Journal of Statistical Planning and Inference 153 (2014) 11–26

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Lasso with long memory regression errors ☆

Abhishek Kaul

Department of Statistics and Probability, Michigan State University, C507, 619 Red Cedar Road, MI 48824, USA

ARTICLE INFO

Article history:
Received 19 June 2013
Received in revised form
22 April 2014
Accepted 6 May 2014
Available online 24 May 2014

Keywords:
Lasso
Sparsity
Long memory dependence
Sign consistency
Asymptotic normality

ABSTRACT

Lasso is a computationally efficient approach to model selection and estimation, and its properties are well studied when the regression errors are independent and identically distributed. We study the case, where the regression errors form a long memory moving average process. We establish a finite sample oracle inequality for the Lasso solution. We then show the asymptotic sign consistency in this setup. These results are established in the high dimensional setup $(p>n)$ where p can be increasing exponentially with n. Finally, we show the consistency, $n^{1/2-d}$–consistency of Lasso, along with the oracle property of adaptive Lasso, in the case where p is fixed. Here d is the memory parameter of the stationary error sequence. The performance of Lasso is also analysed in the present setup with a simulation study.

1. Introduction

In linear regression models, the classical least square approach is not feasible when the number of parameters is larger than the number of observations. However, in various scientific fields such high dimensional data sets are often common, such as in the field of genetics where data is collected for thousands of genes or proteins and financial data where a large number of financial instruments are tracked over time. For more examples see, e.g., the monograph of Bühlmann and van de Geer (2011), and the references therein. To overcome this problem, various parameter shrinkage methods have been proposed in the literature and one of the most successful has been the least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996), due to its desirable finite sample and asymptotic properties, and computational efficiency. Its statistical properties are well studied when the regression errors are independent and identically distributed (i.i.d.) random variables, see, e.g., Knight and Fu (2000), Meinhausen and Bühlmann (2006), Zhao and Yu (2006), Bickel et al. (2009) and Bühlmann and van de Geer (2011).

On the other hand in many problems of practical interest regression models with long memory errors arise naturally in the fields of econometrics and finance, see e.g., Beran (1992), Baillie (1996) and more recent monographs of Giraitis et al. (2012) (GKS), and Beran et al. (2013), and the numerous references therein. It is thus of interest to investigate the behaviour of Lasso in regression models with long memory errors.

Accordingly, let $X_i=(x_{i1},\dots,x_{ip})'$, $i=1,\dots,n$, be vectors of design variables, where for any vector a, a' denotes its transpose. Let Y_i's denote the responses, which are related to X_i's by the relations

$$Y_i=X_i'\beta+\varepsilon_i \quad \text{for some } \beta\in\mathbb{R}^p,\ 1\le i\le n. \tag{1.1}$$

☆ Research supported in part by the NSF-DMS Grant 1205271: PI. Hira L. Koul.
E-mail address: kaulabhi@stt.msu.edu

http://dx.doi.org/10.1016/j.jspi.2014.05.003
0378-3758/

The errors ε_i are assumed to be long memory moving average with i.i.d. innovations, i.e.,

$$\varepsilon_i = \sum_{k=1}^{\infty} a_k \zeta_{i-k} = \sum_{k=-\infty}^{i} a_{i-k}\zeta_k, \tag{1.2}$$

where $a_k = c_0 k^{-1+d}$, $\forall k \ge 1$, $0 < d < \frac{1}{2}$ and some constant $c_0 > 0$, and $a_k = 0$ for $k \le 0$. Also, $\zeta_j, j \in \mathbb{Z} := \{0, \pm 1, \pm 2, \ldots\}$, are i.i.d. r.v.'s with mean zero and variance σ_ζ^2. For notational convenience, we shall assume $c_0 = 1$ and $\sigma_\zeta^2 = 1$, without loss of generality. Also denote $X = (x_{ij})_{n \times p}$ as the design matrix, and $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)'$. Note that $\{\varepsilon_i, i \in \mathbb{Z}\}$ is a stationary process with autocovariance function

$$\gamma_\varepsilon(k) = \sum_{j=1}^{\infty} a_j a_{j+k} = k^{-1+2d} B(d, 1-2d)(1+o(1)), \quad 0 < d < 1/2, \ k \to \infty, \tag{1.3}$$

where $B(a,b) := \int_0^1 u^{a-1}(1-u)^{b-1}\, du$, $a > 0$, $b > 0$, see, e.g., Proposition 3.2.1(ii) in GKS.

Recall, say from Bühlmann and van de Geer (2011), that the Lasso estimate of β is defined as follows:

$$\hat{\beta}^n(\lambda) = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \|Y - X'\beta\|_2^2 + \lambda_n \|\beta\|_1 \right\}, \quad \lambda > 0, \tag{1.4}$$

where $Y = (Y_1, Y_2, \ldots, Y_n)'$ and $\|\beta\|_1 := \sum_{j=1}^{p} |\beta_j|$ denotes l_1 norm of $\beta = (\beta_1, \ldots, \beta_p)'$.

The literature in the area of regularized estimation with dependence considerations is scarce. The first paper dealing with this issue has been that of Alquier and Doukhan (2011). They provide finite sample error bounds under weak dependence structures on the model errors ε. Another recent paper addressing dependence concerns is that of Yoon et al. (2013). Their paper provides asymptotic results in the $n > p$ setup, in a linear regression models with stationary auto-regressive errors. Note that, in that paper the error process is assumed to be an AR(q) process, which is known to be a short memory process, i.e $\sum_{k=1}^{\infty} |\gamma_\varepsilon(k)| < \infty$, see GKS. In this paper we investigate the behaviour of Lasso under a stronger dependence structure and less restrictive model assumptions in comparison to the above-mentioned papers. In particular, we assign a long memory structure on the model errors ε, i.e $\sum_{k=1}^{\infty} |\gamma_\varepsilon(k)| = \infty$. We provide restrictions on the rate of increase of the design variables as well as the rate of increase of the dimension p in order to obtain the corresponding finite sample error bounds. We allow the design variables to grow with the restriction $\sum_{1 \le i \le n} x_{ij}^2 = O(n)$, and hence the results obtained can also easily be extended to the case of Gaussian random designs. Furthermore, all results proved in the high dimensional setup in this paper allow p to grow exponentially with n. In addition, we also discuss the aspect of sign consistency and asymptotic normality of the Lasso estimates.

The three main contributions of this paper are as follows. First, we show that the probability bound for a pre-defined set controlling the stochastic term $\max_{1 \le j \le p} |X_j' \varepsilon|$ can be obtained with a long memory moving average probability structure on ε, under appropriate restrictions on the rate of increase of the design variables and with the proper choice of the regularizer λ_n. Second, we obtain the sign consistency of Lasso under the long memory setup with standard restrictions on the design matrix X. These results are obtained in the high dimensional setup, where p can grow exponentially with n. Lastly, we provide the consistency and $n^{1/2-d}$-consistency of the Lasso in the case where p is fixed and is less than n, under certain assumptions on the design variables X. This proof is also be extended to derive the oracle property for a modified version of Lasso known as the adaptive Lasso. The price that we pay to tackle the persistent correlation among the error sequence is that the rate of increase of the dimension p in the high dimensional setting and the rate of convergence in the $n > p$ setting are slowed down by a factor of n^d.

The paper is organized as follows. Section 2 below investigates the finite sample properties of Lasso under both the non-random design and the random design cases. Section 3 investigates the sign consistency of Lasso. Section 4 provides the asymptotic properties of Lasso and also the oracle property of adaptive Lasso in the $n > p$ setup. Section 5 presents a simulation study to analyse the performance of Lasso in the current setup. Throughout the paper, the design variables X_i's may be triangular arrays depending on n, but we do not exhibit this dependence for the sake of the transparency of the exposition. Also, all limits are taken as $n \to \infty$, unless mentioned otherwise.

2. Results with finite sample

In this section we prove a finite sample oracle inequality for the Lasso solution when the design is non-random, this in turn will imply the consistency as well. Based on the assumptions of the design variables it will soon be clear that the results can easily extended to Gaussian random designs as well. Accordingly, in this subsection we assume that X_i's are non-random. To proceed further, we shall need the following notation. Let

$$W_{nj} = n^{-(1/2+d)} \sum_{i=1}^{n} x_{ij}\varepsilon_i = n^{-(1/2+d)} \sum_{i=1}^{n} \sum_{k=-\infty}^{i} x_{ij} a_{i-k}\zeta_k = \sum_{k=-\infty}^{n} c_{nk,j}\zeta_k, \tag{2.1}$$

where

$$c_{nk,j} := n^{-(1/2+d)} \sum_{i=1}^{n} x_{ij} a_{i-k}, \quad k \in \mathbb{Z}, \ j = 1, \ldots, p,$$

$$c_{n,j}:=\sup_{-\infty<k\le n}|c_{nk,j}|,\quad c_n=\max_{1\le j\le p}c_{n,j}. \tag{2.2}$$

Also, denote by

$$\sigma_{n,j}^2:=\mathrm{Var}(W_{nj}),\quad \sigma_n^2=\max_{1\le j\le p}\sigma_{n,j}^2. \tag{2.3}$$

We shall prove that, with an appropriate choice of λ_n, the Lasso solution obeys the following oracle inequality in the long memory case, with overwhelming probability, i.e. for any $n\ge 1$,

$$\|X(\hat\beta-\beta)\|_2^2/n+\lambda_n\|\hat\beta-\beta\|_1\le\frac{4\lambda_n^2 s_0}{\phi_0^2}$$

Here $\lambda_n=(\mathrm{O}(1))\log p/n^{1/2-d}$, under some conditions on the design matrix. Also, s_0 is the cardinality of the set of nonzero components of β and ϕ_0 is a constant depending on the design matrix X.

As briefly mentioned earlier, the only thing that we require for the proof involves obtaining a probability bound for the set

$$\Lambda=\left\{\max_{1\le j\le p}2n^{-1}\left|\sum_{i=1}^n x_{ij}\varepsilon_i\right|\le\lambda_{0n}\right\}, \tag{2.4}$$

for a proper choice of λ_{0n}. Once this probability bound is obtained, the oracle inequality follows by deterministic arguments (see e.g. Bühlmann and van de Geer (2011)). In fact we have the following:

Proposition 2.1. *Let ε_i be as defined in (1.2) with the innovation distribution satisfying Cramér's condition: for all $k\ge 2$ and some $0<D<\infty$,*

$$E|\zeta_0|^k\le D^{k-2}k!E\zeta_0^2. \tag{2.5}$$

For $t>0$, define

$$\lambda_{0n}=\left\{B_n(t^2+4\log p)+\sqrt{B_n^2(t^2+4\log p)^2+16\sigma_n^2(t^2+4\log p)}\right\}/2n^{1/2-d}, \tag{2.6}$$

where $B_n:=c_nD$. Then, for all $1\le j\le p$ and for all $n\ge 1$,

$$P\left(2\left|n^{-1}\sum_{i=1}^n x_{ij}\varepsilon_i\right|>\lambda_{0n}\right)\le 2\exp\{-(t^2+4\log p)/4\}. \tag{2.7}$$

Consequently,

$$P(\Lambda)\ge 1-2\exp\left(-\frac{t^2}{4}\right),\quad n\ge 1. \tag{2.8}$$

The proof of the above proposition will require several lemmas, hence is postponed to the Appendix. The key to the proof is an application of the Bernstein inequality to finite partial sums and then passing to limit.

We can now proceed to the oracle inequality for the Lasso solution. The corresponding results with i.i.d. errors are proved in Bühlmann and van de Geer (2011, Chapter 6). In what follows, S_0 denotes the collection of indices of the nonzero elements of the true β as defined in (1.1) and s_0 denotes the cardinality of S_0. Also, for any $\delta\in\mathbb{R}^p$, δ_{S_0} denotes the vector of those components of δ which have their indices in S_0. In order to obtain the following inequality we require the 'compatibility condition' on the design matrix X. This condition is as given in Bühlmann and van de Geer (2011), which is restated here for the convenience of the reader.

Definition 2.1. We say that the *compatibility condition* is met for the set S_0, if for some ϕ_0, and for all β satisfying $\|\beta_{S_0^c}\|_1\le 3\|\beta_{S_0}\|_1$,

$$\|\beta_{S_0}\|_1^2\le\frac{(\beta'\hat\Sigma\beta)s_0}{\phi_0^2},$$

with $\hat\Sigma=X'X/n$.

Theorem 2.1. *Assume that the compatibility condition holds for S_0. For some $t>0$ let the regularization parameter be $\lambda_n\ge 2\lambda_{0n}$, where λ_{0n} is given in (2.6). Then with probability at least $1-2\exp(-t^2/4)$, we have*

$$\|X(\hat\beta-\beta)\|_2^2/n+\lambda\|\hat\beta-\beta\|_1\le\frac{4\lambda_n^2 s_0}{\phi_0^2}. \tag{2.9}$$

The proof of Theorem 2.1 is the same as in Bühlmann and van de Geer (2011, Chapter 6), with the value of λ_{0n} changed to the one given in (2.6). This result holds on the set Λ which has the required high probability by Proposition 2.1. □

The only assumptions we have made so far are (i) Cramér's Condition in (2.5) on the innovation distribution and (ii) the compatibility condition in Definition 2.1 on the design variables. It may be of interest to mention that Gaussianity of the error distribution has not been assumed. The price that we have paid for this generality is that λ_{0n} as defined in (2.6) is now itself data driven, i.e. λ_{0n} also depends on the design variables X_i. Thus, keeping in view Theorem 2.1, it is of interest to analyse the rate of convergence of λ_{0n}. The following lemma and remark give additional conditions on the design variables, and the rate of increase of the dimension p, under which λ_{0n} will converge to 0.

Lemma 2.1. *Let $X=(x_{ij})_{n\times p}$ be the design matrix and suppose the following condition holds $\forall\, 1\le j\le p$:*

$$n^{-1}\sum_{i=1}^{n} x_{ij}^2 \le C \quad \text{for some } C<\infty. \tag{2.10}$$

Then with c_n and σ_n^2 as defined in (2.2) and (2.3) respectively, we have $c_n=\mathrm{o}(1)$ and $\sigma_n^2=\mathrm{O}(1)$.

Since $B_n=c_nD$, with D being a fixed constant, the above lemma implies $B_n\to 0$.

Remark 2.1. Now, recall the definition of λ_{0n} from (2.6). Assume that the design variables satisfy condition (2.10). Further assume, $\log p=\mathrm{o}(n^{1/2-d})$, then, $\lambda_{0n}\to 0$.

The following proposition will yield the consistency of the Lasso solution.

Proposition 2.2. *For some $t>0$, let $\lambda_n\ge 2\lambda_{0n}$ where λ_{0n} is defined in (2.6). Then on the set Λ, with probability at least $1-2\exp(-t^2/4)$ we have*

$$2\|X(\hat\beta-\beta)\|_2^2/n \le 3\lambda\|\beta\|_1. \tag{2.11}$$

As mentioned earlier, the proof of this theorem follows deterministic arguments on the set Λ (see Bühlmann and van de Geer, 2011, Chapter 6). The probability of the set Λ is given in Proposition 2.1.

Remark 2.2. Consistency of Lasso: Assume that the following hold:

(i) $\log p/n^{1/2-d}\to 0$,
(ii) Assumption1 holds,
(iii) $\|\beta\|_1=\mathrm{o}(n^{1/2-d}/\log p)$. (2.12)

Then by Remark 2.1 we have $\lambda_{0n}\to 0$. Also, assumption (iii) ensures the right hand side of the inequality (2.11) converges to zero. Hence, Proposition 2.2 along with Lemma 2.1 results in the consistency of the Lasso solution.

Remark 2.3 (*Random design*). There are two assumptions made on the design variables in order to obtain the error bound in Theorem 2.1 and the convergence of λ_{0n} to zero in Remark 2.1. (i) Compatibility condition given in (2.1) and (ii) condition (2.10) which restricts the rate of increase of the design variables. These conditions can be shown to hold in the case of Gaussian random designs with independent rows. Using Theorem 1 of Raskutti et al. (2010), condition (i) can be shown to hold with high probability (increasing to 1 exponentially). If the maximum variance component of the design variables is bounded above by a constant, then (ii) can be shown to hold with high probability using bounds for chi-square distributions given in Johnstone (2001). Hence the above results remain valid with high probability when the design variables are Gaussian with independent rows.

3. Sign consistency of Lasso under long memory

In this section we prove the sign consistency of Lasso for the model (1.1) and (1.2). The results in this section are similar in spirit to Zhao and Yu (2006) and we shall follow the structure of their proofs. They worked in the i.i.d. setup whereas we will be working in the long memory setup. We begin with a definition and some notations.

Definition 3.1. Lasso is said to be strongly sign consistent if there exists $\lambda_n=f(n)$, that is, a function of n and independent of Y^n or X^n such that

$$\lim_{n\to\infty} P(\hat\beta^n(\lambda_n)=_s\beta^n)=1.$$

Here the equality denotes equality in sign, i.e., $\hat\beta^n=_s\beta^n$ if and only if $\mathrm{sign}(\hat\beta^n)=\mathrm{sign}(\beta^n)$, where $\mathrm{sign}(\beta_j)$ assigns a value $+1$ to a positive entry, -1 to a negative entry and 0 to a zero entry.

Assume $\beta^n=(\beta_1^n,\dots,\beta_q^n,\beta_{q+1}^n,\dots,\beta_p^n)'$, where $\beta_j^n\ne 0$, $j=1,\dots,q$, and $\beta_j^n=0$, $j=q+1,\dots,p$. Let $\beta_{(1)}^n=(\beta_1^n,\dots,\beta_q^n)'$ and $\beta_{(2)}^n=(\beta_{q+1}^n,\dots,\beta_p^n)'$. Denote $X(1)$ as the first q columns of X, corresponding to the nonzero components of β^n. Denote $X(2)$

as the last $p-q$ columns of X, corresponding to the zero components of β^n. Let $C^n = n^{-1}X'X$. Then by setting $C_{11}^n = n^{-1}X(1)'X(1)$, $C_{22}^n = n^{-1}X(2)'X(2)$, $C_{12}^n = n^{-1}X(1)'X(2) = (C_{21}^n)'$, C^n can then be expressed as

$$C^n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}.$$

In what follows, we do not exhibit the dependence of β, $\hat{\beta}$ on n for transparency of the exposition. Assuming C_{11}^n is invertible, the Strong Irrepresentable condition as defined by Zhao and Yu is as follows:

Strong Irrepresentable Condition: There exists a vector η, with constant, positive components, such that

$$|C_{21}^n (C_{11}^n)^{-1} \operatorname{sign}(\beta_{(1)})| \le \mathbf{1} - \eta, \tag{3.1}$$

where $\mathbf{1}$ is a $(p-q) \times 1$ vector of ones and the inequality holds element-wise.

The following proposition will serve as a tool to derive the sign consistency in the present setup.

Proposition 3.1. *Assume that the strong irrepresentable condition holds with a vector η, with all components positive. Then*

$$P(\hat{\beta}(\lambda_n) =_s \beta) \ge P(A_n \cap B_n),$$

for

$$A_n = \left\{ |(C_{11}^n)^{-1} W(1)| < n^{1/2-d} \left(|\beta_{(1)}| - \frac{\lambda_n}{2} |(C_{11}^n)^{-1} \operatorname{sign}(\beta_{(1)})| \right) \right\}, \tag{3.2}$$

$$B_n = \left\{ |C_{21}^n (C_{11}^n)^{-1} W(1) - W(2)| \le \frac{\lambda_n}{2} n^{1/2-d} \eta \right\}, \tag{3.3}$$

where

$$W(1) = \frac{X(1)'\varepsilon}{n^{1/2+d}} \quad \textit{and} \quad W(2) = \frac{X(2)'\varepsilon}{n^{1/2+d}}. \tag{3.4}$$

This proposition provides a lower probability bound for the equivalence in sign of the Lasso estimate and the true β vector. The proof is deterministic and hence the conclusion holds with any probabilistic structure on ε. It is also worth mentioning that this proposition holds without any restriction on the dimension p, hence we shall be able to obtain sign consistency under the case where p is increasing with n.

In the following, we shall assume the following conditions on the design matrix and the model parameters. Assume that there exist $0 \le c_1 < c_2 < 1-2d$ and $M_1, M_2, M_3 > 0$, so that

$$\frac{1}{n} X_i' X_i \le M_1 \quad \forall i \in \{1, \dots, n\}, \tag{3.5}$$

$$\alpha' C_{11} \alpha \ge M_2 \quad \forall \alpha \ni \|\alpha\|_2^2 = 1, \tag{3.6}$$

$$q_n = O(n^{c_1}), \tag{3.7}$$

$$n^{1/2-d-c_2/2} \min_{1 \le i \le q} |\beta_i| \ge M_3. \tag{3.8}$$

Under the above assumptions we obtain the following sign consistency result for Lasso in the long memory case.

Theorem 3.1. *Suppose the long memory regression model (1.1) and (1.2) hold, with the innovation distribution satisfying Cramér's condition (2.5). Then under the conditions (3.1), (3.5)–(3.8), if for some $0 < c_3 < c_4 < (c_2 - c_1)/2$, $\lambda_n \propto n^{-(1/2-d-c_4)}$ and $p_n = O(e^{n^{c_3}})$, then*

$$P(\hat{\beta}(\lambda_n) =_s \beta) \to 1. \tag{3.9}$$

The proof is detailed in the Appendix.

4. Asymptotics when $n > p$, p is fixed

4.1. Asymptotic distribution of $X'\varepsilon$

When $n > p$ and p is fixed, the asymptotic properties of Lasso rely critically on the asymptotic distribution of suitably normalized $X'\varepsilon$. This distribution is straightforward to obtain in the case of i.i.d. errors. Here we present the asymptotic distribution of normalized $X'\varepsilon$. This distribution has essentially been obtained in Chapter 4 of GKS, where the authors give

CLT's for weighted sums of a long memory moving average process. Define T_n as the nonnormalized weighted sums W_n as given in (2.1), i.e. $T_n = n^{1/2+d} W_n$. We use T_n instead of W_n to relate the following more closely to GKS. Note that $T_n = X'\varepsilon$.

Our goal is to establish the asymptotic distribution of suitably normalized T_n. This in turn is facilitated by Theorem 4.3.2 of GKS, p. 70. We state a slightly modified version of this theorem which can be proved easily by following the same arguments. In the following denote by $\Sigma_n = \mathrm{Cov}(T_{nj}, T_{nk})_{j,k=1}^p$.

Theorem 4.1. *Let $\{x_{ij}\}_{i=1}^n$, $j = 1, \dots, p$, be p arrays of real weights and $\{\varepsilon_i\}$ be the stationary linear process as defined in (1.2). Assume that the weights $\{x_{ij}\}_{i=1}^n$ satisfy the following condition $\forall j = 1, \dots, p$,*

$$\text{(i)} \max_{1 \le i \le n} |x_{ij}| = o(n^{1/2+d}) \quad \textit{and} \quad \text{(ii)} \sum_{i=1}^n x_{ij}^2 \le C_j n^{1+2d}, \tag{4.1}$$

and for some matrix Σ,

$$n^{-(1+2d)} \Sigma_n \to \Sigma. \tag{4.2}$$

Then, $n^{-(1/2+d)} X'\varepsilon = n^{-(1/2+d)} (T_{n1}, \dots, T_{np}) \to_D \mathcal{N}(0, \Sigma)$.

Corollary 4.1. *Suppose the weights $\{x_{ij}\}$ satisfy (2.10). Then, $n^{-(1+2d)} \Sigma_n = O(1)$ componentwise, for any $0 < d < 1/2$. Moreover, if (4.2) holds then, $n^{-(1/2+d)} X'\varepsilon \to_D \mathcal{N}(0, \Sigma)$.*

Remark 4.1. Theorem 4.1 assumes the convergence (4.2), Corollary 4.1 shows that under a further restriction on the design matrix (2.10) we have $n^{-(1+2d)} \Sigma_n = O(1)$, however, we are unable to show convergence or identify the limit Σ without further assumptions on the design matrix. On the other hand, if we assume the following structure on the design variables, this limit can then be explicitly computed. Let

$$g_j : [0,1] \to \mathbb{R}, \ j = 1, \dots, p, \quad \text{and} \quad X_i = (g_1(i/n), \dots, g_p(i/n))', \ i = 1, \dots, n, \tag{4.3}$$

where we assume that g_j is a continuous function with $\|g_j\|^2 := \int_0^1 g_j^2(u)\, du < \infty$, $\forall 1 \le j \le p$. Under this structure on the design variables, we have $\forall 1 \le j, k \le p$

$$\Sigma_{j,k} := \lim_{n\to\infty} n^{-(1+2d)} \mathrm{Cov}(T_{nj}, T_{nk}) = B(d, 1-2d) \int_0^1 \int_0^1 g_j(u) g_k(v) |u-v|^{-1+2d}\, du\, dv,$$

where $B(d, 1-2d)$ is defined in (1.3) and $\Sigma_{j,k}$ is the (j,k)th component of Σ. This structure on the design variables has been used in Dahlhaus (1995) in the context of polynomial regression with long range dependent regression errors. A short proof is given in the Appendix.

4.2. Asymptotic properties of Lasso

Knight and Fu (2000) proved that in the case of i.i.d. errors, Lasso estimates $\hat{\beta}$ converge in probability to the true coefficient vector β, with an optimal choice of the regularizer λ_n. They also show that Lasso is $\sqrt{n}$-consistent (asymptotic normality). Here we shall present analogous results when the errors are assumed to be long memory moving average. In this section we shall require the following assumption:

$$n^{-1} X'X \to C \quad \text{where } C \text{ is a positive definite matrix.} \tag{4.4}$$

Theorem 4.2. *For the long memory regression model (1.1) and (1.2) assume that the design variables satisfy (4.1), (4.2) and (4.4). Further, if λ_n is such that $\lambda_n \to \lambda_0 \ge 0$, then $\hat{\beta}^n \to_p \arg\min_\phi (Z(\phi))$, where*

$$Z(\phi) = (\phi - \beta)' C (\phi - \beta) + \lambda_0 \sum_{j=1}^p |\phi_j|, \quad \phi \in \mathbb{R}^p.$$

Thus, if $\lambda_n = o(1)$ then $\mathrm{argmin}_\phi (Z(\phi)) = \beta$ and $\hat{\beta}^n(\lambda_n)$ is consistent for β.

Theorem 4.3. *For the long memory regression model (1.1) and (1.2) assume that the design variables satisfy (4.1), (4.2) and (4.4). Suppose $n^{1/2-d} \lambda_n \to \lambda_0 \ge 0$ as $n \to \infty$, then*

$$n^{1/2-d} (\hat{\beta}^n - \beta) \to_D \arg\min_u V(u),$$

where

$$V(u) = -2u'W + u'Cu + \lambda_0 \sum_{j=1}^p [u_j \,\mathrm{sign}(\beta_j) I_{[\beta_j \ne 0]} + |u_j| I_{[\beta_j = 0]}],$$

and W is an $\mathcal{N}_p(0, \Sigma)$ random variable.

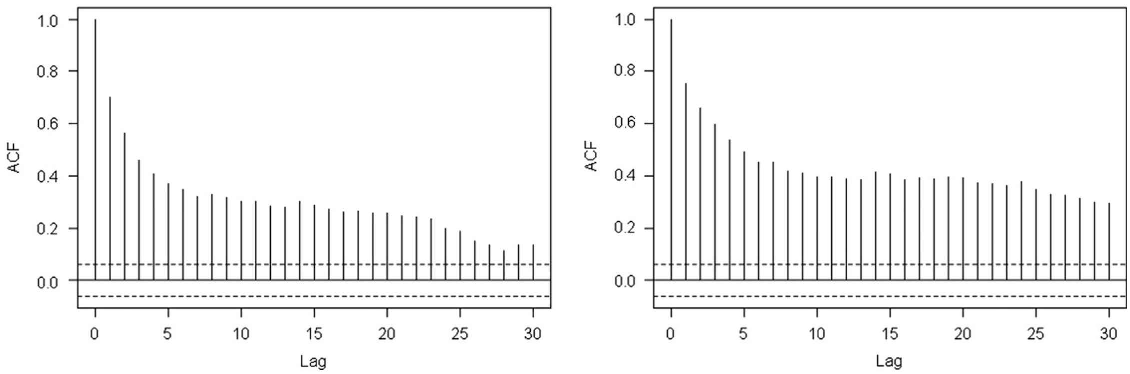

Fig. 1. Lag vs sample auto-correlation function with $d=0.15$ and $d=0.25$.

Note that, when $\lambda_0=0$, arg min $V(u)=C^{-1}W$, where $W\sim\mathcal{N}_p(0,\Sigma)$. The above two theorems highlight the desirable asymptotic properties of Lasso in the current setup. In particular, when $\lambda_0=0$, Theorem 4.2 guarantees estimation consistency, while Theorem 4.3 guarantees the $n^{1/2-d}$–consistency.

The technique used to prove the above theorems is to normalize the dispersion function appropriately in order to use the asymptotic normality of $n^{-(1/2+d)}X'\varepsilon$, in contrast to $n^{-1/2}X'\varepsilon$ in the i.i.d. case. The proof is detailed in the Appendix.

4.3. Adaptive Lasso

The adaptive Lasso differs from Lasso in the way parameters are penalized. To be more precise, for any $\eta>0$, define the weight vector $\hat{w}=1/|(\hat{\beta}^n)|^\eta$, with $\hat{\beta}^n$ being any estimate of β such that $n^{1/2-d}(\hat{\beta}^n-\beta)=O_p(1)$ componentwise. The adaptive Lasso estimates $\tilde{\beta}^n$ are given by

$$\tilde{\beta}^n=\underset{\beta}{\operatorname{argmin}}\left\{\frac{1}{n}\|Y-X\beta\|_2^2+\lambda_n\sum_{j=1}^{p}\hat{w}_j|\beta_j|\right\}. \tag{4.5}$$

Let $\mathcal{A}=\{j:\beta_j\neq 0\}$, $\mathcal{A}_n^\star=\{j:\tilde{\beta}_j^n\neq 0, 1\le j\le p\}$ and $\beta_{\mathcal{A}},\tilde{\beta}_{\mathcal{A}}^n$ be the corresponding vectors with only those components whose indices are in the set $\mathcal{A}$.

As stated in Zuo (2006), an estimator is said to have *oracle property* if the following hold:

1. Asymptotically, the right model is identified, i.e. $\lim_{n\to\infty}P(\mathcal{A}_n^\star=\mathcal{A})=1$.
2. The estimator has an optimal estimation rate, $n^{1/2-d}(\tilde{\beta}_{\mathcal{A}}^n-\beta_{\mathcal{A}})\to_D\mathcal{N}(0,\Sigma^\star)$, for some covariance matrix $\Sigma^\star$.

The adaptive Lasso has an advantage over Lasso, since it possesses a desirable variable selection property under mild assumptions. On the other hand, as seen in Section 3, for Lasso to be sign consistent, we require the strong irrepresentable condition which is a much stronger assumption. The following theorem shows this property of the adaptive Lasso. In other words, the adaptive Lasso enjoys the oracle property in the long memory case. Let $\Sigma_{\mathcal{A}}$ be the limiting covariance matrix in (4.2) with only those components whose indices are in the set $\mathcal{A}\times\mathcal{A}$.

Theorem 4.4. *For the linear model* (1.1), *assume that the design variables satisfy* (4.1), (4.2) *and* (4.4). *Let the regularizer* λ_n *be such that* $n^{1/2-d}\lambda_n\to 0$, *and* $n^{1/2+\eta/2-d-d\eta}\lambda_n\to\infty$. *Then the adaptive Lasso must satisfy the following*:

1. *Variable selection consistency*, $\lim_{n\to\infty}P(\mathcal{A}_n^\star=\mathcal{A})=1$.
2. *Asymptotic normality*, $n^{1/2-d}(\tilde{\beta}_{\mathcal{A}}^n-\beta_{\mathcal{A}})\to_D(C_{11}^n)^{-1}\mathcal{N}(0,\Sigma_{\mathcal{A}})$.

Remark 4.2. For the adaptive weights $\hat{w}=1/|\hat{\beta}|^\eta$, we can choose $\hat{\beta}$ as the ordinary least square estimate. It has already been shown in GKS that $n^{1/2-d}(\hat{\beta}^n-\beta)=O_p(1)$, which is the required condition that the weights must satisfy.

5. Simulation study

In this section we numerically analyse the performance of Lasso under long range dependent setup. We also compare its performance to that in the i.i.d. setup. All simulations were done in R, the estimation of Lasso was done using the package 'glmnet' developed by Friedman et al. (2010) The regularizer λ_n was chosen by five fold cross-validation.

Simulation setup: In this study, β was chosen as a 1000×1 vector, with the first 25 components chosen independently from a uniform distribution over the interval $(-2,5)$, all other components of β were set to zero. The covariates X_i are i.i.d. observations from a 1000 dimensional Gaussian distribution with each component having mean and variance one. We set the pairwise correlation to be $\operatorname{cor}(x_{ij},x_{ik})=0.5^{|j-k|}$. This design matrix has been used by Tibshirani (1996) and many authors

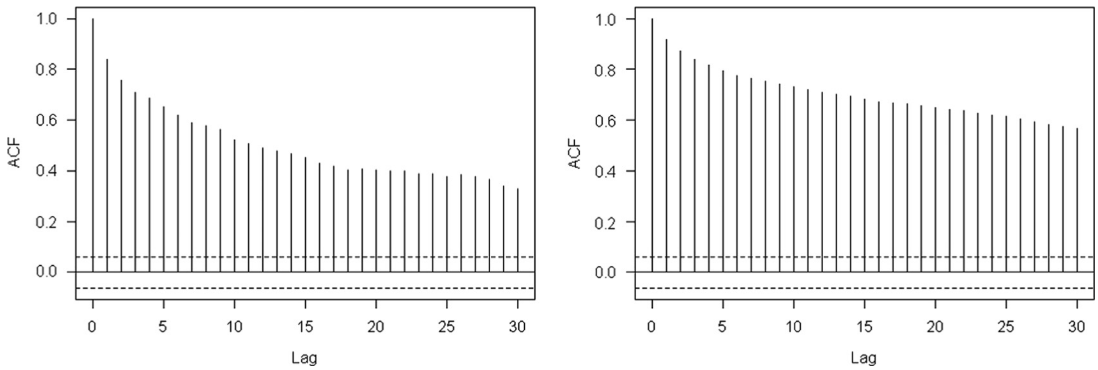

Fig. 2. Lag vs sample auto-correlation function with d=0.35 and d=0.45.

Table 1
Medians of RPE, REE, NZ and IZ of 100 data sets with Gaussian design, long mem. errors.

n	d=0.15				d=0.25				d=0.35				d=0.45			
	REE	RPE	NZ	IZ	REE	RPE	NZ	IZ	REE	RPE	NZ	IZ	REE	RPE	NZ	IZ
100	0.216	0.62	14	33	0.23	0.62	14	33.5	0.24	0.61	14	34.5	0.28	0.46	14	32
200	0.13	0.47	15	30	0.13	0.44	14	32.5	0.14	0.41	14	35	0.18	0.38	14	35
300	0.10	0.39	15	33	0.11	0.36	15	35	0.11	0.38	15	34	0.14	0.31	15	41
400	0.09	0.33	16	39	0.09	0.31	16	40	0.10	0.32	15	36	0.12	0.31	15	41
700	0.05	0.23	20	60	0.06	0.21	19	59.5	0.07	0.23	18	50	0.08	0.22	17	52

Table 2
Medians of RPE, REE, NZ and IZ of 100 data sets with Gaussian design, i.i.d. errors.

n	$\mathrm{Var}(\varepsilon_i)=25.16$				$\mathrm{Var}(\varepsilon_i)=31.98$				$\mathrm{Var}(\varepsilon_i)=47.64$				$\mathrm{Var}(\varepsilon_i)=100.94$			
	REE	RPE	NZ	IZ	REE	RPE	NZ	IZ	REE	RPE	NZ	IZ	REE	RPE	NZ	IZ
200	0.14	0.42	14	29.5	0.15	0.38	14	29	0.18	0.33	14	29	0.25	0.29	14	30
400	0.09	0.28	16	32	0.10	0.26	15	31	0.12	0.22	15	28	0.15	0.18	14	28

since then. The model error vector ε is generated using the definition (1.2) with $c_0=1$ and $d=0.15, 0.25, 0.35, 0.45$, with the innovations being i.i.d. Gaussian random variables as given in (1.2) with mean zero and standard deviation $\sigma_\zeta=3.5$. The simulations were repeated 100 times, i.e. 100 data sets were generated under the above setup with the same parameter vector β.

Since we have chosen d, the corresponding variance of each component of the stationary error process can be computed as $\mathrm{Var}(\varepsilon_i)=\sigma_\zeta^2\sum_{k=1}^{\infty}k^{-2+2d}\ \forall i$, which turns out to be 25.16, 31.98, 47.64 and 100.94 corresponding to d=0.15, 0.25, 0.35, 0.45 respectively.

We begin by illustrating the significant correlation among the components of the regression error vector ε. Figs. 1 and 2 present the sample auto-correlation functions of the error vector ε of the first model of the 100 simulated data sets.

Figs. 1 and 2 exhibit the slow decay of the autocorrelation among the error sequence ε. This slow rate of decay is in coherence with long memory dependence, since $\sum_{k=1}^{\infty}|\gamma_\varepsilon(k)|=\infty$. Also, it is evident from the above two figures that the strength of the dependence is increasing as d increases.

For comparison purposes, we shall also perform the same simulation study with the errors $\varepsilon=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n)$ being i.i.d. Gaussian observations with mean 0 and variances 25.16, 31.98, 47.64, 100.94 which correspond to the variances of the components of the stationary sequence ε under the long memory setup corresponding to $d=0.15, 0.25, 0.35, 0.45$. The reason to choose the same variance of ε_i as in the long memory setup is to maintain the same signal to noise ratio.

Now we proceed to the estimation part. In our study we simulated 100 different realizations of the design matrix X and the error vector ε. Thus leading to 100 data sets with the same parameter vector β. For performance comparison we shall report the Relative Estimation Error (REE), i.e. $\|\hat{\beta}-\beta\|^2/\|\beta\|^2$ and the Relative Prediction Error (RPE) as defined in Zuo (2006), i.e. the empirical estimate of $E\|\hat{Y}-X'\beta\|^2/\sigma_\varepsilon^2$. Also, we shall report the number of correctly estimated non-zero parameters (NZ) and the number of incorrectly estimated zero parameters (IZ). Recall that in the true model there are 25 non-zero and 975 zero parameters. Table 1 summarizes the simulation results under the long memory setup and Table 2 summarizes the results under the i.i.d setup.

Interpretation:

- Lasso is a desirable estimation procedure in our long range dependent setup. It performs accurate estimation at all levels of dependence, from $d=0.15$ to $d=0.45$. It is evident from the simulation results that the estimation becomes increasingly accurate in terms of both REE and RPE as the sample size increases. At $n=400$, the relative error in estimation of β is around 10% at all levels of dependence. As the reader might observe, it was expected that at any fixed sample size, RPE should increase as d increases, however this is not the case, the reason for this is, we use cross validation to choose λ_n and not the theoretical value of λ_n derived earlier.
- In terms of variable selection, Lasso is increasingly successful in choosing the nonzero parameters as the sample size increases. By $n=700$, it identifies around 20 of the nonzero parameters for all levels of dependence. The parameters that Lasso is consistently unable to select are the ones that are too small in size, i.e in our model we have four parameters where $|\beta_j| < 0.65,\ j=3,7,15,19$, and it is these parameters that Lasso is consistently unable to detect, up to the sample size $n=700$. The point here being that this is a known drawback of Lasso connected with assumption (3.8), and it is not due to the long memory dependence structure on the errors. This is confirmed by the results for the i.i.d. case, which exhibits the same problem. The above simulation also brings out another familiar drawback, as the reader might observe, although Lasso manages to correctly estimate a significant portion of the zero parameters (around 95% at $n=700$), however the number of incorrectly estimated zero parameters (IZ) is not decreasing as the sample size increases. This again is not due to the long memory errors but is an inherent drawback of cross validation. This can again be confirmed by the results in the i.i.d. case at the variance levels 47.64 and 100.94 where IZ does not decrease as n increases from 200 to 400.
- Comparing RPE in the long memory case and the i.i.d. case, as expected, we observe that the long memory case requires larger number of observations to reach the same level of accuracy, keeping in mind that the variance of components of ε is similar for both the dependent and independent cases.

6. Discussion

In this paper we discussed the properties of Lasso when used in a regression model exhibiting long range dependent errors generated by a moving average process. The theory in both the non-asymptotic and asymptotic settings was extended to the present setup. The sign consistency of Lasso was established along with the consistency and $n^{1/2-d}$-consistency. Hence showing that all desirable properties of Lasso carry over to the case of long range dependent data. However, the price that we pay to tackle this correlation is a slower rate of increase of the dimension p in the non-asymptotic setting and a slower rate of convergence in the asymptotic setting. The performance of Lasso was also analysed by means of a simulation study, which illustrated its desirable properties in estimation and variable selection. As in the i.i.d. case, the simulation study also brought out the possible weakness of Lasso in identifying all zero parameters successfully in the current setup. A remedy to this in the i.i.d. setup is the adaptive Lasso. The basic oracle property of which was also established in the $n>p$ setup in Section 4. It may be of interest to pursue this further and analyse the theoretical properties of adaptive Lasso in the high dimensional setup, however this has not been pursued here.

Acknowledgement

I wish to thank Professor Hira L. Koul for many useful discussions and comments and for the time that he kindly devoted to me. I also thank two anonymous referees for their patience and constructive comments that helped to improve the paper significantly.

Appendix A

A.1. Proofs for Section 2

The proof of Lemma 2.1, will follow after two key lemmas. To proceed further we require the following notation.

Let r be a finite positive integer and $\forall 1 \le j \le r$, let $h_j = (h_{1j}, h_{2j}, \ldots, h_{nj})'$ be a vector of weights. Further, define $\forall 1 \le j \le r$,

$$W_{n,j} = h_j'\varepsilon = \sum_{i=1}^{n} h_{ij}\varepsilon_i = \sum_{i=1}^{n}\sum_{k=-\infty}^{i} h_{ij}a_{i-k}\zeta_k = \sum_{k=-\infty}^{n} c_{nk,j}\zeta_k, \tag{A.1}$$

where

$$c_{nk,j} = \sum_{i=1}^{n} h_{ij}a_{i-k}$$

Further define

$$c_{n,j} = \sup_{-\infty < k \le n} |c_{nk,j}|, \quad c_n = \max_{1 \le j \le r} c_{n,j}. \tag{A.2}$$

Also, denote by

$$\sigma^2_{n,j} = \mathrm{Var}(W_{n,j}), \quad \sigma^2_n = \max_{1 \le j \le r} \sigma^2_{n,j}. \tag{A.3}$$

Observe that

$$\sigma^2_{n,j} = \sum_{l,m=1}^{n} h_{lj} h_{mj} \gamma_\varepsilon (l-m),$$

furthermore, if we set $h_{ij} = 0 \ \forall i > n$ and $i \le 0$, then under the assumption $\sum_{i=1}^n h^2_{ij} \le M/n^{2d}$, $M < \infty$, we obtain using (1.3) that $\forall 1 \le j \le r$,

$$\sigma^2_{n,j} = c_\gamma \sum_{s=-(n-1), s \ne 0}^{n-1} p(s,j)|s|^{-1+2d} + o(1) = c_\gamma \sum_{m=1}^{n} \sum_{l=1, l \ne m}^{n} h_{mj} h_{lj} |l-m|^{-1+2d} + o(1). \tag{A.4}$$

Here, $p(s,j) := \sum_{i=1}^n h_{ij} h_{(i+s)j}$ and $c_\gamma = B(d, 1-2d)$ as given by (1.3).

Note that, if we replace h_{ij} by $n^{-(1/2+d)} x_{ij}$ in the above definition of W_{nj} then we obtain (2.1). This more general definition of W_{nj} will be essential later in the proof of sign consistency.

Lemma A.1. *For any positive integer r, and for all $1 \le j \le r$, let $h_j = (h_{1j}, \ldots, h_{nj})'$ be any vector of weights such that $\|h_j\|_2^2 = \sum_{i=1}^n h^2_{ij} \le M/n^{2d}$, for some constant $M < \infty$. Let σ^2_n be as defined in (A.3). Then $\sigma^2_n = O(1)$.*

Proof. First

$$\begin{aligned} |p(s,j)| &\le \sum_{i=1}^{n} |h_{ij} h_{(i+s)j}| \le \left(\sum_{i=1}^{n} h^2_{ij} \right)^{1/2} \left(\sum_{i=1}^{n} h^2_{ij} \right)^{1/2} \\ &\le M/n^{2d}, \end{aligned}$$

and hence

$$\begin{aligned} \mathrm{Var}(W_{n,j}) &= c_\gamma \sum_{s=-(n-1), s \ne 0}^{n-1} p(s,j)|s|^{-1+2d} + o(1) \\ &\le c_\gamma \frac{M}{n^{2d}} \sum_{s=-(n-1), s \ne 0}^{n-1} |s|^{-1+2d} + o(1) \\ &\le c_\gamma \frac{M}{n} \sum_{s=-(n-1), s \ne 0}^{n-1} |s/n|^{-1+2d} + o(1) \\ &\to M' \int_{-1}^{1} |t|^{-1+2d} \, dt. \end{aligned} \tag{A.5}$$

Observe that (A.5) is free of j, hence the claim follows. □

Lemma A.2. *For any positive integer r, and for all $1 \le j \le r$, let $h_j = (h_{1j}, \ldots, h_{nj})'$ be any vector of weights such that $\|h_j\|_2^2 = \sum_{i=1}^n h^2_{ij} \le M/n^{2d}$, for some constant $M < \infty$. Then for c_n as defined in (A.2) we have, $c_n = o(1)$.*

Proof. The idea of the proof is borrowed from GKS as part of Proposition 4.3.1, p. 66, where it is used in a different context. First observe, since $\forall 1 \le j \le r$, $\sum_{i=1}^n h^2_{ij} \le M/n^{2d}$, $\Rightarrow 1/\max_{\substack{1 \le i \le n \\ 1 \le j \le r}} |h_{ij}| \ge n^d/\sqrt{M} \to \infty$. Define $K_n := 1/\max_{\substack{1 \le i \le n \\ 1 \le j \le p}} |h_{ij}|$, and consider

$$\begin{aligned} |c_{nk,j}| &\le \sum_{i=1}^{n} |h_{ij} a_{i-k}| \\ &\le \sum_{i=1}^{n} |h_{ij} a_{i-k}| I(|i-k| \ge K_n) \\ &\quad + \sum_{k=1}^{n} |h_{ij} a_{i-k}| I(|i-k| \le K_n) \\ &=: q_{n,1k,j} + q_{n,2k,j} \end{aligned}$$

$$\begin{aligned} q_{n,1k,j} &\le \left(\sum_{k=1}^{n} h^2_{ij} \right)^{1/2} \left(\sum_{k=1}^{n} a^2_{i-k} I(|i-k| \ge K_n) \right)^{1/2} \\ &\le C/n^d \sum_{l \ge K_n}^{n} a^2_l \to 0. \end{aligned} \tag{A.6}$$

$$
\begin{aligned}
q_{n,2k,j} &\le \max_{1\le i\le n, 1\le j\le r} |h_{ij}| \sum_{i=1}^{n} |a_{i-k}| I(|i-k|\le K_n) \\
&\le K_n^{-1} K_n^{1/2} \left(\sum_{l=0}^{\infty} a_l^2\right)^{1/2} \\
&\le C K_n^{-1/2} \to 0.
\end{aligned}
\tag{A.7}
$$

Since the right hand side of (A.6) and (A.7) is free of j, hence we obtain, $c_n = o(1)$. □

Proof of Lemma 2.1. In the above setup $\forall 1 \le j \le p$, let $h_{ij} = n^{-(1/2+d)} x_{ij},\ 1 \le i \le n$. Then, under the assumption (2.10), we have, $\sum_{i=1}^{n} h_{ij}^2 \le C/n^{2d}$. The result now follows from Lemmas A.1 and A.2. □

Remark A.1. Observe from the proof of Lemma A.2, for $h_{ij} = n^{-(1/2+d)} x_{ij}$, we have

$$
|c_{nk,j}| \le \frac{(\sum_{i=1}^{n} x_{ij}^2)^{1/2}}{n^{1/2+d}} \left(\sum_{i=1}^{n} a_{i-k}^2 I(|i-k| \ge K_n)\right)^{1/2} + K_n^{-1/2} \left(\sum_{l=0}^{\infty} a_l^2\right)^{1/2},
$$

where $K_n = \max_{1\le i\le n, 1\le j\le p} |x_{ij}|$. Since $x_{ij} < \infty\ \forall 1 \le i \le n,\ \forall 1 \le j \le p$, and the sequence $\{a_l\}$ is square summable, hence each fixed n, $c_n < \infty$ without the assumption (2.10).

Following are several lemmas that will be required to prove Proposition 2.1. First recall the Bernstein inequality from Doukhan (1994) or Lemma 3.1 from Guo and Koul (2007).

Lemma A.3. *For each $n \ge 1,\ m \ge 1$, let $Z_{mni}, i = -m, \ldots, n$, be an array of mean zero finite variance independent random variables. Assume additionally that they satisfy Cramérs condition: for some $B_{mn} < \infty$,*

$$
E|Z_{mni}|^k \le B_{mn}^{k-2} k! E Z_{mni}^2, \quad k = 2, 3, \ldots,\ i = -m, \ldots, n. \tag{A.8}
$$

Let $T_{mn} = \sum_{i=-m}^{n} Z_{mni},\ \sigma_{mn}^2 = \sum_{i=-m}^{n} \mathrm{Var}(Z_{mni})$. Then, for any $\eta > 0$ and $n \ge 1$,

$$
P(|T_{mn}| > \eta) \le 2 \exp\left\{\frac{-\eta^2}{4\sigma_{mn}^2 + 2B_{mn}\eta}\right\}, \quad \forall m \in \mathbb{Z}^+,\ n \ge 1. \tag{A.9}
$$

We need to apply the above Bernstein inequality p times, jth time to $Z_{mni,j} := c_{ni,j}\zeta_i,\ -m \le i \le n,\ 1 \le j \le p$. In this case then

$$
T_{mnj} = \sum_{i=-m}^{n} c_{ni,j}\zeta_i. \tag{A.10}
$$

For this purpose, we need to verify (A.8) in this case. Let D be as in (2.5) and

$$
B_{mn,j} \equiv B_n := c_n D, \quad c_n = \max_{1\le j\le p} c_{n,j}. \tag{A.11}
$$

Then by assumption (2.5),

$$
\begin{aligned}
|c_{ni,j}|^k E|\zeta_i|^k &\le |c_{ni,j}|^{k-2} D^{k-2} k! c_{ni,j}^2 E\zeta_i^2 \\
&\le B_n^{k-2} k! c_{ni,j}^2 E\zeta_i^2, \quad -m \le i \le n,
\end{aligned}
\tag{A.12}
$$

thereby verifying Cramér's condition (A.8) for $Z_{mni,j}$ for each $1 \le j \le p$ with $B_{mn,j} \equiv B_n$, not depending on m and j.

To proceed further, we need to obtain an upper bound for $\sigma_{mn,j}^2 := \sum_{i=-m}^{n} \mathrm{Var}(Z_{mni,j})$. But

$$
\begin{aligned}
\sigma_{mn,j}^2 &= \sum_{i=-m}^{n} \mathrm{Var}(c_{ni,j}\zeta_i) \le \sum_{i=-\infty}^{n} \mathrm{Var}(c_{ni,j}\zeta_i) \\
&= \mathrm{Var}\left(\sum_{i=-\infty}^{n} c_{ni,j}\zeta_i\right) = \mathrm{Var}\left(\sum_{i=1}^{n} n^{-(1/2+d)} x_{ij}\varepsilon_i\right) \\
&= n^{-(1+2d)} \sum_{k,\ell=1}^{n} x_{kj} x_{\ell j} \gamma_\varepsilon(k-\ell) = \sigma_{n,j}^2 < \infty,
\end{aligned}
\tag{A.13}
$$

From the above discussion we now readily obtain that for all $\eta > 0$ and $1 \le j \le p$,

$$
\begin{aligned}
P\left(\left|\sum_{i=-m}^{n} c_{ni,j}\zeta_i\right| > \eta\right) &\le 2 \exp\left[\frac{-\eta^2}{4\sigma_{mn,j}^2 + 2B_n\eta}\right] \\
&\le 2 \exp\left[\frac{-\eta^2}{4\sigma_n^2 + 2B_n\eta}\right].
\end{aligned}
\tag{A.14}
$$

Remark A.2. By Remark A.1, we see that for each fixed $n \geq 1$, we have $c_n < \infty$ without assumption (2.10). Hence the Bernstein inequality is applicable for every $n \geq 1$ without assumption (2.10).

We are now almost set to derive the probability bound for Λ. Before that, we look at the following preliminary lemma, which will help us to obtain this bound from the truncated sums T_{mnj} defined in (A.10) for W_{nj} defined in (2.1) by taking limit as $m \to \infty$.

Lemma A.4. *For each fixed n, let*

$$A := \left\{ \left| \sum_{i=-\infty}^{n} Y_{ni} \right| > r \right\}, \quad B_m = \left\{ \left| \sum_{i=-m}^{n} Y_{ni} \right| > r - \delta \right\}, \ r > 0, \ \delta > 0, \ m = 1, 2, \ldots$$
$$B = \liminf_{m \to \infty} B_m.$$

If $|\sum_{i=-\infty}^{n} Y_{ni}| < \infty$, *a.s., then, for each fixed n,* $A \subseteq B$

Proof. Let $\omega \in A$. Then $|\sum_{i=-\infty}^{n} Y_{ni}(\omega)| > r$. Also, by assumption, $|\sum_{i=-\infty}^{n} Y_{ni}(\omega)| < \infty$, which implies $\forall \delta > 0 \ \exists N_{\delta,\omega} \ni |\sum_{i=-\infty}^{-m} Y_{ni}(\omega)| < \delta, \ \forall m > N_{\delta,\omega}$. Hence $|\sum_{i=-m}^{n} Y_{ni}(\omega)| > r - \delta, \ \forall m > N_{\delta,\omega}$, which in turn implies

$$\begin{aligned} &\omega \in \bigcap_{m=N_{\epsilon,\omega}}^{\infty} \left\{ \left| \sum_{i=-m}^{n} Y_{ni} \right| > r - \delta \right\} \\ &\Rightarrow \omega \in \bigcup_{m=1}^{\infty} \bigcap_{l=m}^{\infty} \left\{ \left| \sum_{i=-l}^{n} Y_{ni} \right| > r - \delta \right\} \\ &\Rightarrow \omega \in \liminf_{m \to \infty} B_m. \end{aligned} \tag{A.15}$$

Since (A.15) is true for any $\delta > 0$, the claim $A \subseteq B$ follows. □

Before proceeding to the next proposition, we see that the assumption in Lemma A.4 is valid for the series in consideration, which is $\sum_{k=-\infty}^{n} c_{nk,j} \zeta_k$. First, consider the series $\varepsilon_i = \sum_{k=1}^{\infty} a_k \zeta_{i-k}$, since this is an infinite sum of independent zero mean random variables with $\sum_{k=1}^{\infty} \mathrm{Var}(a_k \zeta_{i-k}) < \infty$, hence $\varepsilon_i < \infty$ a.s. (Durrett, 2005, Theorem 1.8.3, p. 62). Now for each fixed n, we have by (2.1), $\sum_{k=-\infty}^{n} c_{nk,j} \zeta_k = n^{-(1/2+d)} \sum_{i=1}^{n} x_{ij} \varepsilon_i$, since this is a finite weighted sum of $\{\varepsilon_i\}$ hence for each fixed n, we have, $\sum_{k=-\infty}^{n} c_{nk,j} \zeta_k < \infty$, a.s. $\forall \, 1 \leq j \leq p$.

Proof of Proposition 2.1. Fix a $1 \leq j \leq p$ and an $n \geq 1$. Recall the definition of $c_{nk,j}$ from (2.2). Let $r_{np} := n^{1/2-d} \lambda_{0n}/2$. Then, for any $0 < \delta < r_{np}$, we have the following inequalities:

$$\begin{aligned} P\left(\left| n^{-(1/2+d)} \sum_{i=1}^{n} x_{ij} \varepsilon_i \right| > r_{np} \right) &= P\left(\left| \sum_{k=-\infty}^{n} c_{nk,j} \zeta_k \right| > r_{np} \right) \\ &\leq P\left(\liminf_{m \to \infty} \left\{ \left| \sum_{k=-m}^{n} c_{nk,j} \zeta_k \right| > r_{np} - \delta \right\} \right) \quad \text{by Lemma A.4,} \\ &\leq \liminf_{m \to \infty} P\left(\left| \sum_{k=-m}^{n} c_{nk,j} \zeta_k \right| > r_{np} - \delta \right) \quad \text{Fatou's lemma,} \\ &\leq \liminf_{m \to \infty} 2 \exp\left[\frac{-(r_{np} - \delta)^2}{4\sigma_n^2 + 2B_n(r_{np} - \delta)} \right], \end{aligned}$$

where the last inequality follows from (A.14). Upon letting $\delta \to 0$ in this bound we thus obtain

$$P\left(\left| n^{-(1/2+d)} \sum_{i=1}^{n} x_{ij} \varepsilon_i \right| > r_{np} \right) \leq 2 \exp\left[\frac{-r_{np}^2}{4\sigma_n^2 + 2B_n r_{np}} \right]. \tag{A.16}$$

Note that r_{np} is a positive solution of the following quadratic equation:

$$\frac{-r_{np}^2}{4\sigma_n^2 + 2B_n r_{np}} = \frac{-(t^2 + 4 \log p)}{4}.$$

Hence, (A.16) and the relation

$$2 \exp\left[\frac{-r_{np}^2}{4\sigma_n^2 + 2B_n r_{np}} \right] = 2 \exp\left[\frac{-(t^2 + 4 \log p)}{4} \right],$$

together imply

$$P\left(2 \left| n^{-1} \sum_{i=1}^{n} x_{ij} \varepsilon_i \right| > \lambda_{0n} \right) = P\left(\left| n^{-(1/2+d)} \sum_{i=1}^{n} x_{ij} \varepsilon_i \right| > r_{np} \right) \leq 2 \exp\left[\frac{-(t^2 + 4 \log p)}{4} \right]. \tag{A.17}$$

This completes the proof of (2.7). □

To prove (2.8), note that

$$1-P(\Lambda)=P\left(\max_{1\le j\le p} 2n^{-1}\left|\sum_{i=1}^{n} x_{ij}\varepsilon_i\right|>\lambda_0\right)$$
$$\le P\left(\bigcup_{j=1}^{p}\left\{2n^{-1}\left|\sum_{i=1}^{n} x_{ij}\varepsilon_i\right|>\lambda_0\right\}\right)$$
$$\le \sum_{j=1}^{p} P\left(2n^{-1}\left|\sum_{i=1}^{n} x_{ij}\varepsilon_i\right|>\lambda_0\right).$$

By (A.17) we get

$$\sum_{j=1}^{p} P\left(2n^{-1}\left|\sum_{i=1}^{n} x_{ij}\varepsilon_i\right|>\lambda_0\right)\le 2p\exp\{-(t^2+4\log p)/4\}=2\exp\left(-\frac{t^2}{4}\right).$$

This completes the proof of Proposition 2.1 □

A.2. Proofs for Section 3

Proof of Proposition 3.1. Let $\hat\beta$ be as defined in (1.4) and let $\hat u=\hat\beta-\beta$. Define

$$V_n(u)=\sum_{i=1}^{n}\frac{1}{n}[(\varepsilon_i-X_i'u)^2-\varepsilon_i^2]+\lambda_n\|u+\beta\|_1.$$

Then $\hat u=\arg\min_u V_n(u)$. Denote the first term in $V_n(u)$ by (I), and the second term by (II). Then (I) can be simplified as

$$\sum_{i=1}^{n}\frac{1}{n}[(\varepsilon_i-X_i'u)^2-\varepsilon_i^2]=\left[-2\sum_{i=1}^{n}\frac{1}{n}u'X_i\varepsilon_i+\sum_{i=1}^{n}\frac{1}{n}(u)'X_iX_i'u\right]$$
$$=\left[\frac{-2u'W}{n^{1/2-d}}+u'C^n u\right], \tag{A.18}$$

where $W=n^{-1/2-d}X'\varepsilon$. Differentiate (A.18) with respect to u to obtain

$$2n^{-(1/2-d)}(C^n(n^{1/2-d}u)-W).$$

Let $\hat u(1)$, $W(1)$ and $\hat u(2)$, $W(2)$ denote the first q and the last $p-q$ entries of $\hat u$, W, respectively. Now note that (Zhao and Yu, 2006)

$$\{\mathrm{sign}(\beta_{(1)})\hat u(1)>-|\beta_{(1)}|\}\subseteq\{\mathrm{sign}(\hat\beta_j)=\mathrm{sign}(\beta_j),\ j=1,2\ldots,q\}. \tag{A.19}$$

Also, by the Karush–Kuhn–Tucker conditions and uniqueness of Lasso, if a solution $\hat u$ exists, then the following conditions must hold:

$$(C_{11}^n(n^{1/2-d}\hat u(1))-W(1))=-\frac{\lambda_n}{2}n^{1/2-d}\ \mathrm{sign}(\beta_{(1)}), \tag{A.20}$$

$$|\hat u(1)|<|\beta_{(1)}|, \tag{A.21}$$

$$|(C_{21}^n(n^{1/2-d}\hat u(1))-W(2))|\le\frac{\lambda_n}{2}n^{1/2-d}\mathbf{1}. \tag{A.22}$$

The set (A.21) is contained in the set on the left of (A.19). Hence (A.20)–(A.22) together imply $\{\mathrm{sign}(\hat\beta_{(1)})=\mathrm{sign}(\beta_{(1)})\}$ and $\hat\beta_{(2)}=\hat u(2)=0$. The condition A_n implies the existence of $\hat u(1)$ which satisfies (A.20) and (A.21) and condition B_n and A_n together imply (A.22). The result follows. □

To maintain clarity of notation in the coming proof, we define the following, for a matrix of weights $h_a=(h_{a1},\ldots,h_{aq})$, where $h_{aj}=(h_{a1j},h_{a2j},\ldots,h_{anj})$, $\forall 1\le j\le q$, define $W_{n,j}^a, c_n^a, \sigma_{an}^2$ as done in (A.1), (A.2), (A.3) respectively. Also define $B_n^a=c_n^a D$. Repeat similarly for a matrix of weights $h_b=(h_{b1},\ldots,h_{b(p-q)})$, with $h_{bj}=(h_{b1j},h_{b2j},\ldots h_{bnj})$, $\forall 1\le j\le(p-q)$.

Proof of Theorem 3.1. Let A_n, B_n be as defined in Proposition 3.1.

$$1-P(A_n\cap B_n)\le P(A_n^c)+P(B_n^c)$$
$$\le\sum_{i=1}^{q}P\left(|z_i|\ge n^{1/2-d}\left(|\beta_i|-\frac{\lambda_n}{2}b_i\right)\right)+\sum_{i=1}^{p-q}P\left(|\kappa_i|\ge\frac{\lambda_n}{2}n^{1/2-d}\eta_i\right),$$

where $z=(z_1,z_2,\ldots,z_q)'=(C_{11}^n)^{-1}W(1)$, $\kappa=(\kappa_1,\kappa_2,\ldots,\kappa_{p-q})'=C_{21}^n(C_{11}^n)^{-1}W(1)-W(2)$, $b=(b_1,b_2,\ldots,b_q)=(C_{11}^n)^{-1}\ \mathrm{sign}(\beta_{(1)})$. Now express $z=h_a'\varepsilon$, where $h_a'=(h_{a1},\ldots,h_{aq})'=(C_{11}^n)^{-1}(n^{-1/2-d}X(1)')$. Then

$$h_a'h_a=(C_{11}^n)^{-1}n^{-2d},$$

and $z_j = h'_{aj}\varepsilon$ with

$$\|h_{aj}\|_2^2 \le \frac{1}{n^{2d}M_2} \quad \forall j=1,\dots,q, \text{ by assumption (3.6)}$$

Similarly write $\kappa = h'_b\varepsilon$, where $h'_b = C_{21}^n (C_{11}^n)^{-1}(n^{-1/2-d}X(1)') - (n^{-1/2-d}X(2)')$. Then

$$h'_b h_b = \frac{1}{n^{1+2d}} X(2)'[I - X(1)(X(1)'X(1))^{-1}X(1)']X(2).$$

Since $[I - X(1)(X(1)'X(1))^{-1}X(1)]$ has eigenvalues between 0 and 1, therefore $\zeta_j^n = h'_{bj}\varepsilon$, with

$$\|h_{bj}\|_2^2 \le M_1/n^{2d} \quad \forall j=1,\dots p-q, \text{ by assumption (3.5)}.$$

Hence the weight vectors $h_{aj}, 1\le j\le q$, and $h_{bj}, 1\le j\le p-q$, both satisfy Lemmas A.1 and A.2 for $r=q$ and $r=p-q$ respectively. Also,

$$|\lambda_n b| = \lambda_n |(C_{11})^{-1}\operatorname{sign}(\beta_{(1)})| \le \frac{\lambda_n}{M_2}\|\operatorname{sign}(\beta_{(1)})\|_2 = \frac{\lambda_n}{M_2}\sqrt{q}. \tag{A.23}$$

Now, $z_j = h'_{aj}\varepsilon = \sum_{i=1}^n h_{aij}\varepsilon_i$. Proceed as done earlier in (A.16). Using (A.23), Lemmas A.1 and A.2 and Bernstein's Inequality as applied in (A.16). We get, for some constants $r_1, r_2 > 0$,

$$\begin{aligned}\sum_{j=1}^q P\left(|z_j| \ge n^{1/2-d}\left(|\beta_j| - \frac{\lambda_n}{2}b_j\right)\right) &\le \sum_{j=1}^q P\left(|z_j| \ge r_1 n^{c_2/2}\right)\\ &\le 2q\exp\left(\frac{-r_1^2 n^{c_2}}{4\sigma_{an}^2 + 2B_n^a r_1 n^{c_2/2}}\right) \to 0.\end{aligned} \tag{A.24}$$

Also

$$\begin{aligned}\sum_{j=1}^{p-q} P\left(|\kappa_j| \ge \frac{\lambda_n}{2}n^{1/2-d}\eta_j\right) &\le (p-q)\exp\left(\frac{-r_2^2(\lambda_n n^{1/2-d})^2}{4\sigma_{bn}^2 + 2B_n^b r_2\lambda_n n^{1/2-d}}\right)\\ &\le (p-q)\exp(-r_2\lambda_n n^{1/2-d}) \quad \text{for } n \text{ large enough},\\ &\le \exp(n^{c_3} - r_2\lambda_n n^{1/2-d}) \to 0.\end{aligned} \tag{A.25}$$

The result follows from (A.24) and (A.25) together. □

A.3. Proofs for Section 4

Proof of Corollary 4.1. Observe that assumption (2.10) implies assumption (4.1)(i) and (4.1)(ii). Hence we only need to show $n^{-(1+2d)}\Sigma_n = O(1)$ componentwise. For each variance component, this has already been shown in (A.5) in the proof of Lemma 2.1, with $h_{ij} = n^{-(1/2+d)}x_{ij}, \forall 1\le j\le p$. The covariance components can be easily dealt with the Cauchy–Schwartz inequality. □

Proof of Remark 4.1. Using (A.4), we obtain

$$\begin{aligned}n^{-1-2d}\operatorname{Cov}(T_{nj}, T_{nk}) &= n^{-1-2d}c_\gamma \sum_{l,m=1, l\ne m}^n g_j\left(\frac{l}{n}\right)g_k\left(\frac{m}{n}\right)|l-m|^{-1+2d} + o(1)\\ &\to c_\gamma \int_0^1\int_0^1 g_j(u)g_k(v)|u-v|^{-1+2d}\,du\,dv. \qquad \square\end{aligned}$$

Proof of Theorem 4.2. Let

$$Z_n(\phi) = \frac{1}{n}\sum_{i=1}^n (Y_i - X'_i\phi)^2 + \lambda_n \sum_{i=1}^p |\phi_i|,$$

then $Z_n(\phi)$ is convex. We need to show the pointwise convergence (in probability) of $Z_n(\phi)$ to $Z(\phi)+k^2$ for some constant k. Clearly,

$$\lambda_n \sum_{i=1}^p |\phi_i| \to \lambda_0 \sum_{i=1}^p |\phi_i|.$$

Consider,

$$\begin{aligned}n^{-1}\sum_{i=1}^n (Y_i - X'_i\phi)^2 &= n^{-1}\sum_{i=1}^n (\varepsilon_i - X'_i(\phi-\beta))^2 = n^{-1}\sum_{i=1}^n \varepsilon_i^2 + n^{-1}\sum_{i=1}^n (\phi-\beta)'X_iX'_i(\phi-\beta) - 2n^{-1}(\phi-\beta)'\sum_{i=1}^n X_i\varepsilon_i\\ &= n^{-1}\sum_{i=1}^n \varepsilon_i^2 + n^{-1}\sum_{i=1}^n (\phi-\beta)'X_iX'_i(\phi-\beta) - 2n^{-1}(\phi-\beta)'X'\varepsilon,\end{aligned}$$

the first term in the above equation converges to k^2 by the ergodic theorem (since $\{\varepsilon_i\}$ form a stationary ergodic sequence), the second term converges to $(\phi-\beta)'C(\phi-\beta)$ and the last term converges to zero in probability (since by Theorem 4.1, $n^{-(1/2+d)}X'\varepsilon$ converges in distribution). This proves the theorem. □

Proof of Theorem 4.3. Define

$$V_n(u) = n^{1-2d}\left[\sum_{i=1}^{n}\frac{1}{n}\left[\left(\varepsilon_i - \frac{X_i'u}{n^{1/2-d}}\right)^2 - \varepsilon_i^2\right] + \lambda_n\sum_{j=1}^{p}\left[\left|\beta_j + \frac{u_j}{n^{1/2-d}}\right| - |\beta_j|\right]\right]. \tag{A.26}$$

Denote the first term in the above equation by (I), and the second term by (II). Then

$$\begin{aligned}(\mathrm{I}) &= n^{1-2d}\left[\sum_{i=1}^{n}\frac{1}{n}\varepsilon_i^2 + \frac{u'\sum_{i=1}^{n}X_iX_i'u}{n.n^{1-2d}} - 2\frac{\sum_{i=1}^{n}u'X_i\varepsilon_i}{n.n^{1/2-d}} - \sum_{i=1}^{n}\frac{1}{n}\varepsilon_i^2\right]\\ &= \left[\frac{u'\sum_{i=1}^{n}X_iX_i'u}{n} - 2\frac{\sum_{i=1}^{n}u'X_i\varepsilon_i}{n^{1/2+d}}\right]\\ &\to u'Cu - 2u'W \quad \text{as } n\to\infty,\end{aligned} \tag{A.27}$$

where W is $\mathcal{N}_p(0,\Sigma)$.

Also,

$$\begin{aligned}(\mathrm{II}) &= n^{1/2-d}\lambda_n\sum_{j=1}^{p}[|n^{1/2-d}\beta_j + u_j| - n^{1/2-d}|\beta_j|]\\ &\to \lambda_0\sum_{j=1}^{p}[u_j\,\mathrm{sign}(\beta_j)I_{[\beta_j\neq 0]} + |u_j|I_{[\beta_j=0]}],\end{aligned} \tag{A.28}$$

The result follows from (A.27) and (A.28) together. □

Proof of Theorem 4.4. The structure of the proof is similar to that of Theorem 2 in Zuo (2006). Define

$$\tilde{V}_n(u) = n^{1-2d}\left[\sum_{i=1}^{n}\frac{1}{n}\left[\left(\varepsilon_i - \frac{X_i'u}{n^{1/2-d}}\right)^2 - \varepsilon_i^2\right] + \lambda_n\sum_{j=1}^{p}\hat{w}_j\left[\left|\beta_j + \frac{u_j}{n^{1/2-d}}\right| - |\beta_j|\right]\right], \tag{A.29}$$

then $\tilde{u}_j = n^{1/2-d}(\tilde{\beta}^n - \beta) = \arg\min \tilde{V}_n(u)$. Expanding $\tilde{V}_n(u)$ as done in (A.27) and (A.28) we get

$$\tilde{V}_n(u) = \frac{u'\sum_{i=1}^{n}X_iX_i'u}{n} - 2\frac{\sum_{i=1}^{n}u'X_i\varepsilon_i}{n^{1/2+d}} + n^{1/2-d}\lambda_n\sum_{j=1}^{p}\hat{w}_j\left[|n^{1/2-d}\beta_j + u_j| - n^{1/2-d}|\beta_j|\right].$$

Recall, $n^{-1}X'X\to C$, and by Theorem 4.1 we have $n^{-(1/2+d)}X'\varepsilon \to_D \mathcal{N}(0,\Sigma)$. Also, since $n^{1/2-d}\lambda_n\to 0$, $n^{1/2+\eta/2-d-d\eta}\lambda_n\to\infty$ and the adaptive weights $\hat{\beta}^n$ are so that $n^{1/2-d}(\hat{\beta}^n - \beta) = O_p(1)$. Hence we obtain $\tilde{V}_n(u)\to\tilde{V}(u)$ where

$$\tilde{V}(u) = \left\{\begin{matrix} u'_{\mathcal{A}}C_{11}u_{\mathcal{A}} - 2u'_{\mathcal{A}}W_{\mathcal{A}} & \text{if } u_j = 0\ \forall j\notin\mathcal{A} \\ \infty & \text{else}\end{matrix}\right\} \tag{A.30}$$

The unique minimum of $\tilde{V}(u)$ is $(C_{11}^{-1}W_{\mathcal{A}}, 0)'$. Hence we obtain

$$\tilde{u}_{\mathcal{A}} = n^{1/2-d}(\tilde{\beta}^n_{\mathcal{A}} - \beta_{\mathcal{A}}) \to_D C_{11}^{-1}W_{\mathcal{A}} \quad \text{and} \quad \tilde{u}_{\mathcal{A}^c}\to_D 0. \tag{A.31}$$

The variable selection part can be obtained by adjusting normalization in the proof of Zuo (2006). From the asymptotic normality obtained in (A.31), we obtain $\forall j\in\mathcal{A}$, $P(j\in\mathcal{A}_n^\star)\to 1$. Let $\mathbf{x}_j{:=}(x_{1j},\ldots,x_{nj})'$ be the jth column of the design matrix X, $1\le j\le p$. Next we show that if $j\notin\mathcal{A}$, then $P(j\notin\mathcal{A}_n^\star)\to 1$. By the KKT conditions for the Lasso solution, we have $|2\mathbf{x}_j'(Y - X\tilde{\beta})| \le n\lambda_n\hat{w}_j$. Consider,

$$\frac{\mathbf{x}_j'(Y - X\tilde{\beta}^n)}{n^{1/2+d}} = \frac{\mathbf{x}_j'Xn^{1/2-d}(\beta - \tilde{\beta}^n)}{n} + \frac{\mathbf{x}_j'\varepsilon}{n^{1/2+d}}$$

using (A.31), the first term on the right side converges to some normal distribution, and by Theorem 4.1 the second term on the right converges to a normal distribution. Also, since $\beta_j = 0$ and $n^{1/2-d}(\beta - \hat{\beta}^n) = O_p(1)$, hence, $n^{1/2-d}\lambda_n\hat{w}_j = n^{1/2-d+\eta-d\eta}\lambda_n 1/|n^{1/2-d}\hat{\beta}_j|^\eta \to\infty$. This implies

$$P(j\notin\mathcal{A}_n^\star) \le P(|2\mathbf{x}_j'(Y - X\tilde{\beta}^n)| \le n\lambda_n\hat{w}_j) \to 1.$$

This completes the proof. □

References

Alquier, P., Doukhan, P., 2011. Sparsity considerations for dependent variables. Electron. J. Statist. 5, 750–774.
Baillie, R.T., 1996. Long memory processes and fractional integration in econometrics. J. Econometrics 73, 5–59.
Beran, J., Feng, Y., Ghosh, S., Kulik, R., 2013. Long Memory Processes. Springer, Heidelberg.
Beran, J., 1992. Statistical methods for data with long-range dependence. Statist. Sci. 7 (4), 404–427.
Bickel, P., Ritov, Y., Tsybakov, A., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37, 1705–1732.

Bühlmann, P., van de Geer, S., 2011. Statistics for High Dimensional Data. Springer, Heidelberg.
Dahlhaus, R., 1995. Efficient location and regression estimation for long range dependent regression models. Ann. Statist. 23, 1029–1047.
Doukhan, P., 1994. Mixing: Properties and Examples. Springer-Verlag, New York.
Durrett, R., 2005. Probability: Theory and Examples, third edition, Brooks/Cole-Thomson Learning, California.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Statist. Software 33, 1–22.
Giraitis, L., Koul, H., Surgailis, D., 2012. Large Sample Inference for Long Memory Processes. Imperial College Press, London.
Guo, H., Koul, H., 2007. Nonparametric regression with heteroscedastic long memory errors. J. Statist. Plann. Inference 137, 379–404.
Johnstone, I.M., 2001. Chi-square Oracle Inequalities, Lecture Notes-Monograph Series, vol. 36, pp. 399–418.
Knight, K., Fu, W., 2000. Asymptotics for Lasso-type estimators. Ann. Statist. 28, 1356–1378.
Meinhausen, N., Bühlmann, P., 2006. High dimensional graphs and variable selection with the Lasso. Ann. Statist. 34, 1436–1462.
Raskutti, G., Wainwright, M., Yu, B., 2010. Restricted eigenvalue properties for correlated Gaussian designs. J. Mach. Learn. Res. 99, 2241–2259.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B 58, 267–288.
Yoon, Y., Park, C., Lee, T., 2013. Penalized regression models with autoregressive error terms. J. Statist. Comput. Simulation 83, 1756–1772.
Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. J. Mach. Learn. Res. 7, 2541–2563.
Zuo, H., 2006. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc., 1418–1429.