Statistics and Probability Letters 125 (2017) 64–70

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Lasso for sparse linear regression with exponentially β-mixing errors

Fang Xie [a], Lihu Xu [a,*], Youcai Yang [b]

[a] Department of Mathematics, Faculty of Science and Technology, University of Macau, China

[b] School of Economics and Management, Qingdao University of Science and Technology, China

ARTICLE INFO

Article history:
Received 6 December 2016
Received in revised form 17 January 2017
Accepted 24 January 2017
Available online 4 February 2017

Keywords:
Lasso
Linear regression models
Consistency
Exponentially β-mixing errors

ABSTRACT

We prove two consistency theorems for the lasso estimators of sparse linear regression models with *exponentially* β*-mixing* errors, in which the number of regressors p is large, even much larger than the sample size n.

1. Introduction

We consider the following linear model:

$$Y_i = X_i'\xi^* + \sigma\epsilon_i, \quad \text{for } i = 1, 2, \dots, n, \tag{1.1}$$

where $\epsilon_1, \dots, \epsilon_n$ are *stationary* and *exponentially* β*-mixing* (see the definition below) with the law F such that $\mathbb{E}\epsilon_i = 0$ and $\mathbb{E}\epsilon_i^2 = 1$. The vector $\xi^* \in \mathbb{R}^p$ is the unknown true parameter, and $\sigma > 0$ is the standard deviation of the error term. In practice, σ is often not known and have to be pre-estimated by some methods such as overlapping batch mean (Alexopoulos et al., 2007). The predictors $X_i = (X_{i1}, \dots, X_{ip})'$ are p-dimensional, with p possibly much larger than the sample size n. The estimation of ξ^* thanks to the following sparsity assumption: $T = \text{supp}(\xi^*)$ has $s < n$ elements. Write $Y = (Y_1, Y_2, \dots, Y_n)'$, $X = (X_1, X_2, \dots, X_n)'$ and $\xi = (\xi_1, \xi_2, \dots, \xi_p)'$, the lasso estimator (Belloni et al., 2011; Efron et al., 2004; Fan and Li, 2001; Lockhart et al., 2014; Meinshausen and Yu, 2006; Benjamin and van de Geer, 2015; Tibshirani, 1996; van de Geer and Bühlmann, 2009; Zou, 2006) is defined by

$$\hat{\xi}^L = \arg\min_{\xi\in\mathbb{R}^p}\left\{\frac{1}{2n}\|Y - X\xi\|_2^2 + \lambda\|\xi\|_1\right\}, \tag{1.2}$$

where $\lambda > 0$ is the penalty coefficient. Bickel et al. (2009) showed that for any $\theta \in (0, 1)$, if $\epsilon_i \sim N(0, 1)$ are independent, by choosing the penalty level $\lambda = c\sigma\sqrt{n}\Phi^{-1}(1 - \theta/2p)$ with $c > 1$ being some constant and Φ being the cumulative

* Corresponding author.
E-mail addresses: fangxie219@gmail.com (F. Xie), lihuxu@umac.mo (L. Xu).

http://dx.doi.org/10.1016/j.spl.2017.01.023

distribution function of standard normal random variable, under appropriate assumptions, the lasso estimator $\hat{\xi}^L$ achieves near-oracle performance in the sense that

$$\frac{1}{n}\left\|X(\hat{\xi}^L-\xi^*)\right\|_2^2 \preceq \sigma^2 s\log(2p/\theta)/n \tag{1.3}$$

with probability at least $1-\theta$. We use $a \preceq b$ to denote $a \leq c'b$ with some $c' > 0$.

In many economic and finance applications, the i.i.d. assumption is unlikely to be satisfied, and it would be desirable to have some theoretical guidance that can be applied to lasso type regressions for time series. Besides, the research on lasso method for time series is getting more and more attentions (Lam and Souza, 2014; Ren and Zhang, 2013; Wang et al., 2007; Xiao et al., in preparation). In this paper, we extend the previous analysis of the lasso with i.i.d. errors to β-*mixing* error models, and provide asymptotic analysis, convergence rate and near-oracle performance. In detail, we focus on the linear model with β-*mixing* error sequence whose elements are either Gaussian or sub-gaussian random variables.

From our consistency theorems below, if the data are normally distributed, by choosing $\lambda = c\sigma\nu_n\sqrt{n}\Phi^{-1}\left(1-\frac{\theta}{2p}\right)$, where ν_n is the long-run correlation parameter that captures the serial correlation, under some design conditions, the inequality (1.3) also holds. For sub-gaussian errors, under additional conditions, similar result can be obtained asymptotically. In Section 3, we conduct two simulations to show the asymptotic performance of lasso estimator under β-*mixing* errors.

2. Definitions and consistency theorems

Firstly, we recall the definitions of β-*mixing* and *sub-gaussian*.

Definition 2.1. Suppose $Z_1, Z_2, \ldots$ is a sequence of random variables (r.v.s). Let $\mathcal{F}_1^a = \sigma(Z_i, 1 \leq i \leq a)$ and $\mathcal{F}_{a+k}^{+\infty} = \sigma(Z_i, a+k \leq i \leq +\infty)$ be two σ-fields. The sequence $\{Z_i, i \geq 1\}$ is called β-*mixing* if $\beta(k) \equiv \sup_{a\geq 1}\mathbb{E}\sup\{|\mathbb{P}(B|\mathcal{F}_1^a)-\mathbb{P}(B)| : B \in \mathcal{F}_{a+k}^{\infty}\} \to 0$ as $k \to \infty$. Furthermore, it is called *exponentially* β-*mixing* if there exist positive constants b_1, b_2, such that $\beta(k) \leq b_1 e^{-b_2 k}$.

Definition 2.2. A random variable η is called *sub-gaussian* if there exists a positive constant K_1 such that $\mathbb{P}(|\eta| > M) \leq \exp\{1-\frac{M^2}{K_1}\}$ for an arbitrary $M \geq 0$.

Remark 2.3. Roughly speaking, β-*mixing* property describes the correlations between the r.v.s in a random sequence, and is not directly related to their specific distributions. For instance, for any sequence of *independent* r.v.s, it is β-*mixing* with $\beta(k) = 0$ for all k.

Notations: For an arbitrary d-dimensional vector $v = (v_1, \ldots, v_d)'$, its l_q norm is defined by $\|v\|_q = (\sum_{i=1}^d |v_i|^q)^{\frac{1}{q}}$ for $q \geq 1$. Further define $\|v\|_\infty := \max_{1\leq i\leq d}|v_i|$ and $\|v\|_0$ the number of non-zero coordinates of v. Given a subset $I \subset \{1, \ldots, d\}$, v_I is the restricted vector on I.

For convenient description, we state assumptions below firstly:

A1. The error sequence $\{\epsilon_i, 1 \leq i \leq n\}$ is *exponentially* β-*mixing* with some constants $b_1, b_2 > 0$.
A2. For each $1 \leq i \leq n$, ϵ_i is *sub-gaussian* with constant K_1.
A3. There exists some positive constant $K_2 < \infty$ such that $\sup_{1\leq i,j\leq n}|X_{ij}| \leq K_2$.
A4. Denoting the sample covariance by $\widehat{\Sigma} = \frac{X^TX}{n}$, we assume $\kappa^2 = \min_{\delta\in\Delta}\frac{\langle\delta,\widehat{\Sigma}\delta\rangle}{\|\delta_T\|_2^2} > 0$, where $\delta = \hat{\xi}^L - \xi^*$ and $\Delta = \{\delta \in \mathbb{R}^p : \|\delta_{T^c}\|_1 \leq 3\|\delta_T\|_1, \delta \neq 0\}$.

Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$, define $\nu_{nj} = \mathrm{Var}(\frac{X_j^T\epsilon}{\sqrt{n}})$ and $\nu_n = \sup_{1\leq j\leq p}\nu_{nj}$.

The following two theorems show that if the error sequence is *exponentially* β-*mixing*, the lasso estimator will be consistent in both Gaussian and sub-gaussian error cases. The corresponding proofs are given in the last section.

Theorem 2.4 (*Gaussian Error*)**.** *Consider the model (1.1) and define $\hat{\xi}^L$ as the lasso estimator in (1.2). Let $F = \Phi$ and the assumptions* **A1**, **A3**, **A4** *be satisfied, if $\lambda = c\sigma\nu_n\sqrt{\frac{\log 2p}{n}}$ with $c > 2\sqrt{2}$ and ν_n defined above, then with probability at least $1-(2p)^{1-c^2/8}$, we have $\|\hat{\xi}^L-\xi^*\|_1 \leq \frac{16c\sigma\nu_n s}{\kappa^2}\sqrt{\frac{\log 2p}{n}}$ and $\frac{1}{n}\|X(\hat{\xi}^L-\xi^*)\|_2^2 \leq \frac{16c^2\sigma^2\nu_n^2 s\log 2p}{n\kappa^2}$.*

Remark 2.5. Comparing with Bickel et al. (2009, Theorem 7.2) which considers the i.i.d. Gaussian error, Theorem 2.4 adds the ν_n, caused by the β-*mixing* errors, into the penalty coefficient λ and obtains similar bounds on $\|\hat{\xi}^L-\xi^*\|_1$ and $\frac{1}{n}\|X(\hat{\xi}^L-\xi^*)\|_2$.