Optimal variable selection and noisy adaptive compressed sensing

Ndaoud, M.* Tsybakov, A.B.†

Abstract
We propose an algorithm for exact support recovery in the setting of noisy compressed sensing when all entries of the design matrix are i.i.d. standard Gaussian. This algorithm achieves the same (optimal) conditions of exact recovery as the exhaustive search, and has an advantage over the latter of being adaptive to all parameters of the problem and computable in polynomial time. We use a more careful approach to study the non-asymptotic minimax Hamming risk where we show that a non-adaptive variant of our method is nearly optimal.

Key Words: Compressed sensing, SLOPE estimator, exact recovery, Hamming loss, variable selection under sparsity, non-asymptotic minimax risk.

1. Statement of the problem

We consider the linear regression model

$$Y = X\beta + \sigma\xi \tag{1}$$

with the outputs $Y \in \mathbb{R}^n$, the unknown vector $\beta \in \mathbb{R}^p$, the design or sensing matrix $X \in \mathbb{R}^{n\times p}$, and $\sigma > 0$. Motivated by the noisy compressed sensing problem, we assume that all entries of X are i.i.d standard Gaussian random variables, and $\xi \sim \mathcal{N}(0, \mathbb{I}_n)$ is a Gaussian noise independent of X where $\mathbb{I}_n$ denotes the identity matrix. The problem of variable selection is stated as follows: Given the observations (X, Y), estimate the binary vector

$$\eta_\beta = (\mathbb{1}\{\beta_1 \neq 0\}, \dots, \mathbb{1}\{\beta_p \neq 0\})$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, and β_j is the jth component of β.

We assume that the vector β is s-sparse and we denote by S_β its support. Consider the following set of s-sparse vectors:

$$\Omega^p_{s,a} := \{\beta \in \mathbb{R}^p : \quad |\beta|_0 = s \quad \text{and} \quad |\beta_i| \geq a, \quad \forall i \in S_\beta\},$$

where $a > 0$, $s \in \{1, \dots, p\}$, and $|\beta|_0$ denotes the number of non-zero components of β.

In order to estimate η_β (and thus the support S_β), we define a selector $\hat\eta = \hat\eta(X, Y)$ as a measurable function of the observations (X, Y) with values in $\{0, 1\}^p$. The performance of selector $\hat\eta$ is measured by the maximal risks

$$\sup_{\beta\in\Omega^p_{s,a}} \mathbb{P}_\beta(\hat\eta \neq \eta_\beta) \quad \text{and} \quad \sup_{\beta\in\Omega^p_{s,a}} \mathbb{E}_\beta|\hat\eta - \eta_\beta|$$

where $|\hat\eta - \eta_\beta|$ stands for the Hamming distance between $\hat\eta$ and η_β, $\mathbb{P}_\beta$ denotes the joint distribution of (X, Y) satisfying (1), and $\mathbb{E}_\beta$ the corresponding expectation. We say that a selector $\hat\eta$ achieves exact support recovery with respect to the above two risks if

$$\lim_{n,p\to\infty} \sup_{\beta\in\Omega^p_{s,a}} \mathbb{P}_\beta(\hat\eta \neq \eta_\beta) = 0, \tag{2}$$

*1 CREST, ENSAE, Université Paris-Saclay, 5, avenue Henry Le Chatelier, 91120 Palaiseau, France
†1 CREST, ENSAE, Université Paris-Saclay, 5, avenue Henry Le Chatelier, 91120 Palaiseau, France

or

$$\lim_{n,p\to\infty} \sup_{\beta\in\Omega^p_{s,a}} \mathbb{E}_\beta|\hat\eta - \eta_\beta| = 0 \tag{3}$$

where the asymptotics are considered as $p \to \infty$ and $n = n(p, s) \to \infty$ such that $n(p, s) \leq p$. For the sake of brevity, n will always stand for $n(p, s)$.

The selector $\hat\eta$ we suggest here is defined by a two step algorithm, which computes at the first step the SLOPE estimator $\hat\beta$ with parameters defined in [1]. At the second step, the components of $\hat\eta$ are obtained by thresholding of debiased estimators of the components of β based on $\hat\beta$.

We now proceed to the formal definition of this selection procedure. Let x_i denote the ith row of matrix X. Split the sample $(x_i, Y_i), i = 1, \dots, n$, in two subsamples with respective sizes n_1 and n_2, such that $n = n_1 + n_2$. For $k = 1, 2$, denote by $(X^{(k)}, Y^{(k)})$ the corresponding submatrices $X^{(k)} \in \mathbb{R}^{n_k\times p}$ and subvectors $Y^{(k)} \in \mathbb{R}^{n_k}$. The SLOPE estimator $\hat\beta$ based on the first subsample $(X^{(1)}, Y^{(1)})$ is defined as follows. Let λ_j be the tuning parameters

$$\lambda_j = A\sigma\sqrt{\frac{\log(\frac{2p}{j})}{n}}, \qquad j = 1, \dots, p,$$

for a constant $A > 4 + \sqrt{2}$. For any $\beta \in \mathbb{R}^p$, let $(\beta^*_1, \dots, \beta^*_p)$ be a non-increasing rearrangement of $|\beta_1|, \dots, |\beta_p|$. Set

$$|\beta|_* = \sum_{j=1}^{p} \lambda_j\beta^*_j, \quad \beta \in \mathbb{R}^p,$$

which is a norm on $\mathbb{R}^p$. Then, the SLOPE estimator $\hat\beta$ is a solution of the minimization problem

$$\hat\beta \in \arg\min_{\beta\in\mathbb{R}^p} \left(\frac{\|Y^{(1)} - X^{(1)}\beta\|^2}{n} + 2|\beta|_* \right).$$

Let $X^{(2)}_i$ denote the ith column of matrix $X^{(2)}$, and $\hat\beta_j$ the jth component of the SLOPE estimator $\hat\beta$. The suggested selector is defined as a vector

$$\hat\eta(X, Y) = (\hat\eta_1(X, Y), \dots, \hat\eta_p(X, Y)) \tag{4}$$

with components

$$\hat\eta_i(X, Y) = \mathbb{1}\Big\{ \Big(X^{(2)}_i, Y^{(2)} - \sum_{j\neq i} X^{(2)}_j \hat\beta_j \Big) > t(X^{(2)}_i)\|X^{(2)}_i\| \Big\} \tag{5}$$

for $i = 1, \dots, p$, where $(\cdot, \cdot)$ denotes the scalar product, $\|\cdot\|$ is the Euclidean norm, and

$$t\,(u) := \frac{a\|u\|}{2} + \frac{\sigma^2 \log\left(\frac{p}{s} - 1\right)}{a\|u\|}, \quad \forall u \in \mathbb{R}^p.$$

2. Non-asymptotic minimax selectors

In this section, we present a non-asymptotic minimax lower bound on the Hamming risk of any selectors as well as a corresponding non-asymptotic upper bound for the Hamming risk of the selector (4) – (5). Set

$$\psi_+ (n, p, s, a, \sigma) = (p - s)\, \mathbb{P}\, (\sigma\varepsilon > t\,(\zeta)) + s\mathbb{P}\, (\sigma\varepsilon > a\|\zeta\| - t\,(\zeta))$$

where ε is a standard Gaussian random variable, and $\zeta \sim \mathcal{N}\,(0, \mathbb{I}_p)$ is a standard Gaussian random vector in $\mathbb{R}^p$ independent of ε. The following minimax lower bound holds.

Theorem 2.1. *For any $a > 0$, $\sigma > 0$ and any integers n, p, s such that $s < p$ we have*

$$\inf_{\tilde{\eta}} \sup_{\beta \in \Omega_{s,a}} \mathbb{E}_\beta |\tilde{\eta} - \eta_\beta| \geq \psi_+(n, p, s, a, \sigma),$$

where $\inf_{\tilde{\eta}}$ denotes the infimum over all selectors $\tilde{\eta}$.

Consider now a quantity close to ψ_+ given by the formula

$$\psi\,(n, p, s, a, \sigma) = (p - s)\,\mathbb{P}\,(\sigma\varepsilon > t\,(\zeta)) + s\mathbb{P}\,(\sigma\varepsilon > (a\|\zeta\| - t(\zeta))_+)\,.$$

Here, we use the notation $x_+ = \max(x, 0)$. Clearly, $\psi\,(n, p, s, a, \sigma) \leq \psi_+\,(n, p, s, a, \sigma)$.

Theorem 2.2. *Let n, p, s, a, σ be as in Theorem 2.1, and let $\hat{\eta}$ be the selector* (4) – (5). *There exists a constant $c_0 > 0$, such that for all $\delta \in [0, 1]$ and $n_1 > \frac{c_0}{\delta} s \log\left(\frac{ep}{s}\right)$ we have*

$$\sup_{\beta \in \Omega_{s,a}} \mathbb{E}_\beta |\hat{\eta} - \eta_\beta| \leq 2\psi(n_2, p, s, a, \sigma\sqrt{1 + \delta^2}) + p\left(\frac{p}{s}\right)^{-s}.$$

Theorems 2.1 and 2.2 imply near optimality of the selector (4) – (5). Their proofs use the arguments close to those developed in [2] combined with a bound for the ℓ_2 risk of the SLOPE estimator from [1].

3. Study of the phase transition

Using the above upper and lower bounds for the minimax risk, we can study the phase transition that gives necessary and sufficient conditions on the sample size in order to achieve exact recovery. A necessary condition for exact recovery using the Hamming risk follows from Theorem 2.1 and has the form

$$n \geq (1 - \epsilon)\frac{\log(p - s) + 7\log(s)}{4\log\left(1 + \frac{a^2}{4\sigma^2}\right)}, \qquad \forall \epsilon \in (0, 1).$$

Sufficient conditions for the selector (4) – (5) to achieve exact recovery are given by the following result.

Theorem 3.1. *Let n, p, s, a, σ be as in Theorem 2.1, and $a \leq \sigma$. Then, the selector $\hat{\eta}$ defined in* (4) – (5) *achieves exact recovery under both risks (Hamming and support recovery) defined in Section 1 if for some $\delta \in [0, 1]$ and $\epsilon > 0$ the following inequalities hold*

$$n_1 > \frac{c_0}{\delta} s \log\left(\frac{ep}{s}\right) \quad \textit{and} \quad n_2 > (1 + \epsilon)\frac{\log(p - s) + \log(s)}{\log\left(1 + \frac{a^2}{4\sigma^2(1+\delta^2)}\right)}.$$

It is interesting to compare these conditions with the best known in the literature (where only the case of support recovery risk was studied). Necessary condition for exact recovery was established in [3], and is given by the inequality

$$n \geq c\left(\frac{s\log(\frac{p}{s})}{\log(1 + s\frac{a^2}{\sigma^2})} \vee \frac{\log(p - s)}{\log(1 + \frac{a^2}{\sigma^2})}\right)$$

for some absolute constant $c > 0$. As shown by [4], this condition is also sufficient provided a and s are such that $a \leq \sigma$, and $a^2 s \geq c'\sigma^2$ for some $c' > 0$, and it is achieved by the exhaustive search selector. However, the exhaustive search selector cannot be computed in polynomial time. The sufficient conditions we establish in Theorem 3.1 are of the same order, with the advantage that our selector can be computed in polynomial time. Nevertheless, the knowledge of parameters s, a and σ is required for the construction. This motivates us to introduce an adaptive version of this selector.

4. Nearly optimal adaptive procedure

It is known, cf. [1], that the SLOPE estimator is adaptive to the sparsity parameter s. The dependence on s and a only appears in the definition of the threshold $t(\cdot)$. We now take an adaptive threshold of the form

$$t(u) = \sigma\sqrt{2\left(\left(\frac{p}{2}\right)^{\frac{2}{n_2}} - 1\right)}\|u\|, \quad \forall u \in \mathbb{R}^p. \tag{6}$$

Then, we have the following result analogous to Theorem 3.1.

Theorem 4.1. *Let n be an even integer. Set $n_1 = n_2 = n/2$, and let the threshold $t(\cdot)$ be as in (6). Then, the selector $\hat{\eta}$ given in (4) – (5) achieves exact recovery under both risks (Hamming and support recovery) if $n \geq 2\left(c_0 s \log\left(\frac{ep}{s}\right) \vee 2\frac{\log(p)}{\log\left(1+\frac{a^2}{8\sigma^2}\right)}\right)$.*

We also generalize this theorem to the case of unknown variance, for which we need to replace σ in (6) by a suitably chosen estimator $\hat{\sigma}$. The main advantage of this algorithm is its adaptivity to all parameters of the problem.

Recall that the results of this work are obtained under the assumption of i.i.d. Gaussian design. Extension of the above approach to more general designs remains an open problem.

References

[1] Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *arXiv preprint arXiv:1605.08651*, 2016.

[2] Cristina Butucea, Natalia A Stepanova, and Alexandre B Tsybakov. Variable selection with hamming loss. *arXiv preprint arXiv:1512.01832*, 2015.

[3] Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Transactions on Information Theory*, 56(6):2967–2979, 2010.

[4] Kamiar Rahnama Rad. Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Transactions on Information Theory*, 57(7):4672–4679, 2011.