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

One World : statistical learning seminars

Pascaline Descoux [1] Claire Boyer [2,3] Julie Josse [4,5,6] Sylvain Sardy [1] Aude Sportisse [2]

[1]University of Geneva, Switzerland

[2]Sorbonne University, France

[3]Ecole Normale Supérieure, France

[4]Ecole Polytechnique, France

[5]INRIA, France

[6]Visiting Researcher Google Brain, France

The sparse corruptions problem

For taking into account occasional corruptions:

Sparse corruptions problem

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

- $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta^0 \in \mathbb{R}^p$, $\omega^0 \in \mathbb{R}^n$
- $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$.
- high-dimension: $p \gg n$, $\mathrm{rank}(X) = n$.
- sparsity: β^0 is s-sparse, ω^0 is k-sparse.

The sparse corruptions problem

For taking into account additional occasional corruptions:

Sparse corruptions problem

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

- $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta^0 \in \mathbb{R}^p$, $\omega^0 \in \mathbb{R}^n$
- $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$.
- high-dimension: $p \gg n$, $\mathrm{rank}(X) = n$.
- sparsity: β^0 is s-sparse, ω^0 is k-sparse.

Sparse linear model with an augmented design matrix and sparse vector

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

The sparse corruptions problem

Face recognition problem

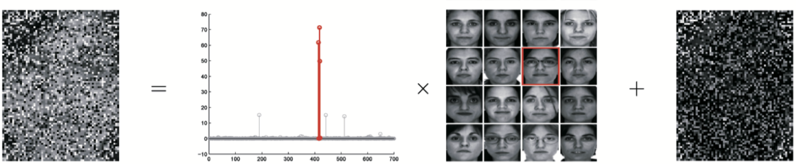


Figure: Corrupted image, a sparse linear combination of all the training images (middle) plus sparse errors (right) due to corruption. Red (darker) coefficients correspond to training images of the correct individual.

Credit: [Wright et al., 2009].

Existing works

With sparse noise ($\omega^0 \neq 0$), without dense noise ($\epsilon = 0$): Justice Pursuit (JP)

$$\min_{\beta\in\mathbb{R}^p,\omega\in\mathbb{R}^n} \|\beta\|_1 + \lambda\|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}\omega,\ \lambda > 0$$

	Condition on (y, X)	Recovery of (β^0, ω^0)
Wright, Ma (2010)	y Gaussian	✓ (support)
Laska et al. (2009) Li et al. (2010)	y, X Gaussian	✓ (exact)
JP with tuned paramater Li (2013) Nguyen and Tran (2013b)	X sub-orthogonal Gaussian design	✓ (exact)

Existing works

With sparse noise ($\omega^0 \neq 0$), without dense noise ($\epsilon = 0$): Justice Pursuit (JP)

$$\min_{\beta \in \mathbb{R}^p, \omega \in \mathbb{R}^n} \|\beta\|_1 + \lambda \|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}\omega, \ \lambda > 0$$

With sparse noise ($\omega^0 \neq 0$) and dense noise ($\epsilon \neq 0$): Robust Lasso

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

	Condition on (y, X)	Recovery of (β^0, ω^0)
Nguyen and Tran (2013b)	X Gaussian invertible covariance matrix	✓ (sign)
ℓ_1-penalized Huber's M-estimator Dalalyan and Thompson (2019)	X Gaussian invertible covariance matrix	✓ (sign)

Existing works

With sparse noise ($\omega^0 \neq 0$), without dense noise ($\epsilon = 0$): Justice Pursuit (JP)

$$\min_{\beta \in \mathbb{R}^p, \omega \in \mathbb{R}^n} \|\beta\|_1 + \lambda \|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}\omega, \ \lambda > 0$$

With sparse noise ($\omega^0 \neq 0$) and dense noise ($\epsilon \neq 0$): Robust Lasso

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

- Our proposal: same problem but different strategy, solving Justice Pursuit and thresholding.

Our strategy: "Overfit, then threshold."

Sparse-linear model

Strategy already introduced for the sparse-linear model $y = X\beta^0 + \epsilon$.

Thresholded Basis Pursuit [Saligrama and Zhao, 2011]

- solving the Basis Pursuit $\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \quad \text{s.t.} \quad y = X\beta.$
- setting the small coefficients to zero.

7 noise generally overfitted.

Our strategy: "Overfit, then threshold."

Sparse-linear model

Strategy already introduced for the sparse-linear model $y = X\beta^0 + \epsilon$.

Thresholded Basis Pursuit [Saligrama and Zhao, 2011]

- solving the Basis Pursuit $\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \quad \text{s.t.} \quad y = X\beta.$
- setting the small coefficients to zero.

7 noise generally overfitted.

Lasso-Zero [Descloux and Sardy, 2018]

- For $k \in \{0, \dots, M\}$
  - use a Gaussian noise dictionary $G^{(k)} \in \mathbb{R}^{n \times q}$, $q > 0$.
  - solve BP problems with the augmented matrix $[X|G^{(k)}]$.
- Aggregate the obtained estimates $\hat{\beta}^{(1)}, \dots, \hat{\beta}^{(M)}$ with the component-wise medians.
- Threshold the aggregated estimator at level $\tau > 0$.

Our strategy: "Overfit, then threshold."

Sparse corruption model

"Thresholded Justice Pursuit"

- solving the Justice Pursuit $\Rightarrow \hat{\beta}_\lambda^{\mathrm{JP}}, \hat{\omega}_\lambda^{\mathrm{JP}}$
- hard-thresholding the solution $\hat{\beta}^{\mathrm{TJP}}_{(\lambda,\tau)} = \eta_\tau(\hat{\beta}_\lambda^{\mathrm{JP}})$ and $\hat{\omega}^{\mathrm{TJP}}_{(\lambda,\tau)} = \eta_\tau(\hat{\omega}_\lambda^{\mathrm{JP}})$.

Robust Lasso-Zero

- For $k \in \{0, \ldots, M\}$
  - use a Gaussian noise dictionary $G^{(k)} \in \mathbb{R}^{n\times q}$, $q > 0$.
  - solve the augmented JP problem

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

- Aggregate the obtained estimates $\hat{\beta}_\lambda^{(1)}, \ldots, \hat{\beta}_\lambda^{(M)}$ with the component-wise medians $\Rightarrow \hat{\beta}_\lambda^{\mathrm{med}}$
- Hard-threshold $\hat{\beta}^{\mathrm{Rlass0}}_{(\lambda,\tau)} := \eta_\tau(\hat{\beta}_\lambda^{\mathrm{med}}) = \hat{\beta}_\lambda^{\mathrm{med}} \mathbf{1}_{(\tau,+\infty)}(|\hat{\beta}_\lambda^{\mathrm{med}}|)$.

Theoretical guarantees on Thresholded Justice Pursuit

Identifiability as a necessary and su□cient condition for consistent sign recovery

☼ see the Robust Lasso-Zero as an extension of the Thresholded Justice Pursuit (TJP). → theoretical results derived for TJP.

Identifiability for the TJP

= Extension of [Tardivel and Bogdan, 2019] for the TBP.

$(\beta^0, \omega^0) \in \mathbb{R}^p \times \mathbb{R}^n$ is identifiable with respect to $X \in \mathbb{R}^{n\times p}$ and $\lambda > 0$ if it is the unique solution to JP when $y = X\beta^0 + \sqrt{n}\omega^0$ (noiseless case).

For a fixed matrix $X \in \mathbb{R}^{n\times p}$ and a sequence $\{(\beta^{(r)}, \omega^{(r)})\}_{r\in\mathbb{N}*}$ assume

- the sign vectors of $\beta^{(r)}$ and $\omega^{(r)}$ are invariant, i.e.

  $\exists\theta \in \{1, -1, 0\}^p$ such that $\text{sign}(\beta^{(r)}) = \theta, \forall r \in \mathbb{N}^*$ (resp. for $\omega^{(r)}$),

- the nonzero coe□cients are large i.e.

$$\lim_{r\to+\infty} \min\{\beta^{(r)}_{\min}, \omega^{(r)}_{\min}\} = +\infty \quad \text{and} \quad \exists q > 0, \frac{\min\{\beta^{(r)}_{\min}, \omega^{(r)}_{\min}\}}{\max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}} \geqslant q,$$

where $\beta_{\min} := \min_{j\in\text{supp}(\beta)} |\beta_j|$.

Theoretical guarantees on Thresholded Justice Pursuit

Identifiability as a necessary and su□cient condition for consistent sign recovery

$(\hat\beta^{\text{TJP}(r)}_{(\lambda,\tau)}, \hat\omega^{\text{TJP}(r)}_{(\lambda,\tau)})$: TJP estimates when $y = y^{(r)} := X\beta^{(r)} + \sqrt{n}\omega^{(r)} + \epsilon$.

Theorem 1 (Descloux, Boyer, Josse, S., Sardy, 2020)

Let $\lambda > 0$ and $X \in \mathbb{R}^{n\times n}$ such that for any $y \in \mathbb{R}^n$, the JP solution is unique. Let $\{(\beta^{(r)}, \omega^{(r)})\}_{r\in\mathbb{N}*}$ be a sequence satisfying assumptions above.

- If the pair of sign vectors $(\theta, \tilde\theta)$ is identifiable w.r.t. X and λ, then $\exists R, \forall r \geqslant R$, there is a threshold $\tau = \tau(r) > 0$ for which

$$\text{sign}(\hat\beta^{\text{TJP}(r)}_{(\lambda,\tau)}) = \theta \quad \text{and} \quad \text{sign}(\hat\omega^{\text{TJP}(r)}_{(\lambda,\tau)}) = \tilde\theta. \tag{1}$$

- Conversely, if for some $\epsilon \in \mathbb{R}^n$ and $r \in \mathbb{N}^*$ there is a threshold $\tau > 0$ such that (1) holds, then $(\theta, \tilde\theta)$ is identifiable w.r.t. X and λ.

Theoretical guarantees on Thresholded Justice Pursuit

Sign consistency of TJP for correlated Gaussian designs

→ How large the coe□cients should scale to be correctly detected?

Assume a correlated Gaussian design, i.e. $X_{i.} \in \mathbb{R}^p \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ and

- the smallest eigenvalue of the covariance matrix Σ is positive,
- the variance of the covariates is equal to one,
- the noise is assumed to be Gaussian.

Theorem 2 (Descloux, Boyer, Josse, S., Sardy, 2020)

Under the correlated Gaussian design above and signal-to-noise ratio high enough, TJP successfully recovers $\text{sign}(\beta^0)$ with high probability, even with a positive fraction of corruptions.

Theoretical guarantees on Thresholded Justice Pursuit

Sign consistency of TJP for correlated Gaussian designs

→ How large the coe□cients should scale to be correctly detected?

Theorem 2 (if Σ is well-conditioned and $p/n \to \delta > 1$)

- Assume that the eigenvalues of Σ are bounded: $0 < \gamma_1 \leqslant \lambda_{\mathrm{min}}(\Sigma) \leqslant \lambda_{\mathrm{max}}(\Sigma) \leqslant \gamma_2$
- Assume $p/n \to \delta > 1$.

TJP achieves sign consistency provided that

$$n = \Omega(s \log p), \; k = \mathcal{O}(n) \text{ and } \beta^0_{\mathrm{min}} = \Omega(\sqrt{n}).$$

(s: sparsity of β^0, k: sparsity of ω^0)

Missing covariates in high dimension

- X partially known, we observe (y, X^{NA}) instead of (y, X)
- very few works for dealing with missing covariates in the high dimensional setting.

(Liu et al., 2016)	multiple imputation	increasingly complex
(Rosenbaum et al., 2013)	modified Dantzig selector	estimation Σ
(Loh and Wainwright, 2012) (Datta and Zou, 2017)	modified LASSO	estimation Σ
(Jiang et al., 2019)	Adaptive Bayesian SLOPE	estimation Σ

7 Parametric assumption on the covariates for the estimation of Σ.

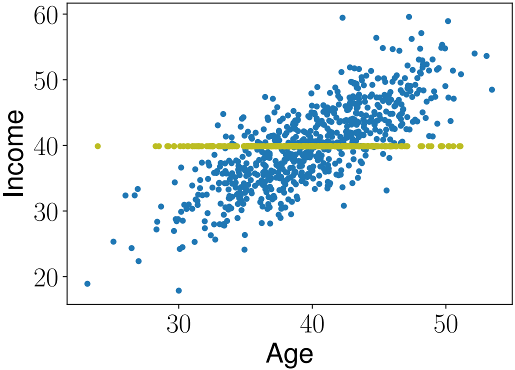
7 Restrictive assumptions on the missing-data mechanism: missingness completely at random.

7 Not suitable in practice.

Missing data as corruptions

- ✓ What we can do (well): impute "naively" the missing entries (by the mean for example).
- 7 Bias in the estimates

ex: Income is missing.



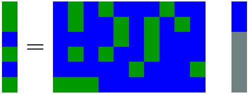
Missing data as corruptions

✓ What we can do (well): impute "naively" the missing entries (by the mean for example).

7 Bias in the estimates

1. Impute "naively" the missing entries in X^{NA} to get $\tilde{X}$ and then correct the imputation error.
2. See the imputation error as a corruption.

$$y = X\beta^0$$



Missing data as corruptions

✓ What we can do (well): impute "naively" the missing entries (by the mean for example).

7 Bias in the estimates

1. Impute "naively" the missing entries in X^{NA} to get $\tilde{X}$ and then correct the imputation error.
2. See the imputation error as a corruption.

How to solve $y = X\beta^0 + \epsilon$ if we observe (X^{NA}, y) ?

- rewrite the model in the form of the sparse corruption model, where

$$\omega^0 := \frac{1}{\sqrt{n}}(X - \tilde{X})\beta^0$$

is the corruption due to imputations.

Robust Lasso-Zero for dealing with missing data

Robust Lasso-Zero for missing data

- Impute "naively" X^{NA} and rescale the imputed matrix X.
- Run Robust Lasso-Zero algorithm with the design matrix X.

✓ without specify a model for the covariates or the missing data mechanism

✓ without estimation of the covariates covariance matrix or of the noise variance,

✓ simple method for the user.

Simulation settings

- $X \sim \mathcal{N}(0, \Sigma)$, $\Sigma \in \mathbb{R}^{200\times 200}$ is a Toeplitz matrix, s.t. $\Sigma_{ij} = \rho^{|i-j|}$.
- noise level $\sigma = 0.5$, coe□cient β^0 drawn uniformly from $\{\pm 1\}$.
- Missing values: MCAR (random missingness, $a = 0$) or MNAR (informative missingness, $a \neq 0$).

$$\mathbb{P}(X_{ij}^{\text{NA}} = \text{NA} \mid X_{ij} = x) = \frac{1}{1 + e^{-a|x|-b}},\ a \geqslant 0 \text{ and } b \in \mathbb{R}.$$

Methods

- Rlass0: Robust Lasso-Zero using $M = 30$ noisy dictionaries. The tuning parameters are obtained using $\lambda = 1$ and selecting τ by quantile universal threshold (QUT) at level $\alpha = 0.05$.
- lass0: Lasso-Zero [Descloux and Sardy, 2018]. The automatic tuning is performed by QUT, at level $\alpha = 0.05$.
- lasso: Lasso [Tibshirani, 1996] performed on the mean-imputed matrix where the regularization parameter is tuned by cross-validation.
- NClasso: the nonconvex ℓ_1 estimator of [Loh and Wainwright, 2012].
- ABSLOPE: Adaptive Bayesian SLOPE of [Jiang et al., 2019].

Simulation settings

- $X \sim \mathcal{N}(0, \Sigma)$, $\Sigma \in \mathbb{R}^{200\times 200}$ is a Toeplitz matrix, s.t. $\Sigma_{ij} = \rho^{|i-j|}$.
- noise level $\sigma = 0.5$, coe□cient β^0 drawn uniformly from $\{\pm 1\}$.
- Missing values: MCAR (random missingness, $a = 0$) or MNAR (informative missingness, $a \neq 0$).

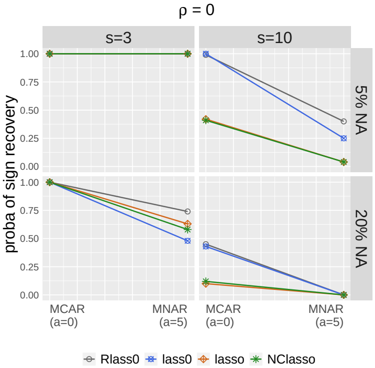
$$\mathbb{P}(X_{ij}^{\mathrm{NA}} = \mathsf{NA} \mid X_{ij} = x) = \frac{1}{1 + e^{-a|x|-b}},\ a \geqslant 0 \text{ and } b \in \mathbb{R}.$$

Performance evaluation

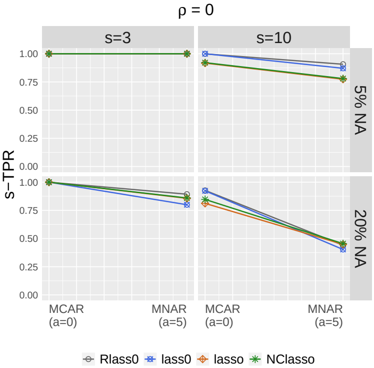
- the Probability of Sign Recovery (PSR), $\mathrm{PSR} = \mathbb{P}(\mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\beta^0))$,
- the signed True Positive Rate (sTPR), the proportion of nonzero coe□cients whose sign is correctly identified;
- the signed False Discovery Rate (sFDR), the proportion of incorrect signs among all discoveries.

Results with s-oracle hyperparameter tuning

Non-correlated case



(a) PSR

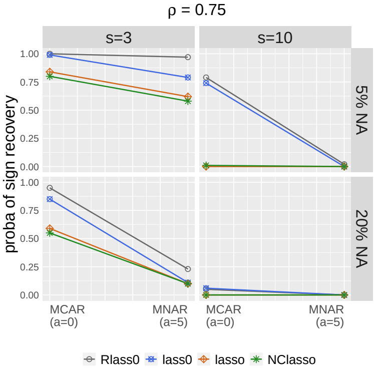


(b) sTPR

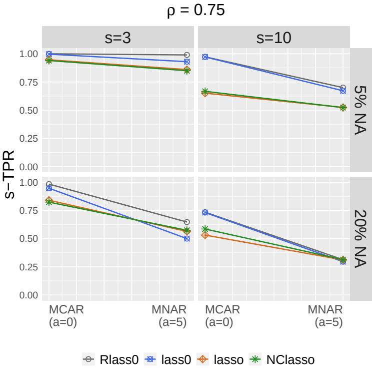
- 5% NA and high sparsity: similar results.
- 20% NA and high sparsity: Robust Lasso-Zero outperforms other methods for MNAR setting.
- lower sparsity: Robust Lasso-Zero and Lasso-Zero generally give the best results.

Results with s-oracle hyperparameter tuning

Correlated case



(a) PSR

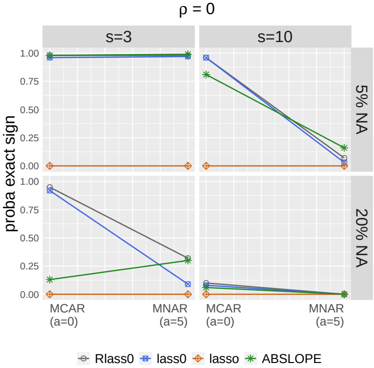


(b) sTPR

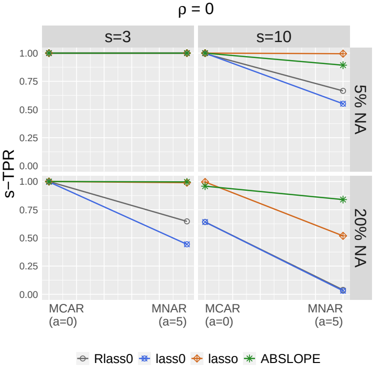
- similar results as in the non-correlated case.
- 5% NA and high sparsity: Robust Lasso-Zero for the MNAR setting behaves very well.

Results with automatic hyperparameter tuning

Non-correlated case



(a) PSR

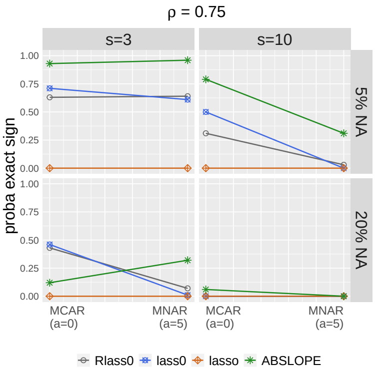


(b) sTPR

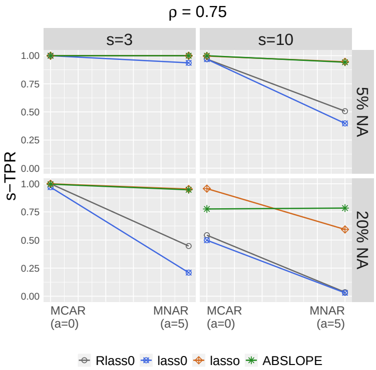
- 5% NA, high sparsity: except the LASSO, good performances.
- 20% NA, high sparsity: Robust Lasso-Zero has the best PSR.
- lower sparsity: except the LASSO, the methods are comparable in terms of PSR.
- ABSLOPE behaves well in term of sTPR.

Results with automatic hyperparameter tuning

Correlated case



(a) PSR

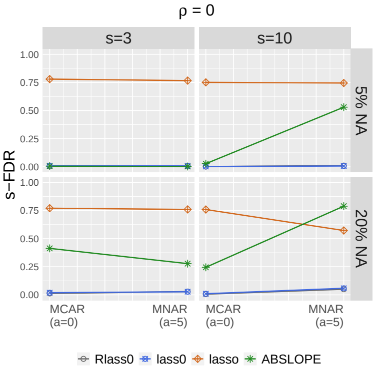


(b) sTPR

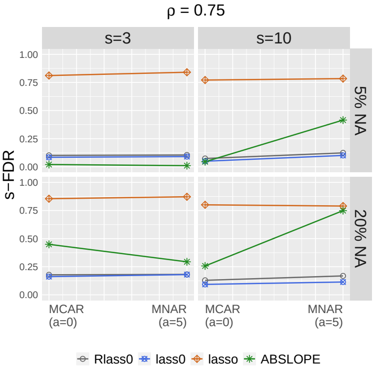
- ABSLOPE generally behaves well.

Results with automatic hyperparameter tuning

sFDR



(a) Non-correlated case



(b) Correlated case

- Robust Lasso-Zero and Lasso-Zero have better performances than ABSLOPE.
- even with low sparsity and 20 % NA, FDR stability in the MCAR and MNAR settings.

Conclusion

- Robust Lasso-Zero: overfit by solving the Justice Pursuit and threshold by handling the overfitting with the use of noise dictionaries.
- Theoretical guarantees for Thresholded Justice Pursuit, a simplified version of the Robust Lasso-Zero.
- Applying Robust Lasso-Zero for dealing with missing data, simple method without parametric assumption.

References I

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

Datta, A. and Zou, H. (2017).
CoCoLasso for high-dimensional error-in-variables regression.
The Annals of Statistics, 45(6):2400–2426.

Descloux, P. and Sardy, S. (2018).
Model selection with lasso-zero: adding straw to the haystack to better find needles.
arXiv:1805.05133 [stat].
arXiv: 1805.05133.

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

References II

Jiang, W., Bogdan, M., Josse, J., Miasojedow, B., Rockova, V., and TraumaBase Group (2019).
Adaptive Bayesian SLOPE – High-dimensional Model Selection with Missing Values.
arXiv e-prints, page arXiv:1909.06631.

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.

Li, X. (2013).
Compressed Sensing and Matrix Completion with Constant Proportion of Corruptions.
Constructive Approximation, 37(1):73–99.

Li, Z., Wu, F., and Wright, J. (2010).
On the systematic measurement matrix for compressed sensing in the presence of gross errors.
In *2010 Data Compression Conference*, pages 356–365. IEEE.

References III

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

Loh, P.-L. and Wainwright, M. J. (2012).
High-Dimensional Regression with Noisy and Missing Data: Provable Guarantees with Nonconvexity.
The Annals of Statistics, 40(3):1637–1664.

Nguyen, N. H. and Tran, T. D. (2013a).
Exact Recoverability From Dense Corrupted Observations via ℓ_1-Minimization.
IEEE Transactions on Information Theory, 59(4):2017–2035.

Nguyen, N. H. and Tran, T. D. (2013b).
Robust Lasso With Missing and Grossly Corrupted Observations.
IEEE Transactions on Information Theory, 59(4):2036–2058.

Rosenbaum, M., Tsybakov, A. B., et al. (2013).
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.

References IV

Saligrama, V. and Zhao, M. (2011).
Thresholded basis pursuit: LP algorithm for order-wise optimal support recovery for sparse and approximately sparse signals from noisy random measurements.
IEEE Transactions on Information Theory, 57(3):1567–1586.

Tardivel, P. and Bogdan, M. (2019).
On the sign recovery by lasso, thresholded lasso and thresholded basis pursuit denoising.

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

Van Buuren, S. (2018).
Flexible imputation of missing data.
Chapman and Hall/CRC.

Wright, J. and Ma, Y. (2010).
Dense Error Correction Via ℓ_1-Minimization.
IEEE Transactions on Information Theory, 56(7):3540–3560.

References V

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009).
Robust Face Recognition via Sparse Representation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):210–227.