Statistical inference with incomplete and high-dimensional data - modeling polytraumatized patients

Wei Jiang

Thèse de doctorat

NNT: 2020UPASM013

Statistical inference with incomplete and high-dimensional data—modeling polytraumatized patients

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 574, École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées
Unité de recherche: Centre de mathématiques appliquées de Polytechnique, UMR 7641 CNRS
Référent: Université Paris-Saclay GS Mathématiques

Thèse présentée et soutenue à Palaiseau, le 21 septembre 2020, par

Wei JIANG

Composition du jury:

Bertrand THIRION Directeur de recherche, Inria Saclay	Président
Adeline SAMSON Professeure, Université Grenoble Alpes	Rapportrice
Daniel YEKUTIELI Professeur, Tel Aviv University	Rapporteur
Pierre NEUVIAL Directeur de recherche, Institut de Mathématiques de Toulouse	Examinateur
Julie JOSSE Professeure, École Polytechnique	Directrice
Marc LAVIELLE Directeur de recherche, Inria Saclay	Codirecteur
Malgorzata BOGDAN Professeure, Uniwersytet Wroclawski	Invitée

Résumé

Le problème des données manquantes existe depuis les débuts de l'analyse des données, car les valeurs manquantes sont liées au processus d'obtention et de préparation des données. Dans les applications des statistiques modernes et de l'apprentissage machine, où la collecte de données devient de plus en plus complexe et où de multiples sources d'information sont combinées, les grandes bases de données présentent souvent un nombre extraordinairement élevé de valeurs manquantes. Ces données présentent donc d'importants défis méthodologiques et techniques pour l'analyse : de la visualisation à la modélisation, en passant par l'estimation, la sélection des variables, les capacités de prédiction et la mise en œuvre par des implémentations. De plus, bien que les données en grande dimension avec des valeurs manquantes soient considérées comme des difficultés courantes dans l'analyse statistique aujourd'hui, seules quelques solutions sont disponibles.

L'objectif de cette thèse est de développer de nouvelles méthodologies pour effectuer des inférences statistiques avec des données manquantes et en particulier pour des données en grande dimension. La contribution la plus importante est de proposer un cadre complet pour traiter les valeurs manquantes, de l'estimation à la sélection d'un modèle, en se basant sur des approches de vraisemblance. La méthode proposée ne repose pas sur un dispositif spécifique du manque, et permet un bon équilibre entre qualité de l'inférence et implémentations efficaces.

Les contributions de la thèse se composent en trois parties. Dans le chapitre 2, nous nous concentrons sur la régression logistique avec des valeurs manquantes dans un cadre de modélisation jointe, en utilisant une approximation stochastique de l'algorithme EM. Nous étudions l'estimation des paramètres, la sélection des variables et la prédiction pour de nouvelles observations incomplètes. Grâce à des simulations complètes, nous montrons que les estimateurs sont non biaisés et ont de bonnes propriétés en termes de couverture des intervalles de confiance, ce qui surpasse l'approche populaire basée sur l'imputation. La méthode est ensuite appliquée à des données pré-hospitalières pour prédire le risque de choc hémorragique, en collaboration avec des partenaires médicaux - le groupe Traumabase des hôpitaux de Paris. En effet, le modèle proposé améliore la prédiction du risque de saignement par rapport à la prédiction faite par les médecins.

Dans les chapitres 3 et 4, nous nous concentrons sur des questions de sélection de modèles pour les données incomplètes en grande dimension, qui visent en particulier à contrôler les fausses découvertes. Pour les modèles linéaires, la version bayésienne adaptative de SLOPE (ABSLOPE) que nous proposons dans le chapitre 3 aborde ces problématiques en intégrant la régularisation triée l_1 dans un cadre bayésien "spike and slab". Dans le chapitre 4, qui vise des modèles plus généraux que celui de la régression linéaire, nous considérons ces questions dans un cadre dit de "model-X", où la distribution conditionnelle de la réponse en fonction des covariables n'est pas spécifiée. Pour ce faire, nous combinons une méthodologie 'knockoff"

et des imputations multiples. Grâce à une étude complète par simulations, nous démontrons des performances satisfaisantes en termes de puissance, de FDR et de biais d'estimation pour un large éventail de scénarios. Dans l'application de l'ensemble des données médicales, nous construisons un modèle pour prédire les niveaux de plaquettes des patients à partir des données pré-hospitalières et hospitalières.

Enfin, dans le chapitre 5, nous fournissons deux logiciels libres avec des tutoriels, afin d'aider la prise de décision dans le domaine médical et les utilisateurs confrontés à des valeurs manquantes.

Abstract

The problem of missing data has existed since the beginning of data analysis, as missing values are related to the process of obtaining and preparing data. In applications of modern statistics and machine learning, where the collection of data is becoming increasingly complex and where multiple sources of information are combined, large databases often have an extraordinarily high number of missing values. These data therefore present important methodological and technical challenges for analysis: from visualization to modeling including estimation, variable selection, predictive capabilities, and implementation through implementations. Moreover, although high-dimensional data with missing values are considered common difficulties in statistical analysis today, only a few solutions are available.

The objective of this thesis is to provide new methodologies for performing statistical inferences with missing data and in particular for high-dimensional data. The most important contribution is to provide a comprehensive framework for dealing with missing values from estimation to model selection based on likelihood approaches. The proposed method doesn't rely on a specific pattern of missingness, and allows a good balance between quality of inference and computational efficiency.

The contribution of the thesis consists of three parts. In Chapter 2, we focus on performing a logistic regression with missing values in a joint modeling framework, using a stochastic approximation of the EM algorithm. We discuss parameter estimation, variable selection, and prediction for incomplete new observations. Through extensive simulations, we show that the estimators are unbiased and have good confidence interval coverage properties, which outperforms the popular imputation-based approach. The method is then applied to pre-hospital data to predict the risk of hemorrhagic shock, in collaboration with medical partners - the Traumabase group of Paris hospitals. Indeed, the proposed model improves the prediction of bleeding risk compared to the prediction made by physicians.

In chapters 3 and 4, we focus on model selection issues for high-dimensional incomplete data, which are particularly aimed at controlling for false discoveries. For linear models, the adaptive Bayesian version of SLOPE (ABSLOPE) we propose in Chapter 3 addresses these issues by embedding the sorted l_1 regularization within a Bayesian spike-and-slab framework. Alternatively, in Chapter 4, aiming at more general models beyond linear regression, we consider these questions in a model-X framework, where the conditional distribution of the response as a function of the covariates is not specified. To do so, we combine knockoff methodology and multiple imputations. Through extensive simulations, we demonstrate satisfactory performance in terms of power, FDR and estimation bias for a wide range of scenarios. In the application of the medical data set, we build a model to predict patient platelet levels from pre-hospital and hospital data.

Finally in Chapter 5, we provide two open-source software packages with tutorials, in order to help decision making in medical field and users facing missing values.

Acknowledgements

D'abord je tiens à remercier Julie, qui m'a encadré tout au long de cette thèse et qui m'a fait partager ses brillantes intuitions. Merci de m'avoir appris la richesse des statistiques et l'importance des expériences de simulation. Merci également pour sa gentillesse, sa disponibilité permanente et pour les nombreux encouragements qu'elle m'a prodigué. Et merci pour tous les bons moments passés en dehors du travail, à Paris et à l'étranger.

Je remercie Marc du fond du cœur pour tout ce qu'il m'a appris. Depuis trois ans, c'est à ses côtés que l'on m'a fait comprendre ce que signifient rigueur et précision. Et merci de m'avoir proposé des sujets et des algorithmes passionants.

J'adresse tous mes remerciements à Adeline Samson et Daniel Yekutieli pour avoir accepté de rapporter ce travail. Daniel, thank you very much for accepting to review this manuscript of thesis and I am so honored that you agreed to be a member of the committee. Également, merci à Karim Lounici, Pierre Neuvial et Bertrand Thirion, je suis très honorée que vous avez accepté la demande de faire partie du jury de ma thèse, même dans une période aussi compliquée.

Thanks to all those with whom I had the chance to collaborate during my thesis: Malgorzata Bogdan, Blazej Miasojedow, Veronika Rockova, Asaf Weinstein and Szymon Majewski. Gosia, thank you for proposing such a nice subject of collaboration and also for offering me the golden opportunity to visit you in Wroclaw University. I cherish so much the time that I spent in Poland with you. Blazej, thank you for supporting solidly the foundation of the work, and for providing relaxing atmosphere during discussion. Veronika, thank you for enhancing and improving the quality of our draft, your opinions are always enlightening. Asaf, thank you for your broad knowledge in model selection, and for questioning always the imprecise part. Szymon, thank you for your rigorous work of the mathematical derivation.

Je tiens aussi à remercier le Groupe TraumaBase de l'APHP, et en particulière à Sophie Hamada, Tobias Gauss et Jean-Denis Moyer, qui m'ont toujours donné des conseils immédiats d'un point de vue médical, et qui améliorent considérablement la qualité du travail. Je suis honorée d'avoir la possibilité de contribuer à la prise de décision sur la base des données de l'hôpital, et d'avoir travaillé avec des personnes gentilles. Merci également à Capgemini pour l'organisation des sessions des travaux, en particulière pour le TrauMatrix plénières et le Data Science workshop de SFAR.

Je tiens à remercier tout particulièrement DIM Math Innov ainsi que FMJH pour leur aide et le financement de ces travaux. Merci à Dominique Wetzel pour l'organisation des événements scientifiques pour nous et pour ses salutations chaleureuses pendant l'épidémie.

Merci du fond du cœur à mes amis qui ont relu et corrigé des parties de cette thèse : Aude et Pavlo. Aude, merci beaucoup pour ton amitié, les organization du group meeting, les pauses café. Pavlo, thank you for your kind emails and your help. I would like to also thanks other nice and brilliant persons who help and support me a lot during the thesis:

Manuel, thank you for letting me realize how essential our work is. Imke, merci beaucoup pour ton amitié et pour ne m'avoir jamais laissé seul. Geneviève, merci d'être toujours au centre et de nous diriger. Patrick, merci pour ton aide et tes idées pendant ces journées en Pologne. Zoltan, thank you for responding my questions with great patience and organizing the sessions for reading group every week.

Merci aussi aux camarades du CMAP, en particulier : Léa, Thomas, Alejandro, Othmane, Ruben, Pierre et Marc Arthur, je suis heureux d'avoir partagé le bureau avec vous pendant la première année. Merci aussi aux camarades du XPOP, en particulière à Belhal et Yao pour vos amitiés. Thanks for other students and post-doc of Julie: Judith, Bénédicte, Tanu, Marine, Antoine, Nicolas, Aravinth and Teresa, no matter how long you have been in the group. Merci aussi à l'équipe administratif au CMAP, en particulier Nasséra, Alexandra, Laura, Maud et Willfried, qui toujours nous aident et nous soignent.

I would like to thank all the nice people who welcomed and helped me in Wroclaw and Warsaw: Michal, Mateusz, MI^2 DataLab team members especially Alicja and Pzymek. Thanks for the warm people who encouraged me during XLV Konferencja Statystyka Matematyczna in Bedlewo. Dziękuję !

最后， 感谢家人的支持和鼓励。感谢父母，尊重我的选择和意愿，给予自由成长的空间，谁言寸草心，报得三春晖。感谢琳，一直以来的陪伴、付出和鼓励，长风破浪会有时， 直挂云帆济沧海。

Scientific production

Articles

- **Jiang W.**, Josse J., Lavielle M., TraumaBase Group (2018). Logistic regression with missing covariates—parameter estimation, model selection and prediction. *published, Computational Statistics & Data Analysis*.

- **Jiang W.**, Bogdan M., Josse J., Miasojedow B., Rockova V., TraumaBase Group (2019). Adaptive Bayesian SLOPE—high-dimensional model selection with missing values. *in revision, Journal of Computational and Graphical Statistics*.

- **Jiang W.**, Majewski S., Bogdan M., Josse J., Weinstein A. (2020). Knockoff with missing values. *preprint*.

Packages

- R package MISAEM (2018), available on *CRAN*, with vignettes.

- R package ABSLOPE (2019), implementation with RCPP, with vignettes.

- Medical application (TraumaBase): MISAEM is used to build a mobile application in the ambulance for predicting hemorrhagic shock.

Awards

- Two-months visiting student researcher (at Wroclaw University) fellowship, Junior Scientific Visibility program, FMJH (2019).

- First prize of young researchers' presentation, Polish Mathematical Statistics Conference (2019).

Communications in scientific congresses

1st year (2017-2018):

- Dec. 2017, London: 10th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2017), oral communication.

- Apr. 2018, Nice: workshop StatLearn, poster.
- Jun. 2018, Paris : Journées de Statistique de la SFdS (JdS2018), oral communication.
- Jun. 2018, Paris : Data Science Summer School (DS3), Ecole Polytechnique, poster.
- Jul. 2018, Rennes : 7e Rencontres R, oral communication.
- Sep. 2018, Wroclaw : Mathematics joint meeting, session "Challenges and Methods of Modern Statistics", oral communication.

2nd year (2018-2019):

- Dec. 2018, Besançon: Proba-Stat seminar, invited presentation.
- May 2019, Paris: workshop Data Science, Société Française d'Anesthésie et de Réanimation (SFAR), presentation of TraumaBase project.
- Mar. 2019, Munich: conference DAGStat, oral communication.
- Jun. 2019, Paris : Data Science Summer School (DS3), Ecole Polytechnique, poster.
- July 2019, Toulouse : conference useR! 2019, oral communication.

3rd year (2019-2020):

- Dec. 2019, Bedlewo, conference Mathematical Statistics, oral communication.
- Dec. 2019, Warsaw & Wroclaw, R enthusiastic meetups, invited presentation.
- May 2020, (Online) International Seminar on Selective Inference, panelist (answering questions on the Q&A board) for talk by M. Bogdan on joint work.
- August 2020, (Online) Bernoulli-IMS One World Symposium, poster.

Teaching

- Teaching assistant for course MAP 536, Ecole Polytechnique - Python/R for Data Science (2017 - 2018).
- Teaching assistant for course MAP 531, Ecole Polytechnique - Statistics Refresher (2018 - 2019).
- Teaching assistant for course MAP 536, Ecole Polytechnique - R (2018 - 2019).
- Two-hours tutorial, Wroclaw University: A missing values tour in R - with a special focus on parameters estimation (Dec. 2019).

Others scientific activities

- Article with special focus on Covid-19: Gu C., **Jiang W.**, Zhao T., Zheng B. (2020). Mathematical recommendations to fight against COVID-19.
- Co-organizing "Group Meeting Missing Values" – statistical seminar, CMAP, École Polytechnique. (2018 - 2020).
- Reviewing for *Journal of Machine Learning Research* (JMLR); *International Conference on Machine Learning* (ICML).
- Supervising five-month internship of Master student from Data science (2019).
- Interviewing for recruitment of Master Data Science for Business joint X-HEC program (2019-2020).

Contents

List of Figures

List of Tables

Essential nomenclature

n	number of rows in a data frame
p	number of columns in a data frame
X	matrix of covariates in $\mathbb{R}^{n\times p}$
$\mathcal{N}(\mu, \Sigma)$	multivariate normal distribution with parameters: the mean vector μ in $\mathbb{R}^p$ and covariance matrix Σ in $\mathbb{R}^{p\times p}$
X_{obs}	covariates which are observed
X_{mis}	covariates which are missing
y	vector of responses in $\mathbb{R}^n$
ε	the noise vector in $\mathbb{R}^n$
X_i	i-th row of X in $\mathbb{R}^p$
$X_{i,\text{obs}}$	the observed elements in X_i (elements may differ from one individual to another)
$X_{i,\text{mis}}$	the missing elements in X_i (elements may differ from one individual to another)
$X_{i,j}$	(i,j)-th entry of X
$X^\top$	transpose of X
β	vector of parameters in a linear regression or generalized linear regression model in $\mathbb{R}^p$
β_j	j-th parameter in β
θ	a set of all model parameters
$\hat{\theta}$	estimates of θ
$\hat{V}$	estimated variance of $\hat{\theta}$
$i.i.d$	independent and identically distributed
$\circ$	Hadamard product (element-wise product)
$\mathcal{L}(\cdot)$	likelihood function
$\ell(\cdot)$	log-likelihood function
$\mathrm{p}(\cdot)$	probability density function
$\|X\|_1$	ℓ_1 norm of X (the sum of entries in absolute value)
$r(\beta)$	rank of β
$\mathbb{1}_n$	identity vector in $\mathbb{R}^n$
$\varnothing$	empty set

Chapter 1

Introduction

Contents

1.1 Overview of state of the art on missing values problematic

Missing data exist in almost all areas of empirical research. There are various reasons why it may occur, including survey non-response, unavailability of measurements, and lost data. Carrying out statistical analysis on data sets with missing data often requires to put additional knowledge on how missing data are generated. The process by which data become incomplete is called the missing data mechanism (Rubin, 1976; Seaman et al., 2013), and include the following three types: *i)* Missing completely at random (MCAR), in which missingness of

the data is independent of both the observed and the missing values; *ii)* Missing at random (MAR), in which data missingness is independent of the missing values, given the observed data. Missing data with MCAR and MAR are referred to as ignorable non-responses, because maximum likelihood (ML) estimation can be obtained while ignoring these mechanisms. *iii)* When the missingness depends on the missing values themselves given the observed data, the process is missing not at random (MNAR), referred to as non-ignorable non-responses because it is often necessary to model the mechanism that generate missing values to do inference.

The most common practice of dealing with missing data, complete case analysis (or listwise deletion), which confines the analysis to the observations with no missing attributes leads to information loss and estimation bias, unless the missing data are MCAR. It really must be stressed that this approach is no longer feasible in a large-scale context. 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." As an example to illustrate the inadequacy of complete case analysis with large data, they imagine a dataset with n observations and p variables where each entry has a probability 1% to be missing independently. If $p = 5$ then complete case analysis can be acceptable since we still have around 95% observations; however, when dimension is much larger such as $p = 300$, only 5% complete rows are retained.

Many statistical methods have been developed to handle missing values (Schafer, 1997; Little and Rubin, 2019; van Buuren, 2018; Josse and Reiter, 2018; Mayer et al., 2019; Mohan et al., 2013) in an inferential framework, i.e. when the aim is to estimate parameters and their variance from incomplete data. One popular approach to handle missing values is imputation, which consists in replacing the missing values by plausible values to get a completed data that can be analyzed by any methods. One can either impute according to a joint model or using a fully conditional modeling approach (van Buuren, 2018). Powerful methods include imputation by random forest (Stekhoven and Buehlmann, 2012) and by low rank methods (Josse and Husson, 2016; Robin, 2019; Udell and Townsend, 2019). More recently, contributions also include imputation methods based on deep learning techniques, such as a variational autoencoder (Mattei and Frellsen, 2019; Ma et al., 2018) and generative adversarial networks (Yoon et al., 2018), however, these methods require a complete dataset to best train the model. Without parametric assumptions, non-parametric Bayesian strategy (Murray and Reiter, 2016) or recent approach using optimal transport (Muzellec et al., 2020) are attempts in this direction. Nevertheless, even if we manage to impute by preserving as well as possible the joint and marginal distribution of the data, a single imputation can not reflect the uncertainty associated to the prediction of missing values. To achieve this goal, multiple imputation (MI) (Rubin, 2009; van Buuren and Groothuis-Oudshoorn, 2011) consists in generating several plausible values for each missing data (to reflect the variance of prediction given observed data and imputation model) leading to different imputed data sets. Then, the analysis is performed on each imputed data sets and results are combined so that the final variance takes into account the supplement variability due to missing values.

An alternative to handle missing values consists in modifying estimation processes so that they can be applied to incomplete data. For example, one can use the EM algorithm (Dempster et al., 1977) to obtain the maximum likelihood estimate (MLE) despite missing values, accompanied by a supplemented EM algorithm (SEM) (Meng and Rubin, 1991) or Louis' formula (Louis, 1982) for estimating the variance. This strategy is valid under MAR

mechanisms. Even though this approach is perfectly suited to specific inference problems with missing values, there are few solutions or implementations available, even for simple models such as logistic regression. This can be explained because, unlike imputation, EM algorithm relies explicitly on strong parametric assumptions and one has to derive an approach for each statistical technique. But the clear advantage of EM is that one can expect better control of the statistical properties of the approach developed. In addition, as it is often not possible to get an explicit form for the EM algorithm, sampling methods were used such as Monte Carlo sampling (Ibrahim et al., 1999), adaptive rejection sampling (Gilks and Wild, 1992), etc but were time consuming which can also be explained why EM based algorithms were not used in practice.

Another part of the literature focuses on statistical learning problems where the objective is to best predict a response variable knowing that the covariates have missing data. For instance, Josse et al. (2019) show the consistency of the simple mean imputation in prediction. Kapelner and Bleich (2015) provide empirical results of predictive performance of decision trees with missing covariates.

Even if there is a multitude of methods to manage missing data (more than 150 packages exist in the R software as reported in Mayer et al. (2019)), surprisingly, there is really very little solution, to make the selection of models and variables with missing data, especially in large dimensions. In this dissertation, we consider the framework of statistical inference with missing covariates and develop new methodologies of parameter estimation and model selection to handle missing values. These works are motivated from a practical problem on a severe trauma registry for decision making.

1.2 Linear regression with missing values

In this section, we focus on an introductory example—linear regression with missing covariates. Let $y \in \mathbb{R}^n$ be the vector of responses and $X \in \mathbb{R}^{n \times p}$ be the design matrix. Consider the following classical linear regression model for complete data, with the vector of regression coefficients $\beta \in \mathbb{R}^p$:

$$y = X\beta + \varepsilon\,, \tag{1.1}$$

where the noise vector $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_n)$ with variance σ^2.

To estimate its parameters β from incomplete covariates, we will describe on two key methods: multiple imputation and EM algorithm. These method rely on assumptions on the missing values mechanism. First, we start by recalling the classical mechanisms that generate the missing values before sketching the methods.

1.2.1 Notations, ignorable missingness

For each individual i, we define the missing data indicator vector $M_i = (M_{ij}, 1 \leq j \leq p)$, with $M_{ij} = 1$ if X_{ij} is missing and $M_{ij} = 0$ otherwise. The matrix $M = (M_i, 1 \leq i \leq n)$ then defines the missing data pattern. The missing data mechanism is characterized by the conditional distribution of M given X and y, with parameter ϕ, *i.e.*, $\mathrm{p}(M_i \mid X_i, y_i, \phi)$. In addition, for one realization m of M, we can denote by $\mathrm{obs(m)}$ (resp. $\mathrm{mis(m)}$) the indices of the zero entries in m (resp. non-zero). Then we can decompose the i-th realization X_i

with missing pattern m into a subset containing the observed data $X_{i,\text{obs(m)}}$ and that for the missing data $X_{i,\text{mis(m)}}$. To lighten notations, when there is no ambiguity, we remove the explicit dependence in m and write, *e.g.*, $X_{i,\text{obs}}$ and $X_{i,\text{mis}}$. Then we let X_{obs} (resp. X_{mis}) be the observed (resp. missing) entries in the entire design matrix X.

Example 1. *To illustrate the problem, we consider a simple example of a data matrix:*

$$X = \begin{pmatrix} 1.3 & 2.4 & 2.3 \\ 2.1 & 3.2 & 3.6 \\ 1.8 & 4.1 & 4.2 \end{pmatrix} .$$

However, only the incomplete design matrix is available, denoted by $\tilde{X}$. Missing values are symbolized with NA, representing "not available" and we assume for all value $x \in \mathbb{R}$, $NA \times x = NA$ and $NA \times 0 = 0$ (Morvan et al., 2020). Then we have: $\tilde{X} = X \circ (1 - M) + NA \circ M$, where the $\circ$ is used for Hadamard product. In summary, the available information can be given for instance:

$$\tilde{X} = \begin{pmatrix} 1.3 & 2.4 & 2.3 \\ NA & 3.2 & 3.6 \\ NA & NA & 4.2 \end{pmatrix} , \quad M = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} .$$

In addition the decomposition of each row of X is give as:

$$X_1 = X_{1,\text{obs}} = (1.3, 2.4, 2.3), \quad X_{1,\text{mis}} = \varnothing ,$$

$$X_2 = (2.1, 3.2), \quad X_{2,\text{obs}} = (3.2, 3.6), \quad X_{2,\text{mis}} = 2.1 ,$$

$$X_3 = (1.8, 4.1), \quad X_{2,\text{obs}} = 4.2, \quad X_{2,\text{mis}} = (1.8, 4.1) .$$

Now we can detail the definition of missing mechanisms. In the following, we use classical notations from Little and Rubin (2019) even if these definitions are often subject to debates (Seaman et al., 2013).

MCAR means that there is no relationship between the missingness of the data and any values, observed or missing. In other words, MCAR implies that:

$$\mathrm{p}(M_i \mid y, X_i, \phi) = \mathrm{p}(M_i|\phi).$$

MAR means that the probability to have missing values may depend on the observed data, but not on the missing data. Thus, the MAR assumption implies that, for each individual i,

$$\mathrm{p}(M_i \mid y_i, X_i; \phi) = \mathrm{p}(M_i \mid y_i, X_{i,\text{obs}}; \phi).$$

Let us consider a likelihood framework to perform regression with missing values. With missing values, we need to consider a probabilistic framework and a distribution for the X. Let us consider a joint Gaussian distribution where $X_i \sim \mathcal{N}(\mu, \Sigma)$. We note $\theta = (\beta, \sigma, \mu, \Sigma)$, the unknown parameters. Let's first recall the likelihood function of θ based on the complete data (y, X):

$$\mathcal{L}(\theta \mid y, X) = \prod_{i=1}^{n} p(y_i, X_i \mid \theta) ,$$

and its logarithm form:

$$\ell(\theta \mid y, X) = \log \mathcal{L}(\theta \mid y, X) = \sum_{i=1}^{n} \log p(y_i, X_i \mid \theta)\,.$$

Then the MLE $\hat{\theta}$ satisfies:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta \mid y, X)\,.$$

With missing data, we want to find θ that maximizes the observed likelihood *i.e.,*

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta \mid y, X_{\text{obs}})\,.$$

Proposition 1 (Ignorable missingness). *The MAR assumption implies that the distribution of M can be ignored when maximizing the observed likelihood (Little and Rubin, 2019). Indeed,*

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

Therefore, to estimate θ, we can maximize $\mathcal{L}(\theta; y, X_{\text{obs}}) = \boldsymbol{p}(y, X_{\text{obs}}; \theta)$.

The strategies suggested in the following sections are only valid under ignorable missingness assumptions, which is common.

Another mechanism MNAR, *i.e.*, for each individual i,

$$\mathrm{p}(M_i \mid y_i, X_i; \phi) = \mathrm{p}(M_i \mid y_i, X_{i,\text{obs}}, X_{i,\text{mis}}; \phi)\,,$$

will result in significantly biased estimation since the observed variables cannot represent the population anymore. Therefore, the missing mechanism should be also be considered. It can be considered explicitly to form the likelihood function, for example the distribution for missing values is often assumed to be logistic; however it requires strong parametric apriori, complicates the inference, burdens the computational cost and is often limited to cases with few MNAR variables. Other works propose methods which don't explicitly model the missing mechanism as in Mohan et al. (2018); Tang et al. (2003). MNAR also raises the problem

of identifiability (the distribution of data is identifiable only if the mechanism is identifiable) and most work focus on it, for instance Nabi et al. (2020) address the characterization of model identifiability based on graphical models, by adopting a causal point of view for MNAR. More discussion for inference with MNAR for specific models is available in Stubbendick and Ibrahim (2003, 2006); Tchetgen et al. (2018); Sportisse et al. (2018, 2019).

1.2.2 Estimation via MLE and EM algorithm

Often there are no explicit solution to the MLE of the observed likelihood and one can resort to an EM algorithm (Dempster et al., 1977). The EM algorithm starts with an initial value θ^0, and iterate as follows until convergence. Letting θ^t be the estimate of θ at t-th iteration, the $(t+1)$-th iteration of EM is processed as follows:

- *E step.* Find the expectation of complete-data log-likelihood, with respect to the conditional distribution of missing variables given the observed ones and if θ were θ^t:

$$Q(\theta \mid \theta^t) = \mathbb{E}\left(\ell(\theta \mid X, y) \mid X_{\text{obs}}, y, \theta^t\right) = \int \ell(\theta \mid X, y)\mathrm{p}(X_{\text{mis}} \mid X_{\text{obs}}, y, \theta^t) dX_{\text{mis}}\,.$$

- *M step.* Determine θ^{t+1} by maximizing this expected log-likelihood

$$\theta^{t+1} = \arg\max_{\theta} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t)\,.$$

In a specific case of joint multivariate Gaussian assumption on X, the calculation can be simplified. The approach consists in considering the joint Gaussian distribution for (y, X).

$$(y, X) \sim \mathcal{N}(\mu_{y,X}, \Sigma_{y,X}) \quad \text{with} \quad \mu_{y,X} = \begin{pmatrix} \mu_y \\ \mu_X \end{pmatrix} \quad \text{and} \quad \Sigma_{y,X} = \begin{pmatrix} \Sigma_y & \Sigma_{y,X} \\ \Sigma_{X,y} & \Sigma_X \end{pmatrix}\,.$$

The E step in an exponential family follows from standard complete-data theory for means of conditional distributions, and the M step uses the identical computational method as MLE from complete data. Once the parameters of the Gaussian distribution are estimated, the parameters of the regression can be directly obtained by plug-in the estimates into the expression

$$\beta = (\mu_y - \Sigma_{y,X}\Sigma_X^{-1}\mu_X\,, \Sigma_{y,X}\Sigma_X^{-1})^\top\,.$$

In the same way, the variance can be estimated via the classical formula as:

$$\hat{V}(\beta) = \mathsf{diag}(C) \quad \text{with} \quad C = \frac{1}{n}\left(\Sigma_y - \beta^\top \Sigma_X \beta\right)(\Sigma_X + \mu_X\mu_X^\top)^{-1}\,. \tag{1.2}$$

Note that there are no differences between the response variable and the explanatory variables, since their role is symmetrical and only the joint distribution matters.

For the prediction on a test set, one can apply the classical formula:

$$\mathbb{E}(y \mid X) = (\mu_y - \Sigma_{y,X}\Sigma_X^{-1}\mu_X) + \Sigma_{y,X}\Sigma_X^{-1}X\,.$$

If the covariates in test set also contain missingness, by applying the decomposition $X = (X_{\text{mis}}, X_{\text{obs}})$ as introduced in Section 1.2.1, we have:

$$\mathbb{E}(y, X_{\text{mis}} \mid X_{\text{obs}}) = (\mu_{y,\text{mis}} - \Sigma_{(y,\text{mis}),\text{obs}}\Sigma_{\text{obs}}^{-1}\mu_{\text{obs}}) + \Sigma_{(y,\text{mis}),\text{obs}}\Sigma_{\text{obs}}^{-1}X_{\text{obs}}\,,$$

with $\mu_{y,\text{mis}}$ (resp. μ_{obs}) the missing (resp. observed) elements of μ for the new observation. The covariance matrix Σ is decomposed in the same way. In this way, y for a new observation is predicted using the observed information and estimated parameters.

One example of implementation of EM with incomplete multivariate Gaussian data is provided in Novo and Schafer (2013) using the SWEEP operator (Schafer, 1997) (see Section 1.6 for details). As far as we know, the linear regression with missing values hasn't yet been implemented. One of the contribution is, based on the SWEEP operator, the implementation of the linear regression with missing values via EM, model selection via BIC, and prediction on test set with missing values. The R package `misaem` (Jiang and Mozharovskyi, 2020) is available in CRAN and is introduced in Chapter 5.

1.2.3 Estimation via multiple imputation

Multiple imputation consists in generating different imputed values for each missing entries leading to say K completed datasets, and then estimating the parameters, here the β from each imputed dataset, by fitting linear regression respectively. Let's denote by $\hat{\beta}_k$, the set of estimate from the k^{th} completed data, and its estimated variance $\hat{V}_k$, for $k \in \{1, \ldots, K\}$. In the final pooling step, results are combined using Rubin's rules (Rubin, 2009). The estimate is obtained by taking the average over $\hat{\beta}_k$ from all K imputed datasets:

$$\hat{\beta} = \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_k \,,$$

and the estimated variance is obtained by combining the within-imputation component $\hat{V}_{\text{within}}$ and between-imputation component $\hat{V}_{\text{between}}$ of overall variance as follows

$$\begin{aligned} \hat{V} &= \hat{V}_{\text{within}} + \left(1 + \frac{1}{K}\right) \hat{V}_{\text{between}} \\ &= \frac{1}{K} \sum_{k=1}^{K} \hat{V}_k + \left(1 + \frac{1}{K}\right) \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})(\hat{\theta}_k - \hat{\theta})^\top \end{aligned}$$

Figure 1.1 illustrates the main steps described above.

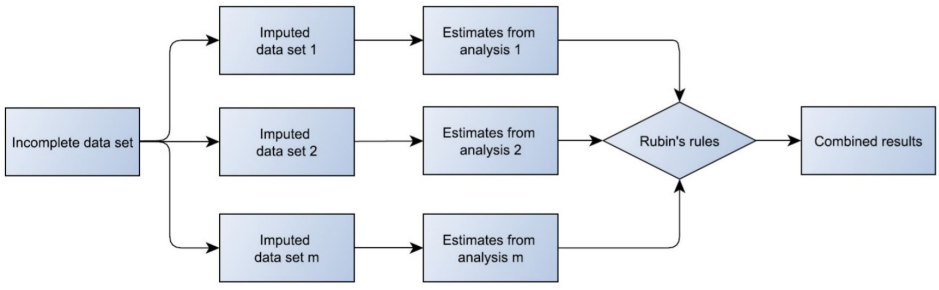


Figure 1.1: Procedure of multiple imputation.

When the imputation model is in agreement with the analysis model (here the regression model) then multiple imputation can lead to unbiased estimates and confidence intervals with

good coverage properties. For instance, here, we can impute the data by assuming a joint Gaussian distribution for (y, X) or by using iterative conditional imputation with regression models of each variables given the others. More details can be found in Murray et al. (2018). For the former approach, an EM algorithm will be performed to estimate $\mu_{y,X}$ and $\Sigma_{y,X}$, then missing values are imputed by drawing K times from their predictive distribution, i.e. a Gaussian distribution of the missing values given observed values and estimated parameters. An additional layer of bootstrapping is added to get what is called proper multiple imputation (Little and Rubin, 2019; Efron, 1994).

Note that even though the aim is to perform a model of y on X, multiple imputation is often carried out including the response variable in practice, which is quite debatable as detailed in the following section.

1.2.4 Which method to use?

People often ask the question in the treatment of missing values, which method should we use to achieve the most optimal estimate. However, we cannot say any of them is the best, since the effectiveness depends on a combination of factors: the distribution of variables, the structure of relationship, the percentage of missing values, and the mechanism underlying the missing values. Therefore, before applying any method on the dataset, researchers are encouraged to explore the data as much as possible, such as summarizing the variables and visualizing the pattern of missing values.

Both EM and multiple imputation have advantages and drawbacks. One drawback of EM is that it cannot provide variance estimate along, but must accompanied by another method such as SEM (Meng and Rubin, 1991) or a bootstrap procedure (Efron, 1982). EM algorithm is a model-based approach so a specific EM algorithm needs to be implemented for each statistical analysis that one wants to apply on a dataset. In addition, such EM algorithms are not necessarily straightforward to establish, in contrast to multiple imputation where once the data have been multiply imputed, one can apply any methods. Multiple imputation is easier to implement and we can expect both unbiased estimate and its variance if a proper imputation model and aggregation rule are taken. However, it's debatable whether to use response variable y in the imputation model, for example, D'Agostino Jr and Rubin (2000) suggested not include the response to avoid conjecturing values for the missing covariates, having already observed in the response values. While according to empirical results, some researches (Donders et al., 2006; Little, 1992) encourage the use of response and claimed that the response contain essential information about the distribution of missing covariates. As a simple example, one can consider two variables positively correlated y and X_1 and there are missing values only on X_1, the aim is to do regression of y on X_1; to impute X_1 it is better to use y than X_1 alone. From empirical study, the performance of EM and imputation can vary in different model, for instance, the likelihood based method performs better in factor analysis studies compared to imputation (Bernaards and Sijtsma, 1999).

The debate for optimal methods lasts long and further abundant discussion is provided in the literature (Pan and Bai, 2015; Little, 1992; Donders et al., 2006).

1.3 Model selection with missing values

Handling missing data within the context of high-dimensional variable selection is a very important problem. Indeed, missing data are omnipresent. For example, genetic data obtained from microarray experiments often contain missing values for several reasons: insufficient resolution, image corruption, manufacturing errors, etc. And a vast number of predictors is available but only a few are deemed relevant for explaining biological phenomena.

1.3.1 Model selection with complete data

First let's review the literature of model selection methods in a common case without any missing value. Traditional regression methods, in which a variable is added or removed from the model based on criteria such as AIC and BIC, have been widely used, but require intensive calculations for large data. In general paradigms of modern statistics, solutions are usually suggested based on sparse recovery. For instance, in the classical linear regression model as described in eq.(1.1) but consider $p >> n$, the parameter vector β is assumed to be sparse. The model selection problem can be viewed as a multiple testing:

$$\mathcal{H}_0 : \beta_j = 0 \quad \leftrightarrow \quad \mathcal{H}_1 : \beta_j \neq 0, \quad j = 1, 2, \cdots, p.$$

A general evaluation on the model selection results of an estimator $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ is based on both power and false discovery rate (FDR), defined as follows:

$$\text{Power} = \mathbb{E}\left(\frac{\#\text{selected true variables}}{\#\text{ all important variables}}\right) = \mathbb{E}\left(\frac{\#\{\text{j} : \hat{\beta}_\text{j} \neq 0 \wedge \beta_\text{j} \neq 0\}}{\max(1, \#\{\text{j} : \beta_\text{j} \neq 0\})}\right),$$

$$\text{FDR} = \mathbb{E}\left(\frac{\#\text{ selected null variables}}{\#\text{ selected variables}}\right) = \mathbb{E}\left(\frac{\#\{\text{j} : \hat{\beta}_\text{j} \neq 0 \wedge \beta_\text{j} = 0\}}{\max(1, \#\{\text{j} : \hat{\beta}_\text{j} \neq 0\})}\right).$$

In this context, the LASSO (Tibshirani, 1996), now a default penalized likelihood method, has proved itself to be successful at simultaneously estimating parameters and covariate sets. The LASSO aims at solving:

$$\hat{\beta}_{\text{LASSO}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|y - X\beta\|^2 + \sigma\lambda\|\beta\|_1 \right\},$$

where the penalty coefficient $\lambda \geq 0$. While LASSO possesses nice theoretical guarantees, it may lead to false discoveries (Su et al., 2017) and it allows to identify the true model only under rather strict irrepresentability conditions (Wainwright, 2009; Tardivel and Bogdan, 2018). The adaptive LASSO variant (Zou, 2006) instead uses a weighted ℓ_1 penalty:

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

where $w = (w_1, w_2, \cdots, w_p)$ is the weighting vector as a function of some initial estimates of regression coefficients. By adjusting regularization, adaptive LASSO reduces bias in estimation and can be consistent for variable selection even when the irrepresentability condition

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 formulation with a spike-and-slab prior:

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

where ϕ_1 serves as a slab distribution for modeling large effects, ϕ_0 as a spike one for modeling negligibly small effects, and $\gamma_j \in \{0,1\}$ indexes all the possible subset models. 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 et al. (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.

Another alternative is Sorted L-One Penalized Estimator (SLOPE) method of Bogdan et al. (2015) by shrink larger coefficient more stringently:

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

where the penalty coefficients $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and the absolute values of elements in β are sorted in a decreasing order $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}$. The main motivation behind SLOPE was the control of the FDR, which is also the central goal of many methodological developments in multiple regression (see *e.g.* Barber et al. (2015); Candes 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).

Other extension of the LASSO targeted on optimal trade-off between false positive and true positive includes LASSO-Zero (Descloux and Sardy, 2018) which, in a fist step, solves the basis pursuit problem (Chen et al., 2001) (relevant for the noiseless case in standard linear models), for an augmented design matrix with a noise dictionary consisting in a random matrix (to fit the noise term), and in a second step the method thresholds the obtained solution to retain only the largest coefficients. In summary, Table 1.1 illustrates the advantages and disadvantages for the representative methods mentioned above.

Table 1.1: Comparison of various model selection methods based on sparse regression.

Methods	Advantages	Disadvantages	Theoretical results
LASSO	default penalized method; good power	many false discoveries; parameters tuning required; strict irrepresentatbility condition	sign recovery; support recovery
Adaptive LASSO	bias reduction compared to LASSO	many false discoveries; depending on the weight function; parameters tuning required	consistency
SSL	parameters tuning less required; bias reduction by providing prior support	many false discoveries	convergence
SLOPE	FDR control	large bias due to over shrinkage	FDR control; convergence

1.3.2 Previous work with missing values

There are only a few methods for selecting a model with missing values, whether to recovering important variables or controlling the FDR. For example, in generalized linear models, Claeskens and Consentino (2008); Ibrahim et al. (2008) 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 is bounded by a constant which depends on the sparsity level rarely known in practice. In other 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 extension based on LASSO also includes convex conditioned LASSO (Datta et al., 2017), with asymptotic sign-consistent selection property, but can handle only MCAR data. More recently, targeted on sign recovery, Descloux et al. (2020) reformulated the missing covariates into a sparse corruption problem and then solves it with robust LASSO-Zero method, which considered, instead of the basis pursuit problem, solving the justice pursuit problem (Laska et al., 2009) (relevant for the noiseless case in the sparse corruption problem), adding a random matrix to account for noise. The method of Descloux et al. (2020) is robust to MNAR data given theoretical guarantees on the sign recovery, however, empirical results are only satisfactory when the sparsity index and/or proportion of missing entries is low.

An alternative to do variable selection with missing values could be to do multiple imputation, apply a variable selection on each imputed dataset and combine the results. However, there is no common solution to aggregate the selected variables, since different imputed dataset can return different models (different sets of variables) and the Rubin's rules (Rubin, 2009) only serve for aggregating estimators (as a regression coefficient). Liu et al. (2016) combined penalized regression techniques with multiple imputation, where they showed in empirical study, good selection performance when the correlation among variables and missing proportion are high. However, aggregating different models for the resulting multiple imputed data sets becomes increasingly complex as the number of data grows.

Despite recent advances, the model selection with missing values remains under-developed. Interesting theoretical guarantees are often obtained under restrictive assumptions. Methodology for specific purpose, such as FDR control or non-parametric regression models, has non been explored yet with missing values.

1.4 TraumaBase project

Our work is motivated by a collaboration with the TraumaBase[1] group at APHP (Public Assistance - Hospitals of Paris), which is dedicated to the management of severely traumatized patients.
Major trauma refers to injuries that cause prolonged disability or endanger a person's life, such as injury from road accidents, interpersonal violence and falls. The World Health Organization has recently reported that major trauma is a prominent source of mortality and morbidity around the world (Hay et al., 2017). In particular, major trauma is the leading cause of mortality and second cause of disability in the 16–45 age group, while hemorrhagic shock and traumatic brain injuries are the two leading causes of early preventable death in severe trauma patients (Dutton et al., 2010; Kauvar and Wade, 2005).
The path of a traumatized patient involves several stages: from *1)* the accident site, where care is typically provided by emergency care teams, to transfer to *2)* resuscitation room of a trauma center, where immediate interventions such as CT-scan assessment, emergency surgery or radiology can be organized, followed by the admission in *3)* intensive-care unit for organ dysfunction support optimization, and finally *4)* a comprehensive care at the hospital, as presented in Figure 1.2. Due to the highly stressful and multi-player environments involved, evidence suggests that the patient management—even in mature trauma systems—often exceeds acceptable time frames (Hamada et al., 2014). In addition, discrepancies may be observed between the diagnoses made by emergency doctors in the ambulance and those made when the patient arrives at the trauma center (Hamada et al., 2015). Such discrepancies can result in poor outcomes like inadequate hemorrhage control and delayed transfusion. To improve decision-making and patient care, 19 French trauma centers have collaborated since 2011 to collect detailed high-quality clinical data from the accident site right through to the hospital. Some centers joined TraumaBase after January 2011. The resulting database, TraumaBase, is a multicenter prospective trauma registry that is continually updated, and now has data up to 20 000 trauma cases. Sociodemographic, clinical, biological and therapeutic data (from the pre-hospital phase to discharge, if hospitalized) are systematically recorded for all trauma patients, and all patients transported to the emergency rooms of

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

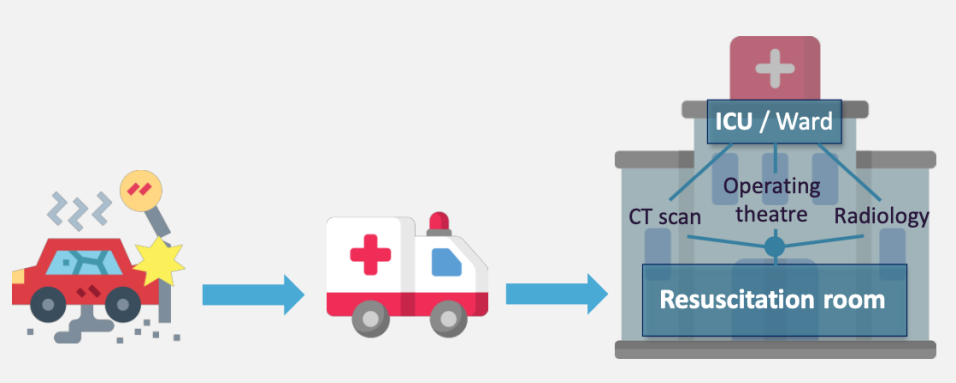


Figure 1.2: Management scheme of a traumatized patient.

participating centers are included in the registry. The sheer quantity of collected data (with more than 250 variables) makes this dataset unique in Europe. However, these data—coming from multiple sources— have high inter-center variability, not to mention the fact that a lot of data are missing, both of which problems make modeling challenging.

One aim of the project is to model the decisions and events taken by the emergency doctors

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73
11	Bicetre	Fall	61	m	84	1.7	29.07	144	105
.....									

	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion .....
1	97	35.6	<NA>	12.7	12	yes
2	100	36.5	4.8	11.1	15	no
3	100	36	3.9	11.4	3	no
4	100	36.7	1.66	13	15	yes
6	100	36	NM	14.4	15	no
7	100	36.6	NM	14.3	15	yes
9	100	37.5	13	15.9	15	yes
10	100	36.9	NM	13.7	15	no
11	100	36.6	1.2	14.2	14	no

Figure 1.3: An extract of TraumaBase dataset with various missing data.

to help them making choices in a very stressful environment and avoid discrepancies between the diagnosis made by the emergency doctors and the one made by the doctors when the patient arrives at the Trauma-center. For instance, we would like to establish predictive models to know whether or not to predict the risk of severe hemorrhage, in order to prepare an appropriate response upon arrival at a trauma center, e.g., a massive transfusion protocol and/or immediate haemostatic procedures.

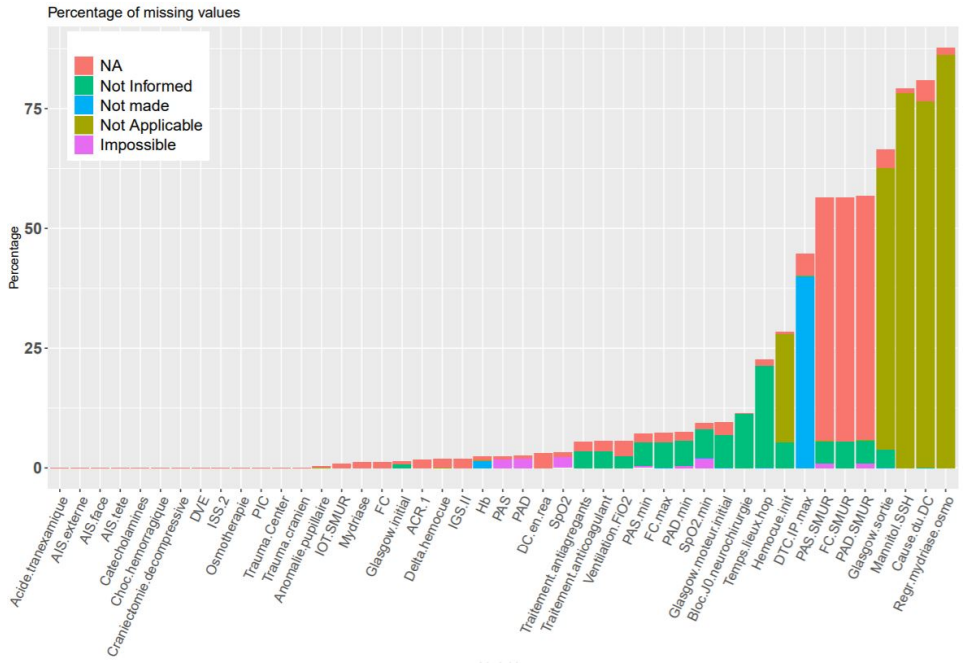


Figure 1.4: Percentage of missing values in each variables in TraumaBase dataset.

From a statistical point of view, the challenges involve performing predictive models such as logistic regressions or regression with many missing values. Other tasks can include model selection to choose the most important measurements to explain the response, in order to help propose an innovative response to the public health challenge of major trauma.

Figure 1.3 shows an extract of dataset, with different coding of missing values (NA for Not Applicable, Imp for impossible, NR for Not Recorded, NM for Not Made), and Figure 1.4 summarizes the percentage of missing values in 45 representative variables among total measures. Reasons why these missingness occurred can be various. For example, when one patient is in a very urgent situation, there is no time left to measure some of the variables (and the medical doctors know, without measuring it, that the values are critical); this case can be considered as MNAR. Other cases include data that have simply been not recorded in the database (Data were measured but not reported in the TraumaBase simply because they are simply forgotten or when they are merged from different sources for example). In addition, as mentioned, some trauma centers have progressively joined the TraumaBase, and they do not have necessarily the same device in each hospital, which results in particular missing data structures with missing columns (corresponding to missing features) for some of the groups. These codes—NR, NM, Imp—can therefore help to understand the nature of the missing data and the reasons for their occurrence. Indeed, even if we will not detail these aspects in this document, the first thing to do when we have missing data is to explore, visualize, make descriptive statistics to understand the missing data. In this work, we have always exchanged with the physicians to see if the hypotheses made seemed plausible. According to the figures, we observe how missingness significantly affects the TraumaBase data, and how essential one needs to design a specific methodology related to missing values.
In this thesis, we investigated a subset of the whole TraumaBase, which contains 7495

individuals logged in the trauma data, included from January 2011 to March 2016, with ages ranging from 12 to 96.

1.5 Summary of contributions

One may realize that the literature of statistical inference with missing values is not abundant enough. Despite EM algorithm is studied thoroughly in these decades, but applications and implementations are limited to the simple models, or with fixed pattern of missingness. As far as we know, none of the available methods address model selection problem to deal with missing values and control FDR simultaneously. The objective of this thesis is to provide efficient and complete statistical methodology to deal with inference problem with existence of missing values, and in particular to the medical application as described in Section 1.4. In addition, user-friendly implementations as R packages are developed. Throughout this dissertation, we assume the MAR mechanism which implies that the missing values mechanism can therefore be ignored when maximizing the likelihood (Little and Rubin, 2019), and we suppose that the missingness occurs only in covariates X but not in the response y. Detailed contributions are listed as follows.

1.5.1 Logistic regression with missing covariates

In Chapter 2, we address the problem of statistical inference for logistic regression model with missing covariates. Surprisingly, there are very few solutions for performing logistic regression with missing values in the covariates, even it's a common model. A complete approach based on a stochastic approximation version of the EM (SAEM) algorithm (Lavielle, 2014; Delyon et al., 1999) is proposed in order to perform statistical inference with missing values, including the estimation of the parameters and their variance, derivation of confidence intervals, and also a model selection procedure. The problem of prediction for new observations on a test set with missing covariate data is also tackled. Supported by a simulation study in which the method is compared to previous ones, it has proved to be computationally efficient, and has good coverage and variable selection properties. The approach is then illustrated on TraumaBase by predicting the occurrence of hemorrhagic shock, a leading cause of early preventable death in severe trauma cases. The aim is to improve the current red flag procedure (Hamada et al., 2018), a binary alert identifying patients with a high risk of severe hemorrhage.

1.5.2 High-dimensional model selection to control FDR

Chapter 3 provides a new methodology to select important variables with missing values, specifically focusing on high-dimensional data where p is comparable to n or even larger than n. We propose a new synergistic procedure—adaptive Bayesian SLOPE (ABSLOPE) – which effectively combines the SLOPE method (sorted l_1 regularization) (Bogdan et al., 2015) together with the Spike-and-Slab LASSO method (Ročková and George, 2018). We position our approach within a Bayesian framework which allows for simultaneous variable selection and parameter estimation, despite the missing values. As with the Spike-and-Slab LASSO, the coefficients are regarded as arising from a hierarchical model consisting of two groups: *1)* the spike for the inactive and *2)* the slab for the active. However, instead of

assigning independent spike priors for each covariate, here we deploy a joint "SLOPE" spike prior which takes into account the ordering of coefficient magnitudes in order to control for false discoveries. Through extensive simulations, we demonstrate satisfactory performance in terms of power, FDR and estimation bias under a wide range of scenarios. Finally, we show excellent performance in predicting platelet levels when analyzing TraumaBase data.

1.5.3 Controlled variable selection with missing values in a model-X framework

Chapter 4 also tackle the problem of model selection with missing values while controlling FDR. However, different from the setting of Chapter 3, we suppose a model-X framework where the conditional distribution of the response given the covariates is not specified, but the joint distribution of covariates is known. Such setting has advantages when the distribution of y given X is complicated such as with non-linear regression model. The newly proposed methodology—missKnockoff is based on the model-X knockoffs method (Candes et al., 2018). Our method uses knockoffs twice: it first substitutes missing values with knockoffs, and then proceeds with a standard application of model-X knockoffs on imputed dataset. In order to account for the uncertainty, multiple imputation is easily incorporated by generating several knockoff copies at the first stage, and we discuss different ways to aggregate the support. We study the performance in terms of power and FDR through extensive simulations.

1.5.4 Implementation and packages

Finally, Chapter 5 provides the instructions about the implementation of the methodologies mentioned above. Two packages are developed to handle the statistical inference with missing values:

- `misaem` is an R (R Core Team, 2017) package to apply statistical inference for linear regression and logistic regression model with missing data. This methodology is based on likelihood, including:
    1. EM-type algorithms to estimate the parameters;
    2. Obtain of variance of estimated parameters;
    3. Model selection procedure based on BIC;
    4. Prediction on test set with missing values.
- `ABSLOPE` is an R package which aims at high-dimensional model selection with missing values via Adaptive Bayesian SLOPE. In addition, an simplified algorithm to accelerate the computing time is also implemented with C++ functions.

1.5.5 Contribution to the TraumaBase

We collaborate with medical partners (the Traumabase group from Paris hospitals) to improve the management and care of severely traumatized patients as detailed in Section 1.4. We have built models with missing values to predict the risk of hemorrhagic shock and level of platelet

given pre-hospital data. Our collaborators, the doctors are extremely satisfied with the results. Indeed, the proposed model of logistic regression with missing values improves the prediction of hemorrhagic risk compared to the prediction made by physicians. Therefore, the objective is now to implement the model in real time, because beyond predictive quality, we must see how physicians will react to such a decision support tool, how to present recommendations with an ergonomic interface and how physicians react to this decision support tool. The results were communicated through French Society of Anesthesia & Intensive Care Medicine (SFAR) meeting and we received constructive comments and strong interest in real time application.

1.6 Supplementary material: sweep operator in EM

In this section, we briefly introduce the sweep operator and then show how to implement EM algorithm to estimate the parameters for multivariate Gaussian distribution using sweep operator.

1.6.1 Definition of sweep operator

Single sweep

Suppose that G is a symmetric matrix of dimension $p \times p$ with element g_{ij}. The sweep operator SWP$[k]$ operates on G by replacing G with another $p \times p$ symmetric matrix H, written as:

$$H = \text{SWP}[k]G\,,$$

where the elements of H are given by:

$$\begin{aligned} h_{kk} &= -\frac{1}{g_{kk}}\,, \\ h_{jk} &= h_{kj} = \frac{g_{jk}}{g_{kk}}, \text{ for } j \neq k\,, \\ h_{jl} &= h_{lj} = g_{jl} - \frac{g_{jk}g_{kl}}{g_{kk}}, \text{ for } j \neq k \text{ and } l \neq k\,. \end{aligned} \tag{1.3}$$

We say that the matrix G is swept on position k.

Successive sweep

Suppose that a symmetric $p \times p$ matrix G can be partitioned as $G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$, where G_{11} is $p_1 \times p_1$. After successive applications of sweeping on positions $1, 2, \cdots, p_1$, the matrix becomes:

$$\text{SWP}[1, 2, \cdots, p_1]G \triangleq \text{SWP}[p_1]\text{SWP}[p_1-1]\cdots\text{SWP}[1]G = \begin{bmatrix} -G_{11}^{-1} & G_{11}^{-1}G_{12} \\ G_{21}G_{11}^{-1} & G_{22} - G_{21}G_{11}^{-1}G_{12} \end{bmatrix}.$$

Remark that the sweep operator is commutative.

Reverse-sweep

It is also convenient to define a reverse-sweep operator that returns a swept matrix to its original form:

$$H = \mathsf{RSW}[k]G\,,$$

where the elements of H are given by:

$$\begin{aligned} h_{kk} &= -\frac{1}{g_{kk}}\,, \\ h_{jk} = h_{kj} &= -\frac{g_{jk}}{g_{kk}}, \text{ for } j \neq k\,, \\ h_{jl} = h_{lj} &= g_{jl} - \frac{g_{jk}g_{kl}}{g_{kk}}, \text{ for } j \neq k \text{ and } l \neq k\,. \end{aligned} \tag{1.4}$$

Remark that the reverse-sweep operator is also commutative. We can verify that:

$$\mathsf{RSW}[k]\mathsf{SWP}[k]G = G\,.$$

1.6.2 EM for a particular pattern

First we consider a simple case of missing pattern for multivariate normal data. Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$. We partition X as $X = (X_1, X_2)$ where X_1 and X_2 are submatrices with dimension $n \times p_1$ and $n \times p_2$ respectively. We know the marginal distribution of X_1 and X_2 are $\mathcal{N}_{p_1}(\mu_1, \Sigma_{11})$ and $\mathcal{N}_{p_2}(\mu_2, \Sigma_{22})$, where $\mu^T = (\mu_1^T, \mu_2^T)$ and $\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$, the block decomposition corresponds to the decomposition of X.

Now we want to implement EM algorithm in a very simple case of missing pattern: all X_1 are observed and all X_2 are missing.

In the setting described above, tth iteration of EM algorithm proceeds as follows:

- *E-step.* We consider another parametrization of the problem:

$$\phi = \begin{bmatrix} -1 & \mu_1^T & \alpha_{2\cdot1}^T \\ \mu_1 & \Sigma_{11} & B_{2\cdot1}^T \\ \alpha_{2\cdot1} & B_{2\cdot1} & \Sigma_{22\cdot1} \end{bmatrix}\,,$$

  where $X_2|X_1 \sim \mathcal{N}_{p_2}(\mu_{2\cdot1}, \Sigma_{22\cdot1})$ with:

$$\begin{aligned} \mu_{2\cdot1} &= \alpha_{2\cdot1} + B_{2\cdot1}X_1\,, \\ \alpha_{2\cdot1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1\,, \\ B_{2\cdot1} &= -\Sigma_{21}\Sigma_{11}^{-1}\,, \\ \Sigma_{22\cdot1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\,. \end{aligned} \tag{1.5}$$

  Given θ^t obtained at the previous iteration, the sweep operator SWP and the reverse sweep operator RSW will allow us to easily compute ϕ^t. Indeed, let

$$\theta^t = \begin{bmatrix} -1 & (\mu^t)^T \\ \mu^t & \Sigma^t \end{bmatrix}\,,$$

we have:

$$\text{SWP}[1,2,\cdots p_1]\theta^t = \left[\begin{array}{ccc|c} & A & & (\alpha^t_{2\cdot1})^T \\ & & & \\ & & & (B^t_{2\cdot1})^T \\ \hline \alpha^t_{2\cdot1} & (B^t_{2\cdot1})^T & & \Sigma^t_{22\cdot1} \end{array}\right],$$

and applying the reverse sweeping operator to the upper left block of the previous matrix gives us:

$$\phi^t = \left[\begin{array}{ccc|c} & RSW[1,\ldots,p_1]A & & (\alpha^t_{2\cdot1})^T \\ & & & \\ & & & (B^t_{2\cdot1})^T \\ \hline \alpha^t_{2\cdot1} & (B^t_{2\cdot1})^T & & \Sigma^t_{22\cdot1} \end{array}\right].$$

In short, With θ^t obtained from last iteration, the E-step apply first the sweep to the full matrix on position $1,2,\cdots p_1$ and then the reverse sweep to the upper-left $(p_1+1)\times(p_1+1)$ submatrix on position $1,2,\cdots p_1$. The result of these two transformation is ϕ^t. Finally we replace the sufficient statistics $T = \begin{bmatrix} n & 1^TX \\ X^T1 & X^TX \end{bmatrix}$ with their expected values:

$$\mathbb{E}(T|X_1,\theta^t) = \begin{bmatrix} n & \mu_1^T & (\mu^t_{2\cdot1})^T \\ \mu_1 & \Sigma_{11}+\mu_1\mu_1^T & X_1(\mu^t_{2\cdot1})^T \\ \mu^t_{2\cdot1} & \mu^t_{2\cdot1}X_1^T & \Sigma^t_{22\cdot1}+\mu^t_{2\cdot1}(\mu^t_{2\cdot1})^T \end{bmatrix},$$

where all the elements can be found in ϕ^t.

- *M-step.* θ^{t+1} may be computed from the sufficient statistics by:

$$\theta^{t+1} = \text{SWP}[0]n^{-1}T\,.$$

1.6.3 EM for general pattern

In general case of missing data, we can specify all possibilities of missing patterns and see how to apply the sweep operator in EM algorithm.

Missing patterns and preliminary manipulations We specify each column of data matrix as $X = (X_1, X_2, \cdot, X_p)$. Let matrix of missingness patterns R be a $S \times p$ matrix of binary indicators with with element r_{sj} where

$$r_{sj} = \begin{cases} 1, \text{ if } X_j \text{ is observed in pattern } s\,, \\ 0, \text{ if } X_j \text{ is missing in pattern } s\,. \end{cases}$$

R is shown as Figure 1.5. For each missingness pattern s, let $O(s)$ and $M(s)$ denote the subsets of column labels $\{1,2,\cdots p\}$ corresponding to variables that are observed and missing respectively:

$$O(s) = \{j : r_{sj} = 1\}\,,$$

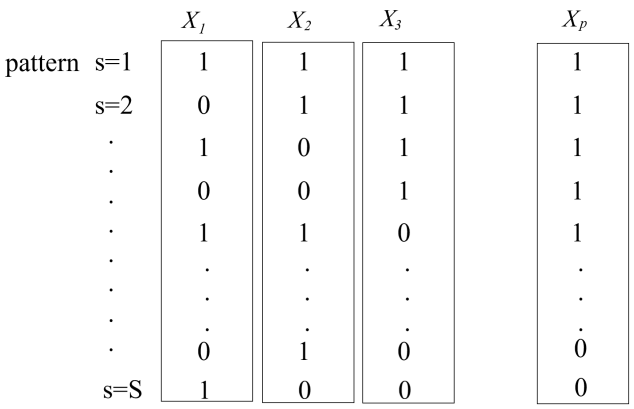


Figure 1.5: Matrix of missingness patterns associated with X with 1 denoting an observed variable and 0 denoting a missing variable

$$M(s) = \{j : r_{sj} = 0\}.$$

Finally let $I(s)$ denote the subset of observation numbers $\{1, 2, \cdots n\}$ corresponding to the rows of data that exhibit pattern s.

Implementation of EM As same as the idea in the simplest case, tth iteration of EM algorithm performs as:

- *E-step.* For each missing pattern, sweep operates on positions related to observed variables $O(s)$, to transform θ^t to ϕ^t. Then calculate the expected value of sufficient statistics $\mathbb{E}(T \mid X_{\text{obs}}, \theta^t)$. We should pay attention here that in the formula of sufficient statistics

$$T = \begin{bmatrix} n & 1^T X \\ X^T 1 & X^T X \end{bmatrix} = \sum_{i \in I(s)} \sum_{s=1}^{S} \begin{bmatrix} 1 & X_{i1} & X_{i2} & \cdots & X_{ip} \\ X_{i1} & X_{i1}^2 & X_{i1}X_{i2} & \cdots & X_{i1}X_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{ip} & X_{ip}X_{i1} & X_{ip}X_{i2} & \cdots & X_{ip}^2 \end{bmatrix}.$$

  the "sum" should first operate for $i \in I(s)$ then $s = 1, 2, \cdots, S$.

- *M-step.* $\theta^{t+1} = \text{SWP}[0] n^{-1} \mathbb{E}(T | X_{\text{obs}}, \theta^t)$.

Now we can implement EM in a general case of missing data for multivariate normal distribution as Algorithm 1.

Here the symbol ":=" indicates the operation of assignment. This implementation require two $(p+1) \times (p+1)$ matrix workspaces: T, into which the expected sufficient statistics are accumulated, and θ, which holds the current estimate of the parameter. For simplicity the rows and columns of these matrices are labeled from 0 to p.

Algorithm 1 Single iteration of EM for multivariate normal data with missingness

Input: $T := T_{obs}$;
  for $s := 1, 2, \ldots, S$ **do**
    for $j := 12, \ldots, p$ **do**
      if $r_{sj} = 1$ and $\theta_{jj} > 0$ **then**
        $\theta := \mathsf{SWP}[j]\theta$
      else if $r_{sj} = 0$ and $\theta_{jj} < 0$ **then**
        $\theta := \mathsf{RSW}[j]\theta$
    for $i \in I(s)$ **do**
      for $j \in M(s)$ **do**
        $c_j := \theta_{0j}$
        for $k \in O(s)$ **do**
          $c_j := c_j + \theta_{kj} X_{ik}$
      for $j \in M(s)$ **do**
        $T_{0j} := T_{0j} + c_j$
        for $k \in O(s)$ **do**
          $T_{kj} := T_{kj} + c_j x_{ik}$
        for $k \in M(s)$ and $k \geq j$ **do**
          $T_{kj} := T_{kj} + \theta_{kj} + c_k c_j$
Output: $\theta := \mathsf{SWP}[0]n^{-1}T$.

In addition, a single vector of length p, denoted by $c = (c_1, c_2, \cdots, c_p)$ is needed as a temporary workspace to hold the values of $X^*_{ij} = a_{0j} + \sum_{k \in O(s)} a_{kj} X_{ik}$, where a_{jk} denote the (j, k) th element of matrix $A = \mathsf{SWP}[O(s)]\theta$.

The iteration begins by setting T equal to T_{obs} which we assume has already been computed. The expectations of X_{ij} and $X_{ij}X_{ik}$ that contribute to T_{mis} are then calculated and added into T, one missingness pattern at a time. In order to calculate these expectations within a missingness pattern s, the θ matrix must be put into the required $\mathsf{SWP}[O(s)]$ condition; for this, we use the convenient book-keeping device that a diagonal element θ_{jj} is negative if and only if θ has been swept on position j. Finally after the expected sufficient statistics are fully accumulated into T, the new parameter estimate is calculated and stored in θ in preparation for the next iteration.

The sweep operator simplifies the notation of EM algorithm for incomplete multivariate normal data and also helps for its implementation. Furthermore, based on sweep operator, the implementation of linear regression with missing values via EM becomes also simplified as presented in Section 1.2.2 and the corresponding implementations as an R package are introduced in Chapter 5.

Chapter 2

Logistic regression with missing covariates

Contents

2.1 Introduction

In this chapter, we focus on the inference problem for logistic regression model with missing values. In this case, classical EM algorithm as introduced in Section 1.2.2 often involves unfeasible computations. In the framework of generalized linear models, Ibrahim et al. (1999, 2005) have suggested using a Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990; McLachlan and Krishnan, 2008), replacing the integral by its empirical sum using Monte Carlo sampling. Ibrahim et al. (1999) also estimated the variance using a Monte Carlo version of Louis' formula, involving Gibbs sampling with an adaptive rejection sampling scheme (Gilks and Wild, 1992). However, their approach has a high computational cost and was only implemented for monotone patterns of missing values and for missing values in only two variables in a dataset. In this chapter, we develop a stochastic approximation version of the EM algorithm (SAEM) (Lavielle, 2014), based on Metropolis-Hastings sampling, to perform statistical inference for logistic regression with incomplete data, where the missing data is found anywhere in the covariates. SAEM uses a stochastic approximation procedure to estimate the conditional expectation of the complete-data likelihood, instead of generating a large number of Monte Carlo samples, which lead to an undeniable computational advantage over MCEM, which we illustrate in the simulations. Note that another solution could be to use a Laplace approximation to compute integrals; this linearizes the likelihood function via differentiation, whereas SAEM supports likelihood-based inference without the intermediate step.

In addition, SAEM allows for model selection using a criterion based on a penalized version of the observed-data likelihood. This is very useful in practice, as there is few available likelihood based methods that can be applied to data with missing values in many columns, due to computational costs and the difficulty of implementation. Claeskens and Consentino (2008); Consentino and Claeskens (2011) suggested an approximation of AIC, Jiang et al. (2015) defined generalized information criteria, while Liu et al. (2016) proposed combining penalized regression techniques with multiple imputation and stability selection. Chow (1979); Fung and Wrobel (1989) also studied the linear discriminant function for logistic regression, using pairs of observed values in different columns to calculate the covariance matrix.

This chapter proceeds as follows: Section 2.2 provided the assumptions and notation used throughout the chapter. In Section 2.3, we derive a SAEM algorithm to obtain the maximum likelihood estimate of parameters in a logistic regression model for continuous covariate data, under the MAR mechanism of missing data. Following parameter estimation, we show how to estimate the Fisher information matrix using a Monte Carlo version of Louis' formula. Section 2.4 describes the model selection scheme, based on a Bayesian information criterion (BIC) with missing values. In addition, we propose an approach to perform prediction for a new observation that includes missing values. Section 2.5 presents a simulation study where the

proposed approach is compared to methods such as multiple imputation (Rubin, 2009) with respect to bias, coverage, and execution time. In Section 2.6, we apply the newly developed approach to predict the occurrence of hemorrhagic shock in patients with blunt trauma in the TraumaBase dataset, where it is crucial to efficiently manage missing data because the percentage of it varies from 0 to 60% depending on the variable. Predictions made using SAEM show an improvement over those made by emergency doctors. Lastly, Section 2.7 discusses the results and provides conclusions.

Our contribution is to give users the ability to perform logistic regression with missing values within a joint-modeling framework, one that combines computational efficiency and a sound theoretical foundation. The methods presented in this chapter are implemented as an R package `misaem`, available on CRAN, which we introduce later in Chapter 5.

2.2 Assumptions and notation

We first introduce the basic notation and assumptions that we use throughout this chapter. The logistic regression model for binary classification can be written as:

$$\mathbb{P}\left(y_i = 1|X_i;\beta\right) = \frac{\exp(\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})}{1+\exp(\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})}, \quad i = 1,\ldots,n, \tag{2.1}$$

where $y = (y_i, 1 \leq i \leq n)$ an n-vector of binary responses coded as $\{0,1\}$. We adopt a probabilistic framework by assuming that $X_i = (X_{i1},\ldots,X_{ip})$ is normally distributed:

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

Let $\theta = (\mu,\Sigma,\beta)$ be the set of parameters of the model. Then, the log-likelihood for the complete data can be written as:

$$\begin{aligned}\ell(\theta;X,y) &= \sum_{i=1}^{n}\ell(\theta;X_i,y_i)\\ &= \sum_{i=1}^{n}\Big(\log(\mathrm{p}(y_i|X_i;\beta)) + \log(\mathrm{p}(X_i;\mu,\Sigma))\Big).\end{aligned}$$

Our main goal is to estimate the vector of parameters $\beta = (\beta_j, 0 \leq j \leq p)$ when missing values exist in the design matrix, i.e., in the matrix X.

Missing values are assumed to be MAR as defined in Chapter 1, which allows to derive MLE by ignoring the missing values mechanism and maximizing the observed-data likelihood $\ell(\theta;y,X_{\text{obs}})$.

2.3 Parameter estimation by SAEM

2.3.1 The EM and MCEM algorithms

We aim to estimate the parameter θ of the logistic regression model by maximizing the observed log-likelihood $\ell(\theta;X_{\text{obs}},y)$. Let us start with the classical EM formulation for ob-

taining the maximum likelihood estimator from incomplete data. Given some initial value θ_0, iteration k updates θ_{k-1} to θ_k with the following two steps:

- **E-step:** Evaluate the quantity

$$\begin{aligned} Q_k(\theta) &= \mathbb{E}[\ell(\theta; X, y)|X_{\text{obs}}, y; \theta_{k-1}] \\ &= \int \ell(\theta; X, y)\mathrm{p}(X_{\text{mis}}|X_{\text{obs}}, y; \theta_{k-1})dX_{\text{mis}}. \end{aligned} \tag{2.2}$$

- **M-step:** Update the estimation of θ: $\theta_k = \arg\max_\theta Q_k(\theta)$.

Since the expectation (2.2) in the E-step for the logistic regression model has no explicit expression, MCEM (Wei and Tanner, 1990; Ibrahim et al., 1999) can be used. The E-step of MCEM generates several samples of missing data from the target distribution $\mathrm{p}(X_{\text{mis}}|X_{\text{obs}}, y; \theta_{k-1})$ and replaces the expectation of the complete log-likelihood by an empirical mean. However, an accurate Monte Carlo approximation of the E-step may require a significant computational effort, as illustrated in Section 2.5.

2.3.2 The SAEM algorithm

To achieve improved computational efficiency, we can derive a SAEM algorithm (Lavielle, 2014) which replaces the E-step (2.2) by a stochastic approximation. Starting from an initial guess θ_0, the kth iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \ldots, n$, draw $X_{i,\text{mis}}^{(k)}$ from

$$\mathrm{p}(X_{i,\text{mis}}|X_{i,\text{obs}}, y_i; \theta_{k-1}). \tag{2.3}$$

- **Stochastic approximation:** Update the function Q according to

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\ell(\theta; X_{\text{obs}}, X_{\text{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right), \tag{2.4}$$

  where (γ_k) is a non-increasing sequence of positive numbers.

- **Maximization:** Update the estimation of θ:

$$\theta_k = \arg\max_\theta Q_k(\theta).$$

The choice of the sequence (γ_k) in (2.4) is important for ensuring the almost sure convergence of SAEM to a maximum of the observed likelihood (Delyon et al., 1999). We will see in Section 2.5 that, in our case, very good convergence is obtained using $\gamma_k = 1$ during the first iterations, followed by a sequence that decreases as $1/k$.

2.3.3 Metropolis-Hastings sampling

In the logistic regression case, the unobserved data cannot in general be drawn exactly from the conditional distribution (2.3), which depends on an integral that is not calculable in closed form. One solution is to use a Metropolis-Hastings (MH) algorithm, which consists of constructing a Markov chain that has the target distribution as its stationary distribution. The states of the chain after S iterations are then used as a sample from the target distribution. To define a proposal distribution for the MH algorithm, we observe that the target distribution (2.3) can be factorized as follows:

$$\mathrm{p}(X_{i,\mathrm{mis}}|X_{i,\mathrm{obs}}, y_i; \theta) \propto \mathrm{p}(y_i|X_i; \beta)\mathrm{p}(X_{i,\mathrm{mis}}|X_{i,\mathrm{obs}}; \mu, \Sigma).$$

We select the proposal distribution as the second term $\mathrm{p}(X_{i,\mathrm{mis}}|X_{i,\mathrm{obs}}, \mu, \Sigma)$, which is normally distributed:

$$X_{i,\mathrm{mis}}|X_{i,\mathrm{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i), \tag{2.5}$$

where

$$\begin{aligned}\mu_i &= \mu_{i,\mathrm{mis}} + \Sigma_{i,\mathrm{mis,obs}}\Sigma^{-1}_{i,\mathrm{obs,obs}}(X_{i,\mathrm{obs}} - \mu_{i,\mathrm{obs}}),\\ \Sigma_i &= \Sigma_{i,\mathrm{mis,mis}} - \Sigma_{i,\mathrm{mis,obs}}\Sigma^{-1}_{i,\mathrm{obs,obs}}\Sigma_{i,\mathrm{obs,mis}},\end{aligned}$$

with $\mu_{i,\mathrm{mis}}$ (resp. $\mu_{i,\mathrm{obs}}$) the missing (resp. observed) elements of μ for individual i. The covariance matrix Σ is decomposed in the same way. The MH algorithm is described further in Section 2.8.1.

2.3.4 Observed Fisher information

After computing the MLE $\hat{\theta}_{\mathrm{ML}}$ with SAEM, we estimate its variance. To do so, we can use the observed Fisher information matrix (FIM): $\mathcal{I}(\theta) = -\frac{\partial^2 \ell(\theta; X_{\mathrm{obs}}, y)}{\partial\theta\partial\theta^T}$. According to Louis' formula (Louis, 1982), we have:

$$\begin{aligned}\mathcal{I}(\theta) = &- \mathbb{E}\left(\frac{\partial^2 \ell(\theta; X, y)}{\partial\theta\partial\theta^T}\Big|X_{\mathrm{obs}}, y; \theta\right)\\ &- \mathbb{E}\left(\frac{\partial \ell(\theta; X, y)}{\partial\theta}\frac{\partial \ell(\theta; X, y)^T}{\partial\theta}\Big|X_{\mathrm{obs}}, y; \theta\right)\\ &+ \mathbb{E}\left(\frac{\partial \ell(\theta; X, y)}{\partial\theta}|X_{\mathrm{obs}}, y; \theta\right)\mathbb{E}\left(\frac{\partial \ell(\theta; X, y)}{\partial\theta}|X_{\mathrm{obs}}, y; \theta\right)^T.\end{aligned}$$

The observed FIM can therefore be expressed in terms of conditional expectations, which can also be approximated using a Monte Carlo procedure. More precisely, given S samples $(X^{(s)}_{i,\mathrm{mis}}, 1 \leq i \leq n, 1 \leq s \leq S)$ of the missing data drawn from the conditional distribution

(2.3), the observed FIM can be estimated as $\hat{\mathcal{I}}_S(\hat{\theta}) = \sum_{i=1}^{n} -(D_i + G_i - \Delta_i \Delta_i^T)$, where

$$\Delta_i = \frac{1}{S}\sum_{s=1}^{S} \frac{\partial \ell(\hat{\theta}; X_{i,\text{mis}}^{(s)}, X_{i,\text{obs}}, y_i)}{\partial \theta},$$
$$D_i = \frac{1}{S}\sum_{s=1}^{S} \frac{\partial^2 \ell(\hat{\theta}; X_{i,\text{mis}}^{(s)}, X_{i,\text{obs}}, y_i)}{\partial \theta \partial \theta^T},$$
$$G_i = \frac{1}{S}\sum_{s=1}^{S} \left(\frac{\partial \ell(\hat{\theta}; X_{i,\text{mis}}^{(s)}, X_{i,\text{obs}}, y_i)}{\partial \theta}\right)\left(\frac{\partial \ell(\hat{\theta}; X_{i,\text{mis}}^{(s)}, X_{i,\text{obs}}, y_i)}{\partial \theta}\right)^T.$$

Here, the gradient and the Hessian matrix can be computed in closed form. The procedure for calculating the observed information matrix is described in Section 2.8.2.

2.4 Model selection with likelihood criteria and prediction

2.4.1 Information criteria

In order to compare different possible covariate models, we can consider penalized likelihood criteria such as BIC. For a given model $\mathcal{M}$ and an estimated parameter $\hat{\theta}_{\mathcal{M}}$, the BIC is defined as:

$$\text{BIC}(\mathcal{M}) = -2\ell(\hat{\theta}_{\mathcal{M}}; X_{\text{obs}}, y) + \log(n) d(\mathcal{M}),$$

where $d(\mathcal{M})$ is the number of estimated parameters in a model $\mathcal{M}$. The distribution of the complete set of covariates $(x_{ij}, 1 \leq i \leq n, 1 \leq j \leq p)$ does not depend on the regression model used for modeling the binary outcomes $(y_i, 1 \leq i \leq n)$; we assume the same normal distribution $\mathcal{N}_p(\mu, \Sigma)$ for all regression models. Thus, the difference between the numbers $d(\mathcal{M})$ of estimated parameters in two models is equivalent to the difference between the numbers of their non-zero coefficients in β. Note that, unlike the approach we suggest, existing methods (Claeskens and Consentino, 2008; Consentino and Claeskens, 2011) use an approximation of the Akaike information criterion (AIC) without estimating the observed likelihood.

2.4.2 Observed log-likelihood

For a given model and parameter θ, the observed log-likelihood is, by definition:

$$\ell(\theta; X_{\text{obs}}, y) = \sum_{i=1}^{n} \log\left(\text{p}(y_i, X_{i,\text{obs}}; \theta)\right).$$

With missing data, the density $\text{p}(y_i, X_{i,\text{obs}}; \theta)$ cannot in general be computed in closed form. We suggest approximating it using an importance sampling Monte Carlo approach. Let g_i be

the density function of the normal distribution defined in (2.5). Then,

$$\begin{aligned}
\mathrm{p}(y_i, X_{i,\mathrm{obs}}; \theta) &= \int \mathrm{p}(y_i, X_{i,\mathrm{obs}} | X_{i,\mathrm{mis}}; \theta) \mathrm{p}(X_{i,\mathrm{mis}}; \theta) dX_{i,\mathrm{mis}} \\
&= \int \mathrm{p}(y_i, X_{i,\mathrm{obs}} | X_{i,\mathrm{mis}}; \theta) \frac{\mathrm{p}(X_{i,\mathrm{mis}}; \theta)}{g_i(X_{i,\mathrm{mis}})} g_i(X_{i,\mathrm{mis}}) dX_{i,\mathrm{mis}} \\
&= \mathbb{E}_{g_i} \left(\mathrm{p}(y_i, X_{i,\mathrm{obs}} | X_{i,\mathrm{mis}}; \theta) \frac{\mathrm{p}(X_{i,\mathrm{mis}}; \theta)}{g_i(X_{i,\mathrm{mis}})} \right).
\end{aligned}$$

Consequently, if we draw M samples from the proposal distribution (2.5):

$$X_{i,\mathrm{mis}}^{(s)} \underset{i.i.d.}{\sim} \mathcal{N}(\mu_i, \Sigma_i), \quad m = 1, 2, \ldots, S,$$

we can estimate $\mathrm{p}(y_i, X_{i,\mathrm{obs}}; \theta)$ by:

$$\hat{\mathrm{p}}(y_i, X_{i,\mathrm{obs}}; \theta) = \frac{1}{S} \sum_{m=1}^{S} \mathrm{p}(y_i, X_{i,\mathrm{obs}} | X_{i,\mathrm{mis}}^{(s)}; \theta) \frac{\mathrm{p}(X_{i,\mathrm{mis}}^{(s)}; \theta)}{g_i(X_{i,\mathrm{mis}}^{(s)})},$$

and derive an estimate of the observed log-likelihood $\ell(\theta; X_{\mathrm{obs}}, y)$.

2.4.3 Prediction on a test set with missing values

In supervised learning, after fitting a model using a training set, a natural step is to evaluate the prediction performance, which can be done with a test set. Assuming $\tilde{X} = (\tilde{x}_{\mathrm{obs}}, \tilde{x}_{\mathrm{mis}})$ is an observation in the test set, we want to predict the binary response $\tilde{y}$. One important point is that test set also contains missing values, since the training set and the test set have the same distribution (i.e., the distribution of covariates and the distribution of missingness). Therefore, we cannot directly apply the fitted model (which uses p coefficients) to predict $\tilde{y}$ from an incomplete observation of the test set $\tilde{X}$.

Our framework offers a natural way to tackle this issue by marginalizing out missing covariates given the observed data. More precisely, with S Monte Carlo samples

$$(\tilde{x}_{\mathrm{mis}}^{(s)}, 1 \leq s \leq S) \sim \mathrm{p}(\tilde{x}_{\mathrm{mis}} \mid \tilde{x}_{\mathrm{obs}}),$$

we estimate directly the response by maximizing its distribution marginalized over missing data given the observed ones:

$$\begin{aligned}
\hat{y} = \arg\max_{\tilde{y}} \mathrm{p}(y \mid \tilde{x}_{\mathrm{obs}}) &= \arg\max_{\tilde{y}} \int \mathrm{p}(y \mid \tilde{X}) \mathrm{p}(\tilde{x}_{\mathrm{mis}} \mid \tilde{x}_{\mathrm{obs}}) \, d\tilde{x}_{\mathrm{mis}} \\
&= \arg\max_{\tilde{y}} \mathbb{E}_{\mathrm{p}_{\tilde{x}_{\mathrm{mis}} | \tilde{x}_{\mathrm{obs}}}} \mathrm{p}(\tilde{y} \mid \tilde{X}) \\
&= \arg\max_{\tilde{y}} \sum_{s=1}^{S} \mathrm{p}\left(\tilde{y} \mid \tilde{x}_{\mathrm{obs}}, \tilde{x}_{\mathrm{mis}}^{(s)}\right).
\end{aligned}$$

Note that in the literature there are very few solutions for dealing with missing values in a test set. In Section 2.6.2, we compare the proposed approach with other methods used in practice, which are based on imputation of the test set.

2.5 Simulation study: estimation bias and variance

2.5.1 Simulation settings

We first generated a design matrix X of size $n = 1000 \times p = 5$ by drawing each observation from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. Then, we generated the response according to the logistic regression model (2.1). We considered as the true parameter values: $\beta = (-0.2, 0.5, -0.3, 1, 0, -0.6)$, $\mu = (1, 2, 3, 4, 5)$, and $\Sigma = \text{diag}(\sigma)C\text{diag}(\sigma)$, where σ is the vector of standard deviations $\sigma = (1, 2, 3, 4, 5)$, and C the correlation matrix

$$C = \begin{bmatrix} 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.3 & 0.6 \\ 0 & 0 & 0.3 & 1 & 0.7 \\ 0 & 0 & 0.6 & 0.7 & 1 \end{bmatrix}. \tag{2.6}$$

Before generating missing values, we performed classical logistic regression on the complete dataset, the results (ROC curve) are provided in Section 2.8.3. We then randomly introduced 10% missing values in the covariates, initially with a missing-completely-at-random (MCAR) mechanism, where each entry has the same probability of being observed.

2.5.2 The behavior of SAEM

The algorithm was initialized with the parameters obtained after mean imputation, i.e., where missing values of a given variable were replaced by the unconditional mean calculated on the available cases, and then logistic regression was applied to the completed data. For the non-increasing sequence (γ_k) in the stochastic approximation step of SAEM, we chose $\gamma_k = 1$

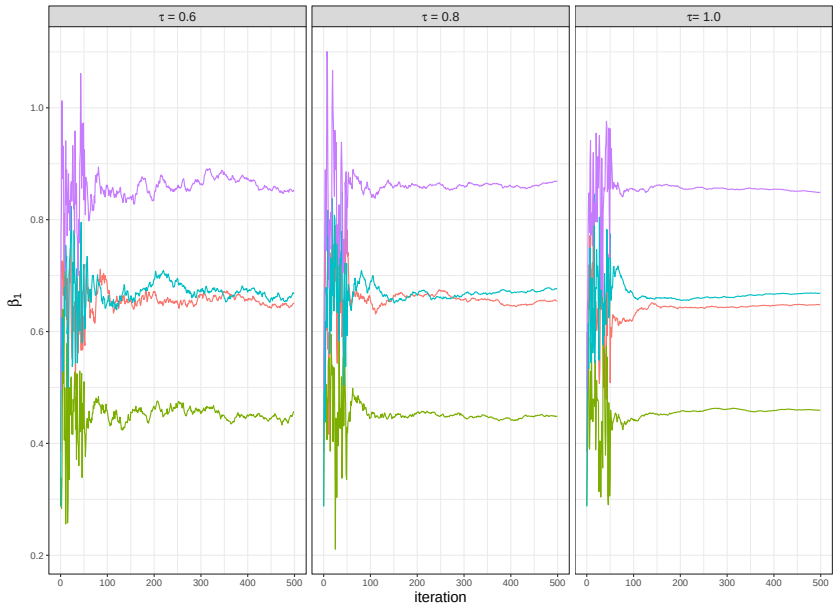


Figure 2.1: Convergence plots for β_1 obtained with three different values of τ (0.6, 0.8, 1.0). Each color represents one simulation. The true value of β_1 is 0.5.

during the first k_1 iterations in order to converge quickly to a neighborhood of the MLE, and from k_1 iterations on, we set $\gamma_k = (k - k_1)^{-\tau}$ to ensure the almost sure convergence of SAEM. In order to study the effect of the sequence of stepsizes (γ_k), we fixed the value of $k_1 = 50$ and used $\tau = (0.6,\ 0.8,\ 1)$ during the next 450 iterations. Representative plots of the convergence of SAEM for the coefficient β_1, obtained from four simulated data sets, are shown in Figure 2.1. For each given simulation, the three sequences of estimates converged to the same solution, but for larger τ, SAEM converged faster and fluctuated less. We therefore use $\tau = 1$ in the following.

2.5.3 Comparison with other methods

We ran 1000 simulations and compared SAEM to several other existing methods, initially in terms of estimation errors for the parameters. We mainly focused on *i)* the complete case (CC) method, i.e., all rows containing at least one unobserved data value were removed, and *ii)* multiple imputation by chained equations (*mice*) with Rubin's combining rules (van Buuren and Groothuis-Oudshoorn, 2011). More precisely, missing values are imputed successively by a series of regression models, where each variable with missing data is modeled conditional upon the other variables. For instance, linear regression is used to model continuous variables and binary variables are modeled using logistic regression. More details can be found in van

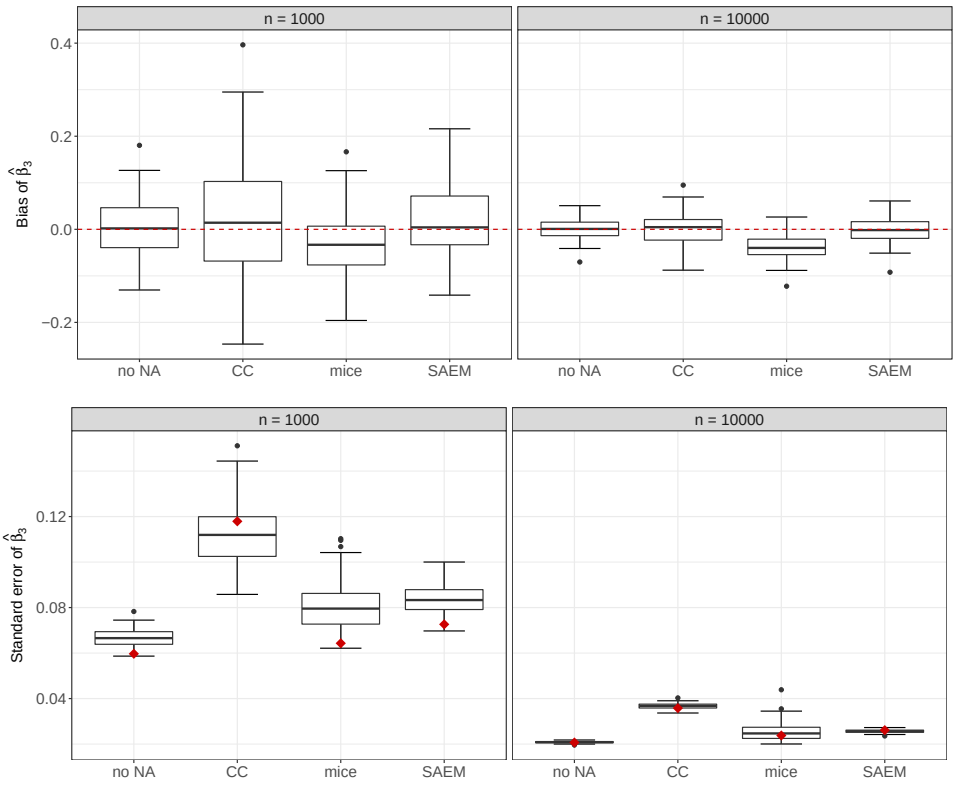


Figure 2.2: Top: Empirical distribution of the bias of $\hat{\beta}_3$. Bottom: Distribution of the estimated standard errors of $\hat{\beta}_3$. For each method, the red point corresponds to the empirical standard deviation of $\hat{\beta}_3$ calculated over the 1000 simulations. Results shown are for 10% MCAR and correlation C.

Buuren and Groothuis-Oudshoorn (2011). Finally, we used the dataset without missing values ("no NA") as a reference, with parameters estimated using the Newton-Raphson algorithm. We varied the number of observations: $n = 200, 1000, 10\,000$, the missing value mechanism: MCAR or MAR, the percentage of missing values: 10% or 30%, and the correlation structure, either using C given by (2.6), or an orthogonal design.

Figure 2.2 (top) displays the distribution of the estimates of β_3 for $n = 1000$ and $n = 10\,000$ under MCAR, with the correlation between covariates given by (2.6). Simulation results for $n = 200$ are presented in Section 2.8.8. This plot is representative of the results obtained with the other components of β. As expected, larger samples yielded less variability. Moreover, we observe that in both cases, the estimation obtained by *mice* could be biased, whereas SAEM provided unbiased estimates with small variances. Figure 2.2 (bottom) shows the empirical distribution of the estimated standard error of $\hat{\beta}_3$. For SAEM it was calculated using the observed Fisher information as described in Section 2.3.4. With larger n, not only the estimated standard errors–but also variance in the estimation–clearly decreased for all methods. In the case where $n = 1000$, SAEM and *mice* slightly overestimated the standard error, while CC underestimated it, on average. Globally, SAEM provided the best results; compared with *mice*, it gave a similar estimate of the standard error, on average, but with much less variance.

Table 2.1 shows the coverage probability of the confidence interval for all parameters and inside the parentheses is the average length of the corresponding confidence interval. We would expect coverage of 95%, corresponding to the nominal 95% level. The simulation margin of error for the coverage results is 1.35%. SAEM had between 94.3% and 95.4% coverage, while *mice* struggled for certain parameters: the coverage rates for two estimates were 89.6% and 86.5%, significantly below the nominal level. Even though CC showed reasonable results in terms of coverage, the widths of its confidence intervals were still too large. Simulations with smaller sample sizes gave similar results—see Section 2.8.8 for $n = 200$.

Table 2.1: Coverage (%) for $n = 10\,000$, correlation C and 10% MCAR, calculated over 1000 simulations. Bold indicates under-coverage. Inside the parentheses is the average length of corresponding confidence interval over 1000 simulations (multiplied by 100).

parameter	no NA	CC	mice	SAEM
β_0	95.2 (21.36)	94.4 (27.82)	95.2 (22.70)	94.9 (22.48)
β_1	96.0 (18.92)	94.7 (24.65)	93.9 (21.77)	95.1 (21.51)
β_2	95.5 (9.53)	94.6 (12.41)	94.0 (10.97)	94.3 (10.83)
β_3	94.9 (8.17)	94.3 (10.66)	**86.5** (9.03)	94.7 (9.03)
β_4	94.6 (4.00)	94.2 (5.21)	96.2 (4.49)	95.4 (4.42)
β_5	95.9 (5.52)	94.4 (7.19)	**89.6** (6.20)	94.7 (6.17)

Lastly, Table 2.2 highlights large differences between the methods in terms of execution time. As an aside, we also implemented the MCEM algorithm (Ibrahim et al., 1999) using adaptive rejection sampling; even with a very small sample size of $n = 200$, MCEM took 5 minutes per simulation on average. In contrast, multiple imputation took less than 1 second per simulation, and SAEM less than 10 seconds, which remains reasonable. However, the bias and standard errors for the SAEM and MCEM estimates were quite similar—see Section

2.8.8. Due to the prohibitive execution time required, for larger sample sizes we did not compare MCEM with the other methods.

Table 2.2: Comparison of execution times between no NA, MCEM, *mice*, and SAEM with correlation C and 10% MCAR, for $n = 200$ and $n = 1000$, calculated over 1000 simulations.

Execution time (seconds) for one simulation	no NA	MCEM	mice	SAEM
$\boldsymbol{n = 1000}$				
min	2.87×10^{-3}	492	0.64	9.96
mean	4.65×10^{-3}	773	0.70	13.50
max	43.50×10^{-3}	1077	0.76	16.79
$\boldsymbol{n = 200}$				
min	1.26×10^{-3}	67.91	0.24	2.64
mean	2.32×10^{-3}	291.47	0.28	3.91
max	21.53×10^{-3}	1003	0.48	6.04

Results obtained for independent covariates are presented in Figure 2.3 (right), for estimation in the orthogonal design case. SAEM was slightly biased since it estimated non-zero terms for the covariance, but still outperformed CC and *mice*.

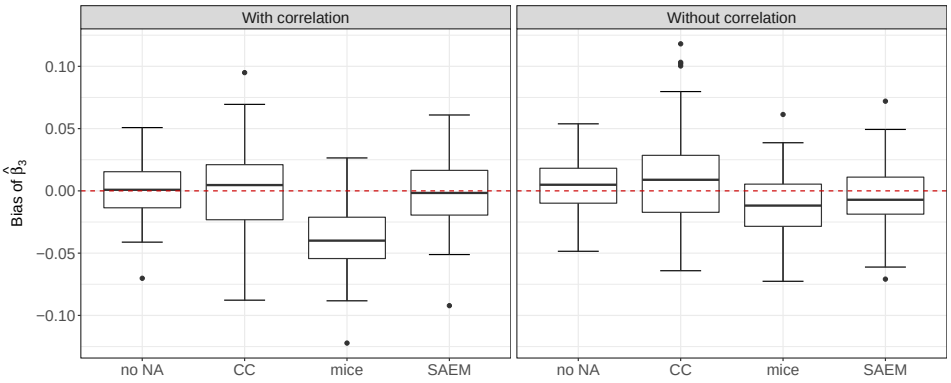


Figure 2.3: Empirical distribution of the estimates of β_3 obtained under MCAR, with $n = 10\,000$ and 10% missing values. Left: the covariates are correlated; right: no correlation between covariates.

2.5.4 Extended simulations

Missing-at-random mechanisms We first simulated the pattern of missingness as a binary vector $\eta = (\eta_1, \eta_2, \ldots, \eta_p)$ from the Bernoulli distribution, where $\eta_j = 0$ indicates that the corresponding variable x_j can be missing while $\eta_j = 1$ indicates it is always observed. Then the probability of having missing data in one variable is calculated by a logistic regression model. For example in our case with the realizations of $\eta = (1, 0, 1, 0, 0)$, the probability that covariates (x_2, x_4, x_5) are missing is calculated by a logistic regression model conditional on x_1 and x_3. The weights in the linear combination of x_1 and x_3 have an effect on the

proportion of missingness. We introduced 10% missing values into the covariates using an MAR mechanism. The results presented in Section 2.8.4 are—as expected—similar to those obtained under MCAR, and show that the parameters are estimated without bias.

Robustness to the normal assumption for covariates First we generated a design matrix of size $n = 1000 \times p = 5$ by drawing each observation from a multivariate Student distribution $t_v(\mu, \Sigma)$ with $v = 5$ or $v = 20$ degrees of freedom, and (μ, Σ) the same as those in the normal distribution in Section 2.5.1. Then, we considered the Gaussian mixture model case by generating half of the samples from $\mathcal{N}(\mu_1, \Sigma)$ and the other half from $\mathcal{N}(\mu_2, \Sigma)$, where $\mu_1 = (1, 2, 3, 4, 5)$ and $\mu_2 = (1, 1, 1, 1, 1)$, with the same Σ as previously. Then, we generated the response according to the same logistic regression model as in Section 2.5.1, and considered either MCAR or MAR mechanisms.

Figure 2.4 illustrates the estimation bias of the parameter β_3, and Section 2.8.5 shows the coverage for all parameters, with the average length of the corresponding confidence interval in parentheses. This experiment shows that the estimation bias for regression coefficients with the proposed method—even based on normal assumptions—is robust to such model misspecification. Indeed, the bias may increase when covariates do not exactly follow a normal distribution, but the increase is negligible compared to the bias of imputation-based methods. We also observe only a small level of under-coverage as compared to *mice*, and a more reasonable length of confidence interval as compared to CC.

Varying the percentage of missing values When the percentage of missing values increases, variability in the results increases, but the suggested method still provide satisfactory results—see Section 2.8.6.

Varying the separability of the classes When the classes are well-separated, SAEM can exhibit bias and large variance, as illustrated in Section 2.8.7. However, logistic regression without missing values also encounters difficulties.

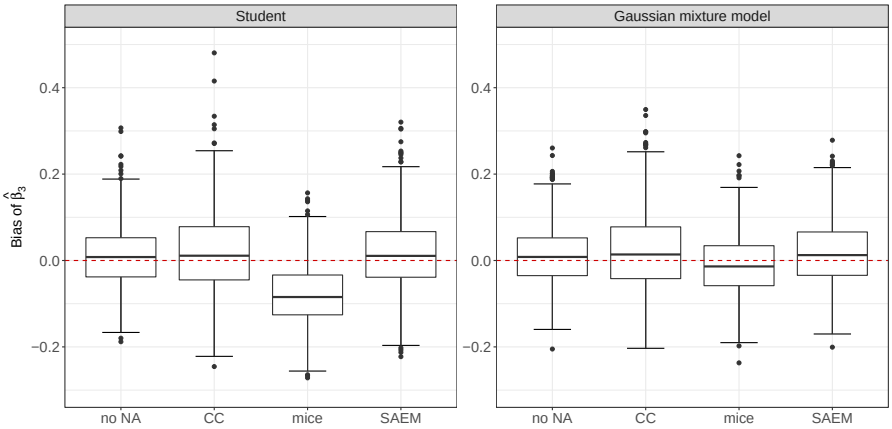


Figure 2.4: Empirical distribution of the bias of $\hat{\beta}_3$ obtained for misspecified models under MCAR, with $n = 1000$. Left: Student's distribution with $v = 5$ degrees of freedom; right: Gaussian mixture model.

In summary, not only did these simulations show that SAEM leads to estimators with limited bias, but also that we obtained accurate inference by taking into account the additional variance due to missing data.

2.5.5 Model selection

To look at the performance of the method in terms of model selection, we considered the same simulation scenarios as in Section 2.5.1, with some parameters set to zero. We now describe the results for the case where all parameters in β are zero except $\beta_0 = -0.2$, $\beta_1 = 0.5$, $\beta_3 = 1$, and $\beta_5 = -0.6$. We compared the BIC_{obs} based on the observed log-likelihood, as described in Section 2.4, to those based on the complete cases BIC_{cc} and obtained from the the original complete data BIC_{orig}.

Table 2.3 shows, with or without correlation between covariates, the percentage of cases where each criterion selects the true model (C), overfits (O)—i.e., selects more variables than there were, or underfits (U)—i.e., selects less variables than there were. In the case where the variables were correlated, the correlation matrix was the same as in Section 2.5.1. These results are representative of those obtained in the other simulation settings.

Table 2.3: For data with or without correlations, the percentage of times that each criterion selects the correct true model (C), overfits (O), or underfits (U).

	Non-Correlated			Correlated		
Criterion	C	O	U	C	O	U
BIC_{obs}	92	3	5	94	2	4
BIC_{orig}	96	2	2	93	0	7
BIC_{cc}	79	1	20	91	0	9

2.5.6 Predictions for a test set with missing values

To evaluate the prediction performance on a test set with missing values, we considered the same simulation scenarios for the training set as in Section 2.5.1 with sample size 1000×5. We also generated a test set of size 100×5. We compared the approach described in Section 2.4.3 with imputation methods. More precisely, we considered single imputation methods on the training set, followed by classical logistic regression and variable selection by BIC on the imputed dataset. The single imputation methods included *i)* imputation by the mean (impMean) *ii)* imputation by PCA (impPCA) (Josse and Husson, 2016), which is based on a low-rank assumption of the data matrix to impute. In addition, we considered multiple imputation using *mice*. Note that Hentges and Dunsmore (1998) showed in a simulation study that imputation methods for MCAR data can perform well when the aim of logistic regression is prediction. For all of the imputation methods, we also imputed the test set independently and then applied the model that had been selected on the training set. Note that this would be a limitation if there was only one individual in the test set, whereas our method does not encounter this issue.

We compared all of these approaches in terms of classical criteria to evaluate the predicted probabilities from the logistic regression. Criteria included AUC (area under the ROC curve),

the Brier score (Brier, 1950) and the logarithmic score (Good, 1952). Figure 2.5 shows that on average, marginalizing over the distribution of missing values gave the best performance: the largest AUCs and logarithmic scores, and the smallest Brier scores.

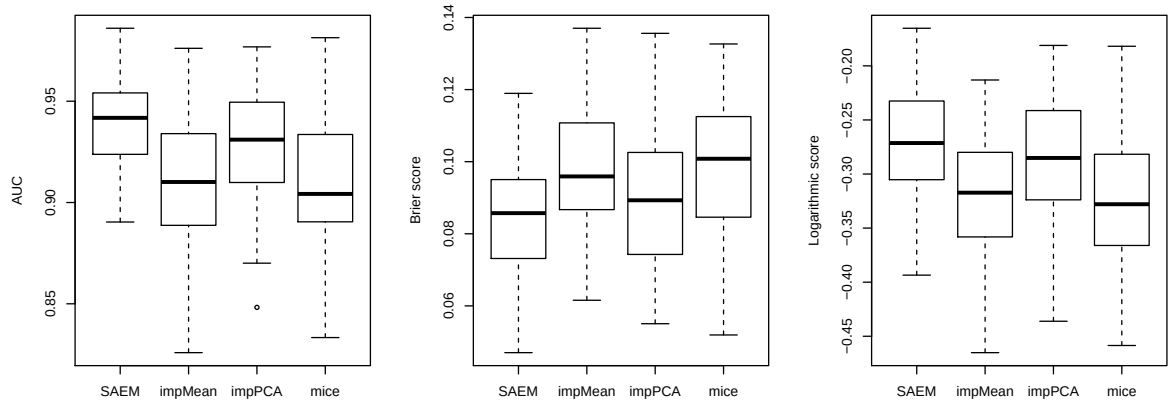


Figure 2.5: Comparisons of the empirical distribution of the AUC, Brier score, and logarithmic score obtained on the test set for the proposed SAEM without imputation method, impMean, impPCA, and *mice*, over 100 simulations.

2.6 Modeling the risk of severe hemorrhage in the TraumaBase context

In the analysis of medical dataset TraumaBase as introduced in Section 1.4, the fundamental goal of our work is to accelerate and simplify the detection of patients presenting with hemorrhagic shock due to blunt trauma in order to speed up management of this, the most preventable cause of death in major trauma cases. Optimized organization is essential to control blood loss as quickly as possible and reduce mortality.

2.6.1 Preprocessing of data

The TraumaBase group decided to focus on patients with blunt trauma so as to be able to compare results with existing prediction rules. Patients with pre-hospital cardiac arrest, penetrating trauma, and missing pre-hospital data, were excluded. This led to 5162 patients being retained in the data set. Based on clinical experience, 16 influential quantitative measurements were included. Detailed descriptions of these and their histograms are shown in Section 2.8.9. These variables were chosen because they were all available to the pre-hospital team, and therefore could be used in real situations.

There was strong collinearity between variables, as can be seen in the variables' PCA factor map (obtained by running an EM-PCA algorithm (Josse and Husson, 2016) which performs PCA with missing values) in Figure 2.6, in particular between the minimum systolic (PAS.min) and diastolic blood pressure (PAD.min). Based on expert advice, the re-

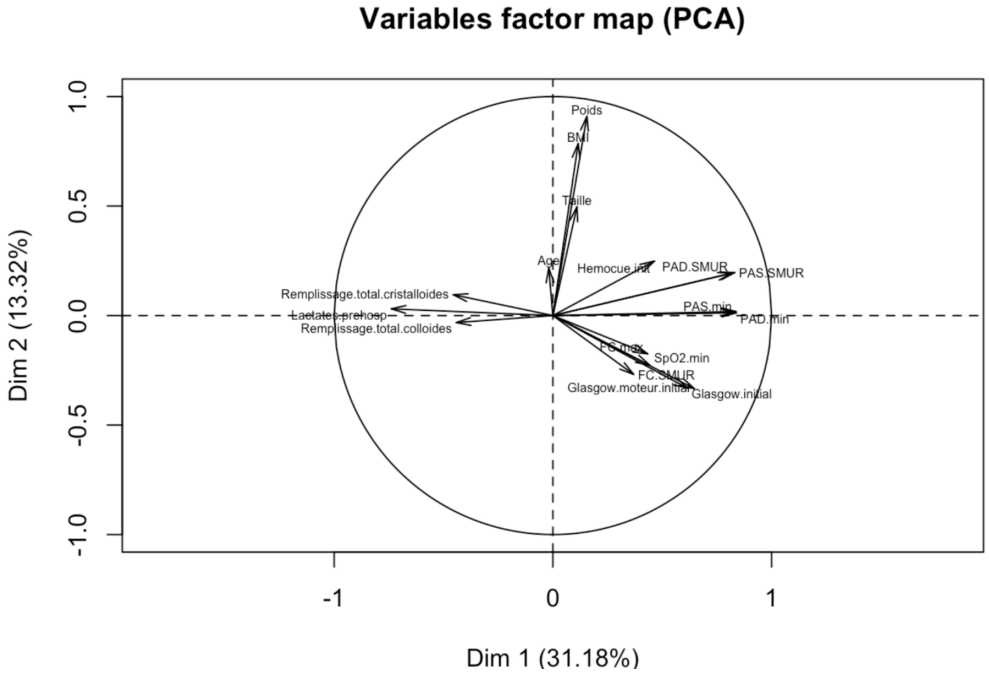


Figure 2.6: The factor map of the variables from PCA.

coded variables, SD.min and SD.SMUR (SD.min = PAS.min − PAD.min; SD.SMUR = PAS.SMUR − PAD.SMUR) were used since they have more clinical significance (Hamada et al., 2018). Thus, we had 14 variables to predict hemorrhagic shock.

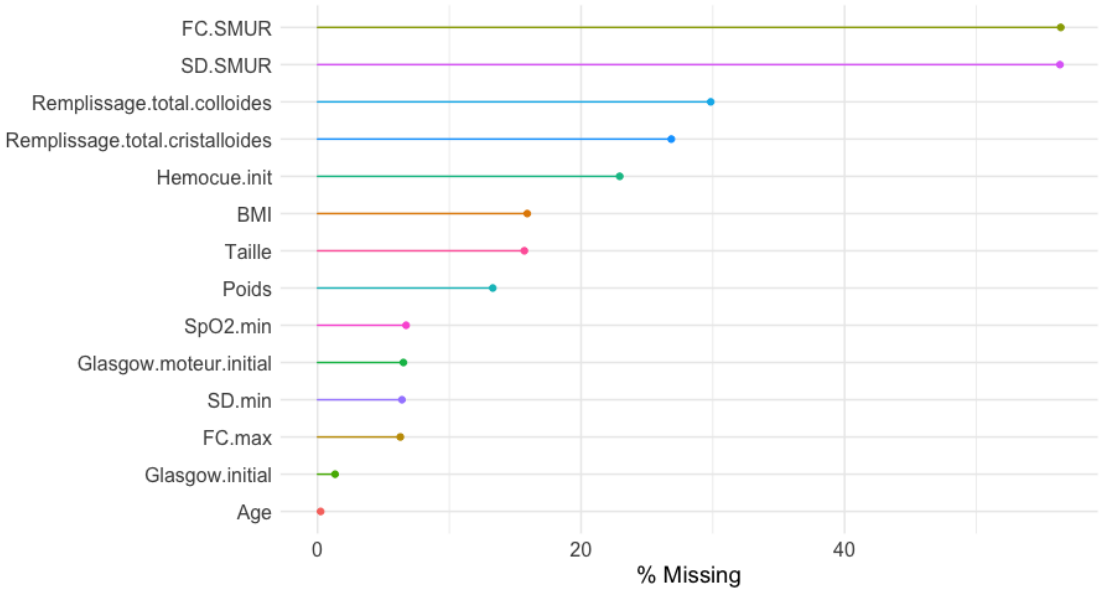


Figure 2.7: Percentage of missing values in each variable.

Figure 2.7 shows the percentage of missingness per variable, varying from 0 to 60%, which demonstrates the importance of taking appropriate account of missing data. Even though there may be many reasons why missingness occurred, overall considering them all to be MAR remains a plausible assumption. For instance, FC.SMUR (heart rate) and SD.SMUR

(the pulse pressure measured when the ambulance arrives at the accident site) contain many missing values because doctors collected these data during transportation. However, many other medical institutes and scientific publications use measurements on arrival at the accident scene. Consequently, doctors decided to record these as well, but this occurred after TraumaBase was set up.

We first applied SAEM for logistic regression with all 14 predictors and for the whole dataset. The estimation obtained by SAEM was broadly similar to that obtained by multiple imputation. Next, we used the model selection procedure described in Section 2.4. There were two observations that led to a very small value for the log-likelihood. Upon closer inspection, we found that for patient number 3302, the BMI was obtained using an incorrect calculation, and for patient number 1144, the weight (200 kg) and height (100 cm) values were likely to be incorrect. Hence, the observed log-likelihood also helped us to identify undetected outliers. In the observations' PCA factor map shown in Figure 2.8, patient number 3302 (circled in blue) is one of the outliers.

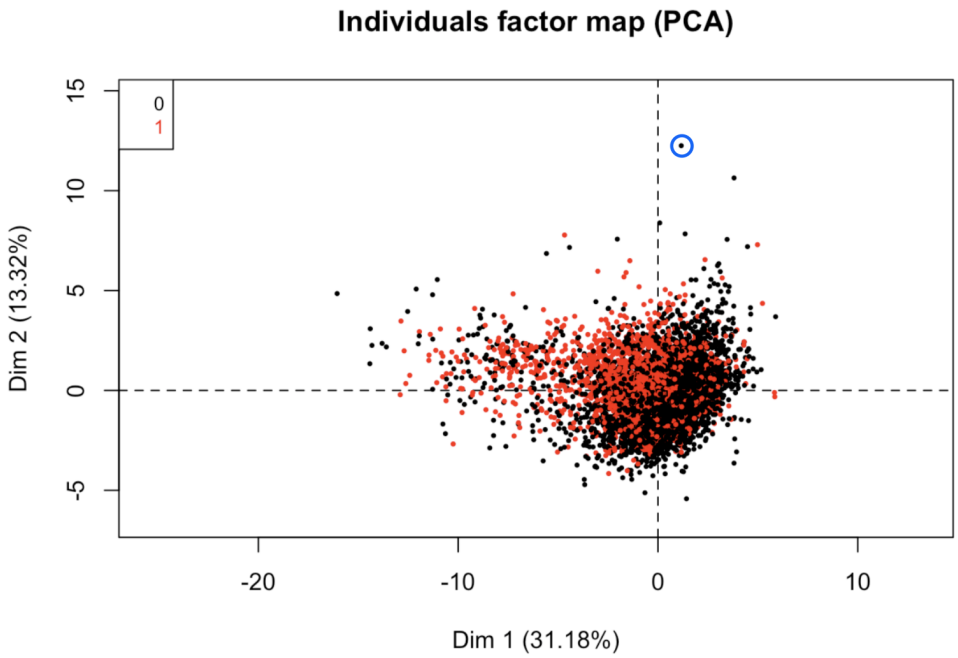


Figure 2.8: The observations' PCA factor map. Red points are hemorrhagic shock patients, and black points those who did not have hemorrhagic shock. Patient number 3302 (circled in blue) has an incorrectly-calculated BMI.

2.6.2 Predictive performance

We divided the dataset into training and test sets. The training set contained a random selection of 70% of the observations, and the test set contained the remaining 30%. In the training set, we selected a model with the approach suggested in Section 2.4, and used forward selection, resulting in a model with 8 variables. The estimates of parameters and their standard errors are shown in Table 2.4.

The TraumaBase medical team indicated to us that the signs of the coefficients were in

Table 2.4: Estimation of β and its standard errors obtained by SAEM, using BIC for model selection.

Variables	Estimate (standard errors)
$(Intercept)$	-0.12 (0.64)
Age	0.017 (0.0037)
$Glasgow.moteur$	-0.22 (0.040)
$FC.max$	0.024 (0.0028)
$Hemocue.init$	-0.26 (0.033)
$RT.cristalloides$	0.00088 (0.00011)
$RT.colloides$	0.0018 (0.00023)
$SD.min$	-0.027 (0.0055)
$SD.SMUR$	-0.018 (0.0061)

agreement with their prior intuition: all things being equal, *a)* Older people are more likely to have hemorrhagic shock; *b)* A low Glasgow score implies little or no motor response, which often is the case for hemorrhagic shock patients; *c)* A typical sign of hemorrhagic shock is rapid heart rate; *d)* The more a patient bleeds, the lower their hemoglobin is and more blood must be transfused. It is then more likely they will end up with hemorrhagic shock; *e)* Therapy involving two types of volume expanders, cristalloides and colloides, can be conducted to treat hemorrhagic shock; *f)* If an extremely low pulse pressure is observed, the cause may be a low stroke volume, which is usually the case in hemorrhagic shock.

Next, we assessed the prediction quality on the test set with the usual metrics based on the confusion matrix (false positive rate, false negative rate, etc.). We need to ensure that the cost of a false negative is much more than that of a false positive, as non-recognition of a potential hemorrhagic shock leads to a higher risk of patient mortality. With this in mind, we define the validation error on the test set as:

$$l(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} w_0 \mathbb{1}_{\{y_i=1, \hat{y}_i=0\}} + w_1 \mathbb{1}_{\{y_i=0, \hat{y}_i=1\}} \tag{2.7}$$

where w_0 and w_1 are user defined weight for the cost of a false negative and false positive respectively, with $w_0 + w_1 = 1$. In this way, we can choose a threshold for the logistic regression by giving values for w_0 and w_1. For instance, we chose $\frac{w_0}{w_1} = 5$, i.e., a false negative is five times more costly than a false positive. This cost function was chosen after discussions with experts. Note that the test set was also incomplete, so we used the strategy described in Section 2.4.3 to perform prediction. The confusion matrix of the predictive performance on the test set is shown in Table 2.5. The associated ROC curve is shown in Figure 2.9, which has an AUC of 0.88.

2.6.3 Comparison with other approaches

Next, we compared the proposed method to other approaches. Similar to in Section 2.6.2, we considered single imputation methods followed by classical logistic regression and variable selection on the imputed training dataset, such as single imputation by PCA (impPCA) (Josse and Husson, 2016), imputation by Random Forest (missForest) (Stekhoven and Buehlmann,

		Predicted outcome	
		1	0
Observed value	1	True Positive (102)	False Negative (46)
	0	False Positive (146)	True Negative (1254)

Table 2.5: Confusion matrix for predictions on the test set.

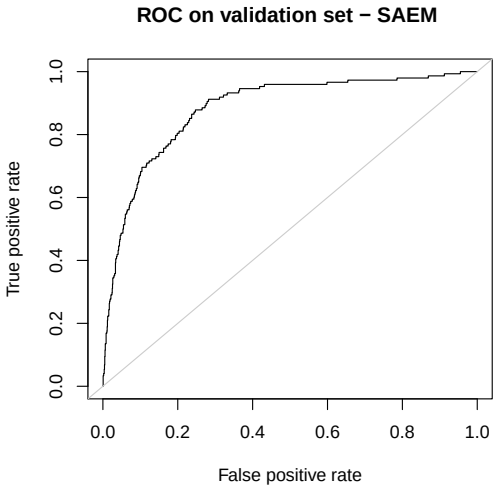


Figure 2.9: ROC curve of the test set predictions.

2012), and mean imputation (impMean). We also compared the logistic regression model with other prediction models such as Random Forest (predRF) and SVM (predSVM), both applied on the Random Forest-imputed (Stekhoven and Buehlmann, 2012) dataset. We also considered *mice*: we applied logistic regression with a classical forward selection method, with the BIC calculated on each imputed data set. However, note that there is no straightforward solution for combining multiple imputation and variable selection; we followed the empirical approach suggested in Wood et al. (2008) where they select variables that appear in at least half of the models selected in each imputed dataset.

We also considered three rules used by doctors to predict hemorrhagic shock: *i)* Doctors' prediction (doctor): the prediction recorded in TraumaBase. This showed whether the doctor considered the patient to be at risk of hemorrhagic shock; *ii)* The assessment of blood consumption score (ABC): this is an examination usually performed when the patient arrives at the trauma center. As such, the score is not exactly pre-hospital but can be computed very early in a hospitalization; *iii)* the trauma associated severe hemorrhage score (TASH): this score was also designed for hemorrhage detection, but at a later time-point since it uses some values that are only available after laboratory tests and radiography.

Figure 2.10 compares the methods in terms of their validation error (2.7). The splitting of data (into training and test sets) was repeated 15 times and we fixed the threshold such that the cost of a false negative was five times that of a false positive, i.e., $\frac{w_0}{w_1} = 5$. On average, SAEM performed well and with low variability between trials, while all of the imputation methods performed similarly to each other even naive mean imputation. In addition, other prediction methods (Random Forest and SVM) did not give smaller errors on the test sets than the logistic regression models. Lastly, the rules used by doctors, even those using more information than pre-hospital data, did not perform as well as SAEM. Section2.8.10 gives further details on classical criteria (AUC, sensitivity, specificity, accuracy and precision) to compare the predictive performance of the methods. The SAEM approach performed well on average, and particularly well for sensitivity, i.e., it rarely misdiagnosed hemorrhagic shock patients, which gels well with the clinical needs of emergency doctors.

More generally, without defining a specific threshold, we show in Figure 2.11 the average predictive loss over 15 replicates as a function of the cost ratio $\{\frac{w_0}{w_1} \mid \frac{w_0}{w_1} > 1\}$ for all methods. SAEM had a small error on the test sets given the value of $\frac{w_0}{w_1}$, especially when we increased

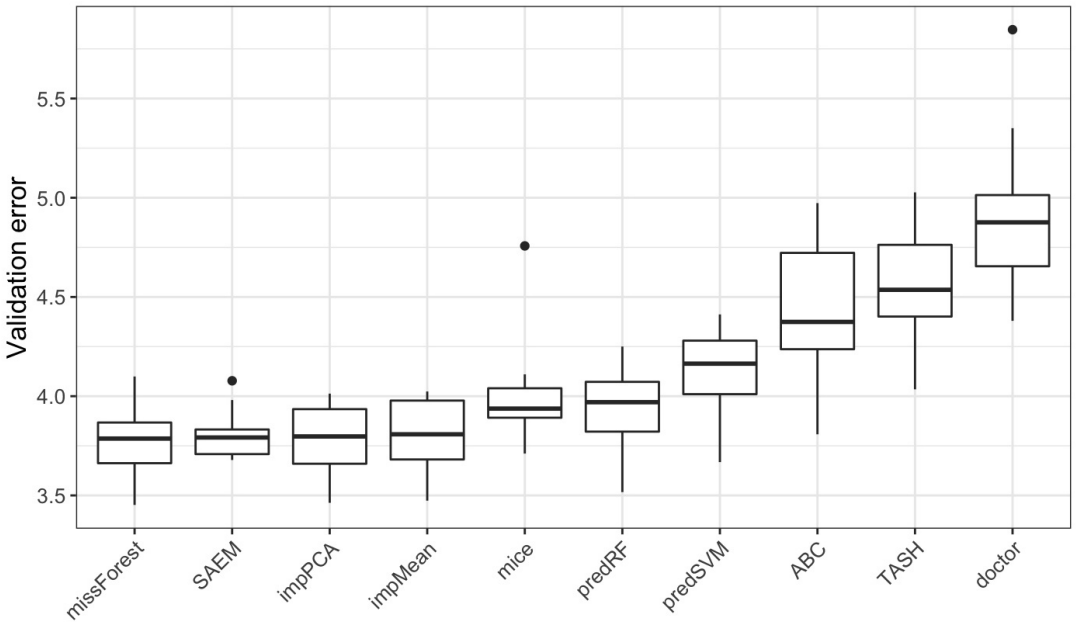


Figure 2.10: Empirical distribution of the prediction errors of different methods over 15 random splits of the TraumaBase data.

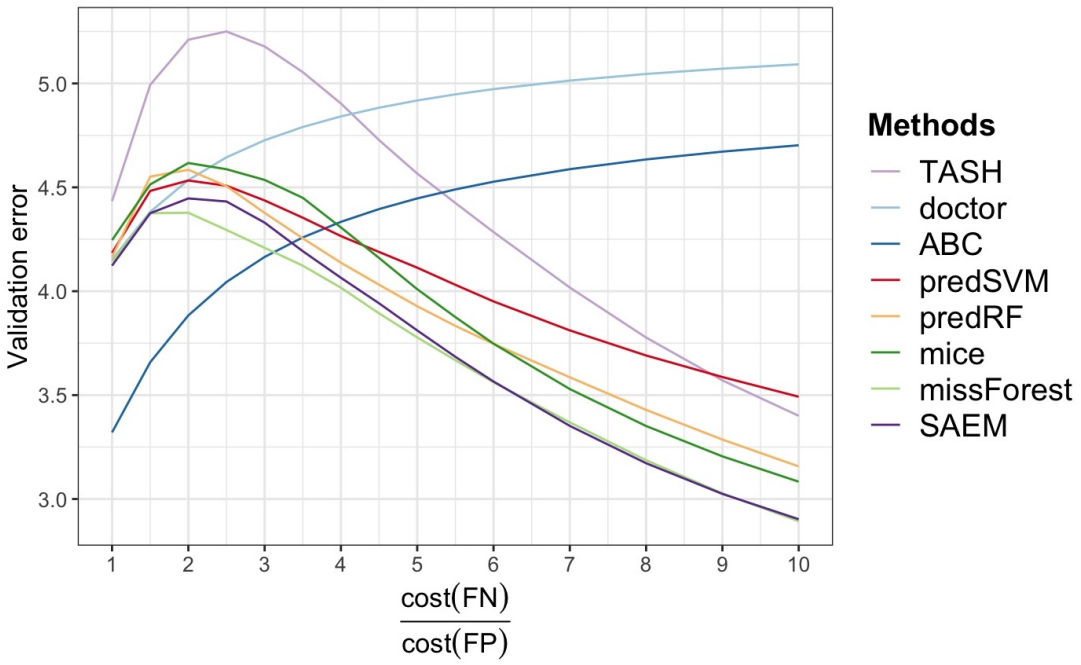


Figure 2.11: Average prediction errors of different methods as a function of the cost ratio $\{\frac{w_0}{w_1} \mid \frac{w_0}{w_1} > 1\}$ taken over 15 random splits of the TraumaBase data.

the cost of false negatives. Note that the errors for the doctors' rules and ABC increased as a function of the cost importance $\frac{w_0}{w_1}$, which means that these rules are more conservative than SAEM is, which may be problematic in this setting. Also, missForest had excellent predictive capabilities which is consistent with the results of Josse et al. (2019). However, it is difficult to interpret the results from random forest in terms of selected variables, which is often crucial for emergency doctors.

Note that even if our proposed methodology is based on the assumption of normally distributed covariates, its performance is better than the predictions made by widely-used medical criterion in terms of prediction error. Further discussion on the normal assumption is provided in Section 2.8.9. In addition, it should be noted that the proposed methodology can be extended to other assumptions about the joint distribution of covariates, such as a mixture of distributions.

In summary, based on the TraumaBase application and comparisons with other methods, we have demonstrated that this new approach has the ability to outperform existing popular methods that deal with missing data.

2.7 Discussion

In this chapter, we have developed a comprehensive joint-modeling framework for logistic regression with missing values. The method is implemented in the R package `misaem`, which we introduce later in Chapter 5. The experiments we have performed indicate that this method is computationally efficient and easy to implement. In addition, compared with multiple imputation—especially in the case of correlation between variables—estimation using SAEM is less biased than other methods and generally leads to interval-estimate coverage that is close to the nominal level. Based on our algorithm, model selection with BIC and missing data can be performed in a natural way. In view of the results reported in this article, we have been invited by emergency-room doctors in one of the TraumaBase centers to implement the missing-data methodology outlined here in a prospective study to evaluate its performance in a real-time clinical setting.

Paths for possible future research include further developing the method to handle both quantitative and categorical data. Since the data have have high inter-center variability, it is also important to take the hospital effect as an explanatory variable. However, modeling with mixed and incomplete data is challenging.

We have begun to explore initial ways of adapting our methodology in the mixed data framework for the logistic regression model. The approach is the following: use a general location model (GLOM) (Olkin et al., 1961) with simplification, then maximize the observed likelihood of the data using the SAEM algorithm. A detailed algorithm and the first results of the implementation are provided in Section 2.8.11. Nevertheless, the efficiency of the proposed algorithm still needs to be improved in order to apply it on more complex cases and on real data.

This chapter focused on making inference with missing values, but we have also suggested a method to predict from a test set with missing values. More work could be done in the direction of supervised learning with missing values, especially when we want to better estimate the variance of predictions. Extensions of the methods of Schafer and Schenker

(2000) could be considered. In addition, in the TraumaBase dataset, it would be reasonable to expect to have both MAR and missing-not-at-random (MNAR) values. MNAR means that missingness is related to the missing values themselves, and therefore a more correct methodology would require incorporating models for missing data mechanisms. As a final note, our proposed method may be quite useful in a causal inference framework, especially for propensity score analysis, which estimates the effect of a treatment, policy, or other intervention. Indeed, inverse probability weighting methods (IPW) are often performed with logistic regression, and the proposed method offers a potential solution for times where there are missing values in the covariates.

2.8 Supplementary materials

2.8.1 Metropolis-Hastings sampling

During SAEM iterations, Metropolis-Hastings sampling is performed as in Algorithm 2, with target distribution $f(X_{i,\text{mis}}) = \text{p}(X_{i,\text{mis}}|X_{i,\text{obs}}, y_i; \theta)$ and proposal distribution $g(X_{i,\text{mis}}) = \text{p}(X_{i,\text{mis}}|X_{i,\text{obs}}; \mu, \Sigma)$.

Algorithm 2 Metropolis-Hastings sampling.

Input: An initial sample $X_{i,\text{mis}}^{(0)} \sim g(X_{i,\text{mis}})$;
for $s = 1, 2, \ldots, S$ **do**
  Generate $X_{i,\text{mis}}^{(s)} \sim g(X_{i,\text{mis}})$;
  Generate $u \sim \mathcal{U}[0, 1]$;
  Calculate the ratio $w = \frac{f(X_{i,\text{mis}}^{(s)})/g(X_{i,\text{mis}}^{(s)})}{f(X_{i,\text{mis}}^{(s-1)})/g(X_{i,\text{mis}}^{(s-1)})}$;
  if $u < w$ **then**
    Accept $X_{i,\text{mis}}^{(s)}$;
  else
    $X_{i,\text{mis}}^{(s)} \leftarrow X_{i,\text{mis}}^{(s-1)}$;
Output: $(X_{i,\text{mis}}^{(s)}, 1 \leq i \leq n, 1 \leq s \leq S)$.

2.8.2 Calculation of the observed information matrix

Procedure 3 shows how we calculate the observed information matrix.

2.8.3 Logistic regression on a simulated complete dataset

Figure 2.12 shows the ROC curve on a simulated complete dataset. The corresponding AUC (for the training set) is 0.8976.

Procedure 3 Calculation of the observed information matrix.

Input: After drawing MH samples $(X_{i,\text{mis}}^{(s)}, 1 \leq i \leq n, 1 \leq s \leq S)$ for unobserved data $(X_{i,\text{mis}}, 1 \leq i \leq n)$, we have imputed observations, noted as $(Z_i^{(s)}, 1 \leq i \leq n, 1 \leq s \leq S)$, where $Z_{ij}^{(s)} = X_{i,\text{obs}}$, if x_{ij} is observed; else $Z_{ij}^{(s)} = X_{i,\text{mis}}^{(s)}$.

for $n = 1, 2, \ldots, n$ **do**

  for $s = 1, 2, \ldots, S$ **do**

    Calculate the gradient:

    $\nabla f_{is} = \frac{\partial \ell(\theta; X_{i,\text{obs}}, X_{i,\text{mis}}^{(s)}, y_i)}{\partial \beta} = Z_i^{(s)} \left(y_i - \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j Z_{ij}^{(s)})}{1 + \exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j Z_{ij}^{(s)})} \right)$;

    Calculate the Hessian matrix:

    $H_{is} = \frac{\partial^2 \ell(\theta; X_{i,\text{obs}}, X_{i,\text{mis}}^{(s)}, y_i)}{\partial \beta \partial \beta^T} = -Z_i^{(s)} Z_i^{(s)^T} \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j Z_{ij}^{(s)})}{\left(1 + \exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j Z_{ij}^{(s)})\right)^2}$;

    $\Delta_i \leftarrow \frac{1}{s}[(s-1)\Delta_i + \nabla f_{is}]$;

    $D_i \leftarrow \frac{1}{s}[(s-1)D_i + H_{is}]$;

    $G_i \leftarrow \frac{1}{s}[(s-1)G_i + \nabla f_{is} \nabla f_{is}^T]$;

  $\hat{\mathcal{I}}_S(\hat{\beta}) \leftarrow \hat{\mathcal{I}}_S(\hat{\beta}) - (D_i + G_i - \Delta_i \Delta_i^T)$;

Output: $\hat{\mathcal{I}}_S(\hat{\beta})$.

2.8.4 Simulation results for missing-at-random data

We consider a missing-at-random mechanism to generate data. Figure 2.13 shows that the biases were very similar to those obtained under a MCAR mechanism, and parameters were estimated without bias.

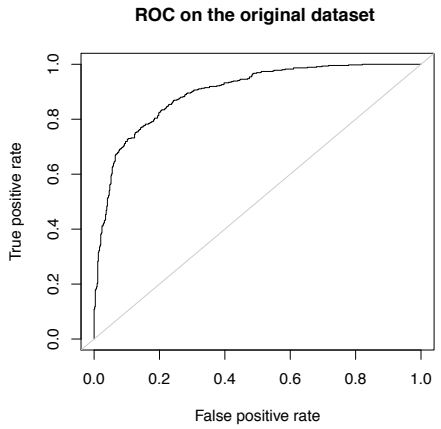


Figure 2.12: ROC curve on a simulated complete dataset.

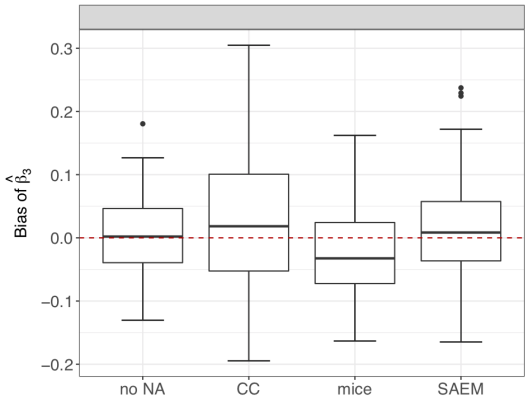


Figure 2.13: Empirical distribution of the bias of $\hat{\beta}_3$ obtained under an MAR mechanism, with $n = 1000$ and 10% missing values.

2.8.5 Simulation results for model misspecification: coverage

Table 2.6 shows the coverage for all parameters, and the average lengths of the corresponding confidence intervals in parentheses.

Table 2.6: Coverage (%) for $n = 1000$, MCAR, and misspecified models, calculated over 1000 simulations. Bold indicates under-coverage. Inside the parentheses is the average length of the corresponding confidence interval over 1000 simulations (multiplied by 100).

parameter	no NA	CC	mice	SAEM
Student distribution:	$(v = 5)$			
β_0	94.7 (68.02)	94.3 (84.14)	94.6 (67.69)	93.8 (68.25)
β_1	95.2 (54.78)	94.2 (72.15)	91.7 (61.96)	93.5 (63.05)
β_2	94.9 (27.66)	94.6 (36.39)	91.4 (31.21)	93.7 (31.84)
β_3	94.9 (26.76)	94.3 (35.24)	**81.5** (30.46)	94.7 (29.98)
β_4	95.2 (11.52)	95.4 (15.16)	95.8 (12.94)	95.5 (12.88)
β_5	93.7 (17.63)	94.9 (23.22)	**83.4** (20.40)	93.3 (19.93)
Gaussian mixture:				
β_0	94.8 (57.54)	95.2 (75.42)	95.4 (61.95)	95.0 (61.33)
β_1	94.7 (58.00)	96.2 (76.05)	95.4 (66.66)	95.3 (66.13)
β_2	94.3 (28.49)	95.3 (37.35)	95.3 (32.65)	94.0 (32.50)
β_3	94.7 (26.16)	94.9 (34.38)	94.9 (28.91)	94.5 (29.10)
β_4	94.4 (12.68)	94.4 (16.60)	94.4 (14.24)	94.7 (14.09)
β_5	95.3 (17.70)	94.7 (23.25)	94.7 (19.86)	95.3 (19.92)

2.8.6 Simulation results varying percentage of missingness

We varied the percentage of missingness from 10% to 30% and results of bias are shown in Figure 2.14.

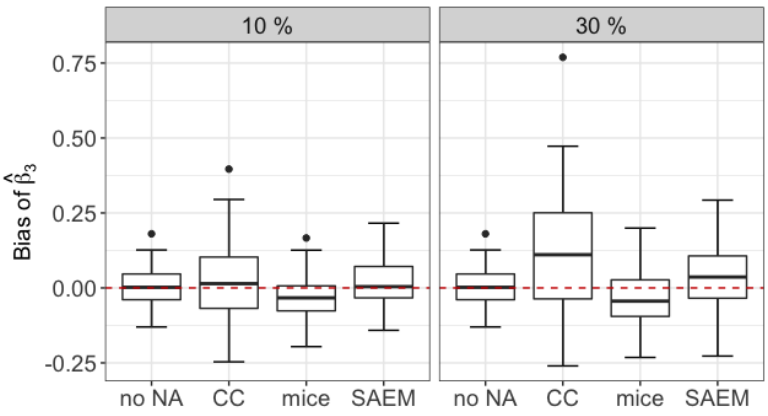
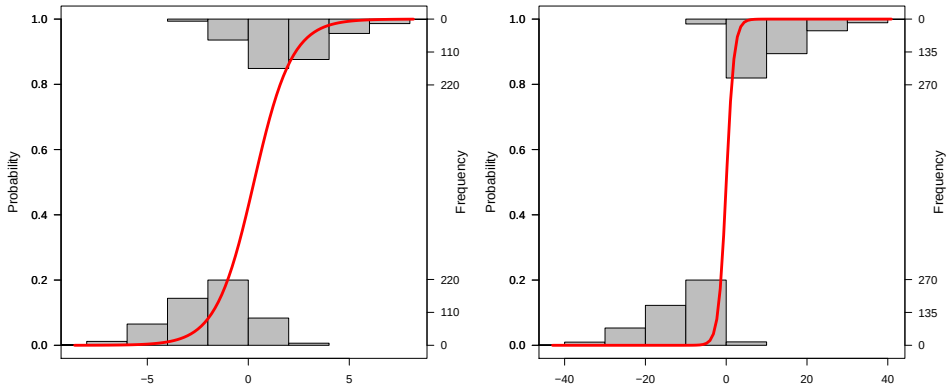


Figure 2.14: Empirical distribution of the bias of $\hat{\beta}_3$ obtained over 1000 simulations, varying the percentage of missingness (left: 10%; right: 30%) under MCAR, with $n = 1000$ with methods no NA, CC, mice and SAEM.

2.8.7 Simulation results varying the separability of classes

We varied the separability of classes by augmenting the value of design matrix $X' = 2X$ or $X' = 5X$ to influence the link function $X'\beta$, where X is the design matrix used in the previous simulation setting in Subsection 6.1. We present the data $(y, X'\beta)$ in Figure 2.15 and the results of bias are shown in Figure 2.16. The left plots represents a case with medium level of separability, where the proposed methodology had a good performance of estimation; while the right plots shows a nearly perfect linear separability, where the performance of mice was strongly affected but the proposed method is still acceptable in comparison to the case without missing values.



(a) $X' = 2X$ (b) $X' = 5X$

Figure 2.15: Logistic regression $(y, X'\beta)$ plot varying the value of link function $X'\beta$.

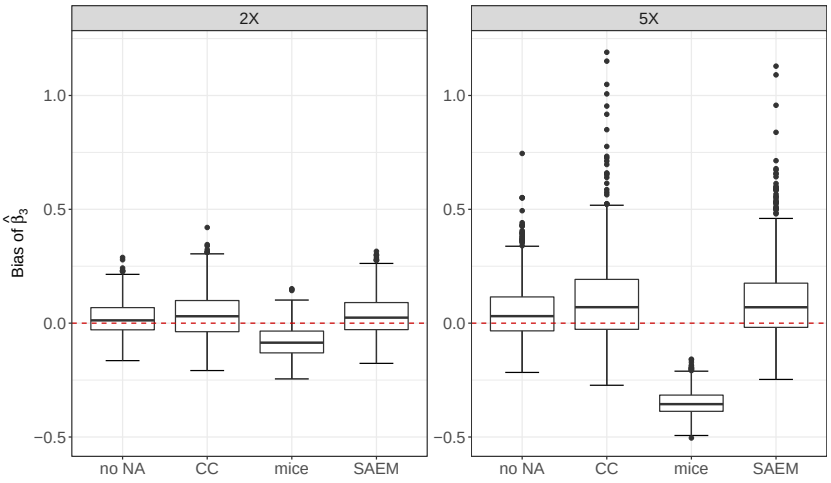


Figure 2.16: Empirical distribution of the bias of $\hat{\beta}_3$ obtained over 1000 simulations, varying the link function (left: $X' = 2X$; right: $X' = 5X$) under MCAR, with $n = 1000$ with methods no NA, CC, mice and SAEM.

2.8.8 Simulation results of comparison with MCEM

We generated a small sample with $n = 200$ in order to illustrate the performance of MCEM, which is computationally intensive. The bias and standard error of estimates over 100 simulations are shown in Figure 2.17.

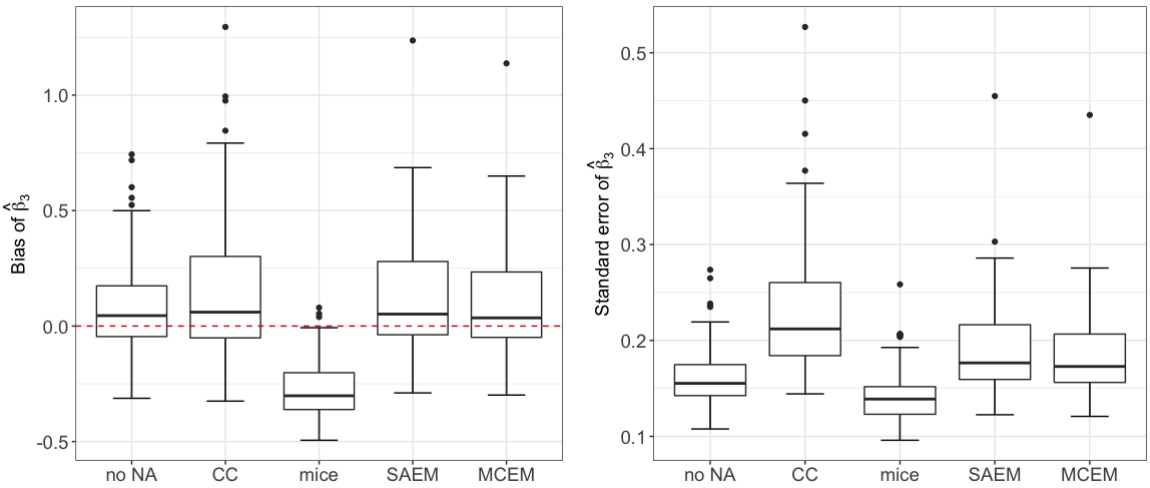


Figure 2.17: Empirical distribution of the bias and standard error of $\hat{\beta}_3$ obtained over 100 simulations, under MCAR, with $n = 200$ and 10% of missing values, with methods no NA, CC, mice, SAEM and MCEM.

Table 2.7: Coverage (%) for $n = 200$, correlation C and 10% MCAR, calculated over 100 simulations. Bold indicates under coverage. Inside the parentheses is the average length of corresponding confidence interval over 100 simulations.

parameter	no NA	CC	mice	SAEM	MCEM
β_0	96 (1.61)	96 (2.20)	97 (1.50)	96 (1.73)	96 (1.71)
β_1	98 (1.44)	95 (1.98)	97 (1.40)	97 (1.70)	99 (1.67)
β_2	97 (0.72)	96 (0.98)	96 (0.69)	97 (0.84)	96 (0.82)
β_3	92 (0.63)	90 (0.90)	**46** (0.56)	89 (0.74)	89 (0.72)
β_4	92 (0.30)	96 (0.41)	95 (0.30)	93 (0.34)	92 (0.34)
β_5	94 (0.43)	94 (0.60)	**54** (0.38)	92 (0.50)	92 (0.49)

Table 2.7 presents the coverage if the confidence interval for all parameters over 100 simulations and inside the parentheses is the average length of corresponding confidence interval over 100 simulations.

2.8.9 Definitions of variables in the TraumaBase dataset

In this section, we define the selected quantitative variables:

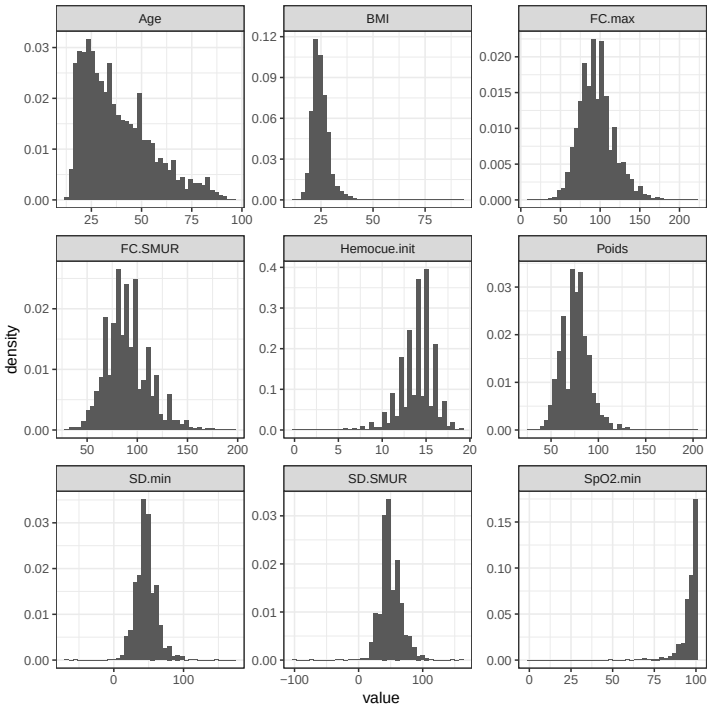
- Age: Age.

- $Poids$: Weight.
- $Taille$: Height.
- BMI: Body Mass index, $BMI = \frac{Weight \text{ in } kg}{(Height \text{ in } m)^2}$
- $Glasgow$: Glasgow Coma Scale.
- $Glasgow.moteur$: Glasgow Coma Scale motor component.
- $PAS.min$: The minimum systolic blood pressure.
- $PAD.min$: The minimum diastolic blood pressure.
- $FC.max$: The maximum number of heart beats per unit time (usually a minute).
- $PAS.SMUR$: Systolic blood pressure at ambulance arrival.
- $PAD.SMUR$: Diastolic blood pressure at ambulance arrival.
- $FC.SMUR$: Heart rate at ambulance arrival.
- $Hemocue.init$: Capillary hemoglobin concentration.
- $SpO2.min$: Oxygen saturation.
- $Remplissage.total.colloides$ (or $RT.colloides$): Fluid expansion colloids.
- $Remplissage.total.cristalloides$ (or $RT.cristalloides$): Fluid expansion cristalloids.
- $SD.min$ $(= PAS.min - PAD.min)$: Pulse pressure for the minimum values of diastolic and systolic blood pressure.
- $SD.SMUR$ $(= PAS.SMUR - PAD.SMUR)$: Pulse pressure at ambulance arrival.

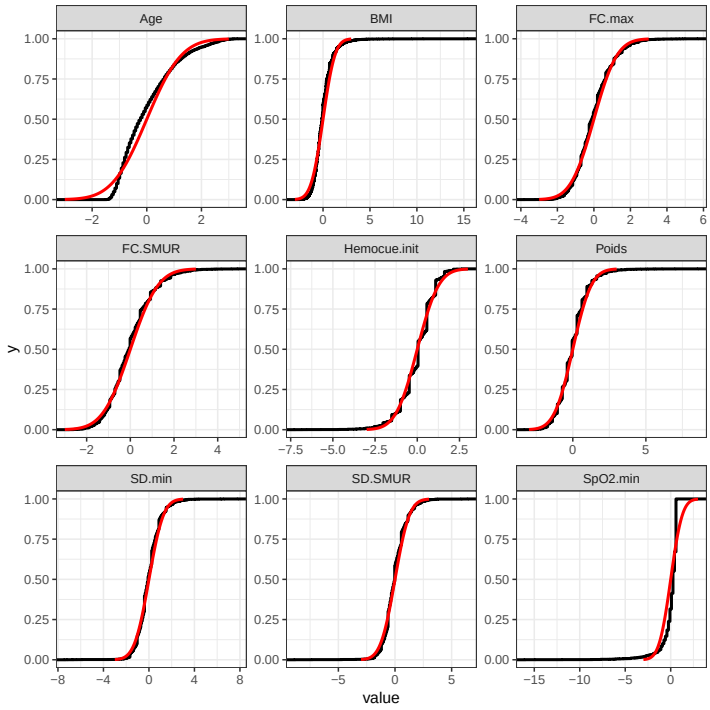
Figure 2.18 shows the histogram and the empirical cumulative distribution function of some of the covariates in TraumaBase. Several of these are not symmetric. In practice, it is possible to consider that suitable transformation of covariates can be approximated by normal distributions. For example, transformations of the form $\log(c + x)$ and $\log(c - x)$, may be appropriate for right-skewed and left-skewed distributions respectively. We applied the proposed methodology to the real dataset after the log-transformation. However, the prediction results from cross-validation did not show any advantage to transforming the variables. Indeed when a log transformation is used as a prepossessing step, it only operates on the observed part, which is appropriate under MCAR values. Consequently, we have decided not to use any transformation on the dataset.

2.8.10 Details on the predictive performance on TraumaBase data

Details on the predictive performance on TraumaBase data are given in Table 2.8.



(a) Histograms of covariates



(b) Empirical cumulative distributions

Figure 2.18: Empirical distributions of variables from TraumaBase. (a) Histograms of covariates; (b) The black line is the empirical cumulative distribution while the red one corresponds to the normal distribution.

Table 2.8: Comparisons of the mean of the predictive performance (values are multiplied by 100) of different methods that can deal with missing data. AUC is the area under the ROC curve; the accuracy is the number of true positives plus true negatives, divided by the total number of observations; the sensitivity is the true positive rate; the specificity is the true negative rate; the precision is the number of true positives over all positive predictions. Best results are shown in bold.

Metrics	SAEM	missForest	impMean	impPCA	mice	predRF	predSVM
AUC	**88.5**	**88.8**	**88.9**	**89.0**	87.7	88.0	80.4
Accuracy	86.9	87.0	87.3	86.7	85.3	87.2	**88.3**
Precision	41.1	41.6	42.2	41.0	37.9	41.6	**44.0**
Sensitivity	**74.6**	74.3	73.2	**75.0**	**75.2**	71.5	66.0
Specificity	88.2	88.4	88.8	87.9	86.4	88.9	**90.6**

2.8.11 First results on logistic regression with mixed and incomplete data

In this section, we briefly discuss the problem of logistic regression with mixed and incomplete data. To solve it, we first suggest modeling the covariates with a simplified general location model then adapt the SAEM algorithm in the mixed data setting.

Covariate model based on GLOM

To specify the joint distribution of mixed covariates, in classification problems, people often refer to the general location model (GLOM) (Olkin et al., 1961), where the categorical variables are marginally distributed as a multinomial distribution with a given number of states (*i.e.,* locations); then given a specific state of the categorical variable, the continuous ones follows conditionally Gaussian distribution, with either homoscedasticity or heteroscedasticity across the locations.

Unfortunately, implementing this kind of model has been hindered in practice by computational issues related to estimation, due to problems associated with the analysis of discrete data. Indeed, when the number of categorical covariates increases and when the latent variables related to missingness involve, solving likelihood problem becomes complicated for such models. Consequently, we take the sparse structure of data into consideration and the proposed model will be a simplified GLOM one.

In the sequel, we assume the same logistic regression model as introduced in Section 2.2. In addition, to correctly identify the continuous variables and the categorical ones contained in X_i, we'll introduce Z_i and U_i to denote respectively the continuous covariates and the categorical ones, such that $X_i = (Z_i, U_i), \quad \forall i \in \{1, \ldots, n\}$. Considering that Z_i is M-dimensional and U_i is L-dimensional, we can go further and write for each subject i: $Z_i = (Z_{im}, m \in \{1, \ldots, M\})$ and $U_i = (U_{i\ell}, \ell \in \{1, \ldots, L\})$, where $Z_{im} \in \mathbb{R}$ and $U_{i\ell} \in \{1, 2, \ldots, K_\ell\}$ (K_ℓ = number of categories associated to the random variable U_ℓ, $\ell \in \{1, \ldots, L\}$).

To make the notations more precise, when a specific distribution depends on parameter θ, we'll explicit the dependence by writing the distribution as follows: $\mathrm{p}(X_i\,;\theta_X)$ for the

distribution of X, or $\mathrm{p}(y_i|X_i\,;\,\theta_{y|X})$ for the conditional distribution of $y|X$ for instance.

To set up the model for the joint distribution of covariates, let's first decompose $\mathrm{p}(X_i\,;\,\theta_X)$ with respect to the continuous covariates Z_i and categorical ones U_i:

$$\begin{aligned}\mathrm{p}(X_i\,;\,\theta_X) &= \mathrm{p}(Z_i, U_i\,;\,\theta_{Z,U}) \\ &= \mathrm{p}(Z_i|U_i\,;\,\theta_{Z|U})\,\mathrm{p}(U_i\,;\,\theta_U)\,.\end{aligned}$$

We can decompose the joint distribution of $X = (Z, U)$ by either opting for the conditional distribution of $(Z|U)$ multiplied by the marginal distribution of U, or the opposite, *i.e.* the conditional distribution of $(U|Z)$ multiplied by the marginal distribution of Z. Our choice to focus on the former one, was supported by the opinions of medical experts in favor of the dependence of the continuous covariates Z on the categorical ones U. For instance, it seems reasonable to assume that the blood pressure or heart rate values (*i.e.*, continuous covariates) observed in a patient will be significantly different depending on whether the kinetics of the trauma was high or low (*i.e.*, categorical covariate). Furthermore, as most of these categorical variables are actually binary, with possible strong correlations between some of them, we could expect an easier modeling.

Simplified GLOM linear model for $\mathrm{p}(Z_i|U_i\,;\,\theta_{Z|U})$

Due to the presence of discrete data, we will first adopt a simplified formulation of the conditional distribution of $(Z|U)$, derived from the most general GLOM approach. More precisely, we assume the following linear model between Z and U:

$$\begin{aligned}
Z_{i1} &= \lambda_{01} + \sum_{\ell=1}^{L}\sum_{k=1}^{K_\ell-1} \lambda_{\ell 1}^k \mathbb{1}_{U_{i\ell}=k} + \varepsilon_{i1}\,, \\
Z_{i2} &= \lambda_{02} + \sum_{\ell=1}^{L}\sum_{k=1}^{K_\ell-1} \lambda_{\ell 2}^k \mathbb{1}_{U_{i\ell}=k} + \varepsilon_{i2}\,, \\
&\vdots \qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\
Z_{iM} &= \lambda_{0M} + \sum_{\ell=1}^{L}\sum_{k=1}^{K_\ell-1} \lambda_{\ell M}^k \mathbb{1}_{U_{i\ell}=k} + \varepsilon_{iM}\,,
\end{aligned} \qquad \forall i \in \{1,\dots,n\}. \tag{2.8}$$

where for each $m \in \{1,\dots,M\}$, the vector of parameters associated to the m^{th} continuous variable Z_m is:

$$\lambda_m = (\lambda_{0m}, \lambda_{1m}^1, \dots, \lambda_{1m}^{K_1-1}, \lambda_{2m}^1, \dots, \lambda_{2m}^{K_2-1}, \dots, \lambda_{Lm}^1, \dots, \lambda_{Lm}^{K_L-1}) \in \mathbb{R}^{\sum_{\ell=1}^{L} K_\ell - L + 1}\,.$$

And a block covariance matrix $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iM}) \sim \mathcal{N}(0_{\mathbb{R}^M}, \Sigma)$ with $\Sigma \in \mathcal{S}_M(\mathbb{R})$ and it denotes the $n-$dimensional vector of residual terms observed within the n individuals when considering the m^{th} continuous variable $Z_m = (Z_{1m}, \dots, Z_{nm})$.

Note that:

- As for each $\ell \in \{1,\dots,L\}$ and $i \in \{1,\dots,n\}$, we get the relationship $\sum_{k=1}^{K_\ell} \mathbb{1}_{U_{i\ell}=k} = 1$, we only need to define $K_\ell - 1$ parameters $(\lambda_{\ell m}^1, \dots, \lambda_{\ell m}^{K_\ell-1})$ associated to each couple of variables (Z_m, U_ℓ) to perfectly define the above model;

- If for each couple of variables (Z_m, U_ℓ), it appears that U_ℓ has no influence on Z_m then we get $(\lambda_{\ell m}^1, \dots, \lambda_{\ell m}^{K_\ell - 1})^\top = 0_{\mathbb{R}^{K_\ell - 1}}$;
- This formulation helps reduce the dimensionality of parameters, because supposing that the ℓ^{th} categorical variable U_ℓ has K_ℓ categories, that makes a total of $\prod_{\ell=1}^L K_\ell$ possible states for the multinomial U. In a classical GLOM model, where the continuous variables are conditionally multivariate normal given each state of the multinomial variable, that implies $M \times \prod_{\ell=1}^L K_\ell$ unknown values to only characterize each mean. With the above hypothesis, there are only $M \times (\sum_{\ell=1}^L K_\ell - L + 1)$ values to determine. As an example, let's take 10 continuous covariates and 15 binary covariates in the mixed setting (which is approximately the values we'll find in the Traumabase pre-hospital measurements), we get 327 680 (classical GLOM) vs. 160 (linear).

Thus if we use the notations $\dot{U}$ for the following design matrix:

$$\dot{U} = \underbrace{\begin{pmatrix} 1 & \mathbb{1}_{U_{11}=1} & \dots & \mathbb{1}_{U_{11}=K_1-1} & \mathbb{1}_{U_{12}=1} & \dots & \mathbb{1}_{U_{12}=K_2-1} & \dots & \dots & \mathbb{1}_{U_{1L}=1} & \dots & \mathbb{1}_{U_{1L}=K_L-1} \\ 1 & \mathbb{1}_{U_{21}=1} & \dots & \mathbb{1}_{U_{21}=K_1-1} & \mathbb{1}_{U_{22}=1} & \dots & \mathbb{1}_{U_{22}=K_2-1} & \dots & \dots & \mathbb{1}_{U_{2L}=1} & \dots & \mathbb{1}_{U_{2L}=K_L-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ 1 & \mathbb{1}_{U_{n1}=1} & \dots & \mathbb{1}_{U_{n1}=K_1-1} & \mathbb{1}_{U_{n2}=1} & \dots & \mathbb{1}_{U_{n2}=K_2-1} & \dots & \dots & \mathbb{1}_{U_{nL}=1} & \dots & \mathbb{1}_{U_{nL}=K_L-1} \end{pmatrix}}_{\text{matrix dimension } = n \times (\sum_{\ell=1}^L K_\ell - L + 1)} ;$$

Then We can write eq. (2.8) in the synthetic linear matrix form:

$$Z_m = \dot{U}\, \lambda_m + \varepsilon_m \,, \qquad \forall m \in \{1, \dots, M\} \,. \tag{2.9}$$

Note that $\dot{U}$ is almost the complete disjunctive table associated to the $L-$dimensional multivariate categorical variable U.

To completely define the aforementioned linear model eq. (2.9), the parameters $\theta_{Z|U}$ to characterize are the components of the matrices $\Lambda = (\lambda_1, \dots, \lambda_M)^\top$ (dimension $= M \times (\sum_{\ell=1}^L K_\ell - L + 1)$) and Σ (dimension $M \times M$). Now given an observation U_i of multinomial variable U for subject i (with $\prod_{\ell=1}^L K_\ell$ possible states for U_i), let U_i^{dum} denote the value taken for subject i by the dummy variable associated to U, that we'll denote by U^{dum}, we can write:

$$U_i^{\text{dum}} = (\mathbb{1}_{U_{i1}=1}, \dots, \mathbb{1}_{U_{i1}=K_1-1}, \mathbb{1}_{U_{i2}=1}, \dots, \mathbb{1}_{U_{i2}=K_2-1}, \dots, \mathbb{1}_{U_{iL}=1}, \dots, \mathbb{1}_{U_{iL}=K_L-1}) \,.$$

If we introduce the notation $\dot{U}_i^{\text{dum}}$ such that: $\dot{U}_i^{\text{dum}} = \left(1, U_i^{\text{dum}}\right)$, we'll then be able to explicit the multivariate normal conditional distribution of $(Z|U)$ thanks to the notations proposed. Indeed we can write:

$$Z_i | U_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}(\Lambda \dot{U}_i^{\text{dum}}, \Sigma) \,.$$

Consequently we get:

$$\mathrm{p}(Z_i | U_i \,;\, \Lambda, \Sigma) = \frac{1}{(2\pi)^{M/2}\, |\Sigma|^{1/2}} \exp\Big(-\frac{1}{2} (Z_i - \Lambda \dot{U}_i^{\text{dum}})^\top \Sigma^{-1} (Z_i - \Lambda \dot{U}_i^{\text{dum}}) \Big) \,.$$

We have therefore defined a GLOM-type model where given a specific state U_i for the discrete variable U in subject i, the conditional distribution of $(Z|U)$ in subject i is multivariate

normal, with mean $\Lambda \dot{U}_i^{\text{dum}}$ and covariance matrix Σ. As we hypothesize that Σ is the same across different locations, we are in a homoscedasticity setting.

Depending on how we hypothesize the structure of the block matrix Σ, *i.e.* depending on the number and sizes of the different blocks, we'll obtain a "clustered" structure for the conditional joint distribution of $(Z_1, \ldots, Z_M)$ given U, that is a partition of the set of variables $(Z_1, \ldots, Z_M)$ in a given number of subsets (= "clusters"), with variables mutually correlated when contained in the same cluster, and conditionally independent of those from other clusters.

Thus, if we define C_Z clusters, it will always be possible to rearrange the order of the continuous components of Z so as to express Σ with the following form:

$$\Sigma = \underbrace{\begin{pmatrix} \underbrace{\boxed{\Sigma_1}}_{\text{size} = |\text{cluster } 1|} & 0 & \ldots & 0 \\ 0 & \underbrace{\boxed{\Sigma_2}}_{\text{size} = |\text{cluster } 2|} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \underbrace{\boxed{\Sigma_{C_Z}}}_{\text{size} = |\text{cluster } C_Z|} \end{pmatrix}}_{\text{size} = M} \tag{2.10}$$

And the conditional distribution $\mathrm{p}(Z_i|U_i\,;\,\theta_{Z|U}) = \mathrm{p}(Z_i|U_i\,;\,\Lambda, \Sigma)$ can then be simplified in:

$$\mathrm{p}(Z_i|U_i\,;\,\Lambda, \Sigma) = \prod_{d=1}^{C_Z} \mathrm{p}(\widetilde{Z}_{id}|U_i\,;\,\Lambda_d, \Sigma_d)\,, \tag{2.11}$$

where $\widetilde{Z}_d$ refers to the variable associated to the d^{th} cluster ($d \in \{1, \ldots, C_Z\}$), which can be univariate or multivariate, depending on the size of the cluster.

Note that we can write:

$$Z = (Z_1, \ldots, Z_M) = (\widetilde{Z}_1, \ldots, \widetilde{Z}_{C_Z})\,.$$

In the eq. (2.11), Σ_d represents the d^{th} block of Σ ($d \in \{1, \ldots, C_Z\}$), and Λ_d represents the submatrix of Λ formed by the lines of Λ corresponding to the variables Z_m specifically contained in $\widetilde{Z}_d$. To illustrate this, let's say we have 5 continuous variables $(Z_1, Z_2, Z_3, Z_4, Z_5)$ grouped in 3 clusters when considering the conditional distribution of $(Z|U)$. If the clusters are such that $\widetilde{Z}_1 = Z_2$, $\widetilde{Z}_2 = (Z_1, Z_3)$ and $\widetilde{Z}_3 = (Z_4, Z_5)$, then we have also $\Lambda_1 = \lambda_2$, $\Lambda_2 = (\lambda_1, \lambda_3)$, and $\Lambda_3 = (\lambda_4, \lambda_5)$.

Clustered multinomial distributions for $\mathrm{p}(U_i\,;\,\theta_U)$

Quite similarly to the above section, we propose a "clustered" dependency structure to model the joint distribution of our categorical variable $(U_1,\ldots,U_L)$, that is a partition of this set of variables in C_U independent categorical variables called $\widetilde{U}_j$ ($j\in\{1,\ldots,C_U\}$), with C_U the number of different independent clusters.

Then we assume that each $\widetilde{U}_j$ follows a multinomial distribution with parameters $\pi_j = (\pi_{j1},\pi_{j2},\ldots,\pi_{jR_j})$ (R_j-dimensional probability vector, R_j being the number of modalities associated to the j^{th} variable $\widetilde{U}_j$, with $\pi_{jr}=\mathbb{P}(\widetilde{U}_{ij}=r)\ \forall i\in\{1,\ldots,n\},\ \forall r\in\{1,\ldots,R_j\}$ and $\sum_{r=1}^{R_j}\pi_{jr}=1$). We can write:

$$\mathrm{p}(U_i\,;\,\theta_U)=\prod_{j=1}^{C_U}\mathrm{p}(\widetilde{U}_{ij}\,;\,\pi_j)\,. \tag{2.12}$$

Naturally, the parameters to be estimated in this part of the model are the components of the parameter vectors π_j, such that $\theta_U=(\pi_j,\ j\in\{1,\ldots,C_U\})$.

Having specified each probability distribution composing our model as well as their parameters, we'll now focus on the different tasks to execute.

Adaptation of SAEM in mixed data setting

We recall that one of the key steps in SAEM at k^{th} iteration is the simulation step as introduced in Section 2.3, *i.e.* being able to draw for each individual i a set of samples $X_{i,\text{mis}}^{(k)}$ from the conditional distribution $\mathrm{p}(X_{i,\text{mis}}\,|\,y_i,X_{i,\text{obs}}\ ;\,\theta^{(k-1)})$

Within mixed data setting, the objective becomes to draw two subsets of unobserved variables $Z_{i,\text{mis}}^{(k)}$ and $U_{i,\text{mis}}^{(k)}$, one continuous and one categorical respectively, from the conditional joint distribution:

$$\mathrm{p}(Z_{i,\text{mis}},U_{i,\text{mis}}\,|\,y_i,Z_{i,obs},U_{i,obs}\ ;\,\theta^{(k-1)})\,.$$

Drawing from such a distribution can be achieved by Gibbs sampling, as detailed in Algorithm 4.

Algorithm 4 Gibbs sampling applied to mixed data: draw $Z_{i,mis}^{(k)}$ and $U_{i,\text{mis}}^{(k)}$, at k^{th} SAEM iteration.

Input: An initial sample $U_{i,\text{mis}}^{(k,0)}=U_{\text{mis}}^{(k-1)}$ obtained from previous $(k-1)^{\text{th}}$ SAEM iteration;
for $t=1,2,\ldots,T$ **do**
Generate

$$Z_{i,\text{mis}}^{(k,t)}\ \sim\ \mathrm{p}(Z_{i,\text{mis}}\,|\,y_i,Z_{i,\text{obs}},U_{i,\text{obs}},U_{i,\text{mis}}^{(k,t-1)}\ ;\,\theta^{(k-1)})\,; \tag{2.13}$$

Generate

$$U_{i,\text{mis}}^{(k,t)}\sim\ \mathrm{p}(U_{i,\text{mis}}\,|\,y_i,Z_{i,\text{obs}},U_{i,\text{obs}},Z_{i,\text{mis}}^{(k,t)}\ ;\,\theta^{(k-1)})\,; \tag{2.14}$$

Output: $Z_{i,\text{mis}}^{(k,T)}=Z_{i,\text{mis}}^{(k)}$ and $U_{i,\text{mis}}^{(k,T)}=U_{i,\text{mis}}^{(k)}\ ,\quad i=1,2,\cdots,n.$

In order to define relevant proposal distributions for MH sampling, let's observe how the target distributions in eq. (2.13) and eq. (2.14) can be factorized:

- For continuous missing covariates sampling:

$$\begin{aligned}\mathrm{p}(Z_{i,\mathrm{mis}} \,|\, y_i, Z_{i,\mathrm{obs}}, U_i \;;\, \theta) &\underset{w.r.t.Z_{mis}}{\propto} \mathrm{p}(y_i \,|\, Z_i, U_i \;;\, \beta) \cdot \mathrm{p}(Z_{i,\mathrm{mis}} \,|\, Z_{i,\mathrm{obs}}, U_i \;;\, \Lambda, \Sigma) \\ &\underset{w.r.t.Z_{\mathrm{mis}}}{\propto} \mathrm{p}(y_i \,|\, X_i \;;\, \beta) \cdot \mathrm{p}(Z_{i,\mathrm{mis}} \,|\, Z_{i,\mathrm{obs}}, U_i \;;\, \Lambda, \Sigma)\,.\end{aligned} \tag{2.15}$$

Then, to draw $Z_{i,\mathrm{mis}}^{(k,t)}$ from $\mathrm{p}(Z_{i,\mathrm{mis}} \,|\, y_i, Z_{i,\mathrm{obs}}, U_{i,\mathrm{obs}}, U_{i,\mathrm{mis}}^{(k,t-1)} \;;\, \theta^{(k-1)})$ in eq. (2.13), we'll select as a proposal distribution for MH algorithm (see Section 2.3.3) the second term in the above factorization, *i.e.* $\mathrm{p}(Z_{i,\mathrm{mis}} \,|\, Z_{i,\mathrm{obs}}, U_i \;;\, \Lambda^{(k-1)}, \Sigma^{(k-1)})$, which is normally distributed:

$$Z_{i,\mathrm{mis}} \,|\, Z_{i,\mathrm{obs}}, U_i \sim \mathcal{N}(\mu_i^*, \Sigma_i^*)\,,$$

where

$$\begin{aligned}\mu_i^* &= (\Lambda^{(k-1)} \dot{U}_i^{\mathrm{dum}})_{i,\mathrm{mis}} \;+\; \Sigma_{i,\mathrm{mis,obs}}^{(k-1)} \left(\Sigma_{i,\mathrm{obs,obs}}^{(k-1)}\right)^{-1} \left(Z_{i,\mathrm{obs}} - (\Lambda^{(k-1)} \dot{U}_i^{\mathrm{dum}})_{i,\mathrm{obs}}\right), \\ \Sigma_i^* &= \Sigma_{i,\mathrm{mis,mis}}^{(k-1)} \;-\; \Sigma_{i,\mathrm{mis,obs}}^{(k-1)} \left(\Sigma_{i,\mathrm{obs,obs}}^{(k-1)}\right)^{-1} \Sigma_{i,\mathrm{obs,mis}}^{(k-1)}\,,\end{aligned}$$

with $(\Lambda^{(k-1)} \dot{U}_i^{\mathrm{dum}})_{i,\mathrm{mis}}$ (resp. $(\Lambda^{(k-1)} \dot{U}_i^{\mathrm{dum}})_{i,\mathrm{obs}}$) the missing (resp. observed) elements, with respect to $Z_{i,\mathrm{mis}}$ (resp. $Z_{i,\mathrm{obs}}$), of $\Lambda^{(k-1)} \dot{U}_i^{\mathrm{dum}}$. The covariance matrix $\Sigma^{(k-1)}$ is decomposed in the same way. We recall that $\Lambda^{(k-1)}$ and $\Sigma^{(k-1)}$ refer to the estimates of the parameters Λ and Σ updated at the end of $(k-1)^{\mathrm{th}}$ SAEM iteration. Note also that actually we should consider $\Lambda_d^{(k-1)}$ and $\Sigma_d^{(k-1)}$ (where $d \in \{1, \ldots, C_Z\}$) rather than $\Lambda^{(k-1)}$ and $\Sigma^{(k-1)}$, due to the "cluster" structure defined in eq. (2.10) but we wanted to preserve readability and avoid overly heavy notations.

Furthermore, the factorization proposed in eq. (2.15) allows us to precise an explicit form for the ratio of the target distribution over the proposal one, which is mandatory in MH structure (as detailed in Section 2.3.3). Here this ratio reads:

$$\begin{aligned}&\frac{\mathrm{p}(Z_{i,\mathrm{mis}}^{(k,t,s)} \,|\, y_i, Z_{i,\mathrm{obs}}, U_{i,\mathrm{obs}}, U_{i,\mathrm{mis}}^{(k,t-1)} \;;\, \theta^{(k-1)})}{\mathrm{p}(Z_{i,\mathrm{mis}}^{(k,t,s)} \,|\, Z_{i,\mathrm{obs}}, U_{i,\mathrm{obs}}, U_{i,\mathrm{mis}}^{(k,t-1)} \;;\, \mu_i^*, \Sigma_i^*)} \\ &= \mathrm{p}(y_i \,|\, X_{i,\mathrm{comp}}^{(k,t,s)} \;;\, \beta^{(k-1)}) \\ &= \begin{cases} \dfrac{\exp\left(\beta^{(k-1)\top} \cdot \left(\dot{X}_{i,\mathrm{comp}}^{(k,t,s)}\right)^{\mathrm{dum}}\right)}{1+\exp\left(\beta^{(k-1)\top} \cdot \left(\dot{X}_{i,\mathrm{comp}}^{(k,t,s)}\right)^{\mathrm{dum}}\right)}\,, & \text{if } y_i = 1\,, \\ 1 - \dfrac{\exp\left(\beta^{(k-1)\top} \cdot \left(\dot{X}_{i,\mathrm{comp}}^{(k,t,s)}\right)^{\mathrm{dum}}\right)}{1+\exp\left(\beta^{(k-1)\top} \cdot \left(\dot{X}_{i,\mathrm{comp}}^{(k,t,s)}\right)^{\mathrm{dum}}\right)}\,, & \text{if } y_i = 0\,, \end{cases}\end{aligned} \tag{2.16}$$

where $X_{i,\mathrm{comp}}^{(k,t,s)} = (Z_{i,\mathrm{obs}}, Z_{i,\mathrm{mis}}^{(k,t,s)}, U_{i,\mathrm{obs}}, U_{i,\mathrm{mis}}^{(k,t-1)})$ is a completed set of observations for individual i, at k^{th} iteration of SAEM, t^{th} iteration of Gibbs sampling, and s^{th} iteration of Metropolis-Hastings sampling; and we denote:

$$\left(X_{i,\mathrm{comp}}^{(k,t,s)}\right)^{\mathrm{dum}} = (Z_{i,\mathrm{obs}}, Z_{i,\mathrm{mis}}^{(k,t,s)}, U_{i,\mathrm{obs}}^{\mathrm{dum}}, (U_{i,\mathrm{mis}}^{(k,t-1)})^{\mathrm{dum}})\,,$$

the dummy observation vector associated with $X_{i,\mathrm{comp}}^{(k,t,s)}$. And finally we have $\left(\dot{X}_{i,\mathrm{comp}}^{(k,t,s)}\right)^{\mathrm{dum}} = \left(1, \left(X_{i,\mathrm{comp}}^{(k,t,s)}\right)^{\mathrm{dum}}\right)$.

- For categorical missing covariates sampling:

$$\begin{aligned} & \mathrm{p}(U_{i,\mathrm{mis}} \,|\, y_i, Z_i, U_{i,\mathrm{obs}} \;;\; \theta) \\ \underset{w.r.t.U_{\mathrm{mis}}}{\propto} & \mathrm{p}(y_i \,|\, Z_i, U_i \;;\; \beta) \cdot \mathrm{p}(Z_i \,|\, U_i \;;\; \Lambda, \Sigma) \cdot \mathrm{p}(U_{i,\mathrm{mis}} \,|\, U_{i,\mathrm{obs}} \;;\; (\pi_j)_{j\in\{1,\dots,C_U\}}) \\ \underset{w.r.t.U_{\mathrm{mis}}}{\propto} & \mathrm{p}(y_i \,|\, X_i \;;\; \beta) \cdot \mathrm{p}(Z_i \,|\, U_i \;;\; \Lambda, \Sigma) \cdot \mathrm{p}(U_{i,\mathrm{mis}} \,|\, U_{i,\mathrm{obs}} \;;\; (\pi_j)_{j\in\{1,\dots,C_U\}}) \,. \end{aligned} \tag{2.17}$$

  Then, to draw $U_{i,\mathrm{mis}}^{(k,t)}$, we'll select the third term in the above factorization: $\mathrm{p}(U_{i,mis} \,|\, U_{i,obs} \;;\; (\pi_j{}^{(k-1)})_{j\in\{1,\dots,C_U\}})$, which can be defined according to the cluster structure defined in eq. (2.12) and by ratios of appropriate probabilities extracted from the set of parameters $\theta_U^{(k-1)} = (\pi_j{}^{(k-1)},\ j \in \{1,\dots,C_U\})$, whose expressions rely on a case-by-case basis.

  Furthermore, as seen in eq. (2.16), the factorization proposed in eq. (2.17) allows us to precise an explicit form for the ratio of distributions:

$$\begin{aligned} & \frac{\mathrm{p}(U_{i,\mathrm{mis}}^{(k,t,s)} \,|\, y_i, Z_{i,\mathrm{obs}}, U_{i,\mathrm{obs}}, Z_{i,\mathrm{mis}}^{(k,t)} \;;\; \theta^{(k-1)})}{\mathrm{p}(U_{i,\mathrm{mis}}^{(k,t,s)} \,|\, U_{i,\mathrm{obs}} \;;\; (\pi_j{}^{(k-1)})_{j\in\{1,\dots,C_U\}})} \\ = & \mathrm{p}(y_i \,|\, X_{i,\mathrm{comp}}^{(k,t,s)} \;;\; \beta^{(k-1)}) \;\cdot\; \mathrm{p}(Z_{i,\mathrm{comp}}^{(k,t)} \,|\, U_{i,\mathrm{comp}}^{(k,t,s)} \;;\; \Lambda^{(k-1)}, \Sigma^{(k-1)}) \,, \end{aligned}$$

  where:

  - $\mathrm{p}(y_i \,|\, X_{i,\mathrm{comp}}^{(k,t,s)} \;;\; \beta^{(k-1)})$ has an explicit form given by eq. (2.16), depending on the value of $y_i \in \{0\,, 1\}$;
  - $\mathrm{p}(Z_{i,\mathrm{comp}}^{(k,t)} \,|\, U_{i,\mathrm{comp}}^{(k,t,s)} \;;\; \Lambda^{(k-1)}, \Sigma^{(k-1)})$ is normally distributed according to:

$$Z_{i,\mathrm{comp}}^{(k,t)} \,|\, U_{i,\mathrm{comp}}^{(k,t,s)} \;\sim\; \mathcal{N}\Big(\Lambda^{(k-1)} \big(\dot{U}_{i,\mathrm{comp}}^{(k,t,s)}\big)^{\mathrm{dum}} \,,\; \Sigma^{(k-1)}\Big) \,,$$

  with $X_{i,\mathrm{comp}}^{(k,t,s)} = (Z_{i,\mathrm{obs}}, Z_{i,\mathrm{mis}}^{(k,t)}, U_{i,\mathrm{obs}}, U_{i,\mathrm{mis}}^{(k,t,s)})$ a completed set of observations for individual i, at k^{th} iteration of SAEM, t^{th} iteration of Gibbs sampling, and s^{th} iteration of Metropolis-Hastings sampling, defined analogously as in eq. (2.16) ; $\big(X_{i,\mathrm{comp}}^{(k,t,s)}\big)^{\mathrm{dum}} = \big(Z_{i,\mathrm{obs}}, Z_{i,\mathrm{mis}}^{(k,t)}, U_{i,\mathrm{obs}}^{\mathrm{dum}}, (U_{i,\mathrm{mis}}^{(k,t,s)})^{\mathrm{dum}}\big)$ the dummy observation vector associated with $X_{i,\mathrm{comp}}^{(k,t,s)}$, and $\big(\dot{X}_{i,\mathrm{comp}}^{(k,t,s)}\big)^{\mathrm{dum}} = \Big(1, \big(X_{i,\mathrm{comp}}^{(k,t,s)}\big)^{\mathrm{dum}}\Big)$; and with $Z_{i,\mathrm{comp}}^{(k,t)} = (Z_{i,\mathrm{obs}}, Z_{i,\mathrm{mis}}^{(k,t)})$ the analogously defined completed continuous observation vector for individual i at k^{th} SAEM iteration and t^{th} Gibbs iteration; and with finally $\big(\dot{U}_{i,\mathrm{comp}}^{(k,t,s)}\big)^{\mathrm{dum}} = \big(1, U_{i,\mathrm{obs}}^{\mathrm{dum}}, (U_{i,\mathrm{mis}}^{(k,t,s)})^{\mathrm{dum}}\big)$ is the "extended-dummy" observation vector associated with completed categorical $U_{i,\mathrm{comp}}^{(k,t,s)} = (U_{i,\mathrm{obs}}, U_{i,\mathrm{mis}}^{(k,t,s)})$.

Simulation study

We first generated the following design matrix $X = (Z, U)$ with the characteristics:

- dimension $n_1 = 200 \times p = 4$;

- 2 categorical variables (U_1, U_2), with 3 possible categories for U_1 , and 2 categories for U_2 (binary), such that $U_{i1} \in \{$"red"; "blue"; "yellow"$\}$ and $U_{i2} \in \{$"big"; "small"$\}$ for each individual $i \in \{1, \ldots, n\}$;
- 2 continuous variables (Z_1, Z_2) such that the conditional distribution of $(Z_1, Z_2|U_1, U_2)$ is bivariate normal;
- we considered as the true parameter values:

$$\beta = (\beta_0, \beta_1^Z, \beta_2^Z, \beta_{11}^U, \beta_{12}^U, \beta_{21}^U)^\top = \begin{bmatrix} 0 & 2 & -3 & 3 & -1 & 2 \end{bmatrix}, \tag{2.18}$$

$$\Lambda = \begin{bmatrix} 0 & 2 & 3 & 2 \\ 0 & 3 & 2 & 3 \end{bmatrix}, \tag{2.19}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{2.20}$$

$$\pi_1 = \begin{bmatrix} \pi_{\text{red}} \\ \pi_{\text{blue}} \\ \pi_{\text{yellow}} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix}, \tag{2.21}$$

$$\pi_2 = \begin{bmatrix} \pi_{\text{big}} \\ \pi_{\text{small}} \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix}. \tag{2.22}$$

- Generative process : we first generate n_1 samples of (U_1, U_2) from their own multinomial distribution of respective parameters (π_1, π_2); given the categories generated, we transform each U_i vector into its dummy equivalent U_i^{dum} and we can then generate values for $(Z_1, Z_2|U_1, U_2)$ from the bivariate normal distribution of parameters $\Lambda \dot{U}_i^{\text{dum}}$ (mean) and Σ (covariance matrix); eventually, we are able to generate the binary response y_i for each individual i according to a logistic regression model with parameters β.

Our objective given these simple settings is to assess how the proposed approach behaves in estimating the true parameters defined above, with a given amount of missing values introduced in the covariates according to a given missingness mechanism. We'll systematically compare the results of SAEM to the other methods as mentioned in Section 2.5. We note that for **MeanImp** method, *i.e.* mean imputation, each missing value in a given column is replaced either by the mean of its belonging column for continuous covariates, or by a sample from the empirical probability mass function for categorical covariates; then a logistic regression estimation is performed on the completed dataset.

In Figure 2.19, one can observe the convergence profiles of the estimates across SAEM iterations, which are quite acceptable.

In Figures 2.20, 2.21, 2.22 and 2.23 presented below, we aimed at comparing the empirical distributions of the relative bias of each estimate, across the five different methods considered, focusing in each figure on the variation of a given configuration parameter:

- Figure 2.20 shows the comparative effects of two different sample sizes:
  - $n_1 = 200$;
  - $n_2 = 1000$;

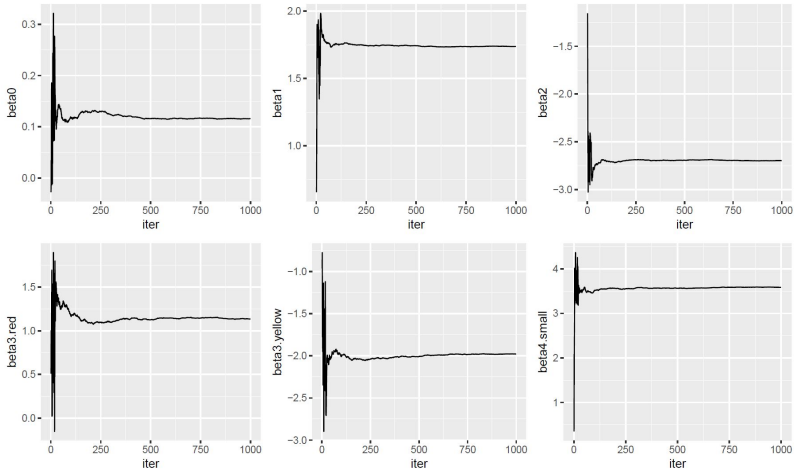


Figure 2.19: Convergence profiles of the parameter estimates across SAEM iterations (sample size $n_1 = 200$, 10% of missingness entry-wise, affecting all variables, MCAR mechanism)

- Figures 2.21 and 2.22 show the comparative effects of three different missingness mechanisms:
  - MCAR, where each entry has the same probability to be observed;
  - MAR with missingness in (Z_2, U_1, U_2) depending on values of fully observed Z_1;
  - MAR with missingness in Z_2 conditionally to fully observed Z_1, and in U_2 conditionally to fully observed (Z_1, U_1);
- Figure 2.23 shows the comparative effects of two different percentages of missingness:
  - 10% entry-wise;
  - 30% entry-wise;

Each estimation task is replicated 100 times per setting. We recall that the relative bias of an estimate, for instance $\hat{\beta}_0$, is defined by:

$$\text{Relative bias}\,(\hat{\beta}_0) = \frac{\hat{\beta}_0 - \beta_0^{\text{true}}}{|\,\beta_0^{\text{true}}\,|} \tag{2.23}$$

As expected, larger samples yielded less variability. Moreover, we observe that SAEM provided unbiased estimates (which is clearly highlighted with $n_2 = 1000$) with small variances. In contrast, estimating with a preliminary mean imputation produced a significant bias. It also appeared that the standard errors observed in CC estimates seemed to be regularly worse than those for SAEM, such a differential effect being more visible when the missingness rate increased (see Figure 2.23), or when the missingness mechanism was assumed to be MAR rather than MCAR (see Figures 2.21 and 2.22). Unsurprisingly, the apparent unbiased property of the same CC method in our different settings also vanished when the missingness assumption changed from MCAR to MAR. Finally it seems in Figure 2.21 that MAR assumption introduced some bias tendency in SAEM estimates, but this may result from a somewhat small sample size, and unfortunately we had not performed yet the simulations with $n_2 = 1000$ to confirm this hypothesis.

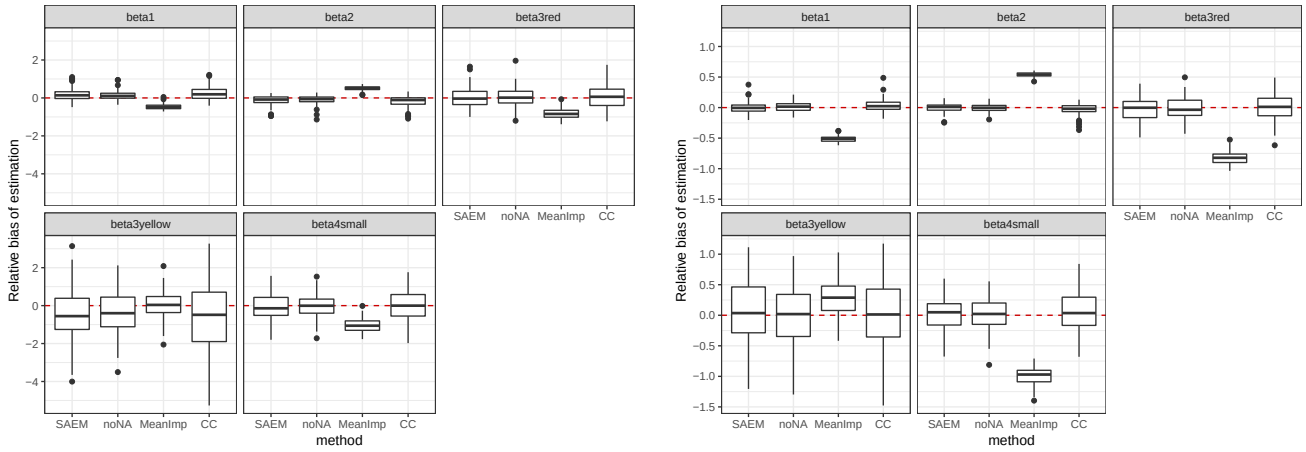


Figure 2.20: Empirical distribution of the relative bias in parameter estimation, across 5 different methods.
Effect of sample size: (left) $n_1 = 200$ (right) $n_2 = 1000$
(10% of missingness entry-wise, MCAR mechanism, $\times$100 replications for each setting)

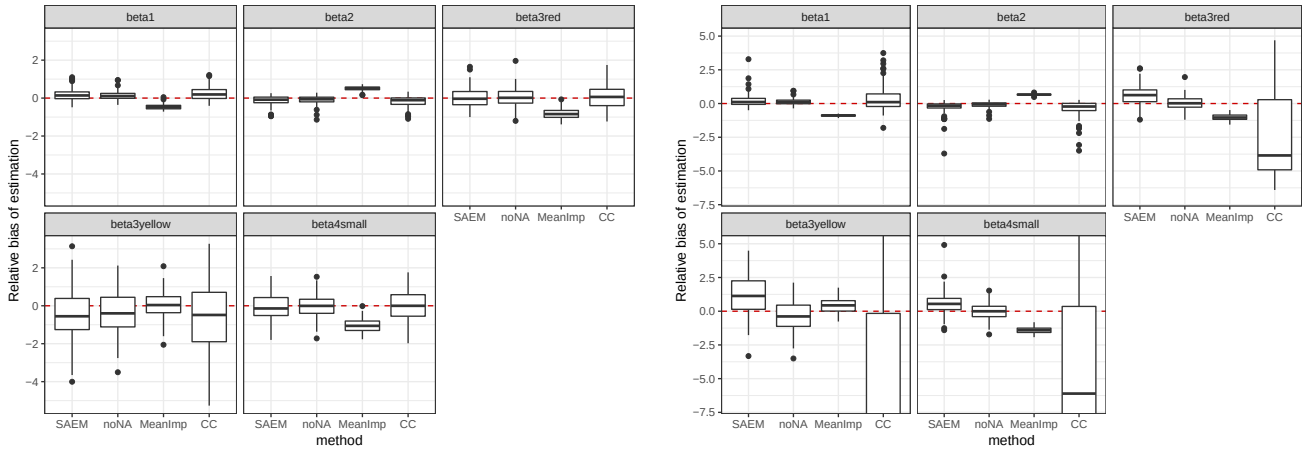


Figure 2.21: Empirical distribution of the relative bias in parameter estimation, across 5 different methods.
Effect of mechanism: (left) MCAR (right) MAR with missingness in (Z_2, U_1, U_2) depending on Z_1
($n_1 = 200$, 10% of missingness entry-wise, $\times$100 replications for each setting)

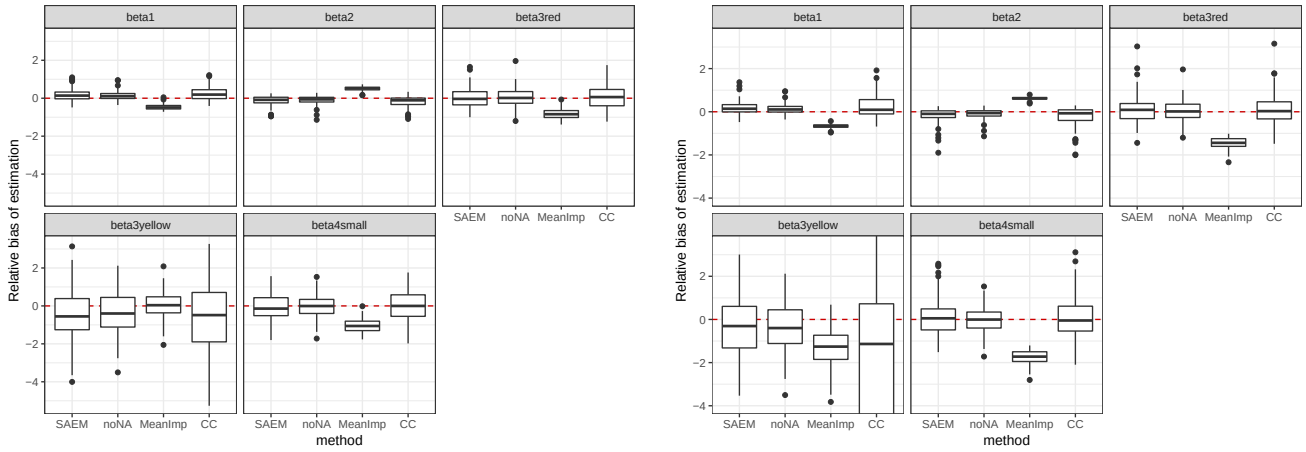


Figure 2.22: Empirical distribution of the relative bias in parameter estimation, across 5 different methods.
Effect of mechanism: (left) MCAR (right) MAR with missingness in Z_2 depending on Z_1, and in U_2 depending on (Z_1, U_1)
($n_1 = 200$, 10% of missingness entry-wise, $\times 100$ replications for each setting)

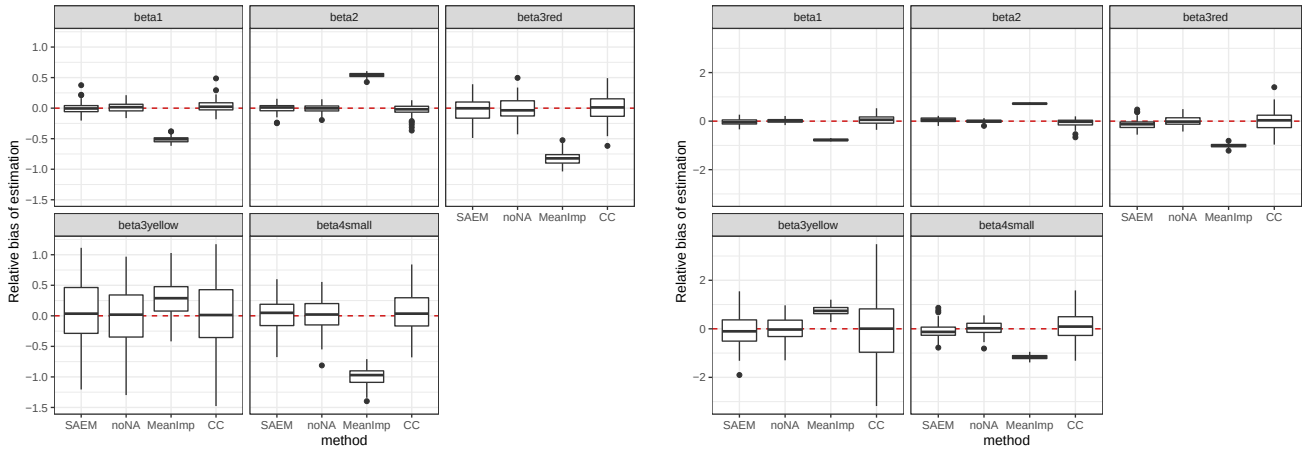


Figure 2.23: Empirical distribution of the relative bias in parameter estimation, across 5 different methods.
Effect of percentage of missingness: (left) 10% entry-wise (right) 30% entry-wise
($n_2 = 1000$, MCAR mechanism, $\times 100$ replications for each setting)

Chapter 3

ABSLOPE—High-dimensional model selection with missing values

Contents

3.1 Introduction

In this chapter, we focus on high-dimensional variable selection with missing values. In particular we are interested in methods that control the FDR. As introduced in Section 1.3, controlling FDR is one of the central goals of many methodological developments in multiple regression (see *e.g.* Barber et al. (2015); Candes 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. Sorted l_1 penalization estimates (SLOPE) is suggested by Bogdan et al. (2015) for the purpose, however, large amounts of shrinkage, needed to keep FDR small, result in large estimation bias of important regression coefficients and thereby poor estimation. On the other hand, when the data contain missing values, to the best of our knowledge, no method exists so far to control FDR on the same time.
To improve the estimation when controlling FDR and to deal with missing values simultaneously, we propose here the adaptive Bayesian version of SLOPE (ABSLOPE) addresses these issues by incorporating aspects of the Spike-and-Slab LASSO and 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 hyper-parameters of our mixture SLOPE prior are iteratively updated using the full Bayesian model in the spirit of stochastic approximation EM (Lavielle, 2014), which can also handle missing data.

Our aim is to develop a complete and efficient methodology for selection of variables with high dimensional data and missing values. The methodology has been implemented in an R (R Core Team, 2017) package `ABSLOPE` (Jiang et al., 2019), which we introduce in details later in Chapter 5. The code that reproduces all our experiments is available from GitHub (Jiang, 2019a).

This chapter is organized as follows: Section 3.2 introduces notation and assumptions about our ABSLOPE model. Section 3.3 describes the stochastic approximation EM algorithm (and its simplified variant) for processing missing data. Section 3.4 evaluates the methodology with a comprehensive simulation study focusing on power, FDR and estimation bias. In Section 3.5, 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 3.6 concludes our work with a discussion.

3.2 Statistical model and assumptions

Let $y = (y_i, 1 \leq i \leq n)$ be a vector of n responses, centered such that $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 0$; and let $X = (X_{ij}, 1 \leq i \leq n, 1 \leq j \leq p)$ be 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 β based on realizations y from the linear regression model:

$$y = X\beta + \varepsilon,$$

where $\beta = (\beta_j, 1 \leq j \leq p)$ is the vector of regression coefficients of length p, for which we assume a sparse structure, and ε is a vector of length n of independent Gaussian errors with mean 0 and variance σ^2, *i.e.* $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_n)$.

3.2.1 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:

$$\hat{\beta}_{\text{SLOPE}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|y - X\beta\|^2 + \sigma \sum_{j=1}^{p} \lambda_j |\beta|_{(j)} \right\} , \tag{3.1}$$

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

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

where $r(\beta, j) \in \{1, 2, \cdots, p\}$ is the rank of β_j among elements in β in a descending order. To solve the convex but non-smooth optimization problem (3.1), 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

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

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

3.2.2 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 λ, which regulates both model selection and shrinkage. Simulation results reported in Bogdan et al. (2015) show that the selection of λ 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 spike component is assigned a large penalty λ_0 (small variance) to weed out noise, while the slab component has a small penalty λ_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 λ_0 relates to the prior mixing weight θ and should reflect the inherent sparsity of the signal (Ročková et al., 2018). The SSL procedure does not choose a single value λ_0 but, similarly as the LASSO, creates a solution path indexed by increasing values of λ_0. 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.

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 with the SLOPE slab prior 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 β 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 β is formally defined as:

$$\mathrm{p}(\beta \mid \gamma, c, \sigma^2; \lambda) \propto c^{\sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)} \prod_{j=1}^{p} \exp\left\{-w_j|\beta_j|\frac{1}{\sigma}\lambda_{r(W\beta,j)}\right\}. \tag{3.2}$$

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 β_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 β_j, *i.e.*, θ indicates the level of sparsity. We assume that θ 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

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

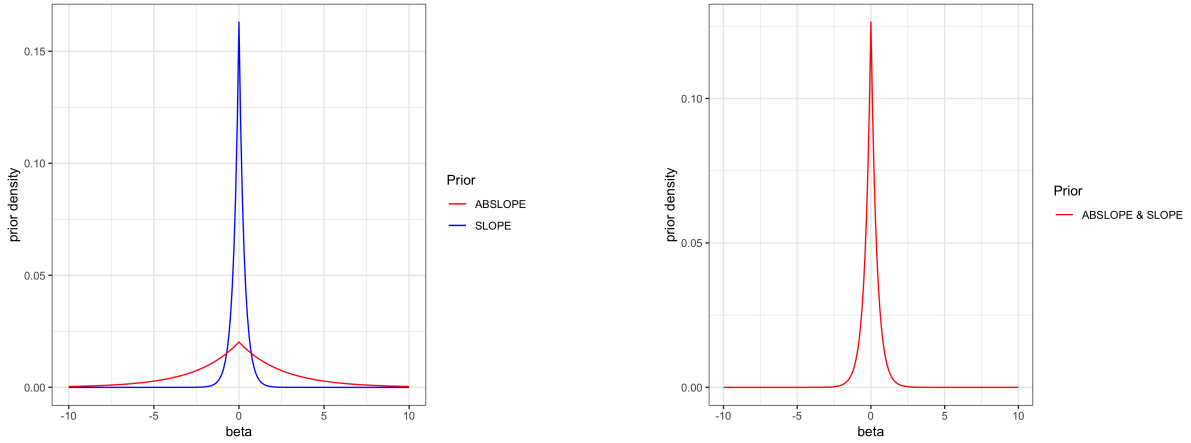
5. For the case when the noise variance σ is unknown, we assume an uninformative prior $\mathrm{p}(\sigma^2) \propto \frac{1}{\sigma^2}$.

3.2.3 Motivation

In Section 3.7.1 it is proved that the prior (3.2) leads to the regular SLOPE prior on the transformed parameter vector $z = W\beta$, i.e.

$$\mathrm{p}(z \mid \sigma^2; \lambda) \propto \prod_{j=1}^{p} \exp\left\{-\frac{1}{\sigma}\lambda_{r(z,j)}|z_j|\right\} \ , \tag{3.3}$$

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 β under the ABSLOPE prior (3.2) can be obtained as a solution to SLOPE (3.1) 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 which will assign substantially larger shrinkage to smaller effects. This is different from the SSL prior, which would shrink all the noise coefficients equally by λ_0. 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 β_j will be assigned a penalty $c\lambda_{r(W\beta,j)}$ that is smaller than $\lambda_{r(\beta,j)}$ obtained under the regular SLOPE. As a result, this



(a) Non-null β (b) Null β

Figure 3.1: Prior distribution of SLOPE and ABSLOPE, on β whose true value is non-null (a) or null (b).

adaptive version is poised to yield more accurate estimation since the l_1 penalty on true signals will be much smaller.

Figure 3.1 shows the difference between the SLOPE prior and the ABLSOPE prior on a single coefficient β_j. On the left, we have a slab prior distribution on an active coefficient β_j which shows that ABSLOPE promotes larger estimates: the mass is greater in the tails compared to SLOPE. On the other hand, for the irrelevant β_j (spike prior depicted on the right), ABSLOPE reduces to the double exponential SLOPE peak to threshold out small effects.

The ABSLOPE prior can be seen as a spike-and-slab prior, where the spike component models regression coefficients close to the noise level and the slab component models large regression coefficients. In fact, the spike-and-slab LASSO prior can be regarded as a special case when one considers the constant sequence of tuning parameters $\lambda_1 = \ldots = \lambda_p = \lambda_0$ for the spike SLOPE component and c as the ratio between spike and slab penalties. The algorithm described in Section 3.3.4 shows that the slab component is destined to de-bias the large regression coefficients while the spike component is aimed at FDR control.

3.2.4 Scaling with existence of missingness

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

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

Missing values are assumed to be either MCAR or MAR as defined in Chapter 1, which allows to derive MLE by ignoring the missing values mechanism and maximizing the observed-data likelihood. Since the covariates should be standardized (as we assumed at the beginning of Section 3.2), we have to reconsider our scaling of X in the light of missing data. When the missing values are MCAR, scaling can be performed as a pre-processing step before performing the analysis. Since observed values represent a random sample from the population, standard

deviations estimated using observed data are unbiased estimates of the population standard deviation even if their variance is larger. When the missing data are MAR, standard deviations estimated using observed data can be severely biased. Indeed, consider the case when 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 downwards. Consequently, its estimation needs to be included in the analysis. In Section 3.7.2, we detail how we update mean and standard deviation at each iteration of the algorithm presented in Section 3.3.

3.2.5 Overview of modeling

Figure 3.2 shows our ABSLOPE graphical model with variables, parameters and their relations. We aim at estimating β and σ^2, treating parameters μ and Σ as nuisance.

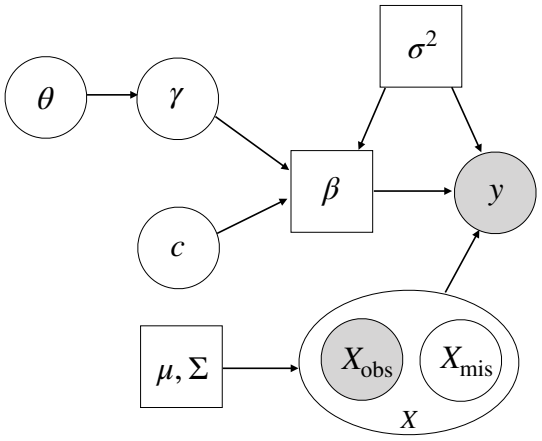


Figure 3.2: ABSLOPE graphical model. Arrows indicate dependencies. White circles are for latent variables, gray ones for observed variables and squares for parameters.

3.3 Model selection by ABSLOPE

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.

3.3.1 Maximizing the observed penalized likelihood

According to the model defined in Section 3.2 and presented in Figure 3.2, the penalized complete-data log-likelihood can be written as:

$$\begin{aligned}\ell_{\text{comp}} &= \log \mathrm{p}(y, X, \gamma, c;\, \beta, \theta, \sigma^2) + pen(\beta) \\ &= \log\left\{\mathrm{p}(X \mid \mu, \Sigma)\, \mathrm{p}(y \mid X;\, \beta, \sigma^2)\, \mathrm{p}(\gamma \mid \theta)\, \mathrm{p}(c)\right\} + pen(\beta) \\ &= -\frac{1}{2}\log(2\pi|\Sigma|) - \frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) - n\log(\sigma) - \frac{1}{2\sigma^2}\|y - X\beta\|^2 \\ &\quad + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1)\log\theta + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 0)\log(1-\theta) - \frac{1}{\sigma}\sum_{j=1}^{p} w_j|\beta_j|\lambda_{r(W\beta,j)}.\end{aligned} \tag{3.4}$$

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_{mis}, γ and c, we estimate β 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 β, and in the meantime, obtain simulated γ to distinguish the true signals from the noise, *i.e.* to select variables. Given the initialization, each iteration t updates β^t to β^{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.*,

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

  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

    $$\mathrm{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^{t-1}, \sigma^{t-1}, \mu^{t-1}, \Sigma^{t-1})\ ; \tag{3.5}$$

  - *Stochastic approximation:* update function Q with

    $$Q^t = Q^{t-1} + \eta_t\left(\ell_{\text{comp}}\Big|_{X^t_{\text{mis}}, \gamma^t, c^t, \theta^t} - Q^{t-1}\right)\ , \tag{3.6}$$

    where η_t is the step-size.

  The step-size (η_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 Σ^{t+1} is estimated as above only when $p << n$. Otherwise we consider a shrinkage estimation as discussed in Remark 1. Indeed, we regard (μ, Σ) 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 (3.4) can be decomposed into several terms: one term for the linear regression part, one term for the covariates distribution and terms for the latent variables γ and c, as illustrated in Figure 3.2. Consequently, one iteration can be divided into tractable sub-problems, as detailed in the following subsections.

3.3.2 Simulation step: sampling the latent variables

To perform the simulation step (3.5), 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:

$$\begin{cases} \gamma \sim Bin\left(\frac{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)+\theta c\exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}\right) ; \\ \theta \sim Beta\left(a+\sum_{j=1}^{p}\mathbb{1}(\gamma_j=1), b+\sum_{j=1}^{p}\mathbb{1}(\gamma_j=0)\right), \text{ with } Beta(a,b) \text{ a prior for } \theta ; \\ c \sim Gamma\left(1+\sum_{j=1}^{p}\mathbb{1}(\gamma_j=1), \quad \frac{1}{\sigma}\sum_{j=1}^{p}|\beta_j|\lambda_{r(W\beta,j)}\mathbb{1}(\gamma_j=1)\right) \text{ truncated to } [0,1]. \end{cases} \tag{3.7}$$

The detailed calculation and interpretation can be found in Section 3.7.3. In addition, to simulate the missing values X_{mis}, we perform a decomposition:

$$\begin{aligned} X_{\text{mis}} &\sim \mathrm{p}(X_{\text{mis}} \mid \gamma, c, y, X_{\text{obs}}, \beta, \sigma, \theta, \mu, \Sigma) \\ &= \mathrm{p}(X_{\text{mis}} \mid y, X_{\text{obs}}, \beta, \sigma, \mu, \Sigma) \\ &\propto \mathrm{p}(y \mid X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma)\, \mathrm{p}(X_{\text{mis}} \mid X_{\text{obs}}, \mu, \Sigma) \ . \end{aligned} \tag{3.8}$$

Here, we observe that the target distribution (3.8) 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_{obs} and x_{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 Σ^{-1} and β_{obs} the observed elements of β.
Let $\tilde{\mu} = (\tilde{\mu}_i)_{i\in\mathcal{M}}$ be the solution of the following system of linear equations:

$$\frac{r\beta_i/\sigma^2 + m_i - u_i}{\tau_i} - \sum_{j\in\mathcal{M}, j\neq i} \frac{\beta_i\beta_j/\sigma^2 + s_{ij}}{\tau_i\tau_j}\tilde{\mu}_j = \tilde{\mu}_i \ , \quad \text{for all } i \in \mathcal{M} \ , \tag{3.9}$$

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 Section 3.7.4.*

3.3.3 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.

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 β.

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

   where $X^t = (X_{\text{obs}}, X^t_{\text{mis}})$. This estimate corresponds to the solution of SLOPE, given the value of W, X_{mis} and σ. 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 then 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. Update σ.

$$\sigma^t = \arg\max_{\sigma} Q^t_2(\sigma) := -n\log(\sigma) - \frac{1}{2\sigma^2}\|y - X^t\beta^t\|^2 - \frac{1}{\sigma}\sum_{j=1}^{p} w^t_j|\beta^t_j|\lambda_{r(W^t\beta^t,j)} \ .$$

Given by the derivative, the solution to estimate σ is:

$$\sigma^t = \frac{1}{2n}\left[\sum_{j=1}^{p}\lambda_{r(W^t\beta^t,j)}w_j^t|\beta_j^t| + \sqrt{\left(\sum_{j=1}^{p}\lambda_{r(W^t\beta^t,j)}w_j^t|\beta_j^t|\right)^2 + 4n\text{RSS}}\right] , \quad (3.10)$$

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

If we omit the penalization term, (3.10) amounts to $\sigma^t = \sqrt{\frac{RSS}{n}}$, which is the classical formula for MLE of σ when β 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 the simulation results in Subsection 3.4.2, often leads to a less biased estimator when most of the true effects are detected by ABSLOPE.

3. Update μ, Σ:

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

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

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

In high dimensional setting, estimation of Σ^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.:*

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

The method boils down to shrinking empirical eigenvalues towards their mean. The parameters ρ_1 and ρ_2 are chosen with asymptotically (as n and p go to infinity) uniformly minimum quadratic risk in its class.

General step-size

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

$$\psi^{t+1} = \psi^t + \eta_t\left[\hat{\psi}_{MLE}^t - \psi^t\right] , \quad (3.11)$$

where $\hat{\psi}_{MLE}^t$ 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 3.3.3 when $\eta_t = 1$. In other words, we apply stochastic approximations on the model parameters, instead of directly operating on the likelihood in (3.6). When the likelihood (3.4) is a linear function of the parameters, the stochastic approximation step in equation (3.6) corresponds exactly to our proposal (3.11). In other situations, it gives good results from an empirical point of view.

3.3.4 SLOBE: Quick version of ABSLOPE

The implementation of SAEM, as described in Subsection 3.3.2 and 3.3.3, can still be costly in terms of computation time, even if the terms of the likelihood decompose well and we use the approximation (3.11). We therefore propose a simplified version of the algorithm, called SLOBE, which instead of drawing samples $(X_{\text{mis}}^t, \gamma^t, c^t, \theta^t)$ from their conditional distribution (3.5) in the simulation step, approximates them by their conditional expectation, i.e.,

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

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 γ_j by:

$$\begin{aligned} \pi &:= \mathbb{E}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) = p(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) \\ &\overset{(3.7)}{=} \frac{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right) + \theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)} \; . \end{aligned} \tag{3.12}$$

2. Approximate θ by:

$$\mathbb{E}(\theta \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, c, \mu, \Sigma, W) = \mathbb{E}(\theta \mid \gamma, \beta, \sigma, W) \overset{(3.7)}{=} \frac{a + \sum_{j=1}^p \mathbb{1}(\gamma_j = 1)}{a + b + p} \; , \tag{3.13}$$

   where a and b are fixed parameters in the prior of θ.

3. Approximate c by:

$$\mathbb{E}(c \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma, W) \overset{(3.18)}{=} \frac{\int_0^1 x^{a'} \exp(-b'x)dx}{\int_0^1 x^{a'-1} \exp(-b'x)dx} \; , \tag{3.14}$$

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

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

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

   which is provided by Proposition 2.

Then, in step M, we maximize the likelihood of the complete data, as in Subsection 3.3.3. 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 Subsection 3.4.5. We provide a summary of ABSLOPE and SLOBE methods in Section 3.7.5.

3.4 Simulation study: FDR and Power

3.4.1 Simulation setting

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

1. A design matrix $X_{n\times p}$ is generated from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. The matrix is standardized, s.t., the mean of each column is 0 and its ℓ_2-norm is 1.

2. The signal magnitude is $c_0\sqrt{2\log p}$[1] when c_0 is large the signal strength is stronger. Only k on the p predictors are non-zero and all equal to $c_0\sqrt{2\log p}$.

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

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

We set the initialization and the hyperparameters as follows.

Initialization Section 3.7.6 provides the default values we have taken for the following simulation studies. The algorithm is not sensitive to the choice of values a and b (3.12), but initial values for β may have a stronger impact. In practice, we use the LASSO estimates based on preliminary mean imputation (missing values replaced by the average of the observed values for each variable) to initialize the coefficients.

Step-size We set $\eta_t = 1$ for the first $t_0 = 20$ iterations to approach the neighborhood of the MLE, then, choose a positive decreasing sequence $\eta_t = \frac{1}{t-t_0}$ to approximate the MLE, with the stochastic approach formula (3.11).

λ sequence A sequence of penalty coefficients λ must be chosen before implementing the algorithm. As introduced in the Subsection 3.2.1, we use a BH sequence inspired by orthogonal designs:

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

3.4.2 Convergence of SAEM

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

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

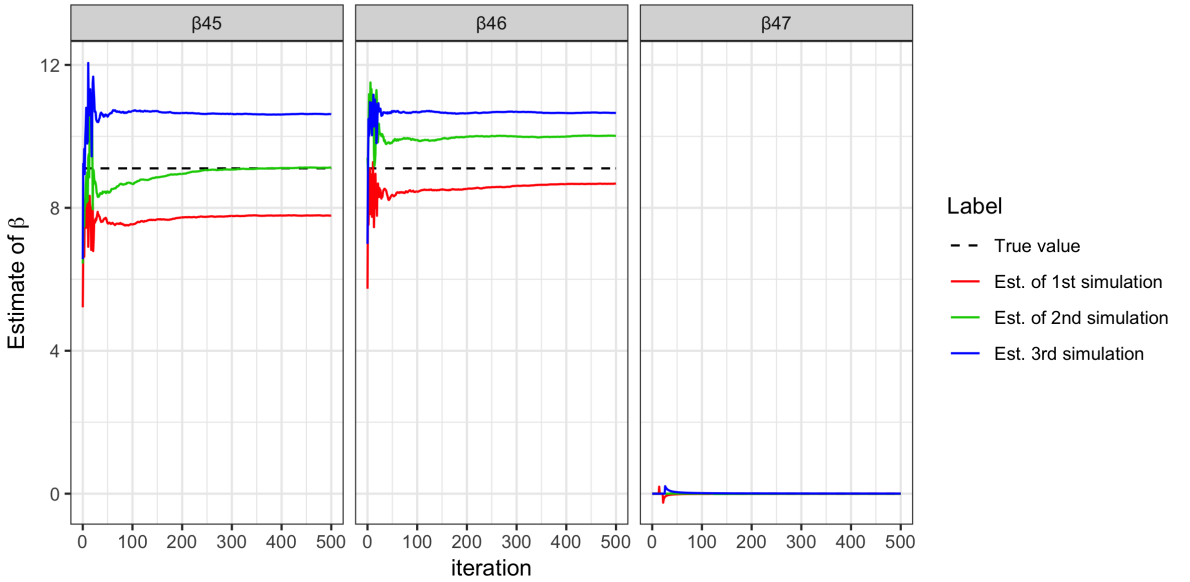


Figure 3.3: Convergence plots for three coefficients with ABSLOPE (colored solid curves). Black dash lines represent the true value for each β. Estimates obtained with three different sets of simulated data are represented by three different colors.

Figure 3.3 shows the convergence of some coefficients with SAEM for three simulated data sets. These graphs are representative of all the observed results. There are large fluctuations during the first t_0 =20 iterations, then after introducing the stochastic approximation at the 20th iteration, convergence is achieved gradually. Due to the existence of a sorted l_1 penalty, the estimates are still slightly biased.

In addition, we also represent the convergence curves for σ with ABSLOPE in Section 3.7.7 in order to compare the estimate of σ by ABSLOPE to the biased MLE estimator without prior knowledge, *i.e.*, $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{RSS}{n}}$. We can see that the estimates of σ with both methods are biased downward, but since ABSLOPE has an additional correction term (3.10), it leads to a less biased estimator.

3.4.3 Behavior of ABSLOPE - SLOBE

We then evaluate ABSLOPE and SLOBE in a different parametrization setting to see how the signal strength, sparsity and other parameters influence their performance.

Criterion We apply ABSLOPE or SLOBE on a synthetic dataset and get estimates for $\hat{\beta}$ and the sampled $\hat{\gamma}$ indicating the model selection results. 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 $= \frac{FP}{FP+TP}$;

- MSE of β (Relative l_2 norm error) $= \frac{\|\hat{\beta}-\beta\|^2}{\|\beta\|^2}$;
- Relative prediction error $= \frac{\|X\hat{\beta}-X\beta\|^2}{\|X\beta\|^2}$.

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

Scenario 1

We first consider $n = p = 100$ and vary:

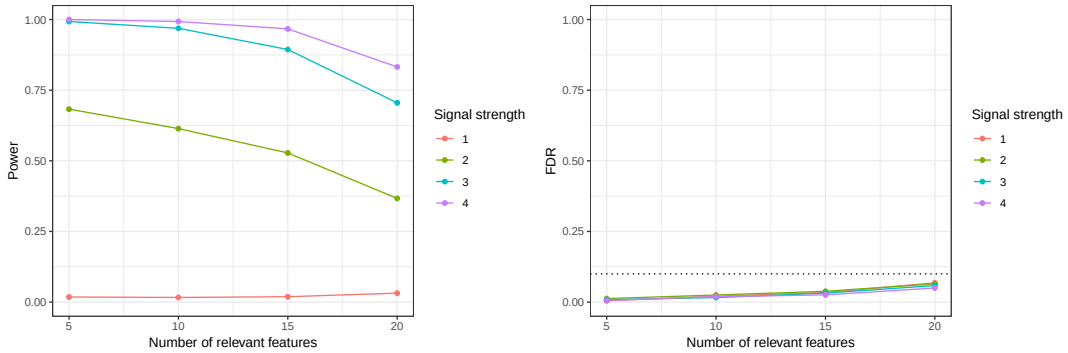
- sparsity: number of true signal $k = 5,\ 10,\ 15,\ 20$;
- signal strength $\sqrt{2\log p}\,, 2\sqrt{2\log p}\,, 3\sqrt{2\log p}\,, 4\sqrt{2\log p}$;
- percentage of missingness $0.1,\ 0.2,\ 0.3$, generated randomly, i.e., MCAR;
- correlation between covariates $\Sigma = \text{toeplitz}(\rho)$[2] where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

Then we applied the Algorithm 5 on each synthetic dataset.

Results 1: no correlation, 10% missingness - vary signal strength According to Figure 3.4:

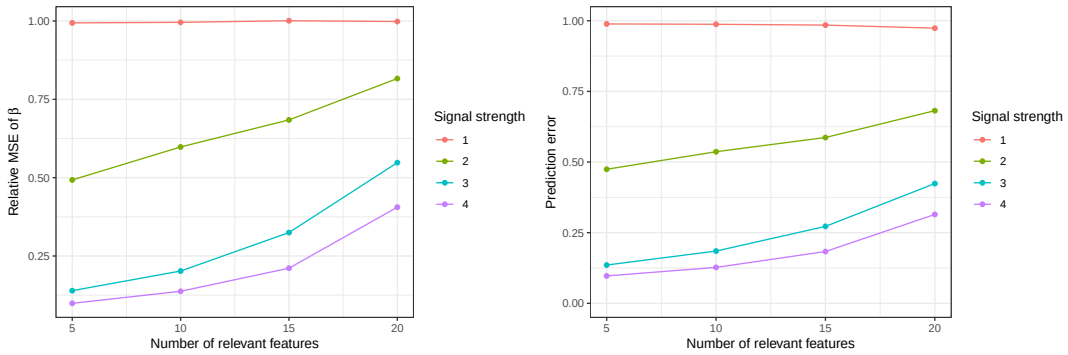
- We observe that FDR is always controlled at the expected level 0.1.
- Power increases and estimation bias decreases with larger sparsity or stronger signal.
- When the signal is too weak (signal strength $= \sqrt{2\log p}$), the power is near 0, which is due to the identifiablility issue that ABSLOPE cannot distinguish the signal from the noise. Indeed, the value $c = \frac{\lambda_1}{\sigma\sqrt{2\log p}}$ is greater than one where λ_1 is the largest penalization coefficient. In addition, the bias is significant. This behaviour can be explained by the fact that we choose the penalty λ to reduce the noise σ; but when the signal is as weak as σ, this choice of λ also "kills" the real signal.

[2]The Toeplitz structure (or auto-regressive structure) for correlation has been introduced for microarry study (Guo et al., 2006), with the form: $\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho^{p-2} & \rho^{p-1} \\ \rho & 1 & \ddots & \cdots & \rho^{p-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{p-2} & \cdots & \ddots & \ddots & \rho \\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1 \end{pmatrix}_{p\times p}$, where $\rho \in [0,1]$ is a constant. For the Toeplitz structure, adjacent pairs of covariates are highly correlated and those further away are less correlated, as in microarry study, genes are correlated due to their distance in the regularity pathway.



(a) Power

(b) FDR

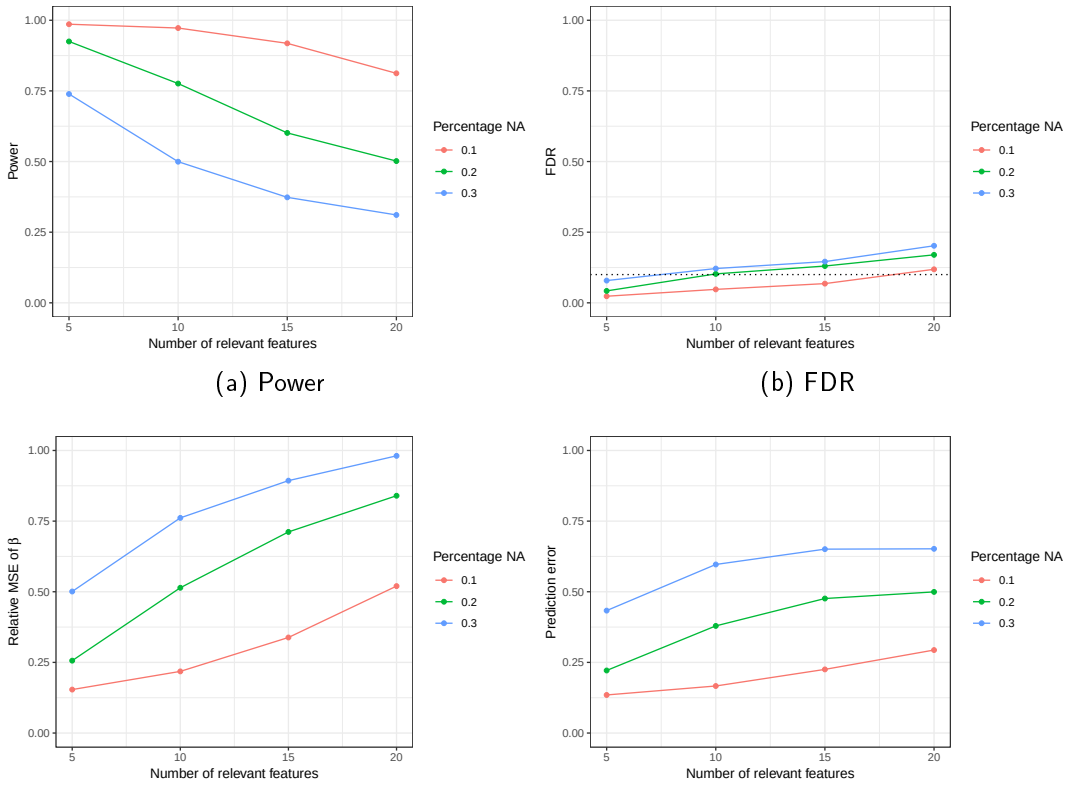


(c) Bias of β

(d) Prediction error

Figure 3.4: Mean of power (a), FDR (b), bias of the estimate for β (c) and prediction error (d), as function of length of true signal, over the 200 simulations. Results for $n = p = 100$, percentage of missingness 10% and Σ orthogonal (no correlation).

Results 2: with correlation, strong signal - vary percentage of missingness Now we add the correlation as $\Sigma = \text{toeplitz}(\rho)$ where $\rho = 0.5$, and also fix a strong signal strength as $3\sqrt{2\log p}$. We then vary the sparsity and percentage of missingness. The results in Figure 3.5 show that:



(a) Power (b) FDR

(c) Bias of β (d) Prediction error

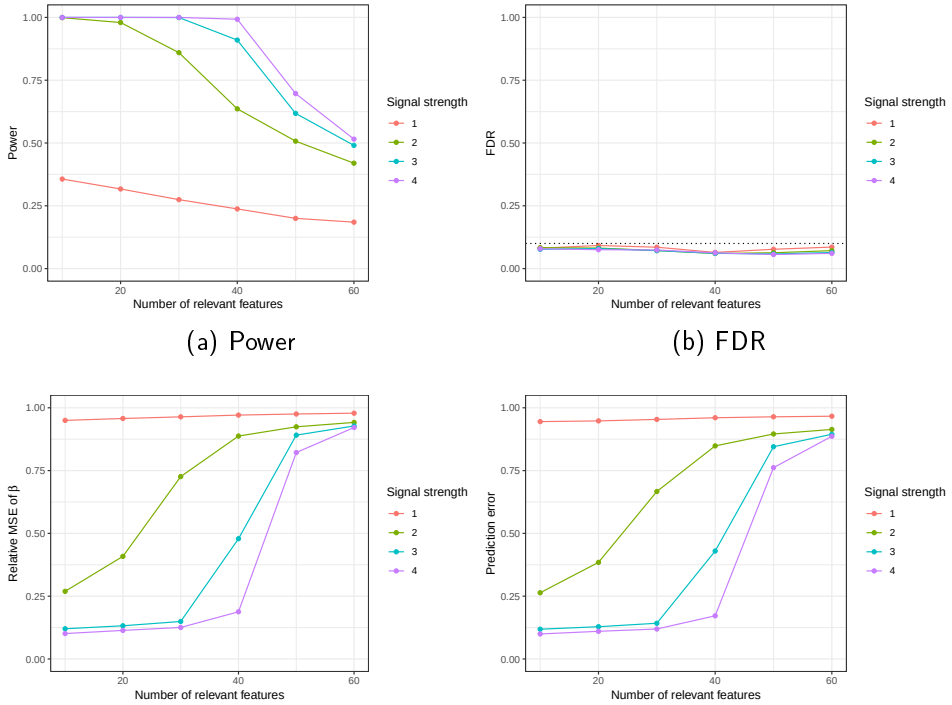
Figure 3.5: Mean of power (a), FDR (b), bias of the estimate for β (c) and prediction error (d), as function of length of true signal over the 200 simulations. Results for $n = p = 100$, with correlation and strong signal.

- The power increases and the estimation bias decreases when the percentage of missing data decreases.
- In the presence of correlation, the FDR control is slightly lost when the number of non-zero coefficients is greater than 10 and the percentage of missing values exceeds 0.2, but is still near the nominal level.

Scenario 2

Now we consider a larger dataset $n = p = 500$ and vary the same parametrization as in Subsection 3.4.3, except the sparsity, for which we take wider range of choices among $k = 10, 20, 30, \cdots, 60$. In this scenario of larger dimension, we have applied the simplified SLOBE algorithm as described in Subsection 3.3.4 to avoid intensive computation.

Results 1: no correlation, 10% missingness - vary signal strength According to Figure 3.6:



(a) Power (b) FDR

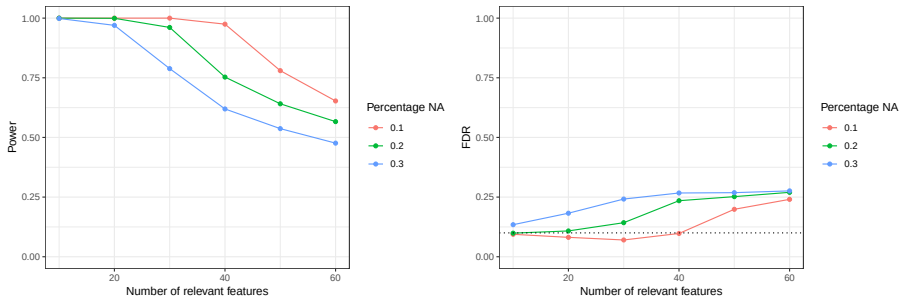
(c) Bias of β (d) Prediction error

Figure 3.6: Mean of power (a), FDR (b), bias of the estimate for β (c) and prediction error (d), as function of length of true signal, over the 200 simulations. Results for $n = p = 500$, percentage of missingness 10% and Σ orthogonal (no correlation).

- FDR is always controlled at expected level 0.1.
- Similar to Figure 3.4, power increases and estimation error decreases with larger sparsity and stronger signal. However in this larger dimension case, we can handle with larger number of relevant features until 30 or 40, at which we observe a phase transition due to the identifiability issue.

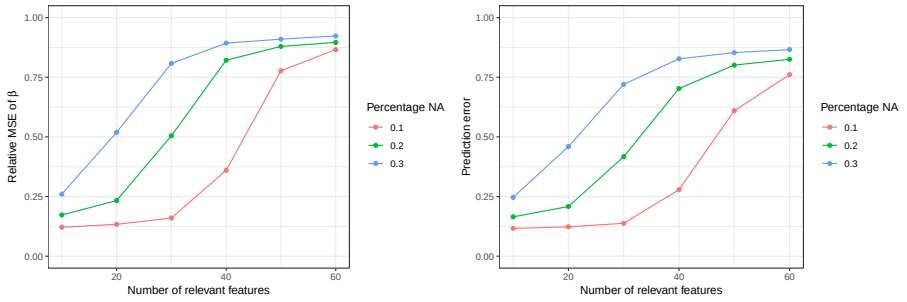
Results 2: with correlation, strong signal - vary percentage of missingness Now we add the correlation as $\Sigma = \text{toeplitz}(\rho)$ where $\rho = 0.5$, and also fix a strong signal strength as $3\sqrt{2 \log p}$. We then vary the sparsity and percentage of missingness. The results in Figure 3.7 show that:

- Similar to Figure 3.5, the power increases and the estimation error decreases when the percentage of missing data decreases.
- Due to the existence of correlation, the FDR control is slight lost, especially in the less sparse and more missing case.



(a) Power

(b) FDR



(c) Bias of β (d) Prediction error

Figure 3.7: Mean of power (a), FDR (b), bias of the estimate for β (c) and prediction error (d), as function of length of true signal over the 200 simulations. Results for $n = p = 500$, with correlation and strong signal.

- With 10% missing values, if the number of relevant features is below 40, then we can always achieve an efficient power and perfect FDR control. With larger percentage of missing values, the sparsity of this changing point will be more conservative.

In addition, we present the results varying the correlations in Section 3.7.8.

3.4.4 Comparison with competitors

We use the same simulation scenario and criteria as those used in Subsection 3.4.3 to compare ABSLOPE and SLOBE to other approaches that can be considered to select variables in the presence of missing data.

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

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

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

We present the results for the case $n = p = 100$ in Section 3.7.9 while Figure 3.8 summarizes the result for the case $n = p = 500$, 10% missingness and with correlation toeplitz(0.5). Lighter colors indicate smaller values.

- ABSLOPE and SLOBE both have strong power and accurate prediction, where FDR is always controlled.
- The power and FDR control achieved by ABSLOPE and SLOBE are better than the case $n = p = 100$. On one hand, correlation helps the generation of missing values. On the other hand, sparsity considered here is less complicated.
- Other methods pay the price of FDR control to achieve good power.

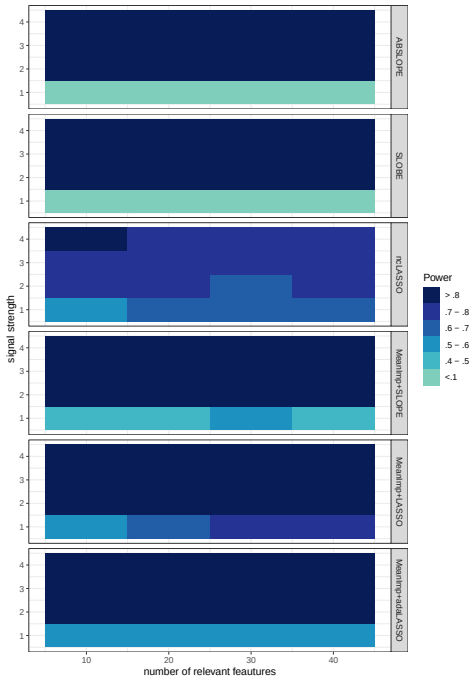
3.4.5 Comparison of computation time

Table 3.1 presents the execution time of the different methods considered in the simulation. In addition, we have implemented our proposed algorithm in C and we use Rcpp (Eddelbuettel and Balamuta, 2017) to integrate these functions within R. In the case $n = p = 100$, we

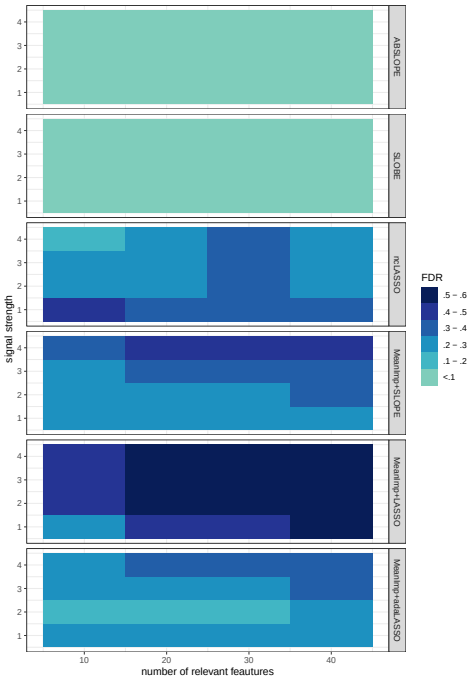
Table 3.1: Comparison of average execution time (in seconds) for one simulation, in the case without correlation and with 10% MCAR, for $n = p = 100$ and $n = p = 500$ calculated over 200 simulations. (MacBook Pro, 2.5 GHz, processor Intel Core i7)

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

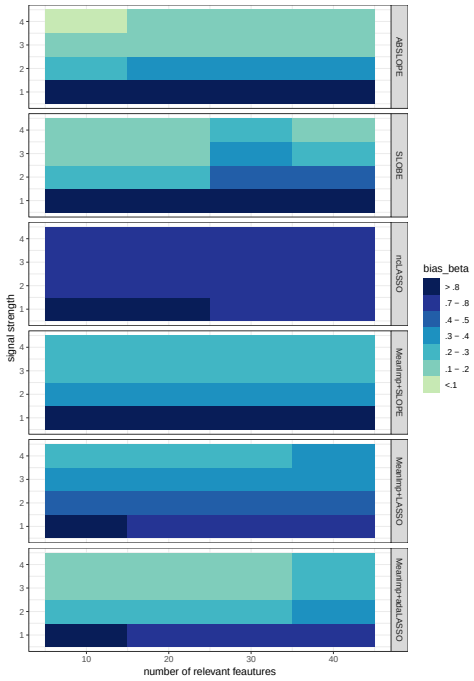
observe that the most time consuming method is ncLASSO, which spent on average 20



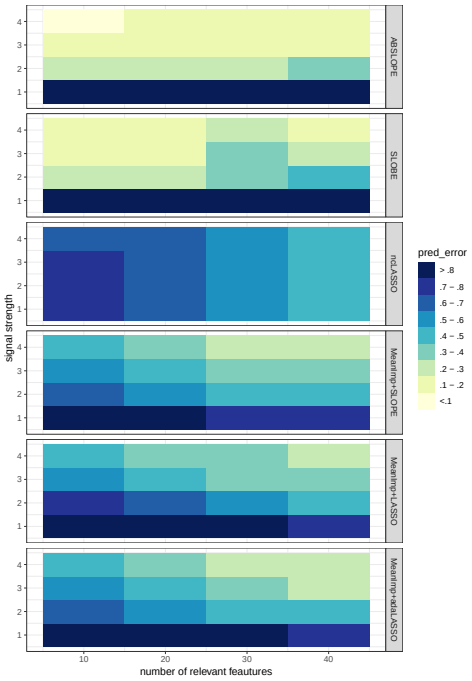
(a) Power



(b) FDR



(c) Bias of β



(d) Prediction error

Figure 3.8: Comparison of power (a), FDR (b), bias of β (c) and prediction error (d) with varying sparsity and signal strength, with 10% missingness over 200 simulations in the case with correlation.

seconds on one simulation. While ABSLOPE also took on average 14 seconds for one run, its simplified version SLOBE reduced this cost to 0.6 seconds, which is comparable to MeanImp + adaLASSO. While when $n = p = 500$, the convergence of ABSLOPE requires much more time but SLOBE helps to simplify the complexity. In addition, the version of C for SLOBE is more accelerated, saving half of the computation time, which makes SLOBE capable of handling larger datasets.

3.5 Modeling the level of placelet in the TraumaBase context

3.5.1 Details on the dataset and preprocessing

In our analysis on medical dataset TraumaBase as introduced in Section 1.4, 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 Section 3.7.10. These variables were included here because they were all available to the pre-hospital team, and therefore could be used in real situations.

Figure 3.9 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.

3.5.2 Model selection results

As is customary in supervised learning, we divide the dataset into training and test sets. The training set contains a random selection of 80% of observations whereas the test set contains the remaining 20%. In the training set, we select a model and estimate the parameters. We apply ABSLOPE and compare it with the same methods than those described in Section 3.4, namely MeanImp + SLOPE, MeanImp + LASSO, MeanImp + adaLASSO, MeanImp + SSL except ncLASSO since we do not known the sparsity and the l_1 bound of coefficients. Moreover, we also include:

- BIC: Mean imputation followed by a stepwise method based on BIC;

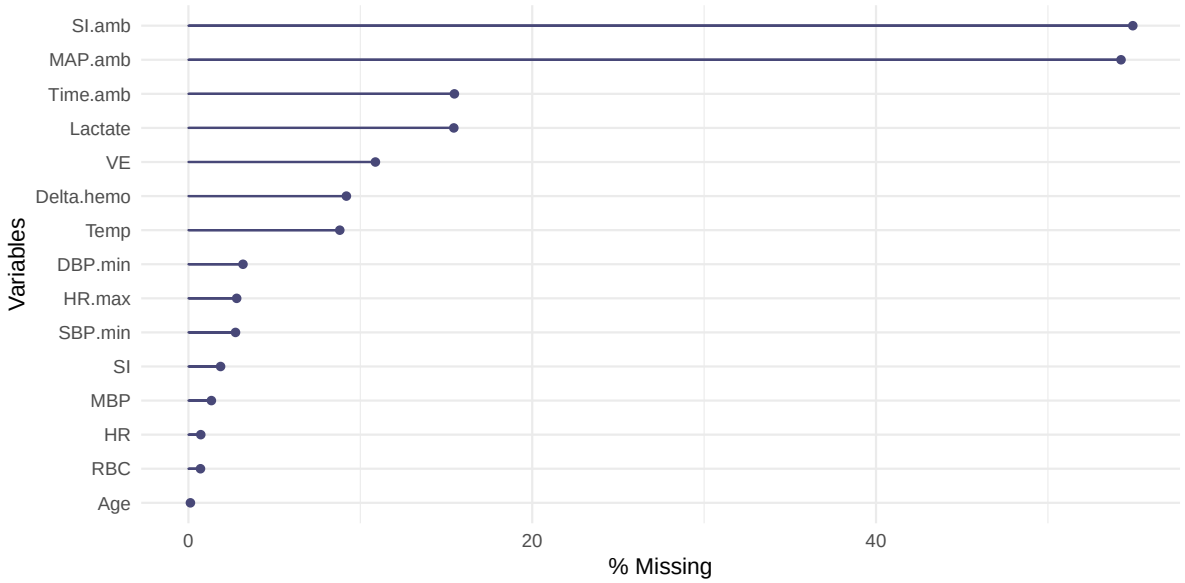


Figure 3.9: Percentage of missing values in each pre-selected variable from TraumaBase.

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

In the SLOPE type methods, we set the penalization coefficient λ 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 3.2 shows that, over 10 replications, how many times each variable is selected. In addition, Table 3.3 reports whether the selected variables by ABSLOPE have on average a positive or negative effect on the platelet.

The TraumaBase medical team indicated that the signs of the coefficients were partially in agreement with their a-priori expectations. Large values of shock index, vascular filling, blood transfusion and lactate give signs of severe bleeding for patients and, thereby, lower levels of platelets. However, the effects of delta Hemocue and the heart rate on the platelet were not entirely in agreement with their opinion.

3.5.3 Prediction performance

In supervised learning, after a model has been fitted on a training set, a natural step is to evaluate the prediction performance on a test set. Assuming an observation $X = (X_{\text{obs}}, X_{\text{mis}})$ in the test set, we want to predict the binary response y. One added difficulty is that the test set also contains missing values, since the training set and the test set have the same distribution (*i.e.*, the distribution of covariates and the distribution of missingness). Therefore, we cannot directly apply the fitted model to predict y from an incomplete observation of the test X.

Our framework offers a natural remedy by marginalizing over the distribution of missing

Table 3.2: Number of times that each variable is selected over 10 replications. Bold numbers indicate which variables are included in the model selected by ABSLOPE.

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

Table 3.3: The effect of the selected variables by *ABSLOPE* on the platelet. "+" indicates positive effect while "−" negative; 0 indicates insignificant variables.

Variable	Effect
Age	−
SI	−
MBP	0
Delta.Hemo	+
Time.amb	0
Lactate	−
Temp	0
HR	+
VE	−
RBC	−
SI.amb	0
MBP.amb	0
HR.max	0
SBP.min	0
DBP.min	0

data, given the observed ones. More precisely, with S Monte Carlo samples $(X_{\text{mis}}^{(s)}, 1 \leq s \leq S) \sim \mathrm{p}(X_{\text{mis}}|X_{\text{obs}})$, we estimate directly the response by maximum a posteriori value:

$$\begin{aligned}\hat{y} = \arg\max_y \mathrm{p}(y|X_{\text{obs}}) &= \arg\max_y \int \mathrm{p}(y|X)\mathrm{p}(X_{\text{mis}}|X_{\text{obs}})dX_{\text{mis}} \\ &= \arg\max_y \mathbb{E}_{\mathrm{p}_{X_{\text{mis}}|X_{\text{obs}}}} \mathrm{p}(y|X) \\ &= \arg\max_y \sum_{s=1}^{S} \mathrm{p}\left(y|X_{\text{obs}}, X_{\text{mis}}^{(s)}\right).\end{aligned}$$

Note that in the literature there are not many solutions to deal with the missing values in the test set (Josse et al., 2019). For those imputation based methods, we imputed the test set with mean imputation and predicted the platelet by $\hat{y} = X^{\text{imp}}\hat{\beta}$. Finally we evaluate the relative l_2 prediction error: $\text{err} = \frac{\|\hat{y}-y\|^2}{\|y\|^2}$. Prediction results obtained are presented in Figure 3.10.

ABSLOPE's performance is comparable to the one of Random Forest and adaptive LASSO, and is slightly better than traditional stepwise regression and LASSO. There is a significant gap between the results of ABSLOPE and those of SLOPE. One of the possible reasons is that the classic version of SLOPE may encounter difficulties in the presence of correlation, while ABSLOPE works well even with correlations (an aspect adopted from the Spike-and-Slab LASSO). Random forests have excellent predictive capabilities which is consistent with the results of Josse et al. (2019) who show good performance of supervised machine learning even in the case of the simple mean imputation. However, it is difficult to interpret results in terms of selected variables, which is often crucial for physicians.

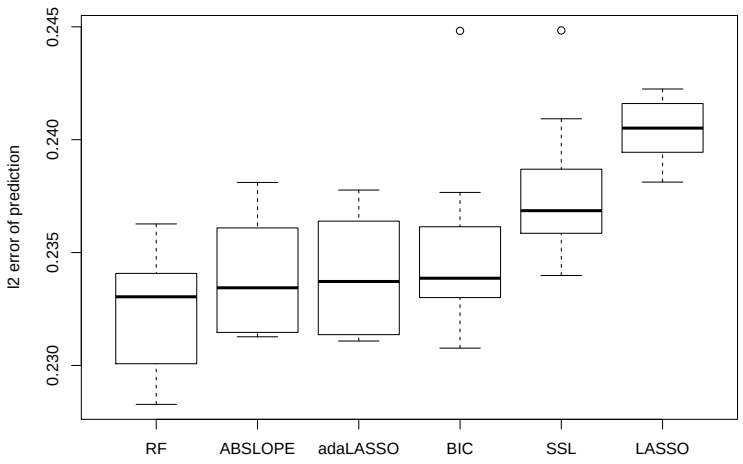


Figure 3.10: Empirical distribution of prediction errors of different methods over 10 replications for the TraumaBase data. Results for SLOPE are not presented due to its large gap compared to others, with a mean of prediction error equals to 0.27.

Figure 3.10 and Table 3.2 show that ABSLOPE and adaLASSO methods, which have the best predictive capabilities, select almost the same variables with adaLASSO selecting MBP (mean blood pressure) and ABSLOPE selecting SI (shock index). These two variables are highly correlated since both are measurements based on the systolic blood pressure.

Finally, we also performed ABSLOPE on the whole standardized dataset without cross-validation, and the formula of regression with model selection was reported as: $\mathrm{Platelet} = -6.92\mathrm{Age} - 7.28\mathrm{SI} + 6.53\mathrm{Delta.hemo} - 8.87\mathrm{Lactate} + 10.05\mathrm{HR} - 3.96\mathrm{VE} - 8.91\mathrm{RBC} + 3.25\mathrm{SBP.min}$. This selection largely agrees with the results from cross-validation presented in Table 3.2. The coefficient values demonstrate the importance of corresponding variables and provide a medical tool to predict the platelet value for a new patient.

3.5.4 Results with Interactions

We also consider a more complete model by adding second order interactions between the covariates, which increases the dimensionality at $p = 55$. We apply the same procedure as before and report the predictive results in Figure 3.11.

Table 3.4 shows which variables are selected more than 5 times out of the 10 replications. Results for SSL and SLOPE are not presented due to their large gap compared to others, with a mean of prediction error equals to 0.35 and 0.40 respectively; BIC is not shown for this case with interactions, because it's computational heavy for this step-wise method with many variables. The average sizes of the variables set selected by ABSLOPE, LASSO and adaLASSO are respectively 6, 7 and 12.

Again, ABSLOPE provides good results in terms of prediction while being sparse. We observe that when interactions are added, age often appears in combination with other variables. LASSO methods tend to include a larger number of variables with a potentially increased

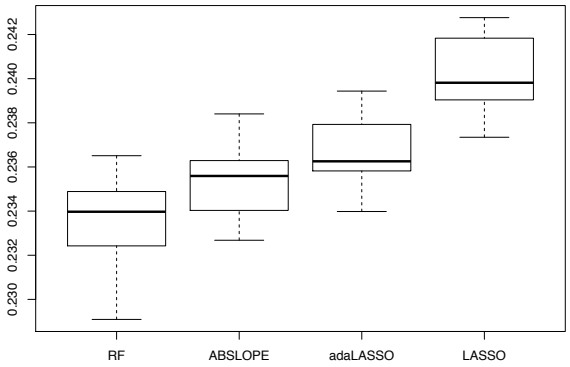


Figure 3.11: Empirical distribution of prediction errors of different methods over 10 replications for the TraumaBase data, with interactions between each pair of variables.

Method	Variables selected
ABSLOPE	Age ∗ MBP.amb, Delta.hemo ∗ Lactate Lactate ∗ RBC, HR ∗ SBP.min
LASSO	RBC, SBP.min, Age ∗ Lactate Age ∗ VE, Delta.hemo ∗ Lactate Delta.hemo ∗ VE , Lactate ∗ RBC
adaLASSO	Age ∗ Time.amb, Age ∗ HR Age ∗ MBP.amb, Age ∗ SBP.min MBP ∗ HR, Delta.hemo ∗ VE Lactate ∗ VE, HR ∗ HR.max HR ∗ SBP.min, VE ∗ RBC

Table 3.4: The variables selected more than 5 times out of the 10 replications, by each method. "∗" indicates the interaction between two variables.

false discovery rate. Note that the prediction properties with interactions are slightly worse than those without interactions, which happened due to the existence of missing values (*e.g.* the interaction term between Age and Lactate will be missing if either of these two variables is unobserved). In conclusion, other methods, apart from ABSLOPE, have a tendency to overfit when interactions are present.

3.6 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.In terms of algorithm efficiency, we can develop a screening rule (Larsson et al., 2020) for ABSLOPE, which allows predictors to be discarded before estimating the model. Alternatively the EMVS procedure of Ročková and George (2014) could be adapted to modify the algorithm, in order to establish a mathematically more rigorous algorithm for the model selection problem with missing values.

3.7 Supplementary materials

3.7.1 Deviation of prior (3.2) started from SLOPE prior

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

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

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 $w_j = c\gamma_j + (1-\gamma_j) = \begin{cases} c, & \gamma_j = 1 \\ 1, & \gamma_j = 0 \end{cases}, \quad j = 1, 2, \cdots, p.$

Then according to the transformation of variables, we have the prior distribution for β:

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

which corresponds to our proposed prior (3.2).

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

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

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

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

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

3.7.3 Details of the simulation step: sampling the latent variables

To perform the simulation step (3.5), 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 γ. According to the dependency between variables presented in Figure 3.2, simulating the element γ_j boils down to:

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

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

$$\begin{aligned}\mathbb{P}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta) &= \frac{\mathbb{P}(\gamma_j = 1 \mid \theta)\mathrm{p}(\beta \mid \gamma_j = 1, \gamma_{-j}, c, \sigma)}{\sum_{\gamma_j \in \{0,1\}} \mathbb{P}(\gamma_j \mid \theta)\mathrm{p}(\beta \mid \gamma_j, \gamma_{-j}, c, \sigma)} \\ &= \left[1 + \frac{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W^0\beta,j)}\right) \times (c)^{\sum_{-j}\mathbb{1}(\gamma_{-j}=1)} \prod_{-j}\exp\left(-w^0_{-j}|\beta_{-j}|\frac{1}{\sigma}\lambda_{r(W^0\beta,-j)}\right)}{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W^1\beta,j)}\right) \times (c)^{\sum_{-j}\mathbb{1}(\gamma_{-j}=1)} \prod_{-j}\exp\left(-w^1_{-j}|\beta_{-j}|\frac{1}{\sigma^t}\lambda_{r(W^1\beta,-j)}\right)}\right]^{-1} \\ &= \left[1 + \frac{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W^0\beta,j)}\right)}{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W^1\beta,j)}\right)} \times \frac{\prod_{-j}\exp\left(-w^0_{-j}|\beta_{-j}|\frac{1}{\sigma}\lambda_{r(W^0\beta,-j)}\right)}{\prod_{-j}\exp\left(-w^1_{-j}|\beta_{-j}|\frac{1}{\sigma}\lambda_{r(W^1\beta,-j)}\right)}\right]^{-1} ,\end{aligned} \tag{3.15}$$

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 (3.15) 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 γ_j from the last iteration, so the difference between the estimates from last and the current iteration will be reduced. Consequently, (3.15) is drawn from:

$$\begin{aligned}\mathbb{P}(\gamma_j = 1 \mid \gamma_{-j}, c, \beta, \sigma, \theta, W) &= \left[1 + \frac{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}\right]^{-1} \\ &= \frac{\theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)}{(1-\theta)\exp\left(-\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right) + \theta c \exp\left(-c\frac{1}{\sigma}|\beta_j|\lambda_{r(W\beta,j)}\right)} ,\end{aligned} \tag{3.16}$$

which can be interpreted as the posterior probability of binary signal indicator for j^{th} variable, given the prior guess $\mathbb{P}(\gamma_j = 1 \mid \theta) = \theta$ and the conditional likelihood of the vector β given $\gamma_j = 1$ and $\gamma_j = 0$, see (3.2).

2. Simulate θ. The update of θ boils down to generate from:

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

where $\mathrm{p}(\gamma \mid \theta)$ is a Bernoulli distribution. In addition, if we also assume a prior for θ 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:

$$Beta\left(a + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1), b + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 0)\right)\ , \tag{3.17}$$

from which we can generate the latent variable θ. The target distribution (3.17) 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.

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

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]:

$$Gamma\left(1 + \sum_{j=1}^{p} \mathbb{1}(\gamma_j = 1),\quad \frac{1}{\sigma} \sum_{j=1}^{p} |\beta_j| \lambda_{r(W\beta, j)} \mathbb{1}(\gamma_j = 1)\right)\ . \tag{3.18}$$

If the signal is strong enough, *i.e.*, β_j is relative large compared to level of noise σ 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.

3.7.4 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_{obs}, and x_{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:

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

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

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

After rearranging terms, with notations:

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

we get:

$$\begin{aligned}
&\mathrm{p}(x^i_{\mathrm{mis}} \mid x_{\mathrm{obs}}, y, \sigma, \beta, \Sigma, \mu, x^{-i}_{\mathrm{mis}}) \\
&\propto \exp\left\{-\frac{1}{2}\left[(x^i_{\mathrm{mis}})^2\left(s_{ii} + \frac{\beta_i^2}{\sigma^2}\right) - 2x^i_{\mathrm{mis}}\left(\frac{r\beta_i}{\sigma^2} + m_i - u_i\right) + 2x^i_{\mathrm{mis}} \sum_{j\in\mathcal{M}, j\neq i}\left(\frac{\beta_i\beta_j}{\sigma^2} + s_{ij}\right)x^j_{\mathrm{mis}}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[x^i_{\mathrm{mis}}\tau_i - \frac{r\beta_i/\sigma^2 + m_i - u_i}{\tau_i} + \sum_{j\in\mathcal{M}, j\neq i} \frac{\beta_i\beta_j/\sigma^2 + s_{ij}}{\tau_i\tau_j} x^j_{\mathrm{mis}}\tau_j\right]^2\right\} \ .
\end{aligned} \tag{3.19}$$

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

$$\mathrm{p}(z_i \mid x_{\mathrm{obs}}, y, \sigma, \beta, \Sigma, \mu, x^{-i}_{\mathrm{mis}}) \propto \exp\left[-\frac{1}{2}\left(z_i - \tilde{\mu}_i + \sum_{j\in\mathcal{M}, j\neq i} B_{ij}\left(z_j - \tilde{\mu}_j\right)\right)^2\right] \ , \tag{3.20}$$

and

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

Combine equations (3.19) and (3.20), 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

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

□

3.7.5 Summary of algorithms

Algorithm 5 Solving ABSLOPE with SAEM.

Input: Initialization β^0, σ^0, c^0, θ^0, X^0_{mis}, μ^0, Σ^0;
for $t = 1, 2, \cdots, \text{Maxit}$ **do**
(Simulation step)

1. Generate γ^t from (3.16);
2. Generate θ^t from Beta distribution (3.17);
3. Generate c^t from truncated Gamma distribution (3.18);
4. Generate X^t_{mis} from Gaussian distribution (3.9);

(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 3.3.3;
2. With step-size $\eta_t = \begin{cases} 1, & \text{if } t \leq 20 \\ \frac{1}{t-20}, & \text{if } t > 20 \end{cases}$, update

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

Update σ, μ and Σ similarly;

if $\|\beta^{t+1} - \beta^t\|^2 < \text{tol}$ **then**
Stop;
Output: Indexes for model selection $\hat{\gamma} \leftarrow \frac{1}{20}\sum_{t'=t-19}^{t} \gamma^{t'}$ (the average of the last 20 iterations), and estimates $\hat{\beta} \leftarrow \beta^t \cdot \hat{\gamma}$.

We propose the ABSLOPE model and solve the problem of the maximization of the penalized likelihood using the SAEM algorithm, described in Algorithm 5. We also give the SLOBE algorithm in Algorithm 6 which is an approximated and accelerated version.

3.7.6 Initialization of ABSLOPE

Here we suggest the following starting values:

- β^0 is obtained from elastic net LASSO (Simon et al., 2011), or Spike and Slab LASSO (Ročková and George, 2018);
- X^0_{mis} are imputed by PCA (imputePCA) (Josse and Husson, 2016), or imputed by the mean of column (imputeMean);
- μ^0 and Σ^0 are estimated with the empirical estimators obtained from the imputed initial data;

Algorithm 6 SLOBE: a quick version of ABSLOPE.

Input: Initialization β^0, σ^0, c^0, θ^0, X^0_{mis}, μ^0, Σ^0;
for $t = 1, 2, \cdots, \text{Maxit}$ **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 (3.12);
2. $\theta^t \leftarrow \mathbb{E}(\theta \mid \gamma, \beta, \sigma, W)$, according to (3.13);
3. $c^t \leftarrow \mathbb{E}(c \mid \gamma, y, X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma, \theta, \mu, \Sigma, W)$, according to (3.14);
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)

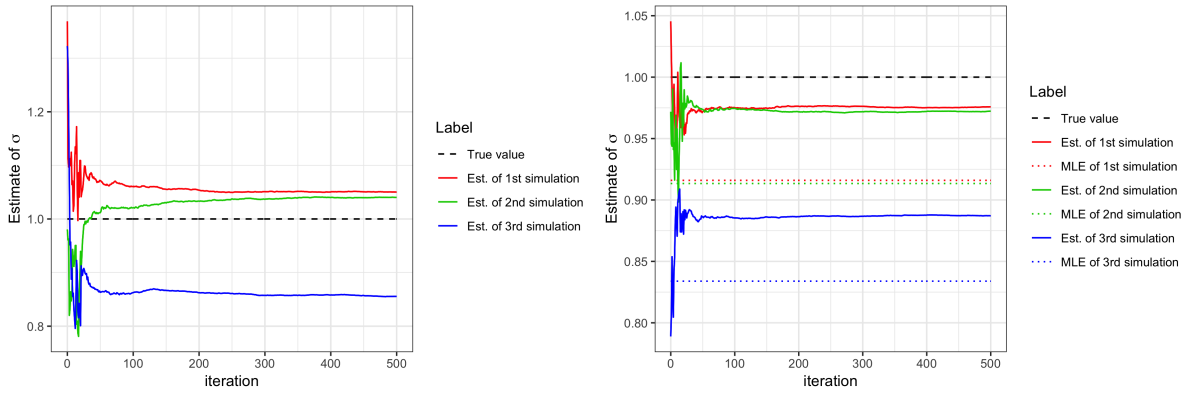
- $(\beta^{t+1}, \sigma^{t+1}, \mu^{t+1}, \Sigma^{t+1}) \leftarrow (\beta^t_{\text{MLE}}, \sigma^t_{\text{MLE}}, \mu^t_{\text{MLE}}, \Sigma^t_{\text{MLE}})$, which are the MLE for complete-data likelihood integrating the imputed missing values by expectation.

if $\|\beta^{t+1} - \beta^t\|^2 < \text{tol}$ **then**
Stop;
Output: Estimates $\hat{\beta} \leftarrow \beta^t$ and indexes for model selection $\{j : \hat{\beta}_j \neq 0\}$.

- σ^0 is given by the standard deviation: $\frac{\|y - X^0_{\text{mis}}\beta^0\|}{\sqrt{n-1}}$;
- $c^0 = \min\left\{\left(\frac{\sum_{j=1}^p \beta_j^0}{\#\{j:|\beta_j^0|>0\}+1}\right)^{-1} \sigma^0 \lambda_{r(\beta^0,1)}, 1\right\}$, where the sign $\#$ means the cardinality of a set. c^0 can be interpreted as the inverse of average magnitude for the true signal, i.e, $\beta_j^0 \neq 0$;
- $\theta^0 = \frac{\#\{j:|\beta_j^0|>0\}+a}{p+b}$ where a and b are known parameters of the prior Beta distribution on θ. Here we choose *i)* $a = \frac{2}{p}$ and $b = 1 - \frac{2}{p}$, such that the prior mean on sparsity is $\frac{2}{p}$; *ii)* $a = 0.01n$ and $b = 0.01n$; *iii)* $a = 1$ and $b = p$. Our estimation results are not sensible to the choice of hyperparameters a and b.

3.7.7 Convergence of SAEM: σ

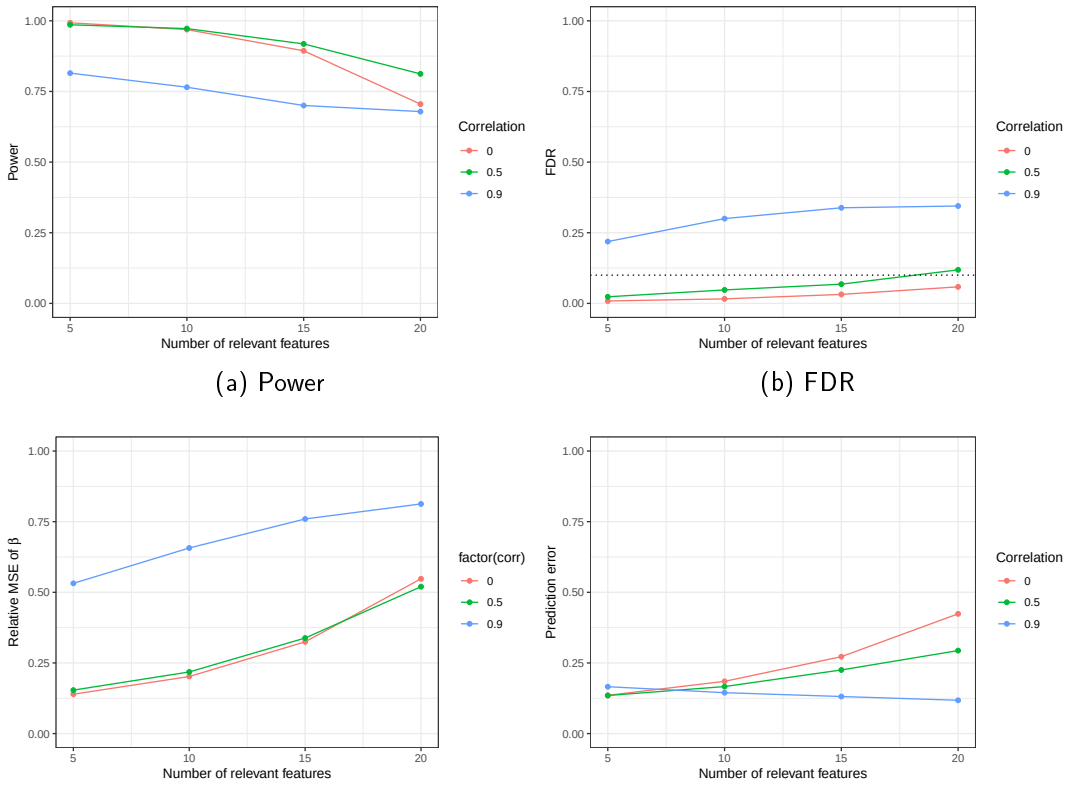
Following the simulation study in Section 3.4, we represent the convergence curves for σ with ABSLOPE in Figure 3.12 (a). The behavior is the same as for the $beta$ coefficients. We also represent convergence in the case without missing values in Figure 3.12 (b), in order to compare the estimate of σ by ABSLOPE (colored solid curves) to the biased MLE estimator without prior knowledge (colored dashed lines), *i.e.*, $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{RSS}{n}}$. We can see that the estimates of σ with both methods are biased downward, but since ABSLOPE has an additional correction term, it leads to a less biased estimator.



(a) With 10% missing values

(b) Without missing values

Figure 3.12: Convergence plots for σ with ABSLOPE (colored solid curves). (a) Case with 10% missing values; (b) Case without missing values. Black dash line represents the true value for σ. In (b) Colored dash lines indicate the biased MLE $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{RSS}{n}}$. Estimates obtained with three different sets of simulated data are represented by three different colors.



(a) Power

(b) FDR

(c) Bias of β

(d) Prediction error

Figure 3.13: Mean of power (a), FDR (b), bias of the estimate for β (c) and prediction error (d), as function of length of true signal, over the 200 simulations. Results for $n = p = 100$, with 10% missingness and strong signal.

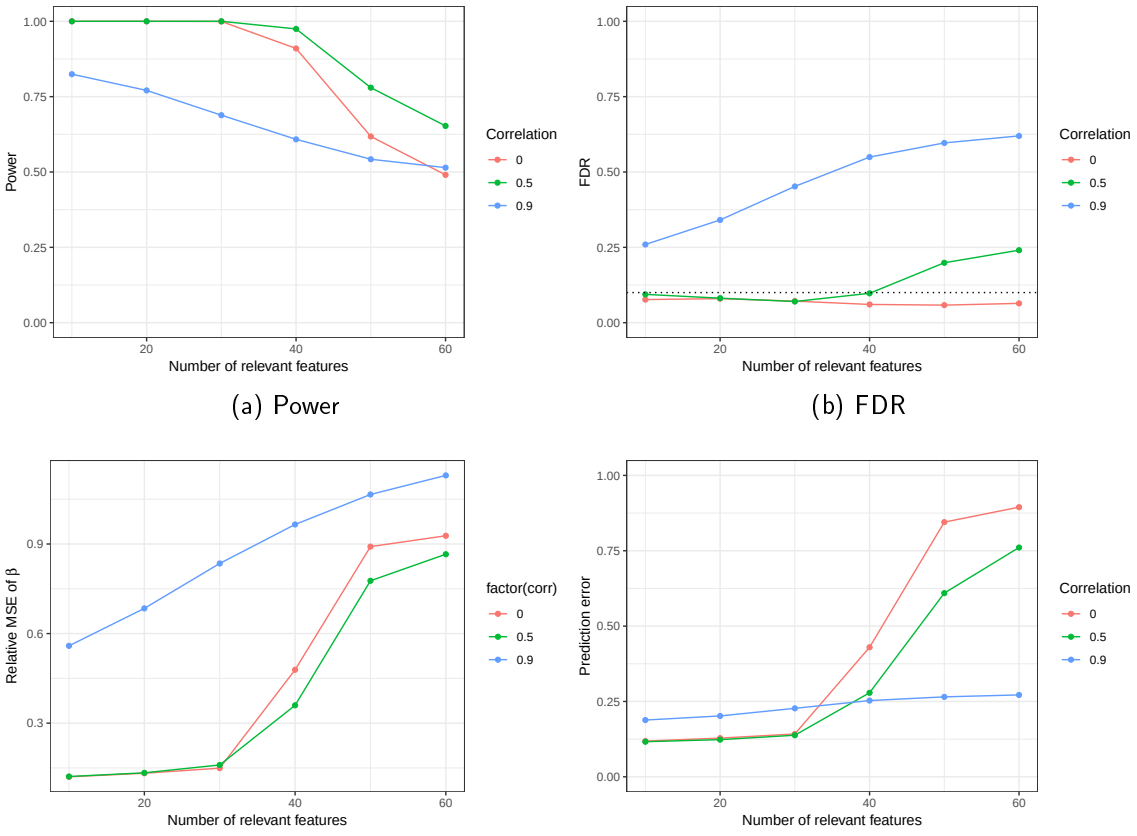
3.7.8 Behavior of ABSLOPE: effect of correlation

$n = p = 100$, 10% missingness, strong signal - vary correlation

Following the simulation study in Section 3.4, we consider additional scenarios varying correlation as follows. We consider a small dataset $n = p = 100$. The signal strength is strong and equals to $3\sqrt{2 \log p}$ and the percentage of missingness is 10%. We then vary the sparsity and correlation. The results in Figure 3.13 show:

- When there is no or little correlation, the FDR is controlled to the desired level of 0.1, but in case of high correlation, the control of the FDR is lost.
- The existence of a correlation can give more power. On on hand, the generation of missing covariates depends on those observed; on the other hand, the structure among covariates improves the prediction performances.

$n = p = 500$, 10% missingness, strong signal - vary correlation



(a) Power (b) FDR

(c) Bias of β (d) Prediction error

Figure 3.14: Mean of power (a), FDR (b), bias of the estimate for β (c) and prediction error (d), as function of length of true signal, over the 200 simulations. Results for $n = p = 500$, with 10% missingness and strong signal.

We consider a larger dataset $n = p = 500$ while the other parameters same as before. We then vary the sparsity and correlation. The results in Figure 3.14 show the same phenomenon as Figure 3.13 for the effect of correlation on FDR control.

3.7.9 Comparison with competitors: $n = p = 100$

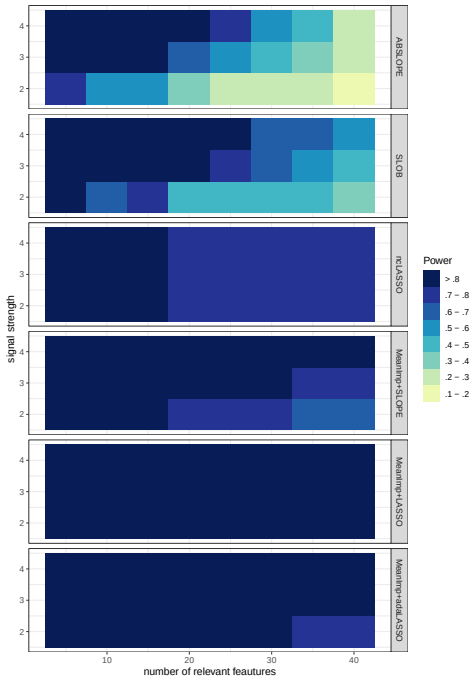
Following the simulation study in Section 3.4, we compare the proposed methodology with its competitors as follows. Figure 3.15 summarizes the result for the case $n = p = 100$, 10% missingness and without correlation. Lighter colors indicate smaller values.

- *ABSLOPE* and *SLOB* both have a strong power and an accurate prediction when the sparsity is large and the signal strength is strong enough;
- FDR is always controlled with *ABSLOPE* or *SLOB*. Other methods pay a price in FDR control to achieve good power.

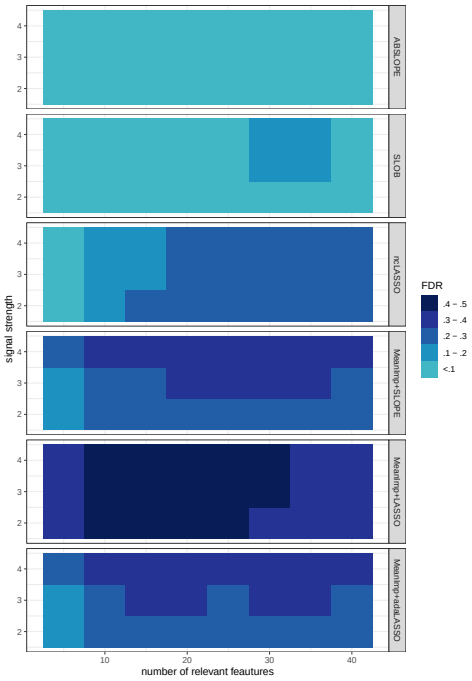
3.7.10 Variables in the TraumaBase dataset and preprocessing

Following the introduction of TraumaBase dataset in Section 3.5, we give the detailed explanation of the variables in the TraumaBase dataset:

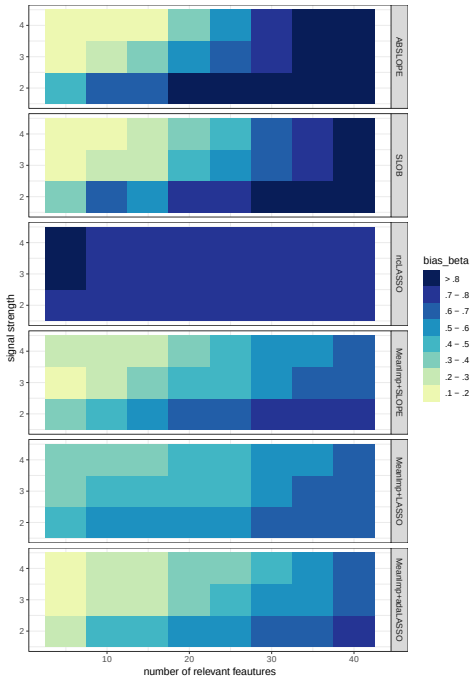
- *Age:* Age
- *SI:* Shock index indicates level of occult shock based on heart rate (HR) and systolic blood pressure (SBP). $SI = \frac{HR}{SBP}$. Evaluated on arrival of hospital.
- *MBP:* Mean arterial pressure is an average blood pressure in an individual during a single cardiac cycle, based on systolic blood pressure (SBP) and diastolic blood pressure (DBP). $MBP = \frac{2DBP+SBP}{3}$. Evaluated on arrival of hospital.
- *Delta.hemo:* The difference between the hemoglobin on arrival at hospital and that in the ambulance.
- *Time.amb:* Time spent in the ambulance *i.e.*, transportation time from accident site to hospital, in minutes.
- *Lactate:* The conjugate base of lactic acid.
- *Temp:* Patient's body temperature.
- *HR:* heart rate measured on arrival of hospital.
- *VE:* A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system.
- *RBC:* A binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed.
- *SI.amb:* Shock index measured on ambulance.



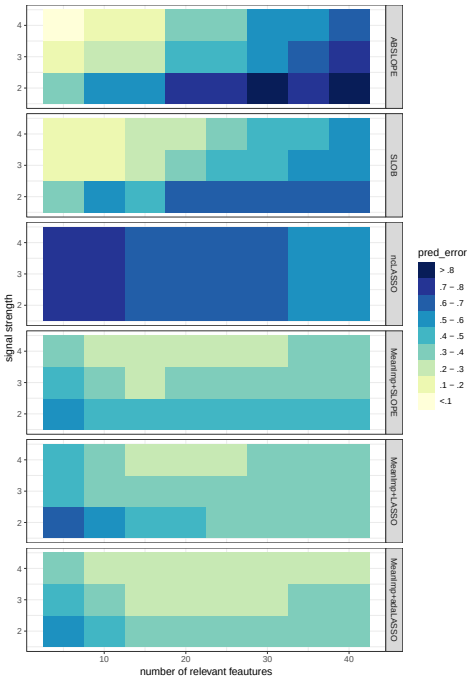
(a) Power



(b) FDR



(c) Bias of β



(d) Prediction error

Figure 3.15: Comparison of power (a), FDR (b), bias of β (c) and prediction error (d) with varying sparsity and signal strength, with 10% missingness over 200 simulations in the case without correlation.

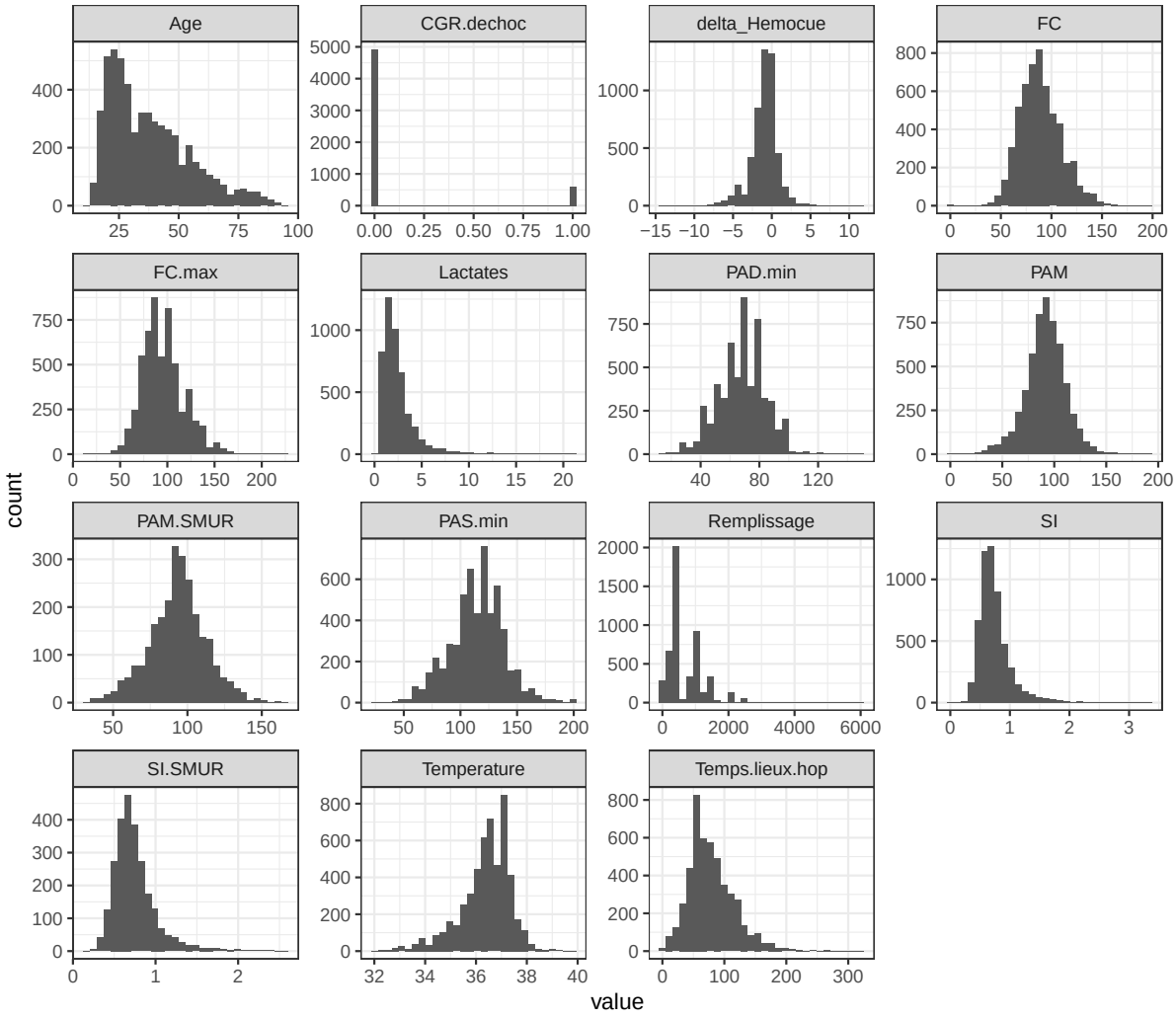
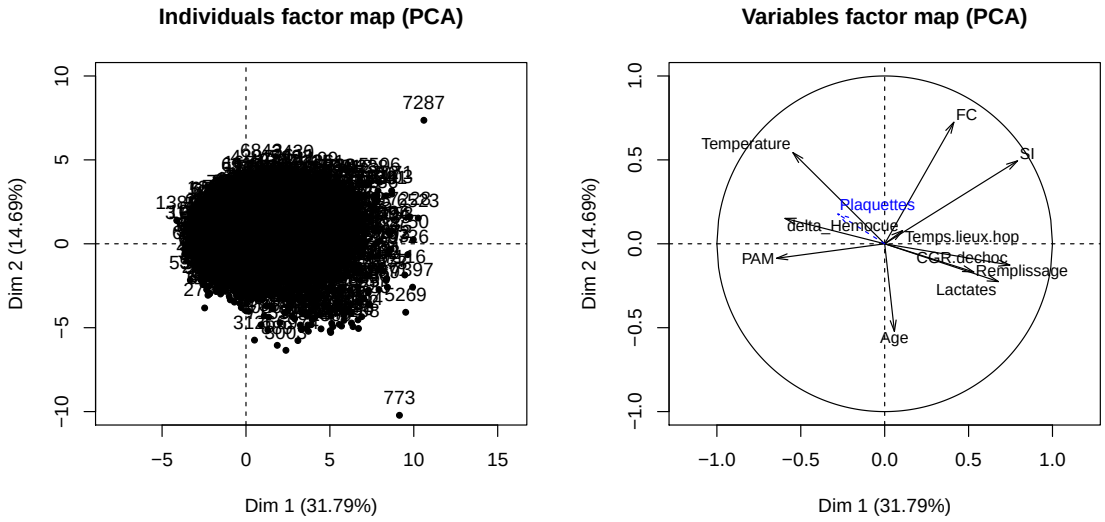


Figure 3.16: Histograms of pre-selected variables from TraumaBase.

- *MAP.amb:* Mean arterial pressure measured in the ambulance.
- *HR.max:* Maximum value of measured heart rate in the ambulance.
- *SBP.min:* Minimum value of measured systolic blood pressure in the ambulance.
- *DBP.min:* Minimum value of measured diastolic blood pressure in the ambulance.

The distribution of each variable is displayed as Figure 3.16.

With PCA, we visualized the individual and variable factor map on the two first dimension. As shown on the left in Figure 3.17, there were two observations regarded as outliers. In details, the temperature of 773^{th} patient was measured as 12.3, while the MBP of 7287^{th} patient was only 38.33, which both stand for a mistake of record.

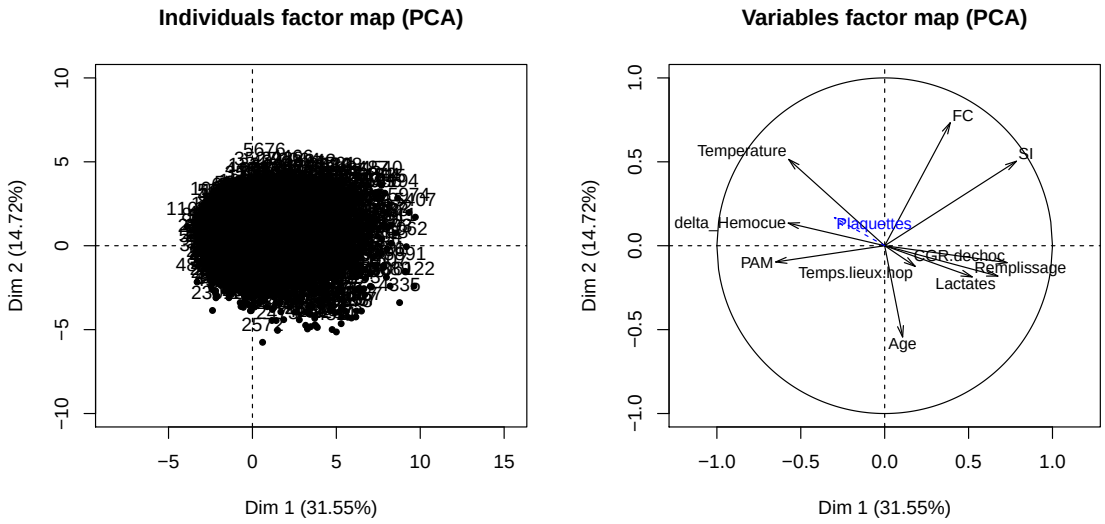


(a) Observation's factor map

(b) Variable's factor map

Figure 3.17: The factor maps from PCA before correction of wrongly recorded entries. (a) Observation's factor map (b) Variable's factor map.

We corrected all the mistakes in the records, for example, converting the value temperature smaller than 34 degree to *NA* and recalculating the MBP with the same unity for SBP. After that, we presented the factor maps from PCA in Figure 3.18, where the distribution of individuals in the principal dimensions were more homogeneous and the outliers disappeared .



(a) Observation's factor map

(b) Variable's factor map

Figure 3.18: The factor maps from PCA after correction of wrongly recorded entries. (a) Observation's factor map (b) Variable's factor map.

Chapter 4

missKnockoff— controlled variable selection with missing values

Contents

4.1 Introduction

In previous chapter, we've developed ABSLOPE for high-dimensional linear regression problem with missing values, and it targets specifically FDR control. In this chapter, we consider a similar problem of controlled model selection with high-dimensional and incomplete data, but without specifying a parametric regression model. To adress this challenge, we will leverage the knockoff methodology.

4.1.1 Problem statement

Consider a setting where observe i.i.d. $(p+1)$-dimensional random vectors

$$(X_{i1}, X_{i2}, \cdots, X_{ip}, y_i) \sim P, \qquad i = 1, ..., n,$$

where P stands for a joint distribution for the covariates $X_i = (X_{i1}, X_{i2}, \cdots, X_{ip})$ and response y_i, for each i. Similar to ABSLOPE in Chapter 3, our aim is to select important features when the design matrix is potentially contaminated by missing values (X_{ij} for some i and j may be missing). However, considering a non-parametric regression framework, here we turn to the knockoff methodology proposed by Candes et al. (2018) which has the advantage of assuming *no knowledge* of the conditional distribution $P_{y|X}$; nevertheless the joint distribution of the covariates P_X is assumed to be known. The objective is to identify a subset $\mathcal{S} \subset [p]$ indexing important variables; in other words, variable X_j with $j \in \mathcal{H}_0 \coloneqq [p]/\mathcal{S}$ is null, if $y \perp\!\!\!\perp X_j \mid X_{-j}$, where X_{-j} denotes the remaining $p-1$ variables excluding X_j. We start by recalling the knockoff method before describing our approach missKnockoff to handle missing values. Our methodology consists in combining the strength of multiple imputation (Rubin, 2009; van Buuren and Groothuis-Oudshoorn, 2011) and recent aggregation strategies that have been suggested to stabilize the knockoff (Holden and Hellton, 2018; Gimenez and Zou, 2018; Nguyen et al., 2019). In addition, we suggest a new aggregation strategy with theoritical guarantees.

4.1.2 Knockoff

When selecting variables, we aim at controlling the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), which can be defined as follows: let $\hat{\mathcal{S}}$ be a model selection outcome through a certain procedure, then:

$$\text{FDR} \coloneqq \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}|}\right] = \mathbb{E}\left[\frac{\text{number of false positives}}{\text{number of selected variables}}\right],$$

where $|\cdot|$ denotes the length of a set. For the parametric models such as penalized linear regression, one available method is SLOPE (Bogdan et al., 2015) which penalizes larger coefficients more stringently. Alternatively, based on the assumption that we are capable to model X rather than y conditionally on X, Candes et al. (2018); Barber et al. (2019) suggest a methodology named as knockoff. Intuitively, knockoff first generates a set of "fake" variables that depend on the original covariates and mimic their correlation structure. Then it returns true variables which are clearly more important than their knockoff copies according to some feature importance measures. More formally, knockoff consists of three steps:

1. First construct a set of "fake" covariates $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p)$ which satisfy:

   (a) *Exchangeability:* for any subset $S \subset \{1, \ldots, p\}$,

   $$(X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \tilde{X}), \tag{4.1}$$

   by swapping the entries X_j and $\tilde{X}_j$ for each $j \in S$.

   (b) *Unimportant variables:* $\tilde{X} \perp\!\!\!\perp y|X$, which can be guaranteed if $\tilde{X}$ is constructed without looking at y.

   A practical approximation solution described in Candes et al. (2018) consists of relaxing the condition (4.1) as to match the first two moments of the distribution (*second-order*

knockoff). For instance, if we consider Gaussian covariates, *i.e.*, $X \sim \mathcal{N}(0, \Sigma)$, then a joint distribution obeying equation (4.1) can be:

$$(X, \tilde{X}) \sim \mathcal{N}(0, G), \quad \text{where } G = \begin{bmatrix} \Sigma & \Sigma - \text{diag}(s) \\ \Sigma - \text{diag}(s) & \Sigma \end{bmatrix}, \tag{4.2}$$

with any choice of the diagonal matrix $\text{diag}(s)$ *s.t.* G is positive semidefinite. Barber et al. (2015) introduce the equicorrelated construction using:

$$s_j = 2\lambda_{\min}(\Sigma) \wedge 1, \, \forall j \, , \tag{4.3}$$

where $\lambda_{\min}$ indicates the smallest eigenvalue. This construction makes $\tilde{X}$ as uncorrelated with X as possible, while all the variable-knockoff pairs have the same correlation. As a result, one possible way to construct a knockoff $\tilde{X}$ is to sample from its conditional distribution:

$$\tilde{X} \mid X \overset{d}{=} \mathcal{N}(\mu_c, \Sigma_c),$$

where

$$\begin{aligned} \mu_c &= X - X\Sigma^{-1}\text{diag}(s) \\ \Sigma_c &= 2\text{diag}(s) - \text{diag}(s)\Sigma^{-1}\text{diag}(s) \, . \end{aligned} \tag{4.4}$$

More general constructions of knockoff can be performed by generative models as suggested in Romano et al. (2018).

2. Once knockoffs built, to use them for variable selection, we need to define some statistics Z_j and $\tilde{Z}_j$ which measure the importance of X_j and $\tilde{X}_j$ respectively. For example, we can perform a supervised learning algorithm, such as cross-validated Lasso regression, on response y and covariates $(X, \tilde{X})$ then obtain fitted coefficient vector as

$$Z_j = |\hat{\beta}_j(\lambda)|, \quad \tilde{Z}_j = |\hat{\beta}_{j+p}(\lambda)| \, .$$

Then we define a Lasso coefficient-difference (LCD) (Candes et al., 2018) statistic $W_j = Z_j - \tilde{Z}_j$ to measure the relative importance of X_j to $\tilde{X}_j$. It's also possible to choose other statistics W_j but intuitively W_j should be large and positive if jth variable is not null; otherwise, W_j should be small and arbitrarily positive or negative.

3. Find knockoff threshold $\tau > 0$ under the constraint of target FDR level q, by setting:

$$\tau = \min \left\{ t > 0 : \frac{\#\{j : W_j \leq -t\} + 1}{\#\{j : W_j \geq t\}} \leq q \right\}$$

in order to establish the estimated support $\hat{\mathcal{S}} = \{j : W_j \geq \tau\}$.

4.2 Knockoff with missing data

A popular strategy to handle missing values is single imputation, which consists in replacing the missing values by plausible values to get a completed data that can be analyzed by any methods, (Little and Rubin, 2019; Mayer et al., 2019). One can either impute according to

a joint model or using a fully conditional modeling approach (van Buuren, 2018). Powerful methods include imputation by random forest (Stekhoven and Buehlmann, 2012) and by principal component analysis (Josse and Husson, 2016). Nevertheless, even if we manage to impute by preserving as well as possible the joint and marginal distribution of the data, a single imputation can not reflect the uncertainty associated to the prediction of missing values. To achieve this goal, multiple imputation (MI) (Rubin, 2009; van Buuren and Groothuis-Oudshoorn, 2011) consists in generating several plausible values for each missing data leading to different imputed data sets. Then, the analysis is performed on each imputed data sets and results are combined. A "proper" MI method need to account for the variability of the imputation model parameters (due to sampling) to appropriately reflect the variance of prediction. It is often done by using bootstrap approaches. In the following subsections, we discuss how to combine the imputation methods and knockoff to achieve the controlled variable selection with missing covariates.

4.2.1 Single imputation – Gaussian covariates

We begin with a simple assumption of Gaussian variables, *i.e.*, $X \sim \mathcal{N}(0, \Sigma)$. With existence of missing values, the eq. (4.2) can be decomposed as follows. For each individual $i = 1, 2, \cdots, n$:

$$(X_{i,\text{obs}}, X_{i,\text{mis}}, \tilde{X}_{i,\text{obs}}, \tilde{X}_{i,\text{mis}}) \sim \mathcal{N}(0, G), \quad \text{where } G = \begin{bmatrix} \Sigma & \Sigma - \text{diag}(s) \\ \Sigma - \text{diag}(s) & \Sigma \end{bmatrix}, \tag{4.5}$$

where $\tilde{X}_{i,\text{obs}}$ (*resp.* $\tilde{X}_{i,\text{mis}}$) are the knockoff copies in $\tilde{X}_i$ corresponding to the observed (*resp.* missing) elements in X_i. Note that we can only access $X_{i,\text{obs}}$, in consequence, the construction of knockoff copies turns into the following two steps sampling:

1. For each individual $i = 1, 2, \cdots, n$, draw values for the missing elements $\hat{X}_{i,\text{mis}}$ from its conditional distribution:

$$\text{p}(X_{i,\text{mis}} \mid X_{i,\text{obs}}) = \mathcal{N}(\mu_i, \Sigma_i), \tag{4.6}$$

   where μ_m and Σ_m:

$$\begin{aligned} \mu_i &= \Sigma_{i,\text{mis,obs}} \Sigma_{i,\text{obs,obs}}^{-1} X_{i,\text{obs}} \\ \Sigma_i &= \Sigma_{i,\text{mis,mis}} - \Sigma_{i,\text{mis,obs}} \Sigma_{i,\text{obs,obs}}^{-1} \Sigma_{i,\text{obs,mis}} \,. \end{aligned} \tag{4.7}$$

   with the decomposition of the covariance matrix Σ corresponding to the missing or observed elements in X_i.

2. Sample knockoff copies $\tilde{X}$ based on the completed data $\hat{X} = (X_{\text{obs}}, \hat{X}_{\text{mis}})$:

$$\text{p}(\tilde{X} \mid X = \hat{X}) = \mathcal{N}(\mu_c, \Sigma_c),$$

   where μ_c and Σ_c are the same as in eq. (4.4).

Once the knockoff copies built, the model selection can be proceeded, as introduced in Step 2 and 3 in Section 4.1.2.

This strategy of single imputation followed by knockoff, controls the FDR as stated in the following theorem.

Theorem 1. *Model-X knockoff procedure with missing values imputed from* $\mathbf{p}(X_{mis}|X_{obs})$ *controls FDR at the level* q.

Proof. We start by observing that the design matrix with imputed missing values satisfies the exchangeability condition

$$(X_{\text{obs}}, \hat{X}_{\text{mis}}, \tilde{X})_{\text{swap}(S)} \stackrel{d}{=} (X, \tilde{X}) \,.$$

Indeed, according to the imputation, we generate values for missing covariates with:

$$\hat{X}_{\text{mis}} \sim \text{p}(X_{\text{mis}} \mid X_{\text{obs}}) \,,$$

so the completed variables share the same distribution with the original ones:

$$(X_{\text{obs}}, \hat{X}_{\text{mis}}) \stackrel{d}{=} X \,.$$

It is also obvious that conditioned on $(X_{\text{obs}}, X_{\text{mis}})$, the knockoff vector $\tilde{X}$ is independent of y. From Lemmas 3.2 and 3.3 in Candes et al. (2018), the signs of W for null variables are *i.i.d* coin flips, and FDR is controlled, according to Theorem 3.4 in Candes et al. (2018). □

Remark 2. *The theorem above is satisfied only when we know the distribution of covariates* X *and especially that we assume knowledge of the covariance matrix.*

Remark 3. *When we generate* $\hat{X}_{i,mis}$ *as eq. (4.6), it entails that* $\hat{X}_{i,mis}$ *is independent of* y *conditionally on* $X_{i,obs}$. *However* $X_{i,mis}$ *(the actual unobserved part of* X_i*) is not in general conditionally independent of* y *(unless all the missing coordinates are null variables). Consequently, the procedure we propose—the construction of the model free knockoffs—leads to the loss of power , but FDR control should be retained. More precisely, for an index* j *of missing components, we immediately have that imputed variables* $\hat{X}_j$ *and knockoff ones* $\tilde{X}_j$ *are completely exchangeable conditionally on* y, *whether or not* j *is a null or a non-null variable. For intuition, assume we have two covariates* X_1 *and* X_2 *correlated and non-null, and 99% of observations in* X_1 *are missing while* X_2 *is complete. In this case, If we sample* X_1 *independent from the response but only conditional on* X_2, *then we have virtually the same power for detecting* X_1 *as when* X_1 *is a null variable. This point would deserve further research in the direction of independence test (Candes et al., 2018) with missing values, as we also mention in Section 4.4.*
Figure 4.1 demonstrates the loss of power when we generate values for missing elements conditional only on X_{obs} *(red boxes), compared to that conditional both on* (X_{obs}, y) *(blue boxes). However, the FDR control is always achieved. The method could be extended to the case where* X_{mis} *are imputed using the information in* y *but this requires additional model assumptions or extensive non-parametric density estimation. Therefore, in simulation studies presented in this section, we restrict attention to the situation where the missing values are imputed using only the information in* X_{obs}. *Note also that this point is reminiscent of the controversy in classical multiple imputation framework where people question about including the response variable* y *in the imputation model, especially when the aim is to perform a regression afterward. Relevant discussion is provided in Section 1.2.4.*

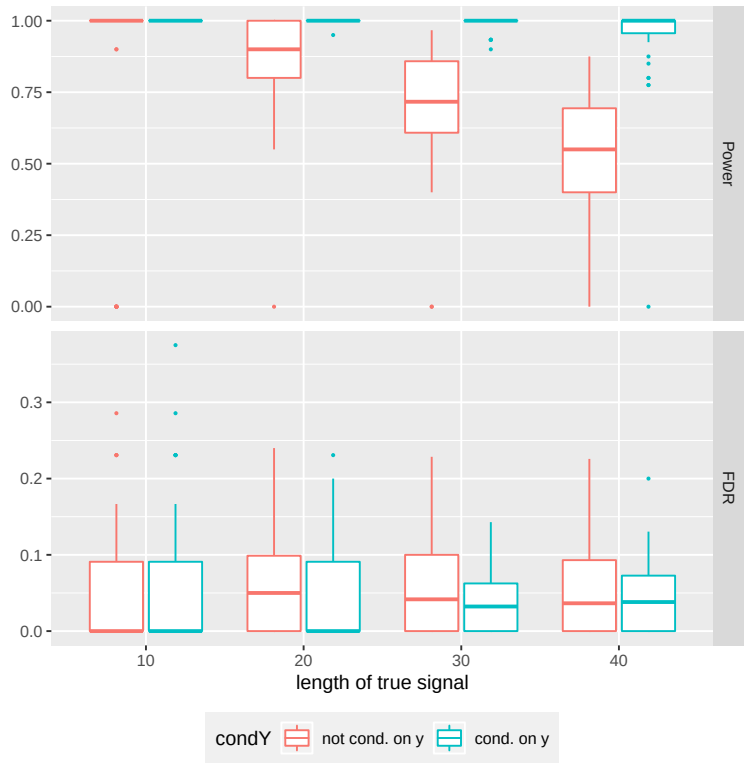


Figure 4.1: Empirical distribution of power (upper) and FDR (lower) when Σ known, grouped by length of true signal, over the 200 simulations. Results for $n = p = 100$, percentage of missingness 10%, correlation as Toeplitz matrix with 0.5 coefficient, signal strength $3\sqrt{2\log p}$.

4.2.2 Multiple imputation – Gaussian covariates

In the first step of sampling which we described in Section 4.2.1, we suggest replacing the single imputation by several possible values drawing from the same distribution, followed by knockoff on each imputed dataset and a rule of aggregation. The detailed stages are described as follows:

1. For each individual $i = 1, 2, \cdots, n$, sample M plausible values $\hat{X}^1_{i,\text{mis}}$, $\hat{X}^2_{i,\text{mis}}$, $\cdots$, $\hat{X}^M_{i,\text{mis}}$ for the missing elements in X_i.

$$X_{i,\text{mis}} \mid X_{i,\text{obs}} \sim \mathcal{N}(\mu_i, \Sigma_i), \tag{4.8}$$

   This step will be modified when we will estimate Σ, as described in Section 4.2.3. with μ_i and Σ_i same as eq. (4.7).

2. Sampling one set of knockoff copies $\tilde{X}^m$ based on each completed dataset:

$$\text{p}(\tilde{X} \mid X = (X_{\text{obs}}, \hat{X}^m_{\text{mis}})) = \mathcal{N}(\mu_c, \Sigma_c), \quad m = 1, 2, \cdots, M\,, \tag{4.9}$$

   where μ_c and Σ_c are the same as eq (4.4) where X is replaced by the completed set $(X_{\text{obs}}, \hat{X}^m_{\text{mis}})$. The vector s remains unchanged as eq. (4.3) when Σ is known.

3. For each pair of imputation and knockoff set $(X_{\text{obs}}, \hat{X}^m_{\text{mis}}, \tilde{X}^m)$, perform a supervised learning algorithm, such as cross-validation LASSO, on response y, then obtain fitted coefficient vectors and statistics:

$$Z_{mj} = |\hat{\beta}_{mj}(\lambda)|, \quad \tilde{Z}_{mj} = |\hat{\beta}_{m,j+p}(\lambda)|\,, \tag{4.10}$$

   However, in order to estimate the knockoff threshold, we can provide several possible ways to define and aggregate the statistics $\{W_{mj}\}$. The methods of aggregation are inspired by multiple knockoff (Holden and Hellton, 2018; Gimenez and Zou, 2018; Nguyen et al., 2019). In these articles with a general case without missing values, they aim at improving the stability of the selected features, and consider running the standard knockoff procedure multiple times in parallel. And they propose different methods in order to aggregates the knockoff statistics $\{W_{mj}\}$, find the knockoff threshold τ, and establish the estimated support. We will use and adapt these approaches of aggregation to deal with the case with missing values via multiple imputation. In the following, we first explain the quantile aggregation method suggested in Nguyen et al. (2019), and show how to adapt it in the framework with missing values. The we propose a new and straightforward method of aggregation that we call *aggregation by averaging the cases*. Both methods work under the assumption of second-order knockoff (Candes et al., 2018).

   - Quantile aggregation of multiple knockoff (Nguyen et al., 2019) first calculates test statistics:

$$W_{mj} = Z_{mj} - \tilde{Z}_{mj}, \qquad m = 1, 2, \cdots, M \text{ and } j = 1, 2, \cdots, p\,,$$

     and converts the test statistics to estimated p-values:

$$\pi_{mj} = \begin{cases} \frac{1 + \#\{k : W_{mk} \leq -W_{mj}\}}{p} & \text{if } W_{mj} > 0 \\ 1 & \text{if } W_{mj} \leq 0 \end{cases}.$$

Then quantile aggregation of these p-values can be defined as follows:

$$\bar{\pi}_j = \min\left\{\frac{q_\gamma\left(\pi_{mj)}\right)}{\gamma}, 1\right\} \qquad j = 1, 2, \cdots, p\,, \tag{4.11}$$

for $\gamma \in (0, 1)$ with q_γ the empirical γ quantile function.
To achieve the FDR control at level q, the p-values $\bar{\pi}_j$ are ordered ascendingly: $\bar{\pi}_{(1)} < \bar{\pi}_{(2)} \cdots < \bar{\pi}_{(p)}$, and then the largest index k such that $\bar{\pi}_{(k)} \leq \frac{kq}{p}$ (Benjamini and Hochberg, 1995) is founded. The corresponding FDR threshold is $\tau = \bar{\pi}_{(k)}$ and the j-th variable is rejected if $\bar{\pi}_j \leq \tau$.

We apply the same aggregation method in the case with missing values. Note that in our situation, we consider only a single knockoff copy for each variable, but multiple imputation for each missing element.

- Aggregation by averaging the cases. Quantile aggregation provably controls FDR but requires very large p to obtain a reasonable power. Below we introduce a new "heuristic" approach, which according to the simulation study controls FDR and has larger power than the quantile aggregation. We also provide some theoretical justification for this procedure.

  Our method relies on two de-randomizaton steps. First, the knockoff threshold value τ is calculated based on many knockoff test statistics:

  $$W_{mj} = Z_{mj} - \tilde{Z}_{mj}, \qquad m = 1, 2, \cdots, M \text{ and } j = 1, 2, \cdots, p\,,$$

  and then estimate the knockoff threshold by the formula:

  $$\tau = \min\left\{t : \frac{1}{M}\sum_{m=1}^{M} \frac{\#\{j : W_{mj} \leq -t\} + c}{\#\{j : W_{mj} \geq t\} \vee 1} \leq q\right\}, \tag{4.12}$$

  where c is a regularizing constant. The choice of c corresponding to Candes et al. (2018) is that $c = \frac{1}{m}$. In the second step for each $j = 1, 2, \cdots, p$, we calculate the median of W_{mj} over $m = 1, 2, \cdots, M$ to obtain $\bar{W}_j$,. Then we reject j-th variable if $\bar{W}_j \leq \tau$.

To justify our procedure, in the following theorem, we first show that average multiple knockoffs aggregation can be used to obtain an upward biased estimator of FDR for the single knockoff procedure which rejects H_{0j} if $W_j > t$.

Theorem 2. *Consider the single knockoffs procedure, which rejects $H_{0j} : \beta_j = 0$ if the feature statistics W_j satisifes $W_j > t$ and let*

$$FDR(t) = \mathbb{E}\left[\frac{\#\{j \in H_0 : W_j \geq t\}}{1 \vee \#\{j : W_j \geq t\}}\right] \quad . \tag{4.13}$$

If for each $i \in 1, \ldots, m$ it holds that the signs of the feature statistics W_{mj} , $j \in \{1, \ldots, p\}$ are i.i.d coin flips then we have:

$$\mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M} \frac{\#\{j : W_{mj} \leq -t\}}{\#\{j : W_{mj} \geq t\} \vee 1}\right) \geq FDR(t)$$

Proof. The proof is presented in Section 4.5.2. □

Remark 4. *The above theorem implies, that for any constant $t > 0$, and integer $M > 1$, the quantity*

$$\widehat{FDP}_M(t) = \frac{1}{M}\sum_{m=1}^{M}\frac{\#\{j : W_{m,j} \leq -t\}}{\#\{j : W_{mj} \geq t\} \vee 1}$$

is an upwards biased estimator of $FDR\ (t)$, with variance which diminishes with M. One may also note that it holds almost surely that

$$\lim_{M\to\infty}\widehat{FDP}_M(t) = \mathbb{E}\left[\widehat{FDP}(t)|X_{obs}, Y\right] \quad ,$$

where the right-hand side represents the conditional expected value of the estimator of the false discovery proportion provided by the single knockoff procedure. Therefore, if we assume that the joint distribution of $(X_{obs}, X_{mis}, \tilde{X})$ is known, and treat X_{obs}, Y as the parameters of conditional distribution $(W_j)_{j=1}^p|(X_{obs}, Y)$, the estimator $\widehat{FDP}_m(t)$ is asymptotically the Rao-Blackwellization of $\widehat{FDP}(t)$.

Remark 5. *Theorem 2 states, that if we care about FDR estimation instead of FDR control the constant c can be equal to 0. It is easy to see that this will also be true if we change the constant threshold t for a random threshold τ, that is independent of W. At the same time when the threshold τ is fully dependent on W (it is a function of W), like in the classical single knockoff procedure, then the constant c should be equal to 1. It is interesting to ask why. One way of thinking about it, is that it is an overcompensation for the error caused by optimal stopping in the choice of τ. Specifically, we can think about the classic choice of τ as starting with $\tau = 0$ and increasing it until the proportion of W_i below $-\tau$ to W_i above τ is small enough. Assuming for the sake of argument that all $|W_i|$ are different, the fraction that we base our choice of τ on, will decrease after we pass a negative W_i, and increase after we pass a positive W_i. Therefore the first time the proportion goes below a certain quantity q, we have just passed a negative W_i. On the other hand, for constant τ (or τ independent of W) one would expect that the frequencies with which the last W_i for which $|W_i| < t$ is positive or negative would be equal - at least if we restrict ourselves to null features. In this case the "+1" in the numerator is unnecessary. The stopping rule is therefore a source of downward bias in $\widehat{FDP}(\tau)$, when we think of it as an estimator of $FDR(\tau)$. An easy fix is to change the bias downwards to bias upwards, by adding this "just removed" negative W_i to the count in the numerator.*

Conjecture 1. *Let us define*

$$V(t, X, Y) = E\left(\frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}\right) = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\frac{\#\{j : W_{m,j} \leq -t\}}{\#\{j : W_{mj} \geq t\} \vee 1}$$

and let

$$\tau(X, Y) = min\{t : V(t, X, Y) \leq q\}$$

. Let us denote by $FDP(\tau(X, Y))$ the False Discovery Proportion of the single knockoff procedure rejecting H_{0j} when $W_j > \tau(X, Y)$. It holds

$$E(FDP(\tau(X, Y)) \leq q \quad .$$

In Section 4.5.1, we also review how the other available approaches of multiple knockoff are normally used for complete data, and then present how to adapt the rules of aggregation to the case with missing values.

4.2.3 High-dimensional covariance estimation with missing values

Both in the step of generating imputed values and sampling knockoff, knowledge of the covariance matrix Σ is required. In many situations, we don't know the true covariance matrix. As a result, we need to estimate it given the fact that we are in high dimension and that there are missing values.

Without missing values In the classical knockoff without missing values, Candes et al. (2018) illustrate that the empirical covariance is a poor estimator in high dimension, but applying the graphical Lasso method (Friedman et al., 2008) on available data results in indistinguishable power and FDR from the case with the true covariance, when the precision matrix of covariates Σ^{-1} is assumed to be sparse.

Alternatively if we have no prior knowledge of the covariance structure, we can advocate the use of shrinkage estimation as detailed in Ledoit and Wolf (2004). More precisely, 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.:

$$\hat{\Sigma} = \rho I_p + (1-\rho)S, \qquad \text{where } \rho = \arg\min_{\rho} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \,. \tag{4.14}$$

The method boils down to shrinking empirical eigenvalues towards their mean. The parameter ρ is chosen with asymptotically (as n and p go to infinity) uniformly minimum quadratic risk in its class, and the resulting parameters depend on the low-dimensional estimate $S = (s_{ij})$ and its variance:

$$\hat{\rho} = \frac{\sum_{i \neq j} Var(s_{ij}) + \sum_i Var(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - 1)^2)} \,. \tag{4.15}$$

In the setting of finite sample size, Schäfer and Strimmer (2005) designed improved shrinkage parameters in eq. (4.14), by replacing the variance of S in eq. 4.15 by its unbiased sample counterparts.

With missing values The simplest way to handle missing values is to first delete all the observations with missingness and then estimate the covariance matrix by graphical Lasso or shrinkage estimator using only the complete cases. However, this could be valid only under MCAR values and it is nearly impossible to do this as complete case analysis would result in deletion of almost all data.

An alternative, suggested by Lounici et al. (2014), in a MCAR homogeneous case where each data entry has a probability δ to be observed independently of the others, is first to impute all the missing values by 0. Then the author proposes adding a correction term related to the probability δ on the empirical covariance matrix of the initially imputed dataset. Furthermore, the author suggests a covariance version of matrix Lasso estimation and it is

demonstrated that, given the covariance matrix is low rank, the estimator is unbiased and valid for large sample size and dimension, typically $n << p$.

Another typical solution to estimate parameters with missing values is to use an EM algorithm (Dempster et al., 1977) on the normal covariates to estimate Σ, which is a valid strategy under a MAR mechanism but not directly designed to handle cases where $n << p$. Even though the regularized EM algorithm (Schneider, 2001) can handle large dimensional data by replacing the submatrix of estimated covariance inverse with a well conditioned matrix, but it is computationally intensive.

Finally, if we are willing to make some assumptions on the shape of covariance matrix as in probabilistic principal component analysis (PPCA) (Tipping and Bishop, 1999), we can estimate the covariance matrix using an EM algorithm. More precisely, they establish a linear model with latent variables and noise vector to explain the observed covariates, then use maximum likelihood estimates to recover the covariance matrix. Nevertheless, Ilin and Raiko (2010) show from a series of simulation results, that parameters are well estimated only when $n >> p$; while for larger scale problem where $n \approx p$, they recommend variational Bayesian PCA (VBPCA), in which they introduce priors over the model parameters to increase regularization and perform the estimation by variational EM algorithm (Neal and Hinton, 1998).
In the simulation study, to estimate Σ, we propose to combine linear shrinkage (Ledoit and Wolf, 2004) with the correction term related to percentage of missing values proposed by Lounici et al. (2014). More formally, with $\hat{\delta}$ an estimation of δ as the proportion of observed entries, the covariance matrix can be estimated by:

$$\hat{\Sigma} = \left(\hat{\delta}^{-1} - \hat{\delta}^{-2}\right) \operatorname{diag}\left(\hat{\Sigma}_n\right) + \hat{\delta}^{-2}\hat{\Sigma}_n \,, \tag{4.16}$$

where $\hat{\Sigma}_n$ is linear shrinkage estimation introduced by Schäfer and Strimmer (2005) based on eq. (4.14) and eq. (4.15), on the empirical covariance matrix of initially imputed dataset by 0.

Note that when we estimate the covariance matrix, the procedure of multiple imputation described as eq. (4.8) needs to be modified to be "proper". We need to have different estimated coefficients to generate multiple values for missingness, in order to reflect the variance of prediction for missing values which contains both the variance of estimation and that of the noise. To achieve this, we bootstrap the observations M times, to obtain different estimation of Σ, denoted by $(\hat{\Sigma}^m)_{m=1,2,\cdots,M}$, and draw the values as:

$$\begin{aligned}
\hat{X}^1_{i,\text{mis}} &\sim \mathcal{N}(\hat{\mu}^1_i, \hat{\Sigma}^1_i) \,, \\
\hat{X}^2_{i,\text{mis}} &\sim \mathcal{N}(\hat{\mu}^2_i, \hat{\Sigma}^2_i) \,, \\
&\cdots \\
\hat{X}^M_{i,\text{mis}} &\sim \mathcal{N}(\hat{\mu}^M_i, \hat{\Sigma}^M_i) \,,
\end{aligned}$$

where $\hat{\mu}^m_i$ and $\hat{\Sigma}^m_i$ are calculated from eq. (4.7) where Σ is replaced by the estimated $\hat{\Sigma}^m$, for all $m = 1, 2, \cdots, M$. For the same reason, the generation of knockoff copies is also modified, such that in eq. (4.7), not only X is replaced by the imputed set $(X_{\text{obs}}, \hat{X}^m_{\text{mis}})$, but the constant s is calculated using the estimated $\hat{\Sigma}^m_i$.

In summary, the entire procedure is illustrated on Figure 4.2 and in Algorithm 7, including

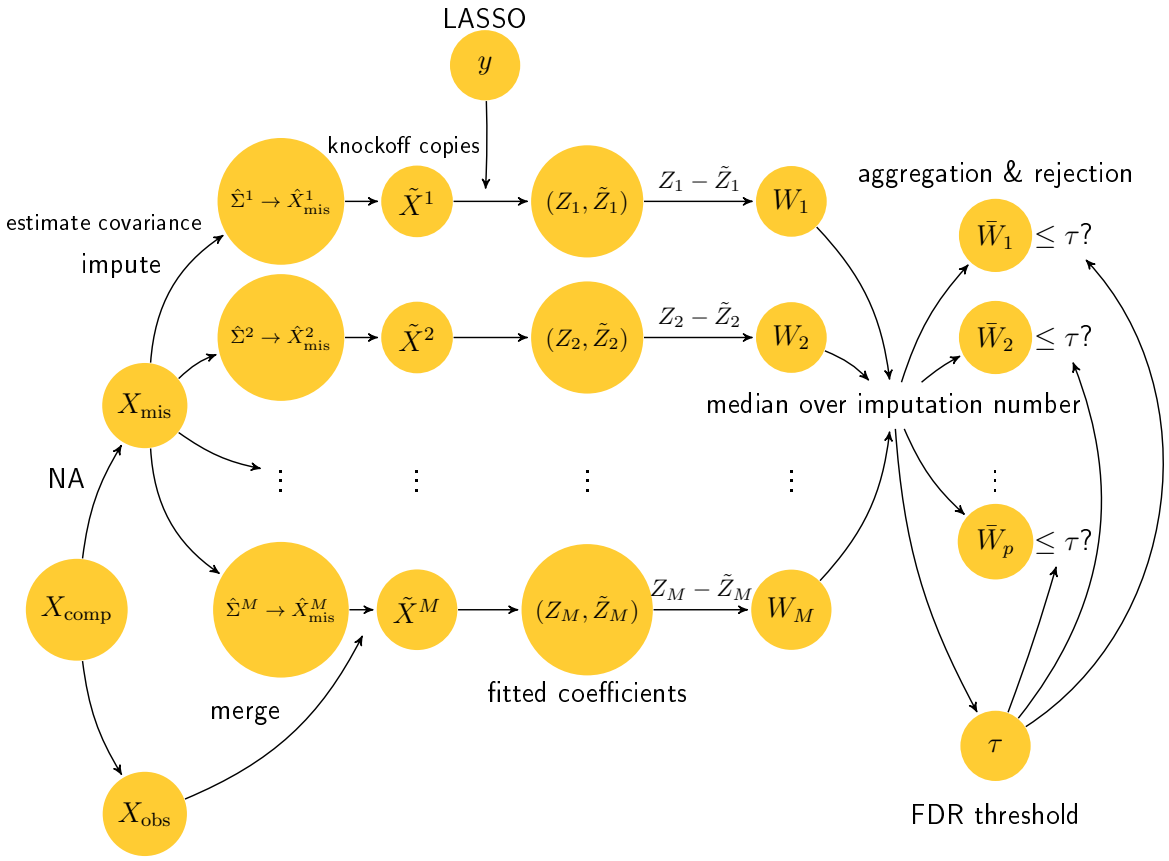


Figure 4.2: Diagram of stages for handling missing values for model selection via missKnockoff (aggregation by averaging the cases).

processing missing values and model selection via multiple knockoff, aggregated by averaging the cases.

Algorithm 7 missKnockoff: multiple knockoff with missing values (aggregation by averaging the cases)

Input: $X = (X_{\text{mis}}, X_{\text{obs}})$;
for $m = 1, 2, \cdots, M$ **do**
(Bootstrap)

1. Bootstrap X with missing values.
2. On bootstrap samples, estimate the covariance $\hat{\Sigma}^m$ from eq. (4.16);
3. Impute missing values $\hat{X_{\text{mis}}}^m$ from eq. (4.8) and generate knockoff copies $\tilde{X}^m$ from eq. (4.9);
4. On the set $(y, X_{\text{obs}}, \hat{X_{\text{mis}}}^m, \tilde{X}^m)$, use LASSO to obtain fitted coefficient vectors and statistics Z_{mj} and $\tilde{Z}_{mj}$ from (4.10), $j = 1, 2, \cdots, p$;
5. Calculate test statistics $W_{mj} = Z_{mj} - \tilde{Z}_{mj}$, $j = 1, 2, \cdots, p$;

(Aggregation)

1. Estimate the knockoff threshold τ by eq. (4.12);
2. Calculate the median of $\{W_{mj}\}$ over $m = 1, 2, \cdots, M$ to obtain $\bar{W}_j$;

if $\bar{W}_j \leq \tau$ **then**
Reject j**-th variable**.
Output: Indexes for model selection $\{j : \bar{W}_j > \tau\}$

Note that in Figure 4.2, for missKnockoff with quantile aggregation, one only need to replace the median $\bar{W}_j$ over imputation number with aggregated p-values $\bar{\pi}_j$ from eq. (4.11).

4.3 Simulation study

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

- A design matrix $X_{n\times p}$ is generated from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$.
- The signal magnitude is $c_0\sqrt{2\log p}$[1] when c_0 is large the signal strength is stronger. Only k on the p predictors are non-zero and all equal to $c_0\sqrt{2\log p}$.
- The response vector is generated from $y = X\beta + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2 I_n)$ and $\sigma = 1$. In Section 4.3.1, nonlinear model is also considered.

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

- Missing values are entered into the design matrix using a MCAR mechanism: we randomly generate a proportion $1-\delta$ of missing cells.

- Correlation between covariates is considered as $\Sigma = \text{toeplitz}(\rho)$[2]. In Section ??, results are presented when all the pairwise correlation equals ρ.

4.3.1 Aggregation by averaging the cases – Effect of parameters

In the following, we illustrate the effect of different parameters setting and present different methods:

- knockoff: classical second-order knockoff on original complete dataset;

- miss1Knockoff: a single imputation to deal with missing values in knockoff, as described in Section 4.2.1;

- miss5Knockoff: multiple imputation to deal with missing values in knockoff, as described in Section 4.2.2. The number of imputed datasets equals to 5 and we apply the proposed method of aggregation by averaging the cases.

- miss10Knockoff: same as described in MI5Knockoff, but with 10 imputed datasets instead.

We consider both the case when the covariance Σ is known and when we estimate Σ using eq. (4.16).

Scenario 1: Linear model

We first consider $n = p = 100$, 10% missingness, correlation as Toeplitz matrix with coefficient 0.5 and the following parameters can be varied:

- sparsity: number of true signal $k = 10,\ 20,\ 30,\ 40$;

- signal strength: from weak to strong $1.3\sqrt{2\log p}\,, 2\sqrt{2\log p}\,, 3\sqrt{2\log p}$

[2]The Toeplitz structure (or auto-regressive structure) for correlation has been introduced for microarry study (Guo et al., 2006), with the form: $\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho^{p-2} & \rho^{p-1} \\ \rho & 1 & \ddots & \cdots & \rho^{p-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{p-2} & \cdots & \ddots & \ddots & \rho \\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1 \end{pmatrix}_{p\times p}$, where $\rho \in [0,1]$ is a constant. For the Toeplitz structure, adjacent pairs of covariates are highly correlated and those further away are less correlated, as in microarry study, genes are correlated due to their distance in the regularity pathway.

Results 1: Effect of sparsity We keep the signal strength strong $3\sqrt{2\log p}$ and unchanged, but vary the number of non-zero elements in coefficients. According to Figure 4.3:

- The classical knockoff on original complete dataset (red boxes) achieves great power and FDR control, regardless of the varying number of non-zero elements.
- When missing data occur, obviously a single imputation (green boxes) cannot reaches the same level of power as the knockoff with complete dataset does, and its variance is huge, However even the power is a bit lower, FDR control is still satisfying on average, under the target level 0.1.

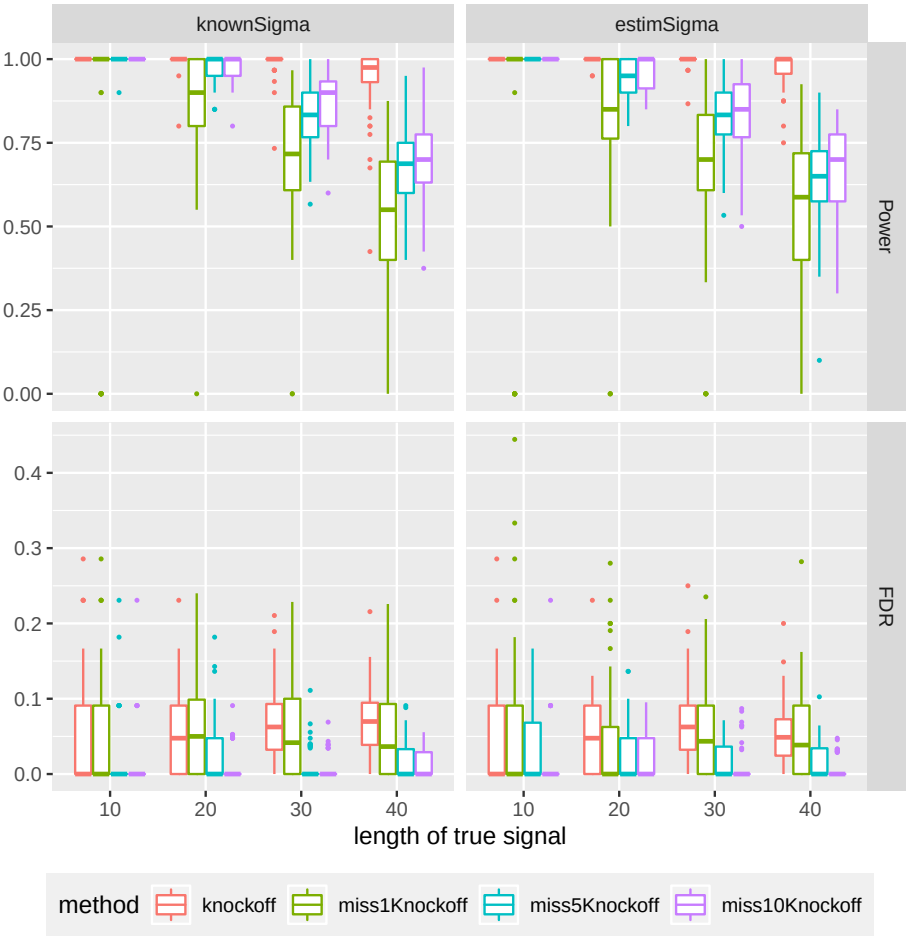


Figure 4.3: Empirical distribution of power (upper) and FDR (lower) when Σ known (left) and when we estimate Σ (right), grouped by length of true signal, over the 200 simulations. Results for $n = p = 100$, percentage of missingness 10%, correlation as Toeplitz matrix with 0.5 coefficient, signal strength $3\sqrt{2\log p}$.

- When multiple imputed datasets are considered (blue and magenta boxes), the selection results improve a lot compared to single imputation, especially in more sparse case (number of true signal equals to 10 or 20). However, when the sparsity decrease, we cannot expect a better power as before. FDR is also controlled, but more conservative, less than 0.03 on average.

- We notice that multiple imputation with 10 replicates (magenta boxes) generally has a better power than that with only 5 replicates (blue boxes). When the number of imputed datasets increases, we observe that the FDR is more conservative. Intuitively, due to the aggregation by averaging the cases, the variability of model selection is reduced.
- When we estimate the covariance matrix, if we compare the subfigures on the right side to the left side, the power of model selection results don't change much for all these methods. In particular with missing values, we almost recover the case when Σ is known. Still, FDR is controlled under the target level 0.1 on average.

Results 2: Effect of signal strength Now we fix the number of true signal equal to 20 but change the signal strength from weak to strong. According to Figure 4.4:

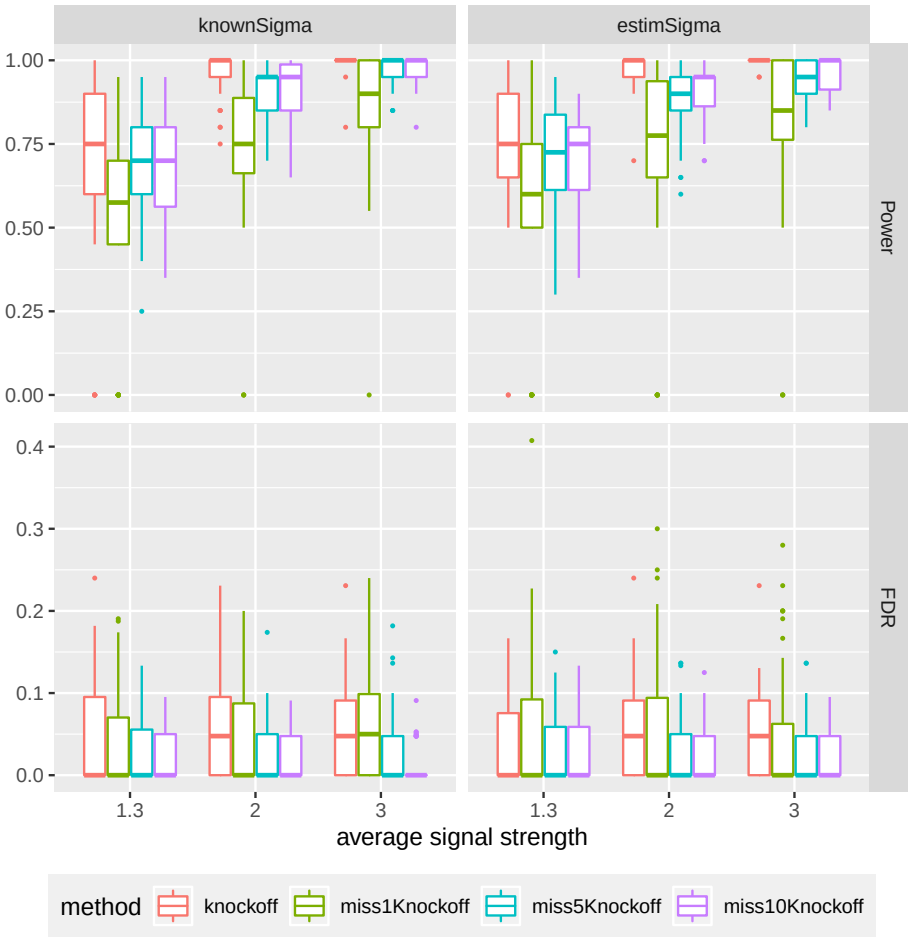


Figure 4.4: Empirical distribution of power (upper) and FDR (lower) when Σ known (left) and when we estimate Σ (right), grouped by average signal strength, over the 200 simulations. Results for $n = p = 100$, percentage of missingness 10%, correlation as Toeplitz matrix with 0.5 coefficient, length of true signal 20.

- The results emphasize what we previously observe in Figure 4.3: perfect FDR control on average, both in the case with or without missing values. In our method, when number

of imputed datasets increase, the power improves generally and the FDR becomes more conservative.

- Weaker signal strength cause more difficulty in achieving good power. However, miss10Knockoff can almost have the same level of power as the original knockoff on complete dataset, even in the case when the covariance Σ need to be estimated. A perfect performance can be observed when signal strength is equal to $3\sqrt{2\log p}$.

Results 3: Effect of covariance estimation To look closer how the estimation of covariance influences the results, now the signal strength is fixed as $3\sqrt{2\log p}$, also length of true signal equal to 20. According to Figure 4.5:

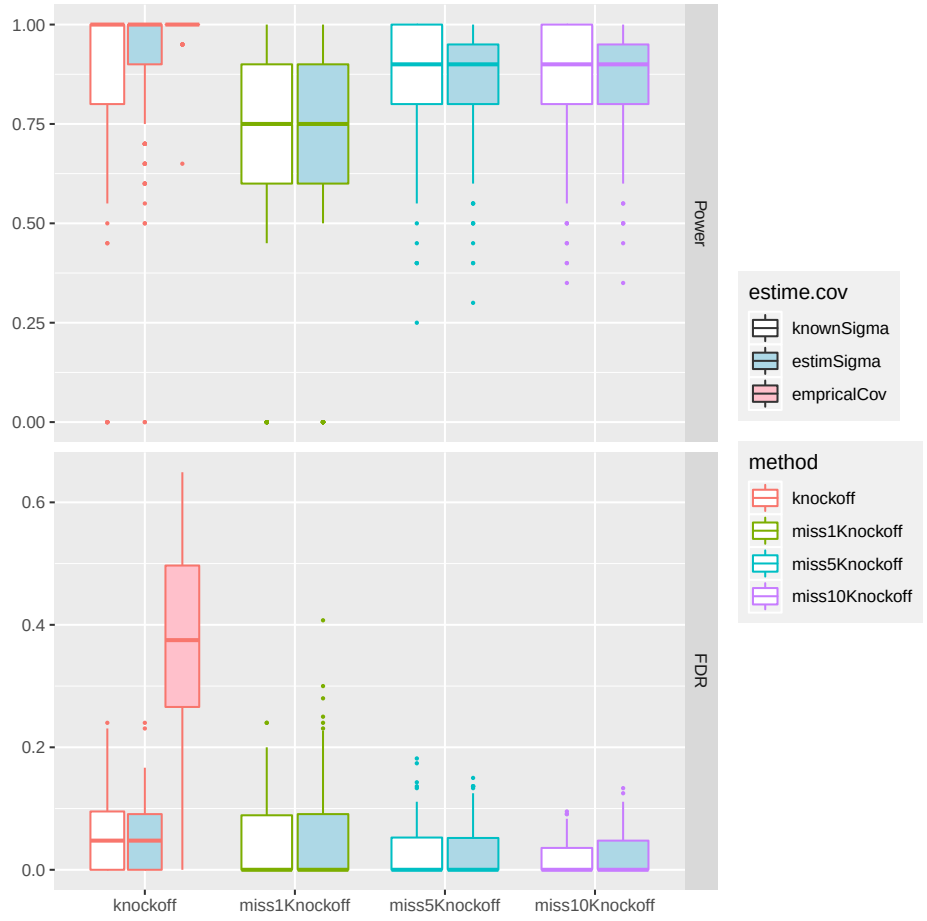


Figure 4.5: Empirical distribution of power (upper) and FDR (lower) when Σ known (boxes without fill), when we estimate Σ using corrected shrinkage estimation as eq. (4.16) (boxes with lightblue fill) or empirical covariance matrix without shrinkage (boxes with pink fill), grouped by four methods, over the 200 simulations. Results for $n = p = 100$, percentage of missingness 10%, correlation as Toeplitz matrix with 0.5 coefficient, length of true signal 20 and signal strength $3\sqrt{2\log p}$.

- On average, both the power and FDR remains at a similar value when we estimate Σ compared to the case when it is known.

- Generally the variability of true positive number is reduced, but the number of false discovery shows slightly more variation.
- Obviously empirical covariance estimation is not appropriate even for the complete dataset, where the FDR control is sacrificed largely to achieve a good power. Note that only for classical knockoff methods, we present the empirical covariance estimates, since for the other cases, using empirical ones results in ill-posed problem without solutions.

Results 4: Effect of percentage of missingness To evaluate how the percentage of missingness influences the results, now the signal strength is fixed as $3\sqrt{2\log p}$, also length of true signal equal to 20. The percentage of missingness is varied from 10 to 40. According to Figure 4.6:

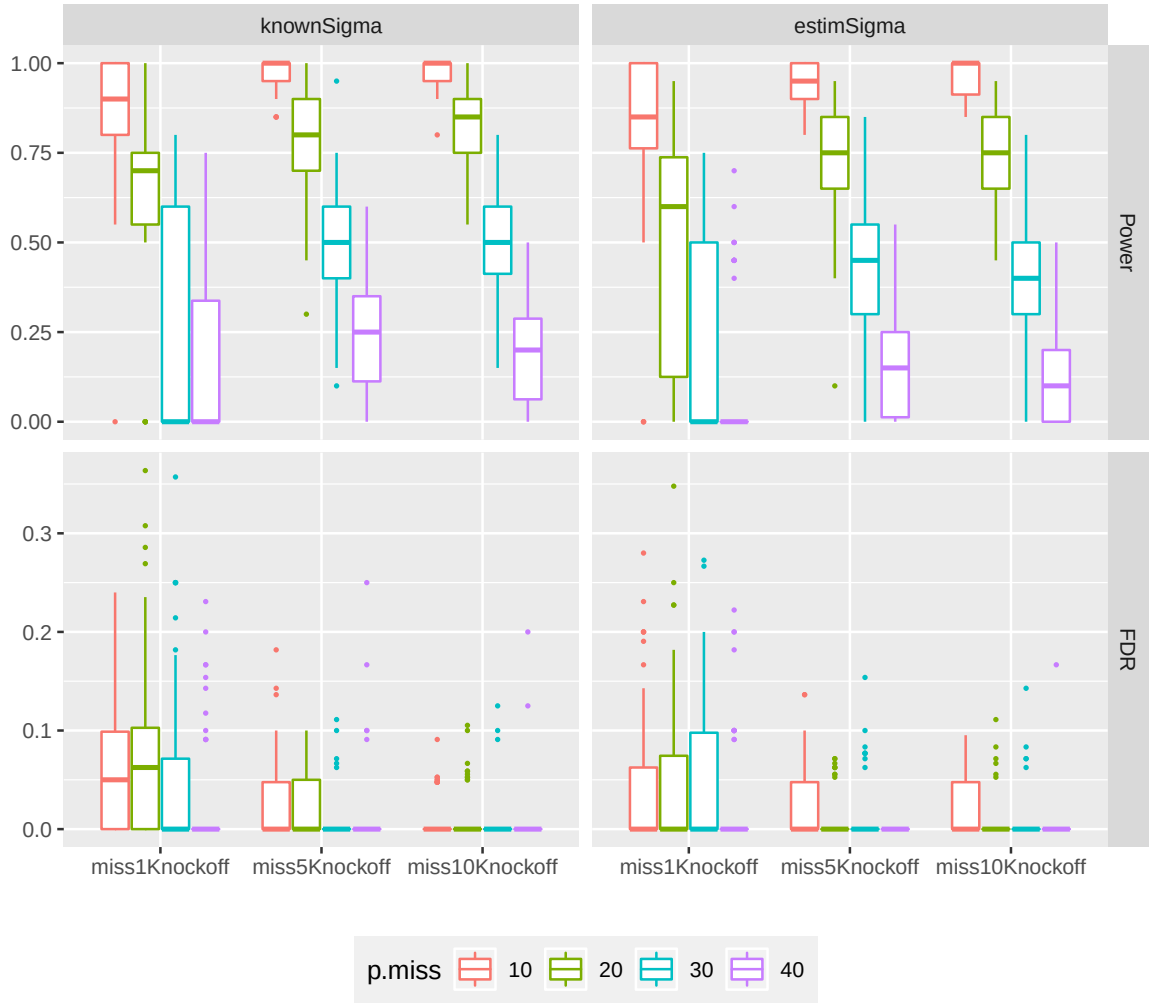


Figure 4.6: Empirical distribution of power (upper) and FDR (lower) when Σ known (left) and when we estimate Σ using corrected shrinkage estimation as eq. (4.16) (right), grouped by three methods with different percentage of missing values, over the 200 simulations. Results for $n = p = 100$, correlation as Toeplitz matrix with 0.5 coefficient, length of true signal 20 and signal strength $3\sqrt{2\log p}$.

- FDR is always control regardless of the percentage of missing values.
- When percentage of missing values is larger, the power obtained is reduced. Multiple imputation, on average, can result in a larger power with less variability.

Scenario 2: Nonlinear model

We consider the same setting as the previous scenario but a more complicated nonlinear model in our simulation study: $y = \exp(\sum_{j=1}^{p} \beta_j X_j + \beta_0) + \varepsilon$, with $\epsilon \sim N(0, \sigma^2 I_n)$ and $\sigma = 1$. in order to demonstrate that the model selection with missing values based on knockoff can be flexible and useful in many applications.

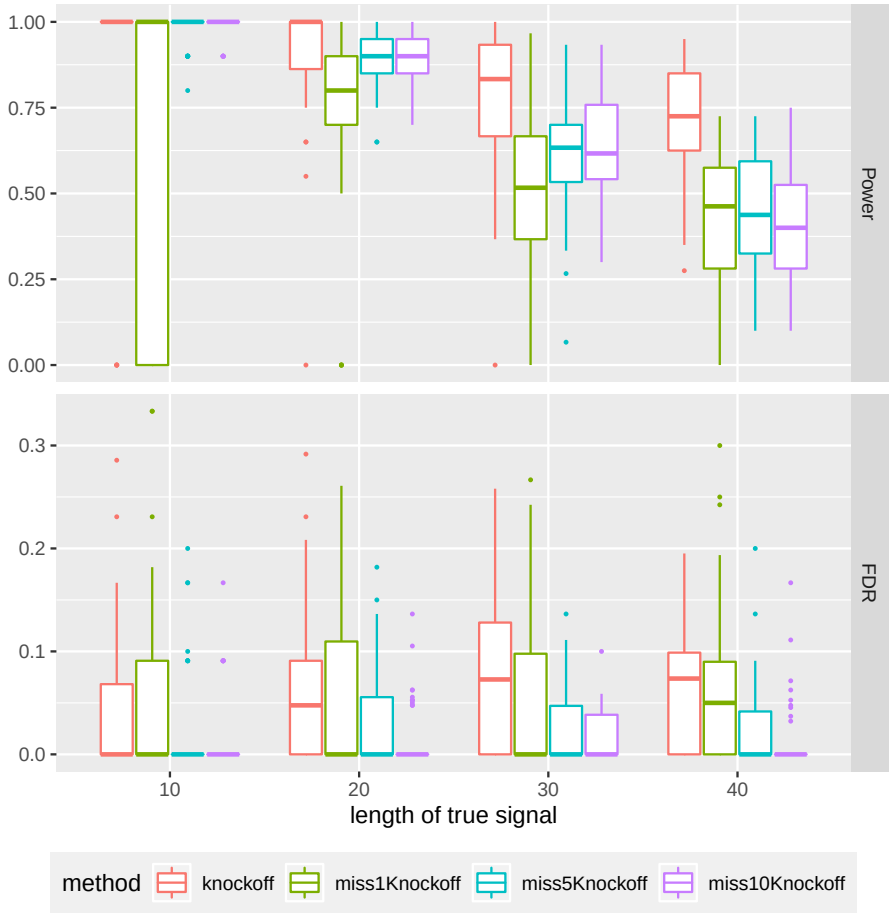


Figure 4.7: Empirical distribution of power (upper) and FDR (lower) when we estimate Σ, grouped by length of true signal, over the 200 simulations. Results for $n = p = 100$, percentage of missingness 10%, correlation as Toeplitz matrix with 0.5 coefficient, signal strength $3\sqrt{2 \log p}$.

In missKnockoff procedure, the regular LASSO described as eq. (4.10) used to compute the statistic is replaced by rank LASSO (Rejchel and Bogdan, 2019) in order to deal with the heavy tail issue caused by exponential function. The ranks of y are defined as $R_i = \sum_{j=1}^{n} \mathbb{I}(y_j \leq y_i), \quad i = 1, \ldots, n$. Then the rank LASSO problem, where the actual values

of the response are replaced by their centered ranks[3], aims at solving:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \quad \frac{1}{2n} \sum_{i=1}^{n} \left(R_i/n - 0.5 - \theta' X_i\right)^2 + \lambda |\theta|_1 \,.$$

Model selection results are presented in Figure 4.7 when $n = p = 100$, percentage of missingness 10%, correlation as Toeplitz matrix with 0.5 coefficient, and strong signal strength $3\sqrt{2 \log p}$.

- The FDR control by missKnockoff still hold well, while sacrificing a bit power.
- In general, multiple imputation outperforms a single one in terms of power, except when length of true signal is too large (40), which is an extreme difficult case for missKnockoff.

4.3.2 Method comparison

In this section, we aim at comparing the proposed method with the other approaches in terms of power and FDR control. The following methods are presented:

- miss10Knockoff, multiple imputation to deal with missing values in knockoff, as described in Section 4.2.2. The number of imputed datasets equals to 10 and we apply the aggregation by averaging the cases. The effect of parameters including the number of imputation for this method is discussed in Section 4.3.1. Here we only present the best case with ten imputations.
- qtKnockoff: instead of the aggregation rule above, using quantile aggregation for multiple imputed datasets in missKnockoff;
- ABSLOPE: high-dimensional model selection with missing values targeted at FDR control based on sorted l_1 penalization, as proposed in Chapter 3;
- SLOB: simplified version of ABSLOPE;
- knockoff: classical knockoff on original complete dataset.

Various relationships between X and y are considered, including linear model and nonlinear ones.

[3]Take a simple example:

y	x_1	x_2
10	6.4	2.6
2	3.3	1.2
4	5.2	1.8

. Then ranks are $R = (3, 1, 2)$. In the objective function of rank LASSO, the corresponding centered ranks $(R_i/n - 0.5)$ are $(\frac{1}{2}, -\frac{1}{6}, \frac{1}{6})$.
Replacing values of response variables by their ranks is a well-known approach in non-parametric statistics and leads to robust procedures. (Rejchel and Bogdan, 2019) shows that, under certain standard assumptions, the support of θ coincides with the support of β. In addition, the method can properly identify relevant predictors, even when the distribution of error terms is unknown and the link function is nonlinear

Scenario 1: Linear model

We first consider $n = p = 100$, 10% missingness, correlation as Toeplitz matrix with coefficient 0.5 and the following parameters can be varied:

- sparsity: number of true signal $k = 10,\ 20,\ 30,\ 40$;
- signal strength: from weak to strong $1.3\sqrt{2\log p}, 2\sqrt{2\log p}, 3\sqrt{2\log p}$

Results 1: Strong signal We keep the signal strength strong $3\sqrt{2\log p}$ or weak as $1.3\sqrt{2\log p}$, and also vary the number of non-zero elements in coefficients. According to Figure 4.8:

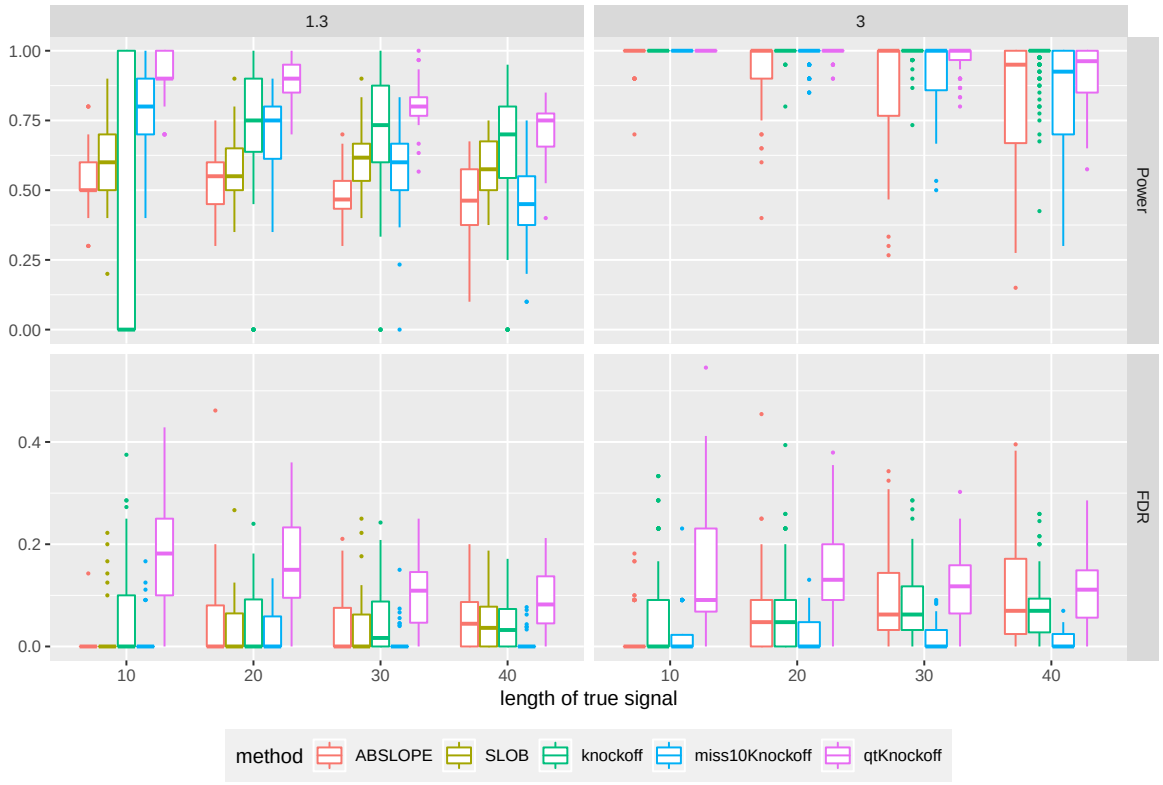


Figure 4.8: Empirical distribution of power (upper) and FDR (lower) when we estimate Σ, when $n = p = 100$, grouped by length of true signal, over the 200 simulations. Results for percentage of missingness 10%, correlation as Toeplitz matrix with 0.5 coefficient, signal strength weak as $1.3\sqrt{2\log p}$ (left) or strong $3\sqrt{2\log p}$ (right).

- In both cases, only the miss10Knockoff (blue boxes) and algorithm ABSLOPE (and its simplified version) (red boxes) controls FDR;
- Even quantile knockoff (magenta boxes) can result in a good power, but FDR control is slightly lost especially when signal is sparse and weak; this can be explained that, with eq. (4.11), the empirical p-value will be bound below by $\frac{1}{p}$; which means if p is not big enough, this p-value will be quite large compare to the FDR threshold after applying the BH correction for controlling FDR.

Scenario 2: Nonlinear model

Result: Strong signal We consider the exponential model as described previously in Section 4.3.1. We take strong signal strength $3\sqrt{2\log p}$. According to Figure 4.9:

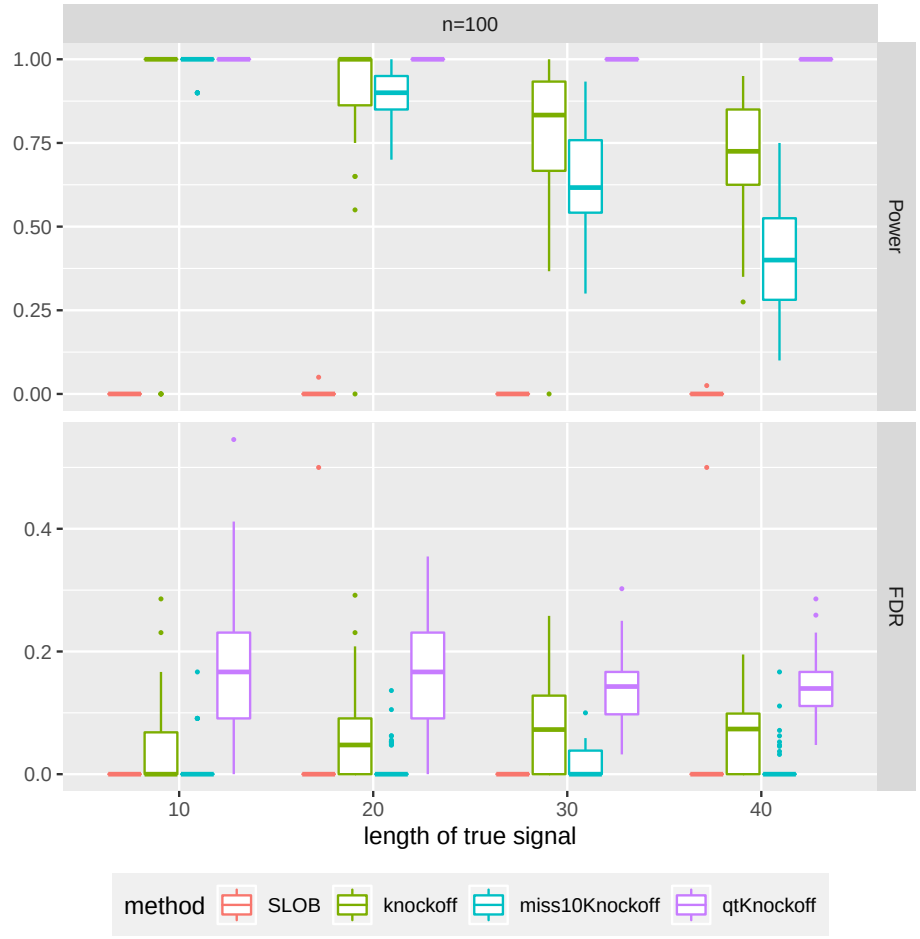


Figure 4.9: Empirical distribution of power (upper) and FDR (lower) when we estimate Σ, when $n = p = 100$, grouped by length of true signal, over the 200 simulations. Results for percentage of missingness 10%, correlation as Toeplitz matrix with 0.5 coefficient, signal strength $3\sqrt{2\log p}$.

- The simplified version ABSLOPE (red boxes) fails to select any variable, since the method was developed only for linear model.
- The knockoff based methods can still reach a good level of power, while the miss10Knockoff (blue boxes) has a good control of FDR. Same as the reasons described for Figure 4.8, quantile aggregation (magenta boxes) slight loses FDR control.

4.4 Discussion

In this chapter, we integrate the model selection for non-parametric regression using knockoff and imputation methods to handle missing values. Both theoretical guarantee and simulated

evaluation are provided for the performance of the proposed methodology. However, due to the ignorance of information provided by the responses, there is still room of improvement for the power. To do this, It would be necessary to return to the issue of testing unconditional independence (Candes et al., 2018) with missing values. Another extension is to consider different sampling methods such as imputation based on neural network (Yoon et al., 2018), and generating knockoff based on deep learning (Romano et al., 2018) to achieve a "model-free" solution both for covariates and responses.

4.5 Supplementary materials

4.5.1 Other approaches of multiple knockoff aggregations

- Multiple model-free knockoffs (Holden and Hellton, 2018) assume the same setting as Candes et al. (2018) (a perfect knowledge of the distribution of X), but replace the typical statistics (absolute value of the estimated Lasso coefficient) by:

$$\begin{aligned} W_{1j} &= Z_{1j} - \frac{1}{M'-1}\sum_{u=M'+1}^{2M'-1} \tilde{Z}_{uj} \\ W_{mj} &= \tilde{Z}_{mj} - \frac{1}{M'-1}\sum_{u=M'+1}^{2M'-1} \tilde{Z}_{uj}\,, \end{aligned} \tag{4.17}$$

  where M' is a reformulation of the replicate numbers of the multiple knockoffs, which satisfies $M = 2M' - 1$. Intuitively a large value of W_{1j} indicates that the j-th variable is significant, while a large value of W_{mj} for $m > 1$ is only due to randomness.

  Then the threshold is defined as:

$$\tau = \min\left\{ t > 0 : \frac{\#\left\{1 < m \leq M', 1 \leq j \leq p : W_{mj} \geq t\right\}}{\#\left\{1 \leq j \leq p : W_{1j} \geq t\right\}(M'-1)} \leq q \right\}. \tag{4.18}$$

  and finally we reject j-th variable if $W_{1j} \leq \tau$. The knockoff randomness is reduced and the power is increased.

  With missing values In order to adapt this procedure in the framework with missing values, we need some modifications, since we have several different coefficients calculated from the multiple imputation, *i.e.*, Z_{mj} are different for each $m = 1, 2, \cdots, M$. The statistic W_{1j} in the eq. (4.17) can be replaced by:

$$W_{1j}^{(m)} = Z_{mj} - \frac{1}{M'-1}\sum_{u=M'+1}^{2M'-1} \tilde{Z}_{uj}\,.$$

  As a result, we take each imputation into account. And we also modify the eq. (4.18) as:

$$\tau = \min\left\{ t > 0 : \frac{\#\left\{1 < m \leq M', 1 \leq j \leq p : W_{mj} \geq t\right\}}{\#\left\{1 \leq m \leq M', 1 \leq j \leq p : W_{1j}^{(m)} \geq t\right\}\frac{M'-1}{M'}} \leq q \right\}.$$

And finally we reject j-th variable if $\text{median}(W_{1j}^{(m)}) \leq \tau$.

- Simultaneous multiple knockoff (Gimenez and Zou, 2018) doesn't neither suggest the absolute value of the estimated Lasso coefficient as the test statistics. They first consider descending order of the coefficient vectors $Z = (Z_{1j}, \tilde{Z}_{1j}, \tilde{Z}_{2j}, \cdots, \tilde{Z}_{Mj})$ for each variable j, as $(Z_j^{(\text{ord})})_{0 \leq \text{ord} \leq M}$. For all $1 \leq j \leq p$, define:

$$\kappa_j = \underset{0 \leq \text{ord} \leq M}{\arg\max} Z_j^{(\text{ord})} \quad \tau_j = Z_j^{(0)} - Z_j^{(1)} .$$

The crucial result is that null κ_j behave uniformly and independently in distribution and can be used to estimate the number of false discoveries:

$$\widehat{FDP} = \frac{\frac{1}{M} + \frac{1}{M} \# \{j : \kappa_j \geq 1, \tau_j \geq t\}}{\# \{j : \kappa_j = 0, \tau_j \geq t\} \vee 1},$$

and hence the threshold can be defined as:

$$\tau = \min \left\{ t > 0 : \frac{\frac{1}{M} + \frac{1}{M} \# \{j : \kappa_j \geq 1, \tau_j \geq t\}}{\# \{j : \kappa_j = 0, \tau_j \geq t\} \vee 1} \leq q \right\} . \tag{4.19}$$

Finally we reject the j-th variable if $\kappa_j = 0$ and $\tau_j \leq \tau$.

With missing values Same as the first method "multiple model-free knockoff", we are able to modify the procedure to adapt to the case with missing values. First we have multiple coefficient vectors $Z'_{mj} = (Z_{mj}, \tilde{Z}_{1j}, \tilde{Z}_{2j}, \cdots, \tilde{Z}_{Mj})$ for each variable j and $m = 1, 2, \cdots, M$ due to multiple imputation. Then descending ordering results in $(Z_{mj}^{(\text{ord})})_{0 \leq \text{ord} \leq M}$. For all $1 \leq j \leq p$ and $1 \leq m \leq M$, define:

$$\kappa_{mj} = \underset{0 \leq \text{ord} \leq M}{\arg\max} Z_{mj}^{(\text{ord})} \quad \tau_{mj} = Z_{mj}^{(0)} - Z_{mj}^{(1)} .$$

And hence the threshold (4.19) can be modified as:

$$\tau = \min \left\{ t > 0 : \frac{1 + \frac{1}{M} \# \{j, m : \kappa_{mj} \geq 1, \tau_{mj} \geq t\}}{\# \{j, m : \kappa_{mj} = 0, \tau_{mj} \geq t\} \vee 1} \leq q \right\} .$$

Finally we select the j-th variable if $\text{median}(\kappa_{mj}) = 0$ and $\text{median}(\tau_{mj}) \geq \tau$.

4.5.2 Proof of theorem 2

Let $H_0 \subseteq \{1, \ldots, p\}$ be the subset of indexes of the null features. Obviously we have

$$\widehat{FDP}(t) = \frac{\#\{j : W_j \leq -t\}}{1 \vee \#\{j : W_j \geq t\}} \geq \frac{\#\{j \in H_0 : W_j \leq -t\}}{1 \vee \#\{j : W_j \geq t\}}.$$

It is therefore enough to prove:

$$\mathbb{E}\left[\frac{\#\{j \in H_0 : W_j \leq -t\}}{1 \vee \#\{j : W_j \geq t\}}\right] \geq \mathbb{E}\left[\frac{\#\{j \in H_0 : W_j \geq t\}}{1 \vee \#\{j : W_j \geq t\}}\right] \tag{4.20}$$

For convenience, let $Z = \sum_{j \notin H_0} \mathbb{1}(W_j \geq t)$. We will prove the inequality (4.20) while conditioning expectations on both sides on $Z = 0$ and $Z > 0$. From such inequalities the desired eq. (4.20) follows directly.

We start with the case $Z > 0$. We would like to prove:

$$\mathbb{E}\left[\frac{\sum_{j \in H_0} \mathbb{1}(W_j \leq -t) - \mathbb{1}(W_j \geq t)}{1 \vee \sum_{j=1}^{n} \mathbb{1}(W_j \geq t)} \mid Z > 0\right] \geq 0 \tag{4.21}$$

Let $S = \{\eta : \eta_j = 1 \text{ for } j \notin H_0, \eta_j \in \{-1, 1\} \text{ for } j \in H_0\}$. Choose an $\eta \in S$. From the assumption, and the fact that Z only depends on non-null variables, we know that vector $(W_1\eta_1, \ldots, W_n\eta_n, Z)$ has the same distribution as $(W_1, \ldots, W_n, Z)$. Therefore, for any $\eta \in S$, eq. (4.21) is equivalent to:

$$\mathbb{E}\left[\frac{\sum_{j \in H_0} \mathbb{1}(\eta_j W_j \leq -t) - \mathbb{1}(\eta_j W_j \geq t)}{Z + \sum_{j \in H_0} \mathbb{1}(\eta_j W_j \geq t)} \mid Z > 0\right] \geq 0$$

and since the right hand side does not depend on $\eta \in S$, the inequality (4.21) is equivalent to:

$$\sum_{\eta \in S} \mathbb{E}\left[\frac{\sum_{j \in H_0} \mathbb{1}(\eta_j W_j \leq -t) - \mathbb{1}(\eta_j W_j \geq t)}{Z + \sum_{j \in H_0} \mathbb{1}(\eta_j W_j \geq t)} \mid Z > 0\right] \geq 0. \tag{4.22}$$

We can look at the sum of two "opposite" terms in the above sum, that is term for a given η and term for η with the sign of η_j flipped for $j \in H_0$. We would get:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sum_{j \in H_0} \mathbb{1}(\eta_j W_j \leq -t) - \mathbb{1}(\eta_j W_j \geq t)}{Z + \sum_{j \in H_0} \mathbb{1}(\eta_j W_j \geq t)} + \frac{\sum_{j \in H_0} \mathbb{1}(\eta_j W_j \geq t) - \mathbb{1}(\eta_j W_j \leq -t)}{Z + \sum_{j \in H_0} \mathbb{1}(\eta_j W_j \leq -t)} \mid Z > 0\right] \\
&= \mathbb{E}\left[\frac{\left(\sum_{j \in H_0} \mathbb{1}(\eta_j W_j \leq -t) - \mathbb{1}(\eta_j W_j \geq t)\right)^2}{(Z + \sum_{j \in H_0} \mathbb{1}(\eta_j W_j \geq t))(Z + \sum_{j \in H_0} \mathbb{1}(\eta_j W_j \leq -t))}\right] \geq 0
\end{aligned}$$

Since the terms in eq. (4.22) can be coupled into $2^{|H_0|-1}$ pairs whose sums are non-negative, therefore eq. (4.21) is established.

To finish, we need to prove eq. (4.20) with both sides conditioned on $Z = 0$. Let $N = \sum_{j \in H_0} \mathbb{1}(|W_j| \geq t)$. For and let $Y \sim B(N, 1/2)$ be a Bernoulli random variable. We have:

$$\mathbb{E}\left[\frac{\sum_{j \in H_0} \mathbb{1}(W_j \leq -t) - \mathbb{1}(W_j \geq t)}{1 \vee \sum_{j=1}^{n} \mathbb{1}(W_j \geq t)} \mid Z = 0, N\right] = \mathbb{E}\left[\frac{N - 2Y}{1 \vee Y}\right],$$

hence to show that eq. (4.20) with both sides conditioned on $Z = 0$ holds, it is enough to prove that:

$$\mathbb{E}\left[\frac{N - Y}{1 \vee Y}\right] \geq \mathbb{E}\left[\frac{Y}{1 \vee Y}\right].$$

The above inequality is trivially true for $N = 0$. For $N > 0$, we notice that $\mathbb{E}\left[\frac{Y}{1 \vee Y}\right] =$

$1 - 2^{-N}$, as well as

$$\mathbb{E}\left[\frac{N-Y}{1 \vee Y}\right] \geq \mathbb{E}\left[\frac{N-Y}{1+Y}\right] = 2^{-N} \sum_{k=0}^{N} \binom{N}{k} \frac{N-k}{1+k}$$
$$= 2^{-N} \sum_{k=1}^{N} \binom{N}{k-1} = 1 - 2^{-N}.$$

This ends the proof.

Remark 6. *Theorem 2 states, that if we care about FDR estimation instead of control the "+1" in the numerator is unnecessary. It is easy to see that this will also be true if we change the constant threshold t for a random threshold τ, that is independent of W. At the same time when the threshold τ is fully dependent on W (it is a function of W) the "+1" is indispensable.*

It is interesting to ask why. One way of thinking about it, is that it is an overcompensation for the error caused by optimal stopping in the choice of τ. Specifically, we can think about the classic choice of τ as starting with $\tau = 0$ and increasing it until the proportion of W_i below $-\tau$ to W_i above τ is small enough. Assuming for the sake of argument that all $|W_i|$ are different, the fraction that we base our choice of τ on, will decrease after we pass a negative W_i, and increase after we pass a positive W_i. Therefore the first time the proportion goes below a certain quantity q, we have just passed a negative W_i. On the other hand, for constant τ (or τ independent of W) one would expepct that the frequencies with which the last W_i for which $|W_i| < t$ is positive or negative would be equal - at least if we restrict ourselves to null features. In this case the "+1" in the numerator is unnecessary. The stopping rule is therefore a source of downard bias in $\widehat{FDP}(\tau)$, when we think of it as an estimator of $FDR(\tau)$. An easy fix is to change the bias downwards to bias upwards, by adding this "just removed" negative W_i to the count in the numerator.

Chapter 5

Implementations, packages and tutorials

Contents

5.1 Tutorial: R package misaem

5.1.1 Introduction of misaem

`misaem` is a CRAN package (Jiang and Mozharovskyi, 2020), to perform linear regression and logistic regression with missing data, under MCAR (Missing completely at random) and MAR (Missing at random) mechanisms. The covariates are assumed to be continuous variables. The methodology implemented is based on maximization of the observed likelihood using EM-types of algorithms. The package includes:

1. Parameters estimation:

- for linear regression, we consider a joint Gaussian distribution for covariates and response, then the `norm` package (Novo and Schafer, 2013) allows to estimate the mean vector and a variance covariance matrix with the EM algorithm and SWEEP operator (Schafer, 1997). Finally we have reshaped the outputs of the `norm` package to obtain the regression coefficient.
- for logistic regression, we use a stochastic approximation version of EM algorithm (SAEM) based on Metropolis-Hasting sampling as introduced in Chapter 2.

2. Estimation of standard deviation for estimated parameters:

- for linear regression, with classical formula 1.2.
- for logistic regression, with Louis formula.

3. Model selection procedure based on BIC.
4. Prediction on test set with missing values.

The package can be installed and loaded with the following commands:

```
# install.packages('misaem')
library(misaem)
```

5.1.2 Linear regression with missing covariates

Synthetic dataset

Let's generate a synthetic example of classical linear regression. We first generate a design matrix of size $n = 50$ times $p = 2$ by drawing each observation from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. We consider as the true values for the parameters:

$$\mu = (1, 1),$$
$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}.$$

Then, we generate the response according to the linear regression model with coefficient $\beta = (2, 3, -1)$ and variance of noise vector $\sigma^2 = 0.25$.

```
set.seed(1)
n <- 50 # number of rows
p <- 2 # number of explanatory variables

# Generate complete design matrix
library(MASS)
mu.X <- c(1, 1)
Sigma.X <- matrix(c(1, 1, 1, 4), nrow = 2)
X.complete <- mvrnorm(n, mu.X, Sigma.X)

# Generate response
b <- c(2, 3, -1) # regression coefficient
sigma.eps <- 0.25 # noise variance
y <- cbind(rep(1, n), X.complete) %*% b + rnorm(n, 0, sigma.eps)
```

Then we randomly introduced 15% of missing values in the covariates according to the MCAR (Missing completely at random) mechanism. To do so, we use the function `ampute` from the R package `mice` (van Buuren and Groothuis-Oudshoorn, 2011). For more details about how to generate missing values of different mechanisms, see Mayer et al. (2019).

```r
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
# Add missing values
yX.miss <- ampute(data.frame(y, X.complete), 0.15, patterns = matrix(
  c(0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0),
  ncol = 3, byrow = TRUE), freq = c(1, 1, 1, 2, 2, 2) / 9,
  mech = "MCAR", bycases = FALSE)
y.obs <- yX.miss$amp[, 1]               # responses
X.obs <- as.matrix(yX.miss$amp[, 2:3]) # covariates with NAs
```

Have a look at the synthetic dataset:

```r
head(X.obs)
```

```
##                X1         X2
## [1,]  0.30528180 -0.1473747
## [2,]  1.59950261  1.2164969
## [3,]  0.22508791 -0.5764402
## [4,]  2.86148303  3.8938533
## [5,]  0.05283648  2.0009229
## [6,] -1.07586521 -0.1496864
```

Check the percentage of missing values:

```r
sum(is.na(X.obs))/(n*p)
```

```
## [1] 0.17
```

Estimation for linear regression with missing values

The main function in our package to fit linear regression with missingness is `miss.lm` function. The function `miss.lm` mimics the structure of widely used function `lm` for the case without missing values. It takes an object of class `formula` (a symbolic description of the model to be fitted) and the data frame as the input. Here we apply this function with its default options.

```r
# Estimate regression using EM with NA
df.obs = data.frame(y, X.obs)
miss.list = miss.lm(y~., data = df.obs)
```

```
## Iterations of EM:
## 1...2...3...4...5...6...7...8...9...10...11...12...13...14...15...16...17...
```

Then it returns an object of self-defined class `miss.lm`, which consists of the estimation of parameters, their standard error and observed log-likelihood. We can print or summarize the obtained results as follows:

```r
print(miss.list)
```

```
##
## Call:  miss.lm(formula = y ~ ., data = df.obs)
##
## Coefficients:
## (Intercept)           X1           X2
##       1.942        3.052       -1.004
## Standard error estimates:
## (Intercept)           X1           X2
##     0.04171      0.03484      0.01936
## Log-likelihood: 31.85
```

```r
print(summary(miss.list))
```

```
##
## Call:
## miss.lm(formula = y ~ ., data = df.obs)
##
## Coefficients:
##              Estimate  Std. Error
## (Intercept)   1.94205   0.04171
## X1            3.05205   0.03484
## X2           -1.00424   0.01936
## Log-likelihood: 31.852
```

```r
summary(miss.list)$coef
```

```
##              Estimate Std. Error
## (Intercept)  1.942050 0.04170924
## X1           3.052050 0.03483511
## X2          -1.004244 0.01936094
```

Self-defined parameters can be also taken such as the maximum number of iterations (`maxruns`), the convergence tolerance (`tol_em`) and the logical indicating if the iterations should be reported (`print_iter`).

```
# Estimate regression using self-defined parameters
miss.list2 = miss.lm(y~., data = df.obs, print_iter = FALSE,
                     maxruns = 500, tol_em = 1e-4)
print(miss.list2)
```

```
##
## Call:  miss.lm(formula = y ~ ., data = df.obs, print_iter = FALSE,
##     maxruns = 500,  tol_em = 1e-04)
##
## Coefficients:
## (Intercept)           X1           X2
##       1.942        3.052       -1.004
## Standard error estimates:
## (Intercept)           X1           X2
##     0.04175      0.03487      0.01938
## Log-likelihood: 31.85
```

Model selection

The function `miss.lm.model.select` adapts a BIC criterion and step-wise method to return the best model selected. We add a null variable with missing values to check if the function can distinguish it from the true variables.

```
# Add null variable with NA
X.null <- mvrnorm(n, 1, 1)
patterns <- runif(n)<0.15 # missing completely at random
X.null[patterns] <- NA
X.obs.null <- cbind.data.frame(X.obs, X.null)

# Without model selection
df.obs.null = data.frame(y, X.obs.null)
miss.list.null = miss.lm(y~., data = df.obs.null)
```

```
## Iterations of EM:
## 1...2...3...4...5...6...7...8...9...10...11...12...13...14...15...
```

```
print(miss.list.null)
```

```
##
## Call:  miss.lm(formula = y ~ ., data = df.obs.null)
##
```

```
## Coefficients:
## (Intercept)           X1           X2       X.null
##     1.88617      3.05883     -1.00391      0.04435
## Standard error estimates:
## (Intercept)           X1           X2       X.null
##     0.05670      0.03438      0.02125      0.02853
## Log-likelihood: 15.87
```

```r
# Model selection
miss.model = miss.lm.model.select(y, X.obs.null)
print(miss.model)
```

```
##
## Call:  miss.lm(formula = Y ~ ., data = df, print_iter = FALSE)
##
## Coefficients:
## (Intercept)           X1           X2
##       1.942        3.052       -1.004
## Standard error estimates:
## (Intercept)           X1           X2
##     0.04171      0.03484      0.01936
## Log-likelihood: 31.85
```

Prediction on test set

In order to evaluate the prediction performance, we generate a test set of size $nt = 20$ times $p = 2$ following the same distribution as the previous design matrix, and we add or not 15% of missing values.

```r
# Prediction
# Generate dataset
set.seed(200)
nt <- 20  # number of new observations
Xt <- mvrnorm(nt, mu.X, Sigma.X)

# Add missing values
Xt.miss <- ampute(data.frame(Xt), 0.15, patterns = matrix(
  c(0, 1, 1, 0),
  ncol = 2, byrow = TRUE), freq = c(1, 1) /2,
  mech = "MCAR", bycases = FALSE)
Xt.obs <- as.matrix(Xt.miss$amp) # covariates with NAs
```

The prediction can be performed for a complete test set:

```r
#train with NA + test no NA
miss.comptest.pred = predict(miss.list2, data.frame(Xt), seed = 100)
print(miss.comptest.pred)
```

```
##  [1]  3.3878210  2.6112345 -0.5562864  6.5926842  2.9231974  8.0234969
##  [7]  0.8286503  3.9363413  6.7515266  3.3517064  6.8156632  2.2406832
## [13]  2.0321507  5.9852215  7.8101528  5.0863422  4.2238612  4.4541193
## [19]  3.5522691  3.0003519
```

And we can also apply the function when both train set and test set have missing values:

```r
#both train & test with NA
miss.pred = predict(miss.list2, data.frame(Xt.obs), seed = 100)
print(miss.pred)
```

```
##  [1]  3.3878210  3.4804264 -0.5562864  6.5926842  2.9231974  8.0234969
##  [7]  0.1435715  3.9363413  6.7515266  3.3517064  6.8156632  2.2406832
## [13]  2.0321507  5.9150631  7.8101528  3.5570286  4.2238612  4.4541193
## [19]  3.5522691  3.0003519
```

5.1.3 Logistic regression with missing covariates

Synthetic dataset

We first generate a design matrix of size $n = 500$ times $p = 5$ by drawing each observation from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. Then, we generate the response according to the logistic regression model.

We consider as the true values for the parameters

$$\begin{aligned}\beta &= (0, 1, -1, 1, 0, -1),\\ \mu &= (1, 2, 3, 4, 5),\\ \Sigma &= \text{diag}(\sigma) C \text{diag}(\sigma),\end{aligned}$$

where the σ is the vector of standard deviations

$$\sigma = (1, 2, 3, 4, 5)$$

,
and C the correlation matrix

$$C = \begin{bmatrix} 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.3 & 0.6 \\ 0 & 0 & 0.3 & 1 & 0.7 \\ 0 & 0 & 0.6 & 0.7 & 1 \end{bmatrix}.$$

```r
# Generate dataset
set.seed(200)
n <- 500  # number of subjects
p <- 5     # number of explanatory variables
mu.star <- 1:p  #rep(0,p)  # mean of the explanatory variables
sd <- 1:p # rep(1,p) # standard deviations
C <- matrix(c(   # correlation matrix
1,   0.8, 0,   0,   0,
0.8, 1,   0,   0,   0,
0,   0,   1,   0.3, 0.6,
0,   0,   0.3, 1,   0.7,
0,   0,   0.6, 0.7, 1), nrow=p)
Sigma.star <- diag(sd)%*%C%*%diag(sd) # covariance matrix
beta.star <- c(1, -1, 1, 1, -1) # coefficients
beta0.star <- 0  # intercept
beta.true = c(beta0.star,beta.star)

# Design matrix
X.complete <- matrix(rnorm(n*p), nrow=n)%*%chol(Sigma.star)+
              matrix(rep(mu.star,n), nrow=n, byrow = TRUE)

# Reponse vector
p1 <- 1/(1+exp(-X.complete%*%beta.star-beta0.star))
y <- as.numeric(runif(n)<p1)
```

Then we randomly introduced 10% of missing values in the covariates according to the MCAR (Missing completely at random) mechanism.

```r
# Generate missingness
set.seed(200)
p.miss <- 0.10
patterns <- runif(n*p)<p.miss # missing completely at random
X.obs <- X.complete
X.obs[patterns] <- NA
```

Have a look at the synthetic dataset:

```r
head(X.obs)
```

```
##            [,1]        [,2]      [,3]         [,4]        [,5]
## [1,] 1.0847563  1.71119812 5.0779956  9.731254821 13.02285225
## [2,] 1.2264603  0.04664033 5.3758000  6.383093558  4.84730504
## [3,] 1.4325565  1.77934455        NA  8.421927692  7.26902254
## [4,] 1.5580652  5.69782193 5.5942869 -0.440749372 -0.96662931
## [5,] 1.0597553 -0.38470918 0.4462986  0.008402997  0.04745022
## [6,] 0.8853591  0.56839374 3.4641522  7.047389616          NA
```

Estimation for logistic regression with missingness

The main function for fitting logistic regression with missing covariates in our package is `miss.glm` function, which mimics the structure of widely used function `glm`. Note that we don't need to specify the binomial family in the input of `miss.glm` function. Here we apply this function with its default options, and then we can print or summarize the obtained results as follows:

```
df.obs = data.frame(y, X.obs)

#logistic regression with NA
miss.list = miss.glm(y~., data = df.obs, seed = 100)
```

```
## Iteration of SAEM:
## 50 100 150 200
```

```
print(miss.list)
```

```
##
## Call:  miss.glm(formula = y ~ ., data = df.obs, seed = 100)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4           X5
##    -0.03659      1.50705     -1.28208      1.12342      1.03435     -1.07691
## Standard error estimates:
## (Intercept)           X1           X2           X3           X4           X5
##      0.3210       0.3446       0.2056       0.1408       0.1240       0.1284
## Log-likelihood: -171.7
```

```
print(summary(miss.list))
```

```
##
## Call:
## miss.glm(formula = y ~ ., data = df.obs, seed = 100)
##
## Coefficients:
##              Estimate  Std. Error
## (Intercept)  -0.03659   0.32104
## X1            1.50705   0.34456
## X2           -1.28208   0.20560
## X3            1.12342   0.14076
## X4            1.03435   0.12396
## X5           -1.07691   0.12843
## Log-likelihood: -171.74
```

```r
summary(miss.list)$coef
```

```
##                Estimate Std. Error
## (Intercept) -0.03659218  0.3210369
## X1           1.50704588  0.3445570
## X2          -1.28208040  0.2056000
## X3           1.12341764  0.1407630
## X4           1.03435057  0.1239566
## X5          -1.07690679  0.1284274
```

Model selection

To perform model selection with missing values, we adapt criterion BIC and step-wise method. The function `miss.glm.model.select` outputs the best model selected. With the current implementation, when p is greater than 20, it may encounter computational difficulties for the BIC based model selection. In the following simulation, we add a null variable with missing values to check if the function can distinguish it from the true variables.

```r
# Add null variable with NA
X.null <- mvrnorm(n, 1, 1)
patterns <- runif(n)<0.10 # missing completely at random
X.null[patterns] <- NA
X.obs.null <- cbind.data.frame(X.obs, X.null)

# Without model selection
df.obs.null = data.frame(y, X.obs.null)
miss.list.null = miss.glm(y~., data = df.obs.null)
```

```
## Iteration of SAEM:
## 50 100 150 200
```

```r
print(miss.list.null)
```

```
##
## Call:  miss.glm(formula = y ~ ., data = df.obs.null)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4           X5
##    -0.08280      1.52860     -1.29067      1.13314      1.05171     -1.09399
##      X.null
##     0.03964
## Standard error estimates:
## (Intercept)           X1           X2           X3           X4           X5
##      0.3585       0.3514       0.2084       0.1417       0.1241       0.1291
##      X.null
```

```
##      0.1666
## Log-likelihood: -171.4
```

```r
# model selection for SAEM
miss.model = miss.glm.model.select(y, X.obs.null)
print(miss.model)
```

```
##
## Call:  miss.glm(formula = Y ~ ., data = df, print_iter = FALSE,
## subsets = subset_choose)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4           X5
##    -0.06956      1.55837     -1.30913      1.14401      1.06008     -1.10143
## Standard error estimates:
## (Intercept)           X1           X2           X3           X4           X5
##      0.3244       0.3500       0.2094       0.1440       0.1279       0.1317
## Log-likelihood: -172
```

Prediction on test set

In order to evaluate the prediction performance, we generate a test set of size $nt = 100$ times $p = 5$ following the same distribution as the design matrix, and without and with 10% of missing values. We evaluate the prediction quality with a confusion matrix.

```r
# Generate test set with missingness
set.seed(200)
nt = 100
X.test <- matrix(rnorm(nt*p), nrow=nt)%*%chol(Sigma.star)+
          matrix(rep(mu.star,nt), nrow = nt, byrow = TRUE)

# Generate the test set
p1 <- 1/(1+exp(-X.test%*%beta.star-beta0.star))
y.test <- as.numeric(runif(nt)<p1)

# Generate missingness on test set
p.miss <- 0.10
X.test[runif(nt*p)<p.miss] <- NA

# Prediction on test set
pr.saem <- predict(miss.list, data.frame(X.test))

# Confusion matrix
pred.saem = (pr.saem>0.5)*1
table(y.test,pred.saem )
```

```
##       pred.saem
## y.test  0  1
##      0 34  8
##      1  6 52
```

5.2 Tutorial: R package ABSLOPE

5.2.1 Introduction of ABSLOPE

ABSLOPE is a package to perform high-dimensional model selection with missing values, under MCAR (Missing completely at random) and MAR (Missing at random) mechanisms. We target at sparse linear model and the objective is to simultaneously perform variable selection and parameter estimation, despite the missing values among the covariates. The implemented method, adaptive Bayesian version of SLOPE (ABSLOPE), as described in Algorithm 5, addresses these issues by embedding the sorted l_1 penalization (Bogdan et al., 2015) (an extension of LASSO (Tibshirani, 1996) within a Bayesian framework. Specifically, the aim of model selection is controlling false discovery rate (FDR).

The package can be installed and loaded with the following commands:

```
# library(devtools)
# install_github("wjiang94/ABSLOPE")
library(ABSLOPE)
```

```
##
## Attaching package: 'ABSLOPE'
```

```
## The following object is masked from 'package:stats':
##
##     power
```

With following example of synthetic data set, we illustrate how to use the package. ## Synthetic dataset Let's generate a synthetic example of linear regression. We first generate a design matrix where the number of observations n and dimension size p are equally large with $n = p = 100$. We generate each observation from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ and then standardize the matrix. We consider all elements in μ equals to 0 and the covariance matrix as a Toeplitz structure, i.e.,

$$\Sigma = \text{toeplitz}(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho^{p-2} & \rho^{p-1} \\ \rho & 1 & \ddots & \cdots & \rho^{p-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{p-2} & \cdots & \ddots & \ddots & \rho \\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1 \end{pmatrix}_{p \times p},$$

where the correlation $\rho = 0.5$. This structure indicates that, adjacent pairs of covariates are highly correlated and those further away are less correlated, for instance, in microarry study (Guo et al., 2006), genes are correlated due to their distance in the regularity pathway.

we generate the response according to the linear regression model with coefficient β. In a sparse setting, we let only the first 10 predictors non-zero and all equal to $2\sqrt{2\log p}$ and the variance of noise vector $\sigma^2 = 1$.

```
set.seed(100)
n <- 100 # number of rows
p <- 100 # number of explanatory variables
mu <- rep(0, p) # mean of covariates distribution
corr <- 0.5 # correlation
Sigma <- toeplitz(corr^(0:(p-1))) # variance of covariates distribution

# Design matrix
X.comp <- matrix(rnorm(n*p), nrow = n) %*% chol(Sigma) +
  matrix(rep(mu,n), nrow = n, byrow = TRUE)
X.comp <- scale(X.comp)/sqrt(n) # Standardization

# Coefficient and response vectors
signallevel <- 3 # signal strength
amplitude <- signallevel*sqrt(2*log(p))  # signal amplitude
nspr <- 10 # number of non-zero predictors
sigma <- 1 # noise variance
nonzero <- sample(p, nspr)
beta <- amplitude * (1:p %in% nonzero) # regression coefficient
y <- X.comp %*% beta +  sigma*rnorm(n)
```

Then we randomly introduced 10% of missing values in the covariates according to the MCAR (Missing completely at random) mechanism.

```
# Add missing values
X.obs <- X.comp
p.miss <- 0.1
patterns <- runif(n*p)< p.miss # missing completely at random
X.obs[patterns] <- NA
```

Have a look at the synthetic dataset:

```
X.obs[1:5, 1:5]
```

```
##              [,1]        [,2]        [,3]         [,4]        [,5]
## [1,] -0.049485626 -0.06877051 -0.02834139 -0.164145080 -0.17609787
## [2,]  0.012600891  0.15429398  0.03219760 -0.065157685 -0.03263146
## [3,] -0.008016932 -0.05707007  0.05397834 -0.057504622  0.10270804
## [4,]  0.086593834  0.14518582  0.10977987  0.001515076 -0.07490049
## [5,]  0.011174444 -0.15181138          NA  0.052723903 -0.04856800
```

Alternatively the function `data.generation` can also help to generate the dataset with missing values.

```
data.list = data.generation(n, p, nspr, p.miss, mu, Sigma,
                            signallevel, mec='MCAR', sigma=sigma)
X.obs = data.list$X.obs
X.comp = data.list$X
y = data.list$y
beta = data.list$beta
```

5.2.2 Estimation and model selection with missing values—Algorithm ABSLOPE

The main function in our package to fit high-dimensional linear regression with missingness is ABSLOPE function. It takes the data frame and regularizing sequence as the inputs. Here we set the regularizing parameters as Benjamini-Hochberg sequence (Benjamini and Hochberg, 1995) in order to achieve FDR control at target level 0.1, and then we apply the main function with its default options.

```
# ABSLOPE
lambda = create_lambda_bhq(ncol(X.obs),fdr=0.10)
list.res = ABSLOPE(X.obs, y, lambda)
```

Then it returns the estimation of parameters. We can print the obtained results as follows:

```
print(list.res$beta)
```

```
##   [1]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
##   [8]  0.000000  0.000000  0.000000  0.000000  0.000000 10.557724  0.000000
##  [15]  0.000000  0.000000  0.000000  7.695746  0.000000 11.443353  8.076250
##  [22]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  7.878916
##  [29]  0.000000  0.000000  0.000000  0.000000  0.000000 10.515769  0.000000
##  [36]  0.000000  0.000000  0.000000  0.000000  9.149652  0.000000  0.000000
##  [43]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
##  [50]  0.000000  0.000000  9.354592  0.000000 10.772740  0.000000  7.363948
##  [57]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
##  [64]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
##  [71]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
##  [78]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
##  [85]  0.000000  0.000000  3.755438  0.000000  0.000000  0.000000  0.000000
##  [92]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
##  [99]  0.000000  0.000000
```

To check which variables are selected:

```
selected = which(list.res$beta!=0)
print(selected)
```

```
##  [1] 13 18 20 21 28 34 40 52 54 56 87
```

Then the power and FDR can be calculated:

```
power(beta, selected)
```

```
## [1] 1
```

```
fdp(beta, selected)
```

```
## [1] 0.09090909
```

Accelarated version of ABSLOPE and implementation with Rcpp

A simplified version of the algorithm has also been developed as described in Algorithm 6, where the sampling procedure in the algorithm ABSLOPE is replaced by conditional expectation. In addition, advanced implementations using Rcpp (Eddelbuettel and Balamuta, 2017) integrate C codes which contributes to efficient solutions to large scale problems.

```
# Accelarated version
lambda = create_lambda_bhq(ncol(X.obs),fdr=0.10)
list.res.approx = SLOBE(X.obs, y, lambda)
```

Summarize the results with power and FDR:

```
selected.approx = which(list.res.approx$beta!=0)
power(beta, selected.approx)
```

```
## [1] 1
```

```
fdp(beta, selected.approx)
```

```
## [1] 0.09090909
```

Compare the calculating time between the original ABSLOPE and approximated version:

```
list.res$time # Execution time for ABSLOPE
```

```
## Time difference of 18.09972 secs
```

```
list.res.approx$time # Execution time for approximated version
```

```
## Time difference of 1.190781 secs
```

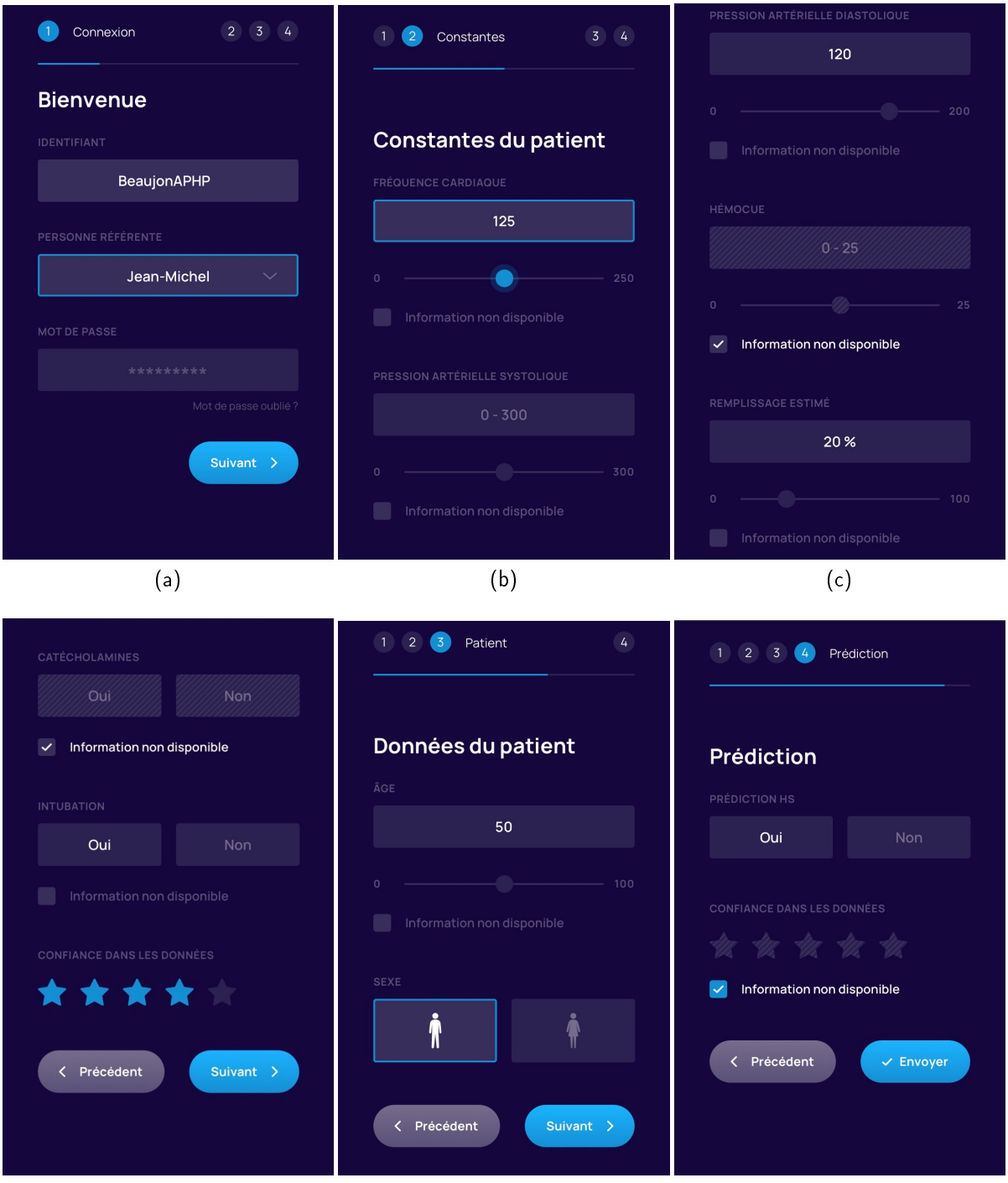


Figure 5.1: Screenshots of TraumaBase mobile application.

5.3 TraumaBase mobile application

The TraumaBase Group has decided to take the algorithms developped into real-time application. In the first step, a mobile application is under developement, as shown in Figure 5.1. This application is designed specifically for the emergency doctors, who can log in with the identifier of the hosipital center as shown in the subplot (a). The purpose of the following secure form is to collect data from severely traumatized patients from both a health and scientific perspective, and then to send them at regular intervals to the APHP and to the TraumaBase dataset. Data colleted include heart rate and systolic blood pressure, diastolic blood pressure, the hemoglobin *etc.* as presented in the suplots (b) to (e). In a final step as shown in the subplot (f), the application asks the doctor whether he thinks the patient will have a hemorragic shock; In the meanwhile, the application will predict the probability that the patient can have a hemorragic shock despite the inavailabitily of some measurements in the form, using the decision tool `misaem` package as described in Section 5.1. Then we can compare the doctor's prediction with the prediction from `misaem`, and finally we will have access to the ground truth. In this way, both medical research and statististical analysis on TraumaBase dataset would be improved. Note that in the future use of the application, the procedure shown in the subplot (f) will moved to the first screen, just after the screen (a). In this way to avoid the influence of `misaem` results, the doctor will first give his opnion then fill in the form.

Chapter 6

Conclusion

The objective of this thesis was to provide an efficient and comprehensive statistical methodology to handle the problem of statistical inference with the existence of missing values, in particular in regression based methods. Even if regression is one of the most widely used basic techniques, and knowing how to carry it out with missing data seems indispensable, both the applications and implementations are restricted to simple models, in low-dimension or with simple missing pattern.

In this thesis, we developed a complete framework for dealing with missing values when performing regression analysis, from estimation to model selection, from prediction to implementation and with the possibility to handle high-dimensional data. The proposed methods provide a good balance between the quality of inference and the efficiency of the calculation. The framework is grounded in the theoretical aspects of the likelihood based methods and EM types algorithm; and for high-dimensional data, we focus on methods that control the false discovery rate.

We started with the problem of fitting logistic regression model with missing covariates. In Chapter 2, we proposed a complete approach based on a stochastic approximation version of the EM (SAEM) algorithm in order to perform statistical inference with missing values, including the estimation of the parameters and their variance, the derivation of confidence intervals, and also a model selection procedure. The problem of predicting new observations on a test set with missing covariate data was also discussed. Supported by a simulation study in which the method is compared to previous ones, it has proven to be computationally efficient, with good coverage and variable selection properties.

Based on the same computational tool—SAEM algorithm, in Chapter 3 we looked beyond data with a regular sample size, but focused on high-dimensional data where the number of features is comparable or even larger than the number of observations. The proposed new procedure—adaptive Bayesian SLOPE (ABSLOPE)—combines the sorted l_1 regularization together with a Bayesian framework which effectively handles latent variable modeling. Through extensive simulations, we demonstrated satisfactory performance in terms of power, false discovery rates and estimation bias in a wide range of scenarios.

Inspired by the problem specified previously in Chapter 3, we tackled also controlled model selection with high-dimensional and incomplete data, but without specifying a parametric regression model. To do so, in Chapter 4, a model-X framework was supposed where the conditional distribution of the response given the covariates was not specified, but the joint

distribution of covariates was known. Knockoff methodology and multiple imputations were combined together with specifically designed aggregation rules. Both theoretical guarantees and simulation evaluations were provided for FDR control despite missing values.

Finally, equipped the new methodologies for estimation and model selection problem with missing values, we implemented two open source R packages `misaem` and `ABSLOPE` available in CRAN. In Chapter 5, detailed instructions and tutorials were provided for users.

Another essential contribution has been to assist decision making in the medical field with missing data. As mentioned almost in all chapters, the developed approaches were illustrated on TraumaBase, for instance, by predicting the occurrence of hemorrhagic shock, a leading cause of early preventable death in severe trauma cases; and by predicting platelet levels using pre-hospital and in-hospital data with a large number of missing values. The logistic regression model proposed with the missing values mentioned above improved the current red flag procedure, to identify patients with a high risk of severe hemorrhage. Moreover, as introduced in Chapter 5, a mobile application designed for emergency doctors is currently under development in order take the algorithms into real-time application.

The contributions of this thesis paved the way for future research in both theoretical and applied directions.

A first extension of the proposed methods would be to deal with mixed covariates with both continuous and categorical, ordinal and binary data. Even if we experimented first ideas using GLOM model as shown in the supplementary materials in Chapter 2, the efficiency of the proposed algorithm still needs to be improved in order to apply it on more complex cases and on real data, especially for high-dimensional data. One alternative could be to model the mixed data using Gaussian copula (Zhao and Udell, 2019) which fits arbitrary marginals for continuous variables and handles ordinal variables with many levels; then efficient EM algorithm can be applied on the Gaussian copula model. However, estimating the marginal distributions with missing values is still challenging even under MAR.

In the same vein to consider other types of variables, since the sorted penalty l_1 can also be used in generalized linear models (Abramovich and Grinshtein, 2018), we can combine the problematic of Chapter 2 and that of Chapter 3, to tackle the problem of logistic regression in high dimensions with missing values and to also show the control of FDR, and then compare it to the method based on knockoff proposed in Chapter 4.

Another important line of research would be to consider another missing mechanism, namely the MNAR case, which is notoriously known to be difficult to handle. Indeed, the suggested methods in this thesis are dedicated to MAR values and MNAR can also be frequent in application. For instance, when a patient's condition is quite critical with extreme low or high value of heart rate, physicians would rather provide emergency care than do measurement. The difficulty with these missing MNAR data is that often it is necessary to specify the model that generated the missing data, and therefore to have strong a priori on the parametric forms. However, part of the literature tries to take into account the mechanism without modeling it. For instance, Mohan et al. (2018) suggested such an approach mainly for a self-masked mechanism, *i.e.*, the lack depends only on the missing variable itself in a regression framework. In recent work (Mohan and Pearl, 2014; Mohan et al., 2013, 2018), the missing data have been treated as a causal inference problem and graph-based procedures for consistently estimating parameters have been proposed to efficiently handle the MNAR case. Further extension of high-dimensional model selection in the MNAR case

needs to be explored.

On the theoretical part, there are improvements to be made to prove the FDR control of ABSLOPE in Chapter 3 and missKnockoff in Chapter 4, with less restricted assumptions on the data. Moreover, we need to consider testing unconditional independence (Candes et al., 2018) with missing values, in order to use the information provided by the responses to improve the power for missKnockoff.

Alternatively, we can also extend the method to handle missing values to other models, such as hierarchical Bayesian model (Yekutieli and Weinstein, 2019) which addresses large-scale inference without parametric assumptions and aims at the empirical distribution of the parameter vector, in a fully Bayesian framework with no previous information on problem.

Last but not least, we hope that the methods developed can help the credibility of the use of statistics and machine learning to improve health care. In order to be used and trusted, the methods which manage missing values must be interpretable, transparent and founded from a theoretical point of view, but also be able to be applied to large modern data.

Appendix A

Synthèse substantiel (en langue française)

État de l'art sur la problématique des valeurs manquantes

Les données manquantes existent dans presque tous les domaines de la recherche empirique. Il existe plusieurs raisons à cela, notamment la non-réponse à une enquête, l'indisponibilité des mesures et la perte de données. Effectuer une analyse statistique sur des ensembles de données comportant des données manquantes nécessite souvent de mettre en place des connaissances supplémentaires sur la manière dont les données manquantes sont générées. Le processus par lequel les données deviennent incomplètes est appelé le mécanisme de données manquantes (Rubin, 1976; Seaman et al., 2013), et comprend les trois types suivants : *i)* Manque complètement au hasard (MCAR), dans lequel le manque de données est indépendant à la fois des valeurs observées et des valeurs manquantes ; *ii)* Manque au hasard (MAR), dans lequel le manque de données est indépendant des valeurs manquantes, compte tenu des données observées. Les données manquantes avec MCAR et MAR sont appelées des non-réponses ignorables, car l'estimation du maximum de vraisemblance peut être obtenue en ignorant ces mécanismes. *iii)* Lorsque l'absence de données dépend des valeurs manquantes elles-mêmes, compte tenu des données observées, le processus est appelé non-réponse non aléatoire (MNAR), appelé non-réponse non ignorable car il est souvent nécessaire de modéliser le mécanisme qui génère les valeurs manquantes pour faire une inférence.

La pratique la plus courante pour traiter les données manquantes, à savoir l'analyse complète des cas (ou la suppression par liste), qui limite l'analyse aux observations sans attributs manquants, entraîne une perte d'informations et un biais d'estimation, sauf si les données manquantes sont MCAR. Il faut vraiment souligner que cette approche n'est plus possible dans un contexte à grande échelle. Comme le dit Zhu et al. (2019) : "L'une des ironies du travail avec de grandes données est que les données manquantes jouent un rôle toujours plus important, et présentent souvent de sérieuses difficultés d'analyse. Pour illustrer l'inadéquation d'une analyse de cas complète avec de grandes données, ils imaginent un ensemble de données avec des observations n et des variables p où chaque entrée a une probabilité de 1% d'être manquante indépendamment. Si $p = 5$, alors l'analyse de cas complète peut être acceptable puisque nous avons encore environ 95% d'observations ;

cependant, lorsque la dimension est beaucoup plus grande, comme $p = 300$, seules 5% des lignes complètes sont retenues.

De nombreuses méthodes statistiques ont été développées pour traiter les valeurs manquantes (Schafer, 1997; Little and Rubin, 2019; van Buuren, 2018; Josse and Reiter, 2018; Mayer et al., 2019; Mohan et al., 2013) dans un cadre inférentiel, c'est-à-dire lorsque l'objectif est d'estimer des paramètres et leur variance à partir de données incomplètes. Une approche populaire pour traiter les valeurs manquantes est l'imputation, qui consiste à remplacer les valeurs manquantes par des valeurs plausibles pour obtenir des données complètes qui peuvent être analysées par n'importe quelle méthode. On peut soit imputer selon un modèle commun, soit utiliser une approche de modélisation entièrement conditionnelle (van Buuren, 2018). Parmi les méthodes puissantes figurent l'imputation par forêt aléatoire (Stekhoven and Buehlmann, 2012) et par des méthodes de rang inférieur (Josse and Husson, 2016; Robin, 2019; Udell and Townsend, 2019). Plus récemment, les contributions comprennent également des méthodes d'imputation basées sur des techniques d'apprentissage approfondi, telles qu'un auto-codeur variationnel (Mattei and Frellsen, 2019; Ma et al., 2018) et des réseaux adversaires générateurs (Yoon et al., 2018), cependant, ces méthodes nécessitent un ensemble de données complet pour former au mieux le modèle. Sans hypothèses paramétriques, la stratégie bayésienne non paramétrique (Murray and Reiter, 2016) ou l'approche récente utilisant le transport optimal (Muzellec et al., 2020) sont des tentatives dans ce sens. Néanmoins, même si nous parvenons à imputer en préservant au mieux la distribution conjointe et marginale des données, une seule imputation ne peut pas refléter l'incertitude associée à la prévision des valeurs manquantes. Pour atteindre cet objectif, l'imputation multiple (MI) (Rubin, 2009; van Buuren and Groothuis-Oudshoorn, 2011) consiste à générer plusieurs valeurs plausibles pour chaque donnée manquante (pour refléter la variance de la prédiction compte tenu des données observées et du modèle d'imputation) conduisant à différents ensembles de données imputées. Ensuite, l'analyse est effectuée sur chaque ensemble de données imputées et les résultats sont combinés de manière à ce que la variance finale tienne compte de la variabilité supplémentaire due aux valeurs manquantes. La figure A.1 illustre les principales étapes décrites ci-dessus.

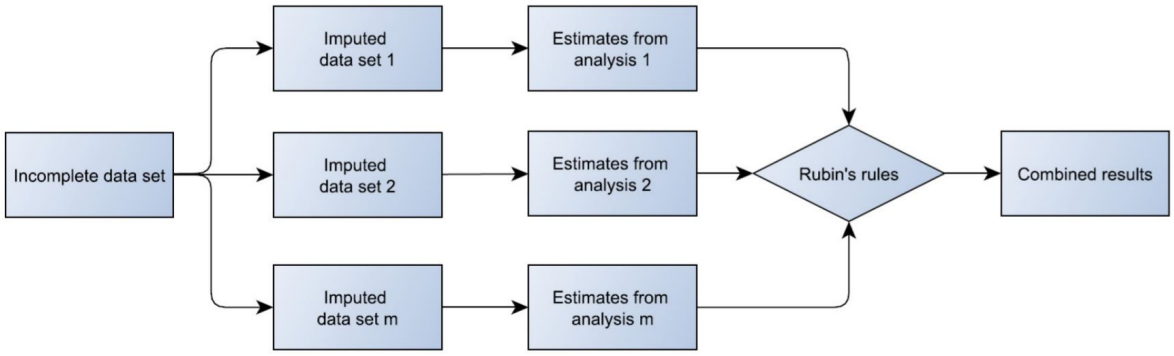


Figure A.1: Procédure d'imputation multiple.

Une alternative pour traiter les valeurs manquantes consiste à modifier les processus d'estimation afin qu'ils puissent être appliqués à des données incomplètes. Par exemple,

on peut utiliser l'algorithme EM (Dempster et al., 1977) pour obtenir l'estimation du maximum de vraisemblance malgré les valeurs manquantes, accompagné d'un algorithme EM supplémenté (Meng and Rubin, 1991) ou la formule de Louis (Louis, 1982) pour estimer la variance.

La figure A.2 illustreles les principales idées de l'EM: l'objectif est de maximiser la courbe bleue, pour ce faire, nous approximons sa limite inférieure, la courbe verte, puis nous mettons itérativement à jour les estimations en maximisant la courbe verte.

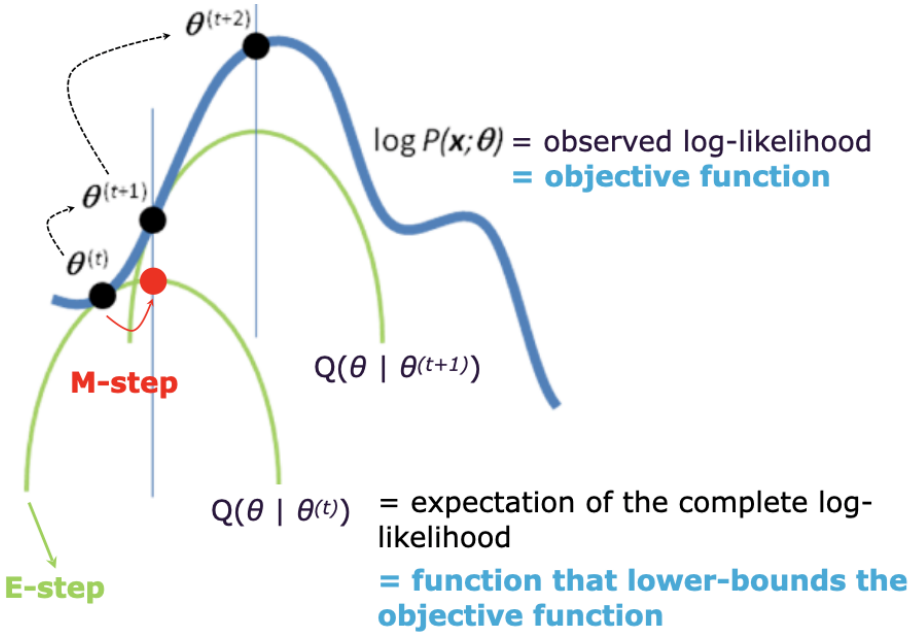


Figure A.2: L'algorithme EM.

Cette stratégie est valable dans le cadre des mécanismes MAR. Même si cette approche est parfaitement adaptée aux problèmes spécifiques d'inférence avec des valeurs manquantes, il existe peu de solutions ou d'implémentations disponibles, même pour des modèles simples tels que la régression logistique.

Cela peut s'expliquer par le fait que, contrairement à l'imputation, l'algorithme EM repose explicitement sur des hypothèses paramétriques fortes et qu'il faut dériver une approche pour chaque technique statistique. Mais l'avantage évident de l'algorithme EM est que l'on peut s'attendre à un meilleur contrôle des propriétés statistiques de l'approche développée. En outre, comme il est souvent impossible d'obtenir une forme explicite pour l'algorithme EM, des méthodes d'échantillonnage ont été utilisées telles que l'échantillonnage Monte Carlo (Ibrahim et al., 1999), l'échantillonnage par rejet adaptatif (Gilks and Wild, 1992), mais elles prenaient beaucoup de temps, ce qui peut également expliquer pourquoi les algorithmes basés sur l'EM n'ont pas été utilisés en pratique.

Une autre partie de la littérature se concentre sur les problèmes d'apprentissage statistique où l'objectif est de prédire au mieux une variable de réponse en sachant que les covariables ont des données manquantes. Par exemple, Josse et al. (2019) montre la cohérence de l'imputation moyenne simple dans la prédiction. Les Kapelner and Bleich (2015) fournissent des résultats empiriques de la performance prédictive des arbres de décision avec des covari-

ables manquantes.

Même s'il existe une multitude de méthodes pour gérer les données manquantes (plus de 150 paquets existent dans le logiciel R, comme indiqué dans Mayer et al. (2019)), il est surprenant de constater qu'il n'existe vraiment que très peu de solutions pour sélectionner des modèles et des variables avec des données manquantes, en particulier dans les grandes dimensions. Dans cette thèse, nous considérons le cadre de l'inférence statistique avec les covariables manquantes et nous développons de nouvelles méthodologies d'estimation des paramètres et de sélection des modèles pour traiter les valeurs manquantes. Ces travaux sont motivés par un problème pratique sur un registre de traumatismes graves pour la prise de décision.

Projet TraumaBase

Notre travail est motivé par une collaboration avec le groupe TraumaBase[1] de l'APHP (Assistance publique - Hôpitaux de Paris), qui se consacre à la prise en charge des patients gravement traumatisés. Les traumatismes majeurs désignent les blessures qui entraînent une invalidité prolongée ou mettent en danger la vie d'une personne, comme les blessures dues aux accidents de la route, aux violences interpersonnelles et aux chutes. L'Organisation mondiale de la santé a récemment indiqué que les traumatismes majeurs sont une source importante de mortalité et de morbidité dans le monde entier (Hay et al., 2017). En particulier, les traumatismes majeurs sont la première cause de mortalité et la deuxième cause d'invalidité dans la tranche d'âge 16–45 ans, tandis que le choc hémorragique et les lésions cérébrales traumatiques sont les deux principales causes de décès précoce évitable chez les patients souffrant de traumatismes graves (Dutton et al., 2010; Kauvar and Wade, 2005).

Le parcours d'un patient traumatisé comporte plusieurs étapes : du *1)* site de l'accident, où les soins sont généralement dispensés par des équipes de soins d'urgence, au transfert vers la *2)* salle de réanimation d'un centre de traumatologie, où des interventions immédiates telles que l'évaluation par scanner, la chirurgie d'urgence ou la radiologie peuvent être organisées, puis à l'admission en *3)* unité de soins intensifs pour l'optimisation du soutien en cas de dysfonctionnement d'un organe, et enfin *4)* une prise en charge complète à l'hôpital, comme le montre la figure A.3.

En raison des environnements très stressants et multi-agents impliqués, il est prouvé que la gestion du patient, même en cas de traumatisme mature dépasse souvent les délais acceptables (Hamada et al., 2014). En outre, des divergences peuvent être observées entre les diagnostics faits par les médecins urgentistes dans l'ambulance et ceux faits lorsque le patient arrive au centre de traumatologie (Hamada et al., 2015). De telles divergences peuvent entraîner de mauvais résultats, comme un contrôle inadéquat des hémorragies et un retard de transfusion.

Pour améliorer la prise de décision et les soins aux patients, 19 centres de traumatologie français ont collaboré depuis 2011 pour collecter des données cliniques détaillées de haute qualité les données du site de l'accident jusqu'à la l'hôpital. Certains centres ont rejoint TraumaBase après janvier 2011. La base de données qui en résulte, TraumaBase, est un registre polycentrique prospectif des traumatismes qui est continuellement mis à jour et qui

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

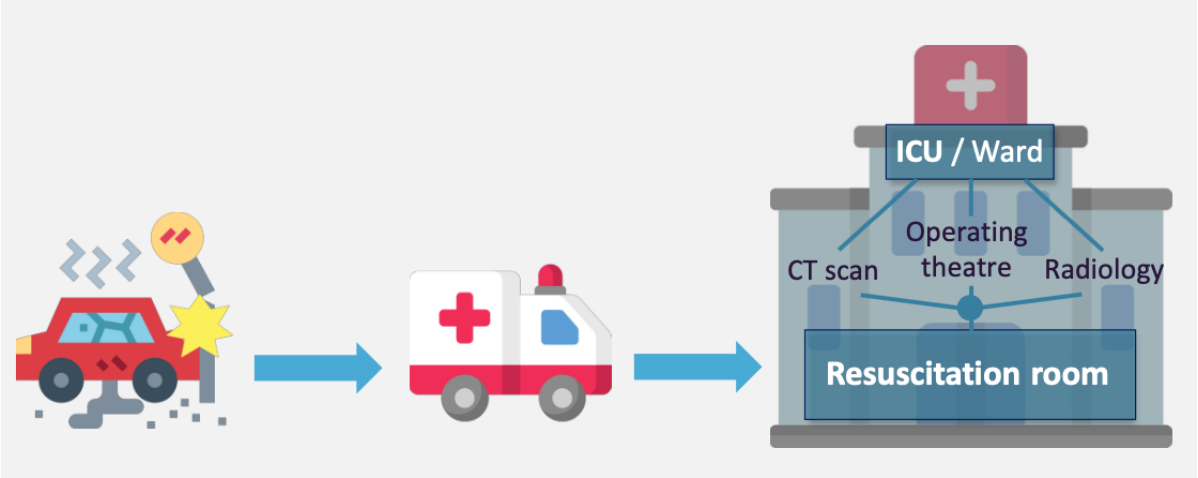


Figure A.3: Schéma de prise en charge d'un patient traumatisé.

contient maintenant des données sur 20 000 cas de traumatismes.

Les données sociodémographiques, cliniques, biologiques et thérapeutiques (de la phase pré-hospitalière à la sortie, en cas d'hospitalisation) sont systématiquement enregistrées pour tous les patients traumatisés, et tous les patients transportés aux urgences des centres participants sont inclus dans le registre. La quantité de données collectées (avec plus de 250 variables) fait de cet ensemble de données un ensemble unique en Europe. Cependant, ces données, provenant de sources multiples, sont présentent une forte variabilité inter-centres, sans parler du fait qu'il manque beaucoup de données, deux problèmes qui rendent la modélisation difficile.

Un des objectifs du projet est de modéliser les décisions et les événements pris par les médecins urgentistes pour les aider à faire des choix dans un environnement très stressant et éviter les divergences entre le diagnostic posé par les médecins urgentistes et celui posé par les médecins à l'arrivée du patient au centre de traumatologie. Par exemple, nous voudrions établir des modèles prédictifs pour savoir s'il faut ou non prévoir le risque d'hémorragie grave, afin de préparer une réponse appropriée à l'arrivée dans un centre de traumatologie, par exemple un protocole de transfusion massive et/ou des procédures hémostatiques immédiates.

D'un point de vue statistique, les défis consistent à réaliser des modèles prédictifs tels que des régressions logistiques ou des régressions avec de nombreuses valeurs manquantes. D'autres tâches peuvent inclure la sélection de modèles afin de choisir les mesures les plus importantes pour expliquer la réponse, afin d'aider à proposer une réponse innovante au défi de santé publique que représente un traumatisme majeur.

La figure A.4 montre un extrait de l'ensemble de données, avec différents codages des valeurs manquantes (NA pour Non Applicable, Imp pour Impossible, NR pour Non Enregistré, NM pour Non Fabriqué), et la figure A.5 résume le pourcentage de valeurs manquantes dans 45 variables représentatives parmi le total des mesures. Les raisons pour lesquelles ces valeurs manquantes se sont produites peuvent être diverses. Par exemple, lorsqu'un patient se trouve dans une situation très urgente, il n'y a plus de temps pour mesurer certaines des variables (et les médecins savent, sans le mesurer, que les valeurs sont critiques) ; ce cas peut être considéré comme un MNAR. D'autres cas incluent des données qui n'ont tout

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73
11	Bicetre	Fall	61	m	84	1.7	29.07	144	105

.....

	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	.....
1	97	35.6	<NA>	12.7	12	yes	
2	100	36.5	4.8	11.1	15	no	
3	100	36	3.9	11.4	3	no	
4	100	36.7	1.66	13	15	yes	
6	100	36	NM	14.4	15	no	
7	100	36.6	NM	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NM	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	

Figure A.4: Un extrait de l'ensemble des données de TraumaBase avec diverses données manquantes.

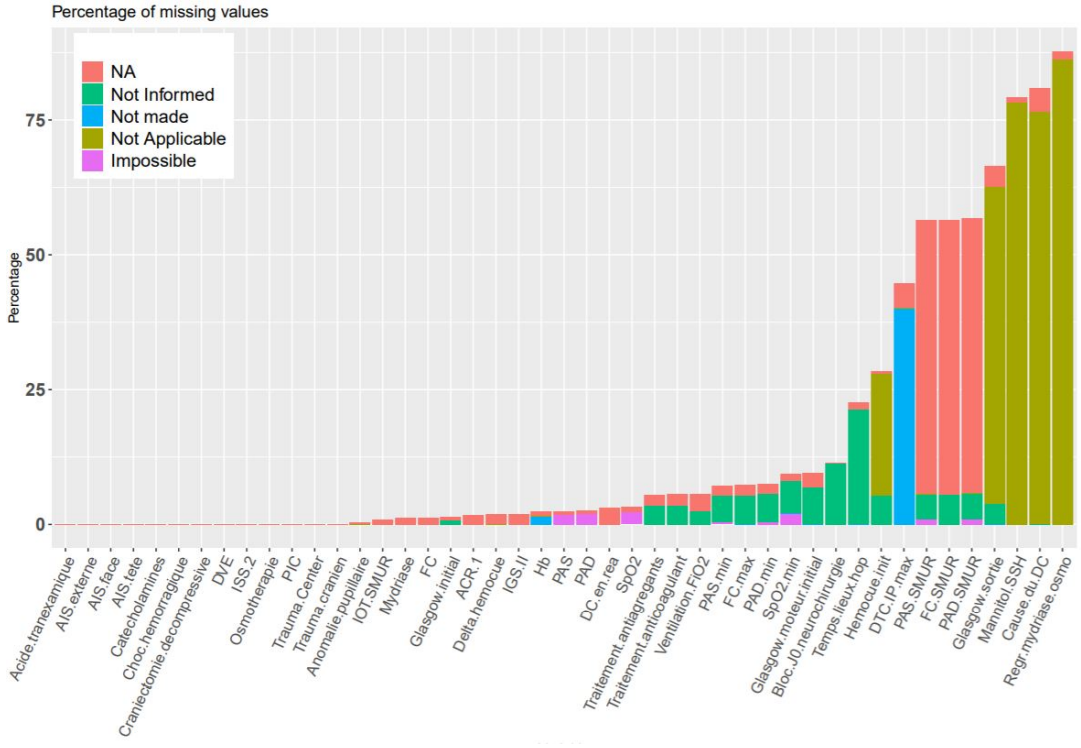


Figure A.5: Pourcentage de valeurs manquantes dans chaque variable de l'ensemble de données TraumaBase.

simplement pas été enregistrées dans la base de données (les données ont été mesurées mais non rapportées dans la TraumaBase simplement parce qu'elles sont oubliées ou lorsqu'elles sont fusionnées à partir de différentes sources par exemple). En outre, comme mentionné, certains centres de traumatologie ont progressivement rejoint la TraumaBase, et ils n'ont pas nécessairement le même dispositif dans chaque hôpital, ce qui se traduit notamment par des structures de données manquantes avec des colonnes manquantes (correspondant à des caractéristiques manquantes) pour certains des groupes. Ces codes—NR, NM, Imp–peuvent donc aider à comprendre la nature des données manquantes et les raisons de leur apparition. En effet, même si nous ne détaillerons pas ces aspects dans ce document, la première chose à faire lorsque nous avons des données manquantes est d'explorer, de visualiser, de faire des statistiques descriptives pour comprendre les données manquantes. Dans ce travail, nous avons toujours échangé avec les médecins pour voir si les hypothèses formulées nous semblaient plausibles. Selon les chiffres, nous observons comment les données manquantes affectent de manière significative les données de TraumaBase, et combien il est essentiel de concevoir une méthodologie spécifique liée aux valeurs manquantes.

Dans cette thèse, nous avons étudié un sous-ensemble de l'ensemble de la TraumaBase, qui contient 7495 individus enregistrés dans les données de traumatisme, inclus de janvier 2011 à mars 2016, avec des âges allant de 12 à 96 ans.

Résumé des contributions

On peut se rendre compte que la littérature sur l'inférence statistique avec des valeurs manquantes n'est pas assez abondante. Bien que l'algorithme EM soit étudié de manière approfondie au cours de ces décennies, les applications et les mises en œuvre sont limitées aux modèles simples, ou avec un schéma fixe de valeurs manquantes. À notre connaissance, aucune des méthodes disponibles n'aborde le problème du choix du modèle pour traiter les valeurs manquantes et contrôler simultanément le taux de fausses découvertes. L'objectif de cette thèse est de fournir une méthodologie statistique efficace et complète pour traiter le problème d'inférence avec l'existence de valeurs manquantes, et en particulier pour l'application médicale. En outre, des mises en œuvre conviviales sous forme de paquets R sont développées. Les contributions détaillées sont énumérées ci-dessous.

Régression logistique avec covariables manquantes

Dans le chapitre 2, nous abordons le problème de l'inférence statistique pour le modèle de régression logistique avec covariables manquantes. Il est surprenant de constater qu'il existe très peu de solutions pour effectuer une régression logistique avec des valeurs manquantes dans les covariables, même s'il s'agit d'un modèle commun. Une approche complète basée sur une version d'approximation stochastique de l'algorithme EM (SAEM) (Lavielle, 2014; Delyon et al., 1999) est proposée afin de réaliser une inférence statistique avec des valeurs manquantes, y compris l'estimation des paramètres et de leur variance, la dérivation des intervalles de confiance, et également une procédure de sélection du modèle. Le problème de la prédiction de nouvelles observations sur un ensemble de tests avec des données de covariables manquantes est également abordé. Soutenue par une étude de simulation dans laquelle la méthode est comparée aux précédentes, elle s'est avérée efficace sur le plan du

calcul, et présente de bonnes propriétés de couverture et de sélection des variables. L'approche est ensuite illustrée sur TraumaBase en prédisant l'apparition d'un choc hémorragique, une des principales causes de décès précoce évitable dans les cas de traumatismes graves. L'objectif est d'améliorer la procédure actuelle d'alerte rouge (Hamada et al., 2018), une alerte binaire identifiant les patients présentant un risque élevé d'hémorragie grave.

Sélection de modèles en haute dimension pour contrôler le taux de fausses découvertes (FDR)

Le chapitre 3 propose une nouvelle méthodologie pour sélectionner les variables importantes comportant des valeurs manquantes, en se concentrant plus particulièrement sur les données à haute dimension où p est comparable à n ou même supérieur à n. Nous proposons une nouvelle procédure synergique – la méthode bayésienne adaptative SLOPE (ABSLOPE) - qui combine efficacement la méthode SLOPE (régularisation triée l_1) (Bogdan et al., 2015) avec la méthode LASSO (Spike-and-Slab LASSO) (Ročková and George, 2018). Nous positionnons notre approche dans un cadre bayésien qui permet la sélection simultanée de variables et l'estimation de paramètres, malgré les valeurs manquantes. Comme pour la méthode LASSO de Spike-and-Slab, les coefficients sont considérés comme provenant d'un modèle hiérarchique composé de deux groupes : *1)* le pic pour les inactifs et *2)* la dalle pour les actifs. Toutefois, au lieu d'assigner des valeurs préalables de pic indépendantes pour chaque covariable, nous déployons ici un pic conjoint "SLOPE" qui prend en compte l'ordre des coefficients afin de contrôler les fausses découvertes. Grâce à des simulations approfondies, nous démontrons une performance satisfaisante en termes de puissance, de FDR et de biais d'estimation dans un large éventail de scénarios. Enfin, nous démontrons une excellente performance dans la prévision des niveaux de plaquettes lors de l'analyse des données de TraumaBase.

Sélection de variables contrôlées avec valeurs manquantes dans un cadre de modèle-X

Le chapitre 4 aborde également le problème de la sélection des modèles avec des valeurs manquantes lors du contrôle des FDR. Cependant, à la différence du cadre du chapitre 3, nous supposons un cadre de modèle-X où la distribution conditionnelle de la réponse donnée aux covariables n'est pas spécifiée, mais où la distribution conjointe des covariables est connue. Un tel cadre présente des avantages lorsque la distribution de y given X est compliquée, comme c'est le cas avec un modèle de régression non linéaire. La nouvelle méthode proposée - missKnockoff est basée sur la méthode de simulation modèle-X (Candes et al., 2018). Notre méthode utilise deux fois les knockoffs : elle substitue d'abord les valeurs manquantes par des knockoffs, puis procède à une application standard des knockoffs du modèle-X sur l'ensemble de données imputées. Afin de tenir compte de l'incertitude, l'imputation multiple est facilement incorporée en générant plusieurs copies imitées au premier stade, et nous discutons de différentes façons d'agréger le support. Nous étudions les performances en termes de puissance et de FDR grâce à des simulations approfondies.

Implémentation et packages

Enfin, le chapitre 5 fournit les instructions relatives à la mise en œuvre des méthodologies mentionnées ci-dessus. Deux progiciels sont développés pour traiter l'inférence statistique avec les valeurs manquantes :

- `misaem` est un progiciel R (R Core Team, 2017) permettant d'appliquer l'inférence statistique pour la régression linéaire et le modèle de régression logistique avec données manquantes. Cette méthodologie est basée sur la vraisemblance, notamment :

  1. EM-type algorithmes pour estimer les paramètres ;
  2. Obtention de la variance des paramètres estimés ;
  3. Procédure de sélection du modèle basée sur le BIC ;
  4. Prédiction sur l'ensemble de test avec des valeurs manquantes.

- `ABSLOPE` est un paquet R qui vise à la sélection de modèles à haute dimension avec des valeurs manquantes via la SLOPE bayésienne adaptative. En outre, un algorithme simplifié pour accélérer le temps de calcul est également mis en œuvre avec des fonctions C++.

Contribution à la TraumaBase

Nous collaborons avec des partenaires médicaux (le groupe TraumaBase des APHP) pour améliorer la prise en charge et les soins des patients gravement traumatisés. Nous avons construit des modèles avec des valeurs manquantes pour prédire le risque de choc hémorragique et le niveau de plaquettes à partir de données pré-hospitalières. Nos collaborateurs, les médecins, sont extrêmement satisfaits des résultats. En effet, le modèle de régression logistique proposé avec les valeurs manquantes améliore la prédiction du risque hémorragique par rapport à la prédiction faite par les médecins. L'objectif est donc maintenant d'implémenter le modèle en temps réel, car au-delà de la qualité prédictive, il faut voir comment les médecins vont réagir à un tel outil d'aide à la décision, comment présenter des recommandations avec une interface ergonomique et comment les médecins réagissent à cet outil d'aide à la décision. Les résultats ont été communiqués lors de la réunion de la Société Française d'Anesthésie et de Réanimation (SFAR) et nous avons reçu des commentaires constructifs et un fort intérêt pour l'application en temps réel.

Bibliography

Abramovich, F. and Grinshtein, V. (2018). High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory*, 65(5):3068–3079.

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

Barber, R. F., Candès, E. J., et al. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537.

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.

Bernaards, C. A. and Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, 34(3):277–313.

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.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

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.

Candes, 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.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159.

Chow, W. K. (1979). A look at various estimators in logistic models in the presence of missing values. Technical report, RAND CORP SANTA MONICA CA.

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

Consentino, F. and Claeskens, G. (2011). Missing covariates in logistic regression, estimation and distribution selection. *Statistical Modelling*, 11(2):159–183.

D'Agostino Jr, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451):749–759.

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.

Descloux, P. and Sardy, S. (2018). Model selection with lasso-zero: adding straw to the haystack to better find needles. *arXiv preprint arXiv:1805.05133*.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091.

Dutton, R. P., Stansbury, L. G., Leone, S., Kramer, E., Hess, J. R., and Scalea, T. M. (2010). Trauma mortality in mature trauma systems: are we doing better? An analysis of trauma mortality patterns, 1997–2008. *Journal of Trauma and Acute Care Surgery*, 69(3):620–626.

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

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426):463–475.

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. (2008). Sparse inverse covariance estimation with the graphical $lasso$. *Biostatistics*, 9(3):432–441.

Fung, K. Y. and Wrobel, B. A. (1989). The treatment of missing values in logistic regression. *Biometrical Journal*, 31(1):35–47.

Gilks, W. R. and Wild, P. P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist*, 41(2):337–348.

Gimenez, J. R. and Zou, J. (2018). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. *arXiv preprint arXiv:1810.11378*.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114.

Guo, Y., Hastie, T., and Tibshirani, R. (2006). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100.

Hamada, S. R., Gauss, T., Duchateau, F.-X., Truchot, J., Harrois, A., Raux, M., Duranteau, J., Mantz, J., and Paugam-Burtz, C. (2014). Evaluation of the performance of french physician-staffed emergency medical service in the triage of major trauma patients. *Journal of Trauma and Acute Care Surgery*, 76(6):1476–1483.

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

Hamada, S. R., Rosa, A., Gauss, T., Desclefs, J.-P., Raux, M., Harrois, A., Follin, A., Cook, F., Boutonnet, M., Attias, A., Ausset, S., Dhonneur, G., Langeron, O., Paugam-Burtz, C., Pirracchio, R., Riou, B., de St Maurice, G., Vigué, B., Rouquette, A., and Duranteau, J. (2018). Development and validation of a pre-hospital "Red Flag" alert for activation of intra-hospital haemorrhage control response in blunt trauma. *Critical Care*, 22(1):113.

Hay, S. I., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulkader, R. S., Abdulle, A. M., Abebo, T. A., Abera, S. F., 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.

Hentges, A. L. and Dunsmore, I. R. (1998). Predictive distributions in binary models with missing data. *Communications in Statistics-Simulation and Computation*, 27(3):735–759.

Holden, L. and Hellton, K. (2018). Multiple model-free knockoffs.

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.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *BIOMETRICS*, 55:591–596.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346.

Ilin, A. and Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, 11(Jul):1957–2000.

Jiang, J., Nguyen, T., and Rao, J. S. (2015). The E-MS algorithm: Model selection with incomplete data. *Journal of the American Statistical Association*, 110(511):1136–1147.

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

Jiang, W. (2019b). Codes and implementations for "Logistic regression with missing covariates – parameter estimation, model selection and prediction within a joint-modeling framework". https://github.com/wjiang94/miSAEM_logReg.

Jiang, W., Miasojedow, B., and Majewski, S. (2019). ABSLOPE: a package for high-dimensional model selection with missing values. https://github.com/wjiang94/ABSLOPE.

Jiang, W. and Mozharovskyi, P. (2020). *misaem: Linear Regression and Logistic Regression with Missing Covariates*. R package version 1.0.0.

Josse, J. and Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31.

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

Josse, J. and Reiter, J. P. (2018). Introduction to the special section on missing data. *Statist. Sci.*, 33(2):139–141.

Kapelner, A. and Bleich, J. (2015). Prediction with missing data via bayesian additive regression trees. *Canadian Journal of Statistics*, 43(2):224–239.

Kauvar, D. S. and Wade, C. E. (2005). The epidemiology and modern management of traumatic hemorrhage: Us and international perspectives. *Critical Care*, 9(5):S1.

Larsson, J., Bogdan, M., and Wallin, J. (2020). The strong screening rule for slope.

Laska, J. N., Davenport, M. A., and Baraniuk, R. G. (2009). Exact signal recovery from sparsely corrupted measurements through the pursuit of justice. In *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pages 1556–1560. IEEE.

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. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

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.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

Lounici, K. et al. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Ma, C., Tschiatschek, S., Palla, K., Hernández-Lobato, J. M., Nowozin, S., and Zhang, C. (2018). Eddi: Efficient dynamic discovery of high-value information with partial vae.

Mattei, P.-A. and Frellsen, J. (2019). MIWAE: Deep generative modelling and imputation of incomplete data sets. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4413–4423, Long Beach, California, USA. PMLR.

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.

McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. ed edition.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

Mohan, K. and Pearl, J. (2014). Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems*, pages 1520–1528.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26*, pages 1277–1285. Curran Associates, Inc.

Mohan, K., Thoemmes, F., and Pearl, J. (2018). Estimation with incomplete data: The linear case. In *IJCAI*, pages 5082–5088.

Morvan, M. L., Prost, N., Josse, J., Scornet, E., and Varoquaux, G. (2020). Linear predictor on linearly-generated data with missing values: non consistency and solutions.

Murray, J. S. et al. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, 33(2):142–159.

Murray, J. S. and Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479.

Muzellec, B., Josse, J., Boyer, C., and Cuturi, M. (2020). Missing data imputation using optimal transport.

Nabi, R., Bhattacharya, R., and Shpitser, I. (2020). Full law identification in graphical models of missing data: Completeness results.

Neal, R. M. and Hinton, G. E. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.

Nguyen, B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2019). Aggregation of multiple knockoffs. *preprint*.

Novo, A. A. and Schafer, J. L. (2013). *norm: Analysis of multivariate normal datasets with missing values*. R package version 1.0-9.5.

Olkin, I., Tate, R. F., et al. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 32(2):448–465.

Pan, W. and Bai, H. (2015). *Propensity score analysis: Fundamentals and developments*. Guilford Publications.

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

Rejchel, W. and Bogdan, M. (2019). Rank-based Lasso—efficient methods for high-dimensional robust model selection. *arXiv preprint 1905.05876*.

Robin, G. (2019). *Low-rank methods for heterogeneous and multi-source data*. PhD thesis.

Ročková, V. et al. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):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.

Romano, Y., Sesia, M., and Candès, E. J. (2018). Deep knockoffs. *arXiv preprint arXiv:1811.06687*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

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

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L. and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, 95(449):144–154.

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of climate*, 14(5):853–871.

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.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by "missing at random"? *Statist. Sci.*, 28(2):257–268.

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

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

Sportisse, A., Boyer, C., and Josse, J. (2018). Imputation and low-rank estimation with missing non at random data. *arXiv preprint arXiv:1812.11409*.

Sportisse, A., Boyer, C., and Josse, J. (2019). Estimation and imputation in probabilistic principal component analysis with missing not at random data.

Stekhoven, D. J. and Buehlmann, P. (2012). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, 59(4):1140–1150.

Stubbendick, A. L. and Ibrahim, J. G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, pages 1143–1167.

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.

Tang, G., Little, R. J., and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90(4):747–764.

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*.

Tchetgen, E. J. T., Wang, L., and Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069–2088.

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

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Udell, M. and Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press LLC.

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

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

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27(17):3227–3246.

Yekutieli, D. and Weinstein, A. (2019). Hierarchical bayes modeling for large-scale inference. *arXiv preprint arXiv:1908.08444*.

Yoon, J., Jordon, J., and Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698.

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

Zhao, Y. and Udell, M. (2019). Missing value imputation for mixed data via gaussian copula.

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.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre: Inférence statistique avec des données incomplètes et de grandes dimensions—modélisation des polytraumatisés graves

Mots clés: Observations manquantes (statistique), Modèles linéaires généralisés, Algorithmes EM, Analyse de régression, Dépendance (statistique), Statistique médicale

Résumé: Le problème des données manquantes existe depuis les débuts de l'analyse des données, car les valeurs manquantes sont liées au processus d'obtention et de préparation des données. Dans les applications des statistiques modernes et de l'apprentissage machine, où la collecte de données devient de plus en plus complexe et où de multiples sources d'information sont combinées, les grandes bases de données présentent souvent un nombre extraordinairement élevé de valeurs manquantes. Ces données présentent donc d'importants défis méthodologiques et techniques pour l'analyse : de la visualisation à la modélisation, en passant par l'estimation, la sélection des variables, les capacités de prédiction et la mise en oeuvre par des implémentations. De plus, bien que les données en grande dimension avec des valeurs manquantes soient considérées comme des difficultés courantes dans l'analyse statistique aujourd'hui, seules quelques solutions sont disponibles. L'objectif de cette thèse est de développer de nouvelles méthodologies pour effectuer des inférences statistiques avec des données manquantes et en particulier pour des données en grande dimension. La contribution la plus importante est de proposer un cadre complet pour traiter les valeurs manquantes, de l'estimation à la sélection d'un modèle, en se basant sur des approches de vraisemblance. La méthode proposée ne repose pas sur un dispositif spécifique du manque, et permet un bon équilibre entre qualité de l'inférence et implémentations efficaces. La méthode est ensuite appliquée aux données pré-hospitalières, en collaboration avec des partenaires médicaux - le groupe Traumabase des hôpitaux de Paris. Enfin, nous fournissons deux logiciels open-source avec des tutoriels, afin d'aider la prise de décision dans le domaine médical et les utilisateurs confrontés à des valeurs manquantes.

Title: Statistical inference with incomplete and high-dimensional data—modeling polytraumatized patients

Keywords: Missing Observations, Generalized Linear Models, EM Algorithms, Regression Analysis, Dependence, Medical Statistics

Abstract: The problem of missing data has existed since the beginning of data analysis, as missing values are related to the process of obtaining and preparing data. In applications of modern statistics and machine learning, where the collection of data is becoming increasingly complex and where multiple sources of information are combined, large databases often have an extraordinarily high number of missing values. These data therefore present important methodological and technical challenges for analysis: from visualization to modeling including estimation, variable selection, predictive capabilities, and implementation through implementations. Moreover, although high-dimensional data with missing values are considered common difficulties in statistical analysis today, only a few solutions are available. The objective of this thesis is to provide new methodologies for performing statistical inferences with missing data and in particular for high-dimensional data. The most important contribution is to provide a comprehensive framework for dealing with missing values from estimation to model selection based on likelihood approaches. The proposed method doesn't rely on a specific pattern of missingness, and allows a good balance between quality of inference and computational efficiency. The method is then applied to pre-hospital data, in collaboration with medical partners - the Traumabase group of Paris hospitals. Finally, we provide two open-source software packages with tutorials, in order to help decision making in medical field and users facing missing values.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France