Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression

To cite this article: Eugene Ndiaye *et al* 2017 *J. Phys.: Conf. Ser.* **904** 012006

Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression

Eugene Ndiaye[†], **Olivier Fercoq**[†], **Alexandre Gramfort**[†], **Vincent Leclère**[*], **and Joseph Salmon**[†]

[†]LTCI, Télécom ParisTech, Université Paris-Saclay, 46 rue Barrault, 75013 Paris, France
[*]Université Paris-Est, Cermics (ENPC), 77455 Marne-la-Vallée, France

Abstract.

In high dimensional settings, sparse structures are crucial for efficiency, both in term of memory, computation and performance. It is customary to consider ℓ_1 penalty to enforce sparsity in such scenarios. Sparsity enforcing methods, the Lasso being a canonical example, are popular candidates to address high dimension. For efficiency, they rely on tuning a parameter trading data fitting versus sparsity. For the Lasso theory to hold this tuning parameter should be proportional to the noise level, yet the latter is often unknown in practice. A possible remedy is to jointly optimize over the regression parameter as well as over the noise level. This has been considered under several names in the literature: Scaled-Lasso, Square-root Lasso, Concomitant Lasso estimation for instance, and could be of interest for uncertainty quantification. In this work, after illustrating numerical difficulties for the Concomitant Lasso formulation, we propose a modification we coined Smoothed Concomitant Lasso, aimed at increasing numerical stability. We propose an efficient and accurate solver leading to a computational cost no more expensive than the one for the Lasso. We leverage on standard ingredients behind the success of fast Lasso solvers: a coordinate descent algorithm, combined with safe screening rules to achieve speed efficiency, by eliminating early irrelevant features.

1. Introduction

In the context of high dimensional regression where the number of features is greater than the number of observations, standard least squares need some regularization to both avoid overfitting and ease the interpretation of discriminant features. Among the least squares with sparsity inducing regularization, the Lasso [27], using the ℓ_1 norm as a regularizer, is the most standard one. It hinges on a regularization parameter governing the trade-off between data fitting and sparsity of the estimator, and requires careful tuning. Though this estimator is well understood theoretically, the choice of the tuning parameter remains an open and critical question in practice as well as in theory. For the Lasso, statistical guarantees [5] rely on choosing the tuning parameter proportional to the noise level, a quantity that is usually unknown. Besides, the noise level is of practical interest since it is required in the computation of model selection criterions such as AIC, BIC, SURE or in the construction of confidence sets. A convenient way to estimate both the regression coefficient and the noise level is to perform a joint estimation.

The most famous road for this joint estimation was inspired by the robust theory developed by Huber [14], particularly in the context of location-scale estimation. Indeed, Owen [20] extended it to handle sparsity inducing penalty, leading to a jointly convex optimization formulation. Since then, his estimator has appeared under various name, and we coined it the Concomitant Lasso.

Later, the same formulation was mentioned in [1], in a response to [23], and was thoroughly analyzed in [25], under the name Scaled-Lasso. Similar results were independently obtained in [4] for the same estimator, though with a different formulation. While investigating pivotal quantities, Belloni *et al* proposed to solve the following convex program: modify the standard Lasso by removing the square in the data fitting term. Thus, they termed their estimator the Square-root Lasso (see also [6]). A second approach leading to this formulation was also proposed by [28] to account for noise in the design matrix in an adversarial scenario. Interestingly their robust construction led exactly to the Square-root Lasso formulation.

Under standard design assumption (see [5]), it is proved that the Scaled/Square-root Lasso reaches optimal rates for sparse regression, with the additional benefit that the regularization parameter is independent of the noise level [4, 25]. Moreover, a practical study [22] has shown that the Concomitant Lasso estimator, or its debiased version (see for instance [3, 15] for a discussion on least-squares refitting), is particularly well suited for estimating the noise level.

Among the solutions to compute the Concomitant Lasso, two roads have been explored. Considering the Scaled-Lasso formulation, Sun and Zhang [24, 25] have proposed an iterative procedure that alternates Lasso and noise estimation steps, the later leading to rescaling the tuning parameter iteratively. On the other hand, considering the Square-root Lasso formulation, Belloni *et al* [4] have leaned on second order cone programming solvers, *e.g.,* TFOCS [2].

Despite the appealing properties listed above, among which the superiority of the theoretical results is the most striking, no consensus for an efficient solver has yet emerged. Our contribution aims at providing a more numerically stable formulation, called the Smoothed Concomitant Lasso. This variant allows to obtain a fast solver based on coordinate descent. Then, we propose dedicated safe rules, as introduced in [9, 12] for the Lasso. We show similar accelerations for the Smoothed Concomitant Lasso, both on real and simulated data. Overall, our method presents the same computational cost as for the Lasso, but enjoys the nice features mentioned earlier in terms of statistical properties.

2. Concomitant estimator

Next we present our estimator and some important properties. Proofs are in [17, Appendix].

Notation For any integer $d \in \mathbb{N}$, we denote by $[d]$ the set $\{1, \dots, d\}$. Our observation vector is $y \in \mathbb{R}^n$ (assumed to be nonzero) and the design matrix $X = [X_1, \dots, X_p] \in \mathbb{R}^{n \times p}$ has p explanatory variables or features. The Euclidean norm is written $\|\cdot\|$, the ℓ_1 norm $\|\cdot\|_1$, the ℓ_∞ norm $\|\cdot\|_\infty$, and the matrix transposition of a matrix Q is denoted by $Q^\top$. We note $\mathcal{B}_\infty$ the unit ball with the ℓ_∞ norm. For real numbers a and b, $a \vee b$ stands for the maximum of a and b. We denote $\mathcal{S}_\tau$ the soft-thresholding operator at level $\tau > 0$, $\mathcal{S}_\tau(x) = \operatorname{sign}(x)(|x| - \tau)_+$. For a closed convex set $\mathcal{C}$, we write $\Pi_\mathcal{C}$ the projection. The sub-gradient of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at x is defined as $\partial f(x) = \{z \in \mathbb{R}^d : \forall y \in \mathbb{R}^d, f(x) - f(y) \geq z^\top (x - y)\}$. We denote by ι_C the indicator function of a set C defined as $\iota_C(x) = 0$ if $x \in C$ and $\iota_C(x) = \infty$ if $x \notin C$. We recall that the sub-differential $\partial\|\cdot\|_1$ of the ℓ_1 norm is the set-valued function $\operatorname{sign}(\cdot)$, defined element-wise for all $j \in [d]$ by $\operatorname{sign}(x_j) = 1$ if $x_j > 0$, by $\operatorname{sign}(x_j) = -1$ if $x_j < 0$ and by $\operatorname{sign}(x_j) = [-1, 1]$ if $x_j = 0$. For a set $S \subset [p]$, we denote by $P_{X,S} = X_S \left(X_S^\top X_S\right)^+ X_S^\top$ the projection operator onto $\operatorname{Span}\{X_j : j \in S\}$, where A^+ represents the Moore-Penrose pseudo-inverse. We note tr(X) the trace of matrix X and $\widehat{\Sigma} = X^\top X / n$ the normalized Gram matrix of X.

2.1. Concomitant Lasso

Let us first introduce the Concomitant Lasso estimator, following the formulation proposed in [20, 25], and present some properties obtained due to convexity and duality.

Definition 1. For $\lambda > 0$, the Concomitant Lasso estimator $\hat{\beta}^{(\lambda)}$ is defined as a solution of the primal optimization problem

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p, \sigma > 0} P_\lambda(\beta, \sigma) := \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 . \tag{1}$$

Theorem 1. *Denoting* $\Delta_{X,\lambda} = \left\{\theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leq 1, \lambda\sqrt{n}\|\theta\| \leq 1\right\}$, *the dual formulation of the Concomitant Lasso reads*

$$\hat{\theta}^{(\lambda)} \in \mathop{\arg\max}_{\theta \in \Delta_{X,\lambda}} D_\lambda(\theta) := \langle y, \lambda\theta \rangle. \tag{2}$$

For an optimal primal vector $\hat{\beta}^{(\lambda)}$, $\hat{\sigma}^{(\lambda)} = \|y - X\hat{\beta}^{(\lambda)}\|/\sqrt{n}$. *Moreover, Fermat's rule reads as the link-equation between primal and dual solutions* $y = n\lambda\hat{\sigma}^{(\lambda)}\hat{\theta}^{(\lambda)} + X\hat{\beta}^{(\lambda)}$ *and the sub-differential inclusion* $X^\top(y - X\hat{\beta}^{(\lambda)}) \in n\lambda\hat{\sigma}^{(\lambda)}\partial \left\|\cdot\right\|_1 (\hat{\beta}^{(\lambda)})$.

As defined in (1), the Concomitant Lasso estimator is ill-defined: the set over which we optimize is not closed and no solution may exist. We circumvent this difficulty by considering instead the Fenchel biconjugate of the objective function (for more details, see [17, Appendix C]). The actual objective function accepts $\sigma \geq 0$ as soon as $y = X\beta$. In the rest of the paper, we will write (1) instead of the minimization of the biconjugate as a slight abuse of notation.

A principled way to estimators regularization parameters of Lasso-type programs is to use cross-validation over a fixed set of parameters. Usually, a geometrical grid $\lambda_t = \lambda_{\max}^{\mathrm{L}} 10^{-\delta(t-1)/(T-1)}, t \in [T]$ is used. `scikit-learn` [10] and `glmnet` [13] set $\delta = 3$.

2.2. Critical parameters for the Concomitant Lasso

As for the Lasso, the null vector is optimal for the Concomitant Lasso problem as soon as the regularization parameter becomes too large, as detailed in the next proposition.

Proposition 1. *We have* $\hat{\beta}^{(\lambda)} = 0$ *for all* $\lambda \geq \lambda_{\max} := \|X^\top y\|_\infty/(\|y\|\sqrt{n})$.

However, for the Concomitant Lasso, there is another extreme. Indeed, there exists a critical parameter $\lambda_{\min}$ such that the Concomitant Lasso is equivalent to the Basis Pursuit for all $\lambda \leq \lambda_{\min}$ and gives an estimate $\hat{\sigma}^{(\lambda)} = 0$. The Basis Pursuit and its dual are given by $\hat{\beta}^{\mathrm{BP}} \in \arg\min_{\beta \in \mathbb{R}^p : y = X\beta} \|\beta\|_1$ and $\hat{\theta}^{\mathrm{BP}} \in \arg\max_{\theta \in \mathbb{R}^n : \left\|X^\top\theta\right\|_\infty \leq 1} \langle y, \theta \rangle$.

Proposition 2. *For any* $\hat{\theta}^{\mathrm{BP}} \in \arg\max_{\theta \in \mathbb{R}^n : \|X^\top\theta\|_\infty \leq 1} \langle y, \theta \rangle$ *and any* $\lambda \leq \lambda_{\min} := 1/(\|\hat{\theta}^{\mathrm{BP}}\|\sqrt{n})$, $(\hat{\beta}^{\mathrm{BP}}, 0)$ *is optimal for* P_λ *and* $\hat{\theta}^{\mathrm{BP}}$ *is optimal for* D_λ.

2.3. Smoothed Concomitant Lasso

We can guarantee the existence of minimizers to the Concomitant Lasso (see [17, Appendix C]), even if $\hat{\sigma}^{(\lambda)} = 0$, but the problem becomes more and more ill-conditioned. The previous proposition shows that for too small λ's, a Basis Pursuit solution will always be found, though numerically this might be challenging to get. Indeed, when λ approaches $\lambda_{\min}$, a coordinate descent algorithm (similar to the one described in Algorithm 1) starts to become significantly slower. To avoid this issue, we propose a slight modification of the objective function by adding a constraint on σ, which corresponds to a noise level limit σ_0. We refer to this method as the Smoothed Concomitant Lasso following the terminology introduced by Nesterov in [19].

Definition 2. For $\lambda > 0$ and $\sigma_0 > 0$, the Smoothed Concomitant Lasso estimator $\hat{\beta}^{(\lambda,\sigma_0)}$ and its associated noise level estimate $\hat{\sigma}^{(\lambda,\sigma_0)}$ are defined as solutions of the primal optimization problem

$$\mathop{\arg\min}_{\beta \in \mathbb{R}^p, \sigma \in \mathbb{R}} P_{\lambda,\sigma_0}(\beta, \sigma) := \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 + \iota_{[\sigma_0, +\infty[}(\sigma). \tag{3}$$

Remark 1. Concurrently to our work, [16] proposed a similar smoothing strategy on the Square-root Lasso formulation, equivalent to our formulation. They also show in [16, Theorem 3.7] that the Smoothed Concomitant Lasso preserves the statistical consistency of the Concomitant Lasso: it achieves the minimax optimal rate of convergence as soon as $\lambda \geq 24\sqrt{\log(p)/n}$ and $\sigma_0 \leq \sigma/4$.

Theorem 2. *The dual formulation of the Smoothed Concomitant Lasso reads*

$$\hat{\theta}^{(\lambda,\sigma_0)} = \underset{\theta \in \Delta_{X,\lambda}}{\arg\max}\, D_{\lambda,\sigma_0}(\theta) := \langle y, \lambda\theta \rangle + \sigma_0 \left(\frac{1}{2} - \frac{\lambda^2 n}{2} \|\theta\|^2 \right), \tag{4}$$

for $\Delta_{X,\lambda} = \left\{\theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leq 1, \|\theta\| \leq 1/(\lambda\sqrt{n})\right\}$. *Associated to an optimal* $\hat{\beta}^{(\lambda,\sigma_0)}$, *we must have* $\hat{\sigma}^{(\lambda,\sigma_0)} = \sigma_0 \vee (\|y - X\hat{\beta}^{(\lambda,\sigma_0)}\|/\sqrt{n})$. *Also, the link-equation* $y = n\lambda\hat{\sigma}^{(\lambda,\sigma_0)}\hat{\theta}^{(\lambda,\sigma_0)} + X\hat{\beta}^{(\lambda,\sigma_0)}$ *and the sub-differential inclusion* $X^\top(y - X\hat{\beta}^{(\lambda,\sigma_0)}) \in n\lambda\hat{\sigma}^{(\lambda,\sigma_0)}\partial\|\cdot\|_1(\hat{\beta}^{(\lambda,\sigma_0)})$ *hold.*

Remark 2. Since $D_{\lambda,\sigma_0}(\theta)$ is strongly concave and $\Delta_{X,\lambda}$ is convex and closed, $\hat{\theta}^{(\lambda,\sigma_0)}$ is unique.

In practice, the choice of σ_0 can be motivated as follows:

- Suppose we have prior information on the minimal noise level expected in the data. Then we can set σ_0 as this bound. Indeed, if $\hat{\sigma}^{(\lambda,\sigma_0)} > \sigma_0$, then the constraint $\sigma \geq \sigma_0$ is not active and the minimizers of Problem (3) are also minimizers of Problem (1). The Smoothed Concomitant Lasso estimator will only be different from the Concomitant Lasso estimator when the prediction given by the Concomitant Lasso violates the a priori information.
- Without prior information we can consider a given accuracy ε, and set $\sigma_0 = \varepsilon$. Then, the theory of smoothing [19] tells us that any $\varepsilon/2$-solution to Problem (3) is an ε-solution to Problem (1). Thus we obtain the same solutions, but as an additional benefit we have a control on the conditioning of the problem.
- If departing slightly from the Concomitant Lasso estimator is not too big of an issue, one can also use an arbitrary proportion of the initial estimation of the noise variance *i.e.*, $\sigma_0 = \|y\|/\sqrt{n} \times 10^{-\alpha}$. This was our choice in practice, and we have set $\alpha = 2$. Indeed, taking a large enough value for σ_0 leads to less numerical issues.

A similar reasoning to Proposition 1 gives the following critical parameter.

Proposition 3. *We have* $\hat{\beta}^{(\lambda,\sigma_0)} = 0$, *for all* $\lambda \geq \lambda_{\max} := \|X^\top y\|_\infty/(n\,(\sigma_0 \vee (\|y\|/\sqrt{n})))$.

2.4. Duality gap and link with the Lasso

By comparing Fermat's rule in Theorems 1 and 2, one can remark that if $\hat{\beta}^{(\lambda,\sigma_0)}$ is a solution of the Smoothed Concomitant Lasso, then it is also a solution of the Lasso with regularization parameter $\lambda\hat{\sigma}^{(\lambda,\sigma_0)}$. The following proposition estimates the quality (in term of duality gap) of a primal-dual vector in the Lasso path compared to Concomitant Lasso path. Denoting $G^{\mathrm{L}}_{\sigma\lambda}(\beta,\theta)$ the duality gap for the standard lasso, one can easily show that

Proposition 4. $\forall \beta \in \mathbb{R}^p, \theta \in \Delta_{X,\lambda}, \sigma \geq \sigma_0$, $G^{\mathrm{L}}_{\sigma\lambda}(\beta,\theta) \leq \sigma G_{\lambda,\sigma_0}(\beta,\sigma,\theta)$.

Hence, as $\forall\lambda,\ \hat{\sigma}^{(\lambda)} \leq \|y\|/\sqrt{n}$, if the duality gap for the Smoothed Concomitant Lasso is small, so is the duality gap for the Lasso with the corresponding regularization parameter.

3. Safe screening rules

In order to achieve greater computational efficiency, we propose new safe screening rules (using the terminology introduced in [9]) for our problem and we compare their performance. The principle underlying safe screening rules is as follows: one can discard inactive features from the optimization problem, thanks to the sub-differential inclusion in Theorem 2 and to a safe region $\mathcal{R}$ such that $\hat{\theta}^{(\lambda,\sigma_0)} \in \mathcal{R}$. Indeed if $\max_{\theta\in\mathcal{R}} |X_j^\top \theta| < 1$ then $|X_j^\top \hat{\theta}^{(\lambda,\sigma_0)}| < 1$ and thus $\hat{\beta}_j^{(\lambda,\sigma_0)} = 0$.

Algorithm 1: CD4SCL – Coordinate Descent for the Smoothed Concomitant Lasso with Gap Safe screening

```
Input  : $X, y, \varepsilon, K, f^{\mathrm{ce}}(=10), \lambda, \sigma_0, \beta, \sigma$
$\mathcal{A} \leftarrow [p]$
for $k \in [K]$ do
    if $k \mod f^{\mathrm{ce}} = 1$ then
        Compute $\theta$ as in Proposition 6
        if $G_{\lambda,\sigma_0}(\beta,\sigma,\theta) = P_{\lambda_t,\sigma_0}(\beta,\sigma) - D_{\lambda_t,\sigma_0}(\theta) \leq \varepsilon$. then      // Stopping criterion
            break
        Update $\mathcal{A}$ thanks to Proposition 5                                      // Screening test
    for $j \in \mathcal{A}$ do                                                           // Loop over coordinates
        $\beta_j \leftarrow \mathcal{S}_{n\sigma\lambda_t/\|X_j\|^2}\left(\beta_j - X_j^\top(X\beta - y)/\|X_j\|^2\right)$   // Soft-thresholding step
        $\sigma \leftarrow \sigma_0 \vee (\|y - X\beta\|/\sqrt{n}))$                      // Noise estimation step
Output: $\beta$, $\sigma$, $\mathcal{A}$
```

Since the dual objective of the Smoothed Concomitant Lasso is $\lambda^2\sigma_0 n$-strongly concave, we can provide a dynamic and converging SAFE sphere region $\mathcal{R}$, following [18].

Proposition 5 (Gap Safe rule)**.** *For all* $(\beta, \sigma, \theta) \in \mathbb{R}^p \times \mathbb{R}_+ \times \Delta_{X,\lambda}$*, then for* $r = \sqrt{2G_{\lambda,\sigma_0}(\beta,\sigma,\theta)/(\lambda^2\sigma_0 n)}$*, we have* $\hat{\theta}^{(\lambda,\sigma_0)} \in \mathcal{B}(\theta, r)$*. Thus, if* $|X_j^\top\theta| + r\,\|X_j\| < 1$ *then* $\hat{\beta}_j^{(\lambda,\sigma_0)} = 0$.

4. Algorithmic details

4.1. Smoothed Concomitant Lasso algorithm (SC)

We first present the inner loop of our main algorithm, *i.e.,* the implementation of coordinate descent for the Smoothed Concomitant Lasso. In Algorithm 1, we denote by $\mathcal{A}$ the active set, *i.e.,* the set of coordinates that we have not been screened out. For safe screening rules, this set is guaranteed to contain the support of the optimal solution.

The fast solver we used for the Smoothed Concomitant Lasso, rely on the two following key features: coordinate descent and Gap Safe screening rules.

4.2. Coordinate descent

The algorithm we consider to compute the Smoothed Concomitant Lasso is coordinate descent, an efficient way to solve Lasso-type problem (even for multiple values of parameters) [13]. Previous attempts mainly focused on iteratively alternating Lasso steps along with noise level estimation [25], or conic programming [2]. In [16], written concurrently to this work, the authors consider ISTA, a first order method using full gradient information at each iteration.

Here we provide a simple and efficient coordinate descent approach, *cf.* Algorithm 1. Our primal objective P_{λ,σ_0} can be written as the sum of a convex differentiable function $f(\beta, \sigma) = \|y - X\beta\|^2/(2n\sigma) + \sigma/2$ and of a separable function $g(\beta, \sigma) = \lambda\|\beta\|_1 + \iota_{[\sigma_0,+\infty[}(\sigma)$. Moreover, for $\sigma \geq \sigma_0 > 0$, the gradient of f is Lipschitz continuous. Hence, we know that the coordinate descent method converges to a minimizer of our problem [29]. We choose to update the variable σ every other iteration because this can be done at a negligible cost.

Our stopping criterion is based on the duality gap defined by $G_{\lambda,\sigma_0}(\beta, \sigma, \theta) = P_{\lambda,\sigma_0}(\beta, \sigma) - D_{\lambda,\sigma_0}(\theta)$. This requires the computation of a dual feasible point as follows.

Proposition 6. *Let* $(\beta_k)_{k\in\mathbb{N}}$ *be a sequence that converges to* $\hat{\beta}^{(\lambda,\sigma_0)}$*. Then* $(\theta_k)_{k\in\mathbb{N}}$ *built as* $\theta_k = \frac{y - X\beta_k}{\lambda n\sigma_0 \vee \left\|X^\top(y - X\beta_k)\right\|_\infty \vee \lambda\sqrt{n}\|y - X\beta_k\|}$ *converges to* $\hat{\theta}^{(\lambda,\sigma_0)}$*. Hence* $G_{\lambda,\sigma_0}(\beta_k, \sigma_k, \theta_k) \rightarrow_{k\rightarrow+\infty} 0$.

5. Numerical experiments

We compare the estimation performance and computation times of standard deviation estimators which are presently the state-of-the-art in high dimensional settings. We refer to [22] for a recent

$\hat{\sigma}_{OR}$	$\hat{\sigma}_{\mathcal{M}-CV}$	$\hat{\sigma}_{\mathcal{M}-LS}$	$\hat{\sigma}_i$	$\widehat{\sigma}_{D2}$
$\frac{\lVert y-P_{X,S^\star} y\rVert}{\sqrt{n-\lvert S^\star\rvert}}$	$\frac{\lVert y-X\hat{\beta}_{\mathcal{M}}^{\lambda_{cv}}\rVert}{\sqrt{n-\lvert\hat{S}_{\mathcal{M}}^{\lambda_{cv}}\rvert}}$	$\frac{\lVert y-P_{X,\hat{S}_{\mathcal{M}}} y\rVert}{\sqrt{n-\lvert\hat{S}_{\mathcal{M}}\rvert}}$	$\frac{\lVert y^{(i')}-P_{X^{(i')},\hat{S}_i} y^{(i')}\rVert}{\sqrt{n/2-\lvert\hat{S}_i\rvert}}$	$\frac{\left(1+\frac{p\widehat{m}_1^2}{(n+1)\widehat{m}_2}\right)\lVert y\rVert^2}{n} - \frac{\widehat{m}_1\lVert X^\top y\rVert^2}{\sqrt{n(n+1)\widehat{m}_2}}$

Table 1: The estimator $\hat{\beta}_{\mathcal{M}}$ are obtained by a method $\mathcal{M}$ and $\mathcal{M}-LS$ is its least square refitting. We note $S^{\odot} = \{j \in [p],\, \beta_j^{\odot} \neq 0\}$, $D_i = (y^{(i)}, X^{(i)})_{i\in[2]}$ is a split in two parts of the observations, and $\hat{S}_i$ the support selected after a cross-validation on the part D_i. The RCV estimator is $\hat{\sigma}_{RCV} = ((\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2)^{1/2}$, and $\widehat{m}_1 = \operatorname{tr}(\widehat{\Sigma})/p$ and $\widehat{m}_2 = \operatorname{tr}(\widehat{\Sigma}^2)/p - (\operatorname{tr}(\widehat{\Sigma}))^2/(pn)$.

comparison. In our simulations[1] we use the common setup: $y = X\beta^\star + \sigma\varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_n)$ and $X \in \mathbb{R}^{n\times p}$ follows a multivariate normal distribution with covariance $\Sigma = (\rho^{|i-j|})_{i,j\in[p]}$. We define $\beta^\star = \alpha\beta$ where the coordinates of β are drawn from a standard Laplace distribution and we randomly set $s\%$ of them to zero. The scalar α is chosen in order to satisfy a prescribed signal to noise ratio denoted snr: $\alpha = \sqrt{\mathrm{snr} \times \sigma^2/\beta^\top\Sigma\beta}$.

The procedures we have compared are summarized in Table 1. Namely, our reference is the oracle estimator (OR) $\hat{\sigma}_{OR}$, the cross-validated estimator (CV) $\hat{\sigma}_{\mathcal{M}-CV}$ whith a parameter λ_{cv} chosen by 5-fold cross-validation, the least-square refitting estimator (LS) $\hat{\sigma}_{\mathcal{M}-LS}$, the refitted cross-validation (RCV) $\hat{\sigma}_{RCV}$ and $\widehat{\sigma}_{D2}$ the estimator introduced in [8].

We run all the following algorithms over the non-increasing sequence $\lambda_t = \lambda_{\max}10^{-\delta(t-1)/(T-1)}, t \in [T]$ with the default value $\delta = 2, T = 100$. The regularization grid for the joint estimations (Scaled-Lasso, with solver from [25] (SZ), Smoothed Concomitant Lasso (SC), Square-root Lasso [4] (SQRT-Lasso) and the estimator introduced in [23] (SBvG)) begins at $\lambda_{\max}$ given in Proposition 3. We set Smoothed Concomitant Lasso with the default value $\sigma_0 = \lVert y\rVert/\sqrt{n} \times 10^{-2}$. As explained in Section 2.3 this choice improves numerical efficiency at the cost of departing slightly from the Concomitant Lasso estimator in the low noise regime. The grid for the Lasso (L) estimators begins with $\lambda_{\max}^{\mathrm{L}} = \lVert X^\top y\rVert_\infty/n$. The Lasso with the universal parameter $\lambda = \sqrt{2\log(p)/n}$ is denoted (L_U) and SZ refers to Concomitant Lasso with the quantile regularization described in [26] in Fig. 1(c).

5.1. Computational performance

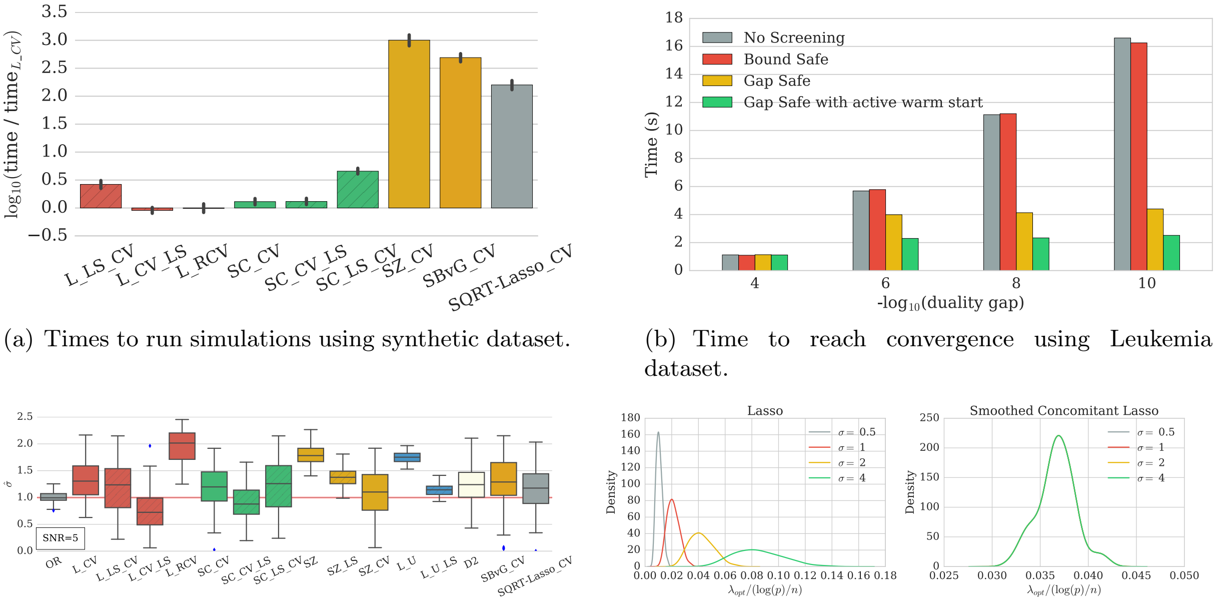
Figure 1(a) presents on the Leukemia dataset the computation times observed for the different CV methods. The Smoothed Concomitant Lasso is based on the coordinate descent algorithm described in this paper and written in Python and Cython to generate low level C code, offering high performance. When a Lasso solver is needed, we have used the one from `scikit-learn`, that is coded similarly. For SZ_CV, computations are quite heavy as one uses the alternating algorithm proposed in [25]. Depending on the regularization parameter (for instance when one approaches $\lambda_{\min}$) the SZ_CV method is quite intractable and the algorithm faces the numerical issues mentioned earlier. The generic solver used for SBvG and SQRT-Lasso, is the `CVXPY` package [7], explaining why these methods are two orders of magnitude slower than a Lasso. In contrast, our solver reaches similar computing time w.r.t. an efficient Lasso solver, with the additional benefit of jointly estimating the coefficients and the standard deviation of the noise.

Figure 1(b) shows the benefit one can obtain thanks to the safe screening rule introduced above. Indeed, the Gap Safe rule greatly benefits from the convergence of the dual vector, leading to smaller and smaller safe sphere as the iterations proceeds [12, 18].

5.2. Performance of standard deviation estimators

As noted earlier in [11], spurious correlations can strongly affect sparse regression and usually lead to large biases. This makes the standard deviation estimation very challenging and affects

[1] Source code available at `https://github.com/EugeneNdiaye/smoothed_concomitant_lasso`



(a) Times to run simulations using synthetic dataset.

(b) Time to reach convergence using Leukemia dataset.

(c) Estimated performance using synthetic dataset.

(d) Estimated distribution of the optimal λ_{opt}.

Figure 1: (a): comparisons of the computational times using different estimation method (time presented relative to the mean time of the Lasso). (b): speed up using screening rules for the Smoothed Concomitant Lasso w.r.t. to duality gap and for $(\lambda_t)_{t\in[100]}$. The dimensions of Leukemia dataset are $(n = 72, p = 7129)$. (c): comparison of quality of different estimators of the noise σ normalized to 1. The synthetic datasets are generated with the settings $(n = 100, p = 500, \rho = 0.6, \text{snr} = 5, s = 0.9, 50 \text{ replications})$. (d): comparisons of the distribution of optimal regularizer λ_{opt} under different levels of noise.

the cross-validation estimator based on the Lasso as they usually underestimate the standard deviation. The phenomenon is amplified when one uses least squares refitting on the cross-validated Lasso, as noticed in [22]. Results are presented as boxplots in Fig. 1(c) (see [17, Appendix B] for additional settings).

We have observed that SC and SZ are efficient in high sparsity settings with low correlations, correcting for the positive bias of the estimator from [23] (SBvG). In [22], it was also argued that a cross-validation estimator based on Lasso is more stable and performs better when the sparsity decreases and when the snr increases. Note that this is not the case when performing cross-validation with the Concomitant Lasso. Here, we show that the latter achieves similar performance as the Lasso. It is worth noting that our method is consistently good over the whole experiments we conducted especially when applying least squares refitting.

Another appealing property of the Smoothed Concomitant Lasso compared to the Lasso is the invariance of the optimal $\lambda_{opt} := \arg\min_{\lambda\in(\lambda_t)_{t\in[T]}} \|X\hat{\beta}^{(\lambda,\sigma_0)} - X\beta^{\star}\|_2$ w.r.t. different levels of noise. We show on Fig. 1(d) a kernel density plot of its distribution on synthetic data with different values of σ. A similar experiment was performed in [16] leading to the same conclusion with an optimal λ chosen by a train/test procedure.

6. Conclusion

We have explored the joint estimation of the coefficients and noise level for ℓ_1 regularized regression. We have corrected some numerical drawbacks of the Concomitant Lasso estimator by proposing a smoother formulation, leading to the Smoothed Concomitant Lasso. A fast algorithm, combining coordinate descent and safe screening rules was investigated achieved the same numerical efficiency than for the Lasso while estimating the noise level. Future research would extend our work to general data-fitting terms [20], and combine sketching techniques [21].

References

[1] A. Antoniadis. Comments on: ℓ_1-penalization for mixture regression models. *TEST*, 19(2):257–258, 2010.
[2] S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3(3):165–218, 2011.
[3] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.
[4] A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.
[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.
[6] S. Chrétien and S. Darses. Sparse recovery with unknown variance: a lasso-type approach. *IEEE Trans. Inf. Theory*, 2011.
[7] S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.*, 2016. To appear.
[8] L. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014.
[9] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 8(4):667–698, 2012.
[10] F. Pedregosa *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.
[11] J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. Roy. Statist. Soc. Ser. B*, 74(1):37–65, 2012.
[12] O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the lasso. In *ICML*, pages 333–342, 2015.
[13] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.
[14] P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., 1981.
[15] J. Lederer. Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions. *arXiv preprint arXiv:1306.0113*, 2013.
[16] X. Li, J. Haupt, R. Arora, H. Liu, M. Hong, and T. Zhao. A first order free lunch for sqrt-lasso. *arXiv preprint arXiv:1605.07950*, 2016.
[17] E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclere, and J. Salmon. Efficient smoothed concomitant Lasso estimation for high dimensional regression. *Arxiv preprint arXiv:1606.02702*, 2016.
[18] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *NIPS*, pages 811–819, 2015.
[19] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.
[20] A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.
[21] V. Pham and L. El Ghaoui. Robust sketching for multiple square-root LASSO problems. In *AISTATS*, 2015.
[22] S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *arXiv preprint arXiv:1311.5274*, 2013.
[23] N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010.
[24] T. Sun and C.-H. Zhang. Comments on: ℓ_1-penalization for mixture regression models. *TEST*, 19(2):270–275, 2010.
[25] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.
[26] T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.*, 14:3385–3418, 2013.
[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
[28] H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. *IEEE Trans. Inf. Theory*, 56(7):3561–3574, 2010.
[29] S. Yun. On the iteration complexity of cyclic coordinate gradient descent methods. *SIAM J. Optim.*, 24(3):1567–1580, 2014.