C. R. Acad. Sci. Paris, Ser. I 350 (2012) 785–788

Contents lists available at SciVerse ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistics

SOCP based variance free Dantzig Selector with application to robust estimation

Sélecteur de Dantzig indépendant de la variance et application à l'estimation robuste

Arnak S. Dalalyan

ENSAE/CREST/GENES, 3, avenue Pierre-Larousse, 92245 Malakoff cedex, France

ARTICLE INFO

Article history:
Received 5 April 2012
Accepted after revision 19 September 2012
Available online 2 October 2012

Presented by the Editorial Board

ABSTRACT

Sparse estimation methods based on ℓ_1 relaxation, such as Lasso and Dantzig Selector, are powerful tools for estimating high dimensional linear models. However, in order to properly tune these methods, the variance of the noise is often used. In this paper, we propose a new approach to the joint estimation of the sparse vector and the noise variance in a high dimensional linear regression. The method is closely related to the maximum a posteriori estimation and has the attractive feature of being computable by solving a simple second-order cone program (SOCP). We establish nonasymptotic sharp risk bounds for the proposed estimator and show how it can be applied in the problem of robust estimation.

RÉSUMÉ

La calibration des méthodes d'estimation parcimonieuses, telles que le Lasso et le sélecteur de Dantzig, nécessite souvent la connaissance a priori de la variance des erreurs. Nous proposons une méthode qui permet de s'affranchir de cette hypothèse, en estimant le vecteur de régression et la variance des erreurs de façon conjointe. L'estimateur qui en découle est calculable de manière efficace en résolvant un programme conique du second ordre. De plus, nous fournissons des garanties de risque pour cet estimateur presque aussi fortes que celles de l'estimateur utilisant la connaissance de la variance des erreurs.

1. Introduction

Consider the classical problem of Gaussian linear regression:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma^*\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}_n(0, \mathbf{I}_n), \tag{1}$$

where $\boldsymbol{Y} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$ are observed, in the neoclassical setting of very large dimensional unknown vector $\boldsymbol{\beta}^*$. Even if the ambient dimensionality p of $\boldsymbol{\beta}^*$ is larger than n, it has proven possible to consistently estimate this vector under the sparsity assumption. The latter states that the number of nonzero elements of $\boldsymbol{\beta}^*$, denoted by s and called intrinsic dimension, is small compared to the sample size n. Most famous methods of estimating sparse vectors, Lasso and Dantzig Selector (DS), rely on convex relaxation of ℓ_0-norm penalty leading to a convex program that involves the ℓ_1-norm of $\boldsymbol{\beta}$. More precisely, for a given $\bar{\lambda} > 0$, Lasso and DS [10,2,3,1] are defined as

E-mail address: arnak.dalalyan@ensae.fr.

1631-073X/$ – see front matter © 2012 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
http://dx.doi.org/10.1016/j.crma.2012.09.016

$$\widehat{\boldsymbol\beta} = \arg\min_{\boldsymbol\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|\boldsymbol Y - \mathbf{X}\boldsymbol\beta\|_2^2 + \bar\lambda\|\boldsymbol\beta\|_1\right\} \qquad \textbf{(Lasso)}$$

$$\widehat{\boldsymbol\beta} = \arg\min\|\boldsymbol\beta\|_1 \quad \text{subject to } \left\|\mathbf{X}^\top(\boldsymbol Y - \mathbf{X}\boldsymbol\beta)\right\|_\infty \leqslant \bar\lambda. \qquad \textbf{(DS)}$$

The performance of these algorithms depends heavily on the choice of the tuning parameter $\bar\lambda$. Several empirical and theoretical studies emphasized that $\bar\lambda$ should be chosen proportionally to the noise standard deviation σ. Unfortunately, in most applications, the latter is unavailable. It is therefore vital to design statistical procedures that estimate $\boldsymbol\beta$ and σ in a joint fashion. This topic received special attention in last years, cf. [7] and the references therein. Most popular σ-adaptive procedures, the square-root Lasso (a.k.a. scaled Lasso), the ℓ_1 penalized log-likelihood minimization and the STIV [6], can be seen as maximum a posteriori (MAP) estimators with some particular choice of prior distribution. The aim of the present work is to present an alternative to these methods, which is closely related to the MAP, but presents some advantages in terms of implementation and more transparent theoretical analysis.

2. Definition of the procedure and finite sample risk bound

Let $\lambda > 0$ and $\mu\in\mathbb{R}$ be tuning parameters. We define $(\widehat{\boldsymbol\beta},\widehat\sigma)$ as a solution of the optimization problem

$$\text{minimize } \|\boldsymbol\beta\|_1 \quad \text{subject to } \begin{cases} \left\|\mathbf{X}^\top(\mathbf{X}\boldsymbol\beta - \boldsymbol Y)\right\|_\infty \leqslant \lambda\sigma \\ \sqrt{\left(\boldsymbol Y^\top\mathbf{X}\boldsymbol\beta\right)^2 + 4n\mu\|\boldsymbol Y\|_2^2\sigma^2} \leqslant 2\|\boldsymbol Y\|_2^2 - \boldsymbol Y^\top\mathbf{X}\boldsymbol\beta. \end{cases} \qquad \textbf{(MAP-DS)}$$

As we demonstrate later, for a fixed and small tolerance level $\delta > 0$, the choice $\lambda = 2\sqrt{n\log(p/\delta)}$ and $\mu = 1$ leads to an estimator that enjoys strong statistical properties. More precisely, as stated in the following theorem, $\widehat{\boldsymbol\beta}$ satisfies an oracle inequality with a rate optimal remainder term, provided that the Gram matrix $\mathbf{X}^\top\mathbf{X}$ satisfies restricted eigenvalue (RE) condition.

Theorem 2.1. *Let us choose a significance level $\delta\in(0,1)$ and set $\lambda = 2\sqrt{n\log(p/\delta)}$. Assume that $\boldsymbol\beta^*$ has at most s nonzero entries, and satisfies*

$$\frac{\|\boldsymbol\beta^*\|_1}{\sigma^*} \leqslant \sqrt{\frac{n}{2\log(1/\delta)}}\left(1 - 2\sqrt{n^{-1}\log(1/\delta)} - \frac{1}{2}\mu\right). \tag{2}$$

If $\mathbf{X}$ satisfies the condition RE$(s,1)$ (cf. [1], page 6) with some $\kappa > 0$, then, with probability at least $1-3\delta$, it holds

$$\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^*\|_1 \leqslant \frac{4}{\kappa^2}(\sigma^* + \widehat\sigma)s\sqrt{\frac{2\log(p/\delta)}{n}}, \qquad \left\|\mathbf{X}\left(\widehat{\boldsymbol\beta} - \boldsymbol\beta^*\right)\right\|_2^2 \leqslant 8\left(\frac{\sigma^* + \widehat\sigma}{\kappa}\right)^2 s\log(p/\delta). \tag{3}$$

If, in addition, $\mathbf{X}$ satisfies the condition RE$(s,s,1)$ (cf. [1], page 7) with $\kappa > 0$, then

$$\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^*\|_2^2 \leqslant 32\left(\frac{\sigma^* + \widehat\sigma}{\kappa^2}\right)^2 \frac{s\log(p/\delta)}{n}. \tag{4}$$

Moreover, with probability at least $1-4\delta$, $\widehat\sigma \leqslant \sigma^*\mu^{-1/2}(3 + \sqrt{2n^{-1}\log(1/\delta)})$ and therefore

$$\left\|\mathbf{X}\left(\widehat{\boldsymbol\beta} - \boldsymbol\beta^*\right)\right\|_2^2 \leqslant 8(\sigma^*)^2\left(\frac{1 + 3\mu^{-1/2} + \sqrt{2(\mu n)^{-1}\log(1/\delta)}}{\kappa}\right)^2 s\log(p/\delta). \tag{5}$$

Remark 1. Bound (5) is a direct consequence of the second inequality in (3) and the bound on $\widehat\sigma$. In a similar manner, one can combine the first inequality in (3) with the bound on $\widehat\sigma$ to get an upper bound on $\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^*\|_1$ independent of the estimator $\widehat\sigma$. Furthermore, (5) suggests that large values of μ will lead to a smaller prediction loss. However, admissible values of μ are constrained by condition (2). A simple choice for this parameter is to take $\mu = 1$, which implies that bound (5) holds true as soon as the signal-to-noise ratio $\|\boldsymbol\beta^*\|_1/\sigma^*$ is smaller than $(1 - O(n^{-1/2}))\sqrt{\frac{n}{8\log(1/\delta)}}$.

An important advantage of the estimator defined by **(MAP-DS)** is that it can be efficiently computed by solving a second-order cone program (SOCP). In fact, if we introduce slack variables $\boldsymbol u\in\mathbb{R}^p$ and $\boldsymbol v\in\mathbb{R}^2$, then **(MAP-DS)** is equivalent to minimizing $\sum_{i=1}^p u_i$ under the constraints

$$|\beta_j| \leqslant u_j; \qquad \left|\boldsymbol X_j^\top(Y - \mathbf{X}\boldsymbol\beta)\right| \leqslant \lambda\sigma; \qquad \|\boldsymbol v\|_2 \leqslant 2\|\boldsymbol Y\|_2^2 - \boldsymbol Y^\top\mathbf{X}\boldsymbol\beta, \quad \boldsymbol Y^\top\mathbf{X}\boldsymbol\beta = v_1; \qquad \sqrt{4n\mu}\|\boldsymbol Y\|_2\sigma = v_2.$$

Denoting by $\boldsymbol\theta$ the vector of unknowns $(\boldsymbol u,\boldsymbol v,\boldsymbol\beta,\sigma)$, all the aforementioned constraints can be written as $\|\boldsymbol A\boldsymbol\theta + \boldsymbol b\|_2 \leqslant \boldsymbol c^\top\boldsymbol\theta + d$, which is the generic form of the constraints characterizing a SOCP. SOCPs can be solved with great efficiency by standard toolboxes such as SeDuMi or TFOCS.

It should also be noted that **(MAP-DS)** and the choice of parameters are tailored to the case of normalized regressors, *i.e.*, when the diagonal elements of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ are all equal to one. If this is not the case, one can either proceed by normalizing the regressors (dividing each column of $\mathbf{X}$ by its Euclidean norm) or modify the optimization problem as follows:

$$\text{minimize} \sum_{j=1}^{p} \|\beta_j \boldsymbol{X}_j\|_2 \quad \text{subject to} \begin{cases} \left|\boldsymbol{X}_j^\top(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y})\right| \leqslant \lambda\|\boldsymbol{X}_j\|_2\sigma, & \forall j \in \{1,\dots,p\} \\ \sqrt{\left(\boldsymbol{Y}^\top\mathbf{X}\boldsymbol{\beta}\right)^2 + 4n\mu\|\boldsymbol{Y}\|_2^2\sigma^2} \leqslant 2\|\boldsymbol{Y}\|_2^2 - \boldsymbol{Y}^\top\mathbf{X}\boldsymbol{\beta}. \end{cases} \qquad \textbf{(MAP-DS')}$$

This form of the optimization problem is better suited for extensions developed in next sections.

3. Relation with MAP

Let us describe now the connections of the proposed estimator to the general MAP methodology taking its roots in Bayesian statistics. First remark that Lasso is a MAP estimator with Gaussian negative log-likelihood

$$\ell(\boldsymbol{Y}|\boldsymbol{\beta},\sigma) = \frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

and the Laplace prior on $\boldsymbol{\beta}$ given by $\pi(\boldsymbol{\beta}) = (0.5\sigma^{-2}\bar{\lambda})^p \exp(-\bar{\lambda}\|\boldsymbol{\beta}\|_1/\sigma^2)$. Choosing $\bar{\lambda}$ proportional to σ, as suggested by both theory and practice, we come up with the following log-density of the posterior probability (in the case of known σ):

$$-\log\pi(\boldsymbol{\beta}|\boldsymbol{Y},\sigma) = \frac{n+p}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1.$$

When σ is unknown, it is natural to introduce a prior on σ and to compute the corresponding posterior. Our proposal here consists in taking the prior defined by the density (w.r.t. Lebesgue measure): $\bar{\pi}(\sigma) = \sigma^\alpha \mathbf{1}_{(0;+\infty)}(\sigma)$. When $\alpha = p$, the resulting MAP estimator coincides with the one proposed by Städler et al. [9]. Another interesting particular case is $\alpha = -2$, which leads to a noninformative prior Kyung et al. [8]. The former strongly penalizes large values of σ, therefore it is likely that the latter outperforms the former in the examples where the true σ is not too small. Thus, using the prior $\bar{\pi}$, we get the following log-density for the posterior probability:

$$-\log\pi(\boldsymbol{\beta},\sigma|\boldsymbol{Y}) = \frac{n+p-\alpha}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1.$$

The right-hand side is a convex function of the parameters $(\phi,\rho) = (\boldsymbol{\beta}/\sigma, 1/\sigma)$ and the KKT conditions take the form (cf. [9, Prop. 1]):

$$\boldsymbol{X}_j^\top(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y}) = -\sigma\lambda\,\mathrm{sign}(\beta_j), \quad \text{if } \beta_j \neq 0,$$
$$\left|\boldsymbol{X}_j^\top(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y})\right| \leqslant -\sigma\lambda, \quad \text{if } \beta_j = 0,$$

and $\sqrt{\boldsymbol{Y}^\top\mathbf{X}\boldsymbol{\beta} + 4(n+p+\alpha)\|\boldsymbol{Y}\|_2^2\sigma^2} = 2\|Y\|_2^2 - \boldsymbol{Y}^\top\mathbf{X}\boldsymbol{\beta}$. One easily checks that the constraints involved in **(MAP-DS)** are the convex relaxation of these KKT conditions. Therefore, to some extent, our proposal is to minimize the ℓ_1 norm under some constraints that are close to the KKT conditions for the MAP estimator.

4. Extension to fused sparsity and application to robust estimation

In some application, the sparsity condition is more likely to be fulfilled for a linear transformation of $\boldsymbol{\beta}$ rather than for $\boldsymbol{\beta}$ itself. We call such a situation "fused sparsity scenario". It means that for a given $q \times p$ matrix $\mathbf{M}$, the vector $\mathbf{M}\boldsymbol{\beta}^*$ is sparse. We will only consider the case $\mathrm{rank}(\mathbf{M}) = q \leqslant p$, which is more relevant for the applications we have in mind (image denoising, robust estimation, etc.). Under this condition, one can find a $(p-q)\times p$ matrix $\mathbf{M}'$ such that the augmented matrix $\widetilde{\mathbf{M}} = [\mathbf{M}^\top \mathbf{M}'^\top]^\top$ is of full rank. Let us denote by $\boldsymbol{m}_j$ the jth column of the matrix $\widetilde{\mathbf{M}}^{-1}$. Using this notation, we define the estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ as a solution of the following optimization problem:

$$\text{minimize} \sum_{j=1}^{q} \|\mathbf{X}\boldsymbol{m}_j\|_2 \left|(\mathbf{M}\boldsymbol{\beta})_j\right| \quad \text{subject to} \begin{cases} \left|\boldsymbol{m}_j^\top\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y})\right| \leqslant \lambda\sigma\|\mathbf{X}\boldsymbol{m}_j\|_2, & \forall j \in \{1,\dots,q\} \\ \boldsymbol{m}_j^\top\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y}) = 0, & \forall j \in \{q+1,\dots,p\} \\ \sqrt{\left(\boldsymbol{Y}^\top\mathbf{X}\boldsymbol{\beta}\right)^2 + 4n\mu\|\boldsymbol{Y}\|_2^2\sigma^2} \leqslant 2\|\boldsymbol{Y}\|_2^2 - \boldsymbol{Y}^\top\mathbf{X}\boldsymbol{\beta}. \end{cases}$$

The recommended values for parameters (λ,μ) in this problem are $\lambda = 2\sqrt{\log q}$ and $\mu = 1$.

This methodology can be applied in the context of robust estimation, *i.e.*, when the relation $Y_i = (\mathbf{A}\boldsymbol{\theta}^*)_i + \xi_i$ holds only for some indices i, called inliers. Following an idea by [4,5], we introduce a new vector $\boldsymbol{\omega}^* \in \mathbb{R}^n$ that serves to characterize the outliers. If an entry ω_i^* of $\boldsymbol{\omega}^*$ is nonzero, then the corresponding observation Y_i is an outlier. This leads to the model $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma^*\boldsymbol{\xi}$, where $\mathbf{X} = [\mathbf{I}_n\ \mathbf{A}]$ and $\boldsymbol{\beta}^* = [\boldsymbol{\omega}^*\boldsymbol{\theta}^*]$. The sparsity assumption, in this context, tells us that $\boldsymbol{\omega}^* = [\mathbf{I}_n\ \mathbf{0}_{n\times p}]\boldsymbol{\beta}^*$ is sparse, *i.e.*, the number of outliers is significantly smaller than the sample size. We are thus in the setting of fused sparsity with $\mathbf{M} = [\mathbf{I}_n\ \mathbf{0}_{n\times p}]$ and $\mathbf{M}' = [\mathbf{0}_{p\times n}\ \mathbf{I}_p]$.

5. Conclusion

We have introduced a new procedure, which allows us to jointly estimate the regression vector and the noise level in a high dimensional linear regression under the sparsity scenario. We have shown that when the ℓ_1-norm of the true regression vector is not too large, then our procedure is as accurate as those exploiting the knowledge of the true noise level. Using more involved arguments, this analysis can be extended to the case of arbitrary regression vector and not necessarily Gaussian noise distributions. This is an ongoing work, which also includes an intensive empirical evaluation of the procedure and its application to aforementioned tasks of computer vision.

References

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

[2] E. Candes, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, Ann. Statist. 35 (6) (2007) 2313–2351.

[3] E.J. Candès, The restricted isometry property and its implications for compressed sensing, C. R. Acad. Sci. Paris, Ser. I 346 (9–10) (2008) 589–592.

[4] A.S. Dalalyan, R. Keriven, L_1-penalized robust estimation for a class of inverse problems arising in multiview geometry, in: NIPS, 2009, pp. 441–449.

[5] A.S. Dalalyan, R. Keriven, Robust estimation for an inverse problem arising in multiview geometry, J. Math. Imaging Vision 43 (1) (2012) 10–23.

[6] E. Gautier, A.B. Tsybakov, High-dimensional instrumental variables regression and confidence sets, Technical report, 2011, arXiv:1105.2454.

[7] C. Giraud, S. Huet, N. Verzelen, High-dimensional regression with unknown variance, Stat. Sci. (2011), in press, arXiv:1109.5587v2 [math.ST].

[8] M. Kyung, J. Gill, M. Ghosh, G. Casella, Penalized regression, standard errors, and Bayesian lassos, Bayesian Anal. 5 (2) (2010) 369–411.

[9] N. Städler, P. Bühlmann, S. van de Geer, ℓ_1-penalization for mixture regression models, TEST 19 (2) (2010) 209–256.

[10] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc. Ser. B 58 (1) (1996) 267–288.