac{1}{20}\sum_{t'=t-19}^{t} \gamma^{t'}$ (the average of the last 20 iterations), with threshold of 0.5 for the selection; and estimate with $\hat{\beta} \leftarrow \beta^t \cdot \hat{\gamma}$.

## 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 Algorithm 1. We also give the SLOBE algorithm in Algorithm 2 which is an approximated and accelerated version.

**Algorithm 2** SLOBE: a quick version of ABSLOPE.

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

**for** $t = 1, 2, \cdots,$ Maxit **do**

*(Imputation by expectation)*

1. **for** $j = 1, 2, \cdots, p$ **do** $\gamma_j^t \leftarrow \mathbb{E}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W)$, according to (14);
2. $\theta^t \leftarrow \mathbb{E}(\theta \mid \gamma, \beta, \sigma, W)$, according to (15);
3. $c^t \leftarrow \mathbb{E}(c \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma, W)$, according to (16);
4. **for** $i = 1, 2, \cdots, n$ **do** $X^t_{i,\text{mis}} \leftarrow \mathbb{E}(X_{i,\text{mis}} \mid y, X_{i,\text{obs}}, \beta, \sigma, \mu, \Sigma)$, according to Proposition 2;

*(Maximization of integrated likelihood)*

- $(\beta^{t+1}, \sigma^{t+1}, \mu^{t+1}, \Sigma^{t+1}) \leftarrow (\beta^t_{\text{MLE}}, \sigma^t_{\text{MLE}}, \mu^t_{\text{MLE}}, \Sigma^t_{\text{MLE}})$, 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 $\{j : \hat{\beta}_j \neq 0\}$.

## A.7 Initialization of ABSLOPE

Here we suggest the following starting values:

- $\beta^0$ is obtained from cross-validated LASSO as implemented in *glmnet* *R* package (Friedman et al., 2010). We recommend using $\lambda$ which minimizes the cross-validation

estimate of prediction error (option *s='lambda.min'* in *cv.glmnet*).

- $X^0_{\text{mis}}$ are imputed by PCA (imputePCA) (Josse and Husson, 2016), MICE (van Buuren and Groothuis-Oudshoorn, 2011) 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: $\frac{\|y - X^0_{\text{mis}}\beta^0\|}{\sqrt{n - \|\beta^0\|_0}}$, where $||\beta^0||_0$ is the number of nonzero elements in $\beta_0$;

- $c^0 = \min\left\{\left(\frac{\sum_{j=1}^p \beta_j^0}{\|\beta^0\|_0 + 1}\right)^{-1} \sigma^0 \lambda_{r(\beta^0,1)}, 1\right\}$;

- $\theta^0 = \frac{\|\beta^0\|_0 + a}{p + b}$ where $a$ and $b$ are known parameters of the prior Beta distribution on $\theta$. Here we choose *i)* $a = 0.01n$ and $b = 0.01n$; *ii)* $a = \frac{2}{p}$ and $b = 1 - \frac{2}{p}$;*iii)* $a = 1$ and $b = p$. Our estimation results are not sensitive to the choice of hyperparameters $a$ and $b$.