Square-root lasso

Yian YU
Yixuan LIU
Huiwen ZHENG

May 24, 2021

Outline

Variable selection in high dimensional linear regression models has become a very active area of research in the last decade. In linear models one observes independent response random variables (outcome) $y_i \in \mathbb{R}, 1 \leq i \leq n$, and assumes that each y_i can be written as a linear function of the i -th observation on a p -dimensional predictor vector $x_i =: (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})$ corrupted by noise.

$$y_i = x_i' \beta_0 + \sigma \varepsilon_i, \ (i = 1, \cdots, n).$$

where $\beta^0 \in \mathbb{R}^p$ is the unknown regression vector, Without loss of generality, normalize the regressors, $\frac{1}{n} \sum_{i=1}^n x_{ij}^2 = 1 \ (j = 1, \cdots, p)$.

$\sigma \geq 0$ is the noise level, and for each $1 \leq i \leq n$, the additive term ϵ_i is a mean zero random noise component. Assume the noise are independent and identically distributed with law F_0 that

$$E_{F_0}(\varepsilon_i) = 0, \ E_{F_0}(\varepsilon_i^2) = 1.$$

For high-dimentional sparse linear regression models, the overall number of regessors p is larger than n, but only s, $s \leq n$, are significant for the sparsity of parameter vector β_0. Obviously, the ordinary least squares estimator is not consistent fot estimating β_0 for $p > n$.

Lasso estimator restore consistency under mild conditions by penalizing the sum of absolute parameter values.

$$\widehat{\beta}_{LASSO} = \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_1,$$

where $\widehat{Q}(\beta) = n^{-1}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$. Under knowning the standard deviation σ of the noise, which are normal, $F_0 = N(0,1)$, if uses the penalty level

$$\lambda = \sigma c 2 n^{1/2}\Phi^{-1}(1-\alpha/2p)$$

for some constant $c > 1$, the estimator achieves the near-oracle performance, as

$$\left\|\widehat{\beta}_{LASSO} - \beta_0\right\|_2 \lesssim \sigma\{s\log(2p/\alpha)/n\}^{1/2}$$

with probability at least $1-\alpha$. If p is polynomial in n, the oracle rate is achieved up to a factor of $\sigma(s/n)^{1/2}$. Results are demonstrated in (Bickel et al., 2009).

However, the estimation of σ is non-trival when p is large. The square-root lasso, eliminates the need to know or to pre-estimate σ.
The square-root lasso estimator of β_0 is defined as the solution to the optimization problem

$$\widehat{\beta}_{SRL} = \arg\min_{\beta \in \mathbb{R}^p} \{\widehat{Q}(\beta)\}^{1/2} + \frac{\lambda}{n}\|\beta\|_1. \tag{1}$$

The solution of the parameter estimator under the orthonomal design, the model selection consistency, risk properties etc will be presented in the following.

The penalty level could be chosen as $\lambda = cn^{1/2}\Phi^{-1}(1-\alpha/2p)$ which is independent of σ. By moderate deviation theory, this proposed penalty, the chosen tuning parameter will also be valid asymptotically without normality imposition.

Besides, the square-root lasso estimator matches the near-oracle performance of lasso. There exist several efficient algorithmic methods, such as interior-point and first order methods for solving the parameters. When the outcome y_i is a vector, then the square-root lasso is modified to the multivariate version. The group square-root lasso method is proposed for high dimensional sparse regression models with group structure.

Outline

The key quantity determining the choice of penalty level is the score. As for lasso, the score $S = \nabla \widehat{Q}(\beta_0) = 2\sigma E_n(x\epsilon)$ is non-pivotal for it dependes on σ. Then, in lasso we need to guess conservative upper bounds on σ or use preliminary estimation of σ using a pilot lasso. Under the situation that p is large, particularly when $p \geqslant n$, estimation of σ is non-trival.

However, in the square-root lasso, none of these is needed.

As the score:

$$\tilde{S} = \nabla \widehat{Q}^{1/2}(\beta_0) = \frac{\nabla \widehat{Q}(\beta_0)}{2\left\{\widehat{Q}(\beta_0)\right\}^{1/2}} = \frac{E_n(x\sigma\epsilon)}{\{E_n(\sigma^2\epsilon^2)\}^{1/2}} = \frac{E_n(x\epsilon)}{\{E_n(\epsilon^2)\}^{1/2}}.$$

The score $\tilde{S}$ does not depend on the unknown standard deviation σ or the unknown true parameter value β_0, and therefore is pivotal with respect to (β_0, σ).
Score summarizes the estimation noise, set the penalty level λ/n to overcome it which is motivated by the choice of penalty level for the lasso (Bickel et al., 2009). Choose the smallest λ such that

$$\lambda \geqslant c\Lambda, \quad \Lambda = n\|\tilde{S}\|_\infty, \tag{2}$$

with a high probability, say $1-\alpha$, where Λ is the maximal score scaled by n, and $c > 1$ is a theoretical constant.

Tuning parameter selection

The rule (2) is not practical since we do not observe Λ directly.

1. If know the distribution of errors exactly, e.g., $F_0 = \Phi$, we propose to set λ as c times the $(1-\alpha)$ quantile of Λ given X. This choice is easy to compute by simulation.
2. When we do not know F_0 exactly, but instead know that F_0 is an element of some family $\mathcal{F}$, we can rely on either finite-sample or asymptotic upper bounds on quantiles of Λ given X. Under some mild conditions on $\mathcal{F}$, $\lambda = cn^{1/2}\Phi^{-1}(1-\alpha/2p)$ is a valid asymptotic choice.

In the normal case, $F_0 = \Phi$, λ can be either of:

$$\text{exact: } \lambda = c\Lambda_{F_0}(1-\alpha \mid X),$$

$$\text{asymptotic: } \lambda = c\Lambda(1-\alpha) = cn^{1/2}\Phi^{-1}(1-\alpha/2p),$$

where $\Lambda_{F_0}(1-\alpha \mid X) = (1-\alpha)$ -quantile of $\frac{n\|E_n(x\epsilon)\|_\infty}{\{E_n(\epsilon^2)\}^{1/2}}$, with independent and identically distributed ϵ_i in law F_0, which can be compute by simulation. c is the constant > 1 which is needed to guarantee a regularization event. Besides, For asymptotic: $\Lambda(1-\alpha) \leqslant \{2n\log(2p/\alpha)\}^{1/2}$.

In normal case (With the setting of λ and $F_0 = \Phi$) **Lemma 1:**

- the exact option implements $\lambda \geqslant c\Lambda$ with probability at least $1 - \alpha$;
- assume $\frac{p}{\alpha} \geq 8$, for any $1 < \ell < \{\frac{n}{\log(1/\alpha)}\}^{1/2}$, the asymptotic option implements $\lambda \geqslant c\Lambda$ with probability at least $1 - \alpha\tau$,
  $\tau = \left\{1 + \frac{1}{\log(p/\alpha)}\right\} \frac{\exp\left[2\log(2p/\alpha)\ell\{\log(1/\alpha)/n\}^{1/2}\right]}{1-\ell\{\log(1/\alpha)/n\}^{1/2}} - \alpha^{\ell^2/4-1}$.

  Under **Condition G :**

  $$\log^2(p/\alpha)\log(1/\alpha) = o(n) \text{ and } \frac{p}{\alpha} \to \infty \text{ as } n \to \infty.$$

  We have $\tau = 1 + o(1)$ by setting $\ell \to \infty$ such that
  $\ell = o\left[n^{1/2}/\left\{\log(p/\alpha)\log^{1/2}(1/\alpha)\right\}\right]$ as $n \to \infty$;
- assume that $\frac{p}{\alpha} > 8$ and $n > 4\log(2/\alpha)$. Then
  $\Lambda_\Phi(1 - \alpha \mid X) \leqslant \nu\Lambda(1 - \alpha) \leqslant \nu\{2n\log(2p/\alpha)\}^{1/2}$, $\nu = \frac{\{1+2/\log(2p/\alpha)\}^{1/2}}{1-2\{\log(2/\alpha)/n\}^{1/2}}$
  where under **Condition G**, $\nu = 1 + o(1)$ as $n \to \infty$.

In the non-normal case, λ can be either of:

$$\text{exact: } \lambda = c\Lambda_F(1-\alpha \mid X),$$

$$\text{semi-exact: } \lambda = c \max_{F\in\mathcal{F}} \Lambda_F(1-\alpha \mid X),$$

$$\text{asymptotic: } \lambda = c\Lambda(1-\alpha) = cn^{1/2}\Phi^{-1}(1-\alpha/2p),$$

The exact option is applicable when $F_0 = F$, as for example in the previous normal case. The semi-exact option is applicable when F_0 is a member of some family $\mathcal{F}$, or whenever the family $\mathcal{F}$ gives a more conservative penalty level. We also assume that $\mathcal{F}$ is either finite or, more generally, that the maximum is well defined.

The asymptotic option is applicable when F_0 and design matrix X satisfy the following moment **Condition M** that:
there exit a finite constant $q > 2$ *such that the law* F_0 *is an element of the family* $\mathcal{F}$ *such that* $\sup_n \geqslant 1 \sup_{F \in \mathcal{F}} E_F\left(|\epsilon|^q\right) < \infty$; *the design* X *obeys* $\sup_{n \geqslant 1, 1 \leqslant j \leqslant p} E_n\left(|x_j|^q\right) < \infty$,

and the restriction on the growth of p relative to n, denoted as **Condition R** that:
as $n \to \infty, p \leqslant \alpha n^{\eta(q-2)/2}/2$ *for some constant* $0 < \eta < 1$, *and* $\alpha^{-1} = o\left[n^{\{(q/2-1)\wedge(q/4)\}\vee(q/2-2)}/(\log n)^{q/2}\right]$, *where* $q > 2$ *is defined in moment condition*.

- the exact option $\lambda \geqslant c\Lambda$ with probability at least $1-\alpha$, if $F_0 = F$;
- the semi-exact option implements $\lambda \geqslant c\Lambda$ with probability at least $1-\alpha$, if either $F_0 \in \mathcal{F}$ or $\Lambda_F(1-\alpha \mid X) \geqslant \Lambda_{F_0}(1-\alpha \mid X)$ for some $F \in \mathcal{F}$;
- the asymptotic option implements $\lambda \geqslant c\Lambda$ with probability at least $1-\alpha-o(\alpha)$;
- the magnitude of the penalty level of the exact and semi-exact options in satisfies the inequality$\max_{F\in\mathcal{F}} \Lambda_F(1-\alpha \mid X) \leqslant \Lambda(1-\alpha)\{1+o(1)\} \leqslant \{2n\log(2p/\alpha)\}^{1/2}\{1+o(1)\}, n \to \infty$

All of the asymptotic conclusions reaches in **Lemma 1** about the penalty level in the Gaussian case continue to hold in the non-Gaussian case under more restricted **Condition M** & **R**. However, **Condition M** & **R** is one possible set of sufficient conditions that guarantees the Gaussian like conclusions of **Lemma 2**, which is derived by the moderate deviation theory. The authors provide an alternative condition based on the use of the self-normalized moderate deviation theory of (Jing et al., 2003), which results in much weaker growth condition on p in relation to n, but requires much stronger conditions on the moments of regressors.

Outline

We shall state convergence rates for $\hat{\delta} = \hat{\beta} - \beta_0$ in the Euclidean norm $||\delta||_2 = (\delta'\delta)^{1/2}$ and also in the prediction norm

$$||\delta||_{2,n} = [E_n(x'\delta)^2]^{1/2} = \left\{\delta' E_n(xx')\delta\right\}^{1/2} \tag{3}$$

The choice of penalty level in turn imply another regularization event, namely that $\hat{\delta}$ belongs to the restricted set Δ_c, where

$$\Delta_{\bar{c}} = \left\{\delta \in \mathbb{R}^p : \|\delta_{T^c}\|_1 \leqslant \bar{c}\,\|\delta_T\|_1, \delta \neq 0\right\}, \quad \bar{c} = \frac{c+1}{c-1} \tag{4}$$

Accordingly, we will state the bounds on estimation errors $||\delta||_{2,n}$ and $||\delta||_2$ in terms of the following restricted eigenvalues of the Gram matrix $E_n(xx')$:

$$\kappa_{\bar{c}} = \min_{\delta \in \Delta_{\bar{c}}} \frac{s^{1/2}\|\delta\|_{2,n}}{\|\delta_T\|_1}, \quad \tilde{\kappa}_{\bar{c}} = \min_{\delta \in \Delta_{\varepsilon}} \frac{\|\delta\|_{2,n}}{\|\delta\|_2} \tag{5}$$

Theorem

Consider the model described in (1)–(4). Let $c \geqslant 1$, $c^- = (c+1)/(c-1)$, *and suppose that* λ *obeys the growth restriction* λF, *for some* $\rho \leqslant 1$. *If* $\lambda \geqslant c\Lambda$, *then*

$$\left\|\widehat{\beta} - \beta_0\right\|_{2,n} \leqslant A_n \sigma \left\{E_n\left(\epsilon^2\right)\right\}^{1/2} \frac{\lambda s^{1/2}}{n}, \quad A_n = \frac{2(1+1/c)}{\kappa_{\bar{c}}\left(1-\rho^2\right)} \tag{6}$$

In particular, if $\lambda \geqslant c\Lambda$ *with probability at least* $1-\alpha$, *and* $E_n(\epsilon^2) \leqslant \omega^2$ *with probability at least* $1-\gamma$, *then with probability at least* $1-\alpha-\gamma$,

$$\tilde{\kappa}_{\bar{c}} \left\|\widehat{\beta} - \beta_0\right\|_2 \leqslant \left\|\widehat{\beta} - \beta_0\right\|_{2,n} \leqslant A_n \sigma \omega \frac{\lambda s^{1/2}}{n}, \tag{7}$$

Proof:

Step 1. We show that $\hat{\delta} = \hat{\beta} - \beta_0 \in \Delta_{\bar{c}}$ under the prescribed penalty level. By definition of $\hat{\beta}$

$$\{\widehat{Q}(\widehat{\beta})\}^{1/2} - \left\{\widehat{Q}(\beta_0)\right\}^{1/2} \leqslant \frac{\lambda}{n}\|\beta_0\|_1 - \frac{\lambda}{n}\|\widehat{\beta}\|_1 \leqslant \frac{\lambda}{n}\left(\left\|\widehat{\delta}_T\right\|_1 - \left\|\widehat{\delta}_{T^c}\right\|_1\right), \quad (8)$$

where the last inequality holds because

$$\|\beta_0\|_1 - \|\widehat{\beta}\|_1 = \|\beta_{0T}\|_1 - \|\widehat{\beta}_T\|_1 - \|\widehat{\beta}_{T^C}\|_1 = \left\|\widehat{\delta}_T\right\|_1 - \left\|\widehat{\delta}_{T^c}\right\|_1, \quad (9)$$

Also, if $\lambda \geqslant cn||\tilde{S}||_\infty$ then

$$\{\widehat{Q}(\widehat{\beta})\}^{1/2} - \left\{\widehat{Q}\left(\beta_0\right)\right\}^{1/2} \geqslant \tilde{S}'\hat{\delta} \geqslant -\|\tilde{S}\|_\infty \|\widehat{\delta}\|_1 \geqslant -\frac{\lambda}{cn}\left(\left\|\widehat{\delta}_T\right\|_1 + \left\|\widehat{\delta}_{T^c}\right\|_1\right), \tag{10}$$

where the first inequality hold by convexity of $\hat{Q}^{1/2}$. Combining (8) with (10) we obtain

$$-\frac{\lambda}{cn}\left(\left\|\widehat{\delta}_T\right\|_1 + \left\|\widehat{\delta}_{T^c}\right\|_1\right) \leqslant \frac{\lambda}{n}\left(\left\|\widehat{\delta}_T\right\|_1 - \left\|\widehat{\delta}_{T^c}\right\|_1\right), \tag{11}$$

that is

$$\left\|\widehat{\delta}_{T^c}\right\|_1 \leqslant \frac{c+1}{c-1}\left\|\widehat{\delta}_T\right\|_1 = \bar{c}\left\|\widehat{\delta}_T\right\|_1. \tag{12}$$

Step 2. We derive bounds on the estimation error. We shall use the following relations:

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) = \|\widehat{\delta}\|_{2,n}^2 - 2E_n\left(\sigma\epsilon x'\widehat{\delta}\right) \tag{13}$$

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) = [\{\widehat{Q}(\widehat{\beta})\}^{1/2} + \left\{\widehat{Q}(\beta_0)\right\}^{1/2}][\{\widehat{Q}(\widehat{\beta})\}^{1/2} - \left\{\widehat{Q}(\beta_0)\right\}^{1/2}], \tag{14}$$

$$2|E_n\left(\sigma\epsilon x'\widehat{\delta}\right)| \leqslant 2\left\{\widehat{Q}(\beta_0)\right\}^{1/2} \|\tilde{S}\|_\infty \|\widehat{\delta}\|_1 \tag{15}$$

$$\|\widehat{\delta}_T\|_1 \leqslant \frac{s^{1/2}\|\widehat{\delta}\|_{2,n}}{\kappa_{\bar{c}}}, \hat{\delta} \in \Delta_{\bar{c}}, \tag{16}$$

where (15) holds by Holder inequality and (16) holds by the definition of $\kappa_{\bar{c}}$. Using (10) and (13)–(16) we obtain

$$\|\widehat{\delta}\|_{2,n}^2 \leqslant 2\left\{\widehat{Q}(\beta_0)\right\}^{1/2} \|\tilde{S}\|_\infty \|\widehat{\delta}\|_1 + \left[\{\widehat{Q}(\widehat{\beta})\}^{1/2} + \right.$$

$$\left\{\widehat{Q}(\beta_0)\right\}^{1/2} \frac{\lambda}{n}\left(\frac{s^{1/2}\|\widehat{\delta}\|_{2,n}}{\kappa_{\bar{c}}} - \left\|\hat{\delta}_{Tc}\right\|_1\right). \tag{17}$$

Also using (10) and (16) we obtain

$$\left\{\widehat{Q}\left(\beta\right)\right\}^{1/2} \leqslant \left\{\widehat{Q}\left(\beta_0\right)\right\}^{1/2} + \frac{\lambda}{n}\left(\frac{s^{1/2}\|\widehat{\delta}\|_{2,n}}{\kappa_{\bar{c}}}\right). \tag{18}$$

Combining inequalities (17) and (18) and since $\lambda \geqslant cn||\tilde{S}||_\infty$ we obtain,

$$\|\widehat{\delta}\|_{2,n}^2 \leqslant 2\left\{\widehat{Q}\left(\beta_0\right)\right\}^{1/2}\|\tilde{S}\|_\infty\left\|\widehat{\delta}_T\right\|_1 + 2\left\{\widehat{Q}\left(\beta_0\right)\right\}^{1/2}\frac{\lambda s^{1/2}}{n\kappa_{\bar{c}}}\|\widehat{\delta}\|_{2,n} + \left(\frac{\lambda s^{1/2}}{n\kappa_{\bar{c}}}\|\widehat{\delta}\|_{2,n}\right)^2 \tag{19}$$

and then using (16) we obtain

$$\left\{1 - \left(\frac{\lambda s^{1/2}}{n\kappa_{\bar{c}}}\right)^2\right\}\|\widehat{\delta}\|_{2,n}^2 \leqslant 2\left(\frac{1}{c}+1\right)\left\{\widehat{Q}\left(\beta_0\right)\right\}^{1/2}\frac{\lambda s^{1/2}}{n\kappa_{\bar{c}}}\|\widehat{\delta}\|_{2,n}. \tag{20}$$

Provided that $(n\kappa_{\bar{c}})^{-1}\lambda s^{1/2} \leqslant \rho < 1$ and solving the inequality above we obtain the bound stated in the theorem.

Corollary

This result provides a finite-sample bound for $\hat{\delta}$ that is similar to that for the lasso estimator with known δ, and this result leads to the same rates of convergence as in the case of lasso.
Theorem 1 implies the following bounds when combined with Lemma 1, Lemma 2, and the concentration property.

Corollary

Consider the model described in (1)-(4). Suppose further that $F_0 = \Phi$, λ *is chosen according to the exact option in (14),* $p/\alpha > 8$, *and* $n > 4\log(2/\alpha)$. *Let* $c > 1$, $\bar{c} = (c+1)/(c-1)$, $\upsilon = \{1 + 2/\log(2p/\alpha)\}^{1/2} / [1 - 2\{\log(2/\alpha)/n\}^{1/2}]$, *and for any* l *such that* $1 < l < \{n/\log(1/\alpha)\}^{1/2}$, *set* $\omega^2 = 1 + l[\log(1/\alpha)/n]^{1/2} + l^2\log(1/\alpha)/(2n)$ *and* $\gamma = \alpha^{l^2/4}$. *If slogp is relatively small as compared to n, namely* $c\upsilon\{2s/\log(2p/\alpha)\}^{1/2} \leqslant \kappa_{\bar{c}}\rho$ *for some* $\rho < 1$, *then with probability at least* $1 - \alpha - \gamma$,

$$\tilde{\kappa}_{\bar{c}} \left\| \widehat{\beta} - \beta_0 \right\|_2 \leqslant \left\| \widehat{\beta} - \beta_0 \right\|_{2,n} \leqslant B_n \sigma \left\{ \frac{2s\log(2p/\alpha)}{n} \right\}^{1/2}, \quad B_n = \frac{2(1+c)\nu\omega}{\kappa_{\bar{c}}(1-\rho^2)}. \tag{21}$$

Corollary

Consider the model described in (1)-(4) and suppose that $F_0 = \Phi$, Conditions RE and G hold, and $(s/n)log(p/\alpha) \to 0$, as $n \to \infty$. Let λ be specified according to either the exact or asymptotic option in (14). There is an o(1) term such that with probability at least $1 - \alpha - \gamma$,

$$\kappa \left\| \widehat{\beta} - \beta_0 \right\|_2 \leqslant \left\| \widehat{\beta} - \beta_0 \right\|_{2,n} \leqslant C_n \sigma \left\{ \frac{2s \log(2p/\alpha)}{n} \right\}^{1/2}, \quad C_n = \frac{2(1+c)}{\kappa\{1 - o(1)\}}. \tag{22}$$

Corollary

Consider the model described in (1)-(4). Suppose that Conditions RE, M, and R hold, and $(s/n)\log(p/\alpha) \to 0$, *as* $n \to \infty$. *Let* λ *be specified according to the asymptotic, exact, or semi-exact option in (36). There is an o(1) term such that with probability at least* $1 - \alpha - \gamma$,

$$\kappa \left\| \widehat{\beta} - \beta_0 \right\|_2 \leqslant \left\| \widehat{\beta} - \beta_0 \right\|_{2,n} \leqslant C_n \sigma \left\{ \frac{2s \log(2p/\alpha)}{n} \right\}^{1/2}, \quad C_n = \frac{2(1+c)}{\kappa\{1 - o(1)\}}. \tag{23}$$

The square-root lasso optimization problem is precisely a conic programming problem with second-order conic constraints. Indeed, we can reformulate (1) as follows:

$$\begin{aligned} &\min_{t,v,\beta^+,\beta^-} \ \frac{t}{n^{1/2}} + \frac{\lambda}{n}\sum_{j=1}^{p}\left(\beta_j^+ + \beta_j^-\right) \\ &s.t. \quad \begin{cases} v_i = y_i - x_i'\beta^+ + x_i'\beta^- \quad (i = 1,\ldots,n) \\ Q^{n+1} = \{(v,t) \in \mathbb{R}^n \times \mathbb{R} : t \geqslant \|v\|\} \\ (v,t) \in Q^{n+1}, \beta^+ \in \mathbb{R}_+^p, \beta^- \in \mathbb{R}_+^p \end{cases} \end{aligned} \tag{24}$$

Furthermore, we can show that this problem admits the following strongly dual problem:

$$\begin{aligned} &\max_{a \in \mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} y_i a_i \\ &s.t. \quad \begin{cases} \left|\sum_{i=1}^{n} x_{ij} a_i / n\right| \leqslant \lambda/n \quad (j = 1,\ldots,p) \\ \|a\| \leqslant n^{1/2}\mathbb{R}_+^p \end{cases} \end{aligned} \tag{25}$$

We have the following theorem.

Theorem

The square-root lasso problem (1) is equivalent to the conic programming problem (24), which admits the strongly dual problem (25). Moreover, if the solution $\hat{\beta}$ to the problem(1) satisfies $Y \neq X\beta$, the solution $\hat{\beta}^+$, $\hat{\beta}^-$, $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_n)$ to (24), and the solution $\hat{a}$ to (25) are related via$\hat{\beta} = \hat{\beta}^+ - \hat{\beta}^-$, and $\hat{a} = n^{1/2}\hat{v}/\|\hat{v}\|$.

The equivalence of the square-root lasso problem (1) and the conic programming problem (24) follows immediately from the definitions. To establish the duality, for $e = (1, ..., 1)'$, we can write (24) in matrix form as

$$\min_{\vec{t},v,\beta^+,\beta^-} -\frac{1}{n^{1/2}} + \frac{\lambda}{n} e'\beta^+ + \frac{\lambda}{n} e'\beta^- : \tag{26}$$
$$s.t. \quad \begin{cases} v_i = y_i - x_i'\beta^+ + x_i'\beta^- \quad (i = 1, \ldots, n) \\ Q^{n+1} = \{(v, t) \in \mathbb{R}^n \times \mathbb{R} : t \geqslant \|v\|\} \\ (v, t) \in Q^{n+1}, \beta^+ \in \mathbb{R}^p_+, \beta^- \in \mathbb{R}^p_+ \end{cases}$$

The constraints $X'a + s^+ = \lambda/n$ and $-X'a + s^- = \lambda/n$ lead to $\|X'a\|_\infty \leqslant \lambda/n$. The conic constraint $\left(s^v, s^t\right) \in Q^{n+1}$leads to$1/n^{l/2} = s^l \geqslant \|s^v\| = \|a\|$. By scaling the variable α by n we obtain the stated dual problem.

Since the primal problem is strongly feasible, strong duality holds by Theorem 3.2.6 of Renegar (2001). Thus, by strong duality, we have $n^{-1}\sum_{i=1}^{r} y_i \vec{a}_i = n^{-l/2}\|Y - X\hat{\beta}\| + n^{-i}\lambda\sum_{j=1}^{p} \mid \hat{\beta}_j$. Since $n^{-i}\sum_{i-1}^{n} x_{ij}\hat{a}_i\hat{\beta}_j = \lambda\left|\hat{\beta}_j\right|/n$, we have

$$\frac{1}{n}\sum_{i=1}^{n} y_i\hat{a}_i = \frac{\|Y - X\hat{\beta}\|}{n^{1/2}} + \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} x_{ij}\hat{a}_i\hat{\beta}_j = \frac{\|Y - X\hat{\beta}\|}{n^{1/2}} + \frac{1}{n}\sum_{i=1}^{n}\hat{a}_i\sum_{j=1}^{p} x_{ij}\hat{\beta}_j \tag{27}$$

Rearranging the terms we have $n^{-1}\sum_{i=1}^{n}\left\{\left(y_i - x_i'\hat{\beta}\right)\hat{a}_i\right\} = \|Y - X\hat{\beta}\|/n^{1/2}$.

If $\mid Y - X\hat{\beta}\| > 0$, since $\|\hat{a}\| \leqslant n^{1/2}$, the equality can only hold for $\hat{a} = n^{1/2}(Y - X\hat{\beta})/\|Y - X\hat{\beta}\| = (Y - X\hat{\beta})/\{\hat{Q}(\hat{\beta})\}^{1/2}$.

Outline

Three different computational methods

The conic formulation and the strong duality demonstrated in Theorem 2 allow us to employ both the interior-point and first-order methods for conic programs to compute the square-root lasso. We will give three different computational methods. They are Interior-point methods, First-order methods and Componentwise Search.

Interior-point method (ipm) solvers typically focus on solving conic programming problems in standard form:

$$\min_{w} c' : Aw = b, w \in K \tag{28}$$

where K is a cone. In order to formulate the optimization problem associated with the lasso estimator as a conic programming problem, specifically, associated with the second-order cone $Q^{n+1} = \{(v, t) \in \mathbb{R}^n \times \mathbb{R} : t \geqslant \|v\|\}$, we let $\beta = \beta^+ - \beta^-$ for $\beta^+ \geqslant 0$ and $\beta^- \geqslant 0$. For any vector $v \in \mathbb{R}^n$ and scalar $t \geqslant 0$, we have that $v'v \leqslant t$ is equivalent to $||(v, (t-1)/2)||_2 \leqslant (t+1)/2$. The latter can be formulated as a second-order cone constraint.

Thus, the lasso problem can be cast as

$$\min_{t,\beta^+,\beta^-,a_1,a_2,v} \frac{t}{n} + \frac{\lambda}{n}\sum_{j=1}^{p}\left(\beta_j^+ + \beta_j^-\right)$$

$$s.t. \quad \begin{cases} v = Y - X\beta^+ + X\beta^- \\ t = -1 + 2a_1, t = 1 + 2a_2 \\ Q^{k+1} = \left\{(v,t) \in \mathbb{R}^{k+1} : t \geqslant \|v\|\right\} \\ (v, a_2, a_1) \in Q^{n+2}, t \geqslant 0, \beta^+ \in \mathbb{R}_+^p, \beta^- \in \mathbb{R}_+^p \end{cases} \tag{29}$$

Recall that the square-root lasso optimization problem can be cast similarly, but without auxiliary variables a_1, a_2:

$$\min_{t,v,\beta^+,\beta^-} \frac{t}{n^{1/2}} + \frac{\lambda}{n}\sum_{j=1}^{p}\left(\beta_j^+ + \beta_j^-\right)$$
$$s.t. \quad \begin{cases} v = Y - X\beta^+ + X\beta^- \\ Q^{n+1} = \{(v,t) \in \mathbb{R}^n \times \mathbb{R} : t \geqslant \|v\|\} \\ (v,t) \in Q^{n+1}, \beta^+ \in \mathbb{R}_+^p, \beta^- \in \mathbb{R}_+^p \end{cases} \tag{30}$$

Modern first-order methods focus on structured convex problems of the form:

$$\min_w f\{A(w)+b\}+h(w) \text{ or } \min_w h(w) : A(w)+b \in K \tag{31}$$

where f is a smooth function and h is a structured function that is possibly non differentiable or having extended values.

Lasso is cast as

$$\min_{w} f\{A(w)+b\}+h(w) \tag{32}$$

where $f(\cdot)=\|\cdot\|^2/n$, $h(\cdot)=(\lambda/n)\|\cdot\|_1$, $A=X$ and $b=-Y$. The projection required to be solved on every iteration for a given current point β^k is

$$\beta\left(\beta^k\right)=\arg\min_{\beta} 2E_n\left\{x\left(y-x'\beta^k\right)\right\}'\beta+\frac{1}{2}\mu\left\|\beta-\beta^k\right\|^2+\frac{\lambda}{n}\|\beta\|_1 \tag{33}$$

where μ is a smoothing parameter. It follows that the minimization in β above is separable and can be solved by soft-thresholding as

$$\beta_j\left(\beta^k\right)=\text{sign}\left[\beta_j^k+\frac{2E_n\left\{x_j\left(y-x'\beta^k\right)\right\}}{\mu}\right]\max\left[\left|\beta_j^k+\frac{2E_n\left\{x_j\left(y-x'\beta^k\right)\right\}}{\mu}\right|-\frac{\lambda}{n\mu},0\right. \tag{34}$$

For the square-root lasso the "conic form" is

$$\min h(w) : A(w) + b \in K. \tag{35}$$

Letting $Q^{n+1} = \{(z, t) \in \mathbb{R}^n \times \mathbb{R} : t \geqslant \|z\|\}$ and $h(w) = f(\beta, t) = t/n^{1/2} + (\lambda/n)||\beta||_1$ we have that

$$\min_{\beta,t} \frac{t}{n^{1/2}} + \frac{\lambda}{n}\|\beta\|_1 : A(\beta, t) + b \in Q^{n+1} \tag{36}$$

where $b = (-Y', 0)'$ and $A(\beta, t) \mapsto (\beta' X', t)'$.

In the associated dual problem, the dual variable $z \in \mathbb{R}^n$ is constrained to be $\|z\| \leqslant 1/n^{1/2}$. Thus we obtain

$$\max_{\|z\| \leqslant 1/n^{1/2}} \inf_{\beta} \frac{\lambda}{n}\|\beta\|_1 + \frac{1}{2}\mu \left\|\beta - \beta^k\right\|^2 - z'(Y - X\beta) \tag{37}$$

Given iterates β^k, z^k, as in the case of lasso, the minimization in β is separable and can be solved by soft-thresholding as

$$\beta_j\left(\beta^k, z^k\right) = \text{sign}\left\{\beta_j^k + \left(X'z^k/\mu\right)_j\right\} \max\left\{\left|\beta_j^k + \left(X'z^k/\mu\right)_j\right| - \lambda/(n\mu), 0\right\}. \tag{38}$$

The dual projection accounts for the constraint $\|z\| \leqslant 1/n^{1/2}$ and solves

$$z\left(\beta^k, z^k\right) = \arg \min_{\|z\| \leqslant 1/n^{1/2}} \frac{\theta_k}{2t_k} \|z - z_k\|^2 + \left(Y - X\beta^k\right)' z \tag{39}$$

which yields

$$z\left(\beta^k, z^k\right) = \frac{z_k + (t_k/\theta_k)\left(Y - X\beta^k\right)}{\|z_k + (t_k/\theta_k)\left(Y - X\beta^k\right)\|} \min\left\{\frac{1}{n^{1/2}}, \left\|z_k + (t_k/\theta_k)\left(Y - X\beta^k\right)\right\|\right\} \tag{40}$$

A common approach to solve unconstrained multivariate optimization problems is to do componentwise minimization, looping over components until convergence is achieved. Consider the following lasso optimization problem:

$$\min_{\beta \in \mathbb{R}^p} E_n \left\{ \left(y - x'\beta \right)^2 \right\} + \frac{\lambda}{n} \sum_{j=1}^{p} \gamma_j \left| \beta_j \right| \tag{41}$$

Under standard normalization assumptions we would have $\gamma_j = 1$ and $E_n(x_j^2) = 1(j = 1, \ldots, p)$. The main ingredient of the componentwise search for lasso is the rule that sets optimally the value of β_j given fixed the values of the remaining variables:
For a current point β, let $\beta_{-j} = (\beta_1, \beta_2, \ldots, \beta_j, 0, \beta_{j+1}, \ldots, \beta_p)$:
If $2E_n \{x_j (y - x'\beta_{-j})\} > \lambda\gamma_j/n$, the optimal choice for β_j is

$$\beta_j = \left[-2E_n \left\{x_j \left(y - x'\beta_{-j}\right)\right\} + \lambda\gamma_j/n\right] / E_n \left(x_j^2\right). \tag{42}$$

If $2E_n \{x_j (y - x'\beta_{-j})\} < -\lambda\gamma_j/n$, the optimal choice for β_j is

$$\beta_j = \left[2E_n \left\{x_j \left(y - x'\beta_{-j}\right)\right\} - \lambda\gamma_j/n\right] / E_n \left(x_j^2\right). \tag{43}$$

If $2|E_n \{x_j (y - x'\beta_{-j})\}| \leqslant \lambda\gamma_j/n$, then $\beta_j = 0$.

Despite the additional square-root, which creates a non-separable criterion function, it turns out that the componentwise minimization for the square-root lasso also has a closed form solution. Consider the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p} E_n[\left\{ \left(y - x'\beta\right)^2 \right\}]^{1/2} + \frac{\lambda}{n} \sum_{j=1}^{p} \gamma_j \left|\beta_j\right| \tag{44}$$

The main ingredient of the componentwise search for square-root lasso is the rule that sets optimally the value of β_j given fixed the values of the remaining variables:

If $E_n\{x_j(y-x'\beta_{-j})\} > (\lambda/n)\gamma_j\left\{\hat{Q}(\beta_{-j})\right\}^{1/2}$, set

$$\beta_j = -\frac{E_n\{x_j(y-x'\beta_{-j})]}{E_n(x_j^2)} + \frac{\lambda\gamma_j}{E_n(x_j^2)}\frac{\left[\widehat{Q}(\beta_{-j}) - \{E_n(x_jy - x_jx'\beta_{-j})\}^2\{E_n(x_j^2)\}^{-1}\right]^{1/2}}{\left[n^2 - \{\lambda^2\gamma_j^2/E_n(x_j^2)\}\right]^{1/2}} \tag{45}$$

If $E_n\{x_j(y-x'\beta_{-j})\} < -(\lambda/n)\gamma_j\left\{\hat{Q}(\beta_{-j})\right\}^{1/2}$, set

$$\beta_j = -\frac{E_n\{x_j(y-x'\beta_{-j})]}{E_n(x_j^2)} - \frac{\lambda\gamma_j}{E_n(x_j^2)}\frac{\left[\widehat{Q}(\beta_{-j}) - \{E_n(x_jy - x_jx'\beta_{-j})\}^2\{E_n(x_j^2)\}^{-1}\right]^{1/2}}{\left[n^2 - \{\lambda^2\gamma_j^2/E_n(x_j^2)\}\right]^{1/2}} \tag{46}$$

If $E_n\{x_j(y-x'\beta_{-j})\} \leqslant -(\lambda/n)\gamma_j\left\{\hat{Q}(\beta_{-j})\right\}^{1/2}$, set $\beta_j = 0$.

Additional computational results and Computational times

The square-root lasso is a convex conic programming problem, which allows us to use conic programming methods to compute the square-root lasso estimator. We shall compare the average running times for solving lasso and the square-root lasso in practical problems.

$n = 100, p = 500$	Componentwise	First Order	Interior Point
lasso	0·2173	10·99	2·545
square-root lasso	0·3268	7·345	1·645
$n = 200, p = 1000$	Componentwise	First Order	Interior Point
lasso	0·6115	19·84	14·20
square-root lasso	0·6448	19·96	8·291
$n = 400, p = 2000$	Componentwise	First Order	Interior Point
lasso	2·625	84·12	108·9
square-root lasso	2·687	77·65	62·86

TABLE 1. We use the same design as in the main text, with $s = 5$ and $\sigma = 1$, we averaged the computational times over 100 simulations.

- **Motivation**: square root lasso has been proposed with a key advantage that the optimal regularization parameter is independent of the noise level in the measurements. Compared to L_1 norm, a proper nonconvex regularization is able to achieve sparse recovery with fewer measurements and faster convergence, and is more robust against noise.
- **Main idea**: a class of nonconvex sparsity-inducing penalties is introduced, the resultant formulation is converted to a nonconvex but multiconvex optimization problem, i.o. it is convex in each block of variables. (Xinyue et al., 2016)

Consider the square-root minimization problem regularized by a nonconvex function $J(\cdot)$:

$$\min_x \lambda J(x) + ||Ax - y||_2 \tag{47}$$

in which $\lambda \geqslant 0$ is the regularization parameter while the sparsity-inducing penalty is defined as

$$J(x) = \sum_{i=1}^{N} F(x_i) \tag{48}$$

where $F(\cdot)$ satisfies the following definition. The scalar function $F : \mathbb{R} \to \mathbb{R}^+$ satisfies
(a) F(0) = 0, F(·) is even and not identically zero;
(b) F(·) is nondecreasing on $[0, +\infty)$;
(c) The function x → F(x)/x is nonincreasing on $(0, +\infty)$;
(d) F(·) is weakly convex on $[0, +\infty)$.

- The concept of weak convexity:

$$f\left(\lambda x_1+(1-\lambda)x_2\right)\leqslant \lambda f\left(x_1\right)+(1-\lambda)f\left(x_2\right)-\lambda(1-\lambda)\rho\left\|x_1-x_2\right\|^2$$

  when ρ is negative, it is weakly convex.
- Basically, it allows us to define $\beta<0$ as the largest quantity such that $H(x)=F(x)-\beta x^2$ is convex. There also exists $\alpha>0$ such that $F(x)/x\to\alpha$ as $x\to 0^+$.
- Nonconvexity of $F(\cdot)$ and $J(\cdot)$:

$$\varsigma=-\beta/\alpha$$

- Example:

$$F(x)=\left(|x|-\zeta x^2\right)\mathbf{1}_{|x|\leq\frac{1}{2\zeta}}(x)+\frac{1}{4\zeta}\mathbf{1}_{|x|>\frac{1}{2\zeta}}(x)$$

- To solve the previous objective function, rewrite as:

$$\min_{x,z} \lambda J(x) + ||z||_2 \quad s.t. Ax - z = y \tag{49}$$

- Problem: the objective function above is nonconvex with respect to x.
- Solution: introduce a slack variable $w \in \mathbb{R}^{M+N}$ and add a quadratic term.
- Equivalent form:

$$\begin{aligned} &\min_{x,z,w} \lambda J(x) + ||z||_2 + \frac{\mu}{2}||[x^T z^T]^T - w||_2^2 \\ &s.t. \quad \begin{cases} [A, -I]w = y \\ [x^T, z^T]^T = w \end{cases} \end{aligned} \tag{50}$$

- Thus, problem (6) is convex with respect to x, z, and w separately when $\varsigma \leqslant \mu/(2\lambda\alpha)$.

First consider When **A** has orthonormal rows.
Plutting the constraint above into the cost function yields the following equivalent problem:

$$\begin{aligned} &\min_{x,z,w} \lambda J(x) + ||z||_2 + \frac{\mu}{2}|||[x^T z^T]^T - w||_2^2 + g(w) \\ &s.t. \quad [x^T z^T]^T = w \end{aligned} \tag{51}$$

where $g(w)$ equals 0 if $[AI]w = y$ holds, and equals positive infinity otherwise.

The augmented Lagrangian is:

$$\begin{aligned} L(x,z,w,\gamma) = \lambda J(x) + ||z||_2 + \frac{\mu}{2}|||[x^T z^T]^T - w||_2^2 \\ + g(w) + \gamma^T([x^T z^T]^T - w) \\ + \frac{\rho}{2}|||[x^T z^T]^T - w||_2^2 \end{aligned} \tag{52}$$

in which $\gamma \in \mathbb{R}^{M+N}$ is the dual variable vector and $\rho>0$ is the penalty parameter. Denote $w^T = [w_1^T w_2^T]$ and $\gamma^T = [\gamma_1^T \gamma_2^T]$, where $w_1, \gamma_1 \in \mathbb{R}^N$ and $w_2, \gamma_2 \in \mathbb{R}^M$.

$$\begin{aligned}\mathbf{x}^{t+1} &= \arg\min_{\mathbf{x}} L\left(\mathbf{x}, \mathbf{z}^t, \mathbf{w}^t, \gamma^t\right) \\ &= \text{prox}_{\frac{\lambda}{\mu+\rho}J(\cdot)}\left(\mathbf{w}_1^t - \frac{\gamma_1^t}{\mu+\rho}\right).\end{aligned} \tag{53}$$

$$\begin{aligned}\mathbf{z}^{t+1} &= \arg\min_{\mathbf{x}} L\left(\mathbf{x}^{t+1}, \mathbf{z}, \mathbf{w}^t, \gamma^t\right) \\ &= \text{prox}_{\frac{\lambda}{\mu+\rho}||\cdot||_2}\left(\mathbf{w}_2^t - \frac{\gamma_2^t}{\mu+\rho}\right).\end{aligned} \tag{54}$$

$$\begin{aligned}\mathbf{w}^{t+1} &= \arg\min_{\mathbf{w}} L\left(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}, \mathbf{w}, \boldsymbol{\gamma}^t\right) \\ &= \Pi_{[\mathrm{A}-\mathrm{I}]\mathrm{w}=\mathrm{y}}\left(\left[\left(\mathrm{x}^{t+1}\right)^{\mathrm{T}}\left(\mathbf{z}^{t+1}\right)^{\mathrm{T}}\right]^{\mathrm{T}} + \frac{\gamma^t}{\mu+\rho}\right)\end{aligned} \tag{55}$$

$$\gamma^{t+1} = \gamma^t + \rho\left(\left[\left(\mathbf{x}^{t+1}\right)^{\mathrm{T}}\left(\mathbf{z}^{t+1}\right)^{\mathrm{T}}\right]^{\mathrm{T}} - \mathbf{w}^{t+1}\right) \tag{56}$$

The stopping criterion of ADMM is that the primal and dual residuals must be small, and for problem (7) the quantities are:

$$r_p^{t+1} = ||r_p^{t+1}||_2,\ r_d^{t+1} = ||\mu r_p^{t+1} + (\mu + \rho) r_d^{t+1}||_2 \tag{57}$$

where

$$r_p^{t+1} = [(x^{t+1})^T (z^{t+1})^T]^T - w^{t+1},\ r_d^{t+1} = w^{t+1} - w^t \tag{58}$$

When the rows of A are not orthonormal, the pseudoinverse of $[A, I]$ does not result in a computationally efficient calculation (used in updating slack variable w). In this section, we propose to efficiently solve (6) with a general A using the linearized ADMM.
The augmented Lagrangian of problem (6) is

$$\begin{aligned} L(x,z,w,\gamma_1,\gamma_2) = \lambda J(x) + ||z||_2 + \frac{\mu}{2}||[x^T z^T]^T - w||_2^2 \\ + \gamma_1^T([A-I]w - y) + \frac{\rho}{2}||[A-I]w - y||_2^2 \\ + \gamma_2^T([x^T z^T]^T - w) + \frac{\rho}{2}||[x^T z^T]^T - w||_2^2 \end{aligned} \tag{59}$$

in which $\gamma_1 \in \mathbb{R}^M$ and $\gamma_2 \in \mathbb{R}^{M+N}$ are vectors and ρ>0is the penalty parameter, respectively. Denote $w^T = [w_1^T w_2^T]$ and $\gamma_2^T = [\gamma_{21}^T \gamma_{22}^T]$, where $w_1, \gamma_{21} \in \mathbb{R}^N$ and $w_2, \gamma_{22} \in \mathbb{R}^M$.

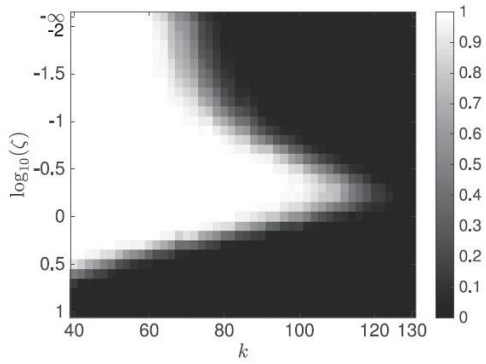

Fig. 1. Recovery probability of Algorithm 1 versus k and ζ when $N = 2^{10}$, $M = 2^8$, and $\lambda = 0.01$. A total of 200 trials are repeated for each point.

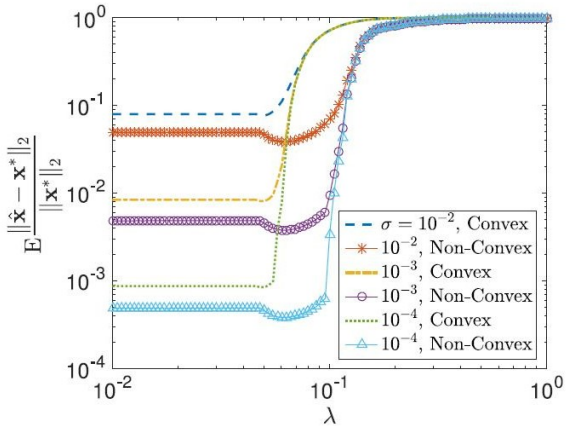

Fig. 2. Mean relative error versus λ when $k = 50$ and $\zeta = 10^{-0.3}$. The convex square-root Lasso is solved by CVX [27], and the nonconvex problem (2) is solved by Algorithm 1. A total of 100 trials are repeated for each point.

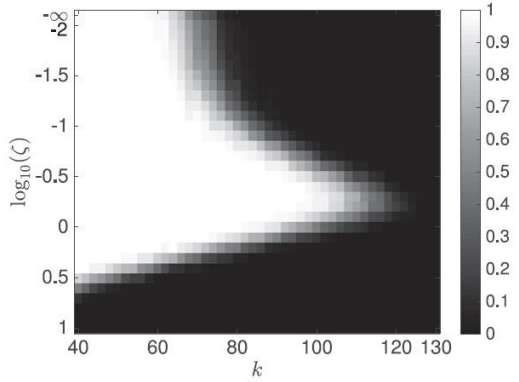

Fig. 3. Recovery probability of Algorithm 2 versus k and ζ when $N = 2^{10}$, $M = 2^8$, and $\lambda = 0.01$. A total of 200 trials are repeated for each point.

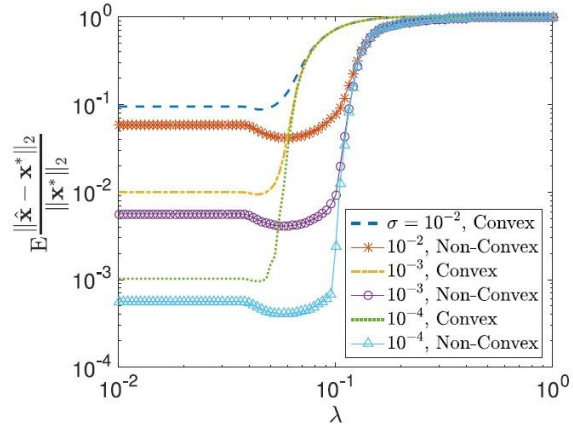

Fig. 4. Mean relative error versus λ when $k = 50$ and $\zeta = 10^{-0.3}$. The convex square-root Lasso is solved by CVX, and the nonconvex problem (2) is solved by Algorithm 2. A total of 100 trials are repeated for each point.

Comparison of CPU Running Time of Different Algorithms

Method \ Size	$N = 2^{12}$, $M = 2^{10}$		$N = 2^{16}$, $M = 2^{14}$	
	$k = 256$	$k = 384$	$k = 4096$	$k = 5632$
CVX [27]	272.10s	—		
ADMM [14]	3.90s	—	58.41s	—
Algorithm 1	0.82s	1.82s	25.98s	66.15s
Algorithm 2	4.46s	9.62s	130.40s	320.54s

The measurements are partial DCT data. The parameters in each algorithm are the same as in the previous experiments. The benchmark algorithms are CVX and ADMM (R Flare package) for the convex square-root Lasso, results are averaged over 10 trials.

Outline

- General form:

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \|Y - X\beta\|_2 + \sum_{j=1}^{q} \lambda_j \left\|\beta^j\right\|_2 \right\} \tag{60}$$

  Where $\lambda_1, \ldots, \lambda_q > 0$ are arbitrary given constants.

- When implement, use the following invariant:

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \|Y - X\beta\|_2 / K + \sum_{j=1}^{q} \lambda_j \left\|\beta^j\right\|_2 \right\} \tag{61}$$

  where K is a fixed, sufficiently large constant.

Scaled thresholding-based iterative selection procedure Abbreviated as **S-TISP**, for solving the general Group Square-Root Lasso problem.

- Scaling step
$$Y \leftarrow Y/K, \quad X \leftarrow X/K$$
- Starting from an arbitrary $\beta(0) \in \mathbb{R}^p$, S-TISP performs the following iterations:
$$\beta^j(t+1) = \vec{\Theta}(\beta^j(t) + \left(X^j\right)'(Y - X\beta(t)); \lambda_j \|\beta(t) - Y\|_2)$$
$$1 \leq j \leq q.$$
- $\vec{\Theta}$ is the multivariate soft-thresholding operator defined through $\vec{\Theta}(0;\lambda) := 0$ and $\vec{\Theta}(a;\lambda) := \Theta\left(\|a\|_2;\lambda\right)/\|a\|_2$ when $a \neq 0$. And $\Theta(t;\lambda) := \text{sign}(t)(|t| - \lambda)_+$ is the soft-thresholding rule.

- Besides the tuning advantage, the regularization selection for SQRT-Lasso type methods is also adaptive to inhomogeneous noise. For example, Han et al. 2015, propose a **multivariate SQRT-Lasso** for sparse multitask learning. Specifically, consider a multitask regression model $Y = X\Theta^* + W$.
  Where Θ^* is solved by a calibrated multivariate regression (CMR) estimator

$$\bar{\Theta}^{\mathrm{CMR}} = \underset{\theta \in \mathbb{R}^{d \times m}}{\operatorname{argmin}} \frac{1}{\sqrt{n}} \sum_{k=1}^{m} \| Y_{*k} - X\Theta_{*k} \|_2 + \lambda_{\mathrm{CMR}} \|\Theta\|_{1,2}$$

- Han et al., 2017 propose a **node-wise SQRT-Lasso** approach for sparse precision matrix Θ estimation. The main idea is to estimate the precision matrix in a column-by-column fashion. For each column, the computation is reduced to a sparse regression problem.

Existing Algorithms for SQRT-Lasso Optimization

- Second order cone program (SOCP): solve by an interior point method with a computational cost of $\mathcal{O}\left(nd^{3.5}\log\left(\epsilon^{-1}\right)\right)$, ϵ is a pre-specified optimization accuracy. (Alexsandre et al., 2011)
- Alternating direction method of multipliers (ADMM) algorithm: computational cost of $\mathcal{O}\left(nd^2/\epsilon\right)$ (Xinguo et al., 2015)
- Coordinate Descent subroutine to accelerate ADMM. (Eugene et al., 2017)
- Proximal gradient descent algorithm and proxmial-Newton algorithm. (Xinguo et al., 2020)

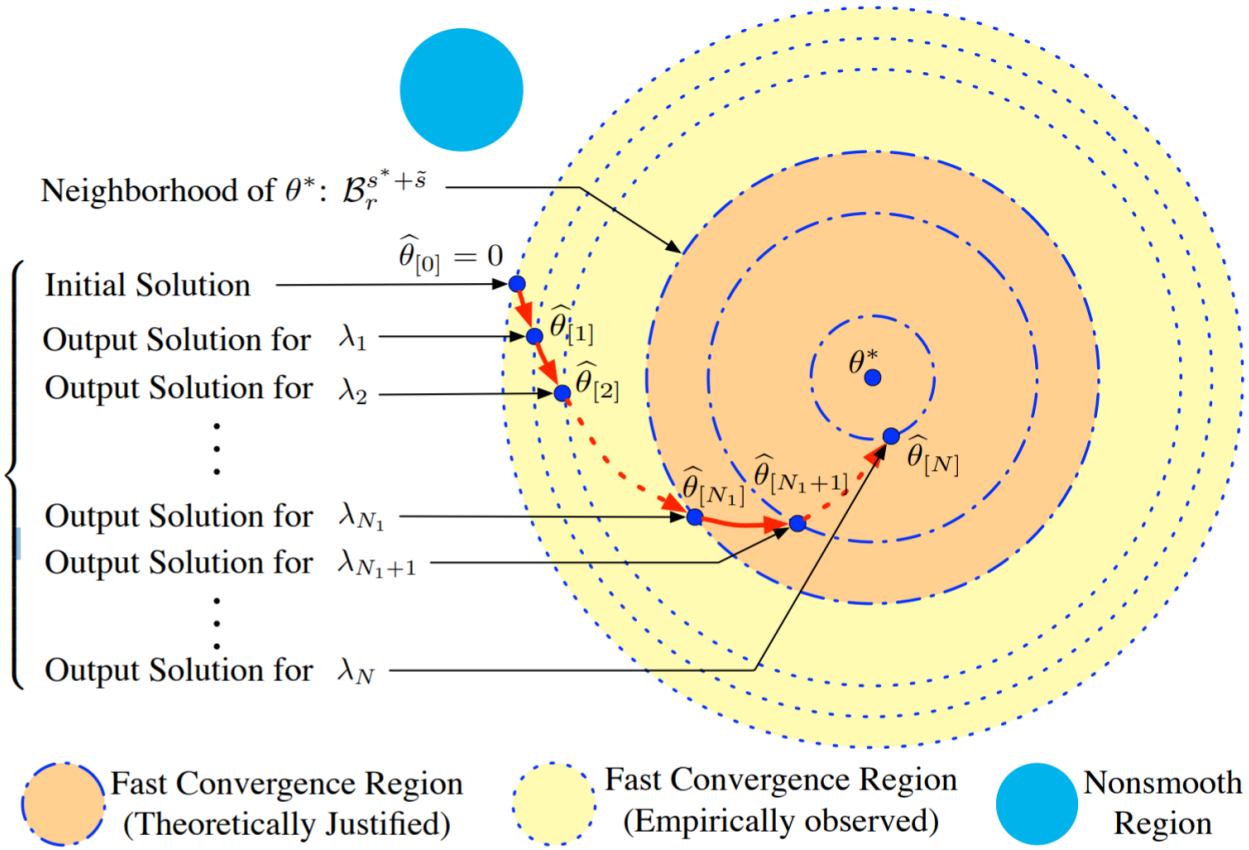

Figure: A geometric illustration for the fast convergence of the proximal algorithms. The proximal algorithms combined with the pathwise optimization scheme suppress the overfitting and yield sparse solutions along the solution path. Non-smooth region is avoided.

Outline

References

1. Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika, 98(4): 791" C806.
2. Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 37(4): 1705-1732.
3. Bunea, F., Lederer, J., and She, Y. (2013). The group square-root lasso: Theoretical properties and fast algorithms. IEEE Transactions on Information Theory, 60(2): 1313-1325.
4. Jing, B.-Y., Shao, Q.-M., Wang, Q., et al. (2003). Self-normalized cramér-type large deviations for independent random variables. The Annals of probability, 31(4): 2167-2215.

5. H. Liu, L. Wang et al., "Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models," Electronic Journal of Statistics, vol. 11, no. 1, pp. 241–294, 2017.
6. H. Liu, L. Wang, and T. Zhao, "Calibrated multivariate regression with application to neural semantic basis discovery." Journal of Machine Learning Research, vol. 16, pp. 1579–1606, 2015.
7. A. Belloni, V. Chernozhukov, and L. Wang, "Square-root Lasso: pivotal recovery of sparse signals via conic programming," Biometrika, vol. 98, no. 4, pp. 791–806, 2011.
8. F. Bunea, J. Lederer and Y. She, "The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms," in IEEE Transactions on Information Theory, vol. 60, no. 2, pp. 1313-1325, Feb. 2014, doi: 10.1109/TIT.2013.2290040.

9. X. Li, T. Zhao, X. Yuan, and H. Liu, "The flare package for high dimensional linear regression and precision matrix estimation in R," The Journal of Machine Learning Research, vol. 16, no. 1, pp. 553–557, 2015.
10. E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclere, ' and J. Salmon, "Efficient smoothed concomitant lasso estimation for high dimensional regression," in Journal of Physics: Conference Series, vol. 904, no. 1. IOP Publishing, 2017, p. 012006.
11. X. Shen, L. Chen, Y. Gu and H. C. So, "Square-Root Lasso With Nonconvex Regularization: An ADMM Approach," in IEEE Signal Processing Letters, vol. 23, no. 7, pp. 934-938, July 2016, doi: 10.1109/LSP.2016.2567482.
12. Stephen Becker1, Emmanuel J. Cand'es2 and Michael Grant1, "Templates for Convex Cone Problems with Applications to Sparse Signal Recovery", Applied and Computational Mathematics, Caltech, Pasadena, CA 91125.

13. Laming Chen and Yuantao Gu, "The Convergence Guarantees of A Non-convex Approach For Sparse Recovery Using Regularized Least Squares", State Key Laboratory on Microwave and Digital Communications.
14. Jim Renegar, "A Framework for Applying First-Order Methods to General Convex Conic Optimization Problems", School of Operations Research Cornell University.
15. Yinyu Ye, "First-Order Methods for Conic Linear Programming", http://www. stanford.edu/ yyye.

Thank you!