Robust Lasso-Zero for sparse corruption and model selection with missing covariates

Pascaline Descloux, Claire Boyer, Julie Josse, Aude Sportisse, Sylvain Sardy

Robust Lasso-Zero for sparse corruption and model selection with missing covariates

Pascaline Descloux[(1)], Claire Boyer[(2,3)], Julie Josse[(4,5)], Aude Sportisse[(2)], and Sylvain Sardy[(1)]

[(1)] *Department of Mathematics, University of Geneva, Switzerland*

[(2)] *LPSM, Sorbonne Université, France,*

[(3)] *DMA, Ecole Normale Supérieure, France*

[(4)] *CMAP, Ecole polytechnique, France,*

[(5)] *Google, France*

May 11, 2020

Abstract

We propose Robust Lasso-Zero, an extension of the Lasso-Zero methodology [Descloux and Sardy, 2018], initially introduced for sparse linear models, to the sparse corruptions problem. We give theoretical guarantees on the sign recovery of the parameters for a slightly simplified version of the estimator, called Thresholded Justice Pursuit. The use of Robust Lasso-Zero is showcased for variable selection

with missing values in the covariates. In addition to not requiring the specification of a model for the covariates, nor estimating their covariance matrix or the noise variance, the method has the great advantage of handling missing not-at random values without specifying a parametric model. Numerical experiments and a medical application underline the relevance of Robust Lasso-Zero in such a context with few available competitors. The method is easy to use and implemented in the R library `lass0`.

Keywords: Lasso-Zero, support recovery, sparse corruptions, incomplete data, informative missing values

1 Introduction

Let us consider the widely used framework of sparse linear models for high dimension,

$$y = X\beta^0 + \epsilon, \tag{1}$$

where $\epsilon \in \mathbb{R}^n$ is a (dense) Gaussian noise vector with variance σ^2, X has a number of columns p larger than the number of rows n, and the parameters of interest $\beta^0 \in \mathbb{R}^p$ is s-sparse (only s out of its p entries are different from zero). To take into account additional occasional corruptions, the sparse corruption problem is

$$y = X\beta^0 + \sqrt{n}\omega^0 + \epsilon, \tag{2}$$

where $\omega^0 \in \mathbb{R}^n$ is a k-sparse corruption vector; see for instance Chen et al. [2013a]. Noting that (2) can be rewritten as

$$y = \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \beta^0 \\ \omega^0 \end{bmatrix} + \epsilon,$$

the sparse corruption model can be seen as a sparse linear model with an augmented design matrix and an augmented sparse vector. We are interested in theoretical guarantees of support recovery for β^0 in (2), with interesting consequences for variable selection with missing covariates.

Related litterature. To recover β^0 when $\epsilon = 0$, several authors proposed *Justice Pursuit* (JP), name coined by Laska et al. [2009], by solving

$$\min_{\beta \in \mathbb{R}^p, \omega \in \mathbb{R}^n} \|\beta\|_1 + \|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}\omega, \tag{3}$$

which is nothing else than the *Basis Pursuit* (BP) problem, with the augmented matrix $\begin{bmatrix} X & I_n \end{bmatrix}$ (modulo the renormalization by $\sqrt{n}$ in (3)) [Wright

et al., 2009]. Wright and Ma [2010] analyzed JP for Gaussian measurements, providing support recovery results when $n \simeq p$ using cross-polytope arguments. Besides, Laska et al. [2009] and Li et al. [2010] proved that if the entries of X are i.i.d. standard Gaussian as well, then the matrix $\begin{bmatrix} X & I_n \end{bmatrix}$ satisfies some restricted isometry property with high probability, implying exact recovery of both β^0 and ω^0, provided that $n \gtrsim (s+k)\log(p)$. However, in these works, the sparsity level k of ω^0 cannot be fixed to a proportion of the sample size n. Therefore, Li [2013] and Nguyen and Tran [2013b] introduced a tuning parameter $\lambda > 0$ and solve

$$\min_{\beta \in \mathbb{R}^p, \omega \in \mathbb{R}^n} \|\beta\|_1 + \lambda \|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \omega. \tag{4}$$

In a sub-orthogonal or Gaussian design, they both proved exact recovery, even for a large proportion of corruption.

In the case of sparse ($\omega^0 \neq 0$) and dense noise ($\epsilon \neq 0$), Nguyen and Tran [2013a] proposed to jointly estimate β^0 and ω^0 by solving

$$\min_{\beta \in \mathbb{R}^p, \omega \in \mathbb{R}^n} \frac{1}{2}\|y - X\beta - \omega\|_2^2 + \lambda_\beta \|\beta\|_1 + \lambda_\omega \|\omega\|_1. \tag{5}$$

In the special case where $\lambda_\beta = \lambda_\omega$, problem (5) boils down to the Lasso [Tibshirani, 1996] applied to the response y and the design matrix $\begin{bmatrix} X & I_n \end{bmatrix}$. Assuming a standard Gaussian design and the invertibility and incoherence properties for the covariance matrix, they obtained sign recovery guarantee for an arbitrarily large fraction of corruption, provided that $n \geq Ck\log(p)\log(n)$. In addition, the required number of samples is proven to be optimal. More recently in the case of a Gaussian design with an invertible covariance matrix, Dalalyan and Thompson [2019] obtained an optimal rate of estimation of β^0 when considering an ℓ_1-penalized Huber's M-estimator, which is actually equivalent to (5) [Sardy et al., 2001].

Contributions. To estimate the support of the parameter vector β^0 in the sparse corruption problem, we study an extension of the Lasso-Zero methodology [Descloux and Sardy, 2018], initially introduced for standard sparse linear models, to the sparse corruptions problem. We provide theoretical guarantees on the sign recovery of β^0 for a slightly simplified version of Robust Lasso-Zero, that we call Thresholded Justice Pursuit (TJP). These garantees are extensions of recent results on Thresholded Basis Pursuit. The first one extends a result of Tardivel and Bogdan [2019], providing a necessary and sufficient condition for consistent recovery in a setting where the design matrix is fixed but the nonzero absolute coefficients tend to infinity. The second one extends a result of Descloux and Sardy [2018], proving sign consistency for correlated Gaussian designs when p, s and k grow with n, allowing a positive fraction of corruptions.

Showing that missing values in the covariates can be reformulated into a sparse corruption problem, we recommand Robust Lasso-Zero for dealing with missing data. For support recovery, this approach requires neither to specify a model for the covariates or the missing data mechanism, nor an estimation of the covariates covariance matrix or of the noise variance, and hence provides a simple method for the user. Numerical experiments and a medical application also underline the effectiveness of Robust Lasso-Zero with respect to few competitors.

Organization. After defining Robust Lasso-Zero in Section 2, we analyse the sign recovery properties of Thresholded Justice Pursuit in Section 2.3. Section 3.1 is dedicated to variable selection with missing values and the selection of tuning parameters is discussed in Section 3.2. Numerical experiments are presented in Section 4 and an application in Section 5.

Notation. Define $[p] := \{1, \dots, p\}$, and the complement of a subset $S \subset [p]$ is denoted $\bar{S}$. For a matrix A of size $u \times v$ and a set $T \subset [v]$, we use A_T to denote the submatrix of size $n \times |T|$ with columns indexed by T. We define the missing value indicator matrix $M \in \mathbb{R}^{n \times p}$ by $M_{ij} = \mathbf{1}_{\{X_{ij}^{\mathrm{NA}} = \texttt{NA}\}}$, and the set of incomplete rows by $\mathcal{M} := \{i \in [n] \mid M_{ij} = 1 \text{ for some } j \in [p]\}$.

2 Robust Lasso-Zero

2.1 Lasso-Zero in a nutshell

Under linear model (1), Thresholded Basis Pursuit (TBP) estimates β^0 by setting the small coefficients of the BP solution to zero. Since BP fits the observations y exactly, noise is generally overfitted. Lasso-Zero [Descloux and Sardy, 2018] alleviates this issue by solving repeated BP problems, respectively fed with the augmented matrices $[X|G^{(k)}]$, where $G^{(k)} \in \mathbb{R}^{n \times n}$, $k = 1, \dots, M$, are different i.i.d. Gaussian noise dictionaries. Hence, some columns of $G^{(k)}$ can be used to fit the noise term. The obtained estimates $\hat{\beta}^{(1)}, \dots, \hat{\beta}^{(M)}$ are then aggregated by taking the component-wise medians, further thresholded at level $\tau > 0$. Descloux and Sardy [2018] show that Lasso-Zero tuned by Quantile Universal Thresholding [Giacobino et al., 2017] achieves a very good trade-off between high power and low false discovery rate compared to competitors.

2.2 Definition of Robust Lasso-Zero

Robust Lasso-Zero arises by applying Lasso-Zero to Justice Pursuit, instead of Basis Pursuit. Consider the sparse corruption model (2), for which S^0 and T^0 denote the respective supports of β^0 and ω^0, and $s := |S^0|$ and

$k := |T^0|$ denote their respective sparsity degrees. To fix notation, we then consider the following parametrization of Justice Pursuit (JP):

$$(\hat{\beta}^{\mathrm{JP}}_\lambda, \hat{\omega}^{\mathrm{JP}}_\lambda) \in \underset{\beta\in\mathbb{R}^p,\ \omega\in\mathbb{R}^n}{\arg\min} \|\beta\|_1 + \lambda\|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}\omega. \tag{6}$$

Renormalization by $\sqrt{n}$ balances the augmented design matrix $\begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix}$: in practice the columns of X are often standardized so that $\|X_j\|_2^2 = n$ for every $j \in [p]$, and this way, all columns of $\begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix}$ have same norm.

Robust Lasso-Zero applied to (6) is fully described in Algorithm 1. Attention has been drawn to the estimation of the support of β^0. However the estimation of the corruption support is also possible by computing the corresponding vectors $\hat{\omega}^{\mathrm{med}}_\lambda$ and $\hat{\omega}^{\mathrm{Rlass0}}_{(\lambda,\tau)}$, at stages 2) and 3).

Algorithm 1 Robust Lasso-Zero

Given data (y, X), for fixed hyper-parameters $\lambda > 0, \tau \geq 0$ and $M \in \mathbb{N}^*$:

1) For $k = 1, \ldots, M$:

   i) generate a matrix $G^{(k)}$ of size $n \times n$ with i.i.d. $\mathcal{N}(0, 1)$ entries

   ii) compute the solution $(\hat{\beta}^{(k)}_\lambda, \hat{\omega}^{(k)}_\lambda, \hat{\gamma}^{(k)}_\lambda)$ to the augmented JP problem

$$\begin{aligned} (\hat{\beta}^{(k)}_\lambda, \hat{\omega}^{(k)}_\lambda, \hat{\gamma}^{(k)}_\lambda) \in \quad & \underset{\beta\in\mathbb{R}^p,\ \omega\in\mathbb{R}^n,\ \gamma\in\mathbb{R}^n}{\arg\min} \quad \|\beta\|_1 + \lambda\|\omega\|_1 + \|\gamma\|_1 \\ & \text{s.t.} \qquad\qquad\quad y = X\beta + \sqrt{n}\omega + G^{(k)}\gamma. \end{aligned} \tag{7}$$

2) Define the vector $\hat{\beta}^{\mathrm{med}}_\lambda$ by

$$\hat{\beta}^{\mathrm{med}}_{\lambda,j} := \mathrm{median}\{\hat{\beta}^{(k)}_{\lambda,j}, k = 1, \ldots, M\} \quad \text{for every } j \in [p].$$

3) Calculate the estimate $\hat{\beta}^{\mathrm{Rlass0}}_{(\lambda,\tau)} := \eta_\tau(\hat{\beta}^{\mathrm{med}}_\lambda)$, where $\eta_\tau(x) = x\mathbf{1}_{(\tau,+\infty)}(|x|)$ hard-thresholds component-wise.

Since the minimization problem (7) in Algorithm 1 can be recast as a linear program, any relevant solver can be used (e.g., proximal methods). Algorithm 1 includes two hyper-parameters: the regularization parameter λ of (6), and the thresholding parameter τ of the Robust Lasso-Zero methodology. Their choice in practice is discussed in Section 3.2.

2.3 Theoretical guarantees on Thresholded Justice Pursuit

Discarding the noise dictionaries in Algorithm 1 amounts to thresholding the solution $(\hat{\beta}^{\mathrm{JP}}_{\lambda}, \hat{\omega}^{\mathrm{JP}}_{\lambda})$ to the Justice Pursuit problem (6). Robust Lasso-Zero can therefore be regarded as an extension of this simpler estimator, which we call *Thresholded Justice Pursuit* (TJP):

$$\hat{\beta}^{\mathrm{TJP}}_{(\lambda,\tau)} = \eta_{\tau}(\hat{\beta}^{\mathrm{JP}}_{\lambda}) \qquad \text{and} \qquad \hat{\omega}^{\mathrm{TJP}}_{(\lambda,\tau)} = \eta_{\tau}(\hat{\omega}^{\mathrm{JP}}_{\lambda}). \tag{8}$$

We present two results about sign consistency of TJP.

2.3.1 Identifiability as a necessary and sufficient condition for consistent sign recovery

First introduced in Tardivel and Bogdan [2019] for the TBP, we propose the following extension of the identifiability notion for the TJP.

Definition 1. The pair $(\beta^0, \omega^0) \in \mathbb{R}^p \times \mathbb{R}^n$ is said to be *identifiable* with respect to $X \in \mathbb{R}^{n\times p}$ and the parameter $\lambda > 0$ if it is the unique solution to JP (6) when $y = X\beta^0 + \sqrt{n}\omega^0$.

It is worth noting that identifiability of (β^0, ω^0) can be shown to depend only on $\mathrm{sign}(\beta^0)$ and $\mathrm{sign}(\omega^0)$, as highlighted in the following result.

Lemma 1. *The pair* $(\beta^0, \omega^0) \in \mathbb{R}^p \times \mathbb{R}^n$ *is identifiable with respect to* $X \in \mathbb{R}^{n\times p}$ *and the parameter* $\lambda > 0$ *if and only if for every pair* $(\beta, \omega) \neq (0, 0)$ *such that* $X\beta + \sqrt{n}\lambda^{-1}\omega = 0$,

$$|\operatorname{sign}(\beta^0)^T\beta + \operatorname{sign}(\omega^0)^T\omega| < \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1.$$

Proof. See Appendix A. □

In order to show that identifiability is necessary and sufficient for TJP to consistently recover $\operatorname{sign}(\beta^0)$ and $\operatorname{sign}(\omega^0)$, assume that for a fixed matrix $X \in \mathbb{R}^{n\times p}$ and a sequence $\{(\beta^{(r)}, \omega^{(r)})\}_{r\in\mathbb{N}^*}$, the following holds:

(i) there exist sign vectors $\theta \in \{1, -1, 0\}^p$ and $\tilde{\theta} \in \{1, -1, 0\}^n$ such that $\operatorname{sign}(\beta^{(r)}) = \theta$ and $\operatorname{sign}(\omega^{(r)}) = \tilde{\theta}$ for every $r \in \mathbb{N}^*$,

(ii) $\lim_{r\to+\infty} \min\{\beta^{(r)}_{\min}, \omega^{(r)}_{\min}\} = +\infty$, where $\beta_{\min} := \min_{j\in\operatorname{supp}(\beta)} |\beta_j|$,

(iii) there exists $q > 0$ such that $\frac{\min\{\beta^{(r)}_{\min}, \omega^{(r)}_{\min}\}}{\max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}} \geq q$.

These assumptions are similar to the ones of Tardivel and Bogdan [2019]. We use the notation $S^0 := \operatorname{supp}(\theta) = \operatorname{supp}(\beta^{(r)})$ and $T^0 := \operatorname{supp}(\tilde{\theta}) = \operatorname{supp}(\omega^{(r)})$. We denote by $(\hat{\beta}^{\mathrm{JP}(r)}_\lambda, \hat{\omega}^{\mathrm{JP}(r)}_\lambda)$ the JP solution when $y = y^{(r)} := X\beta^{(r)} + \sqrt{n}\omega^{(r)} + \epsilon$, and $(\hat{\beta}^{\mathrm{TJP}(r)}_{(\lambda,\tau)}, \hat{\omega}^{\mathrm{TJP}(r)}_{(\lambda,\tau)})$ the corresponding TJP estimates.

Theorem 1. *Let* $\lambda > 0$ *and let* X *be a matrix of size* $n \times p$ *such that for any* $y \in \mathbb{R}^n$, *the solution to JP (6) is unique. Let* $\{(\beta^{(r)}, \omega^{(r)})\}_{r\in\mathbb{N}^*}$ *be a sequence satisfying assumptions* (i)-(iii) *above. If the pair of sign vectors* $(\theta, \tilde{\theta})$ *is identifiable with respect to* X *and* λ, *then for every* $\epsilon \in \mathbb{R}^n$, *there exists* $R = R(\epsilon) > 0$ *such that for every* $r \geq R$ *there is a threshold* $\tau = \tau(r) > 0$ *for which*

$$\operatorname{sign}(\hat{\beta}^{\mathrm{TJP}(r)}_{(\lambda,\tau)}) = \theta \quad \textit{and} \quad \operatorname{sign}(\hat{\omega}^{\mathrm{TJP}(r)}_{(\lambda,\tau)}) = \tilde{\theta}. \tag{9}$$

Conversely, if for some $\epsilon \in \mathbb{R}^n$ *and* $r \in \mathbb{N}^*$ *there is a threshold* $\tau > 0$ *such that (9) holds, then* $(\theta, \tilde{\theta})$ *is identifiable with respect to* X *and* λ.

Proof. See Appendix A. □

Remark 1. One might be interested in recovering the signs of the sparse corruption. If $\omega^{(r)}$ is considered as noise, then only the recovery of $\mathrm{sign}(\beta^{(r)})$ matters. In this case one could weaken assumptions (ii) and (iii) above by replacing $\min\{\beta^{(r)}_{\min}, \omega^{(r)}_{\min}\}$ by $\beta^{(r)}_{\min}$, and identifiability of $(\theta, \tilde{\theta})$ would be sufficient for recovering $\mathrm{sign}(\beta^0)$. However, recovery of both $\mathrm{sign}(\beta^{(r)})$ and $\mathrm{sign}(\omega^{(r)})$ is needed for proving necessity of identifiability.

Identifiability of sign vectors is necessary and sufficient for sign recovery when the nonzero coefficients are large. However, Theorem 1 does not provide a lower bound indicating how large these coefficients should scale to be correctly detected. In the next section, we make this explicit in particular for (correlated) Gaussian designs and prove that sign consistency holds, allowing p, s and k to grow with the sample size n.

2.3.2 Sign consistency of TJP for correlated Gaussian designs

We make the following assumptions:

(iv) the rows of $X \in \mathbb{R}^{n\times p}$ (with $n < p$) are random and i.i.d. $\mathcal{N}(0, \Sigma)$;

(v) The smallest eigenvalue of the covariance matrix Σ is assumed to be positive: $\lambda_{\min}(\Sigma) > 0$,

(vi) the variance of the covariates is equal to one: $\Sigma_{ii} = 1$ for every $i \in [p]$;

(vii) the noise is assumed to be Gaussian: $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Assumptions (iv) and (v) imply that almost surely $\mathrm{rank}\, X = n$.

Theorem 2. *Under Assumptions* (iv)-(vii), *choosing* $\lambda = \frac{1}{\sqrt{\log p}}$ *ensures with probability greater than* $1 - ce^{-c'n} - 1.14^{-n} - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}$, *that there exists a value of* $\tau > 0$ *such that*

$$\operatorname{sign}(\hat{\beta}^{\mathrm{TJP}}_{(\lambda,\tau)}) = \operatorname{sign}(\beta^0),$$

provided that

$$n \geq C \frac{\kappa(\Sigma)}{\lambda_{\min}(\Sigma)} s \log p, \tag{10}$$

$$\frac{n}{k} \geq \max\left\{\frac{1}{C'}, \frac{\kappa(\Sigma)}{C''}\right\}, \tag{11}$$

$$\beta^0_{\min} > \frac{10\sqrt{2}\max\{1,\lambda\}\sigma\sqrt{p+n}}{\left(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1\right)^{1/2}}, \tag{12}$$

where $\kappa(\Sigma) := \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ *is the conditioning number of* Σ, *and* C, C', C'' *are some numerical constants with* $C \geq 144^2$.

Proof. See Appendix B. □

Theorem 2 ensures that, for correlated Gaussian designs and signal-to-noise ratios high enough, TJP successfully recovers $\operatorname{sign}(\beta^0)$ with high probability, even with a positive fraction of corruptions. As a consequence, if Σ is well-conditioned, (i.e. the eigenvalues of Σ are bounded: $0 < \gamma_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \gamma_2$) and $p/n \to \delta > 1$, TJP achieves sign consistency provided that $n = \Omega(s \log p)$, $k = \mathcal{O}(n)$ and $\beta^0_{\min} = \Omega(\sqrt{n})$. The lower-bound required on $\beta^0_{\min}$ in Theorem 2 is of the same order as the one required for TBP in Descloux and Sardy [2018]. One can remark that the analysis of TJP in the sparse corruption setting makes the condition number of Σ come into play in the lower-bounds required on n and k. This quantity seems natural to arise in the sparse corruption problem helping discriminating design instability from corruptions.

3 Model selection with missing covariates

In practice the matrix of covariates X is often partially known due to manual errors, poor calibration, insufficient resolution, etc., and one only observes an incomplete matrix, denoted X^{NA}.

Theoretical guarantees of estimation strategies or imputation methods rely on assumptions regarding the missing-data mechanism, i.e. the cause of the lack of data. Three missing-data mechanisms have been introduced by Rubin [1976]: the restrictive assumptions of data (a) missing completely at random (MCAR), and (b) missing at random (MAR), where the missing data may only depend on the observed variables, and (c) the more general assumption of data missing not at random (MNAR), when data missingness depends on the values of other variables, but also on its own value. Complete case analysis, which discards all incomplete rows, is the most common method for facing missing values in applications. Additionally to the induced estimation bias (especially under the MNAR missing mechanism (c)), with high-dimensional data this procedure has the big disadvantage that missingness of a single entry causes the loss of an entire row, which contains a lot of information when p is large.

High dimensional variable selection with missing values turns out to be a challenging problem and very few solutions are available, not to mention implementations. Available solutions either require strong assumptions on the missing value mechanism, a lot of parameters tuning or strong assumption on the covariates distribution which is hard in high dimensions. They include the Expectation-Maximization algorithm [Dempster et al., 1977] for sparse linear regression [Garcia et al., 2010] and regression imputation methods [Van Buuren, 2018]. A method combining penalized regression techniques with multiple imputation and stability selection has

been developed [Liu et al., 2016]. Yet, aggregating different models for the resulting multiple imputed data sets becomes increasingly complex as the number of data grows. Rosenbaum et al. [2013] modified the Dantzig selector by using a consistent estimation of the design covariance matrix. Following the same idea, Loh and Wainwright [2012] and Datta and Zou [2017] reformulated the Lasso also using an estimate of the design covariance matrix, possibly resulting in a non-convex problem. Chen and Caramanis [2013] presented a variant of orthogonal matching pursuit which recovers the support and achieves the minimax optimal rate. Jiang et al. [2019] proposed Adaptive Bayesian SLOPE, combining SLOPE and Spike-and-Slab Lasso. While some of these methods have interesting theoretical guarantees, they all require an estimation of the design covariance matrix, which is often obtained under the restrictive MCAR assumption.

3.1 Relation to the sparse corruption model

To tackle the problem of estimating the support when the design matrix is incomplete, we suggest an easy-to-implement solution for the user, which consists in imputing the missing entries in X^{NA} with the imputation of his choice to get a completed matrix $\tilde{X}$, and to take into account the impact of the possibly occasional poor imputation as follows. Given the matrix $\tilde{X}$, the linear model (1) can be rewritten in the form of the sparse corruption model (2), where $\omega^0 := \frac{1}{\sqrt{n}}(X - \tilde{X})\beta^0$ is the (unknown) corruption due to imputations. In classical (i.e. non-sparse) regression, one could not say much about ω^0 without any prior knowledge of the distribution of the covariates or the missing data mechanism. Since the key point here is that when β^0 is sparse, then so is ω^0, even if all rows of the design matrix

contain missing entries. Indeed, for every $i \in [n]$,

$$\omega_i^0 = \frac{1}{\sqrt{n}} \sum_{j=1}^{p} (X_{ij} - \tilde{X}_{ij})\beta_j^0 = \frac{1}{\sqrt{n}} \sum_{j \in S^0} (X_{ij} - \tilde{X}_{ij})\beta_j^0, \tag{13}$$

so ω_i^0 is nonzero only if the i^{th} row of X^{NA} contains missing value(s) on the support S^0, since $(X_{ij} - \tilde{X}_{ij}) = 0$ if X_{ij} is observed. So the problem of missing covariates can be rephrased as a sparse corruption problem, as already pointed out in Chen et al. [2013b]. We propose to use Robust Lasso-Zero presented in Section 2.2, which comes with strong theoretical guarantees, to tackle this sparse corruption reformulation, see Algorithm2.

Note that if the i^{th} row of X is fully observed, then $\omega_i^0 = 0$ by (13). Thus the dimension of ω^0 can be reduced by restricting it to the incomplete rows of X^{NA}. The corruption vector ω^0 is now of size $|\mathcal{M}|$ and (2) becomes

$$y = X\beta^0 + \sqrt{n} I_{\mathcal{M}} \omega^0 + \epsilon. \tag{14}$$

Algorithm 2 Robust Lasso-Zero for missing data

Given data (y, X^{NA}), for fixed hyper-parameters $\lambda > 0, \tau \geq 0$ and $M \in \mathbb{N}^*$:

1) Impute X^{NA} and rescale the imputed matrix X such that all columns have Euclidean norm equal to $\sqrt{n}$.

2) Run Algorithm 1 with the design matrix X.

3.2 Selection of tuning parameters

Algorithm 2 required selection of two hyper-parameters. Under the null model, no sparse corruption exists: indeed if $\beta^0 = 0$, so is ω^0 since $\omega^0 = \frac{1}{\sqrt{n}}(X - \tilde{X})\beta^0 = 0$. This property allows us to opt for the Quantile

Universal Threshold (QUT) methodology [Giacobino et al., 2017], generally driven by model selection rather than prediction.

QUT selects the tuning parameter so that under the null model ($\beta^0 = 0$), the null vector $\hat{\beta} = 0$ is recovered with probability $1 - \alpha$. Under the null model, $y = \epsilon$ whatever the missing data pattern is. Then given a fixed value of λ and a fixed imputed matrix $\tilde{X}$, the corresponding QUT value of τ is the upper α-quantile of $\|\hat{\beta}^{\text{med}}_{\lambda}(\epsilon)\|_{\infty}$, where $\hat{\beta}^{\text{med}}_{\lambda}(\epsilon)$ is the vectors of medians obtained at stage 2) of Algorithm 1 applied to $\tilde{X}$ and $y = \epsilon$. To free ourselves from preliminary estimation of the noise level σ, we exploit the noise coefficients $\hat{\gamma}^{(k)}$ of Robust Lasso-Zero to pivotize the statistic $\|\hat{\beta}^{\text{med}}_{\lambda}(\epsilon)\|_{\infty}$, as explained in Descloux and Sardy [2018].

For every $\lambda > 0$, there is a pair of QUT parameters $(\lambda, \tau^{\text{QUT}}_{\alpha}(y; \lambda))$ at level α. The remaining question is how to choose λ. For a fair isotropic penalty on β, ω and γ, we fix $\lambda = 1$.

4 Numerical experiments

We evaluate the performance of Robust-Lasso Zero when missing data affect the design matrix. The code reproducing these experiments is available at `https://github.com/pascalinedescloux/robust-lasso-zero-NA`.

4.1 Simulation settings

Simulation scenarios. We generate data according to model (1) with the covariates matrix obtained by drawing $n = 200$ observations from a Gaussian distribution $\mathcal{N}(0, \Sigma)$, where$\Sigma \in \mathbb{R}^{200\times 200}$ is a Toeplitz matrix, such that $\Sigma_{ij} = \rho^{|i-j|}$; the variance of the noise $\sigma = 0.5$ and the coefficient β^0 are drawn uniformly from $\{\pm 1\}$. We vary the following parameters:

- Correlation structures indexed by ρ with $\rho = 0$ (uncorrelated) and $\rho = 0.75$ (correlated);
- Sparsity degrees indexed by s with $s \in \{3, 10\}$.

Before generating the response vector y, all columns of X are mean-centered and standardized; Missing data are then introduced in X according to two different mechanisms, MCAR or MNAR, and in two different proportions. Any entry of X is missing according to the following logistic model

$$\mathbb{P}(X_{ij}^{\mathrm{NA}} = \texttt{NA} \mid X_{ij} = x) = \frac{1}{1 + e^{-a|x|-b}},$$

where $a \geq 0$ and $b \in \mathbb{R}$. Choosing $a = 0$ yields MCAR data, whereas $a = 5$ leads to MNAR setting in which high absolute entries are more likely to be missing. For a fixed a, the value of b is chosen so that the overall average proportion of missing values is π, with $\pi = 5\%$ and $\pi = 20\%$.

Two sets of simulations are run. The first one is "s-oracle", meaning that the tuning parameters of the different methods are chosen so that the estimated support has correct size s. In the second set, no knowledge of s, β^0 or σ is provided.

Estimators considered. We compare the following estimators:

- **Rlass0:** the Robust Lasso-Zero described in Algorithm 2 using M equal to 30. The tuning parameters are obtained using $\lambda = 1$ and selecting τ by quantile universal threshold (QUT) at level $\alpha = 0.05$.
- **lass0:** the Lasso-Zero proposed in Descloux and Sardy [2018]. The automatic tuning is performed by QUT, at level $\alpha = 0.05$.
- **lasso:** the Lasso [Tibshirani, 1996] performed on the mean-imputed matrix whre the regularization parameter is tuned by cross-validation.

- **NClasso:** the nonconvex ℓ_1 estimator of Loh and Wainwright [2012]. It is only included under the s-oracle setting, as selection of the tuning parameter in practice is not discussed in their work.
- **ABSLOPE:** Adaptive Bayesian SLOPE of Jiang et al. [2019].

Performance evaluation. The performance of each estimator is assessed in terms of the following criteria, averaged over 100 replications:

- the Probability of Sign Recovery (PSR), $\mathrm{PSR} = \mathbb{P}(\mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\beta^0))$,
- the signed True Positive Rate (sTPR), where $\text{s-TPR} = \mathbb{E}(\text{s-TPP})$ with

$$\text{s-TPP} := \frac{|\{j \mid \beta_j^0 > 0, \hat{\beta}_j > 0\}| + |\{j \mid \beta_j^0 < 0, \hat{\beta}_j < 0\}|}{|S^0|}, \tag{15}$$

  which is the proportion of nonzero coefficients whose sign is correctly identified;

- the signed False Discovery Rate (sFDR): $\text{s-FDR} = \mathbb{E}(\text{s-FDP})$ with

$$\text{s-FDP} := \frac{|\hat{S}| - |\{j \mid \beta_j^0 > 0, \hat{\beta}_j > 0\}| - |\{j \mid \beta_j^0 < 0, \hat{\beta}_j < 0\}|}{\max\{1, |\hat{S}|\}}, \tag{16}$$

  which is the proportion of incorrect signs among all discoveries.

4.2 Results

4.2.1 With s-oracle hyperparameter tuning

Under the s-oracle tuning, an s-TPP (15) of one means that the signs of β^0 are exactly recovered, and the s-TPP is related to the s-FDP (16) through $\text{s-FDP} = 1 - \text{s-TPP}$. That is why, in Figure 1, only the average s-TPP and the estimated probability of sign recovery are reported.

Small missingness – High sparsity (5% of NA and $s = 3$). In the non-correlated case, in Figure 1 (a) and (c), MCAR and MNAR results are similar across methods. With correlation, in Figure1 (b) and (d), Rlass0 improves PSR and sTPR, specially with MNAR data.

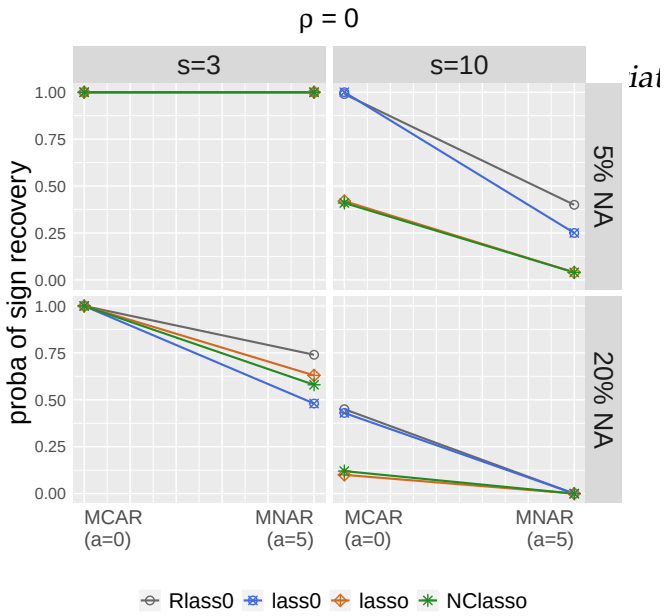
Increasing missingness – High sparsity (20% of NA and $s = 3$). The benefit of Rlass0 is noticeable when increasing the percentage of missing data to 20%, for both performance indicators. Indeed, with no correlation (Figure 1 (a)(c)(bottom left)), the improvement is clear when dealing with MNAR. With correlation (Figure 1 (b)(d)(bottom left)), Rlass0 outperforms the other methods: while the improvement can be marginal when compared to lass0 for MCAR, it becomes significant for MNAR.

Lower sparsity ($s = 10$). The performance of all estimators tends to deteriorate. One can identify two groups of estimators: Rlass0 and lass0 generally outperforms lasso and NClasso, except with a high proportion (20%) of MNAR missing data for which they all behave the same. While comparable when $s = 10$, Rlass0 proves to be better than lass0 in the case of a small proportion of MNAR missing data (5%).

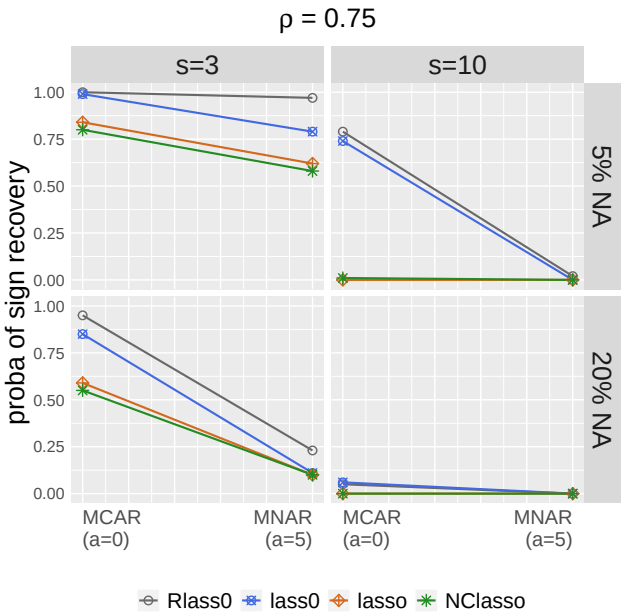
4.2.2 With automatic hyperparameter tuning

Figures 2 and 3 point to the poor performance of lasso in terms of PSR for all experimental settings. The automatic tuning, being done by cross-validation, is known to lead to support overestimation. Indeed, its very good performance in sTPR is made at the cost of a very high sFDR.

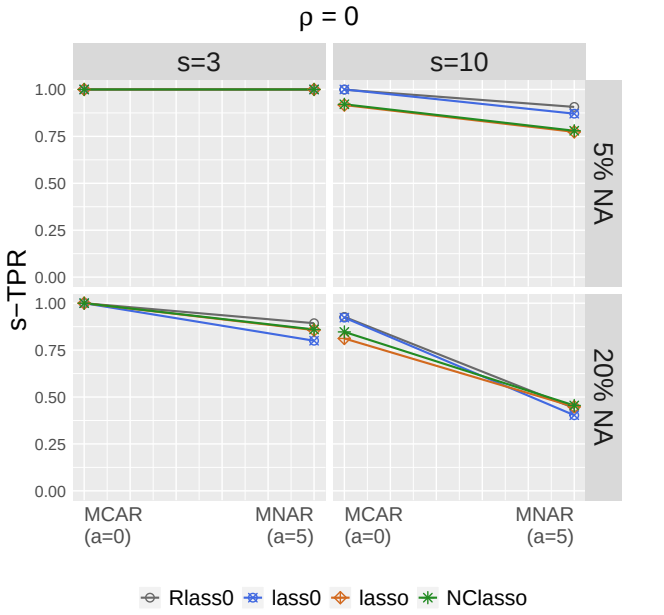
Small missingness – High sparsity (5% of NA and $s = 3$). In Figures 2(a)(top left) and 3(a)(c)(top left), for the non-correlated case, Rlass0, lass0



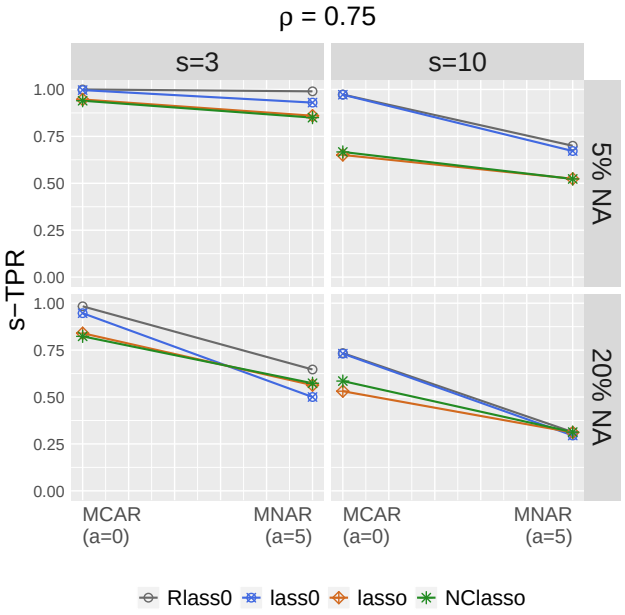
(a) PSR in the non-correlated case



(b) PSR in the correlated case



(c) s-TPR in the non-correlated case



(d) s-TPR in the correlated case

Figure 1: PSR and s-TPR with an s-oracle tuning, for sparsity levels , $s = 3$ and $s = 10$ (subplots columns), proportions of missing values 5% or 20% (subplots rows), and two missing data mechanisms (MCAR vs MNAR).

and ABSlope performs very well, providing a PSR and s-TPR of one, and a s-FDR of zero, either when dealing with MCAR or MNAR data (the lasso being already out of the game). In Figures 2(b)(top left) and 3(b)(d)(top left), adding correlation in the design matrix seems beneficial for ABSlope, at the price of high FDR, however.

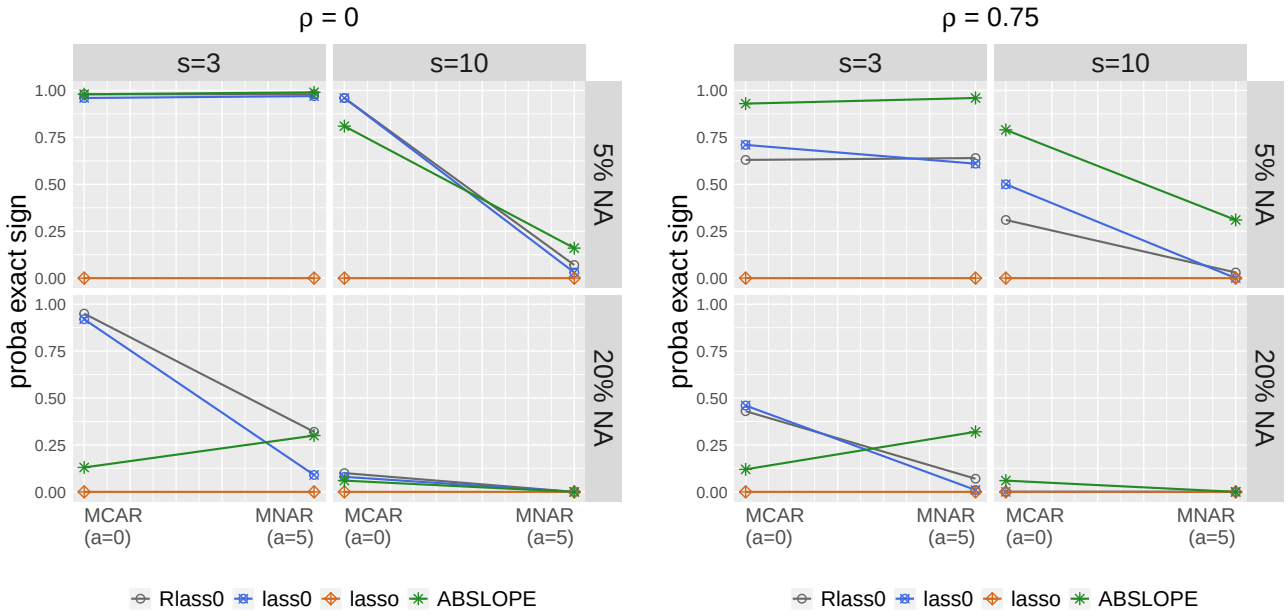
Increasing missingness – High sparsity (20% of NA and $s = 3$). With no correlation, one sees in Figure 2(a)(bottom left) that Rlass0 provides the best PSR, whatever the type of missing data is. One could also note that the performances in terms of PSR of either lass0 or ABSLOPE are extremely variable depending on the type of missing data (MCAR or MNAR) considered: the PSR of lass0 is comparable to the one of Rlass0 when facing MCAR data and is much lower than the one of Rlass0 when facing MNAR data; the converse is true for ABSLOPE.

Regarding the s-TPR and s-FDR results in Figure 3 (a-d)(bottom left), the following observations hold in both correlated or non-correlated cases:

(i) With MCAR data, all the methods behave similarly in terms of s-TPR, identifying correctly signs and coefficient locations in the support of β^0, see Figure 3(a)(b)(bottom left);

(ii) With MNAR data, lasso and ABSLOPE remain stable in terms of s-TPR, providing an s-TPR of one, whereas the s-TPR of Rlass0 deteriorates (to 0.6 and 0.5 respectively for the non-correlated and correlated cases), and even worse for lass0, see Figure 3(a)(b)(bottom left);

(iii) Lasso and ABSLOPE lead to high s-FDR, while lass0 and Rlass0 always give the best s-FDR, see Figure 3(c)(d)(bottom left).



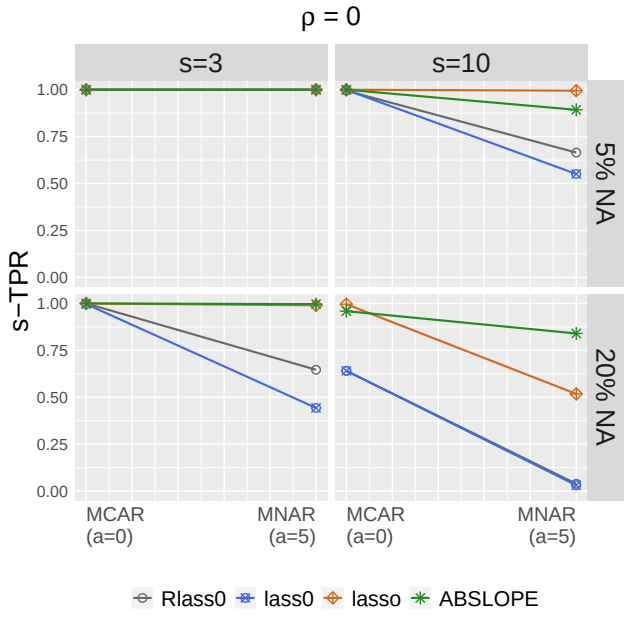
(a) PSR in the non-correlated case (b) PSR in the correlated case

Figure 2: PSR with automatic tuning, for sparsity levels $s = 3$ and $s = 10$ (subplots columns), proportions of missing values 5% or 20% (subplots rows), and two missing data mechanisms (MCAR vs MNAR).

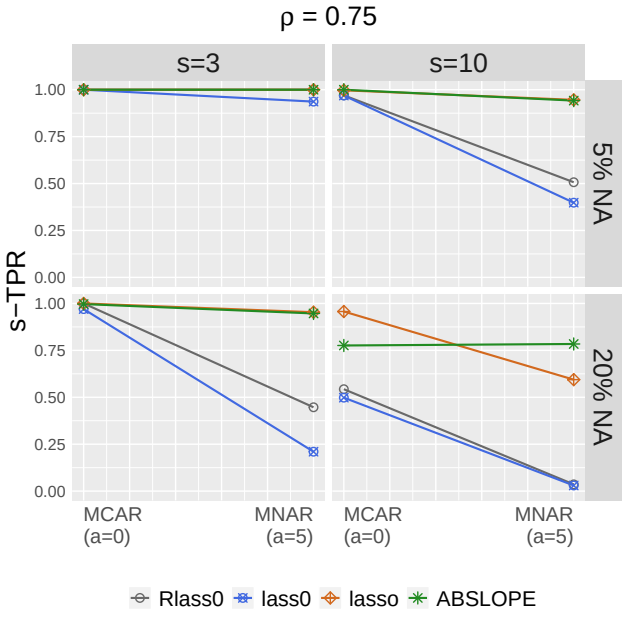
Lower sparsity ($s = 10$). For low missingness (5%), see Figure 2 (a)(b) (top right), ABSLOPE gives high PSR. In terms of s-TPR, lasso and ABSLOPE have high TPR. Moreover Rlass0 improves s-TPR compared to lass0 specially for a small proportion of MNAR missing data. In terms of s-FDR, lass0 and Rlass0 bring very low s-FDR, proving their FDR stability with respect to MCAR/MNAR data, and correlation.

4.2.3 Summary and discussion

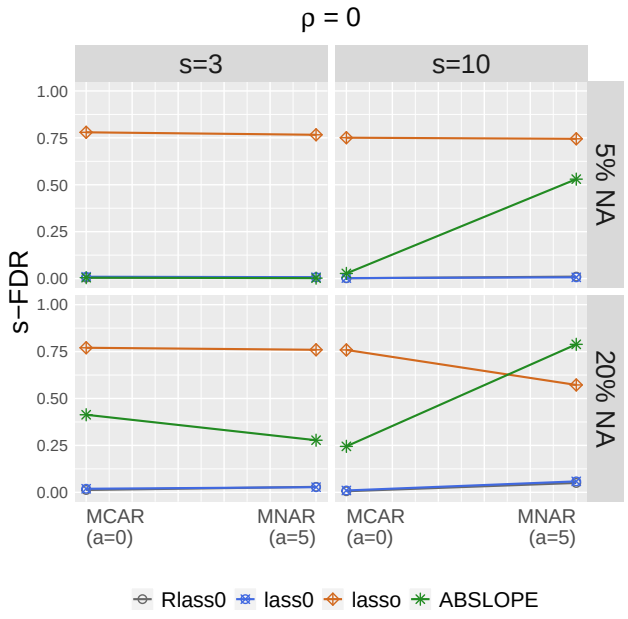
The results of experiments with s-oracle tuning (Section 4.2.1) show that Robust Lasso-Zero performs better than competitors for sign recovery, and is more robust to MNAR data compared to its nonrobust counterpart



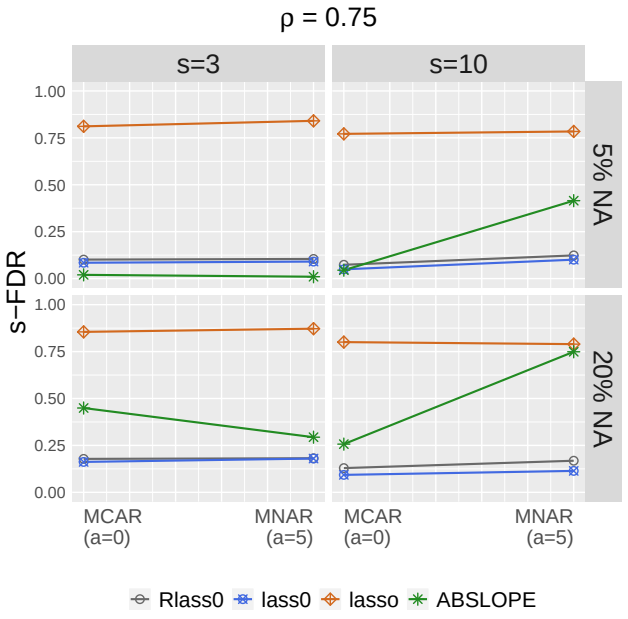
(a) s-TPR in the non-correlated case



(b) s-TPR in the correlated case



(c) s-FDR in the non-correlated case



(d) s-FDR in the correlated case

Figure 3: s-FDR and s-TPR with automatic tuning, for sparsity levels $s = 3$ and $s = 10$ (subplots columns), proportions of missing values 5% or 20% (subplots rows), and two missing data mechanisms (MCAR vs MNAR).

when the sparsity index and/or proportion of missing entries is low. In particular, Robust Lasso-Zero performs better than NClasso, one of the rare existing ℓ_1-estimator designed to handle missing values.

While not designed to handle MNAR data, ABSLOPE appears to be a valid competitor in terms of s-TPR or PSR when the model complexity increases, and when dealing with MNAR data. Its poor performance in FDR in such settings reveals its tendency to overestimate the support of β^0, under higher sparsity degrees, and with informative MNAR missing data.

With automatic tuning (Section 4.2.2), Robust Lasso-Zero is the best method overall. Moreover, our results show that the choice of Robust Lasso-Zero tuned by QUT, with its low s-FDR, is particularly appropriate in cases where one wants to maintain a low proportion of false discoveries.

5 Application to the Traumabase dataset

We illustrate our approach on the public health APHP (Assistance Publique Hopitaux de Paris) TraumaBase® Group for traumatized patients. Effective and timely management of trauma is crucial to improve outcomes, as delays or errors entail high risks for the patient.

In our analysis, we focuse on one specific challenge: selecting a sparse model from data containing missing covariates in order to explain the level of platelet. This model can aid creating an innovative response to the public health challenge of major trauma. Explanatory variables for the level of platelet consist in fifteen quantitative variables containing missing values, which have been selected by doctors. They give clinical measurements on 490 patients. In Figure 4, one sees the percentage of missing values in each variable, varying from 0 to 45% and leading to 20% is the

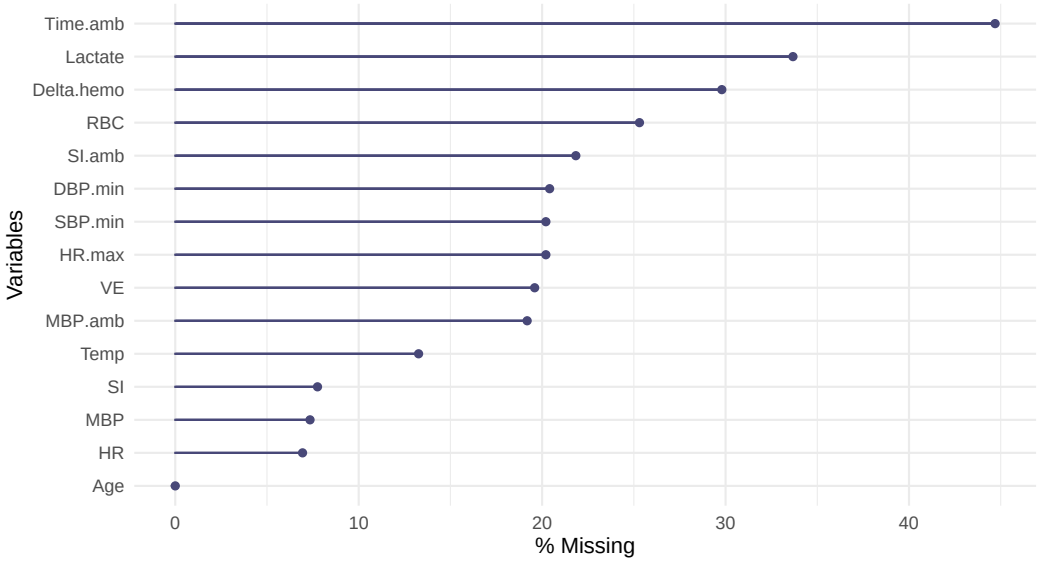


Figure 4: Percentage of missing values in the Traumbase dataset.

Variable	Rlass0	lass0	lasso	ABSLOPE
Age	−	0	−	−
SI	0	0	0	−
Delta.hemo	0	0	0	+
Lactates	0	0	0	+
Temperature	0	0	0	+
VE	−	0	−	0
RBC	−	0	0	−
DBP.min	0	0	−	+
HR.max	0	0	−	0
SI.amb	0	0	0	+

Table 1: Sign of estimated effects on the platelet for Rlass0, lass0, lasso or ABSLOPE. Variables not shown here are not selected by any method.

whole dataset. Based on discussions with doctors, some variables may have informative missingness (M(N)AR variables). Both percentage and nature of missing data demonstrate the importance of taking appropriate account of missing data. More information can be found in Appendix C.

We compare Robust Lasso-Zero to Lasso-Zero, Lasso and ABSLOPE. The signs of the coefficients are shown in Table 1. Lass0 does not select any variable, whereras its robust counterpart selects three. According to doctors, Robust Lasso-Zero is the most coherent. Indeed, a negative effect of age (*Age*), vascular filling (*VE*) and blood transfusion (*RBC*) was expected, as they all result in low platelet levels and therefore a higher risk of severe bleeding. Lasso similarly selects *Age* and *VE*, but also minimum value of diastolic blood pressure *DBP.min* and the maximum heart rate *HR.max*. The effect of *DBP.min* is not what doctors expected. For ABSLOPE, the effects on platelets of delta Hemocue (*Delta.Hemocue*), the lactates (*Lactates*), the temperature (*Temperature*) and the shock index measured on ambulance (*SI.amb*), at odds with the effect of the shock index at hospital (*SI*), are not in agreement with the doctors opinion either.

A Proof of Theorem 1

Lemma 1 implies that under the sign invariance assumption (i), identifiability of $(\beta^{(r)}, \omega^{(r)})$ is equivalent to identifiability of $(\theta, \tilde{\theta})$.

Proof of Lemma 1. Note that $(\hat{\beta}^{\mathrm{JP}}_\lambda, \hat{\omega}^{\mathrm{JP}}_\lambda)$ is a solution to JP (6) if and only if $(\hat{\beta}^{\mathrm{JP}}_\lambda, \hat{\omega}^{\mathrm{JP}}_\lambda) = (\tilde{\beta}, \lambda^{-1}\tilde{\omega})$, where $(\tilde{\beta}, \tilde{\omega})$ is a solution to

$$\min_{(\beta,\, \omega)\in\mathbb{R}^p\times\mathbb{R}^n} \|\beta\|_1 + \|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}\lambda^{-1}\omega. \tag{17}$$

So (β^0, ω^0) is identifiable with respect to X and $\lambda > 0$ if and only if the pair

$(\beta^0, \lambda\omega^0)$ is the unique solution of (17) when $y = X\beta^0 + \sqrt{n}\omega^0$. But (17) is just Basis Pursuit with response vector $y \in \mathbb{R}^n$ and augmented matrix $\begin{bmatrix} X & \sqrt{n}\lambda^{-1} I_n \end{bmatrix}$, so by a result of Daubechies et al. [2010] this is the case if and only if for every $(\beta, \omega) \neq (0, 0)$ such that $X\beta + \sqrt{n}\lambda^{-1}\omega = 0$, we have $|\operatorname{sign}(\beta^0)^T\beta + \operatorname{sign}(\omega^0)^T\omega| < \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1$, which proves our statement. □

We will need the following auxiliary lemma.

Lemma 2. *Under assumptions (i) and (ii), if the pair $(\theta, \tilde{\theta})$ is identifiable with respect to X and λ, then for any $\epsilon \in \mathbb{R}^n$,*

$$\lim_{r \to +\infty} \frac{1}{u_r} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)} \\ \hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where $u_r := \|\beta^{(r)}\|_1 + \lambda\|\omega^{(r)}\|_1$.

Proof. First note that by assumption (ii), $\lim_{r\to+\infty} u_r = +\infty$. Now let $\epsilon \in \mathbb{R}^n$ and denote by $(\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon), \hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon))$ the JP solution when $y = \epsilon$. In particular, one has $\epsilon = X\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon) + \sqrt{n}\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)$, so for every $r \in \mathbb{N}^*$,

$$y^{(r)} = X(\beta^{(r)} + \hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)) + \sqrt{n}(\omega^{(r)} + \hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)).$$

Hence $(\beta^{(r)} + \hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon), \omega^{(r)} + \hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon))$ is feasible for JP when $y = y^{(r)}$, so

$$\begin{aligned} &\frac{\|\hat{\beta}_\lambda^{\mathrm{JP}(r)}\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}(r)}\|_1}{u_r} \\ &\quad\leq \frac{\|\beta^{(r)} + \hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1 + \lambda\|\omega^{(r)} + \hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1}{u_r} \\ &\quad\leq \frac{(\|\beta^{(r)}\|_1 + \lambda\|\omega^{(r)}\|_1) + (\|\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1)}{u_r} \\ &\quad= 1 + \frac{\|\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1}{u_r}. \end{aligned} \tag{18}$$

Therefore

$$\begin{aligned}\frac{1}{u_r}(\|\hat{\beta}_\lambda^{\mathrm{JP}(r)}-\beta^{(r)}\|_1+\lambda\|\hat{\omega}_\lambda^{\mathrm{JP}(r)}-\omega^{(r)}\|_1)&\\ \leq \frac{1}{u_r}((\|\beta^{(r)}\|_1+\lambda\|\omega^{(r)}\|_1)+(\|\hat{\beta}_\lambda^{\mathrm{JP}(r)}\|_1+\lambda\|\hat{\omega}_\lambda^{\mathrm{JP}(r)}\|_1))&\\ = 1+\frac{\|\hat{\beta}_\lambda^{\mathrm{JP}(r)}\|_1+\lambda\|\hat{\omega}_\lambda^{\mathrm{JP}(r)}\|_1}{u_r}&\\ \leq 2+\frac{\|\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1+\lambda\|\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1}{u_r},&\end{aligned} \tag{19}$$

using (18) for last inequality. Since $\lim_{r\to+\infty}\frac{\|\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1+\lambda\|\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1}{u_r}=0$, and since $\begin{bmatrix}\beta\\ \omega\end{bmatrix}\mapsto\|\beta\|_1+\lambda\|\omega\|_1$ defines a norm on $\mathbb{R}^{p+n}$, one deduces that the sequence $\frac{1}{u_r}\begin{bmatrix}\hat{\beta}_\lambda^{\mathrm{JP}(r)}-\beta^{(r)}\\ \hat{\omega}_\lambda^{\mathrm{JP}(r)}-\omega^{(r)}\end{bmatrix}$ is bounded. Therefore we need to check that every convergent subsequence converges to zero. Let

$$\frac{1}{u_{\phi(r)}}\begin{bmatrix}\hat{\beta}_\lambda^{\mathrm{JP}(\phi(r))}-\beta^{(\phi(r))}\\ \hat{\omega}_\lambda^{\mathrm{JP}(\phi(r))}-\omega^{(\phi(r))}\end{bmatrix}$$

(with $\phi:\mathbb{N}^*\to\mathbb{N}^*$ strictly increasing) be an arbitrary convergent subsequence. Since

$$\frac{\|\beta^{(r)}\|_1+\lambda\|\omega^{(r)}\|_1}{u_r}=1 \tag{20}$$

for every r, and by (18), the sequences $\frac{1}{u_r}\begin{bmatrix}\beta^{(r)}\\ \omega^{(r)}\end{bmatrix}$ and $\frac{1}{u_r}\begin{bmatrix}\hat{\beta}_\lambda^{\mathrm{JP}(r)}\\ \hat{\omega}_\lambda^{\mathrm{JP}(r)}\end{bmatrix}$ are bounded as well. Hence without loss of generality (otherwise, reduce the subsequence),

$$\lim_{r\to+\infty}\frac{1}{u_{\phi(r)}}\begin{bmatrix}\beta^{(\phi(r))}\\ \omega^{(\phi(r))}\end{bmatrix}=\begin{bmatrix}\nu_1\\ \nu_2\end{bmatrix}, \tag{21}$$

and

$$\lim_{r\to+\infty}\frac{1}{u_{\phi(r)}}\begin{bmatrix}\hat{\beta}_\lambda^{\mathrm{JP}(\phi(r))}\\ \hat{\omega}_\lambda^{\mathrm{JP}(\phi(r))}\end{bmatrix}=\begin{bmatrix}\nu_1'\\ \nu_2'\end{bmatrix} \tag{22}$$

for some $\begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}, \begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} \in \mathbb{R}^{p+n}$. By (20), one necessarily has

$$\|\nu_1\|_1 + \lambda\|\nu_2\|_1 = 1, \tag{23}$$

and (18) implies that

$$\|\nu_1'\|_1 + \lambda\|\nu_2'\|_1 \leq 1. \tag{24}$$

Now

$$\begin{aligned} \lim_{r\to+\infty} \frac{X(\hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)}) + \sqrt{n}(\hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)})}{u_r} &= \lim_{r\to+\infty} \frac{y^{(r)} - (X\beta^{(r)} + \sqrt{n}\omega^{(r)})}{u_r} \\ &= \lim_{r\to+\infty} \frac{\epsilon}{u_r} = 0, \end{aligned}$$

so one deduces that

$$\lim_{r\to+\infty} \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(\phi(r))}/u_{\phi(r)} \\ \hat{\omega}_\lambda^{\mathrm{JP}(\phi(r))}/u_{\phi(r)} \end{bmatrix} = \lim_{r\to+\infty} \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \beta^{(\phi(r))}/u_{\phi(r)} \\ \omega^{(\phi(r))}/u_{\phi(r)} \end{bmatrix},$$

so by (21) and (22),

$$\begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} = \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}. \tag{25}$$

Assuming for now that (ν_1, ν_2) is identifiable with respect to X and λ, equality (25) together with (23) and (24) imply that $\begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$, hence

$$\lim_{r\to+\infty} \frac{1}{u_{\phi(r)}} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(\phi(r))} - \beta^{(\phi(r))} \\ \hat{\omega}_\lambda^{\mathrm{JP}(\phi(r))} - \omega^{(\phi(r))} \end{bmatrix} = \begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} - \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

It remains to check that (ν_1, ν_2) is identifiable with respect to X and λ, which we will do using Lemma 1. Note that (21) and assumption (i) imply

$$\mathrm{sign}(\nu_1) = \theta - \theta', \tag{26}$$

$$\mathrm{sign}(\nu_2) = \tilde{\theta} - \tilde{\theta}', \tag{27}$$

where $\theta'_j := \theta_j \mathbf{1}_{\{\nu_{1,j}=0, \theta_j \neq 0\}}$, and $\tilde{\theta}'_j = \tilde{\theta}_j \mathbf{1}_{\{\nu_{2,j}=0, \tilde{\theta}_j \neq 0\}}$, and hence

$$\overline{\operatorname{supp}(\nu_1)} = \overline{\operatorname{supp}(\theta)} \sqcup \operatorname{supp}(\theta') = \overline{S^0} \sqcup \operatorname{supp}(\theta'), \tag{28}$$

$$\overline{\operatorname{supp}(\nu_2)} = \overline{\operatorname{supp}(\tilde{\theta})} \sqcup \operatorname{supp} \tilde{\theta}' = \overline{T^0} \sqcup \operatorname{supp} \tilde{\theta}'. \tag{29}$$

Consider a pair $(\beta, \omega) \neq (0, 0)$ such that $X\beta + \sqrt{n}\lambda^{-1}\omega = 0$. By (26) and (27),

$$\begin{aligned} |\operatorname{sign}(\nu_1)^T \beta + \operatorname{sign}(\nu_2)^T \omega| &= |(\theta - \theta')^T \beta + (\tilde{\theta} - \tilde{\theta}')^T \omega| \\ &\leq |\theta^T \beta + \tilde{\theta}^T \omega| + |(\theta')^T \beta| + |(\tilde{\theta}')^T \omega|. \end{aligned} \tag{30}$$

But since $(\theta, \tilde{\theta})$ is identifiable with respect to X and λ, Lemma 1 implies $|\theta^T \beta + \tilde{\theta}^T \omega| < \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1$. Plugging this into (30) gives

$$\begin{aligned} |\operatorname{sign}(\nu_1)^T \beta + \operatorname{sign}(\nu_2)^T \omega| &< \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1 + |(\theta')^T \beta| + |(\tilde{\theta}')^T \omega| \\ &\leq \|\beta_{\overline{S^0}}\|_1 + \|\beta_{\operatorname{supp}(\theta')}\|_1 + \|\omega_{\overline{T^0}}\|_1 + \|\omega_{\operatorname{supp}(\tilde{\theta}')}\|_1 \\ &= \|\beta_{\overline{\operatorname{supp}(\nu_1)}}\|_1 + \|\omega_{\overline{\operatorname{supp}(\nu_2)}}\|_1, \end{aligned}$$

where the equality comes from (28) and (29). By Lemma 1, one concludes that (ν_1, ν_2) is identifiable with respect to X and λ. □

Proof of Theorem 1. Let us assume that $(\theta, \tilde{\theta})$ is identifiable with respect to X and λ, and let $\epsilon \in \mathbb{R}^n$. By Lemma 2,

$$\lim_{r \to +\infty} \frac{1}{u_r} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)} \\ \hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{31}$$

Since

$$\min\{1, \lambda\} \max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\} \leq u_r \leq (|S^0| + \lambda |T^0|) \max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\},$$

(31) is equivalent to $\lim_{r \to +\infty} \frac{1}{\max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)} \\ \hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Therefore there exists $R > 0$ such that for every $r \geq R$,

$$\|\hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)}\|_\infty < \frac{q}{2} \max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\} \tag{32}$$

and

$$\|\hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)}\|_\infty < \frac{q}{2}\max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}. \tag{33}$$

Setting $\tau := \frac{q}{2}\max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}$, (32) implies that $|\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)}| < \tau$ for every $j \notin S^0$, hence $\hat{\beta}_{(\lambda,\tau),j}^{\mathrm{TJP}(r)} = 0$. If $j \in S^0$, assumption (iii) implies

$$|\beta_j^{(r)}| \geq \beta_{\min}^{(r)} \geq 2\tau, \tag{34}$$

and by (32), we have

$$|\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)} - \beta_j^{(r)}| < \tau, \tag{35}$$

so (34) and (35) together imply $|\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)}| > \tau$ and $\mathrm{sign}(\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)}) = \mathrm{sign}(\beta_j^{(r)})$. So we conclude that $\mathrm{sign}(\hat{\beta}_{(\lambda,\tau)}^{\mathrm{TJP}(r)}) = \mathrm{sign}(\beta^{(r)})$. Analogously, (33) implies $\mathrm{sign}(\hat{\omega}_{(\lambda,\tau)}^{\mathrm{TJP}(r)}) = \mathrm{sign}(\omega^{(r)})$.

Conversely, let us assume that for some $\epsilon \in \mathbb{R}^n$, $r \in \mathbb{N}^*$ and $\tau > 0$,

$$\mathrm{sign}(\hat{\beta}_{(\lambda,\tau)}^{\mathrm{TJP}(r)}) = \theta, \quad \mathrm{sign}(\hat{\omega}_{(\lambda,\tau)}^{\mathrm{TJP}(r)}) = \tilde{\theta}. \tag{36}$$

Note that the JP solution $(\hat{\beta}_\lambda^{\mathrm{JP}(r)}, \hat{\omega}_\lambda^{\mathrm{JP}(r)})$ is unique by assumption, hence $(\hat{\beta}_\lambda^{\mathrm{JP}(r)}, \hat{\omega}_\lambda^{\mathrm{JP}(r)})$ is identifiable with respect to X and λ. Now by (36), all nonzero components of θ and $\tilde{\theta}$ must have the same sign as the corresponding entries of $\hat{\beta}_\lambda^{\mathrm{JP}(r)}$ and $\hat{\omega}_\lambda^{\mathrm{JP}(r)}$ respectively. Hence

$$\begin{aligned} \theta &= \mathrm{sign}(\theta) = \mathrm{sign}(\hat{\beta}_\lambda^{\mathrm{JP}(r)}) - \delta, \\ \tilde{\theta} &= \mathrm{sign}(\tilde{\theta}) = \mathrm{sign}(\hat{\omega}_\lambda^{\mathrm{JP}(r)}) - \tilde{\delta}, \end{aligned} \tag{37}$$

where $\delta_j = \mathrm{sign}(\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)})\mathbf{1}_{\{\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)} \neq 0, \theta_j = 0\}}$ and $\tilde{\delta}_i = \mathrm{sign}(\hat{\omega}_{\lambda,i}^{\mathrm{JP}(r)})\mathbf{1}_{\{\hat{\omega}_{\lambda,i}^{\mathrm{JP}(r)} \neq 0, \tilde{\theta}_i = 0\}}$, and

$$\begin{aligned} \overline{S^0} &= \overline{\mathrm{supp}(\theta)} = \overline{\mathrm{supp}(\hat{\beta}_\lambda^{\mathrm{JP}(r)})} \sqcup \mathrm{supp}(\delta) \\ \overline{T^0} &= \overline{\mathrm{supp}(\tilde{\theta})} = \overline{\mathrm{supp}(\hat{\omega}_\lambda^{\mathrm{JP}(r)})} \sqcup \mathrm{supp}(\tilde{\delta}). \end{aligned} \tag{38}$$

In order to apply Lemma 1, let us consider a pair $(\beta, \omega) \neq (0, 0)$ such that $X\beta + \sqrt{n}\lambda^{-1}\omega = 0$. By (37), one has

$$\begin{aligned}
|\theta^T\beta + \tilde{\theta}^T\omega| &= |\operatorname{sign}(\hat{\beta}_\lambda^{\mathrm{JP}(r)})^T\beta - \delta^T\beta + \operatorname{sign}(\hat{\omega}_\lambda^{\mathrm{JP}(r)})^T\omega - \tilde{\delta}^T\omega| \\
&\leq |\operatorname{sign}(\hat{\beta}_\lambda^{\mathrm{JP}(r)})^T\beta + \operatorname{sign}(\hat{\omega}_\lambda^{\mathrm{JP}(r)})^T\omega| + |\delta^T\beta| + |\tilde{\delta}^T\omega| \\
&\leq \|\beta_{\overline{\operatorname{supp}(\hat{\beta}_\lambda^{\mathrm{JP}(r)})}}\|_1 + \|\omega_{\overline{\operatorname{supp}(\hat{\omega}_\lambda^{\mathrm{JP}(r)})}}\|_1 + \|\beta_{\operatorname{supp}(\delta)}\|_1 + \|\omega_{\operatorname{supp}(\tilde{\delta})}\|_1 \\
&= \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1,
\end{aligned}$$

where we have used Lemma 1 and the fact that $(\hat{\beta}_\lambda^{\mathrm{JP}(r)}, \hat{\omega}_\lambda^{\mathrm{JP}(r)})$ is identifiable with respect to X and λ in the last inequality, and (38) for the last equality. Lemma 1 concludes our proof. □

B Proof of Theorem 2

Proof of Theorem 2. We define $\tilde{X} := \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix}$, and $\tilde{\nu} = \begin{bmatrix} \tilde{\beta} \\ \tilde{\omega} \end{bmatrix} := \tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}\epsilon$. We will assume for now that the following properties hold.

a) Every pair (β, ω) such that $X\beta + \sqrt{n}\omega = 0$ satisfies

$$\|\beta_{S^0}\|_1 + \lambda\|\omega_{T^0}\|_1 \leq \frac{1}{3}(\|\beta_{\overline{S^0}}\|_1 + \lambda\|\omega_{\overline{T^0}}\|_1),$$

b) $\|\tilde{\nu}\|_2 \leq \frac{\sqrt{2}\sigma}{\left(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1\right)^{1/2}}$.

Since $\tilde{X}\tilde{\nu} = X\tilde{\beta} + \sqrt{n}\tilde{\omega} = \epsilon$, one can rewrite model (2) as

$$y = X(\beta^0 + \tilde{\beta}) + \sqrt{n}(\omega^0 + \tilde{\omega}).$$

By property a) and Lemma 3 below, one has

$$\|\hat{\beta}_\lambda^{\mathrm{JP}} - (\beta^0 + \tilde{\beta})\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}} - (\omega^0 + \tilde{\omega})\|_1 \leq 4(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1), \qquad (39)$$

and therefore $\|\hat{\beta}^{\mathrm{JP}}_{\lambda} - (\beta^0 + \tilde{\beta})\|_1 \leq 4(\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1)$. Consequently, for any $j \in [p]$ one has

$$\begin{aligned}
|\hat{\beta}^{\mathrm{JP}}_{\lambda,j} - \beta^0_j| &\leq |\hat{\beta}^{\mathrm{JP}}_{\lambda,j} - (\beta^0_j + \tilde{\beta}_j)| + |\tilde{\beta}_j| \leq \|\hat{\beta}^{\mathrm{JP}}_{\lambda} - (\beta^0 + \tilde{\beta})\|_1 + \|\tilde{\beta}\|_1 \\
&\leq 4(\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1) + \|\tilde{\beta}\|_1 \leq 5(\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1) \\
&\leq 5\max\{1,\lambda\}(\|\tilde{\beta}\|_1 + \|\tilde{\omega}\|_1) = 5\max\{1,\lambda\}\|\tilde{\nu}\|_1 \\
&\leq 5\max\{1,\lambda\}\sqrt{p+n}\|\tilde{\nu}\|_2 \leq \frac{5\sqrt{2}\max\{1,\lambda\}\sigma\sqrt{p+n}}{(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1)^{1/2}}
\end{aligned}$$

where we have used property b) in the last inequality. Now setting

$$\tau := \frac{5\sqrt{2}\max\{1,\lambda\}\sigma\sqrt{p+n}}{(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1)^{1/2}},$$

one gets

$$|\hat{\beta}^{\mathrm{JP}}_{\lambda,j} - \beta^0_j| \leq \tau \tag{40}$$

for every $j \in [p]$. If $j \in \overline{S^0}$, we have $|\hat{\beta}^{\mathrm{JP}}_{\lambda,j}| \leq \tau$, hence $\hat{\beta}^{\mathrm{TJP}}_{(\lambda,\tau),j} = 0$. If $j \in S^0$, assumption (12) implies $|\beta^0_j| > 2\tau$, which together with (40) gives $\mathrm{sign}(\hat{\beta}^{\mathrm{TJP}}_{(\lambda,\tau),j}) = \mathrm{sign}(\beta^0_j)$.

It remains to prove that properties a) and b) hold with high probability. First, Lemma 1 in Nguyen and Tran [2013a], implies that with probability greater than $1 - ce^{-c'n}$ the matrix X satisfies the extended restricted eigenvalue property

$$\begin{gathered}
\|\beta_{\overline{S^0}}\|_1 + \lambda\|\omega_{\overline{T^0}}\|_1 \leq 3(\|\beta_{S^0}\|_1 + \lambda\|\omega_{T^0}\|_1) \\
\Downarrow \\
\frac{1}{n}\|X\beta + \sqrt{n}\omega\|_2^2 \geq \gamma^2(\|\beta\|_2^2 + \|\omega\|_2^2),
\end{gathered} \tag{41}$$

with $\gamma^2 = \frac{\min\{\lambda_{\min}(\Sigma),1\}}{16^2}$. Property (41) clearly implies a). Finally, Lemma 4 below proves that b) holds with probability at least $1 - 1.14^{-n} - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}$, which concludes our proof. □

Lemma 3. *Assume that for some sets $S^0 \subset [p]$ and $T^0 \subset [n]$, and some constant $\rho \in (0,1)$, the matrix $X \in \mathbb{R}^{n\times p}$ satisfies*

$$\|\beta_{S^0}\|_1 + \lambda\|\omega_{T^0}\|_1 \leq \rho(\|\beta_{\overline{S^0}}\|_1 + \lambda\|\omega_{\overline{T^0}}\|_1), \tag{42}$$

for every pair $(\beta,\omega) \in \mathbb{R}^p \times \mathbb{R}^n$ such that $X\beta + \sqrt{n}\omega = 0$. Then for every pair $(\tilde{\beta},\tilde{\omega}) \in \mathbb{R}^p \times \mathbb{R}^n$, the solution $(\hat{\beta}^{\mathrm{JP}}_\lambda, \hat{\omega}^{\mathrm{JP}}_\lambda)$ to JP (6) with $y = X\tilde{\beta} + \sqrt{n}\tilde{\omega}$ satisfies

$$\|\hat{\beta}^{\mathrm{JP}}_\lambda - \tilde{\beta}\|_1 + \lambda\|\hat{\omega}^{\mathrm{JP}}_\lambda - \tilde{\omega}\|_1 \leq \frac{2(1+\rho)}{1-\rho}(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1).$$

Proof. This proof is a simple extension of the one of Theorem 4.14 in Foucart and Rauhut [2013]. Let us consider $y = X\tilde{\beta} + \sqrt{n}\tilde{\omega}$ for an arbitrary pair $(\tilde{\beta},\tilde{\omega})$, and define $\beta' := \hat{\beta}^{\mathrm{JP}}_\lambda - \tilde{\beta}$ and $\omega' := \hat{\omega}^{\mathrm{JP}}_\lambda - \tilde{\omega}$. Clearly $X\beta' + \sqrt{n}\omega' = 0$, so by (42),

$$\|\beta'_{S^0}\|_1 + \lambda\|\omega'_{T^0}\|_1 \leq \rho(\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1). \tag{43}$$

We also have

$$\begin{aligned}
\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1 &= \|\tilde{\beta}_{S^0}\|_1 + \|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda(\|\tilde{\omega}_{T^0}\|_1 + \|\tilde{\omega}_{\overline{T^0}}\|_1) \\
&= \|\hat{\beta}^{\mathrm{JP}}_{\lambda,S^0} - \beta'_{S^0}\|_1 + \|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda(\|\hat{\omega}^{\mathrm{JP}}_{\lambda,T^0} - \omega'_{T^0}\|_1 + \|\tilde{\omega}_{\overline{T^0}}\|_1) \\
&\leq \|\hat{\beta}^{\mathrm{JP}}_{\lambda,S^0}\|_1 + \|\beta'_{S^0}\|_1 + \|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda(\|\hat{\omega}^{\mathrm{JP}}_{\lambda,T^0}\|_1 + \|\omega'_{T^0}\|_1 + \|\tilde{\omega}_{\overline{T^0}}\|_1),
\end{aligned}$$

and

$$\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \leq (\|\hat{\beta}^{\mathrm{JP}}_{\lambda,\overline{S^0}}\|_1 + \|\tilde{\beta}_{\overline{S^0}}\|_1) + \lambda(\|\hat{\omega}^{\mathrm{JP}}_{\lambda,\overline{T^0}}\|_1 + \|\tilde{\omega}_{\overline{T^0}}\|_1).$$

Adding the last two inequalities yields

$$\begin{aligned}
\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 + \|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1 \leq \|\hat{\beta}^{\mathrm{JP}}_\lambda\|_1 + \|\beta'_{S^0}\|_1 + 2\|\tilde{\beta}_{\overline{S^0}}\|_1 \\
+ \lambda(\|\hat{\omega}^{\mathrm{JP}}_\lambda\|_1 + \|\omega'_{T^0}\|_1 + 2\|\tilde{\omega}_{\overline{T^0}}\|_1),
\end{aligned}$$

and rearranging terms gives

$$\begin{aligned}
\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \leq (\|\hat{\beta}^{\mathrm{JP}}_\lambda\|_1 + \lambda\|\hat{\omega}^{\mathrm{JP}}_\lambda\|_1) - (\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1) \\
+ (\|\beta'_{S^0}\|_1 + \lambda\|\omega'_{T^0}\|_1) + 2(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1).
\end{aligned}$$

Using (43) and the fact that $\|\hat{\beta}^{\mathrm{JP}}_\lambda\|_1+\lambda\|\hat{\omega}^{\mathrm{JP}}_\lambda\|_1 \leq \|\tilde{\beta}\|_1+\lambda\|\tilde{\omega}\|_1$ by minimality of the JP solution, we get

$$\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \leq \rho(\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1) + 2(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1),$$

hence

$$\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \leq \frac{2}{1-\rho}(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1). \tag{44}$$

Now inequality (43) also implies

$$\begin{aligned}\|\beta'\|_1 + \lambda\|\omega'\|_1 &= \|\beta'_{S^0}\|_1 + \lambda\|\omega'_{T^0}\|_1 + \|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \\ &\leq (1+\rho)(\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1)\end{aligned} \tag{45}$$

and continuing (45) with (44) gives the desired inequality. □

Lemma 4. *Let* $\tilde{X} := \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix}$. *Under assumptions iv), v), vi) and vii),*

$$\|\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}\epsilon\|_2 \leq \frac{\sqrt{2}\sigma}{(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1)^{1/2}},$$

with probability at least $1 - 1.14^{-n} - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}$.

Proof. We have

$$\|\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}\epsilon\|_2^2 = \epsilon^T(\tilde{X}\tilde{X}^T)^{-1}\epsilon \leq \frac{\|\epsilon\|_2^2}{\lambda_{\min}(\tilde{X}\tilde{X}^T)} = \frac{\sigma\|\frac{1}{\sigma}\epsilon\|_2^2}{\lambda_{\min}(\tilde{X}\tilde{X}^T)}.$$

Since $\|\frac{1}{\sigma}\epsilon\|_2^2 \sim \chi^2_n$, it is upper bounded by $2n$ with probability larger than $1 - 1.14^{-n}$ (a corollary of Lemma 1 in Laurent and Massart [2000]). So

$$\mathbb{P}\left(\|\tilde{\nu}\|_2 \leq \frac{\sqrt{2n}\sigma}{\sigma_{\min}(\tilde{X})}\right) \geq 1 - 1.14^{-n}. \tag{46}$$

Let us now bound $\sigma_{\min}(\tilde{X})$. One has

$$\sigma^2_{\min}(\tilde{X}) = \lambda_{\min}(\tilde{X}\tilde{X}^T) = \lambda_{\min}(XX^T + nI_n) = \sigma^2_{\min}(X) + n. \tag{47}$$

One can write $X = G\Sigma^{1/2}$ where $G \in \mathbb{R}^{n\times p}$ with $G_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1)$, thus

$$\sigma_{\min}(X) \geq \sigma_{\min}(G)\sigma_{\min}(\Sigma^{1/2}) = \sigma_{\min}(G)\sqrt{\lambda_{\min}(\Sigma)}. \tag{48}$$

Now it is known (see Rudelson and Vershynin [2010], eq. (2.3)) that

$$\sigma_{\min}(G) \geq \frac{1}{2}(\sqrt{p} - \sqrt{n}) = \frac{\sqrt{n}}{2}(\sqrt{p/n} - 1)$$

with probability at least $1 - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}$. Together with (47) and (48) this gives

$$\mathbb{P}\left(\sigma_{\min}(\tilde{X}) \geq \left(\frac{n\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n} - 1)^2 + n\right)^{1/2}\right) \geq 1 - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}.$$

With (46), this implies

$$\mathbb{P}\left(\|\tilde{\nu}\|_2 \leq \frac{\sqrt{2}\sigma}{(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n} - 1)^2 + 1)^{1/2}}\right) \geq 1 - 1.14^{-n} - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}.$$

□

C Variables in the Traumabase dataset

The variables of the Traumabase dataset are:

- *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.
- *Delta.Hemo*: The difference between the homoglobin on arrival at hospital and that in the ambulance.
- *RBC*: A binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed.

- *SI.amb*: Shock index measured on ambulance.
- *DBP.min*: Minimum value of measured diastolic blood pressure in the ambulance.
- *SBP.min*: Minimum value of measured systolic blood pressure in the ambulance.
- *HR.max*: Maximum value of measured heart rate in the ambulance.
- *VE*: A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system.
- *MBP.amb*: Mean arterial pressure measured in the ambulance.
- *Temp*: Patient's body temperature.
- *SI*: Shock index $SI = HR/SBP$ indicates level of occult shock based on heart rate and systolic blood pressure on arrival at hospital.
- *MBP*: Mean arterial pressure $MBP = (2DBP+SPB)/3$ is an average blood pressure in an individual during a single cardiac cycle.
- *HR*: Heart rate measured on arrival of hospital.
- *Age*: Age.

References

Yudong Chen and Constantine Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In International Conference on Machine Learning, pages 383–391, 2013.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust Sparse Regression under Adversarial Corruption. In International Conference on Machine Learning, pages 774–782, 2013a. URL http://proceedings.mlr.press/v28/chen13h.html.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In International Conference on Machine Learning, pages 774–782, 2013b.

Arnak S Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. arXiv preprint arXiv:1904.06288, 2019.

Abhirup Datta and Hui Zou. CoCoLasso for high-dimensional error-in-variables regression. The Annals of Statistics, 45(6):2400–2426, 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1527. URL https://projecteuclid.org/euclid.aos/1513328577.

Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C. Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. Communications on Pure and Applied Mathematics, 63(1): 1–38, 2010. ISSN 1097-0312. doi: 10.1002/cpa.20303. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20303.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

Pascaline Descloux and Sylvain Sardy. Model selection with lasso-zero: adding straw to the haystack to better find needles. arXiv:1805.05133

[stat], 2018. URL http://arxiv.org/abs/1805.05133. arXiv: 1805.05133.

Simon Foucart and Holger Rauhut. A mathematical introduction to compressive sensing. Number 3. Birkhäuser Basel, 2013. URL http://www.ams.org/bull/2017-54-01/S0273-0979-2016-01546-1/.

Ramon I Garcia, Joseph G Ibrahim, and Hongtu Zhu. Variable selection for regression models with missing data. Statistica Sinica, 20(1):149, 2010.

Caroline Giacobino, Sylvain Sardy, Jairo Diaz-Rodriguez, and Nick Hengartner. Quantile universal threshold. Electronic Journal of Statistics, 11(2):4701–4722, 2017. ISSN 1935-7524. doi: 10.1214/17-EJS1366. URL https://projecteuclid.org/euclid.ejs/1511492459.

Wei Jiang, Malgorzata Bogdan, Julie Josse, Blazej Miasojedow, Veronika Rockova, and TraumaBase Group. Adaptive Bayesian SLOPE – High-dimensional Model Selection with Missing Values. arXiv e-prints, art. arXiv:1909.06631, Sep 2019.

J. N. Laska, M. A. Davenport, and R. G. Baraniuk. 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, 2009. doi: 10.1109/ACSSC.2009.5470141.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. The Annals of Statistics, 28(5):1302–1338, 2000. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1015957395. URL https://projecteuclid.org/euclid.aos/1015957395.

Xiaodong Li. Compressed Sensing and Matrix Completion with Constant Proportion of Corruptions. Constructive Approximation, 37(1):73–99, 2013. ISSN 1432-0940. doi: 10.1007/s00365-012-9176-9. URL https://doi.org/10.1007/s00365-012-9176-9.

Zhi Li, Feng Wu, and John Wright. On the systematic measurement matrix for compressed sensing in the presence of gross errors. In 2010 Data Compression Conference, pages 356–365. IEEE, 2010.

Ying Liu, Yuanjia Wang, Yang Feng, and Melanie M Wall. Variable selection and prediction with incomplete high-dimensional data. The annals of applied statistics, 10(1):418, 2016.

Po-Ling Loh and Martin J. Wainwright. High-Dimensional Regression with Noisy and Missing Data: Provable Guarantees with Nonconvexity. The Annals of Statistics, 40(3):1637–1664, 2012. ISSN 0090-5364. URL http://www.jstor.org/stable/41713688.

N. H. Nguyen and T. D. Tran. Robust Lasso With Missing and Grossly Corrupted Observations. IEEE Transactions on Information Theory, 59 (4):2036–2058, 2013a. ISSN 0018-9448. doi: 10.1109/TIT.2012.2232347.

N. H. Nguyen and T. D. Tran. Exact Recoverability From Dense Corrupted Observations via ℓ_1-Minimization. IEEE Transactions on Information Theory, 59(4):2017–2035, 2013b. ISSN 0018-9448. doi: 10.1109/TIT.2013.2240435.

Mathieu Rosenbaum, Alexandre B Tsybakov, et al. Improved matrix uncertainty selector. In From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner, pages 276–290. Institute of Mathematical Statistics, 2013.

Donald B Rubin. Inference and missing data. Biometrika, 63(3):581–592, 1976.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. arXiv:1003.2990 [math], 2010. URL http://arxiv.org/abs/1003.2990. arXiv: 1003.2990.

Sylvain Sardy, Paul Tseng, and A. G. Bruce. Robust wavelet denoising. IEEE Transactions on Signal Processing, 49:1146–1152, 2001.

Patrick Tardivel and Malgorzata Bogdan. On the sign recovery by lasso, thresholded lasso and thresholded basis pursuit denoising. 2019.

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

Stef Van Buuren. Flexible imputation of missing data. Chapman and Hall/CRC, 2018.

J. Wright and Y. Ma. Dense Error Correction Via ℓ_1-Minimization. IEEE Transactions on Information Theory, 56(7):3540–3560, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2048473.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust Face Recognition via Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):210–227, 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.79.