Statistics and Computing manuscript No.
(will be inserted by the editor)

Supplementary Material for “LASSO-type estimators for Semiparametric Nonlinear Mixed-Effects Models Estimation”

Ana Arribas-Gil · **Karine Bertin** · **Cristian Meza** ·
Vincent Rivoirard

the date of receipt and acceptance should be inserted later

1 Theoretical results for the LASSO-type estimator

1.1 Assumptions

As usual, assumptions on the dictionary are necessary to obtain oracle results for LASSO-type procedures. We refer the reader to van de Geer and Bühlmann (2009) for a good review of different assumptions considered in the literature for LASSO-type estimators and connections between them. The dictionary approach aims at extending results for orthonormal bases. Actually, our assumptions express the relaxation of the orthonormality property. To describe them, we introduce the following notation. For $l \in \mathbb{N}$, we denote

$$\nu_{\min}(l) = \min_{|J| \leq l} \min_{\substack{\lambda \in \mathbb{R}^M \\ \lambda_J \neq 0}} \frac{\|f_{\lambda_J}\|_n^2}{\|\lambda_J\|_{\ell_2}^2} \qquad \text{and} \qquad \nu_{\max}(l) = \max_{|J| \leq l} \max_{\substack{\lambda \in \mathbb{R}^M \\ \lambda_J \neq 0}} \frac{\|f_{\lambda_J}\|_n^2}{\|\lambda_J\|_{\ell_2}^2},$$

where $\|\cdot\|_{\ell_2}$ is the l_2 norm in $\mathbb{R}^M$. The notation λ_J means that for any $k \in \{1,\dots,M\}$, $(\lambda_J)_k = \lambda_k$ if $k \in J$ and $(\lambda_J)_k = 0$ otherwise. Previous quantities correspond to the “restricted” eigenvalues of the Gram matrix $G = (G_{j,j'})$ with coefficients

$$G_{j,j'} = \frac{1}{n}\sum_{i=1}^{n} b_i^2 \varphi_j(x_i)\varphi_{j'}(x_i).$$

Assuming that $\nu_{\min}(l)$ and $\nu_{\max}(l)$ are close to 1 means that every set of columns of G with cardinality less than l behaves like an orthonormal system. We also consider the restricted

Ana Arribas-Gil
Departamento de Estadística, Universidad Carlos III de Madrid,
E-mail: ana.arribas@uc3m.es

Karine Bertin · Cristian Meza
CIMFAV-Facultad de Ingeniería, Universidad de Valparaíso,
E-mail: karine.bertin@uv.cl, E-mail: cristian.meza@uv.cl

Vincent Rivoirard
CEREMADE, CNRS-UMR 7534, Université Paris Dauphine,
E-mail: vincent.rivoirard@dauphine.fr

correlations

$$\delta_{l,l'} = \max_{\substack{|J|\le l\\ |J'|\le l'\\ J\cap J'=\emptyset}} \max_{\substack{\lambda,\lambda'\in\mathbb{R}^M\\ \lambda_J\neq 0,\lambda'_{J'}\neq 0}} \frac{\langle f_{\lambda_J}, f_{\lambda'_{J'}}\rangle}{\|\lambda_J\|_{\ell_2}\|\lambda'_{J'}\|_{\ell_2}},$$

where $\langle f,g\rangle = \frac{1}{n}\sum_{i=1}^n b_i^2 f(x_i)g(x_i)$. Small values of $\delta_{l,l'}$ means that two disjoint sets of columns of G with cardinality less than l and l' span nearly orthogonal spaces. We will use the following assumption considered in Bickel et al (2009).

Assumption 1 *For some integer* $1 \le s \le M/2$*, we have*

$$\nu_{\min}(2s) > \delta_{s,2s}. \tag{A1(s)}$$

Oracle inequalities of the Dantzig selector were established under this assumption in the parametric linear model by Candès and Tao (2007) and for density estimation by Bertin et al (2011). It was also considered by Bickel et al (2009) for nonparametric regression and for the LASSO estimate.

Let us denote

$$\kappa_s = \sqrt{\nu_{\min}(2s)}\left(1-\frac{\delta_{s,2s}}{\nu_{\min}(2s)}\right) > 0, \quad \mu_s = \frac{\delta_{s,2s}}{\sqrt{\nu_{\min}(2s)}}.$$

We will say that $\lambda\in\mathbb{R}^M$ satisfies the Dantzig constraints if for all $j=1,\dots,M$

$$\left|(G\lambda)_j - \hat{\beta}_j\right| \le r_{n,j}, \tag{1}$$

where

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^n b_i\varphi_j(x_i)Y_i.$$

We denote $\mathscr{D}$ the set of λ that satisfies (1). The classical use of Karush-Kuhn-Tucker conditions shows that the LASSO estimator $\hat{\lambda}\in\mathscr{D}$, so it satisfies the Dantzig constraint. Finally, we assume in the sequel

$$M \le \exp(n^{\delta}),$$

for $\delta < 1$. Therefore, if $\|\varphi_j\|_n$ is bounded by a constant independent of n and M, then $r_{n,j} = o(1)$ and oracle inequalities established below are meaningful.

1.2 Oracle inequalities

We obtain the following oracle inequalities.

Theorem 1 *Let* $\tau > 2$*. With probability at least* $1-M^{1-\tau/2}$*, for any integer* $s < n/2$ *such that (A1(s)) holds, we have for any* $\alpha > 0$,

$$\|\hat{f}-f\|_n^2 \le \inf_{\lambda\in\mathbb{R}^M}\inf_{\substack{J_0\subset\{1,\dots,M\}\\ |J_0|=s}}\left\{\|f_\lambda - f\|_n^2 + \alpha\left(1+\frac{2\mu_s}{\kappa_s}\right)^2\frac{\Lambda(\lambda,J_0^c)^2}{s} + 16s\left(\frac{1}{\alpha}+\frac{1}{\kappa_s^2}\right)r_n^2\right\} \tag{2}$$

where

$$r_n = \sup_{j=1,\dots,M} r_{n,j},$$

$$\Lambda(\lambda, J_0^c) = \|\lambda_{J_0^C}\|_{\ell_1} + \frac{\left(\|\hat{\lambda}\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+}{2},$$

for any $x \in \mathbb{R}$ $x_+ := \max(x,0)$ and $\|\cdot\|_{\ell_1}$ is the l_1 norm in $\mathbb{R}^M$.

Theorem 2 *Let $\tau > 2$. With probability at least $1 - M^{1-\tau/2}$, for any integer $s < n/2$ such that (A1(s)) holds, we have for any $\alpha > 0$,*

$$\|\hat{f} - f\|_n^2 \leq \inf_{\lambda \in \mathscr{D}} \inf_{\substack{J_0 \subset \{1,\dots,M\} \\ |J_0| = s}} \left\{ \|f_\lambda - f\|_n^2 + \alpha \left(1 + \frac{2\mu_s}{\kappa_s}\right)^2 \frac{\|\lambda_{J_0^C}\|_{\ell_1} + \|\hat{\lambda}_{J_0^C}\|_{\ell_1}}{s} + 32s\left(\frac{1}{\alpha} + \frac{1}{\kappa_s^2}\right) r_n^2 \right\}. \tag{3}$$

Similar oracle inequalities were established by Bunea et al (2006), Bunea et al (2007a), Bunea et al (2007b), or van de Geer (2010). But in these works, the functions of the dictionary are assumed to be bounded by a constant independent of M and n. Let us comment the right-hand side of inequalities (2) and (3) of Theorems 1 and 2. The first term is an approximation term which measures the closeness between f and f_λ and that can vanish if f is a linear combination of the functions of the dictionary. The second term can be considered as a bias term. In both theorems, the term $\|\lambda_{J_0^C}\|_{\ell_1}$ corresponds to the cost of having λ with a support different of J_0. For a given λ, this term can be minimized by choosing J_0 as the set of largest coordinates of λ. Note that if the function f has a sparse expansion on the dictionary, that is $f = f_\lambda$ where λ is a vector with s non-zero coordinates, then by choosing J_0 as the set of the s non-zero coordinates, the approximation term and the term $\|\lambda_{J_0^C}\|_{\ell_1}$ vanish. In Theorem 1, the term $\left(\|\hat{\lambda}\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+$ will be smaller as the ℓ_1-norm of the LASSO estimator is small and this term is equal to 0 if $\|\hat{\lambda}\|_{\ell_1} \leq \|\lambda\|_{\ell_1}$, which is frequently the case. In Theorem 2, given a vector λ such that f_λ approximates well f, the term $\|\hat{\lambda}_{J_0^C}\|_{\ell_1}$ will be small if the LASSO estimator selects the largest coordinates of λ. The last term can be viewed as a variance term corresponding to the estimation of f as linear combination of s functions of the dictionary (see Bertin et al (2011) for more details). Finally, the parameter α calibrates the weights given for the bias and variance terms.

The following section deals with estimation of sparse functions.

1.3 The support property of the LASSO estimate

Let $\tau > 2$. In this section, we apply the LASSO procedure with $\tilde{r}_{n,j}$ instead of $r_{n,j}$, with

$$\tilde{r}_{n,j} = \sigma \|\varphi_j\|_n \sqrt{\frac{\tilde{\tau} \log M}{n}}, \quad \tilde{\tau} > \tau.$$

We assume that the regression function f can be decomposed on the dictionary: there exists $\lambda^* \in \mathbb{R}^M$ such that

$$f = \sum_{j=1}^{M} \lambda_j^* \varphi_j.$$

We denote S^* the support of λ^*:

$$S^* = \left\{ j \in \{1,\dots,M\} : \quad \lambda_j^* \neq 0 \right\},$$

and by s^* the cardinal of S^*. We still consider the LASSO estimate $\hat{\lambda}$ and, similarly, we denote $\hat{S}$ the support of $\hat{\lambda}$:

$$\hat{S} = \left\{ j \in \{1, \ldots, M\} : \quad \hat{\lambda}_j \neq 0 \right\}.$$

One goal of this section is to show that with high probability, we have:

$$\hat{S} \subset S^*.$$

We have the following result.

Theorem 3 *We define*

$$\rho(S^*) = \max_{k \in S^*} \max_{j \neq k} \frac{| < \varphi_j, \varphi_k > |}{\|\varphi_j\|_n \|\varphi_k\|_n}$$

and we assume that there exists $c \in (0, 1/3)$ *such that*

$$s^* \rho(S^*) \leq c.$$

If we have

$$\frac{\sqrt{\tilde{\tau}} + \sqrt{\tau}}{\sqrt{\tilde{\tau}} - \sqrt{\tau}} \leq \frac{1-c}{2c},$$

then

$$\mathbb{P}\left\{\hat{S} \subset S^*\right\} \geq 1 - 2M^{1-\tau/2}.$$

A similar result was established by Bunea (2008) in a slightly less general model. However, her result is based on strong assumptions on the dictionary, namely each function is bounded by a constant L (see Assumption (A2)(a) in Bunea (2008)). This assumption is mild when considering dictionaries only based on Fourier bases. It is no longer the case when wavelets are considered and Bunea's assumption is satisfied only in the case where L depends on M and n on the one hand and is very large on the other hand. Since L plays a main role in the definition of the tuning parameters of the method, with too rough values for L, the procedure cannot achieve satisfying numerical results for moderate values of n even if asymptotic theoretical results of the procedure are good. In the setting of this paper, where we aim at providing calibrated statistical procedures, we avoid such assumptions.

Finally, we have the following corollary.

Corollary 1 *We suppose that* $A1(s^*)$ *is satisfied and that there exists* $c \in (0, 1/3)$ *such that*

$$s^* \rho(S^*) \leq c.$$

If we have

$$\frac{\sqrt{\tilde{\tau}} + \sqrt{\tau}}{\sqrt{\tilde{\tau}} - \sqrt{\tau}} \leq \frac{1-c}{2c},$$

then, with probability at least $1 - 4M^{1-\tau/2}$,

$$\|\hat{f} - f\|_n^2 \leq \frac{32 s^* \tilde{r}_n^2}{\kappa_{s^*}},$$

where

$$\tilde{r}_n = \sup_{j=1,\ldots,M} \tilde{r}_{n,j}.$$

This corollary is a simple consequence of Theorem 2 with $\lambda = \lambda^*$ and $J_0 = S^*$. Taking $\lambda = \lambda^*$ implies that the approximation term vanishes. Taking $J_0 = S^*$ implies that the bias term vanishes since the support of the LASSO estimator is included in the the support of λ^*. In this case, assuming that $\sup_j \|\varphi_j\|_n < \infty$, the rate of convergence is the classical rate $\frac{s^* \log M}{n}$.

2 The proofs

2.1 Preliminary lemma

Lemma 1 *For $1 \leq j \leq M$, we consider the event $\mathscr{A}_j = \left\{|V_j| < r_{n,j}\right\}$ where $V_j = \frac{1}{n}\sum_{i=1}^n b_i \varphi_j(x_i)\varepsilon_i$. Then,*

$$\mathbb{P}(\mathscr{A}_j) \geq 1 - M^{-\tau/2}.$$

Proof of Lemma 1: We have

$$\begin{aligned}\mathbb{P}\left(\mathscr{A}_j^c\right) &\leq \mathbb{P}\left(\sqrt{n}|V_j|/(\sigma\|\varphi_j\|_n) \geq \sqrt{n}r_{n,j}/(\sigma\|\varphi_j\|_n)\right)\\ &\leq \mathbb{P}\left(|Z| \geq \sqrt{\tau\log M}\right)\\ &\leq M^{-\tau/2}\end{aligned}$$

where Z is a standard normal variable. □

2.2 Proof of Theorem 1

Let $\lambda \in \mathbb{R}^M$ and J_0 such that $|J_0| = s$. We have

$$\|f_\lambda - f\|_n^2 = \|\hat{f} - f\|_n^2 + \|f_\lambda - \hat{f}\|_n^2 + \frac{2}{n}\sum_{i=1}^n b_i^2\left(\hat{f}(x_i) - f(x_i)\right)\left(f_\lambda(x_i) - \hat{f}(x_i)\right).$$

We have $\|f_\lambda - \hat{f}\|_n^2 = \|f_\Delta\|_n^2$ where $\Delta = \lambda - \hat{\lambda}$. Moreover

$$A = \frac{2}{n}\sum_{i=1}^n b_i^2\left(\hat{f}(x_i) - f(x_i)\right)\left(f_\lambda(x_i) - \hat{f}(x_i)\right) = 2\sum_{j=1}^M(\lambda_j - \hat{\lambda}_j)\left[(G\hat{\lambda})_j - \beta_j\right],$$

where

$$\beta_j = \frac{1}{n}\sum_{i=1}^n b_i^2\varphi_j(x_i)f(x_i).$$

Since $\hat{\lambda}$ satisfies the Dantzig constraint, we have with probability at least $1 - M^{1-\tau/2}$, for any $j \in \{1, \ldots, M\}$,

$$|(G\hat{\lambda})_j - \beta_j| \leq |(G\hat{\lambda})_j - \hat{\beta}_j| + |\hat{\beta}_j - \beta_j| \leq 2r_{n,j}$$

and $|A| \leq 4r_n\|\Delta\|_1$. This implies that

$$\|\hat{f} - f\|_n^2 \leq \|f_\lambda - f\|_n^2 + 4r_n\|\Delta\|_1 - \|f_\Delta\|_n^2.$$

Moreover using Lemma 1 and Proposition 1 of Bertin et al (2011) (where the norm $\|\cdot\|_2$ is replaced by $\|\cdot\|_n$), we obtain that

$$\left(\|\Delta_{J_0^C}\|_{\ell_1} - \|\Delta_{J_0}\|_{\ell_1}\right)_+ \leq 2\|\lambda_{J_0^C}\|_{\ell_1} + \left(\|\hat{\lambda}\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+ \tag{4}$$

and

$$\begin{aligned}\|f_\Delta\|_n &\geq \kappa_s\|\Delta_{J_0}\|_{\ell_2} - \frac{\mu_s}{\sqrt{|J_0|}}\left(\|\Delta_{J_0^C}\|_{\ell_1} - \|\Delta_{J_0}\|_{\ell_1}\right)_+\\ &\geq \kappa_s\|\Delta_{J_0}\|_{\ell_2} - 2\frac{\mu_s}{\sqrt{|J_0|}}\Lambda(\lambda, J_0^c).\end{aligned}$$

Note that Proposition 1 of Bertin et al (2011) is obtained using Lemma 2 and Lemma 3 of Bertin et al (2011). In our context, Lemma 2 and Lemma 3 can be proved in the same way by replacing the norm $\|\cdot\|_2$ by $\|\cdot\|_n$ and by considering $P_{J_{01}}$ as the projector on the linear space spanned by $(\varphi_j(x_1),\dots,\varphi_j(x_n))_{j\in J_{01}}$.

Now following the same lines as Theorem 2 of Bertin et al (2011), replacing κ_{J_0} by κ_s and μ_{J_0} by μ_s, we obtain the result of the theorem.

2.3 Proof of Theorem 2

We consider $\hat{\lambda}^D$ defined by

$$\hat{\lambda}^D = \mathrm{argmin}_{\lambda\in\mathbb{R}^M}\|\lambda\|_{\ell_1} \quad \text{such that } \lambda \text{ satisfies the Dantzig constraint (1).}$$

Denote by $\hat{f}^D$ the estimator $f_{\hat{\lambda}^D}$. Following the same lines as in the proof of Theorem 1, it can be obtained that, with probability at least $1-M^{1-\tau/2}$, for any integer $s<n/2$ such that (A1(s)) holds, we have for any $\alpha>0$,

$$\|\hat{f}^D-f\|_n^2 \le \inf_{\lambda\in\mathbb{R}^M}\inf_{\substack{J_0\subset\{1,\dots,M\}\\ |J_0|=s}}\left\{\|f_\lambda-f\|_n^2+\alpha\left(1+\frac{2\mu_s}{\kappa_s}\right)^2\frac{\Lambda(\lambda,J_0^c)^2}{s}+16s\left(\frac{1}{\alpha}+\frac{1}{\kappa_s^2}\right)r_n^2\right\},$$

where here

$$\Lambda(\lambda,J_0^c)=\|\lambda_{J_0^C}\|_{\ell_1}+\frac{\left(\|\hat{\lambda}^D\|_{\ell_1}-\|\lambda\|_{\ell_1}\right)_+}{2}.$$

If the infimum is only taken over the vectors λ that satisfy the Dantzig constraint, then, with the same probability we have

$$\|\hat{f}^D-f\|_n^2 \le \inf_{\lambda\in\mathscr{D}}\inf_{\substack{J_0\subset\{1,\dots,M\}\\ |J_0|=s}}\left\{\|f_\lambda-f\|_n^2+\alpha\left(1+\frac{2\mu_s}{\kappa_s}\right)^2\frac{\|\lambda_{J_0^C}\|_{l_1}^2}{s}+16s\left(\frac{1}{\alpha}+\frac{1}{\kappa_s^2}\right)r_n^2\right\}. \tag{5}$$

Following the same lines as the proof of Theorem 1, replacing λ by $\hat{\lambda}^D$, we obtain, with probability at least $1-M^{1-\tau/2}$,

$$\|\hat{f}-f\|_n^2 \le \|\hat{f}^D-f\|_n^2+4r_n\|\Delta\|_1-\|f_\Delta\|_n^2,$$

with $\Delta=\hat{\lambda}-\hat{\lambda}^D$. Applying (4) where $\hat{\lambda}$ plays the role of λ and $\hat{\lambda}^D$ the role of $\hat{\lambda}$, the vector Δ satisfies

$$\left(\|\Delta_{J_0^C}\|_{\ell_1}-\|\Delta_{J_0}\|_{\ell_1}\right)_+ \le 2\|\hat{\lambda}_{J_0^C}\|_{\ell_1}.$$

Following the same lines as in the proof of Theorem 1, we obtain that for each $J_0\subset\{1,\dots,M\}$ such that $|J_0|=s$

$$\|\hat{f}-f\|_n^2 \le \left\{\|\hat{f}^D-f\|_n^2+\alpha\left(1+\frac{2\mu_s}{\kappa_s}\right)^2\frac{\|\hat{\lambda}_{J_0^C}\|_{l_1}^2}{s}+16s\left(\frac{1}{\alpha}+\frac{1}{\kappa_s^2}\right)r_n^2\right\}. \tag{6}$$

Finally, (5) and (6) imply the theorem.

2.4 Proof of Theorem 3

We first state the following lemma.

Lemma 2 *We have for any $u \in \mathbb{R}^M$,*

$$crit(\hat{\lambda}+u) - crit(\hat{\lambda}) \geq \left\| \sum_{k=1}^{M} u_k \varphi_k \right\|_n^2 .$$

Proof of Lemma 2: Since for any λ,

$$\text{crit}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_i f_\lambda(x_i))^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}|\lambda_j|,$$

$$\begin{aligned}
\text{crit}(\hat{\lambda}+u) - \text{crit}(\hat{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - b_i\sum_{k=1}^{M}\hat{\lambda}_k\varphi_k(x_i) - b_i\sum_{k=1}^{M}u_k\varphi_k(x_i)\right)^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}|\hat{\lambda}_j + u_j| \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - b_i\sum_{k=1}^{M}\hat{\lambda}_k\varphi_k(x_i)\right)^2 - 2\sum_{j=1}^{M}\tilde{r}_{n,j}|\hat{\lambda}_j| \\
&= \frac{1}{n}\sum_{i=1}^{n}b_i^2\left(\sum_{k=1}^{M}u_k\varphi_k(x_i)\right)^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}\left(|\hat{\lambda}_j+u_j| - |\hat{\lambda}_j|\right) \\
&\quad -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - b_i\sum_{k=1}^{M}\hat{\lambda}_k\varphi_k(x_i)\right)b_i\sum_{k=1}^{M}u_k\varphi_k(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}b_i^2\left(\sum_{k=1}^{M}u_k\varphi_k(x_i)\right)^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}\left(|\hat{\lambda}_j+u_j| - |\hat{\lambda}_j|\right) \\
&\quad +\frac{2}{n}\sum_{i=1}^{n}b_i^2\sum_{j=1}^{M}\hat{\lambda}_j\varphi_j(x_i)\sum_{k=1}^{M}u_k\varphi_k(x_i) - \frac{2}{n}\sum_{i=1}^{n}b_iy_i\sum_{k=1}^{M}u_k\varphi_k(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}b_i^2\left(\sum_{k=1}^{M}u_k\varphi_k(x_i)\right)^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}\left(|\hat{\lambda}_j+u_j| - |\hat{\lambda}_j|\right) \\
&\quad +\frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{M}u_k\varphi_k(x_i)\left(b_i^2\sum_{j=1}^{M}\hat{\lambda}_j\varphi_j(x_i) - b_iy_i\right).
\end{aligned}$$

Since $\hat{\lambda}$ minimizes $\lambda \longmapsto \text{crit}(\lambda)$, we have for any k,

$$0 = \frac{2}{n}\sum_{i=1}^{n}\varphi_k(x_i)\left(b_i^2\sum_{j=1}^{M}\hat{\lambda}_j\varphi_j(x_i) - b_iy_i\right) + 2\tilde{r}_{n,k}s(\hat{\lambda}_k),$$

where $|s(\hat{\lambda}_k)| \leq 1$ and $s(\hat{\lambda}_k) = \text{sign}(\hat{\lambda}_k)$ if $\hat{\lambda}_k \neq 0$. So,

$$\frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{M}u_k\varphi_k(x_i)\left(b_i^2\sum_{j=1}^{M}\hat{\lambda}_j\varphi_j(x_i) - b_iy_i\right) = -2\sum_{k=1}^{M}u_k\tilde{r}_{n,k}s(\hat{\lambda}_k)$$

and

$$
\begin{aligned}
\mathrm{crit}(\hat{\lambda}+u)-\mathrm{crit}(\hat{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n} b_i^2\left(\sum_{k=1}^{M} u_k\varphi_k(x_i)\right)^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}\left(|\hat{\lambda}_j+u_j|-|\hat{\lambda}_j|\right)\\
&\quad -2\sum_{k=1}^{M} u_k\tilde{r}_{n,k}s(\hat{\lambda}_k)\\
&= \frac{1}{n}\sum_{i=1}^{n} b_i^2\left(\sum_{k=1}^{M} u_k\varphi_k(x_i)\right)^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}\left(|\hat{\lambda}_j+u_j|-|\hat{\lambda}_j|-u_js(\hat{\lambda}_j)\right)\\
&\geq \frac{1}{n}\sum_{i=1}^{n} b_i^2\left(\sum_{k=1}^{M} u_k\varphi_k(x_i)\right)^2,
\end{aligned}
$$

which proves the result. □

Now, still with $s^*=\mathrm{card}(S^*)$, we consider for $\mu\in\mathbb{R}^{s^*}$

$$
\mathrm{critS}^*(\mu)=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-b_i\sum_{j\in S^*}\mu_j\varphi_j(x_i)\right)^2+2\sum_{j\in S^*}\tilde{r}_{n,j}|\mu_j|,
$$

and

$$
\tilde{\mu}=\arg\min_{\mu\in\mathbb{R}^{s^*}}\mathrm{critS}^*(\mu).
$$

Then we set

$$
\mathscr{S}=\bigcap_{j\notin S^*}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}y_ib_i\varphi_j(x_i)-\sum_{k\in S^*}\tilde{\mu}_k<\varphi_j,\varphi_k>\right|<\tilde{r}_{n,j}\right\}
$$

and we state the following lemma.

Lemma 3 *On the set $\mathscr{S}$, the non-zero coordinates of $\hat{\lambda}$ are included into S^*.*

Proof of Lemma 3: Recall that $\hat{\lambda}$ is a minimizer of $\lambda\longmapsto\mathrm{crit}(\lambda)$. Using standard convex analysis arguments, this is equivalent to say that for any $1\leq j\leq M$,

$$
\begin{cases}
\frac{1}{n}\sum_{i=1}^{n}y_ib_i\varphi_j(x_i)-\sum_{k=1}^{M}\hat{\lambda}_k<\varphi_j,\varphi_k> \;=\tilde{r}_{n,j}\mathrm{sign}(\hat{\lambda}_j) & \text{if } \hat{\lambda}_j\neq 0,\\
\left|\frac{1}{n}\sum_{i=1}^{n}y_ib_i\varphi_j(x_i)-\sum_{k=1}^{M}\hat{\lambda}_k<\varphi_j,\varphi_k>\right|\leq\tilde{r}_{n,j} & \text{if } \hat{\lambda}_j=0.
\end{cases}
$$

Similarly, on $\mathscr{S}$, we have

$$
\begin{cases}
\frac{1}{n}\sum_{i=1}^{n}y_ib_i\varphi_j(x_i)-\sum_{k\in S^*}\tilde{\mu}_k<\varphi_j,\varphi_k> \;=\tilde{r}_{n,j}\mathrm{sign}(\tilde{\mu}_j) & \text{if } j\in S^* \text{ and } \tilde{\mu}_j\neq 0,\\
\left|\frac{1}{n}\sum_{i=1}^{n}y_ib_i\varphi_j(x_i)-\sum_{k\in S^*}\tilde{\mu}_k<\varphi_j,\varphi_k>\right|\leq\tilde{r}_{n,j} & \text{if } j\in S^* \text{ and } \tilde{\mu}_j=0,\\
\left|\frac{1}{n}\sum_{i=1}^{n}y_ib_i\varphi_j(x_i)-\sum_{k\in S^*}\tilde{\mu}_k<\varphi_j,\varphi_k>\right|<\tilde{r}_{n,j} & \text{if } j\notin S^*.
\end{cases}
$$

So, on $\mathscr{S}$, the vector $\hat{\mu}$ such $\hat{\mu}_j=\tilde{\mu}_j$ if $j\in S^*$ and $\hat{\mu}_j=0$ if $j\notin S^*$ is also a minimizer of $\lambda\longmapsto\mathrm{crit}(\lambda)$. Using Lemma 2, we have for any $1\leq i\leq n$:

$$
\sum_{k=1}^{M}(\hat{\lambda}_k-\hat{\mu}_k)\varphi_k(x_i)=0.
$$

So, for $j \notin S^*$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} y_i b_i \varphi_j(x_i) - \sum_{k=1}^{M} \hat{\lambda}_k < \varphi_j, \varphi_k > \right| < \tilde{r}_{n,j}.$$

Therefore, on $\mathscr{S}$, the non-zero coordinates of $\hat{\lambda}$ are included into S^*. □

Lemma 3 shows that we just need to prove that

$$\mathbb{P}\{\mathscr{S}\} \geq 1 - 2M^{1-\tau/2}$$

$$\begin{aligned} \mathbb{P}\{\mathscr{S}^c\} &\leq \sum_{j \notin S^*} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} y_i b_i \varphi_j(x_i) - \sum_{k \in S^*} \tilde{\mu}_k < \varphi_j, \varphi_k > \right| \geq \tilde{r}_{n,j} \right\} \\ &\leq A + B, \end{aligned}$$

with

$$\begin{aligned} A &= \sum_{j \notin S^*} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} [y_i b_i \varphi_j(x_i) - \mathbb{E}(y_i b_i \varphi_j(x_i))] \right| \geq r_{n,j} \right\} \\ &= \sum_{j \notin S^*} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i b_i \varphi_j(x_i) \right| \geq r_{n,j} \right\} \\ &= \sum_{j \notin S^*} \mathbb{P}\left\{ |V_j| \geq r_{n,j} \right\} \end{aligned}$$

(see Lemma 1) and

$$\begin{aligned} B &= \mathbb{P}\left[\bigcup_{j \notin S^*} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(y_i b_i \varphi_j(x_i)) - \sum_{k \in S^*} \tilde{\mu}_k < \varphi_j, \varphi_k > \right| \geq \tilde{r}_{n,j} - r_{n,j} \right\} \right] \\ &= \mathbb{P}\left[\bigcup_{j \notin S^*} \left\{ \left| < \varphi_j, f_{\lambda^*} > - \sum_{k \in S^*} \tilde{\mu}_k < \varphi_j, \varphi_k > \right| \geq \tilde{r}_{n,j} - r_{n,j} \right\} \right] \\ &= \mathbb{P}\left[\bigcup_{j \notin S^*} \left\{ \left| \sum_{k \in S^*} (\lambda_k^* - \tilde{\mu}_k) < \varphi_j, \varphi_k > \right| \geq \tilde{r}_{n,j} - r_{n,j} \right\} \right] \\ &\leq \mathbb{P}\left[\bigcup_{j \notin S^*} \left\{ \rho(S^*) \|\varphi_j\|_n \sum_{k \in S^*} |\lambda_k^* - \tilde{\mu}_k| \, \|\varphi_k\|_n \geq \tilde{r}_{n,j} - r_{n,j} \right\} \right] \end{aligned}$$

since

$$\rho(S^*) = \max_{k \in S^*} \max_{j \neq k} \frac{|< \varphi_j, \varphi_k >|}{\|\varphi_j\|_n \|\varphi_k\|_n}.$$

Using notation of Lemma 3, we have:

$$\begin{aligned} \|f_{\lambda^*} - f_{\hat{\mu}}\|_n^2 &= \| \sum_{k \in S^*} (\lambda_k^* - \hat{\mu}_k) \varphi_k \|_n^2 \\ &= \sum_{k \in S^*} (\lambda_k^* - \hat{\mu}_k)^2 \|\varphi_k\|_n^2 + \sum_{k \in S^*} \sum_{j \in S^*,\ j \neq k} (\lambda_k^* - \hat{\mu}_k)(\lambda_j^* - \hat{\mu}_j) < \varphi_j, \varphi_k >, \end{aligned}$$

and

$$\sum_{k\in S^*}(\lambda_k^*-\hat{\mu}_k)^2\|\varphi_k\|_n^2 \leq \|f_{\lambda^*}-f_{\hat{\mu}}\|_n^2+\rho(S^*)\sum_{k\in S^*}\sum_{j\in S^*,\ j\neq k}|\lambda_k^*-\hat{\mu}_k|\|\varphi_k\|_n\times|\lambda_j^*-\hat{\mu}_j|\|\varphi_j\|_n$$
$$\leq \|f_{\lambda^*}-f_{\hat{\mu}}\|_n^2+\rho(S^*)\left(\sum_{k\in S^*}|\lambda_k^*-\hat{\mu}_k|\|\varphi_k\|_n\right)^2.$$

Finally,

$$\left(\sum_{k\in S^*}|\lambda_k^*-\hat{\mu}_k|\|\varphi_k\|_n\right)^2 \leq s^*\sum_{k\in S^*}(\lambda_k^*-\hat{\mu}_k)^2\|\varphi_k\|_n^2$$
$$\leq s^*\left(\|f_{\lambda^*}-f_{\hat{\mu}}\|_n^2+\rho(S^*)\left(\sum_{k\in S^*}|\lambda_k^*-\hat{\mu}_k|\|\varphi_k\|_n\right)^2\right),$$

which shows that

$$\left(\sum_{k\in S^*}|\lambda_k^*-\hat{\mu}_k|\|\varphi_k\|_n\right)^2 \leq \frac{s^*}{1-\rho(S^*)s^*}\|f_{\lambda^*}-f_{\hat{\mu}}\|_n^2.$$

Now,

$$\frac{1}{n}\sum_{i=1}^n\left(y_i-b_i\sum_{j\in S^*}\tilde{\mu}_j\varphi_j(x_i)\right)^2+2\sum_{j\in S^*}\tilde{r}_{n,j}|\tilde{\mu}_j| \leq$$
$$\frac{1}{n}\sum_{i=1}^n\left(y_i-b_i\sum_{j\in S^*}\lambda_j^*\varphi_j(x_i)\right)^2+2\sum_{j\in S^*}\tilde{r}_{n,j}|\lambda_j^*|.$$

So,

$$\|\sum_{j\in S^*}\tilde{\mu}_j\varphi_j\|_n^2-\frac{2}{n}\sum_{i=1}^n b_iy_i\sum_{j\in S^*}\tilde{\mu}_j\varphi_j(x_i)+2\sum_{j\in S^*}\tilde{r}_{n,j}|\tilde{\mu}_j| \leq$$
$$\|\sum_{j\in S^*}\lambda_j^*\varphi_j\|_n^2-\frac{2}{n}\sum_{i=1}^n b_iy_i\sum_{j\in S^*}\lambda_j^*\varphi_j(x_i)+2\sum_{j\in S^*}\tilde{r}_{n,j}|\lambda_j^*|,$$

and using previous notation,

$$\|f_{\tilde{\mu}}\|_n^2-\frac{2}{n}\sum_{i=1}^n b_iy_i\sum_{j\in S^*}\tilde{\mu}_j\varphi_j(x_i)+2\sum_{j\in S^*}\tilde{r}_{n,j}|\tilde{\mu}_j| \leq$$
$$\|f_{\lambda^*}\|_n^2-\frac{2}{n}\sum_{i=1}^n b_iy_i\sum_{j\in S^*}\lambda_j^*\varphi_j(x_i)+2\sum_{j\in S^*}\tilde{r}_{n,j}|\lambda_j^*|.$$

Therefore,

$$\begin{aligned}
\|f_{\lambda^*}-f_{\hat{\mu}}\|_n^2 &= \|f_{\hat{\mu}}\|_n^2+\|f_{\lambda^*}\|_n^2-2<f_{\hat{\mu}},f_{\lambda^*}> \\
&\leq 2\|f_{\lambda^*}\|_n^2-2<f_{\hat{\mu}},f_{\lambda^*}>+\frac{2}{n}\sum_{i=1}^n b_iy_i\sum_{j\in S^*}(\tilde{\mu}_j-\lambda_j^*)\varphi_j(x_i)+2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*|-|\tilde{\mu}_j|) \\
&= \frac{2}{n}\sum_{i=1}^n b_iy_i(f_{\hat{\mu}}(x_i)-f_{\lambda^*}(x_i))-\frac{2}{n}\sum_{i=1}^n b_i^2 f_{\lambda^*}(x_i)(f_{\hat{\mu}}(x_i)-f_{\lambda^*}(x_i)) \\
&\qquad\qquad +2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*|-|\tilde{\mu}_j|) \\
&= \frac{2}{n}\sum_{i=1}^n b_i(y_i-\mathbb{E}(y_i))(f_{\hat{\mu}}(x_i)-f_{\lambda^*}(x_i))+2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*|-|\tilde{\mu}_j|) \\
&= \frac{2}{n}\sum_{i=1}^n b_i\varepsilon_i(f_{\hat{\mu}}(x_i)-f_{\lambda^*}(x_i))+2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*|-|\tilde{\mu}_j|) \\
&= 2\sum_{j=1}^M V_j(\hat{\mu}_j-\lambda_j^*)+2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*|-|\tilde{\mu}_j|).
\end{aligned}$$

Now let us assume that for any $j\in S^*$, $V_j<r_{n,j}$. Then,

$$\begin{aligned}
\|f_{\lambda^*}-f_{\hat{\mu}}\|_n^2 &< 2\sum_{j\in S^*}(r_{n,j}+\tilde{r}_{n,j})|\hat{\mu}_j-\lambda_j^*| \\
&< 2\sigma\sqrt{\frac{\log M}{n}}(\sqrt{\tau}+\sqrt{\tilde{\tau}})\sum_{j\in S^*}\|\varphi_j\|_n|\hat{\mu}_j-\lambda_j^*|.
\end{aligned}$$

So,

$$\sum_{k\in S^*}|\lambda_k^*-\hat{\mu}_k|\|\varphi_k\|_n < 2\sigma\sqrt{\frac{\log M}{n}}(\sqrt{\tau}+\sqrt{\tilde{\tau}})\frac{s^*}{1-\rho(S^*)s^*}$$

and for any $j\notin S^*$,

$$\begin{aligned}
\rho(S^*)\|\varphi_j\|_n\sum_{k\in S^*}|\lambda_k^*-\hat{\mu}_k|\|\varphi_k\|_n &< 2\sigma\sqrt{\frac{\log M}{n}}\|\varphi_j\|_n(\sqrt{\tau}+\sqrt{\tilde{\tau}})\frac{\rho(S^*)s^*}{1-\rho(S^*)s^*} \\
&< \frac{2\sigma c(\sqrt{\tau}+\sqrt{\tilde{\tau}})}{1-c}\sqrt{\frac{\log M}{n}}\|\varphi_j\|_n \\
&< (\sqrt{\tilde{\tau}}-\sqrt{\tau})\sigma\sqrt{\frac{\log M}{n}}\|\varphi_j\|_n \\
&< \tilde{r}_{n,j}-r_{n,j}.
\end{aligned}$$

Therefore,

$$B\leq\sum_{j\in S^*}\mathbb{P}\left\{|V_j|\geq r_{n,j}\right\}$$

and using Lemma 1, since $\mathbb{P}\{\mathscr{S}^c\}\leq A+B$,

$$\mathbb{P}\{\mathscr{S}\}\geq 1-2M^{1-\tau/2}.$$

2.5 Proof of Corollary 1

First note that λ^* satisfies the Dantzig constraint (1) where $r_{n,j}$ is replaced by $\tilde{r}_{n,j}$ with probability larger than $1-M^{1-\tilde{\tau}/2}$. On the event $\hat{S} \subset S^*$, we have $\lambda^*_{(S^*)^C} = \hat{\lambda}_{(S^*)^C} = 0$, then applying Theorem 2, we obtain that for any $\alpha > 0$

$$\|\hat{f} - f\|_n^2 \leq 32 s^* \left(\frac{1}{\alpha} + \frac{1}{\kappa_{s^*}^2} \right) \tilde{r}_n^2,$$

which implies the result of the theorem.

References

Bertin K, Le Pennec E, Rivoirard V (2011) Adaptive Dantzig density estimation. Annales de l'Institut Henri Poincaré 47:43–74

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of lasso and Dantzig selector. Ann Statist 37(4):1705–1732

Bunea F (2008) Consistent selection via the Lasso for high dimensional approximating regression models. In: Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh, Inst. Math. Stat. Collect., vol 3, Inst. Math. Statist., Beachwood, OH, pp 122–137

Bunea F, Tsybakov AB, Wegkamp MH (2006) Aggregation and sparsity via l_1 penalized least squares. In: Learning theory, Lecture Notes in Comput. Sci., vol 4005, Springer, Berlin, pp 379–391

Bunea F, Tsybakov A, Wegkamp M (2007a) Sparsity oracle inequalities for the Lasso. Electronic Journal of Statistics 1:169–194

Bunea F, Tsybakov AB, Wegkamp MH (2007b) Aggregation for Gaussian regression. The Annals of Statistics 35(4):1674–1697

Candès EJ, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics 35(6):2313–2351

van de Geer S (2010) ℓ_1-regularization in high-dimensional statistical models. In: Proceedings of the International Congress of Mathematicians. Volume IV, Hindustan Book Agency, New Delhi, pp 2351–2369

van de Geer SA, Bühlmann P (2009) On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics 3:1360–1392