Electronic Journal of Statistics
Vol. 5 (2011) 750–774
ISSN: 1935-7524
DOI: 10.1214/11-EJS626

Sparsity considerations for dependent variables

Pierre Alquier*

CREST - ENSAE
3, avenue Pierre Larousse
92240 Malakoff France
and
LPMA - Université Paris 7
175, rue du Chevaleret
75205 Paris Cedex 13 France
e-mail: alquier@ensae.fr
url: http://alquier.ensae.net/

and

Paul Doukhan

Université de Cergy-Pontoise
Laboratoire de Mathématiques - Analyse, G'eométrie, Modélisation
Site de Saint-Martin
2, avenue Adolphe Chauvin
95302 Cergy-Pontoise Cedex France
e-mail: doukhan@u-cergy.fr
url: http://doukhan.u-cergy.fr/

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 this estimator under various dependence assumptions.

AMS 2000 subject classifications: Primary 62J07; secondary 62M10, 62J05, 62G07, 62G08.
Keywords and phrases: Estimation in high dimension, weak dependence, sparsity, deviation of empirical mean, penalization, LASSO, regression estimation, density estimation.

Received February 2011.

Contents

*Research partially supported by the French "Agence Nationale pour la Recherche" under grant ANR-09-BLAN-0128 "PARCIMONIE".
†The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

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 desease, and p of the order of 10^5 or even 10^6 (CGH arrays), see for example [RBV08] and the references therein. Other examples appear in econometrics, we refer the reader to Belloni and Chernozhukov [BC11a, BC11b].

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 there exists 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 desease. Early examples of estimators introduced to deal with this kind of problems include the now famous AIC [Aka73] and BIC [Sch78]. Both can be written

$$\arg\min_{\theta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}(y_i-\theta' x_i)^2+\lambda_n\|\theta\|_0\right\} \tag{1.1}$$

where $\lambda_n > 0$ differs in AIC and BIC. Despite AIC and BIC may give poor results when $p \geq n$ (see [BM01]), taking $\lambda \geq 2\sigma\log(p)$ leads to estimators with very satisfying statistical properties (σ^2 being the variance of the noise). See for example [BM01, BTW07] for such results, and [BGH09] in the case of unknown variance.

The main problem with this so-called ℓ_0 penalization approach is that the effective computation of the estimators defined in (1.1) 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

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

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 Koltchinksii [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 $\varphi_1, \ldots, \varphi_p$. 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, BTWB10] or the iterative feature selection procedure in [Alq08].

One of the common features 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 [BTWB10] 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 $Q : \mathcal{Z} \times \mathbb{R}^p \to \mathbb{R}$ 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 $\overline{\theta}$ 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 $\|\overline{\theta}\|_0 := \text{card}\{j : \overline{\theta}_j \neq 0\}$ 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
$$Y_i = X_i' \theta + \varepsilon_i \tag{1.2}$$
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). 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
$$Q(z, \theta) = \int f_\theta^2(\zeta) d\zeta - 2 f_\theta(z)$$

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 [BTWB10].

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: the first one is about the function Q and is already needed in the iid case. It is usually refered as Restricted Eigenvalue Property. The other hypothesis is more involved, it is specific to the non iid case. It roughly says that we are able to control the deviations of empirical means of dependent variables around their expectations.

In Section 3, we provide several examples of classical assumptions on the observations that can ensure that we have such a control. These assumptions are expressed in terms of weak dependence coefficients, so in the beginning of this section we briefly introduce weak dependence. We also provide some references.

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 in Section 7.

2. Main result

2.1. Assumptions and result

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

Assumption $\mathbf{A}(\kappa)$ *with $\kappa > 0$. As $Q(z, \cdot)$ is a quadratic form, we have the matrix*

$$\mathbf{M} = \frac{\partial^2}{\partial \theta^2} \frac{1}{n} \sum_{i=1}^{n} Q(Z_i, \theta)$$

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), that it is non-random (here again, this is easily checked in the two examples) and that it satisfies

$$\kappa \leq \inf \left\{ \frac{v' \mathbf{M} v}{\sum_{j \in J} v_j^2} \;\middle|\; \begin{array}{l} v \in \mathbb{R}^p, \quad J \subset \{1, \dots, p\}, \quad |J| < \|\overline{\theta}\|_0 \\ \sum_{j \notin J} |v_j| \leq 3 \sum_{j \in J} |v_j| \end{array} \right\}.$$

Note that this condition, usually referred as restricted eigenvalue property (REP), is already required in the iid setting, see [BRT09, vdGB09] for example. In these paper it is also discussed why we cannot hope to get rid of this hypothesis.

We set for simplicity

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

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 $b : \mathbb{R}^p \to \mathbb{R}^p$ so that

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

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*

$$\forall j \in \{1, \dots, p\}, \quad \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} W_i^{(j)}\right| \geq n^{-\frac{1}{2}+\alpha} t\right) \leq \psi(t). \tag{2.1}$$

Let us put

$$\lambda \geq \lambda^* := 4n^{\alpha-\frac{1}{2}}\psi^{-1}\left(\frac{\varepsilon}{p}\right).$$

Then

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

The arguments of the proof of Theorem 2.1 are taken from [BRT09]. The proof is given in Section 7, page 768.

Note that the hypothesis in this theorem heavily depend on the distribution of the variables $Z_1, \dots, Z_n$, and particulary on their type of dependence. Section 3 will provide some examples of situations where this hypothesis is satisfied.

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

$$\mathbb{P}\left\{\begin{array}{c} R(\hat{\theta}_\lambda) - R(\overline{\theta}) \leq \dfrac{64}{\kappa}\dfrac{\|\overline{\theta}\|_0\left[\psi^{-1}(\varepsilon/p)\right]^2}{n^{1-2\alpha}} \\ \text{and} \\ \|\hat{\theta}_\lambda - \overline{\theta}\|_1 \leq \dfrac{8}{\kappa}\dfrac{\|\overline{\theta}\|_0\left[\psi^{-1}(\varepsilon/p)\right]^2}{n^{\frac{1}{2}-\alpha}} \end{array}\right\} \geq 1 - \varepsilon.$$

It is important to remark that the choice $\lambda = 4n^{\alpha-\frac{1}{2}}\psi^{-1}\left(\frac{\varepsilon}{p}\right)$ may be impossible in practice, as the practitionner may not know 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 $\lambda = 4\sigma\sqrt{2n\log(p/\varepsilon)}$. 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 $\overline{\theta}$, 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* λ.
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 this setting. See also the bootstrap method proposed in [BC11b].
3. the LARS algorithm [EHJT04] compute $\hat{\theta}_\lambda$ for any $\lambda > 0$ in a very short time (coordinate descent algorithms [FHHT07] are valuable alternative to LARS).

2.2. Remarks on the density and regression estimation setting

First, note that in the regression setting (Equation 1.2), for any $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,p\}$ we have

$$W_i^{(j)} = (X_j)_i(Y_i - X_i'\overline{\theta}) = (X_j)_i\varepsilon_i.$$

Then, in the density estimation context,

$$W_i^{(j)} = \int \varphi_j(x) f_{\overline{\theta}}(x)dx - \varphi_j(Z_i) = \int \varphi_j(x) f(x)dx - \varphi_j(Z_i)$$
$$= \mathbb{E}[\varphi_j(Z_1)] - \varphi_j(Z_i).$$

So, in both cases, the assumption given by Equation 2.1 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) obsrevations. The idea will be, in the next sections, to apply these results to the processes $W^{(j)} = (W_i^{(j)})_{1\le i\le n}$ for $1 \le j \le p$. 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. We will use the classical notation

$$S_n = \sum_{i=1}^{n} V_i.$$

3.1. Weak dependence ($\alpha = 0$)

We are going to introduce some coefficients in order to control the dependence of the V_i. The first example of such coefficients are the α-mixing coefficients first introduced by Rosenblatt [Ros56],

$$\alpha_V(r) = \sup_{t\in\mathbb{Z}} \sup_{\substack{U \in \sigma(V_i,\, i\le t)\\ U' \in \sigma(V_i,\, i\ge t+r)}} \left|\mathbb{P}(U\cap U') - \mathbb{P}(U)\mathbb{P}(U')\right|.$$

The idea is that the faster $\alpha_V(r)$ decreases to 0, the less dependent are V_i and V_{i+r} for large r. Assumptions on the rate of decay allows to prove laws of large numbers and central limit theorems. Different mixing coefficients were then studied, we refer the reader to [Dou94, Rio00] for more details.

The main problem with mixing coefficients is that they exclude too many processes. It is easy to build a process V satisfying a central limit theorem with constant $\alpha_V(r)$, see [DDL+07] Chapter 1 for an example. This motivated the introduction of *weak dependence* coefficients. The monograph [DDL+07] provides a comprehensive introduction to the various weak dependence coefficients. Our purpose here is not to define all these coefficients, but rather to introduce some examples that allow to satisfy condition (2.1) in Theorem 2.1.

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

$$c_{V,m}(r) = \max_{1\le\ell<m} \sup_{\substack{t_1\le\cdots\le t_m\\ t_{\ell+1}-t_\ell\ge r}} \left|\operatorname{cov}\left(V_{t_1}\cdots V_{t_\ell}, V_{t_{\ell+1}}\cdots V_{t_m}\right)\right|. \tag{3.1}$$

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

Definition 3.2. Let us assume that for any $r\ge 0$, for any g_1 and g_2 respectively L_1 and L_2-Lipschitz, where eg.,

$$L_1 := \sup_{(x_1,\ldots,x_\ell)\neq(y_1,\ldots,y_\ell)} \frac{g_1(y_1,\ldots,y_\ell) - g_1(x_1,\ldots,x_\ell)}{|y_1-x_1|+\cdots+|y_\ell-x_\ell|}.$$

We also assume that for any $t_1\le\cdots\le t_\ell\le t_{\ell+1}\le\cdots\le t_m$ with $t_{\ell+1}-t_\ell\ge r$,

$$\left|\operatorname{cov}\left[g_1(V_{t_1},\ldots,V_{t_\ell}), g_2(V_{t_{\ell+1}},\ldots,V_{t_m})\right]\right| \le \psi(L_1,L_2,\ell,m-\ell)\eta_V(r)$$

with $\psi(L_1,L_2,\ell,\ell') = \ell L_1 + \ell' L_2$. Then V is said to be η-dependent with η-dependence coefficients $(\eta(r), r\ge 0)$.

Remark 3.1. Other functions $\psi(L_1, L_2, \ell, \ell')$ allow to define the λ, κ and ζ-dependence, see [DDL+07].

We finally provide some basic properties, proved in [DDL+07]. The following result allows a comparison between different type of coefficients.

Proposition 3.1. *If* $\sup_i \|V_i\|_\infty \leq M$ *then*

$$\begin{aligned} c_{V,m}(r) &\leq mM^m\eta_V(r) \\ &\leq M^m\alpha_V(r). \end{aligned} \tag{3.2}$$

Finally the following property will be useful in this paper.

Proposition 3.2. *If* V *is* η*-dependent and* f *is* L*-Lipschitz and bounded, then* $f(V)$ *is also* η*-dependent with*

$$\eta_{f(V)}(r) = L\eta_V(r).$$

3.1.1. Moment inequalities

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

$$\exists C \geq 1 \text{ such that:} \qquad c_{V,2q}(r) \leq C(r+1)^{-q}, \quad \forall r \geq 0 \tag{3.3}$$

then Marcinkiewicz-Zygmund inequality follows:

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

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

Proposition 3.3. *Assume that coefficients (3.1) fit the relation (3.3) for some integer* $q \geq 1$*, then Marcinkiewicz-Zygmund inequality follows*

$$\mathbb{E}\left[(V_1 + \cdots + V_n)^{2q}\right] \leq C^q d_{2q}(2q)! n^q$$

where

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

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

Remark 3.2. Sharper constants a_{2q} are also derived in the proof (Equation (7.10), page 771), one may replace the constants $2d_2, 24d_4, 720d_6$ by 1, 4 and 17 and using the recursion (7.11) also improves the above mentioned bounds.

Various inequalities of this type where 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 (3.1).

Theorem 3.1. *[DN07] Let us assume that* $\sup_i \|V_i\|_\infty \leq M$. *Let* $\Psi : \mathbb{N}^2 \to \mathbb{N}$ *be one of the following functions:*

(a) $\Psi(u,v) = 2v$,
(b) $\Psi(u,v) = u+v$,
(c) $\Psi(u,v) = uv$,
(d) $\Psi(u,v) = \alpha(u+v) + (1-\alpha)uv$, *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* $(s_1, \ldots, s_u)$ *and all v-tuples* $(t_1, \ldots, t_v)$ *with* $1 \leq s_1 \leq \cdots \leq s_u \leq t_1 \leq \cdots \leq t_v \leq n$ *the following inequality is fulfilled:*

$$|cov\ (V_{s_1} \cdots V_{s_u}, V_{t_1} \cdots V_{t_v})| \leq K^2 M^{u+v-2} \Psi(u,v) \rho(t_1 - s_u), \tag{3.4}$$

where

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

Then

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

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

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

Remark 3.3. One can easily check that if V is η-dependent then (3.4) is satisfied with Ψ as in *(b)*, $K^2 = M$ and $\rho(r) = \eta(r)$, see Remark 9 page 9 in [DN07]. So if V is η-dependent and $\eta(r)$ decreases fast enough to 0 then we have an exponential inequality.

This result yields convienient bounds for the function ψ. A recent paper by Olivier Wintenberger [Win10] 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 now that V is a centered series satisfies $\sum_i \sup_k |\mathrm{cov}(V_k, V_{k+i})| = \infty$ then $\alpha > 0$ may occur, eg. if

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

for $\beta \in]0, 1]$ then $\mathrm{var}(S_n) \sim n^{2-\beta}$; 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 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 that for each i, j, $G_i \sim \mathcal{N}(0, 1)$ and $(G_i)_i$ is a stationary Gaussian processes with, for some B, β,

$$r(i) = \mathrm{cov}(G_k, G_{k+i}) \sim B i^{-\beta}. \tag{3.5}$$

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

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

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 $(G_1, \ldots, G_n)$:

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

Quoting that

$$n^{k(\beta-1)} \ \mathrm{trace}\ (R_n)^k \to_{n\to\infty} c_k > 0$$

with

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

(B is given by Equation (3.5)), this is thus clear that for small enough $|t| < \tau = \frac{1}{2}\sup_{k\geq 2, j\geq 1}(c_k)^{\frac{1}{k}}$,

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

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

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,

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

for some known σ^2, then we have

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

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 1.2, under Assumption* $\mathbf{A}(\kappa)$*, if the* (ε_i) *are iid and subGaussian with variance upper bounded by* σ^2*, the choice* $\lambda = 4\sigma\sqrt{2\log(p/\varepsilon)/n}$ *leads to*

$$\mathbb{P}\left(R(\hat{\theta}_\lambda) - R(\overline{\theta}) \leq \frac{128\sigma^2}{\kappa}\frac{\|\overline{\theta}\|_0\log\frac{p}{\varepsilon}}{n}\right) \geq 1-\varepsilon.$$

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 3.3 to obtain the following result.

Corollary 4.2. *In the context of Equation 1.2, under Assumption* $\mathbf{A}(\kappa)$*, if the* (ε_i) *satisfy, for some even integer* $2q$*,*

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

the choice

$$\lambda = \frac{4C^{\frac{1}{2}} \max(X)^q}{\sqrt{n}} \left(\frac{d_{2q} q! p}{\varepsilon} \right)^{\frac{1}{2q}}$$

leads to

$$\mathbb{P}\left(R(\hat{\theta}_\lambda) - R(\overline{\theta}) \leq \frac{64C \max(X)^{2q} (d_{2q} q!)^{\frac{1}{q}}}{\kappa} \frac{\|\overline{\theta}\|_0 p^{\frac{1}{q}}}{\varepsilon^{\frac{1}{q}} n} \right) \geq 1 - \varepsilon.$$

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 $\frac{p}{n} \to 0$ the least squares methods apply which make such sparsity algorithms less relevant.

Moreover if $q < 1$ the present method is definitely not efficient. Hence in the case of heavy tails, such as considered in the paper by Bartkiewicz *et al.* [BJMW10], our results are useless. 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 3.1 we prove the following result.

Corollary 4.3. *Let us assume that the* (ε_i) *satisfy the hypothesis of Theorem 3.1: let* $\Psi : \mathbb{N}^2 \to \mathbb{N}$ *be one of the functions of Theorem 3.1, 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* $(s_1, \ldots, s_u)$ *and all v-tuples* $(t_1, \ldots, t_v)$ *with* $1 \leq s_1 \leq \cdots \leq s_u \leq t_1 \leq \cdots \leq t_v \leq n$ *the following inequality is fulfilled:*

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

where

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

Let c be a positive constant and let us put

$$\mathcal{C} := 4K^2 \max(X)^2 \Psi(1,1) L_1 + c2L_2 \max(X)(K \vee M) \left(\frac{2^{\mu+3}}{\Psi(1,1)} \vee 1 \right).$$

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

$$\lambda = 4\sqrt{\frac{\mathcal{C} \log\left(\frac{2p}{\varepsilon}\right)}{n}}$$

we have

$$\mathbb{P}\left\{ R(\hat{\theta}_\lambda) - R(\overline{\theta}) \leq \frac{64\mathcal{C}}{\kappa} \frac{\|\overline{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n} \right\} \geq 1 - \varepsilon.$$

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 shorteness, 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 $j \in \{1, \ldots, p\}$,

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

if we put $\tilde{K} = \max(X)K$ and $\tilde{M} = \max(X)M$. Using Theorem 3.1, 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

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

in other words:

$$\mathbb{P}\left(\left|\sum_i W_i^{(j)}\right| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{C_1 n + C_2 t^{\frac{2\mu+3}{\mu+2}}}\right).$$

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

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i W_i^{(j)}\right| \geq u n^{-\frac{1}{2}}\right) \leq 2\exp\left(-\frac{nu^2/2}{C_1 n + C_2 n^{\frac{2\mu+3}{2\mu+4}} u^{\frac{2\mu+3}{\mu+2}}}\right)$$
$$\leq 2\exp\left(-\frac{u^2/2}{C_1 + C_2 n^{-\frac{1}{2\mu+4}} u^{\frac{2\mu+3}{\mu+2}}}\right).$$

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

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

In this case,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i W_i^{(j)}\right| \geq u n^{-\frac{1}{2}}\right) \leq 2\exp\left(-\frac{u^2/2}{C_1 + C_2 c}\right) = 2\exp\left(-\frac{u^2}{2\mathcal{C}}\right) =: \psi(u)$$

and so

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

So we can take, following Theorem 2.1,

$$\lambda = 4n^{-\frac{1}{2}}\psi^{-1}\left(\frac{\varepsilon}{p}\right) = 4n^{-\frac{1}{2}}\sqrt{\mathcal{C}\log\left(\frac{2p}{\varepsilon}\right)}$$

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,

$$\mathbb{P}\left\{\begin{array}{c} R(\hat{\theta}_\lambda) - R(\overline{\theta}) \leq \dfrac{64\,\mathcal{C}}{\kappa}\dfrac{\|\overline{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n} \\ \text{and,} \\ \|\hat{\theta}_\lambda - \overline{\theta}\|_1 \leq \dfrac{8\,\mathcal{C}}{\kappa}\dfrac{\|\overline{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n^{\frac{1}{2}}} \end{array}\right\} \geq 1-\varepsilon,$$

this ends the proof. □

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 fixed 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 $\Sigma_{i,j} = 0.5^{|i-j|}$. The parameter is given by $\theta = (3, 1.5, 0, 0, 2, 0, 0, \ldots)' \in \mathbb{R}^p$. 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 depent on ϑ. In the experiments,

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

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

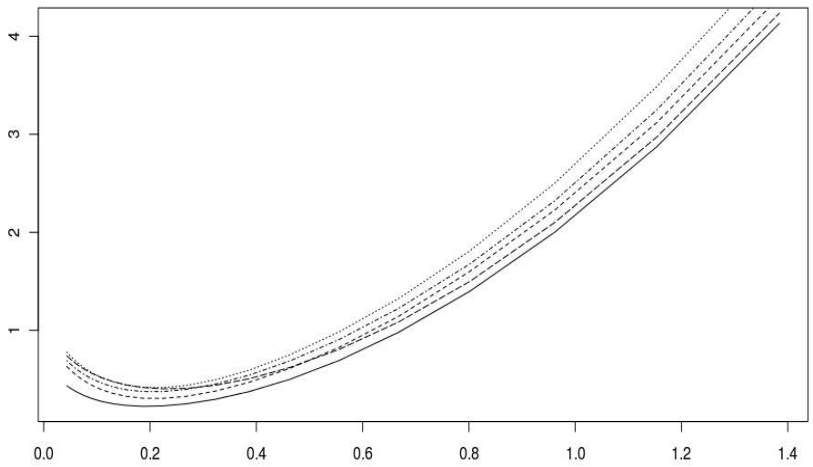

FIG 1. *Results of the experiments. The x-axis gives the value g where* $\lambda = g\sqrt{\log(p)/n}$. *The y-axis gives* $\sum_{i=1}^n (\hat{\theta}'_\lambda X_i - \theta' X_i)^2$ *the error of reconstruction of the signal. The lines code is the following:* $\vartheta = -0.95$*: solid line,* $\vartheta = -0.5$*: short dashed line,* $\vartheta = 0$*: dotted line,* $\vartheta = 0.5$*: dot/dash,* $\vartheta = 0.95$*: long dash.*

in the iid case, it is smaller than the theoretical value given by Theorem 2.1, $\lambda = 4\sigma\sqrt{2\log(p/\varepsilon)/n} \simeq 2.56$ for $\varepsilon = 1/10$, that would correspond to $g \simeq 7.10$, a value that would not event 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* $\lambda = 4B\sqrt{2n\log(2p/\varepsilon)}$ *leads to*

$$\mathbb{P}\left(\int \left(f_{\hat{\theta}_\lambda}(x) - f_{\overline{\theta}}(x) \right)^2 dx \leq \frac{128B^2}{\kappa} \frac{\|\overline{\theta}\|_0 \log\frac{2p}{\varepsilon}}{n} \right) \geq 1 - \varepsilon.$$

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 3.1 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

$$\mathbb{P}\left(\int \left(f_{\hat{\theta}_\lambda}(x) - f_{\overline{\theta}}(x)\right)^2 dx \leq \frac{64\,\mathcal{C}}{\kappa}\frac{\|\overline{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n}\right) \geq 1 - \varepsilon. \tag{5.1}$$

Remark 5.1. The assumption that the φ_j are all L-Lipschitz for a constant L excludes a lot of interesting dictionaries. If we assume that the φ_j are $L(n)$-Lipschitz (this would be the case if we used the first n functions in the Fourier basis for example), then we will suffer a loss in (5.1) when compared to the iid case. However, note that Equation (5.2) below is the starting point of our proof, so we cannot hope to find a simple way to remove this hypothesis when using η-weak dependence. This will be the object of a future work.

Proof. As φ_j is K-Lipschitz, using Proposition 3.2 we have:

$$\eta_{\varphi_j(Z)}(r) \leq L\eta_Z(r). \tag{5.2}$$

So we have

$$\forall k \geq 0, \quad \sum_{k=1}^{\infty}(s+1)^k \eta_{\varphi_j(Z)}(r) \leq LL_1L_2^k(k!)^\mu.$$

Moreover, following Remark 3.3,

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

So we can apply Theorem 3.1 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

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

Here again, if we have

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

then

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

So we take, following Theorem 2.1,

$$\lambda = \frac{4}{\sqrt{n}}\psi^{-1}\left(\frac{\varepsilon}{p}\right) = 4\sqrt{\frac{\mathcal{C}\log\left(\frac{2p}{\varepsilon}\right)}{n}}$$

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

$$\int \left(f_{\hat{\theta}_\lambda}(x) - f_{\overline{\theta}}(x)\right)^2 dx \leq \frac{64\mathcal{C}}{\kappa}\frac{\|\overline{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n}.$$

□

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 adressed 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,

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

and so

$$R(\hat{\theta}_\lambda) - R(\overline{\theta}) \leq \int_{\mathcal{Z}^n} \frac{1}{n}\left\{\sum_{i=1}^{n}\left[Q(z_i, \hat{\theta}_\lambda) - Q(z_i, \overline{\theta})\right]\right\} d\mathbb{P}(z_1, \ldots, z_n) - \frac{1}{n}\sum_{i=1}^{n}\left[Q(Z_i, \hat{\theta}_\lambda) - Q(Z_i, \overline{\theta})\right] + \lambda\left(\|\overline{\theta}\|_1 - \|\hat{\theta}_\lambda\|_1\right). \quad (7.1)$$

Now, as Q is quadratic wrt θ we have, for any z,

$$Q(z,\hat{\theta}_\lambda) = Q(z,\overline{\theta}) + (\hat{\theta}_\lambda - \overline{\theta})' \frac{\partial Q(z,\overline{\theta})}{\partial \theta} + \frac{1}{2}(\hat{\theta}_\lambda - \overline{\theta})' M (\hat{\theta}_\lambda - \overline{\theta}). \tag{7.2}$$

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

$$\frac{\partial R(\overline{\theta})}{\partial \theta} = \int_{\mathcal{Z}^n} \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(z_i,\overline{\theta})}{\partial \theta} d\mathbb{P}(z_1,\ldots,z_n) = 0. \tag{7.3}$$

Pluging (7.2) and (7.3) into (7.1) leads to

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

and then

$$R(\hat{\theta}_\lambda) - R(\overline{\theta}) \leq \|\hat{\theta}_\lambda - \overline{\theta}\|_1 \sup_{1\leq j\leq p} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial Q(Z_i,\overline{\theta})}{\partial \theta_j} \right| + \lambda \left(\|\overline{\theta}\|_1 - \|\hat{\theta}_\lambda\|_1 \right). \tag{7.4}$$

Now, we remind that we have the hypothesis

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

that becomes, with a simple union bound argument,

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

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

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

Also remark that $n^{\alpha-1/2}\psi^{-1}(\varepsilon/p) = \lambda^*/4 \leq \lambda/4$. So until the end of the proof, we will work on the event

$$\left\{ \omega \in \Omega : \sup_{1\leq j\leq p} \left| \frac{1}{n} \sum_{i=1}^n \frac{1}{2} \frac{\partial Q(Z_i(\omega),\overline{\theta})}{\partial \theta_j} \right| \leq \frac{\lambda}{4} \right\}$$

true with probability at least $1-\varepsilon$. Going back to (7.4), we have

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

and then

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

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

$$R(\hat{\theta}_\lambda) - R(\overline{\theta}) + \frac{\lambda}{2}\|\hat{\theta}_\lambda - \overline{\theta}\|_1 \leq 2\lambda\sum_{j:\overline{\theta}_j\neq 0}|(\hat{\theta}_\lambda)_j - \overline{\theta}_j|. \tag{7.5}$$

First, if we remind that $R(\hat{\theta}_\lambda) - R(\overline{\theta}) \geq 0$, (7.5) leads to

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

and so

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

So we can take $v := \hat{\theta}_\lambda - \overline{\theta}$ in Assumption $\mathbf{A}(\kappa)$. So, (7.5) leads to

$$R(\hat{\theta}_\lambda) - R(\overline{\theta}) + \frac{\lambda}{2}\|\hat{\theta}_\lambda - \overline{\theta}\|_1 \leq 2\lambda\sum_{j:\overline{\theta}_j\neq 0}|(\hat{\theta}_\lambda)_j - \overline{\theta}_j| \tag{7.6}$$
$$\leq 2\lambda\left(\|\overline{\theta}\|_0\sum_{j:\overline{\theta}_j\neq 0}[(\hat{\theta}_\lambda)_j - \overline{\theta}_j]^2\right)^{\frac{1}{2}}$$
$$\leq 2\lambda\left(\frac{\|\overline{\theta}\|_0}{\kappa}(\hat{\theta}_\lambda - \overline{\theta})'\frac{\mathbf{M}}{2}(\hat{\theta}_\lambda - \overline{\theta})\right)^{\frac{1}{2}}$$
$$= 2\lambda\left(\frac{\|\overline{\theta}\|_0}{\kappa}\left[R(\hat{\theta}_\lambda) - R(\overline{\theta})\right]\right)^{\frac{1}{2}}. \tag{7.7}$$

We conclude that

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

Now remark that (7.6) to (7.7) states that a convex quadratic function of $\left[R(\hat{\theta}_\lambda) - R(\overline{\theta})\right]$ is negative, so both roots of that quadratic are real. This leads to

$$\|\hat{\theta}_\lambda - \overline{\theta}\|_1 \leq \frac{2\lambda\|\overline{\theta}\|_0}{\kappa}.$$

This ends the proof. □

Proof of Proposition 3.3. First

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

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

$$A_{\ell,n}^{(j)} \le C_{q,\ell,n}^{(j)} + \sum_{m=2}^{\ell-2} A_{m,n}^{(j)} A_{\ell-m,n}^{(j)}, \text{ where} \tag{7.8}$$

$$C_{q,\ell,n}^{(j)} \equiv (p-1)n \sum_{r=0}^{n-1} (r+1)^{\ell-2} c_{W^{(j)},2q}(r). \tag{7.9}$$

Let us now assume the condition (3.3) then

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

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

$$\begin{array}{lllllll} A_{2,n}^{(j)} & \le & Cn, & A_{3,n}^{(j)} & \le & 2Cn, & A_{4,n}^{(j)} \le 4Cn^2 \\ A_{5,n}^{(j)} & \le & 8Cn^2, & A_{6,n}^{(j)} & \le & 17Cn^3. & \end{array} \tag{7.10}$$

Now using precisely condition (3.3) with the relation (7.8) we see that if $a_2 = 1$, and $a_3 = 2$ then the sequence recursively defined as

$$a_m = m - 1 + \sum_{k=2}^{m-2} a_k a_{m-k} \tag{7.11}$$

satisfies $A_m^{(j)} \le a_m C^{[\frac{m}{2}]} n^{[\frac{m}{2}]}$. Remember that

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

hence as in [DL99] we quote that

$$a_m \le d_m$$

is less that the m-th Catalan number, d_m and this ends the proof. □

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. MR2437718

[BC11a] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, 32(11):2011–2055, 2011.

[BC11b] A. Belloni and V. Chernozhukov. High dimensional sparse econometric models: An introduction. In P. Alquier, E. Gautier, and G. Stoltz, editors, *Inverse Problems and High-Dimensional Estimation*. Springer Lecture Notes in Statistics, 2011.

[BGH09] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Annals of Statistics*, 37(2):630–672, 2009. MR2502646

[BJMW10] K. Bartkiewicz, A. Jakubowskin, T. Mikosch, and O. Wintenberger. Infinite variances stable limits for sums of dependent random variables. *Probability Theory and Related Fields*, 2010.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001. MR1848946

[BRT09] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009. MR2533469

[BTW07] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007. MR2351101

[BTWB10] F. Bunea, A. Tsybakov, M. Wegkamp, and A. Barbu. SPADES and mixture models. *Annals of Statistics*, 38(4):2525–2558, 2010. MR2676897

[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. MR2397610

[CDS01] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001. MR1639094

[Che95] S. S. Chen. Basis pursuit, 1995. PhD Thesis, Stanford University.

[CT07] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35, 2007. MR2382644

[DDL+07] J. Dedecker, P. Doukhan, G. Lang, J. R. León R., S. Louhichi, and C. Prieur. *Weak dependence: with examples*

and applications, volume 190 of *Lecture Notes in Statistics*. Springer, New York, 2007. MR2338725

[DL99] P. Doukhan and S. Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84(2):313–342, 1999. MR1719345

[DN07] P. Doukhan and M. H. Neumann. Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, 117(7):878–903, 2007. MR2330724

[Dou94] P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. Properties and examples. MR1312160

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004. MR2060166

[FHHT07] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statist.*, 1(2):302–332, 2007. MR2415737

[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. MR0144363

[Kol] V. Koltchinskii. Sparsity in empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45(1):7–57, 2009. MR2500227

[LP05] H. Leeb and B. M. Pötscher. Sparse estimators and the oracle property, or the return of hodges' 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. MR2117923

[Ros56] M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proc. Nat. Ac. Sc. U.S.A.*, 42:43–47, 1956. MR0074711

[Ros61] M. Rosenblatt. Independence and dependence. *Proceeding 4th. Berkeley Symp. Math. Stat. Prob. Berkeley University Press*, pages 411–443, 1961. MR0133863

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978. MR0468014

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996. MR1379242

[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. MR2576316

[Win10] O. Wintenberger. Deviation inequalities for sums of weakly dependent time series. *Electronic Communications in Probability*, 15:489–503, 2010. MR2733373

[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. MR2137327