Adaptive Dantzig density estimation

Karine Bertin, Erwan Le Pennec, Vincent Rivoirard

Adaptive Dantzig density estimation

K. Bertin[*], E. Le Pennec[†], V. Rivoirard[‡]

Abstract

This paper deals with the problem of density estimation. We aim at building an estimate of an unknown density as a linear combination of functions of a dictionary. Inspired by Candès and Tao's approach, we propose an ℓ_1-minimization under an adaptive Dantzig constraint coming from sharp concentration inequalities. This allows to consider a wide class of dictionaries. Under local or global coherence assumptions, oracle inequalities are derived. These theoretical results are also proved to be valid for the natural Lasso estimate associated with our Dantzig procedure. Then, the issue of calibrating these procedures is studied from both theoretical and practical points of view. Finally, a numerical study shows the significant improvement obtained by our procedures when compared with other classical procedures.

Keywords : Calibration, Concentration inequalities, Dantzig estimate, Density estimation, Dictionary, Lasso estimate, Oracle inequalities, Sparsity.

AMS subject classification : 62G07, 62G05, 62G20

1 Introduction

Various estimation procedures based on l_1 penalization (exemplified by the Dantzig procedure in [13] and the LASSO procedure in [28]) have extensively been studied recently. These procedures are computationally efficient as shown in [17, 24, 25], and thus are adapted to high-dimensional data. They have been widely used in regression models, but only the Lasso estimator has been studied in the density model (see [7, 10, 29]). Although we will mostly consider the Dantzig estimator in the density model for which no result exists so far, we recall some of the classical results obtained in different settings by procedures based on l_1 penalization.

The Dantzig selector has been introduced by Candès and Tao [13] in the linear regression model. More precisely, given

$$Y = A\lambda_0 + \varepsilon,$$

where $Y \in \mathbb{R}^n$, A is a n by M matrix, $\varepsilon \in \mathbb{R}^n$ is the noise vector and $\lambda_0 \in \mathbb{R}^M$ is the unknown regression parameter to estimate, the Dantzig estimator is defined by

$$\hat{\lambda}^D = \arg\min_{\lambda\in\mathbb{R}^M} \|\lambda\|_{\ell_1} \text{ subject to } \|A^T(A\lambda - Y)\|_{\ell_\infty} \le \eta,$$

[*]Supported by Project PBCT 13 laboratorio ANESTOC and Project FONDECYT 1090285. Departamento de Estadística, CIMFAV, Universidad de Valparaíso, Avenida Gran Bretaña 1091, Valparaíso, Chile. Tel 0056-(0)32-2508324. Email: karine.bertin@uv.cl

[†]Laboratoire de Probabilité et Modèles Aléatoires, Université Paris 7, 175 rue de Chevaleret, F-75013 Paris, France. Email: lepennec@math.jussieu.fr

[‡]Laboratoire de Mathématique, U.M.R. C.N.R.S. 8628, Université Paris Sud, 91405 Orsay Cedex, France and Département de Mathématiques et Applications, U.M.R. C.N.R.S. 8553, ENS-Paris, 45 Rue d'Ulm, 75230 Paris Cedex 05, France, Email: vincent.rivoirard@math.u-psud.fr

where $\|\cdot\|_{\ell_\infty}$ is the sup-norm in $\mathbb{R}^M$, $\|\cdot\|_{\ell_1}$ is the ℓ_1 norm in $\mathbb{R}^M$, and η is a regularization parameter. A natural companion of this estimator is the Lasso procedure or more precisely its relaxed form

$$\hat{\lambda}^L = \arg\min_{\lambda \in \mathbb{R}^M} \left\{ \frac{1}{2}\|A\lambda - Y\|_{\ell_2}^2 \ + \eta\|\lambda\|_{\ell_1} \right\},$$

where η plays exactly the exact same role as for the Dantzig estimator. This ℓ_1 penalized method is also called *basis pursuit* in signal processing (see [14, 15]).

Candès and Tao [13] have obtained a bound for the ℓ_2 risk of the estimator $\hat{\lambda}^D$, with large probability, under a global condition on the matrix A (the Restricted Isometry Property) and a sparsity assumption on λ_0, even for $M \geq n$. Bickel et al. [3] have obtained oracle inequalities and bounds of the ℓ_p loss for both estimators under weaker assumptions. Actually, Bickel et al. [3] deal with the non parametric regression framework in which one observes

$$Y_i = f(x_i) + e_i, \quad i = 1, \dots, n$$

where f is an unknown function while $(x_i)_{i=1,\dots,n}$ are known design points and $(e_i)_{i=1,\dots,n}$ is a noise vector. There is no intrinsic matrix A in this problem but for any dictionary of functions $\Upsilon = (\varphi_m)_{m=1,\dots,M}$ one can search f as a weighted sum f_λ of elements of Υ

$$f_\lambda = \sum_{m=1}^{M} \lambda_m \varphi_m$$

and introduce the matrix $A = (\varphi_m(x_i))_{i,m}$, which summarizes the information on the dictionary and on the design. Notice that if there exists λ_0 such that $f = f_{\lambda_0}$ then the model can be rewritten exactly as the classical linear model. However, if it is not the case and if a model bias exists, the Dantzig and Lasso procedures can be after all applied under similar assumptions on A. Oracle inequalities are obtained for which approximation theory plays an important role in [3, 8, 9, 29].

Let us also mention that in various settings, under various assumptions on the matrix A (or more precisely on the associated Gram matrix $G = A^T A$), properties of these estimators have been established for subset selection (see [11, 20, 22, 23, 30, 31]) and for prediction (see [3, 19, 20, 23, 32]).

1.1 Our goals and results

We consider in this paper the density estimation framework already studied for the Lasso estimate by Bunea et al [7, 10] and van de Geer [29]. Namely, our goal is to estimate f_0, an unknown density function, by using the observations of an n-sample of variables $X_1, \dots, X_n$ of density f_0. As in the non parametric regression setting, we introduce a dictionary of functions $\Upsilon = (\varphi_m)_{m=1,\dots,M}$, and search again estimates of f_0 as linear combinations f_λ of the dictionary functions. We rely on the Gram matrix associated with Υ and on the empirical scalar products of f_0 with φ_m

$$\hat{\beta}_m = \frac{1}{n}\sum_{i=1}^{n} \varphi_m(X_i).$$

The Dantzig estimate $\hat{f}^D$ is then obtained by minimizing $\|\lambda\|_{\ell_1}$ over the set of parameters λ satisfying the adaptive Dantzig constraint:

$$\forall\, m \in \{1, \dots. M\}, \quad |(G\lambda)_m - \hat{\beta}_m| \leq \eta_{\gamma,m}$$

where for $m \in \{1, \dots, M\}$, $(G\lambda)_m$ is the scalar product of f_λ with φ_m,

$$\eta_{\gamma,m} = \sqrt{\frac{2\tilde{\sigma}_m^2 \gamma \log M}{n}} + \frac{2\|\varphi_m\|_\infty \gamma \log M}{3n},$$

$\tilde{\sigma}_m^2$ is a sharp estimate of the variance of $\hat{\beta}_m$ and γ is a constant to be chosen. Section 2 gives precise definitions and heuristics for using this constraint. We just mention here that $\eta_{\gamma,m}$ comes from sharp concentration inequalities to give tight constraints. Our idea is that if f_0 can be decomposed on Υ as

$$f_0 = \sum_{m=1}^{M} \lambda_{0,m} \varphi_m,$$

then we force the set of feasible parameters λ to contain λ_0 with large probability and to be as small as possible. Significant improvements in practice are expected.

Our goals in this paper are mainly twofold. First, we aim at establishing sharp oracle inequalities under very mild assumptions on the dictionary. Our starting point is that most of the papers in the literature assume that the functions of the dictionary are bounded by a constant independent of M and n, which constitutes a strong limitation, in particular for dictionaries based on histograms or wavelets (see for instance [6], [7], [8], [9], [11] or [29]). Such assumptions on the functions of Υ will not be considered in our paper. Likewise, our methodology does not rely on the knowledge of $\|f_0\|_\infty$ that can even be infinite (as noticed by Birgé [4] for the study of the integrated $\mathbb{L}_2$-risk, most of the papers in the literature typically assume that the sup-norm of the unknown density is finite with a known or estimated bound for this quantity). Finally, let us mention that, in contrast with what Bunea et al [10] did, we obtain oracle inequalities with leading constant 1, and furthermore these are established under much weaker assumptions on the dictionary than in [10].

The second goal of this paper deals with the problem of calibrating the so-called *Dantzig constant* γ: how should this constant be chosen to obtain good results in both theory and practice? Most of the time, for Lasso-type estimators, the regularization parameter is of the form $a\sqrt{\frac{\log M}{n}}$ with a a positive constant (see [3], [7], [6], [9], [12], [20] or [23] for instance). These results are obtained with large probability that depends on the tuning coefficient a. In practice, it is not simple to calibrate the constant a. Unfortunately, most of the time, the theoretical choice of the regularization parameter is not suitable for practical issues. This fact is true for Lasso-type estimates but also for many algorithms for which the regularization parameter provided by the theory is often too conservative for practical purposes (see [18] who clearly explains and illustrates this point for their thresholding procedure). So, one of the main goals of this paper is to fill the gap between the optimal parameter choice provided by theoretical results on the one hand and by a simulation study on the other hand. Only a few papers are devoted to this problem. In the model selection setting, the issue of calibration has been addressed by Birgé and Massart [5] who considered ℓ_0-penalized estimators in a Gaussian homoscedastic regression framework and showed that there exists a minimal penalty in the sense that taking smaller penalties leads to inconsistent estimation procedures. Arlot and Massart [1] generalized these results for non-Gaussian or heteroscedastic data and Reynaud-Bouret and Rivoirard [26] addressed this question for thresholding rules in the Poisson intensity framework.

Now, let us describe our results. By using the previous data-driven Dantzig constraint, oracle inequalities are derived under local conditions on the dictionary that are valid under classical assumptions on the structure of the dictionary. We extensively discuss these assumptions and we show their own interest in the context of the paper. Each term of these oracle inequalities is easily interpretable. Classical results are recovered when we further assume:

$$\|\varphi_m\|_\infty^2 \leq c_1 \left(\frac{n}{\log M}\right) \|f_0\|_\infty ,$$

where c_1 is a constant. This assumption is very mild and, unlike in classical works, allows to consider dictionaries based on wavelets. Then, relying on our Dantzig estimate, we build an adaptive Lasso procedure whose oracle performances are similar. This illustrates the closeness between Lasso and Dantzig-type estimates.

Our results are proved for $\gamma > 1$. For the theoretical calibration issue, we study the performance of our procedure when $\gamma < 1$. We show that in a simple framework, estimation of the straightforward signal $f_0 = \mathbf{1}_{[0,1]}$ cannot be performed at a convenient rate of convergence when $\gamma < 1$. This result proves that the assumption $\gamma > 1$ is thus not too conservative.

Finally, a simulation study illustrates how dictionary-based methods outperform classical ones. More precisely, we show that our Dantzig and Lasso procedures with $\gamma > 1$, but close to 1, outperform classical ones, such as simple histogram procedures, wavelet thresholding or Dantzig procedures based on the knowledge of $\|f_0\|_\infty$ and less tight Dantzig constraints.

1.2 Outlines

Section 2 introduces the density estimator of f_0 whose theoretical performances are studied in Section 3. Section 4 studies the Lasso estimate proposed in this paper. The calibration issue is studied in Section 5.1 and numerical experiments are performed in Section 5.2. Finally, Section 6 is devoted to the proofs of our results.

2 The Dantzig estimator of the density f_0

As said in Introduction, our goal is to build an estimate of f_0 as a linear combination of functions of $\Upsilon = (\varphi_m)_{m=1,\dots,M}$, where we assume without any loss of generality that, for any m, $\|\varphi_m\|_2 = 1$:

$$f_\lambda = \sum_{m=1}^{M} \lambda_m \varphi_m.$$

For this purpose, we naturally rely on natural estimates of the $\mathbb{L}_2$-scalar products between f_0 and the φ_m's. So, for $m \in \{1, \dots, M\}$, we set

$$\beta_{0,m} = \int \varphi_m(x) f_0(x) dx, \tag{1}$$

and we consider its empirical counterpart

$$\hat{\beta}_m = \frac{1}{n} \sum_{i=1}^{n} \varphi_m(X_i) \tag{2}$$

that is an unbiased estimate of $\beta_{0,m}$. The variance of this estimate is $\mathrm{Var}(\hat{\beta}_m) = \frac{\sigma_{0,m}^2}{n}$ where

$$\sigma_{0,m}^2 = \int \varphi_m^2(x) f_0(x) dx - \beta_{0,m}^2. \tag{3}$$

Note also that for any λ and any m, the $\mathbb{L}_2$-scalar product between f_λ and φ_m can be easily computed:

$$\int \varphi_m(x) f_\lambda(x)dx = \sum_{m'=1}^{M} \lambda_{m'} \int \varphi_{m'}(x)\varphi_m(x)dx = (G\lambda)_m$$

where G is the Gram matrix associated to the dictionary Υ defined for any $1 \leq m, m' \leq M$ by

$$G_{m,m'} = \int \varphi_m(x)\varphi_{m'}(x)dx.$$

Any reasonable choice of λ should ensure that the coefficients $(G\lambda)_m$ are close to $\hat{\beta}_m$ for all m. Therefore, using Candès and Tao's approach, we define the Dantzig constraint:

$$\forall\, m \in \{1, \ldots .M\}, \qquad |(G\lambda)_m - \hat{\beta}_m| \leq \eta_{\gamma,m} \tag{4}$$

and the Dantzig estimate $\hat{f}^D$ by $\hat{f}^D = f_{\hat{\lambda}^{D,\gamma}}$ with

$$\hat{\lambda}^{D,\gamma} = \mathrm{argmin}_{\lambda\in\mathbb{R}^M} \|\lambda\|_{\ell_1} \quad \text{such that } \lambda \text{ satisfies the Dantzig constraint (4)},$$

where for $\gamma > 0$ and $m \in \{1, \ldots, M\}$,

$$\eta_{\gamma,m} = \sqrt{\frac{2\tilde{\sigma}_m^2 \gamma \log M}{n}} + \frac{2\|\varphi_m\|_\infty \gamma \log M}{3n}, \tag{5}$$

with

$$\tilde{\sigma}_m^2 = \hat{\sigma}_m^2 + 2\|\varphi_m\|_\infty \sqrt{\frac{2\hat{\sigma}_m^2 \gamma \log M}{n}} + \frac{8\|\varphi_m\|_\infty^2 \gamma \log M}{n} \tag{6}$$

and

$$\hat{\sigma}_m^2 = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\varphi_m(X_i) - \varphi_m(X_j))^2. \tag{7}$$

Note that $\eta_{\gamma,m}$ depends on the data, so the constraint (4) will be referred as the *adaptive Dantzig constraint* in the sequel. We now justify the introduction of the density estimate $\hat{f}^D$.

The definition of $\eta_{\lambda,\gamma}$ is based on the following heuristics. Given m, when there exists a constant $c_0 > 0$ such that $f_0(x) \geq c_0$ for x in the support of φ_m satisfying $\|\varphi_m\|_\infty^2 = o_n(n(\log M)^{-1})$, then, with large probability, the deterministic term of (5) is negligible with respect to the random one. In this case, the random term is the main one and we asymptotically derive

$$\eta_{\gamma,m} \approx \sqrt{2\gamma \log M \, \frac{\tilde{\sigma}_m^2}{n}}. \tag{8}$$

Having in mind that $\tilde{\sigma}_m^2/n$ is a convenient estimate for $\mathrm{Var}(\hat{\beta}_m)$ (see the proof of Theorem 1), the shape of the right hand term of the formula (8) looks like the bound proposed by Candès and Tao [13] to define the Dantzig constraint in the linear model. Actually, the deterministic term of (5) allows to get sharp concentration inequalities. As often done in the literature, instead of estimating $\mathrm{Var}(\hat{\beta}_m)$, we could use the inequality

$$\mathrm{Var}(\hat{\beta}_m) = \frac{\sigma_{0,m}^2}{n} \leq \frac{\|f_0\|_\infty}{n}$$

and we could replace $\tilde{\sigma}_m^2$ with $\|f_0\|_\infty$ in the definition of the $\eta_{\gamma,m}$. But this requires a strong assumption: f_0 is bounded and $\|f_0\|_\infty$ is known. In our paper, $\mathrm{Var}(\hat{\beta}_m)$ is estimated, which allows not to impose these conditions. More precisely, we slightly overestimate $\sigma_{0,m}^2$ to control large deviation terms and this is the reason why we introduce $\tilde{\sigma}_m^2$ instead of using $\hat{\sigma}_m^2$, an unbiased estimate of $\sigma_{0,m}^2$. Finally, γ is a constant that has to to be suitably calibrated and plays a capital role in practice.

The following result justifies previous heuristics by showing that, if $\gamma > 1$, with high probability, the quantity $|\hat{\beta}_m - \beta_{0,m}|$ is smaller than $\eta_{\gamma,m}$ for all m. The parameter $\eta_{\gamma,m}$ with γ close to 1 can be viewed as the "smallest" quantity that ensures this property.

Theorem 1. *Let us assume that M satisfies*

$$n \leq M \leq \exp(n^{\delta}) \tag{9}$$

for $\delta < 1$. Let $\gamma > 1$. Then, for any $\varepsilon > 0$, there exists a constant $C_1(\varepsilon, \delta, \gamma)$ depending on ε, δ and γ such that

$$\mathbb{P}\left(\forall m \in \{1, \dots, M\}, \quad |\beta_{0,m} - \hat{\beta}_m| \geq \eta_{\gamma,m}\right) \leq C_1(\varepsilon, \delta, \gamma) M^{1-\frac{\gamma}{1+\varepsilon}}.$$

In addition, there exists a constant $C_2(\delta, \gamma)$ depending on δ and γ such that

$$\mathbb{P}\left(\forall m \in \{1, \dots, M\}, \quad \eta_{\gamma,m}^{(-)} \leq \eta_{\gamma,m} \leq \eta_{\gamma,m}^{(+)}\right) \leq C_2(\delta, \gamma) M^{1-\gamma}$$

where, for $m \in \{1, \dots, M\}$,

$$\eta_{\gamma,m}^{(-)} = \sigma_{0,m}\sqrt{\frac{8\gamma \log M}{7n}} + \frac{2\|\varphi_m\|_\infty \gamma \log M}{3n}$$

and

$$\eta_{\gamma,m}^{(+)} = \sigma_{0,m}\sqrt{\frac{16\gamma \log M}{n}} + \frac{10\|\varphi_m\|_\infty \gamma \log M}{n}.$$

This result is proved in Section 6.1. The first part is a sharp concentration inequality proved by using Bernstein type controls. The second part of the theorem proves that, up to constants depending on γ, $\eta_{\gamma,m}$ is of order $\sigma_{0,m}\sqrt{\frac{\log M}{n}} + \|\varphi_m\|_\infty \frac{\log M}{n}$ with high probability. Note that the assumption $\gamma > 1$ is essential to obtain probabilities going to 0.

Finally, let $\lambda_0 = (\lambda_{0,m})_{m=1,\dots,M} \in \mathbb{R}^M$ such that

$$P_\Upsilon f_0 = \sum_{m=1}^{M} \lambda_{0,m}\varphi_m$$

where P_Υ is the projection on the space spanned by Υ. We have

$$(G\lambda_0)_m = \int (P_\Upsilon f_0)\varphi_m = \int f_0 \varphi_m = \beta_{0,m}.$$

So, Theorem 1 proves that λ_0 satisfies the adaptive Dantzig constraint (4) with probability larger than $1 - C_1(\varepsilon, \delta, \gamma) M^{1-\frac{\gamma}{1+\varepsilon}}$ for any $\varepsilon > 0$. Actually, we force the set of parameters λ satisfying the adaptive Dantzig constraint to contain λ_0 with large probability and to be as small as possible. Therefore, $\hat{f}^D = f_{\hat{\lambda}^{D,\gamma}}$ is a good candidate among sparse estimates linearly decomposed on Υ for estimating f_0.

We mention that Assumption (9) can be relaxed and we can take $M < n$ provided the definition of $\eta_{\gamma,m}$ is modified.

3 Results for the Dantzig estimators

In the sequel, we will denote $\hat{\lambda}^D = \hat{\lambda}^{D,\gamma}$ to simplify the notations, but the Dantzig estimator $\hat{f}^D$ still depends on γ. Moreover, we assume that (9) is true and we denote the vector $\eta_\gamma = (\eta_{\gamma,m})_{m=1,\dots,M}$ considered with the Dantzig constant $\gamma > 1$.

3.1 The main result under local assumptions

Let us state the main result of this paper. For any $J \subset \{1,\dots,M\}$, we set $J^C = \{1,\dots,M\} \setminus J$ and define λ_J the vector which has the same coordinates as λ on J and zero coordinates on J^C. We introduce a local assumption indexed by a subset J_0.

- **Local Assumption** Given $J_0 \subset \{1,\dots,M\}$, for some constants $\kappa_{J_0} > 0$ and $\mu_{J_0} > 0$ depending on J_0, we have for any λ,

$$\|f_\lambda\|_2 \geq \kappa_{J_0}\|\lambda_{J_0}\|_{\ell_2} - \mu_{J_0}\left(\|\lambda_{J_0^C}\|_{\ell_1} - \|\lambda_{J_0}\|_{\ell_1}\right)_+ . \qquad (LA(J_0,\kappa_{J_0},\mu_{J_0}))$$

We obtain the following oracle type inequality without any assumption on f_0.

Theorem 2. *Let $J_0 \subset \{1,\dots,M\}$ be fixed. We suppose that $(LA(J_0,\kappa_{J_0},\mu_{J_0}))$ holds. Then, with probability at least $1 - C_1(\varepsilon,\delta,\gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$,*

$$\|\hat{f}^D - f_0\|_2^2 \leq \inf_{\lambda\in\mathbb{R}^M}\left\{\|f_\lambda - f_0\|_2^2 + \beta\frac{\Lambda(\lambda,J_0^c)^2}{|J_0|}\left(1 + \frac{2\mu_{J_0}\sqrt{|J_0|}}{\kappa_{J_0}}\right)^2 + 16|J_0|\left(\frac{1}{\beta} + \frac{1}{\kappa_{J_0}^2}\right)\|\eta_\gamma\|_{\ell_\infty}^2\right\}, \qquad (10)$$

with

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

Let us comment each term of the right hand side of (10). The first term is an approximation term which measures the closeness between f_0 and f_λ. This term can vanish if f_0 can be decomposed on the dictionary. The second term is a price to pay when either λ is not supported by the subset J_0 considered or it does not satisfy the condition $\|\hat{\lambda}^D\|_{\ell_1} \leq \|\lambda\|_{\ell_1}$ which holds as soon as λ satisfy the adaptive Dantzig constraint. Finally, the last term, which does not depend on λ, can be viewed as a variance term corresponding to the estimation on the subset J_0. Indeed, remember that $\eta_{\gamma,m}$ relies on an estimate of the variance of $\hat{\beta}_m$. Furthermore, we have with high probability:

$$\|\eta_\gamma\|_{\ell_\infty}^2 \quad \leq \quad 2\left(\frac{16\sigma_{0,m}^2\gamma\log M}{n} + \left(\frac{10\|\varphi_m\|_\infty\gamma\log M}{n}\right)^2\right).$$

So, if f_0 is bounded then, $\sigma_{0,m}^2 \leq \|f_0\|_\infty$ and if there exists a constant c_1 such that for any m,

$$\|\varphi_m\|_\infty^2 \leq c_1\left(\frac{n}{\log M}\right)\|f_0\|_\infty, \qquad (11)$$

(which is true for instance for a bounded dictionary), then

$$\|\eta_\gamma\|_{\ell_\infty}^2 \leq C\|f_0\|_\infty\frac{\log M}{n},$$

(where C is a constant depending on γ and c_1) and tends to 0 when n goes to ∞. We obtain thus the following result.

Corollary 1. *Let $J_0 \subset \{1, \dots, M\}$ be fixed. We suppose that $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ holds. If (11) is satisfied then, with probability at least $1 - C_1(\varepsilon, \delta, \gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$, for any λ that satisfies the adaptive Dantzig constraint*

$$\|\hat{f}^D - f_0\|_2^2 \le \|f_\lambda - f_0\|_2^2 + \beta c_2 (1 + \kappa_{J_0}^{-2}\mu_{J_0}^2 |J_0|) \frac{\|\lambda_{J_0^C}\|_{\ell_1}^2}{|J_0|} + c_3(\beta^{-1} + \kappa_{J_0}^{-2})|J_0| \|f_0\|_\infty \frac{\log M}{n}, \quad (12)$$

where c_2 is an absolute constant and c_3 depends on c_1 and γ.

The parameter β calibrates the weights given for the bias and variance terms. Remark that if $f_0 = f_{\lambda_0}$ and if $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ holds with $J_0 = J_{\lambda_0}$, under (11), the proof of Theorem 2 yields the more classical inequality

$$\|\hat{f}^D - f_0\|_2^2 \le C' |J_0| \|f_0\|_\infty \frac{\log M}{n},$$

where $C' = c_3 \kappa_{J_0}^{-2}$, with at least the same probability $1 - C_1(\varepsilon, \delta, \gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$.

Assumption $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ is local, in the sense that the constants κ_{J_0} and μ_{J_0} (or their mere existence) may highly depend on the subset J_0. For a given λ, the best choice for J_0 in Inequalities (10) and (12) depends thus on the interaction between these constants and the value of λ itself. Note that the assumptions of Theorem 2 are reasonable as the next section gives conditions for which Assumption $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ holds simultaneously with the same constant κ and μ for all subsets J_0 of the same size.

3.2 Results under global assumptions

As usual, when $M > n$, properties of the Dantzig estimate can be derived from assumptions on the structure of the dictionary Υ. For $l \in \mathbb{N}$, we denote

$$\phi_{\min}(l) = \min_{|J| \le l} \min_{\substack{\lambda \in \mathbb{R}^M \\ \lambda_J \ne 0}} \frac{\|f_{\lambda_J}\|_2^2}{\|\lambda_J\|_{\ell_2}^2} \qquad \text{and} \qquad \phi_{\max}(l) = \max_{|J| \le l} \max_{\substack{\lambda \in \mathbb{R}^M \\ \lambda_J \ne 0}} \frac{\|f_{\lambda_J}\|_2^2}{\|\lambda_J\|_{\ell_2}^2}.$$

These quantities correspond to the "restricted" eigenvalues of the Gram matrix G. Assuming that $\phi_{\min}(l)$ and $\phi_{\max}(l)$ 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 correlations

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

Small values of $\theta_{l,l'}$ mean that two disjoint sets of columns of G with cardinality less than l and l' span nearly orthogonal spaces. We will use one of the following assumptions considered in [3].

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

$$\phi_{\min}(2s) > \theta_{s,2s}. \qquad \text{(A1(s))}$$

  Oracle inequalities of the Dantzig selector were established under this assumption in the parametric linear model by Candès and Tao in [13]. It was also considered by Bunea, Ritov and Tsybakov [3] for non-parametric regression and for the Lasso estimate. The next assumption, proposed in [3], constitutes an alternative to Assumption 1.

- **Assumption 2** For some integers s and l such that

$$1 \le s \le \frac{M}{2}, \quad l \ge s \quad \text{and} \quad s + l \le M, \tag{13}$$

we have

$$l\phi_{\min}(s+l) > s\phi_{\max}(l). \tag{A2(s,l)}$$

If Assumption 2 is true for s and l such that $l \gg s$, then Assumption 2 means that $\phi_{\min}(l)$ cannot decrease at a rate faster than l^{-1} and this condition is related to the "incoherent designs" condition stated in [23].

In the sequel, we set, under Assumption 1,

$$\kappa_1(s) = \sqrt{\phi_{\min}(2s)}\left(1 - \frac{\theta_{s,2s}}{\phi_{\min}(2s)}\right) > 0, \quad \mu_1(s) = \frac{\theta_{s,2s}}{\sqrt{s\phi_{\min}(2s)}}$$

and under Assumption 2,

$$\kappa_2(s,l) = \sqrt{\phi_{\min}(s+l)}\left(1 - \sqrt{\frac{s\phi_{\max}(l)}{l\phi_{\min}(s+l)}}\right) > 0, \quad \mu_2(s,l) = \sqrt{\frac{\phi_{\max}(l)}{l}}.$$

Now, to apply Theorem 2, we need to check $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ for some some subset J_0 of $\{1, \dots, M\}$. Either Assumption 1 or Assumption 2 implies this assumption. Indeed, we have the following result.

Proposition 1. *Let s and l two integers satisfying (13). We suppose that (A1(s)) or (A2(s,l)) is true. Let $J_0 \subset \{1, \dots, M\}$ of size $|J_0| = s$ and $\lambda \in \mathbb{R}^M$, then we have*

$$\|f_\lambda\|_2 \ge \kappa\|\lambda_{J_0}\|_{\ell_2} - \mu\left(\|\lambda_{J_0^C}\|_{\ell_1} - \|\lambda_{J_0}\|_{\ell_1}\right)_+$$

with $\kappa = \kappa_1(s)$ and $\mu = \mu_1(s)$ under (A1(s)) (respectively $\kappa = \kappa_2(s,l)$ and $\mu = \mu_2(s,l)$ under (A2(s,l)). If (A1(s)) and (A2(s,l)) are both satisfied, $\kappa = \max(\kappa_1(s), \kappa_2(s,l))$ and $\mu = \min(\mu_1(s), \mu_2(s,l))$.

Proposition 1 proves that Theorem 2 can be applied under Assumptions 1 or 2. In addition, the constants κ_{J_0} and μ_{J_0} only depend on $|J_0|$. From Theorem 2, we deduce the following result.

Theorem 3. *Let s and l two integers satisfying (13). We suppose that (A1(s)) or (A2(s,l)) is true. Then, with probability at least $1 - C_1(\varepsilon, \delta, \gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$,*

$$\|\hat{f}^D - f_0\|_2^2 \le \inf_{\lambda \in \mathbb{R}^M} \inf_{\substack{J_0 \subset \{1,\dots,M\} \\ |J_0| = s}} \left\{ \|f_\lambda - f_0\|_2^2 + \beta\frac{\Lambda(\lambda, J_0^c)^2}{s}\left(1 + \frac{2\mu\sqrt{s}}{\kappa}\right)^2 + 16s\left(\frac{1}{\beta} + \frac{1}{\kappa^2}\right)\|\eta_\gamma\|_{\ell_\infty}^2 \right\}$$

where

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

Remark that the best subset J_0 of cardinal s in Theorem 3 can be easily chosen for a given λ: it is given by the set of the s largest coordinates of λ. This was not necessarily the case in Theorem 2 for which a different subset may give a better local condition and then may provide a smaller bound. If we further assume the mild assumption (11) on the sup norm of the dictionary introduced in the previous section, we deduce the following result.

Corollary 2. *Let s and l two integers satisfying (13). We suppose that (A1(s)) or (A2(s,l)) is true. If (11) is satisfied, with probability at least $1 - C_1(\varepsilon, \delta, \gamma) M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$, any λ that satisfies the adaptive Dantzig constraint and for the best subset J_0 of cardinal s (that corresponds to the s largest coordinates of λ in absolute value),*

$$\|\hat{f}^D - f_0\|_2^2 \leq \|f_\lambda - f_0\|_2^2 + \beta c_2 (1 + \kappa^{-2}\mu^2 s) \frac{\|\lambda_{J_0^C}\|_{\ell_1}^2}{s} + c_3(\beta^{-1} + \kappa^{-2}) s \|f_0\|_\infty \frac{\log M}{n}, \tag{14}$$

where c_2 is an absolute constant and c_3 depends on c_1 and γ.

Note that, when λ is s-sparse so that $\lambda_{J_0^C} = 0$, the oracle inequality (14) corresponds to the classical oracle inequality obtained in parametric frameworks (see [12] or [13] for instance) or in non-parametric settings. See, for instance [6], [7], [8], [9], [11] or [29] but in these works, the functions of the dictionary are assumed to be bounded by a constant independent of M and n. So, the adaptive Dantzig estimate requires weaker conditions since under (11), $\|\varphi_m\|_\infty$ can go to ∞ when n grows. This point is capital for practical purposes, in particular when wavelet bases are considered.

4 Connections between the Dantzig and Lasso estimates

We show in this section the strong connections between Lasso and Dantzig estimates, which has already been illustrated in [3] for non-parametric regression models. By choosing convenient random weights depending on η_γ for ℓ_1-minimization, the Lasso estimate satisfies the adaptive Dantzig constraint. More precisely, we consider the Lasso estimator given by the solution of the following minimization problem

$$\hat{\lambda}^{L,\gamma} = \mathrm{argmin}_{\lambda \in \mathbb{R}^M} \left\{ R(\lambda) + 2 \sum_{m=1}^{M} \eta_{\gamma,m} |\lambda_m| \right\}, \tag{15}$$

where

$$R(\lambda) = \|f_\lambda\|_2^2 - \frac{2}{n} \sum_{i=1}^{n} f_\lambda(X_i).$$

Note that $R(\cdot)$ is the quantity minimized in unbiased estimation of the risk. For simplifications, we write $\hat{\lambda}^L = \hat{\lambda}^{L,\gamma}$. We denote $\hat{f}^L = f_{\hat{\lambda}^L}$. As said in Introduction, classical Lasso estimates are defined as the minimizer of expressions of the form

$$\left\{ R(\lambda) + 2\eta \sum_{m=1}^{M} |\lambda_m| \right\},$$

where η is proportional to $\sqrt{\frac{\log M}{n}}$. So, $\hat{\lambda}^L$ appears as a data-driven version of classical Lasso estimates.

The first order condition for the minimization of the expression given in (15) corresponds exactly to the adaptive Dantzig constraint and thus Theorem 3 always applies to $\hat{\lambda}^L$. Working along the lines of the proof of Theorem 3 (Replace f_λ by $\hat{f}^D$ and $\hat{f}^D$ by $\hat{f}^L$ in (26) and (27)), one can prove a slightly stronger result.

Theorem 4. *Let us assume that assumptions of Theorem 3 are true. Let $J_0 \subset \{1, \ldots, M\}$ of size $|J_0| = s$. Then, with probability at least $1 - C_1(\varepsilon, \delta, \gamma) M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$,*

$$\left| \|\hat{f}^D - f_0\|_2^2 - \|\hat{f}^L - f_0\|_2^2 \right| \leq \beta \frac{\|\hat{\lambda}^L_{J_0^C}\|_{\ell_1}^2}{s} \left(1 + \frac{2\mu\sqrt{s}}{\kappa}\right)^2 + 16 s \left(\frac{1}{\beta} + \frac{1}{\kappa^2}\right) \|\eta_\gamma\|_{\ell_\infty}^2.$$

To extend this theoretical result, numerical performances of the Dantzig and Lasso estimates will be compared in Section 5.2.

5 Calibration and numerical experiments

5.1 The calibration issue

In this section, we consider the problem of calibrating previous estimates. In particular, we prove that the sufficient condition $\gamma > 1$ is "almost" a necessary condition since we derive a special and very simple framework in which Lasso and Dantzig estimates cannot achieve the optimal rate if $\gamma < 1$ ("almost" means that the case $\gamma = 1$ remains an open question). Let us describe this simple framework. The dictionary Υ considered in this section is the orthonormal Haar system:

$$\Upsilon = \left\{\phi_{jk} : \quad -1 \leq j \leq j_0,\ 0 \leq k < 2^j\right\},$$

with $\phi_{-10} = \mathbf{1}_{[0,1]}$, $2^{j_0+1} = n$, and for $0 \leq j \leq j_0$, $0 \leq k \leq 2^j - 1$,

$$\phi_{jk} = 2^{j/2}\left(1_{[k/2^j,(k+0.5)/2^j]} - 1_{[(k+0.5)/2^j,(k+1)/2^j]}\right).$$

In this case, $M = n$. In this setting, since functions of Υ are orthonormal, the Gram matrix G is the identity. Thus, the Lasso and Dantzig estimates both correspond to the soft thresholding rule:

$$\hat{f}^D = \hat{f}^L = \sum_{m=1}^{M} \mathrm{sign}(\hat{\beta}_m)\left(|\hat{\beta}_m| - \eta_{\gamma,m}\right) 1_{\{|\hat{\beta}_m| > \eta_{\gamma,m}\}} \varphi_m.$$

Now, our goal is to estimate $f_0 = \phi_{-10} = \mathbf{1}_{[0,1]}$ by using $\hat{f}^D$ depending on γ and to show the influence of this constant. Unlike previous results stated in probability, we consider the expectation of the $\mathbb{L}_2$-risk:

Theorem 5. *On the one hand, if $\gamma > 1$, there exists a constant C such that*

$$\mathbb{E}\|\hat{f}^D - f_0\|_2^2 \leq \frac{C \log n}{n}, \tag{16}$$

On the other hand, if $\gamma < 1$, there exists a constant c and $\delta < 1$ such that

$$\mathbb{E}\|\hat{f}^D - f_0\|_2^2 \geq \frac{c}{n^\delta}. \tag{17}$$

This result shows that choosing $\gamma < 1$ is a bad choice in our setting. Indeed, in this case, the Lasso and Dantzig estimates cannot estimate a very simple signal ($f_0 = \mathbf{1}_{[0,1]}$) at a convenient rate of convergence.

A small simulation study is carried out to strengthen this theoretical asymptotic result. Performing our estimation procedure 100 times, we compute the average risk $\overline{R_n}(\gamma)$ for several values of the Dantzig constant γ and several values of n. This computation is summarized in Figure 1 which displays the logarithm of $\overline{R_n}(\gamma)$ for $n = 2^J$ with, from top to bottom, $J = 4, 5, 6, \ldots, 13$ on a grid of γ's around 1. To discuss our results, we denote by $\gamma_{\min}(n)$ the best γ: $\gamma_{\min}(n) = \mathrm{argmin}_{\gamma>0}\overline{R_n}(\gamma)$. We note that $1/2 \leq \gamma_{\min}(n) \leq 1$ for all values of n, with $\gamma_{\min}(n)$ getting closer to 1 as n increases. Taking γ too small strongly deteriorates the performance while a value close to 1 ensures a risk withing a factor 2 of the optimal risk. The assumption $\gamma > 1$ giving a theoretical control on the quadratic error is thus not too conservative. Following these results, we set $\gamma = 1.01$ in our numerical experiments in the next subsection.

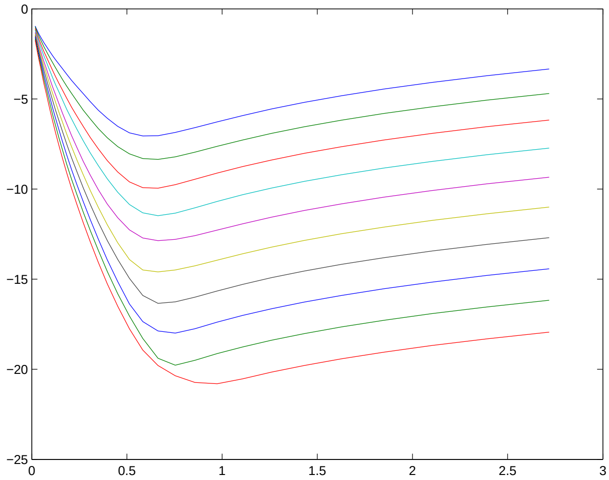

Figure 1: Graphs of $\gamma \mapsto \log_2(\overline{R_n}(\gamma))$ for $n = 2^J$ with, from top to bottom, $J = 4, 5, 6, \ldots, 13$

5.2 Numerical experiments

In this section, we present our numerical experiments with the Dantzig density estimator and their results. We test our estimator with a collection of 6 dictionaries, 4 densities described below and for 2 sample sizes. We compare our procedure with the adaptive Lasso introduced in Section 4 and with a non adaptive Dantzig estimator. We also consider a two-step estimation procedure, proposed by Candès and Tao [13], which improves the numerical results.

The numerical scheme for a given dictionary $\Upsilon = (\varphi_m)_{m=1,\ldots,M}$ and a sample $(X_i)_{i=1,\ldots,n}$ is the following.

1. Compute $\hat{\beta}_m$ for all m,
2. Compute $\hat{\sigma}_m^2$,
3. Compute $\eta_{\gamma,m}$ as defined in (5) by
$$\eta_{\gamma,m} = \sqrt{\frac{2\tilde{\sigma}_m^2 \gamma \log M}{n}} + \frac{2\|\varphi_m\|_\infty \gamma \log M}{3n},$$
with
$$\tilde{\sigma}_m^2 = \hat{\sigma}_m^2 + 2\|\varphi_m\|_\infty \sqrt{\frac{2\hat{\sigma}_m^2 \gamma \log M}{n}} + \frac{8\|\varphi_m\|_\infty^2 \gamma \log M}{n}$$
and $\gamma = 1.01$.
4. Compute the coefficients $\hat{\lambda}^{D,\gamma}$ of the Dantzig estimate, $\hat{\lambda}^{D,\gamma} = \mathrm{argmin}_{\lambda \in \mathbb{R}^M} \|\lambda\|_{\ell_1}$ such that λ satisfies the Dantzig constraint (4)
$$\forall\, m \in \{1, \ldots . M\}, \qquad |(G\lambda)_m - \hat{\beta}_m| \leq \eta_{\gamma,m}$$
with the homotopy-path-following method proposed by Asif and Romberg [2],
5. Compute the Dantzig estimate $\hat{f}^{D,\gamma} = \sum_{m=1}^{M} \hat{\lambda}_m^{D,\gamma} \phi_m$.

Note that we have implicitly assumed that the Gram matrix G used in the definition of the Dantzig constraint has been precomputed.

For the Lasso estimator, the Dantzig minimization of step 4 is replaced by the Lasso minimization (15)

$$\hat{\lambda}^{L,\gamma} = \mathrm{argmin}_{\lambda \in \mathbb{R}^M} \left\{ R(\lambda) + 2 \sum_{m=1}^{M} \eta_{\gamma,m} |\lambda_m| \right\},$$

which is solved using the LARS algorithm. The non adaptive Dantzig estimate is obtained by replacing $\tilde{\sigma}_m^2$ in step 3 by $\|f_0\|_\infty$. The two-step procedure of Candès and Tao adds a least-square step between step 4 and step 5. More precisely, let $\hat{J}^{D,\gamma}$ be the support of the estimate $\hat{\lambda}^{D,\gamma}$. This defines a subset of the dictionary on which the density is regressed

$$\left(\hat{\lambda}^{D+LS,\gamma}\right)_{\hat{J}^{D,\gamma}} = G^{-1}_{\hat{J}^{D,\gamma}} (\hat{\beta}_m)_{\hat{J}^{D,\gamma}}$$

where $G_{\hat{J}^{D,\gamma}}$ is the submatrix of G corresponding to the subset chosen. The values of $\hat{\lambda}^{D+LS,\gamma}$ outside $\hat{J}^{D,\gamma}$ are set to 0 and $\hat{f}^{D+LS,\gamma}$ is set accordingly.

We describe now the dictionaries we consider. We focus numerically on densities defined on the interval $[0,1]$ so we use dictionaries adapted to this setting. The first four are orthonormal systems, which are used as a benchmark, while the last two are "real" dictionaries. More precisely, our dictionaries are

- the Fourier basis with $M = n+1$ elements (denoted "Fou"),
- the histogram collection with the classical number $\sqrt{n}/2 \leq M = 2^{j_0} < \sqrt{n}$ of bins (denoted "Hist"),
- the Haar wavelet basis with maximal resolution $n/2 < M = 2^{j_1} < n$ and thus $M = 2^{j_1}$ elements (denoted "Haar"),
- the more regular Daubechies 6 wavelet basis with maximal resolution $n/2 \leq M = 2^{j_1} < n$ and thus $M = 2^{j_1}$ elements (denoted "Wav"),
- the dictionary made of the union of the Fourier basis and the histogram collection and thus comprising $M = n + 1 + 2^{j_0}$ elements. (denoted "Mix"),
- the dictionary which is the union of the Fourier basis, the histogram collection and the Haar wavelets of resolution greater than 2^{j_0} comprising $M = n+1+2^{j_1}$ elements (denoted "Mix2").

The orthonormal families we have chosen are often used by practitioners. Our dictionaries combine very different orthonormal families, sine and cosine with bins or Haar wavelets, which ensures a sufficiently incoherent design.

We test the estimators of the following 4 functions shown in Figure 2 (with their Dantzig and Dantzig+Least Square estimates with the "Mix2" dictionary):

- a very spiky density

$$f_1(t) = .47 \times (4t \times \mathbf{1}_{t \leq .5} + 4(1-t) \times \mathbf{1}_{t > .5}) + .53 \times \left(75 \times \mathbf{1}_{.5 \leq t \leq .5 + \frac{1}{75}}\right),$$

- a mix of Gaussian and Laplacian type densities

$$f_2(t) = .45 \times \left(\frac{e^{-(t-.45)^2/(2(.125)^2)}}{\int_0^1 e^{-(u-.45)^2/(2(.125)^2)} du} \right) + .55 \times \left(\frac{e^{20|t-.67|}}{\int_0^1 e^{20|u-.67|} du} \right),$$

- a mix of uniform densities on subintervals

$$f_3(t) = .25 \times \left(\frac{1}{.14} \mathbf{1}_{.33 \le t \le .47} \right) + .75 \times \left(\frac{1}{.16} \mathbf{1}_{.64 \le t \le .80} \right),$$

- a mix of a density easily described in the Fourier domain and a uniform density on a subinterval

$$f_4(t) = .45 \times (1 + .9\cos(2\pi t)) + .55 \times \left(\frac{1}{.16} \mathbf{1}_{.64 \le t \le .80} \right).$$

Boxplots of Figures 3 and 4 summarize our numerical experiments for $n = 500$ and $n = 2000$ and 100 repetitions of the procedures. The left column deals with the comparison between Dantzig and Lasso, the center column shows the effectiveness of our data driven constraint and the right column illustrates the improvement of the two-step method. As expected, Dantzig and Lasso estimators are strictly equivalent when the dictionary is orthonormal and very close otherwise. For both algorithms and most of the densities, the best solution appears to be the "Mix2" dictionary, except for the density f_1 where the Haar wavelets are better for $n = 500$. This shows that the dictionary approach yields an improvement over the classical basis approach. One observes also that the "Mix" dictionary is better than the best of its constituent, namely the Fourier basis and the histogram family, which corroborates our theoretical results. The adaptive constraints are much tighter than their non adaptive counterparts and yield to much better numerical results. Our last series of experiments shows the significant improvement obtained with the least square step. As hinted by Candès and Tao [13], this can be explained by the bias common to ℓ_1 methods which is partially removed by this final least square adjustment. Studying directly the performance of this estimator is a challenging task.

6 Proofs

6.1 Proof of Theorem 1

To prove the first part of Theorem 1, we fix $m \in \{1, \ldots, M\}$ and we set for any $i \in \{1, \ldots, n\}$,

$$W_i = \frac{1}{n} \left(\varphi_m(X_i) - \beta_{0,m} \right)$$

that satisfies almost surely

$$|W_i| \le \frac{2\|\varphi_m\|_\infty}{n}.$$

Then, we apply Bernstein's Inequality (see [21] on pages 24 and 26) with the variables W_i and $-W_i$: for any $u > 0$,

$$\mathbb{P}\left(|\hat{\beta}_m - \beta_{0,m}| \ge \sqrt{\frac{2\sigma_{0,m}^2 u}{n}} + \frac{2u\|\varphi_m\|_\infty}{3n} \right) \le 2e^{-u}. \tag{18}$$

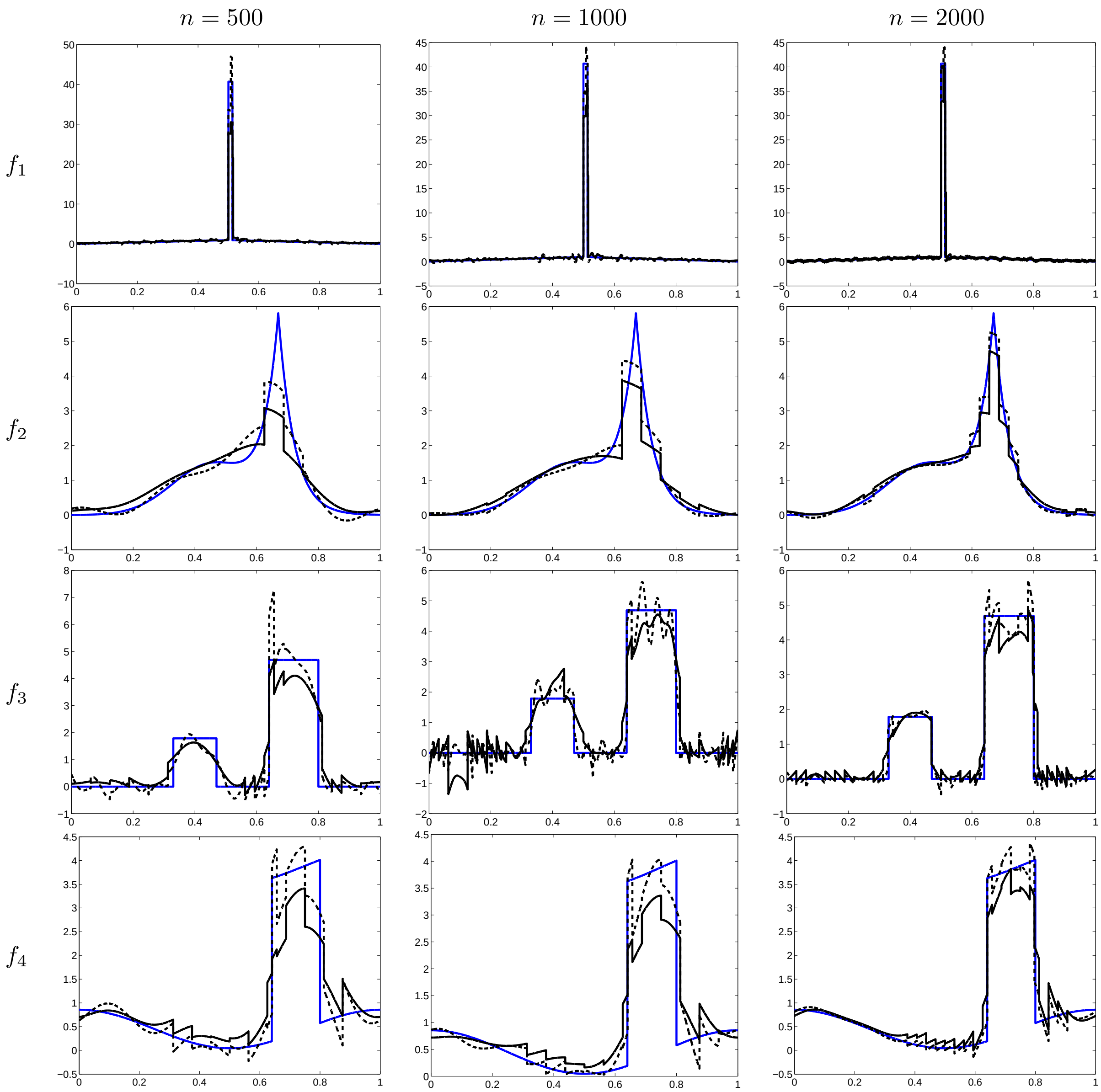

Figure 2: The different densities and their "Mix2" estimates. Densities are plotted in blue while their estimates are plotted in black. The full line corresponds to the adaptive Dantzig studied in this paper while the dotted line corresponds to its least square variant.

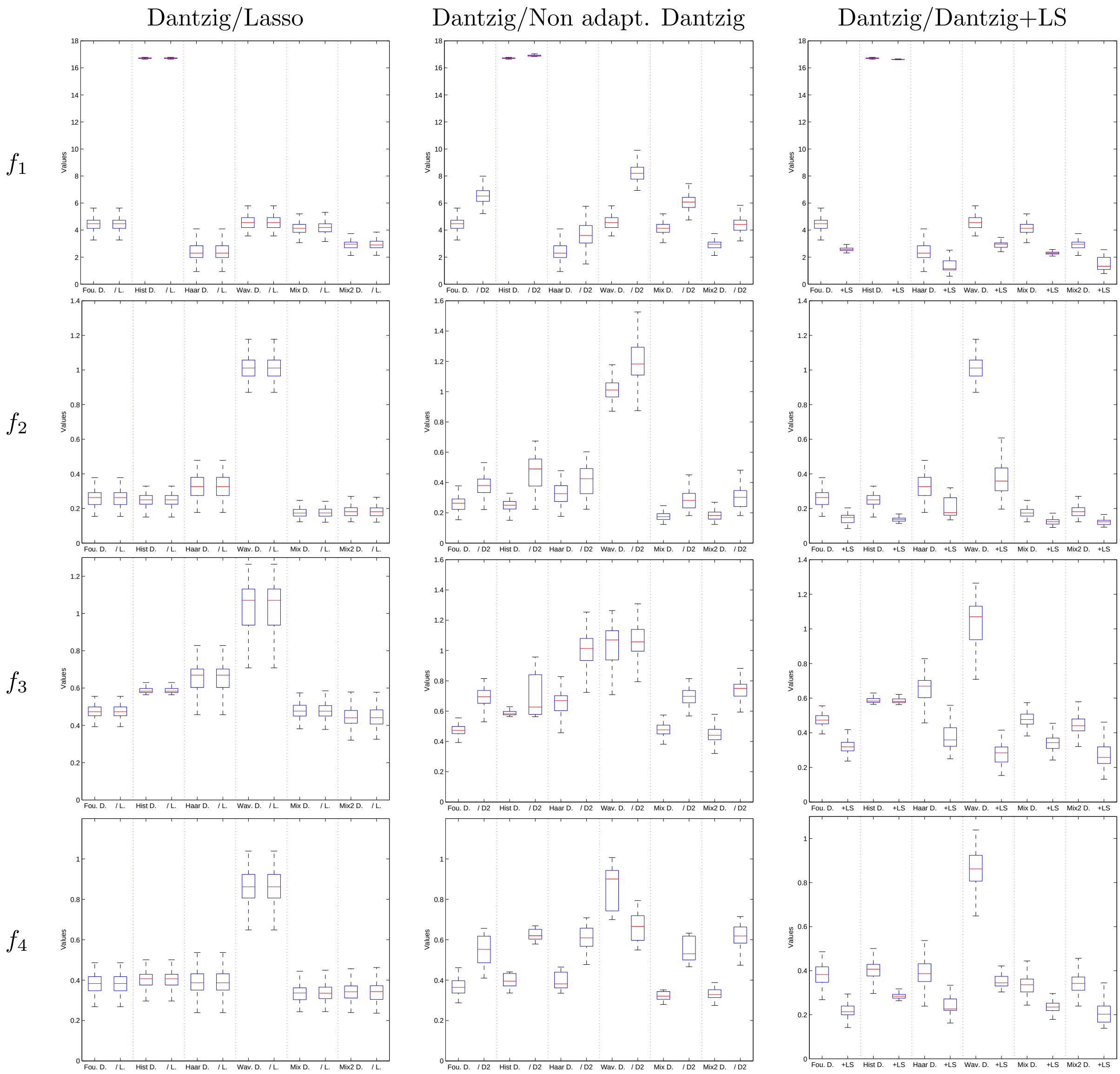

Figure 3: Boxplots for $n = 500$. Left column: Dantzig and Lasso estimates. Center column: Dantzig estimates associated with adaptive and non-adaptive constraints. Right column: Our estimate and the two-step estimate.

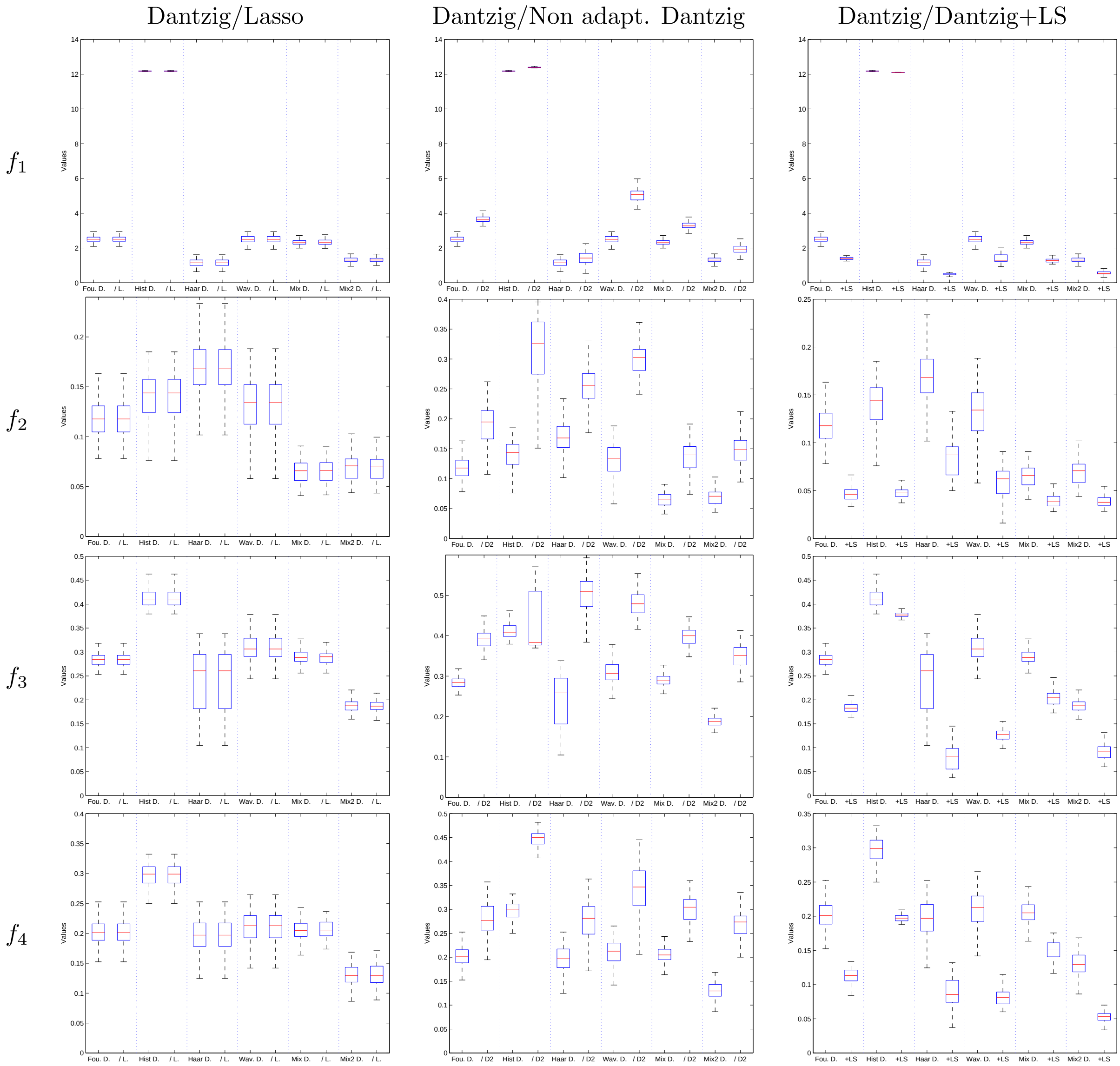

Figure 4: Boxplots for $n = 2000$. Left column: Dantzig and Lasso estimates. Center column: Dantzig estimates associated with adaptive and non-adaptive constraints. Right column: Our estimate and the two-step estimate.

Now, let us decompose $\hat{\sigma}_m^2$ in two terms:

$$\begin{aligned}
\hat{\sigma}_m^2 &= \frac{1}{2n(n-1)} \sum_{i \neq j} (\varphi_m(X_i) - \varphi_m(X_j))^2 \\
&= \frac{1}{2n} \sum_{i=1}^{n} (\varphi_m(X_i) - \beta_{0,m})^2 + \frac{1}{2n} \sum_{j=1}^{n} (\varphi_m(X_j) - \beta_{0,m})^2 \\
&\quad - \frac{2}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\varphi_m(X_i) - \beta_{0,m})(\varphi_m(X_j) - \beta_{0,m}) \\
&= s_n - \frac{2}{n(n-1)} u_n
\end{aligned}$$

with

$$s_n = \frac{1}{n} \sum_{i=1}^{n} (\varphi_m(X_i) - \beta_{0,m})^2 \text{ and } u_n = \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\varphi_m(X_i) - \beta_{0,m})(\varphi_m(X_j) - \beta_{0,m}). \tag{19}$$

Let us first focus on s_n that is the main term of $\hat{\sigma}_m^2$ by applying again Bernstein's Inequality with

$$Y_i = \frac{\sigma_{0,m}^2 - (\varphi_m(X_i) - \beta_{0,m})^2}{n}$$

which satisfies

$$Y_i \leq \frac{\sigma_{0,m}^2}{n}.$$

One has that for any $u > 0$

$$\mathbb{P}\left(\sigma_{0,m}^2 \geq s_n + \sqrt{2 v_m u} + \frac{\sigma_{0,m}^2 u}{3n}\right) \leq e^{-u}$$

with

$$v_m = \frac{1}{n} \mathbb{E}\left(\left[\sigma_{0,m}^2 - (\varphi_m(X_i) - \beta_{0,m})^2\right]^2\right).$$

But we have

$$\begin{aligned}
v_m &= \frac{1}{n}\left(\sigma_{0,m}^4 + \mathbb{E}\left[(\varphi_m(X_i) - \beta_{0,m})^4\right] - 2\sigma_{0,m}^2 \mathbb{E}\left[(\varphi_m(X_i) - \beta_{0,m})^2\right]\right) \\
&= \frac{1}{n}\left(\mathbb{E}\left[(\varphi_m(X_i) - \beta_{0,m})^4\right] - \sigma_{0,m}^4\right) \\
&\leq \frac{\sigma_{0,m}^2}{n} \left(\|\varphi_m\|_\infty + |\beta_{0,m}|\right)^2 \\
&\leq \frac{4\sigma_{0,m}^2}{n} \|\varphi_m\|_\infty^2.
\end{aligned}$$

Finally, with for any $u > 0$

$$S(u) = 2\sqrt{2}\sigma_{0,m} \|\varphi_m\|_\infty \sqrt{\frac{u}{n}} + \frac{\sigma_{0,m}^2 u}{3n},$$

we have

$$\mathbb{P}(\sigma_{0,m}^2 \geq s_n + S(u)) \leq e^{-u}. \tag{20}$$

The term u_n is a degenerate U-statistics that satisfies for any $u > 0$

$$\mathbb{P}(|u_n| \geq U(u)) \leq 6e^{-u}, \tag{21}$$

with for any $u > 0$

$$U(u) = \frac{4}{3}Au^2 + \left(4\sqrt{2} + \frac{2}{3}\right) Bu^{\frac{3}{2}} + \left(2D + \frac{2}{3}F\right) u + 2\sqrt{2}C\sqrt{u},$$

where A, B, C, D and F are constants not depending on u that satisfy

$$\begin{aligned}
&A \leq 4\|\varphi_m\|_\infty^2, \\
&B \leq 2\sqrt{n-1}\|\varphi_m\|_\infty^2, \\
&C \leq \sqrt{\frac{n(n-1)}{2}}\sigma_{0,m}^2, \\
&D \leq \sqrt{\frac{n(n-1)}{2}}\sigma_{0,m}^2,
\end{aligned}$$

and

$$F \leq 2\sqrt{2}\|\varphi_m\|_\infty^2\sqrt{(n-1)\log(2n)}$$

(see [27]). Then, we have for any $u > 0$,

$$\begin{aligned}
\frac{2}{n(n-1)}U(u) \leq &\frac{32}{3}\frac{\|\varphi_m\|_\infty^2}{n(n-1)}u^2 + \left(16\sqrt{2} + \frac{8}{3}\right)\frac{\|\varphi_m\|_\infty^2}{n\sqrt{n-1}}u^{\frac{3}{2}} \\
&+ \left(2\sqrt{2}\frac{\sigma_{0,m}^2}{\sqrt{n(n-1)}} + \frac{8\sqrt{2}}{3}\frac{\sqrt{\log(2n)}\|\varphi_m\|_\infty^2}{n\sqrt{n-1}}\right)u + \frac{4\sigma_{0,m}^2}{\sqrt{n(n-1)}}\sqrt{u}.
\end{aligned}$$

Now, we take u that satisfies

$$u = o(n) \tag{22}$$

and

$$\sqrt{\log(2n)} \leq \sqrt{2u}. \tag{23}$$

Therefore, for any $\varepsilon_1 > 0$, we have for n large enough,

$$\frac{2}{n(n-1)}U(u) \leq \varepsilon_1\sigma_{0,m}^2 + \left(16\sqrt{2} + 8\right)\frac{\|\varphi_m\|_\infty^2}{n\sqrt{n-1}}u^{\frac{3}{2}} + \frac{32}{3}\frac{\|\varphi_m\|_\infty^2}{n(n-1)}u^2.$$

So, for n large enough,

$$\frac{2}{n(n-1)}U(u) \leq \varepsilon_1\sigma_{0,m}^2 + C_1\|\varphi_m\|_\infty^2\left(\frac{u}{n}\right)^{\frac{3}{2}}, \tag{24}$$

where $C_1 = 16\sqrt{2} + 19$. Using Inequalities (20) and (21), we obtain

$$\begin{aligned}
\mathbb{P}\left(\sigma_{0,m}^2 \geq \hat{\sigma}_m^2 + S(u) + \frac{2}{n(n-1)}U(u)\right) &= \mathbb{P}\left(\sigma_{0,m}^2 \geq s_n - \frac{2}{n(n-1)}u_n + S(u) + \frac{2}{n(n-1)}U(u)\right) \\
&\leq \mathbb{P}\left(\sigma_{0,m}^2 \geq s_n + S(u)\right) + \mathbb{P}\left(u_n \geq U(u)\right) \\
&\leq 7e^{-u}.
\end{aligned}$$

Now, using (24), for any $0 < \varepsilon_2 < 1$, we have for n large enough,

$$\begin{aligned}
\hat{\sigma}_m^2 + S(u) + \frac{2}{n(n-1)}U(u) &= \hat{\sigma}_m^2 + 2\sqrt{2}\sigma_{0,m}\|\varphi_m\|_\infty\sqrt{\frac{u}{n}} + \frac{\sigma_{0,m}^2 u}{3n} + \frac{2}{n(n-1)}U(u) \\
&\leq \hat{\sigma}_m^2 + 2\sqrt{2}\sigma_{0,m}\|\varphi_m\|_\infty\sqrt{\frac{u}{n}} + \frac{\sigma_{0,m}^2 u}{3n} + \varepsilon_1\sigma_{0,m}^2 + C_1\|\varphi_m\|_\infty^2\left(\frac{u}{n}\right)^{\frac{3}{2}} \\
&\leq \hat{\sigma}_m^2 + 2\sqrt{2}\sigma_{0,m}\|\varphi_m\|_\infty\sqrt{\frac{u}{n}} + \varepsilon_2\sigma_{0,m}^2 + C_1\|\varphi_m\|_\infty^2\left(\frac{u}{n}\right)^{\frac{3}{2}}.
\end{aligned}$$

Therefore,

$$\mathbb{P}\left((1-\varepsilon_2)\sigma_{0,m}^2 \geq \hat{\sigma}_m^2 + 2\sqrt{2}\sigma_{0,m}\|\varphi_m\|_\infty\sqrt{\frac{u}{n}} + C_1\|\varphi_m\|_\infty^2\left(\frac{u}{n}\right)^{\frac{3}{2}}\right) \leq 7e^{-u}. \tag{25}$$

Now, let us set

$$a = 1-\varepsilon_2, \quad b = \sqrt{2}\|\varphi_m\|_\infty\sqrt{\frac{u}{n}}, \quad c = \hat{\sigma}_m^2 + C_1\|\varphi_m\|_\infty^2\left(\frac{u}{n}\right)^{\frac{3}{2}}$$

and consider the polynomial

$$P(x) = ax^2 - 2bx - c,$$

with roots $\frac{b\pm\sqrt{b^2+ac}}{a}$. So, we have

$$\begin{aligned}
P(\sigma_{0,m}) \geq 0 &\Longleftrightarrow \sigma_{0,m} \geq \frac{b+\sqrt{b^2+ac}}{a} \\
&\Longleftrightarrow \sigma_{0,m}^2 \geq \frac{c}{a} + \frac{2b^2}{a^2} + \frac{2b\sqrt{b^2+ac}}{a^2}.
\end{aligned}$$

It yields

$$\mathbb{P}\left(\sigma_{0,m}^2 \geq \frac{c}{a} + \frac{2b^2}{a^2} + \frac{2b\sqrt{b^2+ac}}{a^2}\right) \leq 7e^{-u},$$

so,

$$\mathbb{P}\left(\sigma_{0,m}^2 \geq \frac{c}{a} + \frac{4b^2}{a^2} + \frac{2b\sqrt{c}}{a\sqrt{a}}\right) \leq 7e^{-u},$$

which means that for any $0 < \varepsilon_3 < 1$, we have for n large enough,

$$\mathbb{P}\left(\sigma_{0,m}^2 \geq (1+\varepsilon_3)\left(\hat{\sigma}_m^2 + C_1\|\varphi_m\|_\infty^2\left(\frac{u}{n}\right)^{\frac{3}{2}} + 8\|\varphi_m\|_\infty^2\frac{u}{n} + 2\sqrt{2}\|\varphi_m\|_\infty\sqrt{\frac{u}{n}}\sqrt{\hat{\sigma}_m^2 + C_1\|\varphi_m\|_\infty^2\left(\frac{u}{n}\right)^{\frac{3}{2}}}\right)\right) \leq 7e^{-u}.$$

Finally, we can claim that for any $0 < \varepsilon_4 < 1$, we have for n large enough,

$$\mathbb{P}\left(\sigma_{0,m}^2 \geq (1+\varepsilon_4)\left(\hat{\sigma}_m^2 + 8\|\varphi_m\|_\infty^2\frac{u}{n} + 2\|\varphi_m\|_\infty\sqrt{2\hat{\sigma}_m^2\frac{u}{n}}\right)\right) \leq 7e^{-u}.$$

Now, we take $u = \gamma\log M$. Under Assumptions of Theorem 1, Conditions (22) and (23) are satisfied. The previous concentration inequality means that

$$\mathbb{P}\left(\sigma_{0,m}^2 \geq (1+\varepsilon_4)\tilde{\sigma}_m^2\right) \leq 7M^{-\gamma}.$$

Now, using (18), we have for n large enough,

$$\begin{aligned}\mathbb{P}\left(|\beta_{0,m}-\hat{\beta}_m|\geq\eta_{\gamma,m}\right)&=\mathbb{P}\left(|\beta_{0,m}-\hat{\beta}_m|\geq\sqrt{\frac{2\tilde{\sigma}_m^2\gamma\log M}{n}}+\frac{2\|\varphi_m\|_\infty\gamma\log M}{3n},\ \sigma_{0,m}^2<(1+\varepsilon_4)\tilde{\sigma}_m^2\right)\\&\quad+\mathbb{P}\left(|\beta_{0,m}-\hat{\beta}_m|\geq\eta_{\gamma,m},\ \sigma_{0,m}^2\geq(1+\varepsilon_4)\tilde{\sigma}_m^2\right)\\&\leq\mathbb{P}\left(|\beta_{0,m}-\hat{\beta}_m|\geq\sqrt{\frac{2\sigma_{0,m}^2\gamma(1+\varepsilon_4)^{-1}\log M}{n}}+\frac{2\|\varphi_m\|_\infty\gamma(1+\varepsilon_4)^{-1}\log M}{3n}\right)\\&\quad+\mathbb{P}\left(\sigma_{0,m}^2\geq(1+\varepsilon_4)\tilde{\sigma}^2\right)\\&\leq 2M^{-\gamma(1+\varepsilon_4)^{-1}}+7M^{-\gamma}.\end{aligned}$$

Then, the first part of Theorem 1 is proved: for any $\varepsilon>0$,

$$\mathbb{P}\left(|\beta_{0,m}-\hat{\beta}_m|\geq\eta_{\gamma,m}\right)\leq C(\varepsilon,\delta,\gamma)M^{-\frac{\gamma}{1+\varepsilon}},$$

where $C(\varepsilon,\delta,\gamma)$ is a constant that depends on ε, δ and γ.

For the second part of the result, we apply again Bernstein's Inequality with

$$Z_i=\frac{(\varphi_m(X_i)-\beta_{0,m})^2-\sigma_{0,m}^2}{n}$$

which satisfies

$$Z_i\leq\frac{(\varphi_m(X_i)-\beta_{0,m})^2}{n}\leq\frac{4\|\varphi_m\|_\infty^2}{n}.$$

One has that for any $u>0$

$$\mathbb{P}\left(s_n\geq\sigma_{0,m}^2+\sqrt{2v_m u}+\frac{4\|\varphi_m\|_\infty^2 u}{3n}\right)\leq e^{-u}$$

with

$$v_m=\frac{1}{n}\mathbb{E}\left(\left[\sigma_{0,m}^2-(\varphi_m(X_i)-\beta_{0,m})^2\right]^2\right)\leq\frac{4\sigma_{0,m}^2}{n}\|\varphi_m\|_\infty^2.$$

So, for any $u>0$,

$$\mathbb{P}\left(s_n\geq\sigma_{0,m}^2+2\sqrt{2}\sigma_{0,m}\|\varphi_m\|_\infty\sqrt{\frac{u}{n}}+\frac{4\|\varphi_m\|_\infty^2 u}{3n}\right)\leq e^{-u}.$$

Now, for any $\varepsilon_5>0$, for any $u>0$,

$$\mathbb{P}\left(s_n\geq(1+\varepsilon_5)\sigma_{0,m}^2+\frac{\|\varphi_m\|_\infty^2 u}{n}\left(\frac{4}{3}+\frac{2}{\varepsilon_5}\right)\right)\leq e^{-u}.$$

Using (21), with

$$\tilde{S}(u)=\frac{\|\varphi_m\|_\infty^2 u}{n}\left(\frac{4}{3}+\frac{2}{\varepsilon_5}\right),$$

$$\begin{aligned}\mathbb{P}\left(\hat{\sigma}_m^2\geq(1+\varepsilon_5)\sigma_{0,m}^2+\tilde{S}(u)+\frac{2}{n(n-1)}U(u)\right)&=\mathbb{P}\left(s_n-\frac{2}{n(n-1)}u_n\geq(1+\varepsilon_5)\sigma_{0,m}^2+\tilde{S}(u)+\frac{2}{n(n-1)}U(u)\right)\\&\leq\mathbb{P}\left(s_n\geq(1+\varepsilon_5)\sigma_{0,m}^2+\tilde{S}(u)\right)+\mathbb{P}\left(-u_n\geq U(u)\right)\\&\leq e^{-u}+6e^{-u}=7e^{-u}.\end{aligned}$$

Using (24),

$$\mathbb{P}\left(\hat{\sigma}_m^2 \geq (1+\varepsilon_1+\varepsilon_5)\sigma_{0,m}^2 + \tilde{S}(u) + C_1\|\varphi_m\|_\infty^2 \left(\frac{u}{n}\right)^{\frac{3}{2}}\right) \leq 7e^{-u}.$$

Since

$$\eta_{\gamma,m} = \sqrt{\frac{2\tilde{\sigma}_m^2\gamma\log M}{n}} + \frac{2\|\varphi_m\|_\infty\gamma\log M}{3n},$$

with

$$\tilde{\sigma}_m^2 = \hat{\sigma}_m^2 + 2\|\varphi_m\|_\infty\sqrt{\frac{2\hat{\sigma}_m^2\gamma\log M}{n}} + \frac{8\|\varphi_m\|_\infty^2\gamma\log M}{n},$$

we have for any $\varepsilon_6 > 0$,

$$\begin{aligned}
\eta_{\gamma,m}^2 &\leq (1+\varepsilon_6)\left(\frac{2\tilde{\sigma}_m^2\gamma\log M}{n}\right) + (1+\varepsilon_6^{-1})\left(\frac{4\|\varphi_m\|_\infty^2(\gamma\log M)^2}{9n^2}\right)\\
&\leq (1+\varepsilon_6)\left(\frac{2\gamma\log M}{n}\right)\left(\hat{\sigma}_m^2 + 2\|\varphi_m\|_\infty\sqrt{\frac{2\hat{\sigma}_m^2\gamma\log M}{n}} + \frac{8\|\varphi_m\|_\infty^2\gamma\log M}{n}\right)\\
&\quad + \frac{4}{9}(1+\varepsilon_6^{-1})\left(\frac{\|\varphi_m\|_\infty\gamma\log M}{n}\right)^2\\
&\leq (1+\varepsilon_6)^2\hat{\sigma}_m^2\left(\frac{2\gamma\log M}{n}\right) + 4\varepsilon_6^{-1}(1+\varepsilon_6)\left(\frac{\|\varphi_m\|_\infty\gamma\log M}{n}\right)^2\\
&\quad + 16(1+\varepsilon_6)\left(\frac{\|\varphi_m\|_\infty\gamma\log M}{n}\right)^2 + \frac{4(1+\varepsilon_6^{-1})}{9}\left(\frac{\|\varphi_m\|_\infty\gamma\log M}{n}\right)^2.
\end{aligned}$$

Finally, with $u = \gamma\log M$, with probability larger than $1 - 7M^{-\gamma}$,

$$\hat{\sigma}_m^2 < (1+\varepsilon_1+\varepsilon_5)\sigma_{0,m}^2 + \tilde{S}(\gamma\log M) + C_1\|\varphi_m\|_\infty^2\left(\frac{\gamma\log M}{n}\right)^{\frac{3}{2}},$$

and

$$\begin{aligned}
\eta_{\gamma,m}^2 &< (1+\varepsilon_6)^2(1+\varepsilon_5+\varepsilon_1)\sigma_{0,m}^2\left(\frac{2\gamma\log M}{n}\right) + (1+\varepsilon_6)^2\left(\frac{\gamma\log M}{n}\right)^2\|\varphi_m\|_\infty^2\left(\frac{8}{3}+\frac{4}{\varepsilon_5}\right)\\
&\quad + 2C_1(1+\varepsilon_6)^2\|\varphi_m\|_\infty^2\left(\frac{\gamma\log M}{n}\right)^{\frac{5}{2}} + \|\varphi_m\|_\infty^2\left(\frac{\gamma\log M}{n}\right)^2\left(4\varepsilon_6^{-1}(1+\varepsilon_6) + 16(1+\varepsilon_6) + \frac{4(1+\varepsilon_6^{-1})}{9}\right).
\end{aligned}$$

Finally, with $\varepsilon_6 = 1$, $\varepsilon_1 = \varepsilon_5 = \frac{1}{2}$, for n large enough,

$$\mathbb{P}\left(\eta_{\gamma,m} \geq 4\sigma_{0,m}\sqrt{\frac{\gamma\log M}{n}} + \frac{10\|\varphi_m\|_\infty\gamma\log M}{n}\right) \leq 7M^{-\gamma}.$$

Note that $\sqrt{32/3+32+8+32+8/9} = 9.1409$.

For the last part, starting from (25) with $u = \gamma\log M$ and $\varepsilon_2 = \frac{1}{7}$, we have for n large enough and with probability larger than $1 - 7M^{-\gamma}$,

$$\begin{aligned}
\frac{6}{7}\sigma_{0,m}^2 &\leq \hat{\sigma}_m^2 + 2\sqrt{2}\sigma_{0,m}\|\varphi_m\|_\infty\sqrt{\frac{\gamma\log M}{n}} + C_1\|\varphi_m\|_\infty^2\left(\frac{\gamma\log M}{n}\right)^{\frac{3}{2}}\\
&\leq \hat{\sigma}_m^2 + \frac{2}{7}\sigma_{0,m}^2 + 7\|\varphi_m\|_\infty^2\frac{\gamma\log M}{n} + C_1\|\varphi_m\|_\infty^2\left(\frac{\gamma\log M}{n}\right)^{\frac{3}{2}}.
\end{aligned}$$

So, for n large enough,
$$\frac{4}{7}\sigma_{0,m}^2 \le \hat{\sigma}_m^2 + 8\|\varphi_m\|_\infty^2 \frac{\gamma \log M}{n} \le \tilde{\sigma}_m^2$$
and
$$\eta_{\gamma,m} > \sigma_{0,m}\sqrt{\frac{8\gamma \log M}{7n}} + \frac{2\|\varphi_m\|_\infty \gamma \log M}{3n}.$$

6.2 Proof of Theorem 2

Let $\lambda = (\lambda_m)_{m=1,\dots,M}$ and set $\Delta = \lambda - \hat{\lambda}^D$. We have

$$\|f_\lambda - f_0\|_2^2 = \|\hat{f}^D - f_0\|_2^2 + \|f_\lambda - \hat{f}^D\|_2^2 + 2\int (\hat{f}^D(x) - f_0(x))(f_\lambda(x) - \hat{f}^D(x))dx. \qquad (26)$$

We have $\|f_\lambda - \hat{f}^D\|_2^2 = \|f_\Delta\|_2^2$. Moreover, with probability at least $1 - C_1(\varepsilon,\delta,\gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, we have

$$\begin{aligned}\left|\int (\hat{f}^D(x) - f_0(x))(f_\lambda(x) - \hat{f}^D(x))dx\right| &= \left|\sum_{m=1}^{M}(\lambda_m - \hat{\lambda}_m^D)\left[(G\hat{\lambda}^D)_m - \beta_{0,m}\right]\right| \qquad (27)\\ &\le \|\Delta\|_{\ell_1} 2\|\eta_\gamma\|_{\ell_\infty},\end{aligned}$$

where the last line is a consequence of the definition of the Dantzig estimator and of Theorem 1. Then, we have
$$\|\hat{f}^D - f_0\|_2^2 \le \|f_\lambda - f_0\|_2^2 + 4\|\eta_\gamma\|_{\ell_\infty}\|\Delta\|_{\ell_1} - \|f_\Delta\|_2^2.$$
We use then the following Lemma:

Lemma 1. *Let $J \subset \{1,\dots,M\}$. For any $\lambda \in \mathbb{R}^M$*

$$\|\Delta_{J^C}\|_{\ell_1} \le \|\Delta_J\|_{\ell_1} + 2\|\lambda_{J^C}\|_{\ell_1} + \left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+,$$

where $\Delta = \hat{\lambda}^D - \lambda$.

Proof.[Proof of Lemma 1] This lemma is based on the fact that

$$\|\hat{\lambda}^D\|_{\ell_1} \le \|\lambda\|_{\ell_1} + \left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+,$$

which implies that

$$\|\Delta_J + \lambda_J\|_{\ell_1} + \|\Delta_{J^C} + \lambda_{J^C}\|_{\ell_1} \le \|\lambda_J\|_{\ell_1} + \|\lambda_{J^C}\|_{\ell_1} + \left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+,$$

and thus

$$\|\lambda_J\|_{\ell_1} - \|\Delta_J\|_{\ell_1} + \|\Delta_{J^C}\|_{\ell_1} - \|\lambda_{J^C}\|_{\ell_1} \le \|\lambda_J\|_{\ell_1} + \|\lambda_{J^C}\|_{\ell_1} + \left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+.$$

■

Note that if λ satisfies the Dantzig condition then by definition of $\hat{\lambda}^D$: $\left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+ = 0$. Using the previous lemma, we have:

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

Using now $\Lambda(\lambda, J_0^c) = \|\lambda_{J_0^C}\|_{\ell_1} + \frac{\left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+}{2}$, so that $\Lambda(\lambda, J_0^c) = \|\lambda_{J_0^C}\|_{\ell_1}$ as soon as λ satisfies the Dantzig condition, we obtain

$$\begin{aligned}
\|f_\Delta\|_2 &\geq \kappa_{J_0}\|\Delta_{J_0}\|_{\ell_2} - \mu_{J_0}\left(\|\Delta_{J_0^C}\|_{\ell_1} - \|\Delta_{J_0}\|_{\ell_1}\right)_+ \\
&\geq \kappa_{J_0}\|\Delta_{J_0}\|_{\ell_2} - 2\mu_{J_0}\Lambda(\lambda, J_0^c)
\end{aligned}$$

and thus

$$\|\Delta_{J_0}\|_{\ell_2} \leq \frac{1}{\kappa_{J_0}}\|f_\Delta\|_2 + 2\frac{\mu_{J_0}}{\kappa_{J_0}}\Lambda(\lambda, J_0^c).$$

We deduce thus

$$\begin{aligned}
\|\Delta\|_{\ell_1} &\leq 2\|\Delta_{J_0}\|_{\ell_1} + 2\Lambda(\lambda, J_0^c) \\
&\leq 2\sqrt{|J_0|}\|\Delta_{J_0}\|_{\ell_2} + 2\|\widetilde{\lambda_{J_0^C}}\|_{\ell_1} \\
&\leq \frac{2\sqrt{|J_0|}}{\kappa_{J_0}}\|f_\Delta\|_2 + 2\Lambda(\lambda, J_0^c)\left(1 + \frac{2\mu_{J_0}\sqrt{|J_0|}}{\kappa_{J_0}}\right)
\end{aligned}$$

and then since

$$4\|\eta_\gamma\|_{\ell_\infty}\frac{2\sqrt{|J_0|}}{\kappa_{J_0}}\|f_\Delta\|_2 \leq \frac{16|J_0|\|\eta_\gamma\|_{\ell_\infty}^2}{\kappa_{J_0}^2} + \|f_\Delta\|_2^2$$

we have

$$\begin{aligned}
4\|\eta_\gamma\|_{\ell_\infty}\|\Delta\|_{\ell_1} - \|f_\Delta\|_2^2 &\leq \frac{16|J_0|\|\eta_\gamma\|_{\ell_\infty}^2}{\kappa_{J_0}^2} + 8\|\eta_\gamma\|_{\ell_\infty}\Lambda(\lambda, J_0^c)\left(1 + \frac{2\mu_{J_0}\sqrt{|J_0|}}{\kappa_{J_0}}\right) \\
&\leq 16|J_0|\left(\frac{1}{\beta} + \frac{1}{\kappa_{J_0}^2}\right)\|\eta_\gamma\|_{\ell_\infty}^2 + \beta\frac{\Lambda(\lambda, J_0^c)^2}{|J_0|}\left(1 + \frac{2\mu_{J_0}\sqrt{|J_0|}}{\kappa_{J_0}}\right)^2,
\end{aligned}$$

which is the result of the theorem.

6.3 Consequences of Assumptions 1 and 2

To prove Proposition 1, we establish Lemmas 2 and 3. In the sequel, we consider two integers s and l such that $1 \leq s \leq M/2$, $l \geq s$ and $s + l \leq M$. We first recall Assumptions 1 and 2. Assumption 1 is stated in a more general form, which allows to unify the statement of the subsequent results.

- **Assumption 1**

$$\phi_{\min}(s+l) > \theta_{l,s+l}.$$

- **Assumption 2**

$$l\phi_{\min}(s+l) > s\phi_{\max}(l).$$

In the sequel, we assume that Assumptions 1 and 2 are both true.

Lemma 2. *Let $J_0 \subset \{1, \ldots, M\}$ with cardinality $|J_0| = s$ and $\Delta \in \mathbb{R}^M$. We denote by J_1 the subset of $\{1, \ldots, M\}$ corresponding to the l largest coordinates of Δ (in absolute value) outside*

J_0 and we set $J_{01} = J_0 \cup J_1$. We denote by $P_{J_{01}}$ the projector on the linear space spanned by $(\varphi_m)_{m\in J_{01}}$. We have:

$$\|P_{J_{01}} f_\Delta\|_2 \geq \sqrt{\phi_{\min}(s+l)}\|\Delta_{J_{01}}\|_{\ell_2} - \min(\mu_1, \mu_2) \|\Delta_{J_0^C}\|_{\ell_1},$$

with

$$\mu_1 = \frac{\theta_{l,s+l}}{\sqrt{l\phi_{\min}(s+l)}} \quad \textit{and} \quad \mu_2 = \sqrt{\frac{\phi_{\max}(l)}{l}}.$$

Proof. For $k > 1$, we denote by J_k the indices corresponding to the coordinates of Δ outside J_0 whose absolute values are between the $((k-1)\times l+1)$–th and the $(k \times l)$–th largest ones (in absolute value). Note that this definition is consistent with the definition of J_1. Using this notation, we have

$$\begin{aligned}\|P_{J_{01}} f_\Delta\|_2 &\geq \|P_{J_{01}} f_{\Delta_{J_{01}}}\|_2 - \|\sum_{k\geq 2} P_{J_{01}} f_{\Delta_{J_k}}\|_2 \\ &\geq \|f_{\Delta_{J_{01}}}\|_2 - \sum_{k\geq 2}\|P_{J_{01}} f_{\Delta_{J_k}}\|_2.\end{aligned}$$

Since J_{01} has $s+l$ elements, we have

$$\|f_{\Delta_{J_{01}}}\|_2 \geq \sqrt{\phi_{\min}(s+l)}\|\Delta_{J_{01}}\|_{\ell_2}.$$

Note that $P_{J_{01}} f_{\Delta_{J_k}} = f_{C_{J_{01}}}$ for some vector $C \in \mathbb{R}^M$. Since,

$$\langle P_{J_{01}} f_{\Delta_{J_k}} - f_{\Delta_{J_k}}, P_{J_{01}} f_{\Delta_{J_k}}\rangle = 0,$$

one obtains that

$$\|P_{J_{01}} f_{\Delta_{J_k}}\|_2^2 = \langle f_{\Delta_{J_k}}, f_{C_{J_{01}}}\rangle$$

and thus

$$\begin{aligned}\|P_{J_{01}} f_{\Delta_{J_k}}\|_2^2 &\leq \theta_{l,s+l}\|\Delta_{J_k}\|_{\ell_2}\|C_{J_{01}}\|_{\ell_2} \leq \theta_{l,s+l}\|\Delta_{J_k}\|_{\ell_2}\frac{\|f_{C_{J_{01}}}\|_2}{\sqrt{\phi_{\min}(s+l)}} \\ &\leq \frac{\theta_{l,s+l}}{\sqrt{\phi_{\min}(s+l)}}\|\Delta_{J_k}\|_{\ell_2}\|P_{J_{01}} f_{\Delta_{J_k}}\|_2.\end{aligned}$$

This implies that

$$\|P_{J_{01}} f_{\Delta_{J_k}}\|_2 \leq \frac{\theta_{l,s+l}}{\sqrt{\phi_{\min}(s+l)}}\|\Delta_{J_k}\|_{\ell_2} = \mu_1\sqrt{l}\|\Delta_{J_k}\|_{\ell_2}.$$

Moreover, using that J_k has less than l elements, we obtain that

$$\|P_{J_{01}} f_{\Delta_{J_k}}\|_2 \leq \|f_{\Delta_{J_k}}\|_2 \leq \sqrt{\phi_{\max}(l)}\|\Delta_{J_k}\|_{\ell_2} = \mu_2\sqrt{l}\|\Delta_{J_k}\|_{\ell_2}.$$

Now using that $\|\Delta_{J_{k+1}}\|_{\ell_2} \leq \|\Delta_{J_k}\|_{\ell_1}/\sqrt{l}$, we obtain

$$\sum_{k\geq 2}\|P_{J_{01}} f_{\Delta_{J_k}}\|_2 \leq \min(\mu_1, \mu_2)\|\Delta_{J_0^C}\|_{\ell_1}$$

and finally

$$\|P_{J_{01}} f_\Delta\|_2 \geq \sqrt{\phi_{\min}(s+l)}\|\Delta_{J_{01}}\|_{\ell_2} - \min(\mu_1, \mu_2)\|\Delta_{J_0^C}\|_{\ell_1}.$$

■

Lemma 3. *We use the same notations as in Lemma 2. For $c \geq 0$, assume that*

$$\|\Delta_{J_0^C}\|_{\ell_1} \leq \|\Delta_{J_0}\|_{\ell_1} + c. \tag{28}$$

Then we have

$$\|P_{J_{01}} f_\Delta\|_2 \geq \max(\kappa_1, \kappa_2) \|\Delta_{J_{01}}\|_{\ell_2} - \min(\mu_1, \mu_2) c,$$

with

$$\kappa_1 = \sqrt{\phi_{\min}(s+l)} \left(1 - \frac{\theta_{l,s+l}}{\phi_{\min}(s+l)} \sqrt{\frac{s}{l}}\right) \quad \text{and} \quad \kappa_2 = \sqrt{\phi_{\min}(s+l)} \left(1 - \sqrt{\frac{s\phi_{\max}(l)}{l\phi_{\min}(s+l)}}\right).$$

Proof. Using Lemma 2 and (28), we obtain that

$$\|P_{J_{01}} f_\Delta\|_2 \geq \sqrt{\phi_{\min}(s+l)} \|\Delta_{J_{01}}\|_{\ell_2} - \min(\mu_1, \mu_2) (\|\Delta_{J_0}\|_{\ell_1} + c).$$

Using $\|\Delta_{J_0}\|_{\ell_1} \leq \sqrt{s}\|\Delta_{J_0}\|_{\ell_2}$, we deduce that

$$\begin{aligned}
\|P_{J_{01}} f_\Delta\|_2 &\geq \left(\sqrt{\phi_{\min}(s+l)} - \sqrt{s}\min(\mu_1, \mu_2)\right) \|\Delta_{J_{01}}\|_{\ell_2} - c\min(\mu_1, \mu_2) \\
&\geq \max(\kappa_1, \kappa_2) \|\Delta_{J_{01}}\|_{\ell_2} - c\min(\mu_1, \mu_2).
\end{aligned}$$

■

6.4 Proof of Theorem 5

The dictionary considered here is the Haar dictionary $(\phi_{jk})_{j,k}$ and is double-indexed. As a consequence, in the following, the quantity $\beta_{0,jk}$, $\hat{\beta}_{jk}$, $\sigma^2_{0,jk}$ $\eta_{\gamma,jk}$, $\tilde{\sigma}^2_{jk}$ and $\hat{\sigma}^2_{jk}$ are defined as in (1), (2), (3), (5), (6) and (7) where φ_m is replaced by ϕ_{jk}. Note that, since $f_0 = \mathbf{1}_{[0,1]}$, we have, for $j \neq -1$, $\beta_{0,jk} = 0$ and for any j, $\sigma^2_{0,jk} = 1$ if $k \in \{0, \ldots, 2^j - 1\}$ and 0 otherwise.

The proof of (16) is provided by using the oracle inequality satisfied by hard thresholding given by Theorem 1 of [27] and the rough control of the soft thresholding estimate by the hard one:

$$\left||\hat{\beta}_{jk}| - \eta_{\gamma,jk}\right| 1_{\{|\hat{\beta}_{jk}| \geq \eta_{\gamma,jk}\}} \leq 2|\hat{\beta}_{jk}| 1_{\{|\hat{\beta}_{jk}| \geq \eta_{\gamma,jk}\}}.$$

An alternative is directly obtained by adapting the oracle results derived for soft thresholding rules in the regression model considered by Donoho and Johnstone [16].

To prove (17), we establish the following lemma.

Lemma 4. *Let $\gamma < 1$. We consider $j \in \mathbb{N}$ such that*

$$\frac{n}{(\log n)^\alpha} \leq 2^j < \frac{2n}{(\log n)^\alpha}, \tag{29}$$

for some $\alpha > 1$. Then for all $\varepsilon > 0$ such that $\gamma + 2\varepsilon < 1$,

$$\sum_{k=0}^{2^j-1} \mathbb{E}\left(\hat{\beta}^2_{jk} \mathbf{1}_{|\hat{\beta}_{jk}| \geq \eta_{\gamma,jk}}\right) \geq \frac{2\gamma(1+\varepsilon)e^{-2}}{\pi} (\log n)^{1-2\alpha} n^{-(\gamma+2\varepsilon)} (1 + o_n(1)).$$

Then, we use the following inequality. For j that satisfies (29), we have for $r > 0$,

$$\begin{aligned}
\mathbb{E}(\|\hat{f}^D - f_0\|_2^2) &\geq \sum_{k=0}^{2^j-1} \mathbb{E}\left(\left(|\hat{\beta}_{jk}| - \eta_{\gamma,jk}\right)^2 \mathbf{1}_{|\hat{\beta}_{jk}| \geq \eta_{\gamma,jk}}\right) \\
&\geq \sum_{k=0}^{2^j-1} \mathbb{E}\left(\left(|\hat{\beta}_{jk}| - \eta_{\gamma,jk}\right)^2 \mathbf{1}_{|\hat{\beta}_{jk}| \geq (1+r)\eta_{\gamma,jk}}\right) \\
&\geq \left(\frac{r}{r+1}\right)^2 \sum_{k=0}^{2^j-1} \mathbb{E}\left(\hat{\beta}_{jk}^2 \mathbf{1}_{|\hat{\beta}_{jk}| \geq (1+r)\eta_{\gamma,jk}}\right) \\
&\geq \left(\frac{r}{r+1}\right)^2 \sum_{k=0}^{2^j-1} \mathbb{E}\left(\hat{\beta}_{jk}^2 \mathbf{1}_{|\hat{\beta}_{jk}| \geq \eta_{jk,(1+r)^2\gamma}}\right).
\end{aligned}$$

So, if r and ε are such that $(1+r)^2\gamma + 2\varepsilon < 1$, then applying Lemma 4, Inequality (17) is proved for any δ such that $(1+r)^2\gamma + 2\varepsilon < \delta < 1$.

Proof. [Proof of Lemma 4] Let j that satisfies (29) and $0 \leq k \leq 2^j - 1$. We have

$$\tilde{\sigma}_{jk}^2 = \hat{\sigma}_{jk}^2 + 2\|\phi_{j,k}\|_\infty \sqrt{2\gamma\hat{\sigma}_{jk}^2 \frac{\log n}{n}} + 8\gamma\|\phi_{j,k}\|_\infty^2 \frac{\log n}{n}.$$

So, for any $0 < \varepsilon < \frac{1-\gamma}{2} < \frac{1}{2}$,

$$\tilde{\sigma}_{jk}^2 \leq (1+\varepsilon)\hat{\sigma}_{jk}^2 + 2\gamma\|\phi_{j,k}\|_\infty^2 \frac{\log n}{n}\left(\varepsilon^{-1} + 4\right).$$

Now,

$$\begin{aligned}
\eta_{\gamma,jk} &= \sqrt{2\gamma\tilde{\sigma}_{jk}^2 \frac{\log n}{n}} + \frac{2\|\phi_{j,k}\|_\infty \gamma \log n}{3n} \\
&\leq \sqrt{2\gamma \frac{\log n}{n}\left((1+\varepsilon)\hat{\sigma}_{jk}^2 + 2\gamma\|\phi_{j,k}\|_\infty^2 \frac{\log n}{n}(\varepsilon^{-1}+4)\right)} + \frac{2\|\phi_{j,k}\|_\infty \gamma \log n}{3n} \\
&\leq \sqrt{2\gamma(1+\varepsilon)\hat{\sigma}_{jk}^2 \frac{\log n}{n}} + \frac{2\|\phi_{j,k}\|_\infty \gamma \log n}{n}\left(\frac{1}{3} + \sqrt{4+\varepsilon^{-1}}\right).
\end{aligned}$$

Furthermore, we have

$$\hat{\sigma}_{jk}^2 = s_{njk} - \frac{2}{n(n-1)} u_{njk},$$

where s_{njk} and u_{njk} are defined as in (19) with φ_m replaced by ϕ_{jk}. This implies that

$$\eta_{\gamma,jk} \leq \sqrt{2\gamma(1+\varepsilon)\frac{\log n}{n} s_{njk}} + \sqrt{2\gamma(1+\varepsilon)\frac{\log n}{n} \times \frac{2}{n(n-1)}|u_{njk}|} + \frac{2\|\phi_{j,k}\|_\infty \gamma \log n}{n}\left(\frac{1}{3} + \sqrt{4+\varepsilon^{-1}}\right).$$

Using (21), with probability larger than $1 - 6n^{-2}$, we have

$$|u_{njk}| \leq U(2\log n),$$

and, since $\sigma^2_{0,jk} = 1$

$$\begin{aligned}\frac{2}{n(n-1)}U(2\log n) &\leq \frac{c_1}{n}\sqrt{\log n} + \frac{c_2}{n}\log n + c_3\|\phi_{j,k}\|^2_\infty\left(\frac{\log n}{n}\right)^{\frac{3}{2}} + c_4\|\phi_{j,k}\|^2_\infty\left(\frac{\log n}{n}\right)^2 \\ &\leq C_1\frac{\log n}{n} + C_2\|\phi_{j,k}\|^2_\infty\left(\frac{\log n}{n}\right)^{\frac{3}{2}},\end{aligned}$$

where c_1, c_2, c_3, c_4, C_1 and C_2 are universal constants. Finally, with probability larger than $1 - 6n^{-2}$, we obtain that

$$\sqrt{2\gamma(1+\varepsilon)\frac{\log n}{n} \times \frac{2}{n(n-1)}|u_{njk}|} \leq \sqrt{2\gamma(1+\varepsilon)C_1}\frac{\log n}{n} + \sqrt{2\gamma(1+\varepsilon)C_2}\|\phi_{j,k}\|_\infty\left(\frac{\log n}{n}\right)^{\frac{5}{4}}.$$

So, since $\gamma < 1$, there exists $w(\varepsilon)$, only depending on ε such that with probability larger than $1 - 6n^{-2}$,

$$\eta_{\gamma,jk} \quad \leq \quad \sqrt{2\gamma(1+\varepsilon)\frac{\log n}{n}s_{njk}} + w(\varepsilon)\|\phi_{jk}\|_\infty\frac{\log n}{n}.$$

We set

$$\widetilde{\eta_{\gamma,jk}} = \sqrt{2\gamma(1+\varepsilon)s_{njk}\frac{\log n}{n}} + w(\varepsilon)\frac{2^{\frac{j}{2}}\log n}{n}$$

so $\eta_{\gamma,jk} \leq \widetilde{\eta_{\gamma,jk}}$. Then, we have

$$\begin{aligned} s_{njk} &= \frac{1}{n}\sum_{i=1}^n(\phi_{jk}(X_i) - \beta_{0,jk})^2 \\ &= \frac{2^j}{n}\sum_{i=1}^n\left(\mathbf{1}_{X_i\in[k2^{-j},(k+0.5)2^{-j}[} - \mathbf{1}_{X_i\in[(k+0.5)2^{-j},(k+1)2^{-j}[}\right)^2 \\ &= \frac{2^j}{n}\left(N^+_{jk} + N^-_{jk}\right),\end{aligned}$$

with

$$N^+_{jk} = \sum_{i=1}^n \mathbf{1}_{X_i\in[k2^{-j},(k+0.5)2^{-j}[}, \quad N^-_{jk} = \sum_{i=1}^n \mathbf{1}_{X_i\in[(k+0.5)2^{-j},(k+1)2^{-j}[}.$$

We consider j such that

$$\frac{n}{(\log n)^\alpha} \leq 2^j < \frac{2n}{(\log n)^\alpha}, \quad \alpha > 1.$$

In particular, we have

$$\frac{(\log n)^\alpha}{2} < n2^{-j} \leq (\log n)^\alpha.$$

Now, we can write

$$\hat\beta_{jk} = \frac{1}{n}\sum_{i=1}^n\phi_{jk}(X_i) = \frac{2^{\frac{j}{2}}}{n}(N^+_{jk} - N^-_{jk}),$$

that implies that

$$\sum_{k=0}^{2^j-1} \mathbb{E}\left(\hat{\beta}_{jk}^2 \mathbf{1}_{|\hat{\beta}_{jk}|\geq \eta_{\gamma,jk}}\right)$$

$$\begin{aligned}
&\geq \sum_{k=0}^{2^j-1} \mathbb{E}\left(\hat{\beta}_{jk}^2 \mathbf{1}_{|\hat{\beta}_{jk}|\geq \widetilde{\eta_{\gamma,jk}}} \mathbf{1}_{|u_{njk}|\leq U(2\mathrm{log}n)}\right)\\
&\geq \sum_{k=0}^{2^j-1} \frac{2^j}{n^2}\mathbb{E}\left((N_{jk}^+ - N_{jk}^-)^2 \mathbf{1}_{|\hat{\beta}_{jk}|\geq \sqrt{2\gamma(1+\varepsilon)s_{njk}\frac{\log n}{n}}+w(\varepsilon)\frac{2^{j/2}\log n}{n}} \mathbf{1}_{|u_{njk}|\leq U(2\mathrm{log}n)}\right).\\
&\geq \sum_{k=0}^{2^j-1} \frac{2^j}{n^2}\mathbb{E}\left((N_{jk}^+ - N_{jk}^-)^2 \mathbf{1}_{\frac{2^{\frac{j}{2}}}{n}|N_{jk}^+-N_{jk}^-|\geq \sqrt{2\gamma(1+\varepsilon)\frac{2^j}{n}\left(N_{jk}^++N_{jk}^-\right)\frac{\log n}{n}}+w(\varepsilon)\frac{2^{j/2}\log n}{n}} \mathbf{1}_{|u_{njk}|\leq U(2\mathrm{log}n)}\right)\\
&\geq \sum_{k=0}^{2^j-1} \frac{2^j}{n^2}\mathbb{E}\left((N_{jk}^+ - N_{jk}^-)^2 \mathbf{1}_{|N_{jk}^+-N_{jk}^-|\geq \sqrt{2\gamma(1+\varepsilon)\left(N_{jk}^++N_{jk}^-\right)\log n}+w(\varepsilon)\log n} \mathbf{1}_{|u_{njk}|\leq U(2\mathrm{log}n)}\right)\\
&\geq \frac{2^{2j}}{n^2}\mathbb{E}\left((N_{j1}^+ - N_{j1}^-)^2 \mathbf{1}_{|N_{j1}^+-N_{j1}^-|\geq \sqrt{2\gamma(1+\varepsilon)\left(N_{j1}^++N_{j1}^-\right)\log n}+w(\varepsilon)\log n} \mathbf{1}_{|u_{njk}|\leq U(2\mathrm{log}n)}\right).
\end{aligned}$$

Now, we consider a bounded sequence $(w_n)_n$ such that for any n, $w_n \geq w(\varepsilon)$ and such that $\frac{\sqrt{v_{nj}}}{2}$ is an integer with

$$v_{nj} = \left(\sqrt{4\gamma(1+\varepsilon)\tilde{\mu}_{nj}\log(n)} + w_n \log(n)\right)^2$$

and $\tilde{\mu}_{nj}$ is the largest integer smaller or equal to $n2^{-j-1}$. We have

$$v_{nj} \sim 4\gamma(1+\varepsilon)\tilde{\mu}_{nj}\log n$$

since

$$\frac{(\log n)^\alpha}{4} - 1 < n2^{-j-1} - 1 < \tilde{\mu}_{nj} \leq n2^{-j-1} \leq \frac{(\log n)^\alpha}{2}.$$

Now, set

$$l_{nj} = \tilde{\mu}_{nj} + \frac{1}{2}\sqrt{v_{nj}}, \quad m_{nj} = \tilde{\mu}_{nj} - \frac{1}{2}\sqrt{v_{nj}},$$

that are positive for n large enough. If $N_{j1}^+ = l_{nj}$ and $N_{j1}^- = m_{nj}$ then we have $N_{j1}^+ - N_{j1}^- = \sqrt{v_{nj}}$. Finally, we obtain that

$$\begin{aligned}
&\sum_{k=0}^{2^j-1} \mathbb{E}\left(\hat{\beta}_{jk}^2 \mathbf{1}_{|\hat{\beta}_{jk}|\geq \eta_{\gamma,jk}}\right)\\
&\geq \frac{2^{2j}}{n^2} v_{nj}\mathbb{P}\left(N_{j1}^+ = l_{nj}, \quad N_{j1}^- = m_{nj}, \quad |u_{njk}| \leq U(2\mathrm{log}n)\right)\\
&\geq v_{nj}(\log n)^{-2\alpha}\left[\mathbb{P}\left(N_{j1}^+ = l_{nj}, \quad N_{j1}^- = m_{nj}\right) - \mathbb{P}\left(|u_{njk}| > U(2\mathrm{log}n)\right)\right]\\
&\geq v_{nj}(\log n)^{-2\alpha}\left[\frac{n!}{l_{nj}!m_{nj}!(n-l_{nj}-m_{nj})!}p_j^{l_{nj}+m_{nj}}(1-2p_j)^{n-(l_{nj}+m_{nj})} - \frac{6}{n^2}\right]\\
&\geq v_{nj}(\log n)^{-2\alpha} \times \left[\frac{n!}{l_{nj}!m_{nj}!(n-2\tilde{\mu}_{nj})!}p_j^{2\tilde{\mu}_{nj}}(1-2p_j)^{n-2\tilde{\mu}_{nj}} - \frac{6}{n^2}\right], \qquad (30)
\end{aligned}$$

where

$$p_j = \int \mathbf{1}_{[2^{-j},(1+0.5)2^{-j}[}(x) f_0(x) dx = \int \mathbf{1}_{[(1+0.5)2^{-j},2^{-j+1}[}(x) f_0(x) dx = 2^{-j-1}.$$

Now, let us study each term of (30). We have

$$\begin{aligned} p_j^{2\tilde{\mu}_{nj}} &= \exp\left(2\tilde{\mu}_{nj}\log(p_j)\right) \\ &= \exp\left(2\tilde{\mu}_{nj}\log(2^{-j-1})\right), \end{aligned}$$

$$\begin{aligned} (1-2p_j)^{n-2\tilde{\mu}_{nj}} &= \exp\left((n-2\tilde{\mu}_{nj})\log(1-2p_j)\right) \\ &= \exp\left(-(n-2\tilde{\mu}_{nj})2^{-j} + o_n(1)\right) \\ &= \exp\left(-n2^{-j}\right)(1+o_n(1)), \end{aligned}$$

and

$$\begin{aligned} (n-2\tilde{\mu}_{nj})^{n-2\tilde{\mu}_{nj}} &= \exp\left((n-2\tilde{\mu}_{nj})\log(n-2\tilde{\mu}_{nj})\right) \\ &= \exp\left((n-2\tilde{\mu}_{nj})\left(\log n + \log\left(1-\frac{2\tilde{\mu}_{nj}}{n}\right)\right)\right) \\ &= \exp\left((n-2\tilde{\mu}_{nj})\log n - \frac{2\tilde{\mu}_{nj}(n-2\tilde{\mu}_{nj})}{n}\right)(1+o_n(1)) \\ &= \exp\left(n\log n - 2\tilde{\mu}_{nj} - 2\tilde{\mu}_{nj}\log n\right)(1+o_n(1)). \end{aligned}$$

Then, using the Stirling relation, $n! = n^n e^{-n}\sqrt{2\pi n}(1+o_n(1))$, we deduce that

$$\begin{aligned} \frac{n!}{(n-2\tilde{\mu}_{nj})!} p_j^{2\tilde{\mu}_{nj}}(1-2p_j)^{n-2\tilde{\mu}_{nj}} &= \frac{e^{n-2\tilde{\mu}_{nj}}}{e^n} \times \frac{n^n}{(n-2\tilde{\mu}_{nj})^{n-2\tilde{\mu}_{nj}}} \times p_j^{2\tilde{\mu}_{nj}}(1-2p_j)^{n-2\tilde{\mu}_{nj}} \times (1+o_n(1)) \\ &= \exp\left(-2\tilde{\mu}_{nj}\right) \times \frac{\exp\left(n\log n\right)}{(n-2\tilde{\mu}_{nj})^{n-2\tilde{\mu}_{nj}}} \times p_j^{2\tilde{\mu}_{nj}}(1-2p_j)^{n-2\tilde{\mu}_{nj}} \times (1+o_n(1)) \\ &= \exp\left(-2\tilde{\mu}_{nj}\right) \times \frac{\exp\left(n\log n + 2\tilde{\mu}_{nj}\log(2^{-j-1}) - n2^{-j}\right)}{\exp\left(n\log n - 2\tilde{\mu}_{nj} - 2\tilde{\mu}_{nj}\log n\right)}(1+o_n(1)) \\ &= \exp\left(2\tilde{\mu}_{nj}\log n + 2\tilde{\mu}_{nj}\log(2^{-j-1}) - n2^{-j}\right)(1+o_n(1)). \end{aligned}$$

It remains to evaluate $l_{nj}! \times m_{nj}!$:

$$\begin{aligned} l_{nj}! \times m_{nj}! &= \left(\frac{l_{nj}}{e}\right)^{l_{nj}} \left(\frac{m_{nj}}{e}\right)^{m_{nj}} \sqrt{2\pi l_{nj}}\sqrt{2\pi m_{nj}}(1+o_n(1)) \\ &= \exp\left(l_{nj}\log l_{nj} + m_{nj}\log m_{nj} - 2\tilde{\mu}_{nj}\right) \times 2\pi\tilde{\mu}_{nj}(1+o_n(1)). \end{aligned}$$

If we set

$$x_{nj} = \frac{\sqrt{v_{nj}}}{2\tilde{\mu}_{nj}} = o_n(1),$$

then

$$l_{nj} = \tilde{\mu}_{nj} + \frac{\sqrt{v_{nj}}}{2} = \tilde{\mu}_{nj}(1+x_{nj}),$$

$$m_{nj} = \tilde{\mu}_{nj} - \frac{\sqrt{v_{nj}}}{2} = \tilde{\mu}_{nj}(1-x_{nj}),$$

and using that

$$\begin{aligned}(1+x_{nj})\log(1+x_{nj}) &= (1+x_{nj})\left(x_{nj}-\frac{x_{nj}^2}{2}+\frac{x_{nj}^3}{3}+O(x_{nj}^4)\right)\\ &= x_{nj}-\frac{x_{nj}^2}{2}+\frac{x_{nj}^3}{3}+x_{nj}^2-\frac{x_{nj}^3}{2}+O(x_{nj}^4)\\ &= x_{nj}+\frac{x_{nj}^2}{2}-\frac{x_{nj}^3}{6}+O(x_{nj}^4),\end{aligned}$$

we obtain that

$$\begin{aligned}l_{nj}\log l_{nj} &= \tilde{\mu}_{nj}(1+x_{nj})\log\left(\tilde{\mu}_{nj}(1+x_{nj})\right)\\ &= \tilde{\mu}_{nj}(1+x_{nj})\log(1+x_{nj})+\tilde{\mu}_{nj}(1+x_{nj})\log\left(\tilde{\mu}_{nj}\right)\\ &= \tilde{\mu}_{nj}\left(x_{nj}+\frac{x_{nj}^2}{2}-\frac{x_{nj}^3}{6}+O(x_{nj}^4)\right)+\tilde{\mu}_{nj}(1+x_{nj})\log\left(\tilde{\mu}_{nj}\right).\end{aligned}$$

Similarly, we obtain that

$$m_{nj}\log m_{nj} = \tilde{\mu}_{nj}\left(-x_{nj}+\frac{x_{nj}^2}{2}+\frac{x_{nj}^3}{6}+O(x_{nj}^4)\right)+\tilde{\mu}_{nj}(1-x_{nj})\log\left(\tilde{\mu}_{nj}\right),$$

that implies that

$$\begin{aligned}l_{nj}\log l_{nj}+m_{nj}\log m_{nj} &= \tilde{\mu}_{nj}\left(x_{nj}^2+O(x_{nj}^4)\right)+2\tilde{\mu}_{nj}\log\left(\tilde{\mu}_{nj}\right)\\ &\leq \tilde{\mu}_{nj}x_{nj}^2+2\tilde{\mu}_{nj}\log(n2^{-j-1})+O(\tilde{\mu}_{nj}x_{nj}^4).\end{aligned}$$

Since

$$\tilde{\mu}_{nj}x_{nj}^2=\frac{v_{nj}}{4\tilde{\mu}_{nj}}\sim\gamma(1+\varepsilon)\log n,$$

we have, for n large enough,

$$\tilde{\mu}_{nj}x_{nj}^2+O(\tilde{\mu}_{nj}x_{nj}^4)\leq(\gamma+2\varepsilon)\log n$$

and

$$l_{nj}\log l_{nj}+m_{nj}\log m_{nj}\leq(\gamma+2\varepsilon)\log n+2\tilde{\mu}_{nj}\log(n2^{-j-1}).$$

Finally, we have

$$\begin{aligned}l_{nj}!\times m_{nj}! &= \exp\left(l_{nj}\log l_{nj}+m_{nj}\log m_{nj}-2\tilde{\mu}_{nj}\right)\times 2\pi\tilde{\mu}_{nj}(1+o_n(1))\\ &\leq \exp\left((\gamma+2\varepsilon)\log n+2\tilde{\mu}_{nj}\log(n2^{-j-1})-2\tilde{\mu}_{nj}\right)\times 2\pi\tilde{\mu}_{nj}(1+o_n(1)).\end{aligned}$$

Since $0<\varepsilon<\frac{1-\gamma}{2}<\frac{1}{2}$, we conclude that there exists $\delta<1$ such that

$$\begin{aligned}&\sum_{k=0}^{2^j-1}\mathbb{E}\left(\hat{\beta}_{jk}^2\mathbf{1}_{|\hat{\beta}_{jk}|\geq\eta_{\gamma,jk}}\right)\\ &\geq v_{nj}(\log n)^{-2\alpha}\left[\frac{\exp\left(2\tilde{\mu}_{nj}\log n+2\tilde{\mu}_{nj}\log(2^{-j-1})-n2^{-j}\right)}{\exp\left((\gamma+2\varepsilon)\log n+2\tilde{\mu}_{nj}\log(n2^{-j-1})-2\tilde{\mu}_{nj}\right)\times 2\pi\tilde{\mu}_{nj}}-\frac{6}{n^2}\right](1+o_n(1))\\ &\geq \frac{v_{nj}(\log n)^{-2\alpha}}{2\pi\tilde{\mu}_{nj}}\left[\exp\left(-(\gamma+2\varepsilon)\log n-2\right)-\frac{6}{n^2}\right](1+o_n(1))\\ &\geq \frac{2\gamma(1+\varepsilon)e^{-2}}{\pi}(\log n)^{1-2\alpha}n^{-(\gamma+2\varepsilon)}(1+o_n(1))\end{aligned}$$

and Lemma 4 is proved. ■

References

[1] Arlot S. and Massart P. (2008) *Data-driven calibration of penalties for least-squares regression*, technical report, arXiv:0802.0837.

[2] Asif M. S. and Romberg J. (2009) *Dantzig selector homotopy with dynamic measurements*, Proceedings of SPIE Computational Imaging VII.

[3] Bickel P., Ritov Y. and Tsybakov A. (2007) *Simultaneous analysis of Lasso and Dantzig selector*, To appear in Annals of Statistics.

[4] Birgé L. (2008) *Model selection for density estimation with $\mathbb{L}_2$-loss.* Submitted.

[5] Birgé L. and Massart P. (2007) *Minimal penalties for Gaussian model selection*, Probab. Theory Relat. Fields, 138(1-2), p. 33-73.

[6] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2006) *Aggregation and sparsity via ℓ_1 penalized least squares*, Proceedings of 19th Annual Conference on Learning Theory (COLT 2006), Lecture Notes in Artificial Intelligence v.4005 (Lugosi, G. and Simon, H.U.,eds.), Springer-Verlag, Berlin-Heidelberg.

[7] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007) *Sparse density estimation with l_1 penalties*, Lecture Notes in Artificial Intelligence, vol 4539, p. 530-543.

[8] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007) *Aggregation for Gaussian regression*, Annals of Statistics 35(4), p. 1674-1697.

[9] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007) *Sparsity Oracle Inequalities for the LASSO*, Electronic Journal of Statistics 1, p. 169-194.

[10] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2009) *Spades and Mixture Models.* Submitted.

[11] Bunea F, (2008) *Consistent selection via the Lasso for high dimensional approximating regression models*, IMS Lecture notes-Monograph Series, vol 3, p.122-137.

[12] Candès E. J. and Plan. Y. (2007) *Near-ideal model selection by l1 minimization.* To appear in Annals of Statistics.

[13] Candès E. J. and Tao T. (2007) *The Dantzig selector: statistical estimation when p is much larger than n.* Annals of Statistics. Volume 35, Number 6, p 2313–2351.

[14] Chen D., Donoho D.L. and Saunders M. (2001) *Atomic decomposition by basis pursuit*, SIAM review, 43, pp 129-159.

[15] Donoho D.L., Elad M. and Temlyakov V. (2006) *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Tran. on information Theory, 52, p.6-18.

[16] Donoho D.L. and Johnstone I.M. (1994) *Ideal spatial adaptation via wavelet shrinkage.*, Biometrika, 81, pp 425–455.

[17] Efron B., Hastie T., Johnstone I. and Tibshirani R. (2004) *Least angle regression*, Ann. Statist. 32, pp 407–499.

[18] Juditsky A. and Lambert-Lacroix S. (2004) *On minimax density estimation on* $\mathbb{R}$, Bernoulli, **10**(2) 187–220.

[19] Knight K. and Fu W. (2000) *Asymptotics for lasso-type estimators*, Ann. Statist., 28, no. 5, pp 1356–1378.

[20] Lounici K. (2008) *Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators*, Electronic Journal of statistics, vol 2.

[21] Massart P. (2007) *Concentration inequalities and model selection.* Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Springer, Berlin

[22] Meinshausen N. and Buhlmann P. (2006) *High dimensional graphs and variable selection with the Lasso*, Ann. Statist. 34, p. 1436-1462.

[23] Meinhausen N. and Yu B. (2009) *Lasso-type recovery of sparse representations for high-dimensional data*, Annals of Statistics, Volume 37, Number 1 , p 246-270.

[24] Osborne M.R., Presnell B. and Turlach B.A. (2000a) *On the Lasso and its dual*, Journal of Computational and Graphical Statistics, 9, 319-337.

[25] Osborne M.R., Presnell B. and Turlach B.A. (2000b) *A new approach to variable selection in least squares problems*, IMA Journal of Numerical Analysis, 20, 389-404.

[26] Reynaud-Bouret, P. and Rivoirard, V. (2009) *Calibration of thresholding rules for Poisson intensity estimation*, Technical report. http://arxiv.org/abs/0904.1148

[27] Reynaud-Bouret P., Rivoirard V. and Tuleau C. (2009) *On the influence of the support of functions for density estimation*, Technical report.

[28] Tibshirani R. (1996) *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistics Society, Series B, 58, 267-288.

[29] van de Geer S. (2008) *High dimensional generalized linear models and the Lasso*, Ann. Statist., 36(2), pp 614-645.

[30] Yu B. and Zhao P. (2006) *On model selection consistency of Lasso estimators*, Journal of Machine Learning Research 7, p. 2541-2567.

[31] Zhang C.H. and Huang J. (2007) *The sparsity and bias of the Lasso selection in high-dimensional linear regression*, To appear in Annals of Statistics.

[32] Zou H. (2006) *The adaptive Lasso and its oracle properties*, Journal of the American Statistical Association 101 n 476, 1418-1429.