Electronic Journal of Statistics
Vol. 10 (2016) 352–379
ISSN: 1935-7524
DOI: 10.1214/16-EJS1108

Performance bounds for parameter estimates of high-dimensional linear models with correlated errors

Wei-Biao Wu*

Department of Statistics
The University of Chicago
5734 S University Ave
Chicago, IL 60637
e-mail: wbwu@galton.uchicago.edu

and

Ying Nian Wu†

Department of Statistics
UCLA
8125 Math Sciences Bldg. Box 951554
Los Angeles, CA
e-mail: ywu@stat.ucla.edu

Abstract: This paper develops a systematic theory for high-dimensional linear models with dependent errors and/or dependent covariates. To study properties of estimates of the regression parameters, we adopt the framework of functional dependence measures ([43]). For the covariates two schemes are addressed: the random design and the deterministic design. For the former we apply the constrained ℓ_1 minimization approach, while for the latter the Lasso estimation procedure is used. We provide a detailed characterization on how the error rates of the estimates depend on the moment conditions that control the tail behaviors, the dependencies of the underlying processes that generate the errors and the covariates, the dimension and the sample size. Our theory substantially extends earlier ones by allowing dependent and/or heavy-tailed errors and the covariates. As our main tools, we derive exponential tail probability inequalities for dependent sub-Gaussian errors and Nagaev-type inequalities for dependent non-sub-Gaussian errors that arise from linear or non-linear processes.

Keywords and phrases: Consistency, dependence-adjusted norm, exponential inequality, functional and predictive dependence measures, high-dimensional time series, impulse response function, Nagaev inequality, predictive persistence, support recovery.

Received March 2015.

*Research partially supported by DMS-1106790 and DMS-1405410.
†Research partially supported by NSF DMS-1310391 and DARPA SIMPLEX N66001-15-C-4035

1. Introduction

During the past two decades there has been a substantial development on high-dimensional linear regression models. Consider the model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + e_i,\ 1 \le i \le n, \tag{1.1}$$

where y_i, $\mathbf{x}_i$ and e_i are the response variable, the $p \times 1$ covariate vector and the error term respectively, and $\boldsymbol{\beta}$ is a p-dimensional regression parameter vector. In matrix notation, we can write it as $Y = X\boldsymbol{\beta} + e$, where Y is the $n \times 1$ response vector, X is the $n \times p$ design matrix, and e is the $n \times 1$ vector of errors. The covariate $\mathbf{x}_i$ can be random or deterministic. Here the dimension p can be much larger than the sample size n. Clearly in this case the classical least squares method fails to estimate $\boldsymbol{\beta}$ since the matrix $X^\top X$ is singular. Under certain sparsity conditions on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, namely if only a small number of components of $\boldsymbol{\beta}$ are non-zero, one can apply the ℓ_1 penalized least squares (Lasso) procedure [37]. A closely related approach is the Dantzig-type estimator [9], which is the optimizer of certain objective function under linear inequality constraints. Other variants include the SCAD estimator [13] and the MCP estimator [47], among others. Theoretical properties of those estimators have been extensively studied in the literature; see for example [4], and the recent book [5] for a thorough treatment and further references.

In most of the theoretical investigations of model (1.1), it is assumed that the errors $(e_i, i = 1, \ldots, n)$ are independent and identically distributed (i.i.d.) Gaussian, sub-Gaussian or sub-exponential random variables which have finite exponential moments. Similar assumptions are also adopted for the covariates $(\mathbf{x}_i, i = 1, ..., n)$ in the case of random design. The associated tools for obtaining performance bounds are the exponential-type concentration inequalities, including, among others, the Bennett, Bernstein and Hoeffding inequalities; see Chapter 14 of [5] for a review. With the help of such inequalities, assuming certain sparseness conditions, one can deal with the case in which p is much larger than n and still obtain consistency under the very mild condition of the type $\log p = o(n)$.

Despite the extensive literature on Lasso and Dantzig-type estimates, there has been very limited research on theoretical properties of the estimates when the errors (e_i) or the covariates $(\mathbf{x}_i)$ are dependent and/or non-sub-Gaussian. If the data are observed over time or space, the independence assumption for the errors (e_i) or the covariates $(\mathbf{x}_i)$ is violated. The sub-Gaussian assumption is also questionable as the errors and the covariates may be heavy-tailed and may not have finite exponential moments. In econometric analysis of vector autoregressive processes, in [34] Sims cautioned that fat tails can affect the validity of the associated statistical inference. In terms of dependent errors, [41] proposed a Lasso estimator when the errors follow an autoregressive model. Gupta [15] analyzed Lasso estimator for weakly dependent errors. Both papers mainly deal with the case where n is greater than p. Ravikumar et al [33] applied the Rosenthal's [35] inequality. Recently [18] studied Lasso with long

memory errors with very light tails such that the Cramér condition is met, Loh [24] considered M-estimators for linear models with i.i.d. data, and in [3] Basu and Michailidis investigated theoretical properties of Lasso estimates for high-dimensional Gaussian processes.

The goal of this paper is threefold: (i) To lay a theoretical foundation for high-dimensional inference in situations in which the errors or the covariates can be dependent; (ii) To develop sharp inequalities for tail probabilities for dependent and/or non-sub-Gaussian processes under mild and easily verifiable conditions; and (iii) To apply our inequalities to Lasso and constrained ℓ_1 minimization estimators for $\boldsymbol{\beta}$ of model (1.1). It is expected that our framework, inequalities and tools will be useful in other high-dimensional inference problems that involve dependent errors.

In our theoretical framework, we shall adopt the dependence concept of [43]. Assume that the errors (e_i) in (1.1) has the form

$$e_i = g(\ldots, \varepsilon_{i-1}, \varepsilon_i), \tag{1.2}$$

where $\varepsilon_i, i \in \mathbb{Z}$, are i.i.d. random variables, and $g(\cdot)$ is a measurable function. It has a clear physical meaning, where (ε_i) are the inputs and (e_i) are the outputs. Such a representation is very natural for modeling time series. It was studied by [42] for representing stationary and ergodic processes, and it is sometimes called nonlinear Wold representation. The framework (1.2) is general enough to include a wide range of stochastic processes ([38, 44]). It subsumes linear processes, their nonlinear transforms, as well as the Volterra processes that involve interactions between the innovations. The representation (1.2) also includes recursive model of the form $e_i = G(e_{i-1}, \varepsilon_i)$, which includes Markov chain models and nonlinear autoregressive models such as threshold autoregressive models, autoregressive models with conditional heteroscedasticity (ARCH), exponential autoregressive models, bilinear autoregressive models etc. Therefore, there is no much loss of generality by assuming representation (1.2).

One advantage of the representation (1.2) is that it enables us to define physically meaningful and easily workable dependent measures of the process (e_i). Since the inputs (ε_i) are i.i.d., all the dependencies among the outputs (e_i) are caused by the input-output transformation $g(\cdot)$. Therefore, we can define the dependence measures in terms of how the outputs are affected by the inputs, or how the change of the inputs leads to the change in the outputs. Specifically, assume $\|e_i\|_q := (\mathbb{E}|e_i|^q)^{1/q} < \infty$, $q \geq 1$, define the functional dependence measure

$$\delta_{i,q} = \|e_i - e_i^*\|_q = \|e_i - g(\mathcal{F}_i^*)\|_q = \|g(\mathcal{F}_i) - g(\mathcal{F}_i^*)\|_q, \tag{1.3}$$

where $\mathcal{F}_i = (\cdots, \varepsilon_{i-1}, \varepsilon_i)$ and the coupled process

$$e_i^* = g(\mathcal{F}_i^*), \ \mathcal{F}_i^* = (\ldots, \varepsilon_{-1}, \varepsilon_0', \varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_i), \tag{1.4}$$

with $\varepsilon_0', \varepsilon_j, j \in \mathbb{Z}$, being i.i.d. Intuitively, $\delta_{i,q}$ measures the dependency of e_i on ε_0, i.e., how replacing ε_0 by an i.i.d. copy while freezing all other innovations affects the output e_i. We can interpret $\delta_{i,q}$ as nonlinear impulse response function.

We shall assume short-range dependence so that

$$\Delta_{m,q} := \sum_{i=m}^{\infty} \delta_{i,q} < \infty. \tag{1.5}$$

Then for fixed m, $\Delta_{m,q}$ measures the cumulative effect of ε_0 on $(e_i)_{i\geq m}$. Condition (1.5) assumes that the cumulative effect is finite. As a closely related concept, we define the predictive dependence measure

$$\theta_{l,q} = \|\mathbb{E}(e_l|\mathcal{F}_0) - \mathbb{E}(e_l|\mathcal{F}_{-1})\|_q = \|\mathcal{P}_0 e_l\|_q, \tag{1.6}$$

where $\mathcal{P}_i \cdot = \mathbb{E}(\cdot|\mathcal{F}_i) - \mathbb{E}(\cdot|\mathcal{F}_{i-1})$ is the projection operator. Note that $\theta_{l,q}$ also measures the input-output dependence in the representation (1.2) by quantifying how much the prediction of e_l changes by concealing ε_0 from $\mathcal{F}_0$. Similar to $\Delta_{m,q}$, we can also define the cumulative predictive dependence measure

$$\Theta_{m,q} := \sum_{l=m}^{\infty} \theta_{l,q}. \tag{1.7}$$

Compared to the mixing conditions such as the α-, β-, ϕ-, and ρ-mixing in the literature, our $\delta_{i,q}$ and $\theta_{l,q}$ are more physically meaningful and easier to use. In many situations theoretical results based on them are optimal or nearly optimal. They lead to natural definitions for norms of random processes by adjusting for dependence; see the dependence-adjusted $\mathcal{L}^q$ norm (2.8) and sub-exponential norm (2.21).

Equipped with the above dependence measures, we shall establish inequalities for tail probabilities, including an exponential inequality and Nagaev-type inequalities for dependent random variables. The latter are generalizations of the classical Nagaev inequality [30] that deals with independent random variables. Using the functional and predictive dependence measures $\delta_{i,q}$ and $\theta_{l,q}$ introduced above, we show that if the dependence does not exceed a threshold, then our Nagaev-type inequality is as sharp as the original Nagaev inequality under independence. If the dependence of (e_i) is stronger, then the tail can be heavier and a correction should be used. Our form of tail probability inequalities is neat, easy-to-use and, in many cases, sharp.

With our probability inequalities as primary tools, we can analyze the properties of the Lasso and the constrained ℓ_1 minimization estimators under dependent and/or non-sub-Gaussian errors and covariates in the context where p is much larger than n. In comparison with the traditional situation where the errors (e_i) are i.i.d. with finite exponential moments, we shall show that the allowed range of the dimension p can be narrower in our setting, though it can still allow the high-dimensional situation with $p > n$. Roughly speaking, p can be at most a power of n if e_i has only finite polynomial moment and the power is related to the moment condition of e_i. Also the convergence can become slower due to the dependencies in errors as well as in the covariates. We shall give a detailed description on how the dependence measures and moment conditions

of the errors and the covariates affect the rates of convergence and the selection consistencies of the estimators.

The rest of the paper is structured as follows. Section 2 presents exponential and Nagaev inequalities, using the framework of functional and predictive dependence measures. Section 3 deals with the constrained ℓ_1 minimization estimators in the random design scheme in which the covariate process $(\mathbf{x}_i)$ in (1.1) is a high-dimensional stationary process. Lasso estimators with deterministic covariates are treated in Section 4. In both sections we present rates of convergence, model selection consistency and support recovery results. A simulation study is carried out in Section 5.

We now introduce some notation. For a matrix $A = (a_{ij})_{i\le I, j\le J}$, we define $|A|_q = (\sum_{i=1}^I \sum_{j=1}^J |a_{ij}|^q)^{1/q}$, $q > 0$, $|A|_\infty = \max_{ij} |a_{ij}|$, and $|A|_0 = \#\{(i,j) : a_{ij} \neq 0\}$. Define the matrix norm $|||A|||_q = \max_{\mathbf{x}\neq 0} |A\mathbf{x}|_q/|\mathbf{x}|_q$. Hence $|||A|||_1 = \max_{j\le J} \sum_{i=1}^I |a_{ij}|$, and $|||A|||_2$ is the spectral norm. For a random variable W, we write $W \in \mathcal{L}^q$, $q \ge 1$, if $\|W\|_q := [\mathbb{E}(|W|^q)]^{1/q} < \infty$. We use $C, C_1, C_2, \ldots$ to denote constants that do not depend on p and n and they may vary from place to place.

2. Probability inequalities under dependence

Exponential inequalities play a fundamental role in high dimensional inference. In this section we shall present new and powerful inequalities for tail probabilities of weighted sums of dependent and/or non-sub-Gaussian random variables. In Sections 2.1 and 2.2 we provide Nagaev inequalities (cf. Theorems 1 and 2) for linear and nonlinear processes, respectively. The processes can be non-sub-Gaussian ones that do not have finite exponential moments. If the error process satisfies stronger moment condition that it has finite moments of all orders, then under suitable dependence conditions we can have an exponential inequality which is optimal in view of [21]; cf Theorem 3 in Section 2.3. The functional dependence measure provides a convenient framework and it greatly facilitates the formulation of such inequalities.

The Nagaev inequality for tail probability is a useful result in probability theory. However, it appears little known in statistical community. As a result, some of the performance bounds obtained by the Markov inequality in the statistical literature under polynomial moment conditions are not sharp. Let $X_1, \ldots, X_n$ be mean 0 independent random variables with $\|X_i\|_q = [\mathbb{E}(|X_i|^q)]^{1/q} < \infty$, $q > 2$; let $S_n = \sum_{i=1}^n X_i$, $\mu_{n,q} = \sum_{i=1}^n \mathbb{E}(|X_i|^q)$ and $c_q = 2e^{-q}(q+2)^{-2}$. By Corollary 1.7 in [30], for $x > 0$, the tail probability

$$\mathbb{P}(|S_n| \ge x) \le (1 + 2/q)^q \frac{\mu_{n,q}}{x^q} + 2\exp\left(-\frac{c_q x^2}{\mu_{n,2}}\right). \tag{2.1}$$

Inequality (2.1) implies two types of bounds for $\mathbb{P}(|S_n| \ge x)$. For moderate deviation, if x^2 is around the variance $\mu_{n,2}$, then one can use the Gaussian type tail. For large deviation, namely if x^2 is much bigger than the variance $\mu_{n,2}$, then

the polynomial tail dominates. If one applies the Markov and the Rosenthal [35] inequalities, one has

$$\mathbb{P}(|S_n| \geq x) \leq \frac{\mathbb{E}(|S_n|^q)}{x^q} \leq c'_q \frac{\mu_{n,q} + \mu_{n,2}^{q/2}}{x^q}, \tag{2.2}$$

for some constant c'_q only depends on q. Simple calculations show that, up to a multiplicative constant, the upper bound (2.1) is sharper than the one in (2.2), especially when x is big. For example, when X_i are i.i.d., if x is big, (2.1) yields the bound $O(n/x^q)$, while (2.2) gives a worse bound $O(n^{q/2}/x^q)$.

Here we shall present probability inequalities for dependent errors. Consider the weighted sum of the form $S_n = a_1 e_1 + \ldots + a_n e_n$, where $a_1, \ldots, a_n$ are fixed coefficients. Write $a = (a_1, \ldots, a_n)^\top$. In Theorems 1, 2 and 3 below we assume that $|a|_2^2 = \sum_{i=1}^n a_i^2 = n$. Recall that $|a|_q = (\sum_{i=1}^n |a_i|^q)^{1/q}$.

2.1. Nagaev inequality for linear processes

To begin with, in Theorem 1, we assume that (e_i) follow a linear process

$$e_i = \sum_{j=0}^{\infty} f_j \varepsilon_{i-j}, \tag{2.3}$$

where ε_j, $j \in \mathbb{Z}$, are i.i.d. with mean zero and $\varepsilon_j \in \mathcal{L}^q$, $q > 2$, and f_j are real coefficients with $|f|_2^2 := \sum_{j=0}^{\infty} f_j^2 < \infty$, so that by Kolmogorov's three series theorem e_i exists. Linear processes are widely used in practice and they include the popular ARMA processes.

Theorem 1 (Nagaev inequalities for linear processes)**.** *Assume (2.3).*

(i) (Short-range dependence) Let $c_q = 2e^{-q}(q+2)^{-2}$. *If* $|f|_1 := \sum_{j=0}^{\infty} |f_j| < \infty$, *then*

$$\mathbb{P}(|S_n| \geq x) \leq (1 + 2/q)^q \frac{|a|_q^q |f|_1^q \|\varepsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{n|f|_1^2 \|\varepsilon_0\|_2^2}\right). \tag{2.4}$$

(ii) (Long-range dependence). Assume $K = \sup_{j \geq 0} |f_j|(1+j)^\beta < \infty$, *where* $1/2 < \beta < 1$. *Then there exists constants* C_1, C_2, *only depend on* q *and* β *such that*

$$\mathbb{P}(|S_n| \geq x) \leq C_1 \frac{K^q |a|_q^q n^{q(1-\beta)} \|\varepsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{C_2 x^2}{n^{3-2\beta}\|\varepsilon_0\|_2^2 K^2}\right). \tag{2.5}$$

Proof of Theorem 1. Let $f_j = 0$ for $j < 0$. We write $S_n = \sum_{j \in \mathbb{Z}} b_j \varepsilon_j$, where $b_j = \sum_{i=1}^n a_i f_{i-j}$. Let $q' = q/(q-1)$. By Hölder's inequality,

$$\sum_{j \in \mathbb{Z}} |b_j|^q \leq \sum_{j \in \mathbb{Z}} \left(\sum_{i=1}^n |a_i|^q |f_{i-j}|\right) \left(\sum_{i=1}^n |f_{i-j}|\right)^{q/q'}$$

$$\leq \sum_{j\in\mathbb{Z}}\left(\sum_{i=1}^{n}|a_i|^q|f_{i-j}|\right)|f|_1^{q/q'} \leq |a|_q^q|f|_1^q. \tag{2.6}$$

Clearly $\sum_{j\in\mathbb{Z}}|b_j|^2 \leq |a|_2^2|f|_1^2$. Hence (2.4) follows from the original Nagaev inequality (2.1). (ii). Let $f_m^* = \max_{j\geq m}|f_j|$ and $F_n = \sum_{j=0}^n |f_j|$. Then $F_n \leq K\sum_{l=1}^{n+1} l^{-\beta} \leq K(n+1)^{1-\beta}/(1-\beta)$. By (2.6), we have

$$\sum_{j=1-n}^{n}|b_j|^q \leq \sum_{j=1-n}^{n}\left(\sum_{i=1}^{n}|a_i|^q|f_{i-j}|\right)F_{2n}^{q/q'} \leq |a|_q^q F_{2n}^q. \tag{2.7}$$

If $j \leq -n$, $|b_j| \leq |a|_1 f_{1-j}^*$. Then $\sum_{j\leq -n}|b_j|^q \leq |a|_1^q\sum_{j\leq -n}(f_{1-j}^*)^q$. Note that $|a|_1 \leq n^{1-1/q}|a|_q$ and $\sum_{j\leq -n}(f_{1-j}^*)^q \leq C_3 n(Kn^{-\beta})^q$. Hence by (2.7) $\sum_{j\in\mathbb{Z}}|b_j|^q \leq C_4|a|_q^q K^q n^{q(1-\beta)}$. Similarly $\sum_{j\in\mathbb{Z}}|b_j|^2 \leq C_5|a|_2^2K^2n^{2(1-\beta)}$. So (2.5) follows. □

2.2. Nagaev inequality for nonlinear processes

For nonlinear processes, with functional dependence measure $\Delta_{m,q}$ in (1.5), we have Theorem 2, a Nagaev-type inequality for $S_n = a_1e_1 + \ldots + a_ne_n$. The special case of Theorem 2 with $a_1 = \ldots = a_n = 1$ was treated in [23]. As an important improvement, in the stronger dependence case with slow decay of $\Delta_{m,q}$, Theorem 2 provides a sharper bound than the one in the latter paper. To account for dependence, for the process $e_\cdot = (e_i)_{i=-\infty}^{\infty}$ we introduce the following *dependence adjusted norm* (DAN)

$$\|e_\cdot\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^\alpha\Delta_{m,q} = \sup_{m\geq 0}(m+1)^\alpha\sum_{i=m}^{\infty}\delta_{i,q}, \quad \alpha \geq 0. \tag{2.8}$$

It can happen that, due to dependence, $\|e_\cdot\|_{q,\alpha} = \infty$ while $\|e_i\|_q < \infty$. Since $e_0 = \sum_{l=-\infty}^{0}\mathcal{P}_l e_0$, we have $\Delta_{0,q} = \|e_\cdot\|_{q,0}$ and

$$\|e_0\|_q \leq \sum_{i=0}^{\infty}\|\mathcal{P}_{-i}e_0\|_q = \sum_{i=0}^{\infty}\|\mathcal{P}_0 e_i\|_q \leq \sum_{i=0}^{\infty}\|e_i - e_i^*\|_q = \Delta_{0,q},$$

by stationarity, Jensen's inequality and the fact that $\mathcal{P}_0e_i = \mathbb{E}(e_i - e_i^*|\mathcal{F}_0)$. If $e_i \in \mathcal{L}^q$ are i.i.d. with mean 0, then $\delta_{i,q} = 0$ for all $i \geq 1$, and $\delta_{0,q} = \|e_0 - e_0'\|_q$. Note that in this case the dependence-adjusted norm $\|e_\cdot\|_{q,\alpha}$ and the $\mathcal{L}^q$ norm $\|e_0\|_q$ are equivalent in the sense that $\|e_0\|_q \leq \delta_{0,q} \leq \|e_0\|_q + \|e_0'\|_q = 2\|e_0\|_q$.

Theorem 2. *Assume that $\|e_\cdot\|_{q,\alpha} < \infty$, where $q > 2$ and $\alpha > 0$, and $\sum_{i=1}^n a_i^2 = n$. Let $\varpi_n = 1$ (resp. $(\log n)^{1+2q}$ or $n^{q/2-1-\alpha q}$) if $\alpha > 1/2 - 1/q$ (resp. $\alpha = 0$ or $\alpha < 1/2 - 1/q$). Then for all $x > 0$,*

$$\mathbb{P}(|S_n| \geq x) \leq C_1\frac{\varpi_n|a|_q^q\|e_\cdot\|_{q,\alpha}^q}{x^q} + C_2\exp\left(-\frac{C_3x^2}{n\|e_\cdot\|_{2,\alpha}^2}\right), \tag{2.9}$$

where C_1, C_2, C_3 are constants that only depend of q and α.

The Nagaev inequality of form (2.9) provides a very natural extension of the classical one (2.1) in that the dependence-adjusted qth norm $\|e.\|_{q,\alpha}$ plays the role of the $\mathcal{L}^q$ norm $\|X_i\|_q$, while the dependence-adjusted 2nd order norm $\|e.\|_{2,\alpha}$ play the role of the $\mathcal{L}^2$ norm $\|X_i\|_2$.

Proof of Theorem 2. In the proof the constants C_1, C_2, ..., may change from line to line. They only depend on q and α, and are independent of n, $(a_i)_{i=1}^n$ and x. It suffices to deal with the case in which $x \geq \sqrt{n}\|e.\|_{2,\alpha}$ since otherwise (2.9) trivially holds. Let $L = \lfloor (\log n)/(\log 2) \rfloor$, $\tau_l = 2^l$ if $1 \leq l < L$ and $\tau_L = n$. Define $e_{i,\tau} = \mathbb{E}(e_i | \varepsilon_{i-\tau}, \varepsilon_{i-\tau+1}, \ldots, \varepsilon_i)$, $\tau \geq 0$, and

$$M_{i,l} = \sum_{k=1}^{i} a_k (e_{k,\tau_l} - e_{k,\tau_{l-1}}). \tag{2.10}$$

Let $S_{n,m} = \sum_{k=1}^n a_k e_{k,m}$ and write

$$S_n = S_{n,0} + (S_n - S_{n,n}) + \sum_{l=1}^{L} M_{n,l}. \tag{2.11}$$

The proof is based on the above decomposition. Note that the summands $a_k e_{k,0}$ of $S_{n,0}$ are independent. By the Nagaev inequality (2.1),

$$\mathbb{P}(|S_{n,0}| \geq x) \leq c_q \frac{|a|_q^q \|e_0\|_q^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{n\|e_0\|^2}\right), \tag{2.12}$$

where c_q is a constant only depending on q, and it may vary at each occurrence. By the Burkholder inequality [7],

$$\|S_n - S_{n,n}\|_q \leq \sum_{m=n}^{\infty} \|S_{n,m+1} - S_{n,m}\|_q \leq \sum_{m=n}^{\infty} c_q n^{1/2} \delta_{m+1,q} = c_q n^{1/2} \Delta_{n+1,q},$$

which in view of the Markov inequality implies

$$\mathbb{P}(|S_n - S_{n,n}| \geq x) \leq \frac{\|S_n - S_{n,n}\|_q^q}{x^q} \leq c_q \frac{n^{q/2} \Delta_{n+1,q}^q}{x^q}. \tag{2.13}$$

Let $\tilde{\delta}_{l,q} = \sum_{t=1+\tau_{l-1}}^{\tau_l} \delta_{t,q}$ and $\tilde{\delta}_{l,2} = \sum_{t=1+\tau_{l-1}}^{\tau_l} \delta_{t,2}$. For $1 \leq i < i' \leq n$, we have by the Burkholder inequality and the Hölder inequality that

$$\begin{aligned}
\|M_{i',l} - M_{i,l}\|_q &\leq \sum_{t=1+\tau_{l-1}}^{\tau_l} \left\| \sum_{k=i+1}^{i'} a_k (e_{k,t} - e_{k,t-1}) \right\|_q \\
&\leq \sum_{t=1+\tau_{l-1}}^{\tau_l} c_q \left(\sum_{k=i+1}^{i'} a_k^2 \delta_{t,q}^2 \right)^{1/2} \\
&\leq \frac{c_q (\sum_{k=i+1}^{i'} a_k^q)^{1/q}}{(i'-i)^{1/q-1/2}} \tilde{\delta}_{l,q}.
\end{aligned} \tag{2.14}$$

By definition the summands $D_k = a_k(e_{k,\tau_l} - e_{k,\tau_{l-1}})$ of $M_{n,l}$ are τ_l-dependent. Let $\mathcal{A} = \{2\tau_l i + j : i \in \mathbb{Z}, 1 \le j \le \tau_l\}$, $\mathcal{B} = \{2\tau_l i + j : i \in \mathbb{Z}, 1 + \tau_l \le j \le 2\tau_l\}$, $A_n = \sum_{k \le n, k \in \mathcal{A}} D_k$ and $B_n = \sum_{k \le n, k \in \mathcal{B}} D_k$. Then $\mathbb{E}(A_n^2) \le n\tilde{\delta}_{l,2}^2$ and $\mathbb{E}(B_n^2) \le n\tilde{\delta}_{l,2}^2$. By the τ_l-dependence, (2.14) and (2.1),

$$\mathbb{P}\left(|A_n| \ge y\right) \le c_q \frac{|a|_q^q}{y^q} \tau_l^{q/2-1} \tilde{\delta}_{l,q}^q + 2\exp\left(-\frac{c_q y^2}{n\tilde{\delta}_{l,2}^2}\right).$$

A similar inequality holds for $\mathbb{P}\left(|B_n| \ge y\right)$. Since $M_{n,l} = A_n + B_n$,

$$\mathbb{P}\left(|M_{n,l}| \ge 2y\right) \le 2c_q \frac{|a|_q^q}{y^q} \tau_l^{q/2-1} \tilde{\delta}_{l,q}^q + 4\exp\left(-\frac{c_q y^2}{n\tilde{\delta}_{l,2}^2}\right). \tag{2.15}$$

Let $c = q/2 - 1 - \alpha q$; let $\lambda_l = l^{-2}/(\pi^2/3)$ if $1 \le l \le L/2$ and $\lambda_l = (L+1-l)^{-2}/(\pi^2/3)$ if $L/2 < l \le L$. Then $\sum_{l=1}^L \lambda_l < 1$. Noting that $\tilde{\delta}_{l,q} \le \Delta_{1+\tau_{l-1},q} \le \|e.\|_{q,\alpha}(2+\tau_{l-1})^{-\alpha}$ and $\tilde{\delta}_{l,2} \le \|e.\|_{2,\alpha}(2+\tau_{l-1})^{-\alpha}$. Then (2.15) implies

$$\begin{aligned}
\mathbb{P}(|\sum_{l=1}^L M_{n,l}| \ge x) &\le \sum_{l=1}^L \mathbb{P}(|M_{n,l}| \ge \lambda_l x) \\
&\le c_q \frac{|a|_q^q}{x^q} \sum_{l=1}^L \frac{\tau_l^{q/2-1}\tilde{\delta}_{l,q}^q}{\lambda_l^q} + 2\sum_{l=1}^L \exp\left(-\frac{c_q x^2 \lambda_l^2}{n\tilde{\delta}_{l,2}^2}\right) \\
&\le C_4 \frac{|a|_q^q \|e.\|_{q,\alpha}^q}{x^q} \sum_{l=1}^L \frac{\tau_l^c}{\lambda_l^q} \\
&\quad + C_5 \sum_{l=1}^L \exp\left(-C_6 n^{-1} x^2 \lambda_l^2 \tau_l^{2\alpha} / \|e.\|_{2,\alpha}^2\right).
\end{aligned} \tag{2.16}$$

By the definitions of τ_l and λ_l, we have $\phi := \min_{l \ge 1} \lambda_l^2 \tau_l^{2\alpha} > 0$. By elementary manipulations, there exists a constant $C_7 > 1$ such that for all $u \ge 1$, we have

$$\sum_{l=1}^L \exp\left(-C_6 u \lambda_l^2 \tau_l^{2\alpha}\right) \le C_7 \exp(-C_6 u \phi). \tag{2.17}$$

We shall apply (2.17) with $u = x^2/(n\|e.\|_{2,\alpha}^2)$. Observe that, if $c > 0$, we have $\sum_{l=1}^L \tau_l^c/\lambda_l^q \le C_8 \tau_L^c = C_8 n^c$. If $c < 0$, then $\sum_{l=1}^L \tau_l^c/\lambda_l^q \le C_9$. Hence, by (2.11), (2.12), (2.13), (2.16) and (2.17), both cases with $c < 0$, $c = 0$ and $c > 0$ of Theorem 2 follow. □

Remark 1. In the stronger dependence case $0 < \alpha < 1/2 - 1/q$, when $a_1 = \ldots = a_n = 1$, [23] obtained the following inequality

$$\mathbb{P}(|S_n| \ge x) \le C_1' \frac{n^{q/2-\alpha q}}{x^q} + C_2' \exp[-C_3' x^2 n^{(1+2\alpha q-2q)/(1+q)}], \tag{2.18}$$

where C_1', C_2', C_3' are constants that may depend on the dependence condition $\Delta_{m,q} = O(m^{-\alpha})$. Since $\alpha < 1/2 - 1/q$, $(1 + 2\alpha q - 2q)/(1 + q) < -1$. Then the neater and simpler bound $\exp(-C_3 x^2/(n\|e.\|_{2,\alpha}^2))$ in Theorem 2(ii) is sharper than the one in (2.18). Additionally our form (2.9) is easier to use since the constants C_1, C_2, C_3 therein only depend on α and q. □

Remark 2. To appreciate the sharpness of inequality (2.9), we assume that all $a_i > 0$, e_i are i.i.d. with mean 0, variance 1 and, for some constant $h_0 > 0$,

$$\mathbb{P}(e_i \geq x) = \frac{h_0}{x^q}(1 + o(1)) \text{ as } x \to \infty. \tag{2.19}$$

By Theorem 2.1 in [32], we have for all $x \geq \sqrt{n}$ that

$$\mathbb{P}(S_n \geq x) = \frac{h_0 + o(1)}{x^q} \sum_{i=1}^{n} a_i^q + (1 + o(1))(1 - \Phi(x/\sqrt{n})), \tag{2.20}$$

where Φ is the standard normal cumulative distribution function. Let $t_n = [n \log(n^{-q/2}|a|_q^q)]^{1/2}$. If $x \leq c_1 t_n$ with $c_1 < \sqrt{2}$, then the Gaussian part in (2.20) dominates, while for large x with $x \geq c_2 t_n$ with $c_2 > \sqrt{2}$, the power decaying term $h_0|a|_q^q x^{-q}$ dominates. Note that (2.20) is asymptotically exact. Hence inequality (2.9) provides a nearly optimal bound for both large and small x. □

2.3. Exponential tail bounds

If e_i satisfies stronger moment condition than the existence of finite qth moment, we expect that a stronger form than (2.9) exists. Indeed, we have the following Theorem 3, an exponential inequality, which is a generalization and an improvement of Theorem 2 in [43] by allowing weights and by providing an explicit close-form upper bound. Write $\Theta_q = \Theta_{0,q}$. We shall assume stronger moment condition by allowing $\Theta_q < \infty$ for all $q > 0$, and we further assume that Θ_q increases slower than Cq^ν for some $\nu \geq 0$ in the following sense:

$$\|e.\|_{\psi_\nu} := \sup_{q \geq 2} q^{-\nu} \Theta_q < \infty. \tag{2.21}$$

In view of its definition, we can interpret Θ_q as the *predictive persistence* of the process (e_i). In the very special case in which e_i are i.i.d., we have $\Theta_q = \|e_0\|_q$ and $\nu = 1$ (resp. $\nu = 1/2$) if e_i are sub-exponential (resp. sub-Gaussian). In this case $\|e.\|_{\psi_\nu}$ is the sub-Gaussian or sub-exponential norms of a random variable; see for example Section 5.2.3 in [39]. Hence $\|e.\|_{\psi_\nu}$ can be naturally interpreted as *dependence-adjusted sub-exponential or sub-Gaussian norm.*

Theorem 3. *Assume (2.21). Let* $Z_n = S_n/\sqrt{n}$ *and* $\alpha = 2/(1 + 2\nu)$. *Then*

$$m(t) := \sup_{n \in \mathbb{N}} \mathbb{E}[\exp(t|Z_n|^\alpha)] \leq 1 + c_\alpha (1 - t/t_0)^{-1/2} t/t_0 \tag{2.22}$$

holds for $0 \le t < t_0$ *with* $t_0 = (e\alpha\gamma_0^\alpha)^{-1}$, *where* $\gamma_0 = \|e.\|_{\psi_\nu}$, c_α *is a constant only depending on* α. *Consequently, letting* $t = t_0/2$, *for* $u > 0$, *we have*

$$\mathbb{P}(Z_n > u) \le \exp(-tu^\alpha)m(t) \le (1 + c_\alpha/\sqrt{2})e^{-(u/\gamma_0)^\alpha/(2e\alpha)}. \tag{2.23}$$

Proof of Theorem 3. Note that $1/2-1/\alpha = -\nu$. We adopt the argument in [43]. Let $M_{n,l} = \sum_{i=1}^n a_i \mathcal{P}_{i-l} e_i$, $l \ge 0$. Then $M_{n,l}$ is a martingale. By Burkholder's inequality, we have for all $q \ge 2$ that

$$\|M_{n,l}\|_q^2 \le (q-1)\sum_{i=1}^n \|a_i \mathcal{P}_{i-l} e_i\|_q^2 = (q-1)n\theta_{l,q}^2. \tag{2.24}$$

Hence $\|Z_n\|_q \le (q-1)^{1/2}\Theta_q$ in view of the decomposition $S_n = \sum_{l=0}^\infty M_{n,l}$. Write the negative binomial expansion $(1-s)^{-1/2} = 1 + \sum_{k=1}^\infty a_k s^k$, where $|s| < 1$ and $a_k = (2k)!/(2^{2k}(k!)^2)$. By Stirling's formula, as $k \to \infty$, $a_k \sim (k\pi)^{-1/2}$. Hence $k! \sim \sqrt{2}(k/e)^k a_k^{-1}$, and there exists absolute constants $c_1, c_2 > 0$ such that $c_1(k/e)^k a_k^{-1} \le k! \le c_2(k/e)^k a_k^{-1}$ holds for all $k \ge 1$. By (2.21), if $k\alpha \ge 2$, we have $\Theta_{\alpha k} \le \gamma_0(\alpha k)^\nu$ and hence by elementary manipulations

$$\frac{t^k\|Z_n^\alpha\|_k^k}{k!} \le \frac{t^k(\alpha k-1)^{\alpha k/2}\Theta_{\alpha k}^{\alpha k}}{c_1(k/e)^k a_k^{-1}} \le \frac{a_k t^k}{c_1 t_0^k}\frac{(\alpha k-1)^{\alpha k/2}}{(\alpha k)^{\alpha k/2}} \le \frac{a_k t^k}{c_1 t_0^k\sqrt{e}}. \tag{2.25}$$

If $k\alpha < 2$, then $\|Z_n\|_{\alpha k} \le \|Z_n\|_2 \le 2^\nu\gamma_0$. Using $e^x = \sum_{k=0}^\infty x^k/k!$, we obtain

$$\begin{aligned} m(t) &\le 1 + \sum_{1\le k<2/\alpha} \frac{t^k(2^\nu\gamma_0)^{\alpha k}}{k!} + \sum_{k\ge 2/\alpha} \frac{a_k}{c_1\sqrt{e}}\frac{t^k}{t_0^k} \\ &\le 1 + c'_\alpha \sum_{k=1}^\infty a_k \frac{t^k}{t_0^k} \le 1 + c_\alpha \frac{t/t_0}{(1-t/t_0)^{1/2}}, \end{aligned}$$

where constants $c_\alpha, c'_\alpha > 0$ only depend on α. Clearly (2.23) follows from the Markov inequality. □

Remark 3. Note that condition (2.21) is equivalent to

$$\gamma := \limsup_{q\to\infty} q^{-\nu}\Theta_q < \infty. \tag{2.26}$$

Let $t'_0 = (e\alpha\gamma^\alpha)^{-1}$. By the argument in the proof of Theorem 3, we have $m(t) < \infty$ if $0 \le t < t'_0$ in view of

$$\limsup_{k\to\infty} \frac{t\|Z_n^\alpha\|_k}{(k!)^{1/k}} \le \limsup_{k\to\infty} \frac{t(\alpha k-1)^{\alpha/2}\Theta_{\alpha k}^\alpha}{k/e} \le te\alpha\gamma^\alpha < 1. \tag{2.27}$$

Since $\gamma \le \gamma_0$, the range $t < t'_0$ is wider than the one $t < t_0$ in Theorem 3. □

Remark 4. The exponential inequality in Theorem 3 is optimal for martingale differences with finite exponential moment. Let $D_i, i \in \mathbb{Z}$, be a stationary

martingale difference sequence with $\mathbb{E}(D_i^2) = 1$ and finite exponential moment $\mathbb{E}\exp(|D_0|) < \infty$. Then $\mathbb{E}(|D_0|^k) = o(k!)$ as $k \to \infty$. Note that $\mathcal{P}_i D_0 = 0$ if $i \geq 1$. Then $\Theta_q = \|D_0\|_q$. Hence (2.21) holds with $1/2 - 1/\alpha = -1$, or $\alpha = 2/3$. Let $S_n = \sum_{i=1}^n D_i$. By (2.23), there exists $c_1, c_2 > 0$ such that for all $x > 0$,

$$\mathbb{P}(|S_n| \geq nx) \leq c_1 \exp(-c_2 n^{1/3} x^{2/3}), \tag{2.28}$$

by letting $u = \sqrt{n}x$. In comparison with Theorem 3.2 in [21], (2.28) is optimal up to a constant. They also proved that the inequality $\mathbb{P}(|S_n| \geq n) \leq c_1 \exp(-c_2 n^{1/3})$ is the best possible under the condition $\mathbb{E}\exp(|D_0|) < \infty$. If D_i is bounded (say by $b > 0$), using Azuma's inequality [1], one can have the bound $\mathbb{P}(|S_n| \geq nx) \leq 2\exp(-nx^2/(2b^2))$. Again in this case our inequality (2.23) is sharp up to a multiplicative constant by letting $\alpha = 2$ since $\Theta_q = \|D_0\|_q \leq b$ for all q, so that (2.21) holds with $\gamma = b$. □

3. Constrained ℓ_1 minimization estimator with random design

Using the constrained ℓ_1 minimization approach, one estimates $\boldsymbol{\beta}$ in (1.1) by

$$\hat{\boldsymbol{\beta}} = \arg\min |\boldsymbol{\beta}|_1 \text{ subject to } |X^\top X\boldsymbol{\beta} - X^\top Y|_\infty \leq \lambda, \tag{3.1}$$

where $\lambda > 0$ is a thresholding parameter. Recall here for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, the 1-norm $|\boldsymbol{\beta}|_1 = \sum_{j=1}^p |\beta_j|$. Properties of the Dantzig estimator (3.1) has been extensively studied in the literature; see [9]. In most of the previous papers it is assumed that the error sequence (e_i) is i.i.d. and/or sub-Gaussian. Following [8], we call $\hat{\boldsymbol{\beta}}$ the *Clime estimator*.

In our random design setting we assume that in model (1.1) the covariate process $(\mathbf{x}_i, i = 1, ..., n)$ is high-dimensional stationary of the form

$$\mathbf{x}_i = \mathbf{h}(\mathcal{F}_i), \quad \mathcal{F}_i = (\ldots, \varepsilon_{i-1}, \varepsilon_i), \tag{3.2}$$

where (ε_i) are i.i.d. random vectors, and $\mathbf{h}(\cdot) = (h_1(\cdot), \ldots, h_p(\cdot))^\top$ is a measurable function in $\mathbb{R}^p$. With $\mathbf{x}_i$ defined in (3.2), letting ε_i be i.i.d. random vectors, we can allow models with homogeneous or heteroscedastic errors; see Example 2. In the former homogeneous case, the covariance process $(\mathbf{x}_i)$ and the error process (e_i) can be independent to each other. Similar to $\delta_{i,q}$ in (1.3), assume that $\mathbf{x}_i \in \mathcal{L}^\iota$, $\iota > 2$, and we define the functional dependence measure

$$\phi_{i,\iota} = \max_{1 \leq j \leq p} \|h_j(\mathcal{F}_i) - h_j(\mathcal{F}_i^*)\|_\iota, \tag{3.3}$$

Similar to $\Delta_{m,q}$, we can define and assume

$$\Phi_{m,\iota} := \sum_{i=m}^{\infty} \phi_{i,\iota} < \infty. \tag{3.4}$$

Two important cases of vector autoregressive processes and linear stochastic models are given in Examples 1 and 2 below and they are widely used in practice.

Example 1 (Vector autoregressive model)**.** Let $Z_i = (Z_{i1}, \ldots, Z_{ip})^\top$ be i.i.d. mean 0 random vectors; let $A_1, \ldots, A_d$ be $p \times p$ coefficient matrices such that the vector AR(d) process $W_i = (W_{i1}, \ldots, W_{ip})^\top$ given by

$$W_i = A_1 W_{i-1} + \ldots + A_d W_{i-d} + Z_i, \tag{3.5}$$

has a stationary solution. Let $\mathcal{F}_i = (\ldots, Z_{i-1}, Z_i)$ and $\mathbf{x}_i = (W_{i-1}^\top, \ldots, W_{i-d}^\top)^\top$ be a $pd \times 1$ vector obtained by stacking lag vectors $W_{i-l}, 1 \le l \le d$. Then $\mathbf{x}_i$ is $\mathcal{F}_{i-1}$-measurable. For each $j \le p$, consider the linear regression model

$$W_{i,j} = \mathbf{x}_i^\top \mathbf{b}_j + Z_{i,j}, \tag{3.6}$$

which is of the form (1.1) with $Z_{i,j}$ and $\mathbf{x}_i$ being independent. Here $\mathbf{b}_j$ is a $pd \times 1$ parameter vector. Observe that constrained ℓ_1 minimization estimation of the coefficient matrices $A_1, \ldots, A_d$ of model (3.5) can be decomposed into the p sub-problems of estimating $\mathbf{b}_j, 1 \le j \le p$, of (3.6), thus allowing for parallel computation; see also [8, 16] for similar treatments. □

Example 2 (Linear stochastic models with heteroscedastic errors)**.** Let $\varepsilon_i = (\xi_i, \eta_i)$, where $\eta_i, i \in \mathbb{Z}$, are i.i.d. with mean 0, variance 1, $\xi_i, i \in \mathbb{Z}$, are also i.i.d. random vectors and (η_i) and (ξ_i) are independent. Let

$$\mathbf{x}_i = \mathbf{h}_0(\ldots, \xi_{i-1}, \xi_i) \text{ and } e_i = \sigma(\ldots, \xi_{i-1}, \xi_i)\eta_i, \tag{3.7}$$

where $\mathbf{h}_0(\cdot)$ and $\sigma(\cdot)$ are measurable functions such that $\mathbf{x}_i$ and e_i are properly defined. It is clear that by choosing appropriate $\mathbf{h}(\cdot)$ and $g(\cdot)$, $\mathbf{x}_i$ and e_i can be written in the form of (3.2) and (1.2), respectively. If $\sigma(\cdot) \equiv$ a constant, then $\mathbf{x}_i$ and e_i are independent. If $\sigma(\cdot)$ is not a constant function, then the errors e_i and $\mathbf{x}_i$ are dependent and thus (1.1) is a model with heteroscedastic errors. □

Assume that $(\mathbf{x}_i)$ is centered with $\mathbb{E}(\mathbf{x}_i) = 0$. Further assume that, to avoid collinearity, the covariance matrix $\Sigma_X = \mathbb{E}(\mathbf{x}_i \mathbf{x}_i^\top)$ is non-singular. Denote its inverse by $\Omega_X = \Sigma_X^{-1}$. Write $x_{\cdot j} = (x_{l,j})_{l \in \mathbb{Z}}$ and $e_\cdot = (e_l)_{l \in \mathbb{Z}}$. For two nonnegative sequences (a_n) and (b_n), we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \le C b_n$ holds for all large n. Theorem 4 imposes conditions on dependence-adjusted norms of the processes $x_{\cdot j}$ and $e_\cdot$.

Theorem 4. *(i) Assume that* $\max_{j \le p} \|x_{\cdot j}\|_{\iota, \alpha_X} = N_X < \infty$ *and* $\|e_\cdot\|_{p, \alpha_e} = N_e < \infty$*, where* $q > 2$*,* $\iota > 4$ *and* $\alpha_X, \alpha_e > 0$*. Let* $\chi = 1$ *if* $\alpha_X > 1/2 - 2/\iota$ *and* $\chi = \iota/4 - \alpha_X \iota/2$ *if* $\alpha_X < 1/2 - 2/\iota$*. Assume* $\tau = q\iota/(q + \iota) > 2$ *and let* $\alpha = \min(\alpha_X, \alpha_e)$*. Define* $\pi = 1$ *if* $\alpha > 1/2 - 1/\tau$ *and* $\pi = \tau/2 - \alpha\tau$ *if* $\alpha < 1/2 - 1/\tau$*. Then for all* $a > 0$ *and* $b \ge \lambda/n$*, the inequality*

$$\begin{aligned} \mathbb{P}[|\Sigma_X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_\infty \ge 2(a|\boldsymbol{\beta}|_1 + b)] \quad \lesssim \quad & \frac{p n^\pi N_X^\tau N_e^\tau}{(nb)^\tau} + p e^{-C_2 n b^2/(N_X^2 N_e^2)} \\ & + \frac{p^2 n^\chi N_X^\iota}{(na)^{\iota/2}} + p^2 e^{-C_1 n a^2/N_X^4} \end{aligned} \tag{3.8}$$

holds, where C_1, C_2 *and the constant in* $\lesssim$ *only depend on* q, ι, α_X *and* α_e*. (ii) If* $X_i \in \mathcal{L}^q$ *and* $e_i \in \mathcal{L}^q$ *holds for all* $q > 2$ *and, for some* $\nu, \varrho \ge 0$*,*

$$K_\nu := \sup_{q \ge 2} q^{-\nu} \Delta_{0,q} < \infty, \quad L_\varrho := \sup_{q \ge 2} q^{-\varrho} \Phi_{0,q} < \infty, \tag{3.9}$$

then for all $a > 0, b \geq \lambda/n$,

$$\mathbb{P}[|\Sigma_X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_\infty \geq 2(a|\boldsymbol{\beta}|_1 + b)] \lesssim pe^{-C_2(\sqrt{n}b/(K_\nu L_\varrho))^{2/(1+2\nu+2\varrho)}} + p^2 e^{-C_1(\sqrt{n}a/L_\varrho^2)^{2/(1+4\varrho)}}, \tag{3.10}$$

where $C_1, C_2 > 0$ *and the constant in* $\lesssim$ *only depend on* ν *and* ϱ.

Before proving Theorem 4, we shall provide two examples of high-dimensional time series for which one can bound $N_X = \max_{j\leq p} \|x_{\cdot j}\|_{\iota,\alpha_X}$, a key step in applying this theorem.

Example 3 (High-dimensional linear process)**.** Let $\zeta_{ij}, i, j \in \mathbb{Z}$, be i.i.d. random variables with mean 0, variance 1 and having finite ιth moment, $\iota > 2$; let $A_0, A_1, \ldots,$ be $p \times p$ matrices with real entries such that $\sum_{j=0}^\infty \operatorname{tr}(A_j A_j^\top) < \infty$. Write $\varepsilon_i = (\zeta_{i1}, \ldots, \zeta_{ip})^\top$. Then by Kolmogorov's three series theorem (see for example Corollary 5.1.3 in [11]) the p-dimensional linear process

$$\mathbf{x}_i = \sum_{l=0}^\infty A_l \varepsilon_{i-l} \tag{3.11}$$

is well-defined. The above process is a special case of (3.2) with a linear functional $\mathbf{h}(\cdot)$. Let $A_{l,j}$ be the jth row of A_l. Then by Burkholder's inequality, $\|A_{i,j}\varepsilon_0\|_\iota \leq (\iota-1)^{1/2}|A_{i,j}|_2\|\zeta_{00}\|_\iota$. If there exist $\theta > 1$ and $K > 0$ such that $\max_{j\leq p}|A_{i,j}|_2 \leq K(i+1)^{-\theta}$ hold for all $i \geq 0$, then with $\alpha = \theta - 1$ we have $N_X \leq CK\|\zeta_{00}\|_\iota$, where the constant c only depends on θ and ι. □

Example 4 (High-dimensional ARCH process)**.** Let $\zeta_{ij}, i, j \in \mathbb{Z}$, be i.i.d. random variables with mean 0, variance 1 and having finite ιth moment, $\iota > 2$; let

$$x_{ij} = \zeta_{ij}(b_j^2 + \mathbf{x}_{i-1}^\top A_j \mathbf{x}_{i-1})^{1/2} =: G_{\varepsilon_i}^{(j)}(\mathbf{x}_{i-1}), \quad j = 1, \ldots, p, \tag{3.12}$$

where $A_1, \ldots, A_p$ are $p \times p$ nonnegative-definite matrices and $b_1, \ldots, b_p$ are real numbers. Let $G_{\varepsilon_i}(\cdot) = (G_{\varepsilon_i}^{(1)}(\cdot), \ldots, G_{\varepsilon_i}^{(p)}(\cdot))^\top$ and we abbreviate (3.12) as $\mathbf{x}_i = G_{\varepsilon_i}(\mathbf{x}_{i-1})$. Let λ_j be the spectral norm of A_j. Note that, for $\mathbf{x}, \mathbf{w} \in \mathbb{R}^p$,

$$|(b_j^2 + \mathbf{x}^\top A_j \mathbf{x})^{1/2} - (b_j^2 + \mathbf{w}^\top A_j \mathbf{w})^{1/2}|^2 \leq (\mathbf{x}-\mathbf{w})^\top A_j(\mathbf{x}-\mathbf{w}) \leq \lambda_j|\mathbf{x}-\mathbf{w}|_2^2.$$

Hence $|G_{\varepsilon_i}(\mathbf{x}) - G_{\varepsilon_i}(\mathbf{w})|_2 \leq |\mathbf{x}-\mathbf{w}|_2(\sum_{j=1}^p \lambda_j\zeta_{ij}^2)^{1/2}$. Assume that

$$\|L_0\|_q := [\mathbb{E}(L_0^q)]^{1/q} < 1, \text{ where } L_i = (\sum_{j=1}^p \lambda_j \zeta_{ij}^2)^{1/2}, \tag{3.13}$$

holds for some $0 < q \leq \iota$. Let $K_i = (\sum_{j=1}^p b_j^2\zeta_{ij}^2)^{1/2} = |G_{\varepsilon_i}(\mathbf{0})|_2$. By the arguments for Theorem 2 in [45], recursion (3.12) allows a stationary solution $(\mathbf{x}_i)_i$ and its functional dependence measure

$$\|\mathbf{x}_i - \mathbf{x}_i^*\|_q \leq c_q\|K_0\|_q\|L_0\|_q^i/(1 - \|L_0\|_q), \tag{3.14}$$

where the constant c_q only depends on q. Hence by (3.12) the functional dependence measure for the jth component process $(x_{ij})_i = x_{\cdot j}$ satisfies

$$\|x_{ij} - x_{ij}^*\|_q \le c_q \lambda_j^{1/2} \|\zeta_{00}\|_q \|K_0\|_q \|L_0\|_q^i / (1 - \|L_0\|_q). \tag{3.15}$$

Since $\|L_0\|_q < 1$, the corresponding dependence adjusted norm $\|x_{\cdot j}\|_{\iota,\alpha} < \infty$ for all $\alpha \ge 0$. □

Proof of Theorem 4. Write $\hat{\Sigma} = (\hat{\sigma}_{jk})_{1\le j,k\le p} = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$, $\Sigma = \Sigma_X = (\sigma_{jk})_{1\le j,k\le p}$ and define the event

$$A = \{|\hat{\Sigma} - \Sigma|_\infty \le a\} = \left\{\max_{j,k\le p} |\hat{\sigma}_{jk} - \sigma_{jk}| \le a\right\} \tag{3.16}$$

Write $\mathbf{x}_l = (x_{l1}, \ldots, x_{lp})^\top$ and $T_{j,n} = \sum_{l=1}^n x_{lj}e_l$. We now compute the functional dependence measure for the process $(x_{lj}e_l)_{l\in\mathbb{Z}}$ for fixed j. Similarly as e_i^* in (1.3), we define the coupled process $\mathbf{x}_i^*$. By Hölder's inequality, we have for $m \ge 0$ that

$$\begin{aligned}\sum_{l=m}^\infty \|x_{lj}e_l - x_{lj}^*e_l^*\|_\tau &\le \sum_{l=m}^\infty [\|x_{lj}(e_l - e_l^*)\|_\tau + \|(x_{lj} - x_{lj}^*)e_l^*\|_\tau] \\ &= \sum_{l=m}^\infty (\|x_{lj}\|_\iota \|e_l - e_l^*\|_q + \|x_{lj} - x_{lj}^*\|_\iota \|e_l^*\|_q).\end{aligned}$$

Since $\alpha = \min(\alpha_X, \alpha_e)$, the dependence-adjusted norm

$$\|x_{\cdot j}e_\cdot\|_{\tau,\alpha} \le \|x_{\cdot j}\|_\iota \|e_\cdot\|_{q,\alpha_e} + \|x_{\cdot j}\|_{\iota,\alpha_X} \|e_\cdot\|_q \le 2N_eN_X. \tag{3.17}$$

For the process $(x_{lj}x_{lk})_{l\in\mathbb{Z}}$, since $|x_{lj}x_{lk} - x_{lj}^*x_{lk}^*| \le |(x_{lj} - x_{lj}^*)x_{lk}| + |x_{lj}^*(x_{lk} - x_{lk}^*)|$, we similarly have

$$\|x_{\cdot j}x_{\cdot k} - \sigma_{jk}\|_{\iota/2,\alpha_X/2} \le 2N_X^2. \tag{3.18}$$

Let event $B = \{n^{-1}|X^\top e|_\infty \le b\}$. On the event $A \cap B$, since $b \ge \lambda/n$, we have

$$\begin{aligned}|\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_\infty &\le |(\Sigma - \hat{\Sigma})\hat{\boldsymbol{\beta}}|_\infty + |\hat{\Sigma}\hat{\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}|_\infty \\ &\le |\Sigma - \hat{\Sigma}|_\infty |\hat{\boldsymbol{\beta}}|_1 + \lambda/n + |n^{-1}X^\top Y - \Sigma\boldsymbol{\beta}|_\infty \\ &\le a|\boldsymbol{\beta}|_1 + \lambda/n + |\hat{\Sigma}\boldsymbol{\beta} - \Sigma\boldsymbol{\beta}|_\infty + n^{-1}|X^\top e|_\infty \\ &\le 2a|\boldsymbol{\beta}|_1 + 2b.\end{aligned} \tag{3.19}$$

Hence, by (3.17) and (3.18), (3.8) follows by applying Theorem 2 to $\mathbb{P}(A^c)$ and $\mathbb{P}(B^c)$, respectively, and the Bonferroni technique.

We now prove (ii). Let $Z_l = x_{lj}e_l$, $\iota = \tau(1 + \varrho/\nu)$ and $q = \tau(1 + \nu/\varrho)$. Let $\mu_\iota = \max_{j\le p} \|x_{lj}\|_\iota$ and $\kappa_q = \|e_l\|_q$. Then $\kappa_q \le \Delta_{0,q}$ and $\mu_\iota \le \Phi_{0,\iota}$ and

$$\sum_{l=0}^\infty \|x_{lj}e_l - x_{lj}^*e_l^*\|_\tau \le \sum_{l=0}^\infty (\|x_{lj}\|_\iota \|e_l - e_l^*\|_q + \|x_{lj} - x_{lj}^*\|_\iota \|e_l^*\|_q)$$

$$\leq \mu_\iota \Delta_{0,q} + \kappa_q \Phi_{0,\iota} \leq 2\Delta_{0,q}\Phi_{0,\iota}. \tag{3.20}$$

By (3.9) and the definition of q and ι, we have

$$\sup_{\tau\geq 2} \frac{\Delta_{0,q}\Phi_{0,\iota}}{\tau^{\nu+\varrho}} \leq K_\nu L_\varrho \sup_{\tau\geq 2} \frac{q^\nu \iota^\varrho}{\tau^{\nu+\varrho}} = K_\nu L_\varrho C_3, \tag{3.21}$$

where $C_3 = (1+\nu/\varrho)^\varrho(1+\varrho/\nu)^\nu$. Then by Theorem 3,

$$\mathbb{P}(B) \leq C_4 p \exp\{-C_5[\sqrt{n}b/(K_\nu L_\varrho)]^{2/(1+2\nu+2\varrho)}\}, \tag{3.22}$$

where constants C_4, C_5 only depend on ν and ϱ. Similarly, by (3.18), we have

$$\sup_{\iota\geq 2} \iota^{-2\varrho} \sum_{l=0}^{\infty} \|x_{lj}x_{lk} - x^*_{lj}x^*_{lk}\|_{\iota/2} \leq 2 \sup_{\iota\geq 2} \iota^{-2\varrho}\Phi^2_{0,\iota} = 2L^2_\varrho$$

which again by Theorem 3 implies $\mathbb{P}(A) \leq C_6 p^2 \exp\{-C_7(\sqrt{n}a/L^2_\varrho)^{2/(1+4\nu)}\}$. Hence (3.10) follows in view of the arguments in (i). □

Remark 5. The argument in the proof of Theorem 4 implies that, if $\lambda = \lambda_n$ is chosen such that $(pn^\pi)^{1/\tau} = o(\lambda_n)$ and $\lambda_n \geq C(n\log p)^{1/2}$ for a sufficiently large constant C, then for the true parameter $\boldsymbol{\beta}$, $|X^\top X\boldsymbol{\beta} - X^\top Y|_\infty \leq \lambda_n$ holds with probability going to 1. Namely, for event B with $b = \lambda/n$, $\mathbb{P}(B) \to 1$. □

Remark 6. If the two processes (e_l) and $(\mathbf{x}_l)$ are independent of each other, then we can let $\tau = \min(q, \iota)$. Additionally for the model (3.6) in Example 1, if $Z_i \in \mathcal{L}^q$ with $q > 2$, then $\iota = q$ and we can let $\tau = q$ instead of $q/2$, since $e_l x_{lj} \in \mathcal{L}^q$, and the functional dependence measure for $(e_l x_{lj})$ decays exponentially. □

The bound (3.8) reveals two different decay behaviors: if a is small, let $\chi = 1$, then the Gaussian-type bound $e^{-C_1 na^2}$ dominates. On the other hand, if it is large, then the dominating term is the polynomial tail $p^2 n/(na)^{\iota/2}$. A similar claim can be made for the term involving b. The borderline of this phase-transition phenomenon is at $a = a_n$ and $b = b_n$, where a_n (resp. b_n) is the solution to the equation $n(na)^{-\iota/2} = e^{-na^2}$ (resp. $n(nb)^{-\tau} = e^{-nb^2}$).

Theorem 4 immediately leads to the following result on the rate of convergence and support recovery.

Corollary 1. *Let* $\Omega_X = \Sigma_X^{-1}$ *and* $t_n = |||\Omega_X|||_1(a_n|\boldsymbol{\beta}|_1 + b_n)$, *where (i)*

$$a_n = (n^{-1}\log p)^{1/2} + p^{4/\iota} n^{2/\iota - 1} \text{ and } b_n = (n^{-1}\log p)^{1/2} + p^{1/\tau} n^{1/\tau - 1} \tag{3.23}$$

under Theorem 4(i) with $\alpha_X > 1/2 - 2/\iota$ *and* $\min(\alpha_X, \alpha_e) > 1/2 - 1/\tau$, *or (ii)*

$$a_n = \frac{(\log p)^{1/2+2\varrho}}{\sqrt{n}} \text{ and } b_n = \frac{(\log p)^{1/2+\varrho+\nu}}{\sqrt{n}} \tag{3.24}$$

under Theorem 4(ii). Then we have the convergence rate

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_\infty = O_{\mathbb{P}}(t_n). \tag{3.25}$$

In particular, if $\boldsymbol{\beta}$ is sparse such that $t_n = o(\min_{j:\beta_j\neq 0}|\beta_j|)$, then the support of $\boldsymbol{\beta}_0$ can be recovered by that of $\hat{\boldsymbol{\beta}}$ with probability going to 1.

Proof of Corollary 1. Since $\Omega_X\Sigma_X = \mathrm{Id}$, we have

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_\infty \le \||\Omega_X\||_1 |\Sigma_X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_\infty.$$

Then (3.25) follows from Theorem 4. □

The setting in our Theorem 4 and Corollary 1 is very general as it allows dependent and/or non sub-Gaussian error processes and it also allows heteroscedasticity in that the error process and the covariance process can be dependent. Han and Liu [16] considered the special case of the estimation of $A_1, \ldots, A_d$ of model (3.5) under the assumption that Z_i are i.i.d. Gaussian. Sims [34] mentioned several challenges in the inference of vector autoregressive models: the possibility of fat tails in the innovations and the low degrees of freedom due to the estimation of possibly extremely many parameters. The latter problem has been widely recognized in the analysis of vector autoregressive processes; see for example [17, 20, 2] among others. Our setting allows both fat tails and large number of parameters. Additionally, by checking non-zero entries in the estimate $\hat{\boldsymbol{\beta}}$, we can infer economic relations between variables, a theory-free principle that was advocated in [34].

4. LASSO estimators with deterministic design

Following [37], one can estimate the unknown parameter vector $\boldsymbol{\beta}$ by minimizing the criterion function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \frac{1}{n}|y - X\boldsymbol{\beta}|_2^2 + \lambda|\boldsymbol{\beta}|_1, \tag{4.1}$$

where $\lambda > 0$ is the regularization parameter. In this section we assume that $\mathbf{x}_i$ is deterministic and (e_i) is of the form (1.2). For convenience, we scale the diagonal entries of the Gram matrix $\Psi_n = X^\top X/n$ to be 1. Then $|X|_2 = (np)^{1/2}$. Let $\hat{\boldsymbol{\beta}}$ be the minimizer of (4.1). Consistency properties of $\hat{\boldsymbol{\beta}}$ are discussed in [4, 5], where e_i are i.i.d. and sub-Gaussian.

4.1. Convergence rate of the Lasso estimator

Equipped with the probability inequalities established in Section 2, we shall study properties of the Lasso estimator $\hat{\boldsymbol{\beta}}$ in (4.1), in particular the rate of convergence of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. Let X be the design matrix and write (1.1) as $Y = X\boldsymbol{\beta} + e$. We assume that the true parameter vector $\boldsymbol{\beta}$ has at most s non-zero entries. We shall also assume that the restricted eigenvalue assumption $\mathrm{RE}(s, 3)$ in [4] holds with constant $\kappa = \kappa(s, 3)$, namely

$$\kappa(s, c_0) := \min_{J\subseteq\{1,\ldots,p\},\,|J|\le s}\ \min_{u\neq 0,\,|u_{J^c}|_1\le c_0|u_J|_1} \frac{|Xu|_2}{\sqrt{n}|u_J|_2} > 0 \tag{4.2}$$

holds with $c_0 = 3$, where u_J is defined as a modification of u by setting its elements outside J to zero. This condition is weaker than the restricted isometry property of [9]. See also [6, 28, 48, 49] for related conditions.

Theorem 5 shows how the rate of convergence of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1$ and the prediction error $|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_2^2$ depend on q and α, which correspond to the moment condition and the dependence condition respectively. Let the regularization parameter $\lambda = 2r$.

Theorem 5. *Assume (4.2). Assume that the error sequence (e_i) has finite qth moment, $q > 2$, and dependence-adjusted norm $\|e.\|_{q,\alpha} < \infty$, $\alpha \geq 0$. (i) Assume $\alpha > 1/2 - 1/q$. Let*

$$r = \max(A(n^{-1}\log p)^{1/2}\|e.\|_{2,\alpha},\ B\|e.\|_{q,\alpha}|X|_q/n). \tag{4.3}$$

Then with probability at least $1 - C_1 B^{-q} - C_2 p^{1-C_3 A^2}$, we have

$$|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_2^2/n \leq 16sr^2/\kappa^2, \tag{4.4}$$

and

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \leq 16sr/\kappa^2. \tag{4.5}$$

(ii) Assume $\alpha < 1/2 - 1/q$. Let

$$r = \max(A(n^{-1}\log p)^{1/2}\|e.\|_{2,\alpha},\ Bn^{-1/2-1/q-\alpha}|X|_q\|e.\|_{q,\alpha}), \tag{4.6}$$

then (4.4) and (4.5) hold with the same probability as in (i).

Proof. As in the proof of Lemma B.1 in [4], since $\hat{\boldsymbol{\beta}}$ minimizes (4.1), we have

$$2r|\hat{\boldsymbol{\beta}}|_1 + |X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_2^2/n \leq 2r|\boldsymbol{\beta}|_1 + 2e^\top X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/n. \tag{4.7}$$

Let $\delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. On the event

$$\mathcal{A} = \bigcap_{j=1}^{p}\{2|V_j| \leq r\},\ \text{where } V_j = \frac{1}{n}\sum_{i=1}^{n} e_i x_{ij}, \tag{4.8}$$

inequality (4.7) implies that

$$r|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 + |X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_2^2/n \leq 4r|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J|_1 \leq 4r\sqrt{s}|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J|_2. \tag{4.9}$$

Hence $|\delta_{J^c}|_1 \leq 3|\delta_J|_1$, which by (4.2) entails $|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_2^2/n \geq \kappa^2|\delta_J|_2^2$. Then

$$\mathbb{P}[|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|_2^2/n \leq 16sr^2/\kappa^2] + \mathbb{P}[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \leq 16sr/\kappa^2] \leq \mathbb{P}(\mathcal{A}^c). \tag{4.10}$$

So (4.4) and (4.5) follow if we can control the probability $\mathbb{P}(\mathcal{A})$. For (i), by Inequality (2.9) of Theorem 2 with $\varpi_n = 1$, we have

$$\mathbb{P}(|V_j| \geq r) \leq C_1\|e.\|_{q,\alpha}^q \frac{\sum_{i=1}^n |x_{ij}|^q}{(nr)^q} + C_2\exp(-C_3 nr^2/\|e.\|_{2,\alpha}^2). \tag{4.11}$$

Hence

$$\mathbb{P}(\mathcal{A}^c) \le \sum_{j=1}^{p} C_1 \|e.\|_{q,\alpha}^q \frac{\sum_{i=1}^n |x_{ij}|^q}{(nr)^q} + C_2 p \exp(-C_3 nr^2/\|e.\|_{2,\alpha}^2)$$
$$= C_1 \|e.\|_{q,\alpha}^q \frac{|X|_q^q}{(nr)^q} + C_2 p \exp(-C_3 nr^2/\|e.\|_{2,\alpha}^2). \tag{4.12}$$

Under our choice of r, we have

$$\mathbb{P}(\mathcal{A}^c) \le C_1 B^{-q} + C_2 p^{1-C_3 A^2}. \tag{4.13}$$

Following the argument in [4], with probability at least $1 - C_1 B^{-q} - C_2 p^{1-C_3 A^2}$, we have (4.4) and (4.5). Case (ii) can be similarly proved. □

Theorem 5 indicates how the dimension breaks down if the moment condition weakens or the dependence becomes stronger. Assume that $|X|_q \asymp (np)^{1/q}$ and $\|e.\|_{q,\alpha} < \infty$, $\alpha > 1/2 - 1/q$, then by (4.3), the requirement $r \to 0$ implies necessarily that $(np)^{1/q}/n \to 0$, or $p = o(n^{q-1})$. In comparison, if e_i are i.i.d. sub-Gaussian, then the condition $\log p = o(n)$ suffices. In the stronger dependence case (ii), the more restrictive condition $p = o(n^{q\alpha+q/2})$ is needed. The latter range on p is substantially narrower.

It is interesting to compare the two terms in r. Assume that $|X|_q \asymp (np)^{1/q}$. In the relatively low dimensional case with $p \le n^{q/2-1}(\log n)^{q/2}$, the Gaussian part, which corresponds to $(n^{-1}\log p)^{1/2}$, is larger. On the other hand, if the dimension p is large such that $p > n^{q/2-1}(\log n)^{q/2}$, then the tail part $n^{-1}|X|_q \asymp n^{1/q-1}p^{1/q}$ dominates and it is larger than the penalty of the form $A(n^{-1}\log p)^{1/2}$ that is used in the Gaussian errors case.

For sub-Gaussian errors, we have the following theorem.

Theorem 6. *Assume that the error sequence (e_i) satisfies (2.21). Let*

$$r = An^{-1/2}(\log p)^{1/\alpha}\|e.\|_{q,\alpha}. \tag{4.14}$$

Then with probability at least $1 - C_1 p^{1-C_2 A^\alpha}$, we have bounds (4.4) and (4.5).

The proof of this theorem is similar to the corresponding result in [4], and is omitted.

Example 5 (Nonparametric trend estimation)**.** Consider the model

$$y_i = \mu(i/n) + e_i, \quad 1 \le i \le n, \tag{4.15}$$

where $\mu(\cdot)$ is a trend function, and e_i are stationary noises. Let $f_1(\cdot), f_2(\cdot), \ldots$ be basis functions on $L[0,1]$. We approximate $\mu(\cdot)$ by $\mu_{\boldsymbol{\beta}}(u) = \sum_{j=1}^p \beta_j f_j(u)$ and the coefficients $\beta_1, \ldots, \beta_p$ are estimated by (4.1). With the estimated $\hat\beta_1, \ldots, \hat\beta_p$, let $\mu_{\hat{\boldsymbol{\beta}}}(u) = \sum_{j=1}^p \hat\beta_j f_j(u)$. Assume that $\|f_j\|_n^2 := n^{-1}\sum_{i=1}^n f_j^2(i/n) = 1$ for all $j \le p$. Applying the arguments of Theorem 6.1 in [4] with $\varepsilon = 4$ therein, we have

for probability at least $1 - C_1 B^{-q} - C_2 p^{1-C_3 A^2}$ (resp. $1 - C_1 p^{1-C_2 A^\alpha}$) under settings in Theorem 5 (resp. Theorem 6) that

$$\|\mu_{\hat{\boldsymbol{\beta}}} - \mu\|_n^2 \le \inf_{\boldsymbol{\beta}\in\mathbb{R}^p, |\boldsymbol{\beta}|_0 \le s} \left[5\|\mu_{\boldsymbol{\beta}} - \mu\|_n^2 + \frac{36|\boldsymbol{\beta}|_0 r^2}{\kappa^2(s,4)} \right], \tag{4.16}$$

where $|\boldsymbol{\beta}|_0 = \#\{j \le p : \beta_j \ne 0\}$. □

4.2. Model selection consistency

In this section we shall consider sign consistency for model selection based on the Lasso. Zhao and Yu [46] introduced the concept of sign consistency; see also [40, 27]. We write $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s, \beta_{1+s}, \ldots, \beta_p)^\top$, where $\beta_i \ne 0$ if $i \le s$ and $\beta_i = 0$ if $i > s$. Correspondingly we write $X = (X_1, X_2)$, where X_1 is the $n \times s$ sub-matrix that corresponds to the predictors with non-zero coefficients, and X_2 is the remaining $n \times (p-s)$ sub-matrix. We scale the diagonal entries of the Gram matrix $\Psi_n = X^\top X/n$ to be 1. We have the following Theorem 7 which extends the result in [46] to models with dependent errors. Note that, even in the independence case, our bound is sharper. We use the same conditions as those in [46]. The quantity $\eta \in (0,1)$ in Theorem 7 is from the strong irrepresentable condition in [46]. Namely

$$|X_2^\top X_1 (X_1^\top X_1)^{-1} \mathrm{sign}(\boldsymbol{\beta}_{1:s})|_\infty \le 1 - \eta, \text{ where } \boldsymbol{\beta}_{1:s} = (\beta_1, \ldots, \beta_s)^\top.$$

Here the sign function $\mathrm{sign}(u) = 1$ if $u > 0$, -1 if $u < 0$ and 0 if $u = 0$.

Theorem 7. *Let $M_2 = 1/|||n(X_1^\top X_1)^{-1}|||_2$ and $L = \min_{i\le s} |\boldsymbol{\beta}_i|$. Assume that $\lambda \le nM_2 L/\sqrt{s}$. (i) (Polynomial Tail Bound) Assume that (e_i) satisfies $\|e.\|_{q,\alpha} < \infty$, $\alpha > 1/2 - 1/q$. Then the sign consistency probability $\mathbb{P}(\hat{\boldsymbol{\beta}} =_s \boldsymbol{\beta})$ is at least*

$$\begin{aligned} 1 - \Bigg[& \frac{C_1 |H_1|_q^q \|e.\|_{q,\alpha}^q}{(\sqrt{n}L)^q} + C_2 s \exp\left(-C_3 n L^2 M_2 / \|e.\|_{2,\alpha}^2\right) \\ & + \frac{C_4 |H_2|_q^q \|e.\|_{q,\alpha}^q}{(\eta\lambda/\sqrt{n})^q} + C_5 p \exp\left(-C_6 \eta^2 \lambda^2 / (n\|e.\|_{2,\alpha}^2)\right) \Bigg], \end{aligned} \tag{4.17}$$

where the matrices $H_1 = \sqrt{n}(X_1^\top X_1)^{-1} X_1$ and $H_2 = n^{-1/2} X_2^\top [X_1 (X_1^\top X_1)^{-1} \times X_1^\top - I_n]$ and constants $C_1, \ldots, C_6$ only depend on α and q. Assume that

$$|H_1|_q \|e.\|_{q,\alpha} + \frac{\|e.\|_{2,\alpha}\sqrt{\log s}}{\sqrt{M_2}} + \sqrt{s}\frac{\|e.\|_{2,\alpha}\sqrt{\log p} + |H_2|_q \|e.\|_{q,\alpha}}{M_2 \eta} = o(\sqrt{n}L) \tag{4.18}$$

and choose $\lambda \le nM_2 L/\sqrt{s}$ such that $\|e.\|_{2,\alpha}\sqrt{\log p} + |H_2|_q \|e.\|_{q,\alpha} = o(\lambda\eta/\sqrt{n})$. Then the quantity (4.17) converges to 1. (ii) Assume (2.21). Then

$$\mathbb{P}(\hat{\boldsymbol{\beta}} \ne_s \boldsymbol{\beta}) \le C_1 s \exp(-C_2 (\frac{\sqrt{n}L\sqrt{M_2}}{\gamma_0})^\alpha) + C_3 p \exp(-C_4 (\frac{\eta\lambda}{\sqrt{n}\gamma_0})^\alpha), \tag{4.19}$$

where $\gamma_0 = \|e.\|_{\psi_\nu}$, and constants $C_1, \ldots, C_4 > 0$ only depend on α.

Proof. (i) Write the matrices $H_1 = (a_{ij})_{i\le s,\, j\le n}$ and $H_2 = (b_{ij})_{i\le p-s,\, j\le n}$. Let $(b_1,\ldots,b_s)^\top = n(X_1^\top X_1)^{-1}\mathrm{sign}(\boldsymbol{\beta}_{1:s})$, $(Z_1,\ldots,Z_s)^\top = H_1 e$ and $(\zeta_1,\ldots,\zeta_{p-s})^\top = H_2 e$. By Proposition 1 in [46], we have

$$\begin{aligned}\mathbb{P}(\hat{\boldsymbol{\beta}} \neq_s \boldsymbol{\beta}) \le{}& \mathbb{P}(\cup_{1\le i\le s}\{|Z_i| \ge \sqrt{n}(|\beta_i| - \lambda|b_i|/(2n))\}) \\ &+ \mathbb{P}(\cup_{1\le i\le p-s}\{|\zeta_i| \ge \eta\lambda/(2\sqrt{n})\}). \end{aligned} \tag{4.20}$$

For $i \le s$, note that $|b_i| \le \sqrt{s}/M_2$, by the condition $\lambda\sqrt{s} \le nM_2L$, we have $|\beta_i| - \lambda|b_i|/(2n) \ge |\beta_i|/2 \ge L/2$. Then $\mathbb{P}(|Z_i| \ge \sqrt{n}(|\beta_i| - \lambda|b_i|/(2n))) \le \mathbb{P}(|Z_i| \ge L/2)$. Also note that $\sum_{j=1}^n a_{ij}^2 \le 1/M_2$ for all $i \le s$ and $\sum_{j=1}^n b_{ij}^2 \le 1$ for all $i \le p-s$. Then (4.17) follows by applying the Nagaev inequality (2.9) of Theorem 2 to $\mathbb{P}(|Z_i| \ge z)$ and $\mathbb{P}(|\zeta_i| \ge z)$ via the sub-additivity of probability measures.

Under (4.18), we have $\|e.\|_{2,\alpha}\sqrt{\log p} + |H_2|_q\|e.\|_{q,\alpha} = o(\sqrt{n}M_2L\eta/\sqrt{s})$. Hence there exists λ such that $\lambda \le nM_2L/\sqrt{s}$ and $\|e.\|_{2,\alpha}\sqrt{\log p} + |H_2|_q\|e.\|_{q,\alpha} = o(\lambda\eta/\sqrt{n})$. For such λ, it is easily seen that the quantity in (4.17) converges to 1 under (4.18).

(ii) By Theorem 3 and the arguments in (i), (4.19) follows from (4.20). □

If $|H_1|_q^q \asymp nsn^{-q/2}$ and $|H_2|_q^q \asymp n(p-s)n^{-q/2}$, which hold if a_{ij} and b_{ij} are typically of order $n^{-1/2}$, $M_2 \asymp 1$ and $\eta \asymp 1$, then (4.18) reduces to $\sqrt{s}\sqrt{\log p} + \sqrt{s}n^{1/q-1/2}p^{1/q} = o(\sqrt{n}L)$. If additionally $s = O(n^{c_1})$ and $L \asymp n^{(c_2-1)/2}$ for some $0 \le c_1 \le c_2 \le 1$, then by Theorem 7 the valid regularization parameter λ has the range $(pn)^{1/q} \ll \lambda \ll n^{(c_2-c_1+1)/2}$. In other words, existence of such λ requires the dimension $p \ll n^{q(c_2-c_1+1)/2-1}$. In comparison, [46] has a narrower range $p \ll n^{q(c_2-c_1)/2}$ since $q(c_2-c_1+1)/2 - 1 > q(c_2-c_1)/2$ as $q > 2$. Their range is invalid if $c_1 = c_2$.

In the special case in which e_i are i.i.d., a slightly improved version of Theorem 7 can be obtained. Let $\mu_q = \|e_i\|_q$, $\Gamma_1 = (\sum_{j=1}^n \max_{i\le s}|a_{ij}|^q)^{1/q}$ and $\Gamma_2 = (\sum_{j=1}^n \max_{i\le p-s}|b_{ij}|^q)^{1/q}$. Note that $\Gamma_1 \le |H_1|_q$ and $\Gamma_2 \le |H_2|_q$.

Theorem 8. *Let $M_2 = 1/\|n(X_1^\top X_1)^{-1}\|_2$ and $L = \min_{i\le s}|\boldsymbol{\beta}_i|$. Assume that*

$$\sqrt{n}L \ge K_*(M_2^{-1/2}\mu_2\sqrt{\log s} + \Gamma_1\mu_q \log s), \tag{4.21}$$

$$\eta\lambda/\sqrt{n} \ge K_*(\mu_2\sqrt{\log p} + \Gamma_2\mu_q \log p), \tag{4.22}$$

where K_* is an absolute constant, and that $\lambda \le nM_2L/\sqrt{s}$. Then the sign consistency probability $\mathbb{P}(\hat{\boldsymbol{\beta}} =_s \boldsymbol{\beta})$ is at least

$$\begin{aligned} 1 - \Bigg[&\exp(-nL^2K_1M_2/\mu_2^2) + \frac{K_q\Gamma_1^q\mu_q^q}{(\sqrt{n}L)^q} \\ &+ \exp(-K_2\eta^2\lambda^2/(n\mu_2^2)) + \frac{K_q\Gamma_2^q\mu_q^q}{(\eta\lambda n^{-1/2})^q}\Bigg], \end{aligned} \tag{4.23}$$

where K_1, K_2 are absolute constants and K_q is a constant only depending on q.

Proof. By Lemma A.3 in [10], there exists an absolute constant K such that

$$\mathbb{E}[\max_{i\le s}|Z_i|] \le K[(\max_{i\le s}\sum_{j=1}^{n} a_{ij}^2)^{1/2}\mu_2\sqrt{\log s} + \|\max_{i\le s}\max_{j\le n}|a_{ij}e_j|\|_2 \log s]$$
$$\le K(M_2^{-1/2}\mu_2\sqrt{\log s} + \Gamma_1\mu_q \log s). \tag{4.24}$$

Choose K_* in (4.21) to be $16(K+1)$. Then $\sqrt{n}L/2 \ge 2\mathbb{E}[\max_{i\le s}|Z_i|] + \sqrt{n}L/4$. By Lemma A.2 in [10], there exists an absolute constant $K_1 > 0$ and a constant K_q only depending on q such that

$$\mathbb{P}(\cup_{1\le i\le s}\{|Z_i| \ge \frac{\sqrt{n}L}{2}\}) \le \exp(-n\frac{L^2K_1M_2}{\mu_2^2}) + \frac{K_q\sum_{j=1}^{n}\|\max_{i\le s}|a_{ij}e_j|\|_q^q}{(\sqrt{n}L)^q}$$
$$= \exp(-nL^2K_1M_2/\mu_2^2) + \frac{K_q}{(\sqrt{n}L)^q}\Gamma_1^q\mu_q^q. \tag{4.25}$$

To deal with $\max_{i\le p-s}|\zeta_i|$, we have similarly the moment inequality

$$\mathbb{E}[\max_{i\le p-s}|\zeta_i|] \le K[(\max_{i\le p-s}\sum_{j=1}^{n} b_{ij}^2)^{1/2}\mu_2\sqrt{\log p} + \|\max_{i\le p-s}\max_{j\le n}|b_{ij}e_j|\|_2 \log p]$$
$$\le K(\mu_2\sqrt{\log p} + \Gamma_2\mu_q \log p). \tag{4.26}$$

Thus a similar version of (4.25) holds for $\mathbb{P}(\max_{i\le p-s}|\zeta_i| \ge \eta\lambda/(2\sqrt{n}))$, which implies (4.23) by (4.20). □

5. A simulation study

This section presents a simulation study to illustrate the effects of heavy tails and dependencies of the error and/or the covariate processes on the performance of the Clime and the Lasso estimators, with stochastic and deterministic designs, respectively. For the former, we consider model (1.1) with the residual process

$$e_i = (1-\rho^2)^{1/2}(1-2/\kappa)^{1/2}e_i^\circ, \quad e_i^\circ = \rho e_{i-1}^\circ + \eta_i, \tag{5.1}$$

where η_i are i.i.d. Student t_κ random variables with degrees of freedom $\kappa > 2$ and $-1 < \rho < 1$. Note that the parameters ρ and κ controls the dependence and the heaviness of the tails of (e_i), respectively. Observe that the e_i has mean 0 and variance 1. For the regressor process $\mathbf{x}_i$, we let $\mathbf{x}_i = \Sigma^{-1/2}\mathbf{x}_i^\circ$,

$$\mathbf{x}_i^\circ = A\mathbf{x}_{i-1}^\circ + \varepsilon_i = A\mathbf{x}_{i-1}^\circ + (\varepsilon_{i1},\ldots,\varepsilon_{ip})^\top, \tag{5.2}$$

where ε_{ij} are i.i.d. t_ν random variables with degrees of freedom ν, and $\Sigma = \mathrm{cov}(\mathbf{x}_i^\circ)$. Then the covariance matrix of $\mathbf{x}_i$ is identity. For the coefficient matrix A, we let $A = (a_{jk})_{j,k\le p}/\sqrt{4p}$, where a_{jk} are i.i.d. $N(0,1)$ variables. We choose a realization of A such that its spectral norm $\|A\|_2 < 1$, so that (5.2) has a stationary solution. Once A is simulated, we keep it throughout the simulation.

We choose $n = 25$, $p = 100$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ with $\beta_j = 1$ if $j \le 5$ and $\beta_j = 0$ if $j > 5$. In our simulation we use the R `flare` package by [23] to compute the Clime estimate. To study how the dependence and the heavy tails affect the convergence speed, we consider the tail probability ratio function

$$R_1(t) = \frac{\mathbb{P}(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \ge t)}{\mathbb{P}(|\hat{\boldsymbol{\beta}}^\dagger - \boldsymbol{\beta}|_1 \ge t)}, \tag{5.3}$$

where $\hat{\boldsymbol{\beta}}^\dagger$ is the Clime estimator of $\boldsymbol{\beta}$ in model (1.1) with $\mathbf{x}_i$ being i.i.d. $\mathbb{R}^p$ standard normal random vectors, and e_i are i.i.d. $\mathrm{N}(0,1)$ random variables, and $\hat{\boldsymbol{\beta}}$ is the Clime estimator with error process (5.1) and regressor process (5.2) with different dependence and tail conditions. The denominator in (5.3) can be viewed as benchmark probabilities. The `flare` program suggests that the threshold value λ is around 0.6. Hence in our simulation we use $\lambda = 0.6$. In the benchmark setting, based on 10^6 repetitions, the 99% and 99.9% quantiles of the $\mathcal{L}_1$ distance $|\hat{\boldsymbol{\beta}}^\dagger - \boldsymbol{\beta}|_1$ are estimated as 11.781 and 12.495, respectively.

Table 1 presents the simulated values of the tail probability ratio function $R_1(t)$ with $t = 11.781$ and $t = 12.495$, which correspond to the ratio between $\mathbb{P}(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \ge t)$ under various dependence and moment conditions and the benchmark tail probabilities 0.01 and 0.001, respectively. For each different combinations of (ρ, ν, κ), the tail probability $\mathbb{P}(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \ge t)$ is estimated by the proportions of the 5000 values of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1$ that are larger than t. Table 1 suggests the following phenomena, as expected from our theoretical results: (i) heavier tails (smaller ν or κ) can lead to larger $\mathbb{P}(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \ge t)$, thus inflating the tail probability ratio R_1; (ii) the upper tail probability $\mathbb{P}(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \ge t)$ with larger t is affected more than the one with smaller t. For example, with $(\rho, \nu, \kappa) = (-.75, 3, 3)$, the latter probability can be $R_1(t) = 61.8$ times larger than the nominal level 0.001, as obtained based on i.i.d. standard normal distributions.

For Lasso estimation with fixed design, we shall also focus on the tail behavior of the ℓ_1 error of the estimated parameters. We set $p = 800$, and $n = 100, 200$, and 400. In each of the three (n, p) combinations, we generate the $n \times p$ design

TABLE 1

Simulated values of the tail probability ratio function $R_1(t)$. Left (resp. right) panel: $t = 11.781$ (resp. $t = 12.495$) is the 99% (resp. 99.9%) quantile of $|\hat{\boldsymbol{\beta}}^\dagger - \boldsymbol{\beta}|_1$

ρ	ν	$\kappa = 3$	$\kappa = 12$
-.75	3	10.9	10.5
-.5	3	10.2	10.0
0	3	10.0	8.1
.5	3	9.1	8.3
.75	3	7.4	6.5
-.75	12	4.3	3.2
-.5	12	3.9	2.2
0	12	4.0	1.7
.5	12	2.8	1.3
.75	12	2.3	1.3

ρ	ν	$\kappa = 3$	$\kappa = 12$
-.75	3	61.8	49.4
-.5	3	57.6	40.8
0	3	49.2	33.2
.5	3	47.8	41.0
.75	3	40.6	28.0
-.75	12	22.6	9.4
-.5	12	21.6	5.2
0	12	21.6	2.6
.5	12	15.0	3.0
.75	12	13.2	3.2

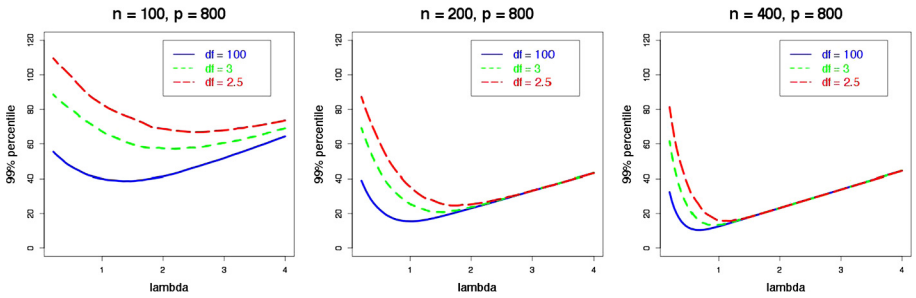

FIG 1. *The* 99% *quantiles of the* ℓ_1 *error* $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1$ *with i.i.d. errors.*

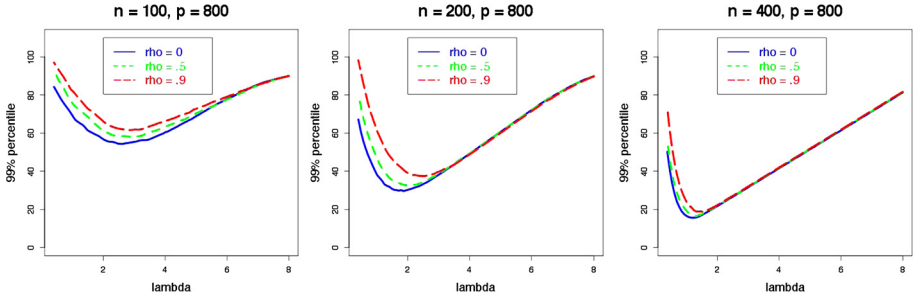

FIG 2. *The* 99% *quantiles of* $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1$ *with errors from a non-linear autoregressive process.*

matrix X whose entries x_{ij} are iid $\mathrm{N}(0,1)$ random variables. We fix the design matrix for all the 3,000 repetitions of the simulation study. For the true value of the parameter vector $\boldsymbol{\beta}$, we let the first s elements be non-zero, and the rest of the elements be zero. We set $s = 10$. For the non-zero elements, with $1/2$ probability, $\boldsymbol{\beta}_j = b$, and with $1/2$ probability, $\boldsymbol{\beta}_j = -b$. We set $b = 10$. We fix the marginal variance of e_i at $\mathrm{Var}(e_i) = 5^2$. We use the R `glmnet` package for the Lasso computation [14]. We adopt the option that the intercept is set to 0.

We first experiment with independent errors, where e_i are i.i.d. student t_ν, which has a polynomial tail. Note that $\mathrm{Var}(e_i) = \nu/(\nu - 2)$, which we use to normalize the variance of e_i. We examine the performance of the Lasso over a range of values of the regularization parameter λ. For each value of λ, we compute the 99% quantile of the ℓ_1 error $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1$. The quantiles are estimated from 3,000 repetitions. Figure 1 shows the results for $\nu = 100, 3$ and 2.5, for different values of n ($n = 100$, 200, 400 respectively). t_{100} is close to normal. It can be seen that the tail of the ℓ_1 error becomes heavier as ν decreases.

We then continue to study the behavior of the Lasso for dependent and heavy tailed errors. We consider a non-linear autoregressive model. We first generate a Gaussian autoregressive process $\tilde{e}_i = \rho \tilde{e}_{i-1} + \sqrt{1-\rho^2}\epsilon_i$, where ϵ_i are i.i.d. $\mathrm{N}(0,1)$. So marginally $\tilde{e}_i \sim \mathrm{N}(0,1)$. We then let $e_i = g(\tilde{e}_i)$. The non-linear transformation $g(x) = F^{-1}(\Phi(z))$ transforms $\tilde{e}_i$ into a random variable e_i that follows t_ν with $\nu = 2.5$, where F is the cdf of t_ν, and Φ is the standard normal cdf. We then normalize the marginal variance so that $\mathrm{Var}(e_i) = 5^2$. Figure 2

shows the results for $\rho = 0, .5, .9$. As the autocorrelation becomes stronger or the sample size n gets smaller, the tail of the ℓ_1 error becomes heavier.

Our simulation studies show that the performance of the Lasso deteriorates if the errors have a heavy tail distribution and if there are dependencies among the errors. This is qualitatively consistent with the theoretical results we have obtained, although it is difficult for the simulation studies to capture the rate of the tail decay quantitatively.

Acknowledgments

We thank two anonymous referees, an Associate Editor and the Editor for their helpful comments that have improved the paper.

References

[1] Azuma, K. (1967) Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, **19**, 357–367. MR0221571

[2] Barnett, W. A., Chae, U. and Keating, J. (2012) Forecast design in monetary capital stock measurement. *Global Journal of Economics*, **1**, 1250005.

[3] Basu, S. and Michailidis, G. (2015) Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, **43**, 1535–1567. MR3357870

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

[5] Bühlmann, P. and van de Geer, S. (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer. MR2807761

[6] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007) Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, **1**, 169–194. MR2312149

[7] Burkholder, D. L. (1973) Distribution function inequalities for martingales. *Annals of Probability*, **1**, 19–42. MR0365692

[8] Cai, T., Liu, W. and Luo, X. (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of American Statistical Association*, **106**, 594–607. MR2847973

[9] Candes E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, **35**, 2313–2404. MR2382644

[10] Chernozhukov, V., Chetverikov, D. and Kato, K. (2014) Testing many moment inequalities. http://arxiv.org/abs/1312.7614

[11] Chow, Y. S. and Teicher, H. (1988). *Probability Theory*, 2nd ed. Springer, New York. MR0953964

[12] Davis, R. A., Zang, P., and Zheng, T. (2015) Sparse Vector Autoregressive Modeling. *Journal of Computational and Graphical Statistics*.
[13] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, **96**, 1348–1360. MR1946581
[14] Friedman, J. H., Hastie, T., and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.
[15] Gupta, S. (2012) A note on the asymptotic distribution of LASSO estimator for correlated data. *Sankhya A*, **74**, 10–28. MR3010290
[16] Han F. and Liu, H. (2013) A direct estimation of high dimensional stationary vector autoregressions. http://arxiv.org/abs/1307.0293
[17] Hsiao, C. (1979) Autoregressive modeling of canadian money and income data. *Journal of the American Statistical Association*, **74**, 553–560.
[18] Kaul, A. (2014) Lasso with long memory regression errors. *Journal of Statistical Planning and Inference*, **153**, 11–26. MR3229019
[19] Kock, A. and Callot, L. (2012) Oracle inequalities for high dimensional vector autoregressions, Research Paper 12, CREATES, Aarhus University.
[20] Krolzig, H. M. (2003) General-to-specific model selection procedures for structural vector autoregressions. *Oxford Bulletin of Economics and Statistics*, **65**, 769–801.
[21] Lesigne, E. and Volný, D. Large deviations for martingales. *Stochastic Processes and their Applications*, **96**, 143–159, 2001. MR1856684
[22] Liu, H. and Wang, L. (2012) TIGER: A Tuning-Insensitive Approach for Optimally Estimating Gaussian Graphical Models. http://arxiv.org/abs/1209.2437
[23] Liu, W., Xiao, H., and Wu, W. B. (2013) Probability and moment inequalities under dependence. *Statistica Sinica*, **23**, 1257–1272. doi:10.5705/ss.2011.287. MR3114713
[24] Loh, P.-L. (2015) Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. http://arxiv.org/abs/1501.00312
[25] Loh, P.-L. and Wainwright, M. J. (2012), High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Stat.*, **40**, 1637–1664. MR3015038
[26] Lütkepohl, H. (2005), New introduction to multiple time series analysis, Springer. MR2172368
[27] Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, **34**, 1436–1462. MR2278363
[28] Meinshausen, N. and Yu, B. (2009) Lasso–type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, **37**, 246–270. MR2488351
[29] Merlevede, F., Peligrad, M., and Rio, E. (2011) A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, **141**, 435–474. MR2851689

[30] Nagaev, S. V. (1979). Large deviations of sums of independent random variables. *Annals of Probability*, **7**, 745–789. MR0542129

[31] Nardi, Y. and Rinaldo, A. (2011) Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, **102**, 528–549. MR2755014

[32] Peligrad, M., Sang, H., Zhong, Y., and Wu, W. B. (2014) Exact moderate and large deviations for linear processes, *Statistica Sinica*, **24**, 957–1969. MR3235407

[33] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011) High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, **5**, 935–980. MR2836766

[34] Sims, C. (1980) Macroeconomics and reality. *Econometrica*, **48**, 1–48.

[35] Rosenthal, H. P. (1970) On the subspaces of L_p $(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, **8**, 273–303. MR0271721

[36] Song, S. and Bickel, P. J. (2011) Large vector autoregressions. http://arxiv.org/abs/1106.3915

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

[38] Tong, H. (1990) *Nonlinear time series analysis: A dynamic approach*, Oxford University Press, Oxford. MR1079320

[39] Vershynin, R. (2012) *Introduction to the non-asymptotic analysis of random matrices.* In Chapter 5 of: *Compressed Sensing, Theory and Applications.* pp. 210–268. Edited by Y. Eldar and G. Kutyniok. Cambridge University Press. MR2963170

[40] Wainwright, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, **55**, 2183–2202. MR2729873

[41] Wang, H., Li, G., and Tsai, C.-L. (2007) Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of Royal Statistical Society, B*, **69**, 63–78. MR2301500

[42] Wiener, N. (1958) *Nonlinear Problems in Random Theory.* MIT Press. MR0100912

[43] Wu, W. B. (2005) Nonlinear system theory: another look at dependence. *Proceedings of National Academy of Science*, **102**, 14150–14154. MR2172215

[44] Wu, W. B. (2011) Asymptotic theory for stationary processes, *Statistics and Its Interface*, **4**, 207–226. MR2812816

[45] Wu, W. B. and Shao, X. (2004) Limit theorems for iterated random functions. *Journal of Applied Probability*, **41**, 425–436. MR2052582

[46] Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541–2567. MR2274449

[47] Zhang, C. H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942. MR2604701

[48] Zhang, C. H. and Huang, J. (2008) The sparsity and bias of the Lasso selection in high dimensional linear regression. *Annals of Statistics*, **36**, 1567–1594. MR2435448

[49] Zhang, T. (2009) Some sharp performance bounds for least squares regression with l1 regularization. *Annals of Statistics*, **37**, 2109–2144. MR2543687