Support recovery and sup-norm convergence rates for sparse pivotal estimation

Mathurin Massias*
Université Paris-Saclay
Inria, CEA
Palaiseau, France

Quentin Bertrand*
Université Paris-Saclay
Inria, CEA
Palaiseau, France

Alexandre Gramfort
Université Paris-Saclay
Inria, CEA
Palaiseau, France

Joseph Salmon
IMAG
Univ. Montpellier, CNRS
Montpellier, France

Abstract

In high dimensional sparse regression, pivotal estimators are estimators for which the optimal regularization parameter is independent of the noise level. The canonical pivotal estimator is the square-root Lasso, formulated along with its derivatives as a "non-smooth + non-smooth" optimization problem. Modern techniques to solve these include smoothing the datafitting term, to benefit from fast efficient proximal algorithms. In this work we show minimax sup-norm convergence rates for non smoothed and smoothed, single task and multitask square-root Lasso-type estimators. Thanks to our theoretical analysis, we provide some guidelines on how to set the smoothing hyperparameter, and illustrate on synthetic data the interest of such guidelines.

1 Introduction

Since the mid 1990's and the development on the Lasso (Tibshirani, 1996), a vast literature has been devoted to sparse regularization for high dimensional regression. Statistical analysis of the Lasso showed that it achieves optimal rates (up to log factor, Bickel et al. 2009); see also Bühlmann and van de Geer (2011) for an extensive review. Yet, this estimator requires a specific calibration to achieve such an appealing rate: the regularization parameter must be proportional to the noise level. This quantity is generally unknown to the practitioner, hence the development of methods which are adaptive *w.r.t.* the noise level. An interesting candidate with such a property is the square-root Lasso

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. *denotes equal contribution.

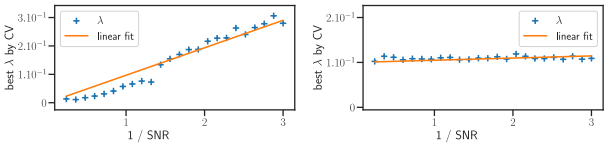

Figure 1: Lasso (left) and square-root Lasso (right) optimal regularization parameters λ determined by cross validation on prediction error (blue), as a function of the noise level on simulated values of y. As indicated by theory, the Lasso's optimal λ grows linearly with the noise level, while it remains constant for the square-root Lasso.

($\sqrt{\text{Lasso}}$, Belloni et al. 2011) defined for an observation vector $y \in \mathbb{R}^n$, a design matrix $X \in \mathbb{R}^{n \times p}$ and a regularization parameter λ by

$$\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{\sqrt{n}} \|y - X\beta\|_2 + \lambda \|\beta\|_1 \ . \tag{1}$$

It has been shown to be *pivotal* with respect to the noise level by Belloni et al. (2011): the optimal regularization parameter of their analysis does not depend on the true noise level. This feature is also encountered in practice as illustrated by Figure 1 (see details on the framework in Section 4.1).

Despite this theoretical benefit, solving the square-root Lasso requires tackling a "non-smooth + non-smooth" optimization problem. To do so, one can resort to conic programming (Belloni et al., 2011) or primal-dual algorithms (Chambolle and Pock, 2011) for which practical convergence may rely on hard-to-tune hyper-parameters. Another approach is to use variational formulations of norms, *e.g.,* expressing the absolute value as $|x| = \min_{\sigma > 0} \frac{x^2}{2\sigma} + \frac{\sigma}{2}$ (Bach et al. 2012, Sec. 5.1, Micchelli et al. 2010). This leads to *concomitant estimation* (Huber and Dutter, 1974), that is, optimization problems over the regression parameters and an additional variable. In sparse regression, the seminal concomitant approach is the con-

comitant Lasso (Owen, 2007):

$$\underset{\beta\in\mathbb{R}^p,\sigma>0}{\arg\min}\ \frac{1}{2n\sigma}\|y-X\beta\|_2^2+\frac{\sigma}{2}+\lambda\|\beta\|_1\ , \qquad (2)$$

which yields the same estimate $\hat{\beta}$ as Problem (1) whenever $y-X\hat{\beta}\neq 0$. Problem (2) is more amenable: it is jointly convex, and the datafitting term is differentiable. Nevertheless, the datafitting term is still not smooth, as σ can approach 0 arbitrarily: proximal solvers cannot be applied safely. A solution is to introduce a constraint $\sigma\geq\underline{\sigma}$ (Ndiaye et al., 2017), which amounts to *smoothing* (Nesterov, 2005; Beck and Teboulle, 2012) the square-root Lasso, *i.e.,* replacing its non-smooth datafit by a smooth approximation (see details in Section 1.4).

There exist a straightforward way to generalize the square-root Lasso to the multitask setting (observations $Y\in\mathbb{R}^{n\times q}$): the multitask square-root Lasso,

$$\underset{\mathrm{B}\in\mathbb{R}^{p\times q}}{\arg\min}\ \frac{1}{\sqrt{nq}}\|Y-X\mathrm{B}\|_F+\lambda\|\mathrm{B}\|_{2,1}\ , \qquad (3)$$

where $\|\mathrm{B}\|_{2,1}$ is the ℓ_1 norm of the ℓ_2 norms of the rows. Another extension of the square-root Lasso to the multitask case is the multivariate square-root Lasso[1] (van de Geer, 2016, Sec. 3.8):

$$\underset{\mathrm{B}\in\mathbb{R}^{p\times q}}{\arg\min}\ \frac{1}{\sqrt{nq(n\wedge q)}}\|Y-X\mathrm{B}\|_*+\lambda\|\mathrm{B}\|_{2,1}\ . \qquad (4)$$

It is also shown by van de Geer (2016) that when $Y-X\hat{\mathrm{B}}$ is full rank, Problem (4) also admits a concomitant formulation, this time with an additional matrix variable:

$$\underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ S\succ 0}}{\arg\min}\ \frac{1}{2nq}\|Y-X\mathrm{B}\|_{S^{-1}}^2+\frac{1}{2n}\operatorname{Tr}(S)+\lambda\|\mathrm{B}\|_{2,1}\ . \qquad (5)$$

In the analysis of the square-root Lasso (1), the non-differentiability at 0 can be avoided by excluding the corner case where the residuals $y-X\hat{\beta}$ vanish. However, analysis of the multivariate square-root Lasso through its concomitant formulation (5) has a clear weakness: it requires excluding rank deficient residuals cases, which is far from being a corner case. As illustrated in Figure 2, the full rank assumption made by van de Geer and Stucky (2016, Lemma 1) or Molstad (2019, Rem. 1) is not realistic, even for $q\geq n$ and high values of λ (see Section 4 for the setting's details). Motivated by numerical applications, Massias et al. (2018) introduced a lower bound on the smallest eigenvalue of S ($S\succeq\underline{\sigma}\,\mathrm{Id}_n$) in Problem (5) to circumvent this issue. As observed by Bertrand et al. (2019, Sec 3.1), this amounts to smoothing the nuclear norm.

[1] modified here with a row-sparse penalty instead of ℓ_1

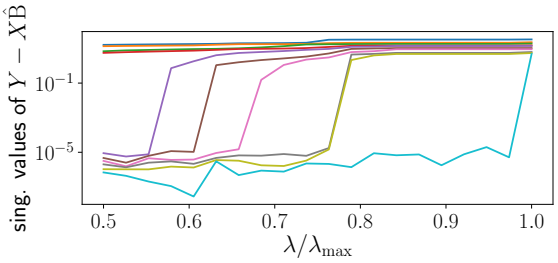

Figure 2: Singular values of the residuals $Y-X\hat{\mathrm{B}}$ of the multivariate square-root Lasso ($n=10, q=20, p=30$), as a function of λ. The observation matrix Y is full rank, but the residuals are rank deficient even for high values of the regularization parameter, invalidating the classical assumptions needed for statistical analysis.

Our goal is to prove sup-norm convergence rates and support recovery guarantees for the estimators introduced above, and their smoothed counterparts.

Related works The statistical properties of the Lasso have been studied under various frameworks and assumptions. Bickel et al. (2009) showed that with high probability, $\|X(\hat{\beta}-\beta^*)\|_2$ vanishes at the minimax rate (prediction convergence), whereas Lounici (2008) proved the sup-norm convergence and the support recovery of the Lasso (estimation convergence), *i.e.,* controlled the quantity $\|\hat{\beta}-\beta^*\|_\infty$. The latter result was extended to the multitask case by Lounici et al. (2011).

Since then, other Lasso-type estimators have been proposed and studied, such as the square-root Lasso (Belloni et al., 2011) or the scaled Lasso (Sun and Zhang, 2012). In the multitask case, Liu et al. (2015) introduced the Calibrated Multivariate Regression, and van de Geer and Stucky (2016); Molstad (2019) studied the multivariate square-root Lasso. These estimators have been proved to converge *in prediction*. However, apart from Bunea et al. (2014) for a particular group square-root Lasso, we are not aware of other works showing sup-norm convergence[2] of these estimators.

Within the framework introduced by Lounici (2008), our contributions are the following:

- We prove sup-norm convergence and support recovery of the multitask square-root Lasso and its smoothed version.
- We prove sup-norm convergence and support recovery of the *multivariate square-root Lasso* (van de Geer and Stucky, 2016, Sec. 2.2), and a smoothed version of it.

[2] of particular interest: combined with a large coefficients assumption, it implies support identification

- Theoretical analysis leads to guidelines for the setting of the smoothing parameter $\underline{\sigma}$. In particular, as soon as $\underline{\sigma} \leq \sigma^*/\sqrt{2}$, the "optimal" λ and the sup-norm bounds obtained do not depend on $\underline{\sigma}$.
- We show on synthetic data the support recovery performances are little sensitive to the smoothing parameter $\underline{\sigma}$ as long as $\underline{\sigma} \leq \sigma^*/\sqrt{2}$.

Our contributions with respect to the existing literature are summarized in Table 1.

Notation Columns and rows of matrices are denoted by $A_{:i}$ and $A_{i:}$ respectively. For any $\mathrm{B} \in \mathbb{R}^{p\times q}$ we define $\mathcal{S}(\mathrm{B}) \triangleq \{j \in [p] : ||\mathrm{B}_{j:}||_2 \neq 0\}$ the row-wise support of B. We write $\mathcal{S}_*$ for the row-wise support of the true coefficient matrix $\mathrm{B}^* \in \mathbb{R}^{p\times q}$. For any $\mathrm{B} \in \mathbb{R}^{p\times q}$ and any subset $\mathcal{S}$ of $[p]$ we denote $\mathrm{B}_\mathcal{S}$ the matrix in $\mathbb{R}^{p\times q}$ which has the same values as B on the rows with indices in $\mathcal{S}$ and vanishes on the complement $\mathcal{S}^c$. The estimated regression coefficients are written $\hat{\mathrm{B}}$, their difference with the true parameter B^* is noted $\Delta \triangleq \hat{\mathrm{B}} - \mathrm{B}^*$. The residuals at the optimum are noted $\hat{\mathrm{E}} \triangleq Y - X\hat{\mathrm{B}}$. The infimal convolution between two functions f_1 and f_2 from $\mathbb{R}^d$ to $\mathbb{R}$ is denoted by $f_1 \,\square\, f_2$ and is defined for any x as $\inf\{f_1(x-y) + f_2(y) : y \in \mathbb{R}^d\}$. For $a < b$, $[x]_a^b \triangleq \max(a, \min(x, b))$ is the clipping of x at levels a and b. The Frobenius and nuclear norms are denoted by $\|\cdot\|_F$ and $\|\cdot\|_*$ respectively. For matrices, $\|\cdot\|_{2,1}$ and $\|\cdot\|_{2,\infty}$ are the row wise $\ell_{2,1}$ and $\ell_{2,\infty}$ norms, *i.e.,* respectively the sum and maximum of rows norms. The subdifferential of a function f is denoted ∂f, and its Fenchel conjugate is written f^*, equal at u to $\sup_x \langle u, x\rangle - f(x)$. For a symmetric definite positive matrix S, $\|x\|_S = \sqrt{\mathrm{Tr}\, x^\top S x}$.

Model Consider the multitask[3] linear regression model:

$$Y = X\mathrm{B}^* + \mathrm{E} \ , \tag{6}$$

where $Y \in \mathbb{R}^{n\times q}$, $X \in \mathbb{R}^{n\times p}$ is the deterministic design matrix, $\mathrm{B}^* \in \mathbb{R}^{p\times q}$ are the true regression coefficients and $\mathrm{E} \in \mathbb{R}^{n\times q}$ models a centered noise.

For an estimator $\hat{\mathrm{B}}$ of B^*, we aim at controlling $\|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty}$ with high probability, and showing support recovery guarantees provided the non-zero coefficients are large enough. To prove such results, the following assumptions are classical: Gaussianity and independence of the noise, and *mutual incoherence*.

Assumption 1. The entries of $\mathrm{E}_1, \dots, \mathrm{E}_n$ are i.i.d. $\mathcal{N}(0, {\sigma^*}^2)$ random variables.

[3]Results simplify in the single task case, where $q = 1$, $\mathrm{B} = \beta \in \mathbb{R}^p$, $\|\cdot\|_{2,1} = \|\cdot\|_1$, $\|\cdot\|_{2,\infty} = \|\cdot\|_\infty$. We state these simpler results in Appendix C.

Assumption 2 (Mutual incoherence)**.** The *Gram matrix* $\Psi \triangleq \frac{1}{n} X^\top X$ satisfies

$$\Psi_{jj} = 1 \ , \text{ and } \max_{j' \neq j} |\Psi_{jj'}| \leq \tfrac{1}{7\alpha s}, \forall j \in [p] \ , \tag{7}$$

for some integer $s \geq 1$ and some constant $\alpha > 1$.

Mutual incoherence of the design matrix (Assumption 2) implies the Restricted Eigenvalue Property introduced by Bickel et al. (2009).

Lemma 3 (Restricted Eigenvalue Property, Lounici (2008, Lemma 2))**.** If Assumption 2 is satisfied, then:

$$\min_{\substack{\mathcal{S}\subset[p]\\|\mathcal{S}|\leq s}} \min_{\substack{\Delta\neq 0\\ \|\Delta_{\mathcal{S}^c}\|_{2,1} \leq 3\|\Delta_\mathcal{S}\|_{2,1}}} \frac{1}{\sqrt{n}} \frac{\|X\Delta\|_F}{\|\Delta_\mathcal{S}\|_F} \geq \sqrt{1 - \tfrac{1}{\alpha}} > 0 \ . \tag{8}$$

In particular, with the choice $\Delta \triangleq \hat{\mathrm{B}} - \mathrm{B}^*$, if $\left\|\Delta_{\mathcal{S}_*^c}\right\|_{2,1} \leq 3 \left\|\Delta_{\mathcal{S}_*}\right\|_{2,1}$, the following bound holds:

$$\frac{1}{n} \|X\Delta\|_F^2 \geq \left(1 - \frac{1}{\alpha}\right) \|\Delta_{\mathcal{S}_*}\|_F^2 \ . \tag{9}$$

1.1 Motivation and general proof structure

Structure of all proofs We prove results of the following form for several estimators $\hat{\mathrm{B}}$ (summarized in Table 1): for some parameter λ independent of the noise level σ^*, with high probability,

$$\frac{1}{q} \|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty} \leq C \frac{1}{\sqrt{nq}} \sqrt{\frac{\log p}{q}} \sigma^* \ . \tag{10}$$

Then, assuming a signal strong enough such that

$$\min_{j \in \mathcal{S}^*} \tfrac{1}{q} \|\mathrm{B}^*_{j:}\|_2 > 2C \frac{1}{\sqrt{nq}} \sqrt{\frac{\log p}{q}} \sigma^* \ , \tag{11}$$

on the same event,

$$\hat{\mathcal{S}} \triangleq \{j \in [p] : \tfrac{1}{q} \|\hat{\mathrm{B}}_{j:}\|_2 > C(3 + \eta)\lambda\sigma^*\} \tag{12}$$

matches the true sparsity pattern: $\hat{\mathcal{S}} = \mathcal{S}^*$.

We explain here the general sketch proofs for all the estimators. We assume that Assumption 2 holds and then place ourselves on an event $\mathcal{A}$ such that $\|X^\top Z\|_{2,\infty} \leq \lambda/2$ (for a $Z \in \partial f(\mathrm{E})$, where f is the datafitting term) in order to use Lemma 4 *ii)*, which links the control of $\|\Psi(\hat{\mathrm{B}} - \mathrm{B}^*)\|_{2,\infty}$ to the control of $\|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty}$. To obtain sup-norm convergence it remains for each estimator to:

- control the probability of the event $\mathcal{A}$ with classical concentration inequalities.
- control the quantity $\|\Psi(\hat{\mathrm{B}} - \mathrm{B}^*)\|_{2,\infty}$, with:

- first order optimality conditions, which provide a bound on $\|X^\top Z\|_{2,\infty}$: $\|X^\top \hat{Z}\|_{2,\infty} \leq \lambda$ for a $\hat{Z} \in \partial f(\hat{\mathrm{E}})$,
- the definition of the event $\mathcal{A}$,
- for some estimators, an additional assumption (Assumption 7).

Next, we detail the lemmas used in this strategy.

1.2 Preliminary lemma

We now provide conditions leading to $\|\Delta_{\mathcal{S}_*^c}\|_{2,1} \leq 3\|\Delta_{\mathcal{S}_*}\|_{2,1}$, to be able to apply Lemma 3. In this section we consider estimators of the form

$$\hat{\mathrm{B}} \triangleq \underset{\mathrm{B}\in\mathbb{R}^{p\times q}}{\arg\min}\, f(Y - X\mathrm{B}) + \lambda \|\mathrm{B}\|_{2,1} \quad , \tag{13}$$

for a proper, lower semi-continuous and convex function $f : \mathbb{R}^{n\times q} \to \mathbb{R}$ (see the summary in Table 1).

Fermat's rule for Problem (13) reads:

$$0 \in X^\top \partial f(\hat{\mathrm{E}}) + \lambda \partial \|\cdot\|_{2,1}(\hat{\mathrm{B}}) \quad , \tag{14}$$

Hence, we can find $\hat{Z} \in \partial f(\hat{\mathrm{E}})$ such that

$$\|X^\top \hat{Z}\|_{2,\infty} \leq \lambda \quad . \tag{15}$$

Lemma 4. Consider an estimator based on Problem (13), and assume that there exists $Z \in \partial f(\mathrm{E})$ such that $\|X^\top Z\|_{2,\infty} \leq \lambda/2$. Then:

i) $\left\|\Delta_{\mathcal{S}_*^c}\right\|_{2,1} \leq 3\|\Delta_{\mathcal{S}_*}\|_{2,1}$,

ii) if Ψ and α satisfy Assumption 2,

$$\|\Delta\|_{2,\infty} \leq \left(1 + \frac{16}{7(\alpha - 1)}\right) \|\Psi\Delta\|_{2,\infty} \quad .$$

Proof. For Lemma 4 *i)*, we use the minimality of $\hat{\mathrm{B}}$:

$$f(\hat{\mathrm{E}}) - f(\mathrm{E}) \leq \lambda\|\mathrm{B}^*\|_{2,1} - \lambda\|\hat{\mathrm{B}}\|_{2,1} \quad . \tag{16}$$

We upper bound the right hand side of Equation (16), using $\|\hat{\mathrm{B}}\|_{2,1} = \|\hat{\mathrm{B}}_{\mathcal{S}_*}\|_{2,1} + \|\hat{\mathrm{B}}_{\mathcal{S}_*^c}\|_{2,1}$, $\mathrm{B}^*_{\mathcal{S}_*^c} = 0$ and with the triangle inequality:

$$\begin{aligned}
\|\mathrm{B}^*\|_{2,1} - \|\hat{\mathrm{B}}\|_{2,1} &= \|\mathrm{B}^*_{\mathcal{S}_*}\|_{2,1} - \|\hat{\mathrm{B}}_{\mathcal{S}_*}\|_{2,1} - \|\hat{\mathrm{B}}_{\mathcal{S}_*^c}\|_{2,1} \\
&= \|\mathrm{B}^*_{\mathcal{S}_*}\|_{2,1} - \|\hat{\mathrm{B}}_{\mathcal{S}_*}\|_{2,1} - \|\Delta_{\mathcal{S}_*^c}\|_{2,1} \\
&\leq \|(\mathrm{B}^* - \hat{\mathrm{B}})_{\mathcal{S}_*}\|_{2,1} - \|\Delta_{\mathcal{S}_*^c}\|_{2,1} \\
&\leq \|\Delta_{\mathcal{S}_*}\|_{2,1} - \|\Delta_{\mathcal{S}_*^c}\|_{2,1} \quad . \qquad (17)
\end{aligned}$$

We now aim at finding a lower bound of the left hand side of Equation (16). By convexity of f, $\partial f(\mathrm{E}) \neq \varnothing$. Picking $Z \in \partial f(\mathrm{E})$ such that $\|X^\top Z\|_{2,\infty} \leq \frac{\lambda}{2}$ yields:

$$\begin{aligned}
f(Y - X\hat{\mathrm{B}}) - f(Y - X\mathrm{B}^*) &\geq -\left\langle Z, X(\hat{\mathrm{B}} - \mathrm{B}^*)\right\rangle \\
&\geq -\left\langle X^\top Z, \Delta\right\rangle \\
&\geq -\left\|X^\top Z\right\|_{2,\infty} \|\Delta\|_{2,1} \\
&\geq -\frac{1}{2}\lambda \|\Delta\|_{2,1} \quad .
\end{aligned}$$

Combining Equations (18), (16) and (17) leads to:

$$\begin{aligned}
-\frac{1}{2}\|\Delta\|_{2,1} &\leq \|\Delta_{\mathcal{S}_*}\|_{2,1} - \|\Delta_{\mathcal{S}_*^c}\|_{2,1} \\
\left\|\Delta_{\mathcal{S}_*^c}\right\|_{2,1} &\leq 3\|\Delta_{\mathcal{S}_*}\|_{2,1} \quad . \qquad (18)
\end{aligned}$$

Proof of Lemma 4 *ii)* is a direct application of Lemmas 4 *i)* and A.1 *iii)*. □

Equipped with these Assumptions and Lemmas, we will show that the considered estimators reach the minimax lower bounds, which we recall in the following.

1.3 Minimax lower bounds

As said in Section 1.1, our goal is to provide convergence rates on the quantity $\|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty}$. To show that our bounds are "optimal" we recall that the considered estimators achieve minimax rate (up to a logarithmic factor). Indeed, under some additional assumptions controlling the conditioning of the design matrix, one can show (Lounici et al., 2011) minimax lower bounds.

Assumption 5. For all $\Delta \in \mathbb{R}^{p\times q}\backslash\{0\}$ such that $|\mathcal{S}(\Delta)| \leq 2|\mathcal{S}^*|$:

$$\underline{\kappa} \leq \frac{\|X\Delta\|_F^2}{n\|\Delta\|_F^2} \leq \bar{\kappa} \quad . \tag{19}$$

Provided Assumptions 1 and 5 hold true, Lounici et al. (2011, Thm. 6.1) proved the following minimax lower bound (with an absolute constant R):

$$\inf_{\hat{\mathrm{B}}} \sup_{\substack{\mathrm{B}^* \text{ s.t.} \\ |\mathcal{S}(\mathrm{B}^*)|\leq s}} \mathbb{E}\left(\tfrac{1}{q}\|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty}\right) \geq \tfrac{R\sigma^*}{\bar{\kappa}\sqrt{n}}\sqrt{1 + \tfrac{\log(ep/s)}{q}} \quad .$$

1.4 Smoothing

Some of the pivotal estimators studied here are obtained via a technique called smoothing. For $L > 0$, a convex function ϕ is L-smooth (*i.e.,* its gradient is L-Lipschitz) if and only if its Fenchel conjugate ϕ^* is $\frac{1}{L}$-strongly convex (Hiriart-Urruty and Lemaréchal, 1993, Thm 4.2.1). Therefore, given a smooth function ω, a principled way to smooth a function f is to add the strongly convex ω^* to f^*, thus creating a strongly

Table 1: Summary of estimators (MT: multitask, MV: multivariate)

Name	$f(\mathrm{E})$	Sup-norm cvg	Pred. cvg
MT $\sqrt{\text{Lasso}}$ (3)	$\frac{1}{\sqrt{nq}}\|\mathrm{E}\|_F$	Bunea et al. (2014)	Bunea et al. (2014)
MT concomitant Lasso	$\min_{\sigma>0} \frac{1}{2nq\sigma}\|\mathrm{E}\|_F^2 + \frac{\sigma}{2}$	us	Li et al. (2016)
MT smooth. conco. Lasso (21)	$\min_{\sigma>\underline{\sigma}} \frac{1}{2nq\sigma}\|\mathrm{E}\|_F^2 + \frac{\sigma}{2}$	us	Li et al. (2016)
MV $\sqrt{\text{Lasso}}$ (4)	$\frac{1}{n}\|\mathrm{E}/\sqrt{q}\|_*$	us	Molstad (2019)
MV conco. $\sqrt{\text{Lasso}}$ (5)	$\min_{S\succ 0} \frac{1}{2nq}\|\mathrm{E}\|_{S^{-1}}^2 + \frac{1}{2n}\operatorname{Tr}(S)$	us	Molstad (2019)
MV SGCL (34)	$\min_{\bar{\sigma}\succeq S\succeq \underline{\sigma}} \frac{1}{2nq}\|\mathrm{E}\|_{S^{-1}}^2 + \frac{1}{2n}\operatorname{Tr}(S)$	us	

convex function, whose Fenchel transform is a smooth approximation of f. Formally, given a smooth convex function ω, the ω-smoothing of f is $(f^* + \omega^*)^*$. By properties of the Fenchel transform, the latter is also equal to $f \,\square\, \omega$ whenever f is convex (Bauschke and Combettes, 2011, Prop. 13.21).

Proposition 6. Let $\omega_{\underline{\sigma}} = \frac{1}{2\underline{\sigma}}\|\cdot\|_F^2 + \frac{\underline{\sigma}}{2}$. The $\omega_{\underline{\sigma}}$-smoothing of the Frobenius norm is equal to:

$$\left(\omega_{\underline{\sigma}} \,\square\, \|\cdot\|_F\right)(Z) = \begin{cases} \|Z\|_F \ , \text{ if } \|Z\|_F \leq \underline{\sigma} \ , \\ \frac{1}{2\underline{\sigma}}\|Z\|_F^2 + \frac{\underline{\sigma}}{2} \ , \text{ if } \|Z\|_F \geq \underline{\sigma} \ . \end{cases}$$
$$= \min_{\sigma \geq \underline{\sigma}} \frac{1}{2\sigma}\|Z\|_F^2 + \frac{\sigma}{2} \ . \tag{20}$$

2 Multitask square-root Lasso

It is clear that the multitask square-root Lasso (Pb. (3)) suffers from the same numerical weaknesses as the square-root Lasso. A more amenable version has been introduced by Bertrand et al. (2019, Prop. 21). The smoothed multitask square-root Lasso is obtained by replacing the non-smooth function $\|\cdot\|_F$ with a smooth approximation, depending on a parameter $\underline{\sigma} > 0$:

$$\operatorname*{arg\,min}_{\mathrm{B}\in\mathbb{R}^{p\times q}} \left(\|\cdot\|_F \,\square\, \left(\tfrac{1}{2\underline{\sigma}}\|\cdot\|^2 + \tfrac{\underline{\sigma}}{2}\right)\right)\left(\tfrac{Y - X\mathrm{B}}{\sqrt{nq}}\right) + \lambda\|\mathrm{B}\|_{2,1} \ . \tag{21}$$

Plugging the expression of the smoothed Frobenius norm (20), the problem formulation becomes:

$$(\hat{\mathrm{B}}, \hat{\sigma}) \in \operatorname*{arg\,min}_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q} \\ \sigma \geq \underline{\sigma}}} \frac{1}{2nq\sigma}\|Y - X\mathrm{B}\|_F^2 + \frac{\sigma}{2} + \lambda\|\mathrm{B}\|_{2,1} \ , \tag{22}$$

where the datafitting term is $(nq\underline{\sigma})^{-1}$-smooth *w.r.t.* B. We show that estimators (3) and (21) reach the minimax lower bound, with a regularization parameter independent of σ^*. For that, another assumption is needed.

Assumption 7 (van de Geer (2016, Lemma 3.1))**.** There exists $\eta > 0$ verifying

$$\lambda\|\mathrm{B}^*\|_{2,1} \leq \eta\sigma^* \ . \tag{23}$$

Proposition 8. Let $\hat{\mathrm{B}}$ denote the multitask square-root Lasso (3) or its smoothed version (21). Let Assumption 1 be satisfied, let α and η satisfy Assumptions 2 and 7. For $C = \left(1 + \frac{16}{7(\alpha-1)}\right)$, $A > \sqrt{2}$ and $\lambda = \frac{2\sqrt{2}}{\sqrt{nq}}\left(1 + A\sqrt{(\log p)/q}\right)$, if $\underline{\sigma} \leq \frac{\sigma^*}{\sqrt{2}}$ then with probability at least $1 - p^{1-A^2/2} - (1+e^2)e^{-nq/24}$,

$$\tfrac{1}{q}\|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty} \leq C(3+\eta)\lambda\sigma^* \ . \tag{24}$$

Moreover provided that

$$\min_{j\in\mathcal{S}^*} \tfrac{1}{q}\|\mathrm{B}^*_{j:}\|_2 > 2C(3+\eta)\lambda\sigma^* \ , \tag{25}$$

then, with the same probability, the estimated support

$$\hat{\mathcal{S}} \triangleq \{j \in [p] : \tfrac{1}{q}\|\hat{\mathrm{B}}_{j:}\|_2 > C(3+\eta)\lambda\sigma^*\} \tag{26}$$

recovers the true sparsity pattern: $\hat{\mathcal{S}} = \mathcal{S}^*$.

Proof. We first bound $\|\Psi\Delta\|_{2,\infty}$. Let $\mathcal{A}_1$ be the event

$$\mathcal{A}_1 \triangleq \left\{\tfrac{\|X^\top\mathrm{E}\|_{2,\infty}}{\sqrt{nq}\|\mathrm{E}\|_F} \leq \tfrac{\lambda}{2}\right\} \cap \left\{\tfrac{\sigma^*}{\sqrt{2}} < \tfrac{\|\mathrm{E}\|_F}{\sqrt{nq}} < 2\sigma^*\right\} \ . \tag{27}$$

By Lemma B.2 *viii)*, $\mathbb{P}(\mathcal{A}_1) \geq 1 - p^{1-A^2/2} - (1 + e^2)e^{-nq/24}$. For both estimators, on $\mathcal{A}_1$ we have:

$$\begin{aligned} n\|\Psi\Delta\|_{2,\infty} &= \|X^\top(\hat{\mathrm{E}} - \mathrm{E})\|_{2,\infty} \\ &\leq \|X^\top\hat{\mathrm{E}}\|_{2,\infty} + \|X^\top\mathrm{E}\|_{2,\infty} \\ &\leq \|X^\top\hat{\mathrm{E}}\|_{2,\infty} + \lambda nq\sigma^* \ , \end{aligned} \tag{28}$$

hence we need to bound $\|X^\top\hat{\mathrm{E}}\|_{2,\infty}$. We do so using optimality conditions, that yield for Problem (3), with $\hat{\mathrm{E}} \neq 0$,

$$\begin{aligned} \|X^\top \tfrac{\hat{\mathrm{E}}}{\|\hat{\mathrm{E}}\|_F}\|_{2,\infty} &\leq \lambda\sqrt{nq} \\ \tfrac{1}{nq}\|X^\top\hat{\mathrm{E}}\|_{2,\infty} &\leq \lambda\tfrac{\|\hat{\mathrm{E}}\|_F}{\sqrt{nq}} \ , \end{aligned} \tag{29}$$

and the last equation is still valid if $\hat{\mathrm{E}} = 0$. For Problem (21), the optimality conditions yield:

$$\begin{cases} \frac{1}{nq}\|X^\top \hat{\mathrm{E}}\|_{2,\infty} \leq \lambda \frac{\|\hat{\mathrm{E}}\|_F}{\sqrt{nq}} \ , & \text{if } \frac{\|\hat{\mathrm{E}}\|_F}{\sqrt{nq}} \geq \underline{\sigma} \ , \\ \frac{1}{nq}\|X^\top \hat{\mathrm{E}}\|_{2,\infty} \leq \lambda \underline{\sigma} \ , & \text{otherwise} \ . \end{cases} \tag{30}$$

Therefore,

$$\frac{1}{nq}\|X^\top \hat{\mathrm{E}}\|_{2,\infty} \leq \lambda \max\left(\frac{\|\hat{\mathrm{E}}\|_F}{\sqrt{nq}}, \underline{\sigma}\right) \ . \tag{31}$$

It now remains to bound $\|\hat{\mathrm{E}}\|_F$ for both estimators, which is done with Assumption 7: for Problem (3), by minimality of the estimator,

$$\begin{aligned} \frac{1}{\sqrt{nq}}\|\hat{\mathrm{E}}\|_F + \lambda\|\hat{\mathrm{B}}\|_{2,1} &\leq \frac{1}{\sqrt{nq}}\|\mathrm{E}\|_F + \lambda\|\mathrm{B}^*\|_{2,1} \\ \frac{1}{\sqrt{nq}}\|\hat{\mathrm{E}}\|_F &\leq \frac{1}{\sqrt{nq}}\|\mathrm{E}\|_F + \lambda\|\mathrm{B}^*\|_{2,1} \\ &\leq 2\sigma^* + (1+\eta)\sigma^* \\ &\leq (3+\eta)\sigma^* \ , \end{aligned} \tag{32}$$

and we can obtain the same bound in the case of Problem (21) (see Lemma A.2). Combining Equations (28), (29), (31) and (32) we have in both cases:

$$\frac{1}{q}\|\Psi\Delta\|_{2,\infty} \leq (3+\eta)\lambda\sigma^* \ . \tag{33}$$

Finally we exhibit an element of $\partial f(\mathrm{E})$ to apply Lemma 4 *ii)*. Recall that $f = \frac{1}{\sqrt{nq}}\|\cdot\|_F$ for Problem (3), and $f = \|\cdot\|_F \square \left(\frac{1}{2\underline{\sigma}}\|\cdot\|_F^2 + \frac{\underline{\sigma}}{2}\right)(\frac{\cdot}{\sqrt{nq}})$ for Problem (21). On $\mathcal{A}_1$, $\partial f(\mathrm{E})$ is a singleton for both estimators, whose element is $\mathrm{E}/(\|\mathrm{E}\|_F\sqrt{nq})$.

Additionally, on $\mathcal{A}_1$ the inequality $\frac{1}{\sqrt{nq}}\frac{\|X^\top \mathrm{E}\|_{2,\infty}}{\|\mathrm{E}\|_F} \leq \frac{\lambda}{2}$ holds, meaning we can apply Lemma 4 *ii)* with $Z = \mathrm{E}/(\|\mathrm{E}\|_F\sqrt{nq})$. This proves the bound on $\|\Delta\|_{2,\infty}$. Then, the support recovery property easily follows from Lounici et al. (2009, Cor. 4.1). □

Single task case For the purpose of generality, we proved convergence results for the multitask versions of the square-root/concomitant Lasso and its smoothed version, but the results are also new in the single task setting. Refined bounds of Proposition 8 in the single-task case are in Appendix C.1.

3 Multivariate square-root Lasso

Here we show that the multivariate square-root Lasso[4] and its smoothed version also reach the minimax rate. Recall that the multivariate square-root Lasso is Problem (4). For the numerical reasons mentioned above, as well as to get rid of the invertibility assumption of $\hat{\mathrm{E}}^\top\hat{\mathrm{E}}$, we consider the smoothed estimator of Massias et al. (2018):

$$\underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q} \\ \bar{\sigma}\,\mathrm{Id}_n \succeq S \succeq \underline{\sigma}\,\mathrm{Id}_n}}{\arg\min} \ \frac{1}{2nq}\|Y - X\mathrm{B}\|_{S^{-1}}^2 + \frac{\mathrm{Tr}\,S}{2n} + \lambda\|\mathrm{B}\|_{2,1} \ . \tag{34}$$

The variable introduced by concomitant formulation is now a matrix S, corresponding to the square root of the noise covariance estimate. The multivariate square-root Lasso (4) and its concomitant formulation (5) have the same solution in B provided $\hat{\mathrm{E}}^\top\hat{\mathrm{E}}$ is invertible. In this case, the solution of Problem (5) in S is $\hat{S} = (\frac{1}{q}\hat{\mathrm{E}}\hat{\mathrm{E}}^\top)^{\frac{1}{2}}$.

Problem (34) is actually a small modification of Massias et al. (2018), where we have added the second constraint $S \preceq \bar{\sigma}\,\mathrm{Id}_n$. $\bar{\sigma}$ can for example be set as $\|(\frac{1}{q}YY^\top)^{1/2}\|_2$, as Figure 2 illustrates that this is the order of magnitude of $\|\hat{S}\|_2$. Because of these constraints, the solution in S is different from that of Problem (5). We write a singular value decomposition of $\frac{1}{\sqrt{q}}\hat{\mathrm{E}}$: $UDV^\top$, with $D = \mathrm{diag}(\gamma_i) \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{q\times n}$ such that $U^\top U = V^\top V = \mathrm{Id}_n$. Then the solution in S to Problem (34) is $\hat{S} = U\,\mathrm{diag}\left([\gamma_i]_{\underline{\sigma}}^{\bar{\sigma}}\right)U^\top$ (this result is easy to derive from Massias et al. (2018, Prop. 2)). $\hat{S}$ can be used to bound $\|X^\top\hat{\mathrm{E}}\|_{2,\infty}$:

Lemma 9. (Proof in Lemma A.3) For the concomitant multivariate square-root Lasso (5) and the smoothed concomitant multivariate square-root (34) we have:

$$\|X^\top\hat{\mathrm{E}}\|_{2,\infty} \leq \|\hat{S}\|_2\|X^\top\hat{S}^{-1}\hat{\mathrm{E}}\|_{2,\infty} \ . \tag{35}$$

We can prove the minimax sup-norm convergence of these two estimators, using the following assumptions.

Assumption 10. For the multivariate square-root Lasso, $\hat{\mathrm{E}}^\top\hat{\mathrm{E}}$ is invertible, and there exists η such that $\|(\frac{1}{q}\hat{\mathrm{E}}^\top\hat{\mathrm{E}})^{\frac{1}{2}}\|_2 \leq (2+\eta)\sigma^*$.

We get rid of this very strong hypothesis for the smoothed version, as the estimated noise covariance is invertible because of the constraint $S \succeq \underline{\sigma}\,\mathrm{Id}_n$, and we can control its operator norm via the constraint $S \preceq \bar{\sigma}\,\mathrm{Id}_n$. We still need an assumption on $\underline{\sigma}$ and $\bar{\sigma}$.

Assumption 11. $\underline{\sigma}$, $\bar{\sigma}$ and η verify: $\underline{\sigma} \leq \frac{\sigma^*}{\sqrt{2}}$ and $\bar{\sigma} = (2+\eta)\sigma^*$ with $\eta \geq 1$.

Proposition 12. For the multivariate square-root Lasso (4) (*resp.* its smoothed version (34)), let Assumption 1 be satisfied, let α satisfy Assumption 2 and let η satisfy Assumption 10 (*resp.* let $\underline{\sigma}, \bar{\sigma}, \eta$ satisfy Assumption 11). Let $C = \left(1 + \frac{16}{7(\alpha-1)}\right)$, $A \geq \sqrt{2}$,

[4] we keep the name of van de Geer (2016), although a better name in our opinion would be the (multitask) trace norm Lasso, but the name is used by Grave et al. (2011) when the nuclear norm is used as a regularizer

and $\lambda = \frac{2\sqrt{2}}{\sqrt{nq}}(1 + A\sqrt{(\log p)/q})$. Then there exists $c \geq 1/64$ such that with probability at least $1 - p^{1-A^2/2} - 2ne^{-cq/n}$,

$$\frac{1}{q}\|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty} \leq C(3+\eta)\lambda\sigma^* \ . \tag{36}$$

Moreover if

$$\min_{j \in \mathcal{S}^*} \frac{1}{q}\|\mathrm{B}^*_{j:}\|_2 > 2C(3+\eta)\lambda\sigma^* \ , \tag{37}$$

then with the same probability:

$$\hat{\mathcal{S}} \triangleq \{j \in [p] : \frac{1}{q}\|\hat{\mathrm{B}}_{j:}\|_2 > C(3+\eta)\lambda\sigma^*\} \tag{38}$$

correctly estimates the true sparsity pattern: $\hat{\mathcal{S}} = \mathcal{S}^*$.

Proof. Let $\mathcal{A}_2$ be the event:

$$\left\{\frac{\|X^\top \mathrm{E}\|_{2,\infty}}{nq} \leq \frac{\lambda\sigma^*}{2\sqrt{2}}\right\} \cap \{2\sigma^* \operatorname{Id}_q \succ (\tfrac{\mathrm{E}^\top \mathrm{E}}{n})^{\frac{1}{2}} \succ \tfrac{\sigma^*}{\sqrt{2}} \operatorname{Id}_q\} \ . \tag{39}$$

By Lemma B.2 *ix)*, $\mathbb{P}(\mathcal{A}_2) \geq 1 - p^{1-A^2/2} - 2ne^{-cq/n}$ ($c \leq 1/64$). When the multivariate square-root Lasso residuals are full rank, the optimality conditions for Problems (4) and (34) read the same, but with differents $\hat{S}$ (introduced above):

$$\|X^\top \hat{S}^{-1}\hat{\mathrm{E}}\|_{2,\infty} \leq \lambda q n \ . \tag{40}$$

With Lemma 9 and Eq. (40) and Assumption 10 for the multivariate square-root Lasso (or Assumption 11 for its smoothed version):

$$\begin{aligned} n\|\Psi\Delta\|_{2,\infty} &= \|X^\top(\mathrm{E} - \hat{\mathrm{E}})\|_{2,\infty} \\ &\leq \|X^\top\hat{\mathrm{E}}\|_{2,\infty} + \|X^\top\mathrm{E}\|_{2,\infty} \\ &\leq \lambda q n\|\hat{S}\|_2 + \|X^\top\mathrm{E}\|_{2,\infty} \\ &\leq \lambda(2+\eta)qn\sigma^* + \|X^\top\mathrm{E}\|_{2,\infty} \ . \end{aligned} \tag{41}$$

Then on the event $\mathcal{A}_2$:

$$\begin{aligned} \frac{1}{q}\|\Psi\Delta\|_{2,\infty} &\leq \lambda(2+\eta)\sigma^* + \tfrac{1}{nq}\|X^\top\mathrm{E}\|_{2,\infty} \\ &\leq (3+\eta)\,\lambda\sigma^* \ . \end{aligned} \tag{42}$$

Finally we exhibit an element of $\partial f(\mathrm{E})$ to apply Lemma 4 *ii)*. Recall that $f = \frac{1}{n\sqrt{q}}\|\cdot\|_*$ for Problem (5), and $f = \min_{\bar{\sigma}\operatorname{Id}_n \succeq S \succeq \underline{\sigma}\operatorname{Id}_n} \frac{1}{2nq}\|\cdot\|^2_{S^{-1}} + \frac{\operatorname{Tr} S}{2n}$ for Problem (34). We also recall that for a full rank matrix $A \in \mathbb{R}^{n\times q}$ (Koltchinskii et al., 2011, Sec. 2):

$$\partial\|A\|_* = \{(AA^\top)^{-1/2}A\} \ . \tag{43}$$

On $\mathcal{A}_2$, $\partial f(\mathrm{E})$ is a singleton for both estimators, whose element is $(\mathrm{E}\mathrm{E}^\top)^{-1/2}\mathrm{E}/(n\sqrt{q})$. Additionally on $\mathcal{A}_2$, using the same proof as in Lemma A.3:

$$\begin{aligned} \tfrac{1}{n\sqrt{q}}\|X^\top(\mathrm{E}\mathrm{E}^\top)^{-1/2}\mathrm{E}\|_{2,\infty} &\leq \tfrac{1}{nq}\|X^\top\mathrm{E}\|_{2,\infty}\|(\tfrac{\mathrm{E}\mathrm{E}^\top}{q})^{-1/2}\|_2 \\ &\leq \frac{\lambda\sigma^*}{2\sqrt{2}} \times \frac{\sqrt{2}}{\sigma^*} \leq \frac{\lambda}{2} \ , \end{aligned} \tag{44}$$

meaning we can apply Lemma 4 *ii)* with $Z = \mathrm{E}(\mathrm{E}^\top\mathrm{E})^{-1/2}/n\sqrt{q}$. This proves the bound on $\|\Delta\|_{2,\infty}$. Then, the support recovery property easily follows from Lounici et al. (2009, Cor. 4.1). □

4 Experiments

We first describe the setting of Figures 1 and 2. Then we show that empirically that results given by Propositions 8 and 12 hold in practice. The signal-to-noise ratio (SNR) is defined as $\frac{\|X\mathrm{B}^*\|_F}{\|Y - X\mathrm{B}\|_F}$.

4.1 Pivotality of the square-root Lasso

In this experiment the matrix X consists of the 10 000 first columns of the *climate* dataset ($n = 864$). We generate β^* with 20 non-zero entries. Random Gaussian noise is added to $X\beta^*$ to create y, with a noise variance σ^* controlling the SNR.

For each SNR value, both for the Lasso and the square-root Lasso, we compute the optimal λ on a grid between $\lambda_{\max}$ (the estimator specific smallest regularization level yielding a 0 solution), using cross validation on prediction error on left out data. For each SNR, results are averaged over 10 realizations of y.

Figure 1 shows that, in accordance with theory, the optimal λ for the Lasso depends linearly on the noise level, while the square-root Lasso achieves pivotality.

4.2 Rank deficiency experiment

For $(n, q, p) = (10, 20, 30)$, we simulate data: entries of X are i.i.d. $\mathcal{N}(0, 1)$, B^* has 5 non zeros rows, and Gaussian noise is to $X\mathrm{B}^*$ added to result in a SNR of 1. We reformulate Problem (4) as a Conic Program, and solve it with the SCS solver of `cvxpy` (O'Donoghue et al., 2016; Diamond and Boyd, 2016) for various values of λ ($\lambda_{\max}$ is the smallest regularization value yielding a null solution). We then plot the singulars values of the residuals at optimum, shown on Figure 2.

Since the problem is reformulated as a Conic Program and solved approximately (precision $\epsilon = 10^{-6}$), the residuals are not exact; however the sudden drop of singular values of $Y - X\hat{\mathrm{B}}$ must be interpreted as the singular value being exactly 0. One can see that even for very high values of λ, the residuals are rank deficient while the matrix Y is not. This is most likely due to the trace penalty on S in the equivalent formulation of Problem (5), encouraging singular values to be 0. Therefore, even on simple toy data, the hypothesis used by van de Geer and Stucky (2016); Molstad (2019) does not hold, justifying the need for smoothing approaches, both from practical and theoretical point of views.

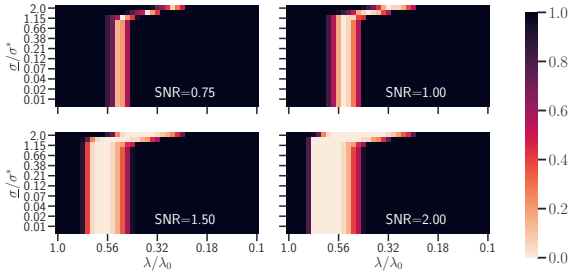

Figure 3: (Synthetic data, $n = 50$, $p = 1000$, $q = 20$) Hard recovery loss for different values of SNR for the multitask SCL.

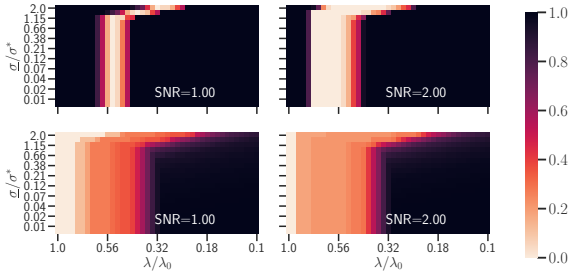

Figure 4: (Synthetic data, $n = 50$, $p = 1000$, $q = 20$) Hard recovery loss (top) and percent of non-zeros coefficients (bottom) for different values of SNR: SNR $= 1$ (left), SNR $= 2$ (right) for the multitask SCL.

4.3 (Multitask) smoothed concomitant Lasso

Here we illustrate, as indicated by theory, that when the smoothing parameter $\underline{\sigma}$ is sufficiently small, the multitask SCL is able to recover the true support (Proposition 8). More precisely, when $\underline{\sigma} \leq \sigma^*/\sqrt{2}$, there exist a λ, independent of $\underline{\sigma}$ and σ^*, such that the multitask SCL recovers the true support with high probability. We use $(n, q, p) = (50, 50, 1000)$. The design X is random with Toeplitz-correlated features with parameter $\rho_X = 0.5$ (correlation between $X_{:i}$ and $X_{:j}$ is $\rho_X^{|i-j|}$), and its columns have unit Euclidean norm. The true coefficient B^* has 5 non-zeros rows whose entries are i.i.d. $\mathcal{N}(0, 1)$.

Comments on Figures 3 and 4 The multitask SCL relies on two hyperparameters: the penalization coefficient λ and the smoothing parameter $\underline{\sigma}$, whose influence we study here. The goal is to show empirically that when $\underline{\sigma} \leq \sigma^*/\sqrt{2}$ the optimal λ does not depend on the smoothing parameter $\underline{\sigma}$. We vary λ and $\underline{\sigma}$ on a grid: for each pair $(\lambda, \underline{\sigma})$ we solve the multitask SCL. For each solution $\hat{\mathrm{B}}^{(\lambda,\underline{\sigma})}$ we then compute a metric, the hard recovery (Figure 3) or the size of the support (Figure 4). The metrics are averaged over 100 realizations of the noise. Figure 3 shows the latter graph for different values of SNR. We can see that when $\underline{\sigma} \leq \sigma^*$, support recovery is achieved for λ independent of $\underline{\sigma}$. As soon as $\underline{\sigma} > \sigma^*$ the optimal λ depends on $\underline{\sigma}$. When $\underline{\sigma}$ reaches a large enough value (*i.e.,* σ^*) then the recovery profile is modified: the optimal λ decreases as $\underline{\sigma}$ grows. This is logical, since as soon as the constraint is saturated, the (multitask) SCL boils down to a multitask Lasso with regularization parameter $\lambda\underline{\sigma}$:

$$\hat{\mathrm{B}} \triangleq \operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times q}} \frac{1}{2nq} \|Y - X\mathrm{B}\|_F^2 + \lambda\underline{\sigma} \|\mathrm{B}\|_{2,1} \quad . \tag{45}$$

Figure 4 shows that with a fixed λ higher values of $\underline{\sigma}$ may lead to smaller support size, see *e.g.,* $\lambda/\lambda_0 = 0.32$.

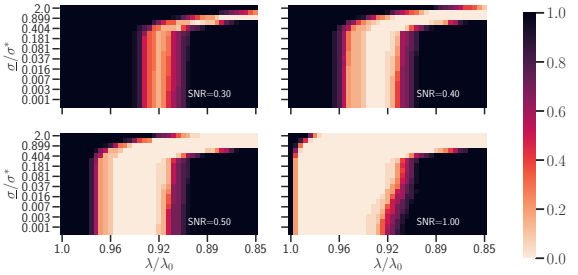

Figure 5: (Synthetic data, $n = 150$, $p = 500$, $q = 100$) Hard recovery loss for different values of SNR for the SGCL.

4.4 Smoothed generalized concomitant Lasso (SGCL)

The experimental setting is the same as before, except here we used $(n, q, p) = (150, 100, 500)$. Figure 5 illustrates Proposition 12. When $\underline{\sigma} \leq \sigma^*$, there exist a λ that does not depend on $\underline{\sigma}$ and such that SGCL finds the true support $\mathcal{S}^*$. However, as before, when $\underline{\sigma} \geq \sqrt{2}\sigma^*$, λ depends on $\underline{\sigma}$.

Conclusion We have proved sup norm convergence rates and support recovery for a family of sparse estimators derived from the square-root Lasso. We showed that they are pivotal too: the optimal regularization parameter does not depend on the noise level. We showed that their smoothed versions retain these properties while being simpler to solve, and requiring more realistic assumptions to be analyzed. These findings were corroborated numerically, in particular for the influence of the smoothing parameter.

Acknowledgments This work was funded by ERC Starting Grant SLAB ERC-StG-676943. We would like to thank Karim Lounici for numerous discussions, suggestions and pointers.

References

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Foundations and Trends in Machine Learning*, 4(1): 1–106, 2012.

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011.

A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM J. Optim.*, 22 (2):557–580, 2012.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Q. Bertrand, M. Massias, A. Gramfort, and J. Salmon. Handling correlated and repeated measurements with the smoothed multivariate square-root lasso. *NeurIPS*, 2019.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

F. Bunea, J. Lederer, and Y. She. The group square-root Lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theory*, 60(2):1313–1325, 2014.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.*, 17(83):1–5, 2016.

C. Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

A. Gittens and J. A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. *arXiv preprint arXiv:1104.4513*, 2011.

E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. In *NeurIPS*, pages 2187–2195, 2011.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. II*, volume 306. Springer-Verlag, Berlin, 1993.

P. J. Huber and R. Dutter. Numerical solution of robust regression problems. In *Compstat 1974 (Proc. Sympos. Computational Statist., Univ. Vienna, Vienna, 1974)*, pages 165–172. Physica Verlag, Vienna, 1974.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39 (5):2302–2329, 2011.

X. Li, J. Haupt, R. Arora, H. Liu, M. Hong, and T. Zhao. On fast convergence of proximal algorithms for sqrt-lasso optimization: Don't worry about its nonsmooth loss function. *arXiv preprint arXiv:1605.07950*, 2016.

H. Liu, L. Wang, and T. Zhao. Calibrated multivariate regression with application to neural semantic basis discovery. *J. Mach. Learn. Res.*, 16:1579–1606, 2015.

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

K. Lounici, M. Pontil, A. Tsybakov, and S. van de Geer. Taking Advantage of Sparsity in Multi-Task Learning. *arXiv preprint arXiv:0903.1468*, 2009.

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011.

M. Massias, O. Fercoq, A. Gramfort, and J. Salmon. Generalized concomitant multi-task lasso for sparse multimodal regression. In *AISTATS*, volume 84, pages 998–1007, 2018.

C. A. Micchelli, J. M. Morales, and M. Pontil. A family of penalty functions for structured sparsity. In *NeurIPS*, pages 1612–1623, 2010.

A. J. Molstad. Insights and algorithms for the multivariate square-root lasso. *arXiv preprint arXiv:1909.05041*, 2019.

E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. *Journal of Physics: Conference Series*, 904(1):012006, 2017.

Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016.

A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58 (1):267–288, 1996.

S. van de Geer. *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, 2016. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour.

S. van de Geer and B. Stucky. χ 2-confidence sets in high-dimensional regression. In *Statistical analysis for high-dimensional data*, pages 279–306. Springer, 2016.