



Sparsity considerations for dependent observations

Pierre Alquier, Paul Doukhan

► To cite this version:

Pierre Alquier, Paul Doukhan. Sparsity considerations for dependent observations. 2011. hal-00564291v1

HAL Id: hal-00564291

<https://hal.science/hal-00564291v1>

Preprint submitted on 8 Feb 2011 (v1), last revised 6 Aug 2011 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sparsity considerations for dependent observations

Pierre Alquier^(1,2) & Paul Doukhan⁽³⁾

(1) Université Paris 7, LPMA

(2) CREST-ENSAE, Laboratoire de Statistique

(3) Université de Cergy

Abstract

The aim of this paper is to provide a comprehensive introduction for the study of ℓ_1 -penalized estimators in the context of dependent observations. We define a general ℓ_1 -penalized estimator for solving problems of stochastic optimization. This estimator turns out to be the LASSO [Tib96] in the regression estimation setting. Powerful theoretical guarantees on the statistical performances of the LASSO were provided in recent papers, however, they usually only deal with the iid case. Here, we study our estimator under various dependence assumptions.

Contents

1	Introduction	2
1.1	Sparsity in high dimensional estimation problems	2
1.2	General setting and ℓ_1 -penalized estimator	3
1.3	Overview of the paper	4
2	Main result	4
2.1	Assumptions and result	4
2.2	Remarks on the density and regression estimation setting	6
3	Models fitting conditions of Theorem 2.1	7
3.1	Weak dependence ($\alpha = 0$)	7
3.1.1	Moment inequalities	8
3.1.2	Exponential inequalities	8
3.2	Long range dependence ($\alpha \in]0, \frac{1}{2}[$)	9
3.2.1	Power decays	9
3.2.2	Gaussian case	9
3.2.3	Non subGaussian tails	9
4	Application to regression estimation	10
4.1	Regression in the iid case	10
4.2	Regression estimation in the dependent case	11
4.2.1	Marcinkiewicz-Zygmund type inequalities	11

4.2.2	Exponential inequalities	12
4.3	Simulations	14
5	Application to density estimation	15
5.1	Density estimation in the iid case	15
5.2	Density estimation in the dependent case	16
6	Conclusion	17
7	Proofs	17

1 Introduction

1.1 Sparsity in high dimensional estimation problems

In the last few years, statistical problems in large dimension received a lot of attention. That is, estimation problems where the dimension of the parameter to be estimated, say p , is larger than the size of the sample, usually denoted by n . This setting is motivated by modern applications such as genomics, where we often have $n \leq 100$ the number of patients with a very rare disease, and p of the order of 10^5 or even 10^6 (CGH arrays), see [RBV08] and the references therein for example. Other examples appear in econometrics, see Belloni and Chernozhukov [BC09, BC10] for example.

Probably the most famous example is high dimensional regression estimation: one observes pairs (x_i, y_i) for $1 \leq i \leq n$ with $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$ and one wants to find a $\theta \in \mathbb{R}^p$ such that for a new pair (x, y) , $\theta'x$ would be a good prediction for y . If $p \geq n$, it is well known that a good estimation cannot be performed unless we make an additional assumption. Very often, it is quite natural to assume that most one can find such a θ that is sparse: most of its coordinates are equal to 0. If we let $\|\theta\|_0$ denote the number of non-zero coordinates in θ , this means that $\|\theta\|_0 \ll p$. In the genomics example, it means that only a few genes are relevant to explain the disease. The first estimators to deal with this kind of problems where the now famous AIC [Aka73] and BIC [Sch78]. Both can be written

$$\arg \min_{\theta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n (y_i - \theta'x_i)^2 + \lambda_n \|\theta\|_0 \right\} \quad (1)$$

where $\lambda_n > 0$ differs in AIC and BIC. These estimators were proved to have satisfying statistical properties, see for example [BTW07] where a nice oracle inequality for BIC is provided.

The main problem with this so-called ℓ_0 penalization approach is that the problem of the effective computation of the estimators defined in (1) is very time consuming. In practice, these estimators cannot be used for p more than a few tens. This motivated the study of the LASSO introduced by Tibshirani [Tib96]. This estimator is defined by

$$\arg \min_{\theta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n (y_i - \theta'x_i)^2 + \lambda_n \|\theta\|_1 \right\}. \quad (2)$$

The convexity of this minimization problem ensures that the estimator can be computed for very large p , see Efron *et al.* [EHJT04] for example. This

motivated a lot of theoretical studies on the statistical performances of this estimator. The results with the weakest hypothesis can be found in the work of Bickel *et al.* [BRT09] or Koltchinskii [Kol]. See also very nice reviews in the paper by Van de Geer and Bühlmann [vdGB09] or in the PhD Thesis of Hebiri [Heb09]. Also note that a quantity of variants of the idea of ℓ_1 -penalization were studied simultaneously to the LASSO: among others the basis pursuit [Che95, CDS01], the Dantzig Selector [CT07], the Elastic Net [ZH05]...

Another problem of estimation in high dimension is the so-called problem of sparse density estimation. In this setting, we observe n random variables with (unknown) density f and the purpose is to estimate f as a linear combination of some functions $\varphi_1, \dots, \varphi_p$. If $p \geq n$ and

$$f(\cdot) \simeq \sum_{j=1}^p \theta_j \varphi_j(\cdot)$$

we can use the SPADES (for sparse density estimator) by Bunea *et al.* [BWT07] or the iterative feature selection procedure in [Alq08].

One of the common feature of all the theoretical studies of sparse estimators is that they focus only on the case where the observations are independent. For example, for the density estimation case, in [BWT07] and [Alq08] the observations are assumed to be iid. The purpose of this paper is to propose a unified framework. Namely, we define a general stochastic optimization problem that contains as a special case regression and density estimation. We then define a general ℓ_1 -penalized estimator for this problem, in the special case of regression estimation this estimator is actually the LASSO and in the case of density estimation it is SPADES. Finally, we provide guarantees on the statistical performances of this estimator in the spirit of [BRT09], but we do not only consider independent observations: we want to study the case of dependent observations, and prove that we can still recover the target θ in this case, under various hypothesis.

1.2 General setting and ℓ_1 -penalized estimator

We now give the general setting and notations of our paper. Note that the cases of regression and density estimation will appear as particular cases.

We observe n random variables in $\mathcal{Z} : Z_1, \dots, Z_n$. Let \mathbb{P} be the distribution of (Z_1, \dots, Z_n) . We have a function $Q : \mathcal{Z} \times \mathbb{R}^p \rightarrow \mathbb{R}$ such that for any $z \in \mathcal{Z}$, $\theta \in \mathbb{R}^p \mapsto Q(z, \theta)$ is a quadratic function. The objective is the estimation of a value $\bar{\theta}$ that minimizes the following expression which only depends on n and θ :

$$R(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbb{E} Q(Z_i, \theta) = \int_{\mathcal{Z}^n} \frac{1}{n} \sum_{i=1}^n Q(z_i, \theta) d\mathbb{P}(z_1, \dots, z_n).$$

All the results that will follow are intended to be interesting in the case $p > n$ on the condition that $\|\bar{\theta}\|_0 := \text{card}\{j : \bar{\theta}_j \neq 0\}$ is small.

We use the following estimator:

$$\arg \min_{\theta \in \mathbb{R}^p} \left[\frac{1}{n} \sum_{i=1}^n Q(Z_i, \theta) + \lambda \|\theta\|_1 \right]$$

and $\hat{\theta}_\lambda$ denotes any solution of this minimization problem.

We now detail the notations in the two examples of interest:

1. in the regression example, $Z_i = (X_i, Y_i)$ with the $X_i \in \mathbb{R}^p$ deterministic, and

$$Y_i = X_i' \theta + \varepsilon_i \quad (3)$$

where $\mathbb{E}(\varepsilon_i) = 0$ (the ε_i are not necessarily iid, they may be dependent and have different distribution). Here we take $Q((x, y), \theta) = (y - x' \theta)^2$. In this example, $\hat{\theta}_\lambda$ is known as the LASSO estimator [Tib96].

2. in the density estimation case, $Z_i \in \mathbb{R}$ have the same density wrt Lebesgue measure (but they are not necessarily independent) and we have a family of functions $(\varphi_i)_{i=1}^p$ and we want to estimate the density f of Z_i by functions of the form

$$f_\theta(\cdot) = \sum_{i=1}^p \theta_i \varphi_i(\cdot).$$

In this case we take

$$Q(z, \theta) = \int f_\theta^2(\zeta) d\zeta - 2f_\theta(z)$$

and note that this leads to

$$R(\theta) = \int (f_\theta(x) - f(x))^2 dx - \int f^2(x) dx = \int (f_\theta(x) - f(x))^2 dx - \text{cst}.$$

Then $\hat{\theta}_\lambda$ is the estimator known as SPADES [BWT07].

1.3 Overview of the paper

In Section 2 we provide a sparsity inequality that extend the one of Bickel *et al.* [BRT09] to the case of non iid variables. This result involves two assumptions: one of them is about the function Q and is already needed in the iid case. The other one is more involved, it is specific to the non iid case.

In Section 3, we provide examples of classical assumptions on the observations that can ensure that this hypothesis is satisfied. These assumptions are expressed in terms of dependence coefficients.

We apply the results of Sections 2 and 3 to regression estimation in Section 4 and to density estimation in Section 5.

Finally the proofs are given Section 7.

2 Main result

2.1 Assumptions and result

First, we need an assumption on the quadratic form $R(\cdot)$.

Assumption A(κ) with $\kappa > 0$. As $R(\cdot)$ is a quadratic form, we have the matrix

$$\mathbf{M} = \frac{\partial^2 R(\theta)}{\partial \theta^2}$$

that does not depend on θ , and we assume that the matrix \mathbf{M} has only 1 on its diagonal (actually, this just means that we renormalize the observations X_i in the regression case, or the function φ_j in the density estimation case) and that it satisfies

$$\kappa \leq \inf_{\substack{v \in \mathbb{R}^p, J \subset \{1, \dots, p\} \\ \sum_{j \notin J} |v_j| \leq 3 \sum_{j \in J} |v_j| \\ |J| < \|\bar{\theta}\|_0}} \frac{v' \mathbf{M} v}{\sum_{j \in J} v_j^2}.$$

Note that this condition is already required in the iid setting, see [BRT09] and [vdGB09] for example. In these paper it is also discussed why we cannot hope to get rid of this hypothesis.

We set for simplicity

$$W_i^{(j)} = \frac{1}{2} \frac{\partial Q(Z_i, \bar{\theta})}{\partial \theta_j}, \quad i \in \{1, \dots, n\}, \quad j \in \{1, \dots, p\} \quad (4)$$

Recall that as $Q(z, \theta)$ is a quadratic function it may be written as $Q(z, \theta) = \theta' A(z) \theta + b(z)' \theta + c(z)$ for a $p \times p$ -matrix valued function A on \mathbb{R}^p and a vector function $b : \mathbb{R}^p \rightarrow \mathbb{R}^p$ so that

$$W_i^{(j)} = (A(Z_i) \bar{\theta})_j + \frac{1}{2} (b(Z_i))_j$$

Theorem 2.1 *Let us assume that Assumption $\mathbf{A}(\kappa)$ is satisfied. Let us assume that the distribution \mathbb{P} of (Z_1, \dots, Z_n) is such that there is a constant $\alpha \in [0, \frac{1}{2}]$ and a decreasing continuous function $\psi(\cdot)$ with*

$$\forall j \in \{1, \dots, p\}, \quad \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n W_i^{(j)} \right| \geq n^{-\frac{1}{2} + \alpha} t \right) \leq \psi(t). \quad (5)$$

Let us put

$$\lambda \geq \lambda^* := 2n^{\alpha - \frac{1}{2}} \psi^{-1} \left(\frac{\varepsilon}{p} \right).$$

Then

$$\mathbb{P} \left\{ \begin{array}{l} R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{4\lambda^2 \|\bar{\theta}\|_0}{\kappa} \\ \text{and, simultaneously} \\ \|\hat{\theta}_\lambda - \bar{\theta}\|_1 \leq \frac{8\lambda \|\bar{\theta}\|_0}{\kappa} \end{array} \right\} \geq 1 - \varepsilon.$$

The arguments of the proof of Theorem 2.1 are taken from [BRT09]. The proof is given in Section 7, page 17.

Note that the hypotheses of this theorem heavily depend on the distribution of the variables Z_1, \dots, Z_n , and particularly on their type of dependence.

Also note that the upper bound in the inequality is minimized if we make the choice $\lambda = \lambda^*$. Then

$$\mathbb{P} \left\{ \begin{array}{l} R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{16 \|\bar{\theta}\|_0 [\psi^{-1}(\varepsilon/p)]^2}{\kappa n^{1-2\alpha}} \\ \text{and} \\ \|\hat{\theta}_\lambda - \bar{\theta}\|_1 \leq \frac{16 \|\bar{\theta}\|_0 [\psi^{-1}(\varepsilon/p)]^2}{\kappa n^{\frac{1}{2}-\alpha}} \end{array} \right\} \geq 1 - \varepsilon.$$

It is important to remark that the choice $\lambda = 2n^{\alpha-\frac{1}{2}}\psi^{-1}\left(\frac{\varepsilon}{p}\right)$ may be impossible in practice, as the practitioner may ignore α and $\psi(\cdot)$. Moreover, this choice is not necessarily the best one in practice: in the regression case with iid noise $\mathcal{N}(0, \sigma^2)$, we will see that this choice leads to $\lambda = 2\sigma\sqrt{2n\log(p/\varepsilon)}$. This choice requires the knowledge of σ . Moreover it is not usually the best choice in practice, see for example the simulations in [Heb09]. Even in the iid case, the choice of a good λ in practice is still an open problem. However, note that

1. the question is in some sense meaningless. For example the value of λ that minimizes the quadratic risk $R(\hat{\theta}_\lambda)$ is not the same than the value of λ that may ensure, under some supplementary hypothesis, that $\hat{\theta}_\lambda$ identifies correctly the non-zero coordinates in $\bar{\theta}$, see for example Leeb and Pötscher [LP05] on that topic. One has to be careful to what one means when one say a good choice for $\lambda_{\hat{\theta}}$.
2. some popular methods like cross-validation seem to give good results for the quadratic risk, at least in the iid case. An interesting open question is to know if one can prove theoretical results for cross validation in the non iid setting.
3. the LARS algorithm [EHJT04] compute $\hat{\theta}_\lambda$ for any $\lambda > 0$ in a very short time.

Our point of view here is that we should use Theorem 2.1 in the following way: we fix λ large enough (depending on what we know about the data, for example, $\lambda = \mathcal{L}n^{-\frac{1}{2}}\log(p/\varepsilon)$ with \mathcal{L} , a large enough constant). Then the theorem ensures the sparsity oracle inequality as soon as the data satisfy some suitable conditions. These conditions are discussed below.

2.2 Remarks on the density and regression estimation setting

First, note that in the regression setting (Equation 3), for any $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$ we have

$$W_i^{(j)} = (X_j)_i(Y_i - X_i'\bar{\theta}) = (X_j)_{i\varepsilon_i}. \quad (6)$$

Then, in the density estimation context,

$$\begin{aligned} W_i^{(j)} &= \int \varphi_j(x) f_{\bar{\theta}}(x) dx - \varphi_j(Z_i) = \int \varphi_j(x) f(x) dx - \varphi_j(Z_i) \\ &= \mathbb{E}[\varphi_j(Z_1)] - \varphi_j(Z_i). \end{aligned} \quad (7)$$

So, in both cases, the assumption given by Equation 5 is satisfied if we have a control of the deviation of empirical means to their expectation. In the next sections, we discuss some conditions to obtain such controls with dependent variables.

3 Models fitting conditions of Theorem 2.1

In this section, we give some results that allow to control the deviation of empirical means to their expectations for general (non iid) observations. The idea will be, in the next sections, to apply these results to the processes $W^{(j)} = (W_i^{(j)})_{1 \leq i \leq n}$ for $1 \leq j \leq p$. For the sake of simplicity, in this section, we deal with a generic process $V = (V_i)_{i \in \mathbb{Z}}$ and the applications are given in the next sections. Various examples of pairs (α, ψ) are given.

3.1 Weak dependence ($\alpha = 0$)

Definition 3.1 We put, for any process $(V_i)_{i \in \mathbb{Z}}$,

$$c_{V,m}(r) = \max_{1 \leq l < m} \sup_{\substack{t_1 \leq \dots \leq t_m \\ t_{l+1} - t_l \geq r}} |\text{cov}(V_{t_1} \dots V_{t_l}, V_{t_{l+1}} \dots V_{t_m})| \quad (8)$$

We precise in §-3.1.1 and in §-3.1.2 that suitable decays of those coefficients yield the conditions in Theorem 2.1. Those two sections will provide quite different forms of the function ψ .

Cases of bounded or non-bounded random variables are to be considered separately. Eg. weak dependence conditions from Dedecker *et al.* [DDL⁺07] imply that the coefficients (8) are bounded by a constant only depending on m multiplied by $\sup_i \|V_i\|_\infty^m \varepsilon(r)$. Set $M = \sup_i \|V_i\|_\infty$, then for example if η , or κ weak dependence or if strong mixing hold then

$$\begin{aligned} c_{V,m}(r) &\leq mM^m \eta_V(r) \\ &\leq m^2 M^{m-1} \kappa_V(r) \\ &\leq M^m \alpha_V(r). \end{aligned} \quad (9)$$

Now the previous strong mixing condition is completely hereditary through measurable images but this is not the case for the weak dependence conditions. If the random process (V_i) satisfies a weak dependence condition then $f(V_i)$ too if for some $L > 0$, $a > 1$,

$$\forall t, z : |f(z+t) - f(z)| \leq L(|z|^{a-1} \vee 1)|t| \quad (10)$$

and $K = \sup_j \|V_j\|_s < \infty$ for some $s > a$ then looking more precisely at the proofs in [DDL⁺07] page 13 yields

$$\begin{aligned} \eta_{f(V)}(r) &\leq 5LK \eta_V(r)^{\frac{s-a}{s-1}} \\ \kappa_{f(V)}(r) &\leq 5LK \kappa_V(r)^{\frac{s-a}{s+a-2}}. \end{aligned} \quad (11)$$

3.1.1 Moment inequalities

In Doukhan and Louhichi [DL99] it is proved that if for an even integer $2q$ we have

$$\exists C \geq 1 \text{ such that: } c_{V,2q}(r) \leq C(r+1)^{-q}, \quad \forall j \in [1,p], \quad \forall r \geq 0 \quad (12)$$

then Marcinkiewicz-Zygmund inequality follows:

$$\mathbb{E}((V_1 + \dots + V_n)^{2q}) = \mathcal{O}(n^q)$$

and thus $\alpha = 0$ and $\psi(t)$ is of the order of $1/t^{2q}$. However, explicit constants are needed in Theorem 2.1. We actually have the following result.

Proposition 3.1 *Assume that coefficients (8) fit the relation (12) for some integer $q \geq 1$, then Marcinkiewicz-Zygmund inequality follows*

$$\mathbb{E}[(V_1 + \dots + V_n)^{2q}] \leq C^q d_{2q}(2q)! n^q \quad (13)$$

where

$$d_m \equiv \frac{1}{m} \frac{(2m-2)!}{((m-1)!)^2}, \quad m = 2, 3, \dots$$

The proof follows [DL99], it is given in Section 7.

Remark 3.1 *Sharper constants a_{2q} are also derived in the proof (Equation 26, page 20), one may replace the constants $2d_2, 24d_4, 720d_6$ by 1, 4 and 17) and using the recursion (27) also improves the above mentioned bounds.*

Various inequalities of this type were derived for alternative dependences (see Doukhan [Dou94], Rio [Rio00] and Dedecker *et al.* [DDL⁺07] for an extensive bibliography which also covers the case of non integer exponents.

3.1.2 Exponential inequalities

Using the previous inequality, Doukhan and Louhichi [DL99] proved exponential inequalities that would lead to $\psi(t)$ in $\exp(-\sqrt{t})$. Doukhan and Neumann [DN07] use alternative cumulant techniques to get $\psi(t)$ in $\exp(-t^2)$ for suitable bounds of the previous covariances (8).

Theorem 3.2 [DN07] *Let $\Psi : \mathbb{N}^2 \rightarrow \mathbb{N}$ be one of the following functions:*

- (a) $\Psi(u, v) = 2v,$
- (b) $\Psi(u, v) = u + v,$
- (c) $\Psi(u, v) = uv,$
- (d) $\Psi(u, v) = \alpha(u + v) + (1 - \alpha)uv, \quad \text{for some } \alpha \in (0, 1).$

We assume that there exist constants $K, L_1, L_2 < \infty$, $\mu \geq 0$, and a nonincreasing sequence of real coefficients $(\rho(n))_{n \geq 0}$ such that, for all u -tuples (s_1, \dots, s_u) and all v -tuples (t_1, \dots, t_v) with $1 \leq s_1 \leq \dots \leq s_u \leq t_1 \leq \dots \leq t_v \leq n$ the following inequality is fulfilled:

$$|\text{cov}(X_{s_1} \dots X_{s_u}, X_{t_1} \dots X_{t_v})| \leq K^2 M^{u+v-2} \Psi(u, v) \rho(t_1 - s_u), \quad (14)$$

where

$$\sum_{s=0}^{\infty} (s+1)^k \rho(s) \leq L_1 L_2^k (k!)^\mu \quad \forall k \geq 0. \quad (15)$$

Then

$$\mathbb{P}(S_n \geq t) \leq \exp \left(- \frac{t^2/2}{A_n + B_n^{1/(\mu+2)} t^{(2\mu+3)/(\mu+2)}} \right), \quad (16)$$

where A_n can be chosen as any number greater than or equal to $\sigma_n^2 := \text{Var}(V_1 + \dots + V_n)$ and

$$B_n = 2(K \vee M) L_2 \left(\left(\frac{2^{4+\mu} n K^2 L_1}{A_n} \right) \vee 1 \right).$$

This result yields convenient bounds for the function ψ . A review paper by Olivier Wintenberger [Win09] is also of interest: it directly yields alternative results from our main result. In this paper, we do not intend to provide the reader with encyclopedic references but mainly to precise some ideas and techniques so that this will be developed in further papers.

3.2 Long range dependence ($\alpha \in]0, \frac{1}{2}[$)

3.2.1 Power decays

Assume that a centered series (V_i) satisfies $\sum_i \sup_k |\text{cov}(V_k, V_{k+i})| = \infty$ then $\alpha > 0$ may occur, eg. if

$$r(i) \equiv \sup_k |\text{cov}(V_k, V_{k+i})| \sim i^{-\beta}$$

for $\beta \in]0, 1]$ then $\text{var}(S_n) \sim n^{2-\beta}$; then $\alpha = (1 - \beta)/2$ holds.

3.2.2 Gaussian case

In the special case of Gaussian processes $(V_i)_i$, tails of $S_n^{(j)}$ are classically described because $S_n \sim \mathcal{N}(0, \sigma_n^2)$ and here $\psi(t) = \exp(-t^2)$. We thus may obtain simultaneously subGaussian tails and $\alpha = (1 - \beta)/2 > 0$.

3.2.3 Non subGaussian tails

Assume that for each i, j $G_i \sim \mathcal{N}(0, 1)$ and $(G_i)_i$ is a stationary Gaussian processes with

$$r(i) = \text{cov}(G_k, G_{k+i}) \sim ci^{-\beta}$$

Let $V_i = P(G_i)$ for a function with Hermite rank $m \geq 1$, and since

$$\text{cov}(H_m(G_0), H_m(G_i)) = m! (r(i))^m$$

their covariance series is non m -th summable in case $\beta \in]\frac{1}{m}, 1[$.

The case $P(x) = x^2 - 1$ and $\beta \in]\frac{1}{2}, 1[$ is investigated by using the following

expansion in the seminal work by Rosenblatt [Ros61].

Set R_n for the covariance matrix of the Gaussian random vector (G_1, \dots, G_n) :

$$\begin{aligned}\mathbb{E}e^{tn^{\beta-1}S_n} &= e^{-tn^{\beta}} \det^{-\frac{1}{2}}(I_n - 2tn^{\beta-1}R_n) \\ &= \exp\left(\frac{1}{2} \sum_{k=2}^{\infty} \frac{1}{k} (2tn^{\beta-1})^k \text{trace}(R_n)^k\right)\end{aligned}$$

Quoting that

$$n^{k(\beta-1)} \text{trace}(R_n)^k \rightarrow_{n \rightarrow \infty} c_k > 0$$

with

$$c_k = c^k \int_0^1 \cdots \int_0^1 |x_1 - x_2|^{-\beta} |x_2 - x_3|^{-\beta} \cdots |x_{k-1} - x_k|^{-\beta} |x_k - x_1|^{-\beta} dx_1 \cdots dx_k$$

this is thus clear that for small enough $|t| < \tau = \frac{1}{2} \sup_{k \geq 2, j \geq 1} \left(c_k^{(j)}\right)^{\frac{1}{k}}$,

$$\mathbb{E}e^{tn^{\beta-1}S_n^{(j)}} \rightarrow_{n \rightarrow \infty} \exp\left(\frac{1}{2} \sum_{k=2}^{\infty} (2t)^k \frac{c_k}{k}\right)$$

Here the conditions in the main theorem hold with $\psi(t) = e^{-t}$ and $\alpha = \frac{1}{2} - \beta > 0$ for any $M > 1/\tau$.

Extension to higher order Gaussian subordinated models. Analogously if $V_i = P(G_i)$ for a function with Hermite rank $m \geq 3$, and for Gaussian processes $(G_i)_i$ with non m -th summable covariance series. The above mentioned arguments from analytic functions theory however don't work anymore in this case; indeed the corresponding Laplace transforms are it is no more an analytic function around zero in this case. Anyway we guess that here $\psi(t) = \exp(-t^{\frac{2}{m}})$.

4 Application to regression estimation

In this section we apply Theorem 2.1 and the various examples of Section 3 to obtain results for regression estimation. Note that the results in the iid setting are already known, they are only given here for the sake of completeness, in order to provide comparison with the other cases.

Let us remind that in the regression case, we want to apply the results of Section 3 to

$$W_i^{(j)} = (X_j)_{i\varepsilon_i}.$$

For the sake of simplicity, in this whole session dedicated to regression, let us put

$$\max(X) := \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |(X_i)_j|.$$

4.1 Regression in the iid case

Under the usual assumption that the ε_i are iid and subGaussian,

$$\forall s, \quad \mathbb{E}[\exp(s\varepsilon_i^2)] \leq \exp\left(\frac{s^2}{2\sigma^2}\right)$$

for some known σ^2 , then we have

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n W_i^{(j)} \right| \geq \frac{t}{\sqrt{n}} \right) \leq \psi(t) = \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

So we can apply Theorem 2.1 in order to obtain the following well known result:

Corollary 4.1 ([BRT09]) *In the context of Equation 3, under Assumption $\mathbf{A}(\kappa)$, if the (ε_i) are iid and subGaussian with variance upper bounded by σ^2 , the choice $\lambda = 2\sigma\sqrt{2\log(p/\varepsilon)/n}$ leads to*

$$\mathbb{P} \left(R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{32\sigma^2 \|\bar{\theta}\|_0 \log \frac{p}{\varepsilon}}{\kappa n} \right) \geq 1 - \varepsilon.$$

4.2 Regression estimation in the dependent case

4.2.1 Marcinkiewicz-Zygmund type inequalities

Let us remark that, for any $1 \leq j \leq p$,

$$c_{W^{(j)},m}(r) \leq c_{\varepsilon,m}(r) \left(\max_{i,j} |(X_j)_i| \right)^m = \max(X)^m c_{\varepsilon,m}(r).$$

Thus, we apply Theorem 2.1 and Proposition 3.1 to obtain the following result.

Corollary 4.2 *In the context of Equation 3, under Assumption $\mathbf{A}(\kappa)$, if the (ε_i) satisfy, for some even integer $2q$,*

$$\exists C \geq 1 \text{ such that: } \quad \forall r \geq 0, \quad c_{\varepsilon,2q}(r) \leq C(r+1)^{-q},$$

the choice

$$\lambda = \frac{2C^{\frac{1}{2}} \max(X)^q}{\sqrt{n}} \left(\frac{d_{2q} q! p}{\varepsilon} \right)^{\frac{1}{2q}}$$

leads to

$$\mathbb{P} \left(R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{16C \max(X)^{2q} (d_{2q} q!)^{\frac{1}{q}} \|\bar{\theta}\|_0 p^{\frac{1}{q}}}{\kappa \varepsilon^{\frac{1}{q}} n} \right) \geq 1 - \varepsilon.$$

Remark 4.1 *This result aims at filling a gap for non subGaussian and non iid random variables.*

The result still allows to deal with the sparse case $p > n$ in case $q > 1$. In this case we deal with the case $p = n^{q/2}$ and we get a rate of convergence in probability $\mathcal{O}(1/\sqrt{n})$.

If $q = 1$ and $\frac{p}{n} \rightarrow 0$ the least squares methods apply which make such sparsity algorithms less relevant.

Moreover if $q < 1$ the present method is definitely not efficient. Hence the case of heavy tails such as considered in the paper by Bartkiewicz et al. [BJMW10] should neither use our result. Anyway, using least squares for heavy tailed models (without second order moments) does not look to be a good idea!

4.2.2 Exponential inequalities

Using Theorem 2.1 and Theorem 3.2 we prove the following result.

Corollary 4.3 *Let us assume that the (ε_i) satisfy the hypothesis of Theorem 3.2: let $\Psi : \mathbb{N}^2 \rightarrow \mathbb{N}$ be one of the functions of Theorem 3.2, we assume that there are constants $K, L_1, L_2 < \infty$, $\mu \geq 0$, and a nonincreasing sequence of real coefficients $(\rho(n))_{n \geq 0}$ such that, for all u -tuples (s_1, \dots, s_u) and all v -tuples (t_1, \dots, t_v) with $1 \leq s_1 \leq \dots \leq s_u \leq t_1 \leq \dots \leq t_v \leq n$ the following inequality is fulfilled:*

$$|\text{cov}(\varepsilon_{s_1} \dots \varepsilon_{s_u}, \varepsilon_{t_1} \dots \varepsilon_{t_v})| \leq K^2 M^{u+v-2} \Psi(u, v) \rho(t_1 - s_u),$$

where

$$\sum_{s=0}^{\infty} (s+1)^k \rho(s) \leq L_1 L_2^k (k!)^\mu \quad \forall k \geq 0.$$

Let c be a positive constant and let us put

$$\mathcal{C} := 4K^2 \max(X)^2 \Psi(1, 1) L_1 + c 2L_2 \max(X) (K \vee M) \left(\frac{2^{\mu+3}}{\Psi(1, 1)} \vee 1 \right).$$

If we assume that Let us assume that $\varepsilon > 0$, p and n are such that

$$p \leq \frac{\varepsilon}{2} \exp \left(\frac{c^2 n^{\frac{1}{\mu+2}}}{\mathcal{C}} \right)$$

then for

$$\lambda = 2 \sqrt{\frac{\mathcal{C} \log \left(\frac{2p}{\varepsilon} \right)}{n}}$$

we have

$$\mathbb{P} \left\{ R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{16\mathcal{C}}{\kappa} \frac{\|\bar{\theta}\|_0 \log \left(\frac{2p}{\varepsilon} \right)}{n} \right\} \geq 1 - \varepsilon.$$

So the rate is the same than in the iid case. The only difference is in the constant, and a restriction for very large values of p .

Proof: For the sake of shortness, let us put

$$C_1 = 4K^2 \max(X)^2 \Psi(1, 1) L_1 \text{ and } C_2 = 2L_2 \max(X) (K \vee M) \left(\frac{2^{\mu+3}}{\Psi(1, 1)} \vee 1 \right)$$

and note that $\mathcal{C} = C_1 + cC_2$. First, note that for any $j \in \{1, \dots, p\}$,

$$\begin{aligned} & \left| \text{cov} \left(W_{s_1}^{(j)} \dots W_{s_u}^{(j)}, W_{t_1}^{(j)} \dots W_{t_v}^{(j)} \right) \right| \\ & \leq \left(\sup_i X_i^{(j)} \right)^{u+v} K^2 M^{u+v-2} \Psi(u, v) \rho(t_1 - s_u) \\ & \leq \tilde{K}^2 \tilde{M}^{u+v-2} \Psi(u, v) \rho(t_1 - s_u) \end{aligned}$$

if we put $\tilde{K} = \max(X)K$ and $\tilde{M} = \max(X)M$. Using Theorem 3.2, we obtain for any j ,

$$\mathbb{P} \left(\left| \sum_i W_i^{(j)} \right| \geq t \right) \leq 2 \exp \left(- \frac{t^2/2}{A_n + B_n^{\frac{1}{\mu+2}} t^{\frac{2\mu+3}{\mu+2}}} \right)$$

where $A_n = \sigma_n^2 \leq 2n\tilde{K}^2\Psi(1,1)L_1$ and

$$B_n = 2(K \vee M)L_2 \left(\left(\frac{2^{4+\mu}nK^2L_1}{A_n} \right) \vee 1 \right),$$

in other words:

$$\mathbb{P} \left(\left| \sum_i W_i^{(j)} \right| \geq t \right) \leq 2 \exp \left(- \frac{t^2/2}{C_1n + C_2t^{\frac{2\mu+3}{\mu+2}}} \right).$$

Now, let us put $u = t/\sqrt{n}$, we obtain

$$\begin{aligned} \mathbb{P} \left(\left| \frac{1}{n} \sum_i W_i^{(j)} \right| \geq un^{-\frac{1}{2}} \right) \\ \leq 2 \exp \left(- \frac{nu^2/2}{C_1n + C_2n^{\frac{2\mu+3}{2\mu+4}}u^{\frac{2\mu+3}{\mu+2}}} \right) \\ \leq 2 \exp \left(- \frac{u^2/2}{C_1 + C_2n^{-\frac{1}{2\mu+4}}u^{\frac{2\mu+3}{\mu+2}}} \right). \end{aligned}$$

Remark that we cannot in general compute explicitly the inverse of this function but we decide to upper-bound the range for u :

$$u \leq c \cdot n^{\frac{1}{2\mu+4}}$$

In this case,

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_i W_i^{(j)} \right| \geq un^{-\frac{1}{2}} \right) \leq 2 \exp \left(- \frac{u^2/2}{C_1 + C_2c} \right) = 2 \exp \left(- \frac{u^2}{2\mathcal{C}} \right) =: \psi(u)$$

and so

$$\psi^{-1}(y) = \sqrt{\mathcal{C} \log \left(\frac{2}{y} \right)}.$$

So we can take, following Theorem 2.1,

$$\lambda = 2n^{-\frac{1}{2}}\psi^{-1} \left(\frac{\varepsilon}{p} \right) = 2n^{-\frac{1}{2}}\sqrt{\mathcal{C} \log \left(\frac{2p}{\varepsilon} \right)}$$

as soon as $\psi^{-1}(\varepsilon/p) < n^{1/(2\mu+4)}$. For example, for a fixed number of observations n and a fixed confidence level ε , we have the restriction:

$$p \leq \frac{\varepsilon}{2} \exp \left(\frac{cn^{\frac{1}{\mu+2}}}{\mathcal{C}} \right).$$

Under this condition we have, by Theorem 2.1,

$$\mathbb{P} \left\{ \begin{array}{c} R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{16\mathcal{C}}{\kappa} \frac{\|\bar{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n} \\ \text{and,} \\ \|\hat{\theta}_\lambda - \bar{\theta}\|_1 \leq \frac{16\mathcal{C}}{\kappa} \frac{\|\bar{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n^{\frac{1}{2}}} \end{array} \right\} \geq 1 - \varepsilon,$$

this ends the proof. \square

4.3 Simulations

In order to illustrate the results, we propose a very short simulation study. The purpose of this study is not to show the good performances of the estimator in practice or to give recipes for the choice of λ . The aim is more to show that the performances of the iid setting are likely to be obtained in the dependent setting if the dependence coefficients are small.

We use the following model:

$$Y_i = \theta' X_i + \varepsilon_i, \quad 1 \leq i \leq n = 30$$

where the X_i 's will be treated as a random design, but in practice will be iid vectors in \mathbb{R}^p with $p = 50$, with distribution $\mathcal{N}_p(0, \Sigma)$ where Σ is given by $\Sigma_{i,j} = 0.5^{|i-j|}$.

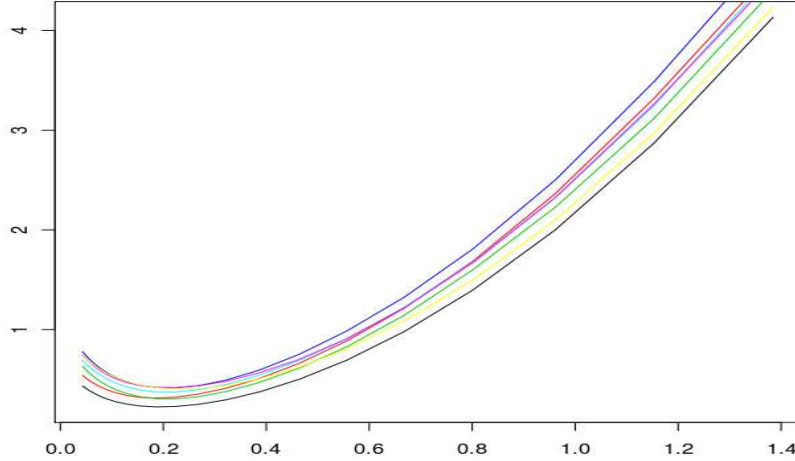


Figure 1: results of the experiments. The x -axis gives the value g where $\lambda = g\sqrt{\log(p)/n}$. The y -axis gives $\sum_{i=1}^n (\hat{\theta}'_\lambda X_i - \theta' X_i)^2$ the error of reconstruction of the signal. The color code is the following: $\vartheta = -0.95$: black, $\vartheta = -0.75$: red, $\vartheta = -0.5$: green, $\vartheta = 0$: blue, $\vartheta = 0.5$: cyan, $\vartheta = 0.75$: violet, $\vartheta = 0.95$: yellow.

The parameter is given by $\theta = (3, 1.5, 0, 0, 2, 0, 0, \dots)' \in \mathbb{R}^p$. This is the toy example used by Tibshirani [Tib96]. Let $\vartheta \in]-1, 1[$.

The noise satisfies $\varepsilon_i = \vartheta \varepsilon_{i-1} + \eta_i$, for $i \geq 2$, where the η_i are iid $\mathcal{N}(0, 1 - \vartheta^2)$ and $\varepsilon_1 \sim \mathcal{N}(0, 1)$. Note that this ensure that $\mathbb{E}(\varepsilon_i^2) = 1$ for any i , so the noise level does not depend on ϑ . In the experiments,

$$\vartheta \in \{-0.95, -0.75, -0.5, 0, 0.5, 0.75, 0.95\}.$$

We fixed a grid of values $\mathcal{G} \subset]0, 1.5[$ and we computed, for every experiment, the LASSO estimator with $\lambda = g\sqrt{\log(p)/n}$ for all $g \in \mathcal{G}$. We have repeated the experiment 25 times for every value of ϑ and report the results in Figure 1.

We can remark that all the curves are very similar. The minimum reconstruction error is obtained for $g \simeq 0.2$, that corresponds to $\lambda \simeq 0.072$. Note that in the iid case, it is smaller than the theoretical value given by Theorem 2.1, $\lambda = 2\sigma\sqrt{2\log(p/\varepsilon)/n} \simeq 1.28$ for $\varepsilon = 1/10$, that would correspond to $g \simeq 3.55$, a value that would not even stand in the figure!

5 Application to density estimation

Here we apply Theorem 2.1 and Section 3 to the context of density estimation. Let us remind that in this setting,

$$W_i^{(j)} = \mathbb{E}[\varphi_j(Z_1)] - \varphi_j(Z_i).$$

5.1 Density estimation in the iid case

If the Z_i are iid with density f and if $\|\varphi_j\|_\infty < B$ for any $j \in \{1, \dots, p\}$ then we can apply Hoeffding inequality [Hoe63] to upper bound

$$\left| \int \varphi_j(x) f(x) dx - \frac{1}{n} \sum_{i=1}^n \varphi_j(Z_i) \right|.$$

We obtain

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n W_i^{(j)} \right| \geq \frac{t}{\sqrt{n}} \right) \leq \psi(t) = 2 \exp(-\frac{t^2}{2B^2}).$$

So we can apply Theorem 2.1.

Corollary 5.1 *In the context of density estimation, under Assumption $\mathbf{A}(\kappa)$, if the Z_i are iid with density f and if $\|\varphi_j\|_\infty < B$ for any $j \in \{1, \dots, p\}$, the choice $\lambda = 2B\sqrt{2n \log(p/\varepsilon)}$ leads to*

$$\mathbb{P} \left(\int \left(f_{\hat{\theta}_\lambda}(x) - f_{\bar{\theta}}(x) \right)^2 dx \leq \frac{32B^2}{\kappa} \frac{\|\bar{\theta}\|_0 \log \frac{p}{\varepsilon}}{n} \right) \geq 1 - \varepsilon.$$

This result is essentially known, see [BWT07].

5.2 Density estimation in the dependent case

Note that if as previously we work with bounded $\varphi_j(\cdot)$, we automatically have moments of any order. So we will only state a result based on exponential inequality.

So, using Theorem 2.1 and Theorem 3.2 we obtain:

Corollary 5.2 *Let us assume that there are $L > 0$ and $B \geq 1$ such that $\varphi_j(\cdot)$ is L -Lipschitz and $\|\varphi_j\|_\infty < B$ for any $j \in \{1, \dots, p\}$. Let us assume that Z_1, \dots, Z_n satisfy*

$$\forall k \geq 0, \quad \sum_{s=0}^{\infty} (s+1)^k \eta_Z(s) \leq L_1 L_2^k (k!)^\mu$$

for some $L_1, L_2, \mu > 0$. Let us put a $c > 0$, define

$$\mathcal{C} := 4BLL_1 + (2^{3+\mu}BL_1)^{1/(\mu+2)}c$$

and assume that p, n and the confidence level ε are such that

$$p \leq \frac{\varepsilon}{2} \exp\left(\frac{c^2 n^{\frac{1}{\mu+2}}}{\mathcal{C}}\right).$$

Then

$$\mathbb{P}\left(\int \left(f_{\hat{\theta}_\lambda}(x) - f_{\bar{\theta}}(x)\right)^2 dx \leq \frac{16\mathcal{C}}{\kappa} \frac{\|\bar{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n}\right) \geq 1 - \varepsilon.$$

Proof: As $\varphi(j)$ is K -Lipschitz, we have:

$$\eta_{\varphi_j(Z)}(r) \leq L\eta_Z(r).$$

So we have

$$\forall k \geq 0, \quad \sum_{k=1}^{\infty} (s+1)^k \eta_{\varphi_j(Z)}(r) \leq LL_1 L_2^k (k!)^\mu.$$

Moreover, following Remark 9 page 9 in Doukhan and Neumann [DN07],

$$|\text{cov}(\varphi_j(Z_{s_1}) \cdots \varphi_j(Z_{s_u}), \varphi_j(Z_{t_1}) \cdots \varphi_j(Z_{t_v}))| \leq B^{u+v-1}(u+v)L \cdot \eta_Z(r).$$

So we can apply Theorem 3.2 with $\Psi(u, v) = u + v$ and we obtain

$$\mathbb{P}\left(\left|\sum_{i=1}^n W_i^{(j)}\right| > t\right) \leq 2 \exp\left(\frac{-t^2/2}{A_n + B_n^{\frac{1}{\mu+2}} t^{\frac{2\mu+3}{\mu+2}}}\right)$$

with $A_n = 4nBLL_1$ and

$$B_n = 2^{3+\mu}BL_1,$$

in other words

$$\mathbb{P}\left(\left|\sum_{i=1}^n W_i^{(j)}\right| > t\right) \leq 2 \exp\left(\frac{-t^2/2}{4nBLL_1 + (2^{3+\mu}BL_1)^{\frac{1}{\mu+2}} t^{\frac{2\mu+3}{\mu+2}}}\right).$$

We then put $u = t\sqrt{n}$ to obtain

$$\begin{aligned} \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n W_i^{(j)} \right| > \frac{u}{\sqrt{n}} \right) \\ \leq 2 \exp \left(\frac{-nu^2/2}{4BLL_1 + (2^{3+\mu}BL_1)^{\frac{1}{\mu+2}} n^{-\frac{1}{2\mu+4}} u^{\frac{2\mu+3}{\mu+2}}} \right). \end{aligned}$$

Here again, if we have

$$u \leq cn^{1/(2\mu+4)}$$

then

$$\begin{aligned} \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n W_i^{(j)} \right| > \frac{u}{\sqrt{n}} \right) \\ \leq 2 \exp \left(\frac{-nu^2/2}{4BLL_1 + (2^{3+\mu}BL_1)^{\frac{1}{\mu+2}} c} \right) 2 \exp \left(-\frac{nu^2}{2C} \right) =: \psi(u). \end{aligned}$$

So we take, following Theorem 2.1,

$$\lambda = \frac{2}{\sqrt{n}} \psi^{-1} \left(\frac{\varepsilon}{p} \right) = 2 \sqrt{\frac{C \log \left(\frac{2p}{\varepsilon} \right)}{n}}$$

and we obtain, with probability at least $1 - \varepsilon$,

$$\int \left(f_{\hat{\theta}_\lambda}(x) - f_{\bar{\theta}}(x) \right)^2 dx \leq \frac{16C}{\kappa} \frac{\|\bar{\theta}\|_0 \log \left(\frac{2p}{\varepsilon} \right)}{n}.$$

□

6 Conclusion

In this paper, we showed how the LASSO and other ℓ_1 -penalized methods can be extended to the case of dependent random variables.

An open and ambitious question to be addressed later is to find a good data-driven way to calibrate the regularization parameter λ when we don't know in advance the dependence coefficients of our observations.

Anyway this first step with sparsity in the dependent setting is done for accurate applications and our brief simulations let us think that such techniques are reasonable for time series.

Here again extensions to random fields or to dependent point processes seem plausible.

7 Proofs

Proof of Theorem 2.1: By definition,

$$\frac{1}{n} \sum_{i=1}^n Q(Z_i, \hat{\theta}_\lambda) + \lambda \|\hat{\theta}_\lambda\|_1 \leq \frac{1}{n} \sum_{i=1}^n Q(Z_i, \bar{\theta}) + \lambda \|\bar{\theta}\|_1$$

and so

$$R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \int_{\mathcal{Z}^n} \frac{1}{n} \left\{ \sum_{i=1}^n [Q(z_i, \hat{\theta}_\lambda) - Q(z_i, \bar{\theta})] \right\} d\mathbb{P}(z_1, \dots, z_n) \\ - \frac{1}{n} \sum_{i=1}^n [Q(Z_i, \hat{\theta}_\lambda) - Q(Z_i, \bar{\theta})] + \lambda (\|\bar{\theta}\|_1 - \|\hat{\theta}_\lambda\|_1). \quad (17)$$

Now, as Q is quadratic wrt θ we have

$$Q(z, \hat{\theta}_\lambda) = Q(z, \bar{\theta}) + (\hat{\theta}_\lambda - \bar{\theta})' \frac{\partial Q(z, \bar{\theta})}{\partial \theta} + \frac{1}{2} (\hat{\theta}_\lambda - \bar{\theta})' M (\hat{\theta}_\lambda - \bar{\theta}). \quad (18)$$

Moreover, as $\bar{\theta}$ is the minimizer of $R(\cdot)$, we have the relation

$$\frac{\partial R(\bar{\theta})}{\partial \theta} = \int_{\mathcal{Z}^n} \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(Z_i, \bar{\theta})}{\partial \theta} d\mathbb{P}(z_1, \dots, z_n) = 0. \quad (19)$$

Plugging (18) and (19) into (17) leads to

$$R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq (\hat{\theta}_\lambda - \bar{\theta})' \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(z, \bar{\theta})}{\partial \theta} + \lambda (\|\bar{\theta}\|_1 - \|\hat{\theta}_\lambda\|_1)$$

and then

$$R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \|\hat{\theta}_\lambda - \bar{\theta}\|_1 \sup_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(z, \bar{\theta})}{\partial \theta_j} \right| + \lambda (\|\bar{\theta}\|_1 - \|\hat{\theta}_\lambda\|_1). \quad (20)$$

Now, we remind that we have the hypothesis

$$\forall j \in \{1, \dots, p\}, \quad \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(Z_i, \bar{\theta})}{\partial \theta_j} \right| \geq n^{\alpha - \frac{1}{2}} t \right) \leq \psi(t)$$

that becomes, with a simple union bound argument,

$$\mathbb{P} \left(\sup_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(Z_i, \bar{\theta})}{\partial \theta_j} \right| \geq n^{\alpha - \frac{1}{2}} t \right) \leq p\psi(t)$$

and so, if we put $t = \psi^{-1}(\varepsilon/p)$,

$$\mathbb{P} \left(\sup_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(Z_i, \bar{\theta})}{\partial \theta_j} \right| \geq n^{\alpha - \frac{1}{2}} \psi^{-1} \left(\frac{\varepsilon}{p} \right) \right) \leq \varepsilon.$$

Also remark that $n^{\alpha - 1/2} \psi^{-1}(\varepsilon/p) = \lambda^*/2 \leq \lambda/2$. So until the end of the proof, we will work on the event

$$\left\{ \omega \in \Omega : \sup_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(Z_i(\omega), \bar{\theta})}{\partial \theta_j} \right| \leq \frac{\lambda}{2} \right\}$$

true with probability at least $1 - \varepsilon$. Going back to 20, we have

$$R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{\lambda}{2} \|\hat{\theta}_\lambda - \bar{\theta}\|_1 + \lambda (\|\bar{\theta}\|_1 - \|\hat{\theta}_\lambda\|_1)$$

and then

$$\begin{aligned}
R(\hat{\theta}_\lambda) - R(\bar{\theta}) + \frac{\lambda}{2} \|\hat{\theta}_\lambda - \bar{\theta}\|_1 &\leq \lambda \left(\|\hat{\theta}_\lambda - \bar{\theta}\|_1 + \|\bar{\theta}\|_1 - \|\hat{\theta}_\lambda\|_1 \right) \\
&= \lambda \left(\sum_{j=1}^p |(\hat{\theta}_\lambda)_j - \bar{\theta}_j| + \sum_{j=1}^p (|\bar{\theta}_j| - |(\hat{\theta}_\lambda)_j|) \right) \\
&= \lambda \left(\sum_{j: \bar{\theta}_j \neq 0} |(\hat{\theta}_\lambda)_j - \bar{\theta}_j| + \sum_{j: \bar{\theta}_j \neq 0} (|\bar{\theta}_j| - |(\hat{\theta}_\lambda)_j|) \right)
\end{aligned}$$

that leads to the following inequality that will play a central role in the end of the proof:

$$R(\hat{\theta}_\lambda) - R(\bar{\theta}) + \frac{\lambda}{2} \|\hat{\theta}_\lambda - \bar{\theta}\|_1 \leq 2\lambda \sum_{j: \bar{\theta}_j \neq 0} |(\hat{\theta}_\lambda)_j - \bar{\theta}_j|. \quad (21)$$

First, if we remind that $R(\hat{\theta}_\lambda) - R(\bar{\theta}) \geq 0$, (21) leads to

$$\|\hat{\theta}_\lambda - \bar{\theta}\|_1 \leq 4 \sum_{j: \bar{\theta}_j \neq 0} |(\hat{\theta}_\lambda)_j - \bar{\theta}_j|$$

and so

$$\sum_{j: \bar{\theta}_j = 0} |(\hat{\theta}_\lambda)_j - \bar{\theta}_j| \leq 3 \sum_{j: \bar{\theta}_j \neq 0} |(\hat{\theta}_\lambda)_j - \bar{\theta}_j|.$$

So we can take $v := \hat{\theta}_\lambda - \bar{\theta}$ in Assumption $\mathbf{A}(\kappa)$. So, (21) leads to

$$\begin{aligned}
R(\hat{\theta}_\lambda) - R(\bar{\theta}) + \frac{\lambda}{2} \|\hat{\theta}_\lambda - \bar{\theta}\|_1 &\leq 2\lambda \sum_{j: \bar{\theta}_j \neq 0} |(\hat{\theta}_\lambda)_j - \bar{\theta}_j| \\
&\leq 2\lambda \left(\|\bar{\theta}\|_0 \sum_{j: \bar{\theta}_j \neq 0} [(\hat{\theta}_\lambda)_j - \bar{\theta}_j]^2 \right)^{\frac{1}{2}} \\
&\leq 2\lambda \left(\frac{\|\bar{\theta}\|_0}{\kappa} (\hat{\theta}_\lambda - \bar{\theta})' \frac{\mathbf{M}}{2} (\hat{\theta}_\lambda - \bar{\theta}) \right)^{\frac{1}{2}} \\
&= 2\lambda \left(\frac{\|\bar{\theta}\|_0}{\kappa} [R(\hat{\theta}_\lambda) - R(\bar{\theta})] \right)^{\frac{1}{2}}. \quad (23)
\end{aligned}$$

We conclude that

$$R(\hat{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{4\lambda^2 \|\bar{\theta}\|_0}{\kappa}.$$

Remark that plugging this result into (22) to (23) gives

$$\|\hat{\theta}_\lambda - \bar{\theta}\|_1 \leq \frac{8\lambda \|\bar{\theta}\|_0}{\kappa}.$$

This ends the proof. \square

Proof of Proposition 3.1: First

$$\left| \mathbb{E}(W_1^{(j)} + \dots + W_n^{(j)})^\ell \right| \leq \ell! A_{\ell,n}^{(j)} \equiv \ell! \sum_{1 \leq k_1, \dots, k_\ell \leq n} \left| \mathbb{E} W_{k_1}^{(j)} \dots W_{k_\ell}^{(j)} \right|.$$

The same combinatorial arguments as in [DL99] yield for $p \leq 2q$

$$A_{\ell,n}^{(j)} \leq C_{q,\ell,n}^{(j)} + \sum_{m=2}^{\ell-2} A_{m,n}^{(j)} A_{\ell-m,n}^{(j)}, \quad \text{where} \quad (24)$$

$$C_{q,\ell,n}^{(j)} \equiv (p-1)n \sum_{r=0}^{n-1} (r+1)^{\ell-2} c_{W^{(j)}, 2q}(r). \quad (25)$$

Let us now assume the condition (12) then

$$\begin{aligned} C_{q,\ell,n}^{(j)} &\leq C(\ell-1)n \sum_{r=0}^{n-1} (r+1)^{\ell-2-q} \\ &\leq C(\ell-1)n \int_2^{n+1} x^{\ell-2-q} dx, & \text{if } \ell \geq q+2 \\ &\leq C(q+1)(n+1)^2, & \text{if } \ell = q+2 \\ &\leq C \frac{\ell-1}{\ell-q-2} (n+1)^{\ell-q}, & \text{if } \ell \geq q+2 \\ &\leq C(\ell-1)n \int_1^n x^{\ell-2-q} dx, & \text{if } \ell < q+2 \\ &\leq C \frac{\ell-1}{q+2-\ell} n, \end{aligned}$$

A rough bound is thus $C_{q,\ell,n}^{(j)} \leq C(\ell-1)n^{(\ell-q) \vee 1}$ and we thus derive

$$\begin{aligned} A_{2,n}^{(j)} &\leq Cn, & A_{3,n}^{(j)} &\leq 2Cn, & A_{4,n}^{(j)} &\leq 4Cn^2 \\ A_{5,n}^{(j)} &\leq 8Cn^2, & A_{6,n}^{(j)} &\leq 17Cn^3. \end{aligned} \quad (26)$$

Now using precisely condition (12) with the relation (24) we see that if $a_2 = 1$, and $a_3 = 2$ then the sequence recursively defined as

$$a_m = m-1 + \sum_{k=2}^{m-2} a_k a_{m-k} \quad (27)$$

satisfies $A_m^{(j)} \leq a_m C^{\lceil \frac{m}{2} \rceil} n^{\lfloor \frac{m}{2} \rfloor}$. Remember that

$$d_m \equiv \frac{1}{m} \frac{(2m-2)!}{((m-1)!)^2}, \quad m = 2, 3, \dots \quad (28)$$

hence as in [DL99] we quote that

$$a_m \leq d_m$$

is less than the m -th Catalan number, d_m and this ends the proof. \square

References

- [Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Budapest: Akademia Kiado, 1973.
- [Alq08] P. Alquier. Density estimation with quadratic loss, a confidence intervals method. *ESAIM: P&S*, 12:438–463, 2008.
- [BC09] A. Belloni and V. Chernozhukov. ℓ_1 -penalized quantile regression in high-dimensional sparse models. preprint arXiv:0904.2931 to appear in the *Annals of Statistics*, 2009.
- [BC10] A. Belloni and V. Chernozhukov. High dimensional sparse econometric models: An introduction. to appear in the proceedings of ”Stats in the Chateau”, 2010.
- [BJMW10] K. Bartkiewicz, A. Jakubowski, T. Mikosch, and O. Wintenberger. Infinite variances stable limits for sums of dependent random variables. *Proba. Theory and Related Fields*, 2010.
- [BRT09] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.
- [BTW07] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007.
- [BWT07] F. Bunea, M. Wegkamp, and A. Tsybakov. Sparse density estimation with ℓ_1 penalties. *Proceedings of 20th Annual Conference on Learning Theory (COLT 2007) - Springer*, pages 530–543, 2007.
- [CDS01] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- [Che95] S. S. Chen. Basis pursuit, 1995. Stanford University.
- [CT07] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n . *The Annals of Statistics*, 35, 2007.
- [DDL⁺07] J. Dedecker, P. Doukhan, G. Lang, S. León R., J. R. and Louhichi, and C. Prieur. *Weak dependence: with examples and applications*, volume 190 of *Lecture Notes in Statistics*. Springer, New York, 2007.
- [DL99] P. Doukhan and S. Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.*, 84(2):313–342, 1999.
- [DN07] P. Doukhan and M. H. Neumann. Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Process. Appl.*, 117(7):878–903, 2007.
- [Dou94] P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. Properties and examples.

- [EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.
- [Heb09] M. Hebiri. Quelques questions de selection de variables autour de l’estimateur lasso, 2009. PhD Thesis, Université Paris 7 (in english).
- [Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
- [Kol] V. Koltchinskii. Sparsity in empirical risk minimization. *Annales de l’Institut Henri Poincaré, Probability and Statistics* (to appear).
- [LP05] H. Leeb and B. M. Pötscher. Sparse estimators and the oracle property, or the return of hedges’ estimator. Cowles Foundation Discussion Papers 1500, Cowles Foundation, Yale University, 2005.
- [RBV08] F. Rapaport, E. Barillot, and J.-P. Vert. Classification of array-CGH data using fused SVM. *Bioinformatics*, 24(13):1375,1382, 2008.
- [Rio00] E. Rio. *Théorie asymptotique pour des processus aléatoire faiblement dépendants*. SMAI, Mathématiques et Applications 31, Springer, 2000.
- [Ros61] M. Rosenblatt. Independence and dependence. *Proceeding 4th. Berkeley Symp. Math. Stat. Prob. Berkeley University Press*, pages 411–443, 1961.
- [Sch78] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.
- [Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.
- [vdGB09] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- [Win09] O. Wintenberger. Deviation inequalities for sums of weakly dependent time series. preprint arXiv:0911.1682, 2009.
- [ZH05] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2):301–320, 2005.