Statistica Sinica Preprint No: SS-2022-0249	
Title	A Bernstein-type Inequality for High Dimensional Linear Processes with Applications to Robust Estimation of Time Series Regressions
Manuscript ID	SS-2022-0249
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0249
Complete List of Authors	Linbo Liu and Danna Zhang
Corresponding Authors	Danna Zhang
E-mails	daz076@ucsd.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

A Bernstein-type Inequality for High Dimensional Linear Processes with Applications to Robust Estimation of Time Series Regressions

Linbo Liu and Danna Zhang

Department of Mathematics, University of California, San Diego

Abstract: Time series regression models are commonly used in time series analysis. However, in modern real-world applications, serially correlated data with an ultra-high dimension and fat tails are prevalent. This presents a challenge in developing new statistical tools for time series analysis. In this paper, we propose a novel Bernstein-type inequality for high-dimensional linear processes and apply it to investigate two high-dimensional robust estimation problems: (1) time series regression with fat-tailed and correlated covariates and errors, and (2) fat-tailed vector autoregression. Our proposed approach allows for exponential increases in dimension with sample size under mild moment and dependence conditions, while ensuring consistency in the estimation process.

Key words and phrases: Bernstein-type Inequality; High Dimensional Time Series; Fat-tailed Data; Robust Estimation.

1. Introduction

High dimensional data analysis has become increasingly important in the era of big data, due to the explosion of massive datasets. High-dimensional linear regression has acquired significant relevance and attention. Specifically, consider the linear regression models

$$Y_i = X_i^\top \beta + \xi_i, \quad i = 1, \ldots, n$$

where Y_i, X_i and ξ_i are the response, covariate and error variables, respectively. Various regularization methods have been widely used for estimating the p-dimensional regression parameter vector, including Tibshirani (1996), Zou and Hastie (2005), Fan and Li (2001), Bickel et al. (2009), Meinshausen and Yu (2009) and many others; see Bühlmann and Van De Geer (2011) for a comprehensive overview. Most investigations assume that the covariates X_i (if it is a random design) and errors ξ_i are i.i.d. Gaussian or sub-Gaussian random variables, which can be too restrictive in many real-world applications.

On one hand, serial correlation might occur when the data are collected over time. Linear regression with time series regressors and autore-

gressive errors is often considered (Harvey (1990), Tsay (1984), Shumway et al. (2000)). On the other hand, many applications involving time series data are concerned with high dimensional objects and fat-tailed distributions, ranging from quantitative finance (Cont (2001)) and portfolio allocation (Kim et al. (2012)) to risk management(Koopman and Lucas (2008)), brain network (Friston (2011)) and geophysical dynamic studies (Kondrashov et al. (2005)).

Previous literature has made progress in linear regression with correlated errors. Specifically, the Lasso estimator was studied for linear regression with autoregressive errors by Wang et al. (2007) and Yoon et al. (2013), weakly dependent errors by Gupta (2012) and long memory errors by Kaul (2014). However, these investigations mainly focus on cases where the dimension p is smaller than the sample size n or where the Gaussian assumption is imposed on the error process. More recently, Wu and Wu (2016) and Chernozhukov et al. (2021) used the framework of functional dependence measures to account for both dependent covariates and errors in linear regression, allowing p to increase with n at a polynomial rate while maintaining consistency. However, a narrow range is still restricted for the dimension in the presence of non-Gaussian and dependent errors. To address the ultra-high dimensional cases where p can grow exponentially

with n, various robust methods have been investigated for linear regression with i.i.d. fat-tailed errors, including penalized Huber M-estimation (Fan et al. (2017), Loh (2017, 2021)), sparse least trimmed squares (Alfons et al. (2013)), ESL-Lasso (Wang et al. (2013)), among others. In this paper, we aim to consider robust estimation of time series regression allowing ultra-high dimensions and fat-tailed and correlated errors.

Vector autoregression (VAR) is another widely used linear model to describe the evolution of a set of variables over time. In recent years, there has been significant progress in estimating high-dimensional VAR models. Inspired from the development in high-dimensional linear regression, Hsu et al. (2008), Nardi and Rinaldo (2011) and Basu and Michailidis (2015) considered the Lasso estimator using ℓ_1 penalty. Kock and Callot (2015) established oracle inequalities for high-dimensional vector autoregressive models. Han et al. (2015) adopted a Dantzig-type penalization. Guo et al. (2016) proposed a Bayesian information criterion based on residual sums of least squares estimator to estimate high dimensional banded autoregression. However, most of these studies require the Gaussian assumption or the existence of finite exponential moment. In econometric analysis, Sims (1980) raised the concern that fat tails in VAR models can affect the validity of statistical inference and may lead to low degrees of freedom due to

the estimation of possibly a large number of parameters. Therefore, there is a need to investigate robust estimation methods for high-dimensional fat-tailed VAR models.

In summary, our work will focus on tackling the challenges posed by high dimensional time series analysis with time series covariates, possibly correlated errors, fat tails, and ultra high dimension. This requires the development of new statistical tools that are tailored to the specific characteristics of these datasets. One of our key contributions is a novel Bernstein-type inequality for the sum of a bounded transformation of high dimensional linear processes. This inequality will be instrumental in obtaining consistent estimators under mild conditions, such as $\log p = o(n^c)$ for some $c > 0$.

The paper is organized as follows. In Section 2, we introduce the framework of high dimensional linear processes and the important quantities that can characterize temporal and cross-sectional dependence. We then present a new Bernstein type inequality for high dimensional linear processes. In Section 3, we investigate two robust estimation problems: (1) time series linear regression with correlated and fat-tailed covariates and errors, and (2) autoregressive models with fat-tailed errors. We also provide some simulation results in Section 4 to assess the empirical performance of robust

estimators. All proofs are relegated to the supplementary material.

We first introduce some notation. For a vector $\beta = (\beta_1, \ldots, \beta_p)^\top$, let $|\beta|_1 = \sum_i |\beta_i|$, $|\beta|_2 = (\sum_i \beta_i^2)^{1/2}$, $|\beta|_0 = |\{i : \beta_i \neq 0\}|$ and $|\beta|_\infty = \max_i |\beta_i|$. Let $\mathrm{Supp}(\beta)$ be the support of β. For a matrix $A = (a_{ij})_{1 \le i,j \le p} \in \mathbb{R}^{p \times p}$, let λ_i, $i = 1, \ldots, p$, be its eigenvalues and $\lambda_{\max}(A) = \max_i |\lambda_i|$ be the spectral radius, $\lambda_{\min}(A) = \min_i |\lambda_i|$. Let $\kappa(A)$ denote the condition number of A. Denote $|A|_1 = \sum_{i,j} |a_{ij}|$, $\|A\|_1 = \max_j \sum_i |a_{ij}|$, $\|A\|_\infty = \max_i \sum_j |a_{ij}|$, spectral norm $\|A\| = \|A\|_2 = \sup_{|x|_2 \neq 0} |Ax|_2 / |x|_2$ and Frobenius norm $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$. Moreover, let $\mathrm{tr}(A)$ be the trace of A, $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ be the entry-wise maximum norm, $|A|$ be a matrix after taking absolute value of A, i.e. $|A| = (|a_{ij}|)_{i,j}$. For a random variable X and $q > 0$, define $\|X\|_q = (\mathbb{E}[|X|^q])^{1/q}$. For two real numbers x, y, set $x \vee y = \max(x, y)$. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists some constant $C > 0$, such that $a_n / b_n \le C$ as $n \to \infty$, and also write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We use $c_0, c_1, \ldots$ and $C_0, C_1, \ldots$ to denote some universal positive constants whose values may vary in different context. Throughout the paper, we consider the high dimensional regime, allowing the dimension p to grow with the sample size n, that is, we assume $p = p_n \to \infty$ as $n \to \infty$.

2. Bernstein-type Inequality for High Dimensional Linear Processes

We consider a general framework of p-dimensional stationary linear process

$$X_i = (X_{i1}, \ldots, X_{ip})^\top = \mu + \sum_{k=0}^{\infty} A_k \boldsymbol{\varepsilon}_{i-k} \tag{2.1}$$

where $\mu \in \mathbb{R}^p$ is the mean vector, $A_0 = I_p$, A_k, $k \geq 1$, are $p \times p$ coefficient matrices with real entries such that $\sum_{k=0}^{\infty} \mathrm{tr}(A_k^\top A_k) < \infty$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})^\top$, and ε_{ij}, $i \in \mathbb{Z}$, $1 \leq j \leq p$, are i.i.d. random variables with zero mean and finite variance. Kolmogorov's three-series theorem ensures that the linear process (2.1) is well-defined. Many researchers have recently worked on this model, including Bhattacharjee and Bose (2014, 2016), Liu et al. (2015), and Chen et al. (2016), among others. One special case of (2.1) is the stationary Gaussian process. If $A_k = 0$ for $k > d$, it becomes a vector moving average process of order d (Reinsel (2003), Lütkepohl (2005), Brockwell and Davis (2009)). Another important class of models covered by (2.1) is the vector autoregressive (VAR) model, which has been widely used in economics and finance (e.g., Sims (1980), Stock and Watson (2001), Tsay (2005), Fan et al. (2011)).

The linear process (2.1) is a flexible multivariate model for correlated

data in that the coefficient matrices A_k capture both temporal and cross-sectional (spatial) dependence. Previous research has explored different structural conditions on the matrices A_k. For example, Liu et al. (2015) worked on a restrictive class of linear processes with matrices A_k that are simultaneously diagonalizable, which implies the absence of spatial dependence among the components. Bhattacharjee and Bose (2016) assumed that $\lim p^{-1}\mathrm{tr}(\Pi)$ exists and is finite for any polynomial Π in $\{A_k, A_k^\top\}$, a joint convergence assumption that is difficult to verify. In this work, we shall impose a condition on the decay rate of the spectral norms of A_k, which allows for more general dependence structures and is easier to check in practice. Assume that there exist $0 < \rho_p < 1$ and $1 \leq \gamma_p < \infty$ such that

$$\|A_k\| = \sup_{|x|_2 \neq 0} \frac{|A_k x|_2}{|x|_2} \leq \gamma_p \cdot \rho_p^k \tag{2.2}$$

for all $k \geq 0$. It implies short-range dependence in the sense that the auto-covariance matrices $\mathrm{Cov}(X_0, X_j) = \sum_{k=0}^{\infty} A_k A_{k+j}^\top$ are absolutely summable. The proposed quantities ρ_p and γ_p can capture temporal and spatial dependence of the underlying high-dimensional process. In particular, ρ_p can depict the strength of temporal dependence, with smaller values indicating faster decay rates and weaker temporal dependence. And the magnitude of γ_p can naturally quantify spatial dependence. A notable feature is that both γ_p and ρ_p may depend on p in the high-dimensional regime. For exam-

ple, when p is large, ρ_p may be a relatively large rate close to 1, indicating slow decay speed. In fact, there exists an absolute constant, independent of p and strictly smaller than 1, such that (2.2) can be rephrased as

$$\|A_k\| \leq \gamma_p \cdot \rho_0^{k/\tau_p} \text{ for some } \tau_p \geq 1. \tag{2.3}$$

Particularly, we define $\tau_p \equiv 1$ if there exists ρ_0 such that $\rho_p \leq \rho_0 < 1$, and $\tau_p = \log \rho_0 / \log \rho_p$ for ρ_0 satisfying $0 < \rho_0 \leq \rho_p$ if ρ_p is large and increase with p. In the latter case, it could happen that $\tau := \tau_p$ is an unbounded function in terms of the dimension p. It is worth noting that measures of dependence quantified by the dimension p have been rarely explored in previous literature, despite their high relevance in analyzing high-dimensional time series. This feature is illustrated by the high dimensional vector autoregressive model in Example 2.1. Thereafter, for notational simplicity, we omit the dimension subscript in γ_p, τ_p, and refer them as γ, τ. And we assume $\tau \leq n$; otherwise there may exist very strong temporal dependence in the sense that $\|A_k\|$ is decaying at a rate no faster than $\rho_0^{1/n}$.

Example 2.1 (High Dimensional Vector Autoregressive Models). Consider the VAR(1) model

$$X_i = AX_{i-1} + \boldsymbol{\varepsilon}_i, \tag{2.4}$$

where $A \in \mathbb{R}^{p \times p}$ is the transition matrix, and $\boldsymbol{\varepsilon}_i$, $i \in \mathbb{Z}$, are i.i.d. error vectors

with mean 0 and covariance matrix I_p. Equivalently it can be represented by the moving average model: $X_i = \sum_{k=0}^{\infty} A^k \boldsymbol{\varepsilon}_{i-k}$, a special case of (2.1) with $A_k = A^k$. The process is stable (and hence stationary) if and only if the spectral radius $\lambda_{\max}(A) < 1$ (Lütkepohl (2005)). If A is symmetric, as $\lambda_{\max}(A) = \|A\|$, condition (2.2) can be easily verified with $\rho_p = \lambda_{\max}(A)$ and $\gamma = 1$. For asymmetric A, it has a better interpretation when we look into condition (2.3), and it could happen that τ may increase with the dimension p. Consider the design $A = (a_{ij})_{i,j=1}^p$ with $a_{ij} = \lambda^{j-i+1}\mathbf{1}\{0 \le j-i \le B-1\}$ for some $0 < \lambda < 1$ and $1 \le B \le p$. Here B depicts how many variables at most in X_{i-1} that have spatial effect on X_{ij}. Figure 1 delivers the plot of $\|A^k\|$ under the numerical setup $\lambda = 0.55$, $B = 3, 4$ and $p = 20, 25, 30$. As can be seen, $\|A^k\|$ decays truly after a certain lag that is moving forward as p is getting larger. This lag can be defined as τ in condition (2.3), so τ increases with p in this design. Additionally the geometric decay (its existence is to be shown later) occurs at a slow speed, viewed as another evidence of large ρ_p (or large τ equivalently). For example, when $B = 3$, $p = 30$, $\|A^k\|$ roughly drops from 1.35 to 0.1 over a broad lag range from 30 to 60. The peak of $\|A^k\|$ before decay is defined as γ, indicating the strength of spatial dependence; by comparing the two plots, we can tell that stronger spatial dependence with a larger B results to larger γ.

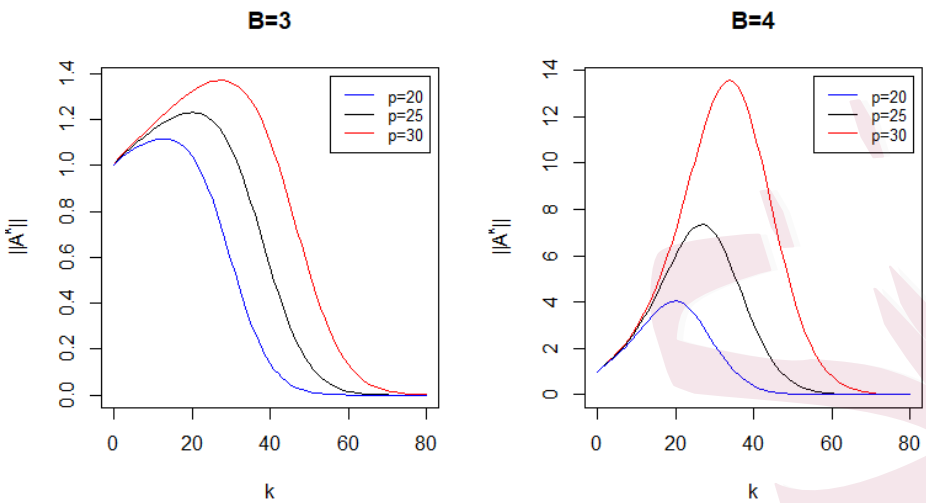

Figure 1: The graph of $\|A^k\|$ for $B = 3, 4$ and $p = 20, 25, 30$.

Concentration inequalities play an important role in the study of sums of random variables. A number of inequalities have been derived for independent random variables; see Bühlmann and Van De Geer (2011) for a review. Bernstein's inequality (Bernstein (1946)) is one of the powerful tools when analyzing the concentration behavior by providing an exponential inequality for sums of independent random variables which are uniformly bounded. To fix the idea, let $Y_1, \ldots, Y_n$ be i.i.d. random variables such that $\mathbb{E}Y_i = 0$, $\mathrm{Var}(Y_i) = \sigma^2 < \infty$, and $|Y_i| \leq M$ for all i. Then for any $x > 0$, one has

$$\mathbb{P}\Big(\sum_{i=1}^{n} Y_i \geq x\Big) \leq \exp\Big\{-\frac{x^2}{2n\sigma^2 + 2Mx/3}\Big\}, \tag{2.5}$$

which suggests two types of bound for tail probability: sub-Gaussian-type

tail $\exp\{-x^2/(Cn\sigma^2)\}$ in terms of the variance of $\sum_{i=1}^n Y_i$ and sub-exponential-type tail $\exp\{-x/(CM)\}$ in terms of the uniform bound M. Bernstein type inequalities have been developed for Markov chains or temporally dependent processes with an additional order ($\log n$ in some constant powers) in the sub-exponential-type tail; see, for example, Adamczak et al. (2008), Merlevède et al. (2009), Hang and Steinwart (2017) and Zhang (2021). The problem of generalizing to high dimensional time series is quite challenging and very few results have been obtained. Our first goal is to establish a new Bernstein type inequality for the sum of a bounded transformation of the high dimensional linear processes in (2.1).

Theorem 2.1. *Let X_i be the linear process generated from (2.1) with $\mathbb{E}\varepsilon_{ij} = 0$, $\mathbb{E}\varepsilon_{ij}^2 = \sigma^2 < \infty$ and condition (2.3) be satisfied. Let $G : \mathbb{R}^p \to \mathbb{R}$ be a function with $|G(u)| \le M$ for all $u \in \mathbb{R}^p$. Suppose there exists a vector $g = (g_1, \ldots, g_p)^\top$ with $g_i \ge 0$ and $\sum_{i=1}^p g_i = 1$ such that the following Lipschitz condition holds: for all $u = (u_1, \ldots, u_p)^\top$ and $v = (v_1, \ldots, v_p)^\top$,*

$$|G(u) - G(v)| \le \sum_{i=1}^p g_i |u_i - v_i|. \tag{2.6}$$

Then for any $x > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^n G(X_i) - \mathbb{E}G(X_i) \ge x\Big) \le 2\exp\Big\{-\frac{x^2}{C_1 n\sigma^2\tau^2\gamma^2 + C_2\tau M x}\Big\}, \tag{2.7}$$

where the constants C_1 and C_2 are given by

$$C_1 = \frac{16\mathrm{e}^2}{\sqrt{2\pi}\rho_0^4[\log(1/\rho_0)]^3}, \quad C_2 = \frac{8\mathrm{e}}{\log(1/\rho_0)}.$$

Remark 2.2. Equipped with our new inequality (2.7), one can investigate the concentration properties of sums of bounded transformations of high dimensional linear processes that exhibit both temporal and cross-sectional dependence, characterized by τ and γ respectively. In the special case where the processes are one-dimensional, denoted by $X_i \in \mathbb{R}$, and $\tau = 1$ and γ is of a constant order that satisfies condition (2.2), our probability inequality (2.7) is just as sharp as the classical Bernstein's inequality (2.5). It is worth mentioning that our inequality is strictly sharper than the existing Bernstein-type inequalities for univariate time series established by Merlevède et al. (2009) and Zhang (2021). To fix the idea, we shall recall that Merlevède et al. (2009) derived a concentration inequality for univariate strong mixing process (X_i) of mean 0 and upper bounded by M in magnitude:

$$\mathbb{P}\Big(\sum_{i=1}^n X_i \geq x\Big) \leq \exp\Big\{-\frac{Cx^2}{nv^2 + M^2 + M(\log n)^2 x}\Big\}, \tag{2.8}$$

where v^2 is the asymptotic variance of $\sum_{i=1}^n X_i/\sqrt{n}$. Zhang (2021) obtained a similar bound with v^2 represented in terms of functional dependence measures. In our framework of linear processes with condition (2.2)

satisfied, $v^2 \asymp \sigma^2\gamma^2$ can be computed for one-dimensional cases. Notably, our inequality is sharper by removing the additional factor $(\log n)^2$ in the sub-exponential type bound.

To study high dimensional time series, an important class of transformations is linear combinations of transformed component processes, that is, $G(X_i) = \sum_{j=1}^p a_j h_j(X_{ij})$, where $\sum_{j=1}^n |a_j| = 1$, $h_j : \mathbb{R} \to \mathbb{R}$ are univariate functions satisfying $|h_j(x)| \leq M$ and $|h_j(x) - h_j(y)| \leq 1$ for any $x, y \in \mathbb{R}$, thus condition (2.6) is satisfied with $g_j = |a_j|$. As a special case, when $G(X_i) = h_j(X_{ij})$, for a fixed $1 \leq j \leq p$, the result provides a concentration inequality for sums of each component process $(X_{ij})_{i\in\mathbb{Z}}$ after the transformation h_j. This is useful in the application of estimating the mean vector of high-dimensional linear processes in a robust way, as discussed at the end of this section. In Remark 2.3, we highlight that our inequality yields a rate of ℓ_∞ norm convergence for the robust mean estimator that is as sharp as the optimal rate for i.i.d. processes.

Condition (2.3) requires $\|A_k\|$ geometrically decayed up to the quantity γ and the decay speed is controlled by τ. Chen et al. (2016) worked on the same linear model under a weaker condition allowing polynomial decay, namely, $\|A_k\| = O((1 \vee k)^{-\alpha})$ for some $\alpha > 1$, under which, it is noteworthy that an exponential type probability inequality does not hold in general

even if it is one dimensional process with a uniform bound. That is to say, if we relax the condition (2.2) to a polynomial decay, the concentration inequality delivers an exact rate with algebraic decay for one dimensional linear process; see Theorem 14 in Chen and Wu (2018).

In Theorem 2.1, the existence of a finite variance of ε_{ij} is assumed. If it is relaxed to the existence of finite exponential moment, a similar bound can be achieved with G not necessarily bounded; see Theorem 2.2 below.

Theorem 2.2. *In the model (2.1), assume that* $\mathbb{E}\varepsilon_{ij} = 0$, $\mathbb{E}\exp(c_0|\varepsilon_{ij}|) = \theta < \infty$ *for some constant* $c_0 > 0$ *and condition (2.3) is met. Then for* G *satisfying (2.6), it holds that*

$$\mathbb{P}\Big(\sum_{i=1}^{n} G(X_i) - \mathbb{E}G(X_i) \geq x\Big) \leq 2\exp\Big\{-\frac{x^2}{C_3 n\theta^2\tau^2\gamma^2 + C_4\gamma\tau x}\Big\}, \quad (2.9)$$

where the constants C_3 *and* C_4 *depend on* ρ_0 *and* c_0.

One immediate application of Theorem 2.1 is to estimate the mean vector for high dimensional fat-tailed linear processes. From an M-estimation viewpoint, we apply Huber's estimator (Huber (1964)) of the mean vector, denoted by $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)^\top$, with $\hat{\mu}_j$ as the solution of a to the equation

$$\sum_{i=1}^{n} \phi_\nu(X_{ij} - a) = 0,$$

where $\phi_\nu(x) = (x \wedge \nu) \vee (-\nu)$ is the Huber function with the robustification parameter $\nu > 0$.

Theorem 2.3. *Let X_i be generated from model (2.1) with $\mathbb{E}\varepsilon_{ij} = 0$, $Var(\varepsilon_{ij}) = 1$, $\mu = \mathbb{E}X_i$ and $\max_{1\le j\le p} Var(X_{ij}) = \mu_2^2 < \infty$. Choose $\nu \asymp \mu_2\sqrt{n/\log p}$. With probability at least $1 - 4p^{-c}$ for some $c > 0$, it holds that*

$$|\hat{\mu} - \mu|_\infty \le C(\gamma + \mu_2)\tau\sqrt{\frac{\log p}{n}} \tag{2.10}$$

under the scaling condition $(\gamma + \mu_2)\tau\sqrt{\log p/n} \to 0$, where C is a positive constant depending on c and the constants C_1, C_2 in Theorem 2.1.

Remark 2.3. Theorem 2.3 delivers the rate of ℓ_∞ norm convergence for the robust mean estimator $\hat{\mu}$ and it involves a delicate interplay with the cross-sectional dependence, temporal dependence and the variance of the process. If γ, μ_2 and τ are all of a constant order, it follows that

$$|\hat{\mu} - \mu|_\infty = O_{\mathbb{P}}(\sqrt{\log p/n}), \tag{2.11}$$

under the scaling condition $\log p/n \to 0$. We shall remark that (2.11) is as sharp as the optimal rate provided in Theorem 5 of Fan et al. (2017) concerning the concentration of the mean estimation for the i.i.d. case. And it is strictly sharper than the results using existing Bernstein type inequalities for time series such as the ones in Merlevède et al. (2009), Hang and Steinwart (2017) and Zhang (2021).

3. Robust Estimation of Time Series Regression

In this section, we shall investigate robust estimation of high dimensional time series linear regression and autoregression with fat-tailed covariates and errors. It is expected that our framework of high dimensional linear processes and these Bernstein type inequalities will be useful in other high-dimensional estimation and inference problems that involve dependent and non-sub-Gaussian random variables.

3.1 Estimating Time Series Regression with Correlated Errors

We work on linear regression models with random design that involve time dependent covariates and errors:

$$Y_i = X_i^\top \beta^* + \xi_i, \tag{3.1}$$

with more justification provided as follows.

Assumptions.

(A1) X_i is generated from the p-dimensional linear process $X_i = \sum_{k=0}^{\infty} A_k \boldsymbol{\varepsilon}_{i-k}$ where the components of $\boldsymbol{\varepsilon}_i$ are i.i.d. random variables with $\mathbb{E}(\varepsilon_{ij}) = 0$ and $\mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2 < \infty$. Condition (2.3) is satisfied with γ and τ, which may depend on p.

(A2) $\xi_i = \sum_{k=0}^{\infty} b_k \eta_{i-k}$ is the error process, where η_i are i.i.d. random vari-

ables with $\mathbb{E}(\eta_i) = 0$ and $\mathrm{Var}(\eta_i) = \sigma_\eta^2 < \infty$, and $b_k \leq C\rho^k$ for universal constants $0 < \rho < 1$ and $C < \infty$.

(A3) X_i is strictly exogenous in the sense that $(\boldsymbol{\varepsilon}_i)_i$ are independent of $(\eta_i)_i$, where $(\boldsymbol{\varepsilon}_i)_i$ and $(\eta_i)_i$ are error processes of X_i and ξ_i respectively as defined in (A1) and (A2).

The framework (3.1) is quite general as the linear process includes a wide range of commonly used time series models. For linear regression models with dependent errors, earlier work mainly dealt with fixed design or i.i.d. covariates. Wang et al. (2007) and Yoon et al. (2013) considered the case where ξ_i follows an autoregressive process, one type of linear processes. Gupta (2012) concerned weakly dependent ξ_i introduced by Doukhan and Louhichi (1999) and specifically discussed the AR(1) and ARMA cases. Alfons et al. (2013) adopted the same format of moving average errors but assumed long memory dependence. More generally, Wu and Wu (2016) and Chernozhukov et al. (2021) considered the nonlinear Wold representation with $X_i = g(\ldots, \boldsymbol{\varepsilon}_{i-1}, \boldsymbol{\varepsilon}_i)$ and $\xi_i = h(\ldots, \eta_{i-1}, \eta_i)$.

We form a modified ℓ_1-regularized Huber estimator of β, given by

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \Phi_\nu((Y_i - X_i^\top \beta) w(X_i)) + \lambda |\beta|_1,$$

where Φ_ν is Huber loss function (Huber (1964))

$$\Phi_\nu(x) = \begin{cases} x^2/2, & \text{if } |x| \le \nu, \\ \nu|x| - \nu^2/2, & \text{if } |x| > \nu, \end{cases}$$

defined with respect to the robustification parameter $\nu > 0$. More properties of Huber regression are referred to Huber (1973), Yohai and Maronna (1979), Mammen (1989), Sun et al. (2020), Fan et al. (2017) among others. Motivated by Loh (2021), $w(x) : \mathbb{R}^p \to \mathbb{R}$ is a weight function defined by

$$w(x) = \min\left\{1, \frac{b}{|Bx|_2}\right\}$$

where $b \in \mathbb{R}$ is a fixed constant and $B \in \mathbb{R}^{p\times p}$ is a provided positive definite matrix. With such a choice of $w(x)$, it always holds that $|w(x)x|_2 \le b/\lambda_{\min}(B) =: b_0$. Different from the regular Huber regression concerning well-behaved X_i (e.g., Gaussian or sub-Gaussian), an additional weight function is incorporated on the covariate process to account for the fat tails of X_i. In Section S1, we conduct a simulation study for robust time series regression estimation and look into the effect of $w(x)$.

As a popular convention, β^* is assumed to be sparse in the sense that $|\beta^*|_0 = s$. Denote the condition number of B as $\kappa(B) = \lambda_{\max}(B)/\lambda_{\min}(B)$. Theorem 3.1 below concerns the estimation consistency of $\hat{\beta}$.

Theorem 3.1. *Let Assumptions (A1) (A2) (A3) be satisfied. Assume*

$$b_0(b_0 + \kappa(B)\gamma\sigma_\varepsilon)\tau\sqrt{s}\sqrt{(\log p)^3/n} \to 0. \tag{3.2}$$

Choose $\nu \asymp \sigma_\eta(n/\log p)^{1/2}$ *and* $\lambda \asymp b_0\sigma_\eta(\log p/n)^{1/2}$. *With probability at least* $1 - 8p^{-c}$ *for some* $c > 0$, *it holds that that*

$$|\widehat{\beta} - \beta|_2 \leq C\frac{b_0\sigma_\eta}{\lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2}X_iX_i^\top])}\sqrt{\frac{s\log p}{n}}. \tag{3.3}$$

The scaling condition (3.2) to ensure consistency indicates a subtle interplay with the dimensionality parameters (s, p, n), internal parameters $(\tau, \gamma, \sigma_\varepsilon)$, and the known values b_0 and $\kappa(B)$ associated with the weight function $w(x)$. The convergence rate (3.3) scales inversely with the quantity $\lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2}X_iX_i^\top])$ and it suggests that we can not shrink the covariates too aggressively. If X_i is well-behaved with the existence of finite exponential moment, one may eliminate the weight function and replace the factor by the larger quantity $\lambda_{\min}(\mathbb{E}[X_iX_i^\top])$.

In the extensively studied regression setting with i.i.d. covariates, Fan et al. (2017) delivered an optimal convergence rate of $|\widehat{\beta} - \beta|_2$ for weakly sparse model under the fat tails (the same as the minimax rate in Raskutti et al. (2011)). In the special exact sparse case, their convergence rate is $\sqrt{s(\log p)/n}$ and it relies on the sub-Gaussian tail assumption for the covariates X_i. Loh (2021) allowed broader classes of distributions for X_i by

inserting a weight function to control the Euclidean norm of X_i, but required the errors drawn i.i.d. from a symmetric distribution and thus selected ν at a fixed constant order (cf. Theorem 1), while Fan et al. (2017) waived the symmetry requirement by allowing ν to diverge in order to reduce the bias induced by the Huber loss when the distribution of ξ_i is asymmetric. We borrow the ideas from both and further account for time dependent covariates and errors. Compared to Loh (2021) with i.i.d. covariates and i.i.d. errors, our result requires a stronger scaling condition (3.2) in terms of the dependence quantities γ, τ and a larger power of $\log p$, by concerning both dependent covariates and errors.

Applying ℓ_1 regularization in time series regression, Wu and Wu (2016) (cf. Theorem 5) dealt with correlated covariates and errors and allowed a wider class of stationary processes in a causal form. The linear error process in our consideration falls in the weaker dependence range within their framework. If $\gamma, \tau, \sigma_\eta = O(1)$, $p = o(n^{q-1})$ is required for their regular estimator without accounting for robustness, where $q > 2$ is the order of finite moments for ξ_i. Chernozhukov et al. (2021) considered the Lasso estimator for a system of time series regression equations with one regression equation as a special case, for which the allowed dimension is still of a polynomial rate to ensure consistency by looking into the performance bound with re-

spect to the prediction norm (cf. Corollary 5.4). In comparison with the above two work, we can allow a much wider range for the dimension p under mild conditions.

The tuning parameter ν plays a key role by adapting to errors with fat tails. In practical applications, the optimal values of the tuning parameters ν and λ can be chosen by a two-dimensional grid search using cross-validation or information-based criterion such as AIC or BIC. We leave theoretical investigation on selecting the tuning parameters as important future work.

3.2 Estimating Transition Matrix in VAR Models

To study the evolution of a set of endogenous variables over time, a popular choice is vector autoregression. Interpretations of large vector autoregressive models have been developed in various applications such as policy analysis (Sims (1992)), financial systemic risk analysis (Gourieroux and Jasiak (2011)), portfolio selection (Ledoit and Wolf (2003)), functional genomics (Shojaie et al. (2012)) and brain networks (Sameshima and Baccala (2014)).

As a general VAR model of order d can be reformulated as a VAR(1) model by appropriately redefining the random vectors, much work (Han et al. (2015), Guo et al. (2016)) considered the model with lag 1 as given

in (2.4). Among the work concerning high dimensional vector autoregressive models, most investigations require the Gaussian assumption (Kock and Callot (2015), Basu and Michailidis (2015), Han et al. (2015)) or some structure assumption stronger than the minimal requirement $\lambda_{\max}(A) < 1$; for example, Han et al. (2015) imposed $\|A\| < 1$, and Guo et al. (2016) considered banded A with some decay condition on $\|A^k\|$ free of p. For many VAR designs (Example 2.1 is one such), it could happen that $\|A\| \geq 1$ and the dimension p, as the size of A, can play a role in measuring the temporal and cross-sectional dependence. Basu and Michailidis (2015) proposed stability measures to capture temporal and cross-sectional dependence. From a different viewpoint, we try to fill in the gap between the spectral radius of a matrix and its spectral norm. Intuition can be gained from the proposition below. It provides a sufficient and necessary condition for $\lambda_{\max}(A) < 1$ by relating to the spectral norm.

Proposition 3.2. *For any matrix A, it holds that $\lambda_{\max}(A) < 1$ if and only if there exists some finite integer k such that $\|A^k\| \leq \rho_0$ given any universal constant $0 < \rho_0 < 1$.*

Letting $\tau = \min\{k \in \mathbb{Z}^+ : \|A^k\| \leq \rho_0\}$ and $\gamma = \rho_0^{-1} \max_{0 \leq k \leq \tau - 1} \|A^k\|$, condition (2.3) holds for the model (2.4) without extra requirement. We now introduce the notation. Let $\boldsymbol{a}_{j\cdot}^\top$ be the j-th row of A and s_j be the

cardinality of the support set of $\boldsymbol{a}_{j\cdot}$, i.e., $s_j = |\text{supp}(\boldsymbol{a}_{j\cdot})| = |\{i : a_{ij} \neq 0\}|$. Denote $s = \max_{1\leq j\leq p} s_j$ and $\mathcal{S} = \sum_{i=j}^{p} s_j$. For robustness, we first truncate the data by obtaining $\tilde{X}_i = \phi_\nu(X_i)$, where ν is the truncation parameter and is to be determined later. For notational convenience, we assume X_0 is also observed. Based on the truncated sample $\widetilde{X}_i$ and tuning parameter $\lambda > 0$, we propose to estimate A by solving the following Lasso problem:

$$\widehat{A} = \arg\min_{B\in\mathbb{R}^{p\times p}} \frac{1}{n}\sum_{i=1}^{n} |\widetilde{X}_i - B\widetilde{X}_{i-1}|_2^2 + \lambda|B|_1, \tag{3.4}$$

which is equivalent to solving the p sub-problems:

$$\widehat{\boldsymbol{a}}_{j\cdot} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} (\widetilde{X}_{ij} - \boldsymbol{b}^\top \widetilde{X}_{i-1})^2 + \lambda|\boldsymbol{b}|_1. \tag{3.5}$$

Before proceeding, we state the key assumptions on the process (2.4) and some scaling conditions to guarantee consistency of the robust estimator $\hat{A}$.

Assumptions.

(B1) $\mathbb{E}\varepsilon_{ij} = 0$; $\mathbb{E}\varepsilon_{ij}^2 = 1$; $\max_{1\leq j\leq p} \|X_{ij}\|_q = \mu_q < \infty$ for some $q > 2$.

(B2) $\mu_q\gamma\tau s^2[(\log p)/n]^{(q-2)/(2q-2)} \to 0$.

(B2') $\mu_q\gamma\tau\mathcal{S}^2[(\log p)/n]^{(q-2)/(2q-2)} \to 0$.

Assumption (B1) imposes polynomial moment conditions for the underlying VAR process. Assumption (B2) (or (B2')) assumes a vanishing scaling property. If μ_q, τ and γ are of a constant order, (B2) is reduced to the scaling condition that involves s (or $\mathcal{S}$), n and p only.

Theorem 3.3. *Let Assumptions (B1) and (B2) be satisfied. Choose the truncation parameter* $\nu \asymp \mu_q(n/\log p)^{1/(2q-2)}$. *Let* $\widehat{A}$ *be the solution of (3.4) with* $\lambda \asymp \mu_q\gamma\tau(\|A\|_\infty + 1)[(\log p)/n]^{(q-2)/(2q-2)}$. *It holds that*

$$\|\widehat{A} - A\|_\infty \leq C\mu_q\gamma\tau(\|A\|_\infty + 1)s\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{1}{2q-2}} \tag{3.6}$$

with probability at least $1 - 8p^{-c}$ *for some constant* $c > 0$. *If Assumption (B2') is further satisfied, it also holds that*

$$\|\widehat{A} - A\|_F \leq C'\mu_q\gamma\tau(\|A\|_\infty + 1)\sqrt{\mathcal{S}}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{1}{2q-2}} \tag{3.7}$$

with probability at least $1 - 8p^{-c}$ *for some constant* $c > 0$.

The obtained rates of convergence are governed by two sets of parameters: (i) dimensionality parameters: the dimension p, sparseness parameter s (or $\mathcal{S}$), and the sample size n; (ii) internal parameters: the moment μ_q, dependence quantities τ, γ, and the maximum absolute row sum $\|A\|_\infty$. If the internal parameters are assumed to be of a constant order, we can get

$$\|\widehat{A} - A\|_F = O_{\mathbb{P}}\Big(\sqrt{\mathcal{S}}\big(\frac{\log p}{n}\big)^{\frac{1}{2}-\frac{1}{2q-2}}\Big).$$

To ensure consistency, the dimension p can be allowed to increase exponentially with n in view of the mild scaling condition. Compared to Guo et al. (2016) with the same constant order of internal parameters, they can only

allow the narrower range $p = o(n^c)$ for some $0 < c < (q-4)/8$ (cf. Condition 4(i)). For Gaussian autoregressive models, proposition 4.1 of Basu and Michailidis (2015) suggests the order in terms of dimension parameters as

$$\|\widehat{A} - A\|_F = O_{\mathbb{P}}\Big(\sqrt{\mathcal{S}}\sqrt{\frac{\log p}{n}}\Big).$$

In the presence of fat tails with the existence of finite q-th moment, our result yields a slightly slower convergence rate characterized by the moment order q and it is closer to their bound when q gets larger.

As an alternative method, the idea of Dantzig-type estimation (Candes et al. (2007), Cai et al. (2011), Han et al. (2015)) can be modified in the robust way. Let Σ_k denote the autocovariance matrix of the process (X_i) at lag k. The celebrated Yule-Walker equation $A = \Sigma_0^{-1}\Sigma_1$ suggests that a good estimate $\widehat{A}$ should have a small error in terms of $\|\Sigma_0\widehat{A} - \Sigma_1\|_{\max}$. Without direct access to the autocovariance matrices Σ_0 and Σ_1, a natural approach is to find nice estimators for them. Han et al. (2015) used sample autocovariance matrices and enjoyed a nice performance bound under Gaussianity. For fat-tailed cases, we consider the robust estimators of the autocovariance matrices based on the truncated sample:

$$\widehat{\Sigma}_k = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{i-k}\widetilde{X}_i^\top, \text{ for } k = 0, 1.$$

The Dantzig- type estimator is then modified to solving the following convex

programming:

$$\widehat{A} = \arg\min_{B\in\mathbb{R}^{p\times p}} |B|_1 \quad \text{s.t.} \quad \|\widehat{\Sigma}_0 B - \widehat{\Sigma}_1\|_{\max} \le \lambda, \tag{3.8}$$

where $\lambda > 0$ is a tuning parameter. Observe that problem (3.8) can be solved in parallel, namely, (3.8) is equivalent to p subproblems:

$$\widehat{\boldsymbol{a}}_{\cdot j} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} |\boldsymbol{b}|_1 \quad \text{s.t.} \quad |\widehat{\Sigma}_0 \boldsymbol{b} - \widehat{\Sigma}_1 u_j|_\infty \le \lambda, \quad j = 1, \ldots, p \tag{3.9}$$

for any unit vector u_j. Let $\boldsymbol{a}_{\cdot 1}, \boldsymbol{a}_{\cdot 2}, \ldots, \boldsymbol{a}_{\cdot p}$ be columns of A and denote $s^* = \max_{1\le j\le p} |\text{supp}(\boldsymbol{a}_{\cdot j})|$. We can obtain $\widehat{A}$ by simply concatenating all the columns $\widehat{\boldsymbol{a}}_{\cdot j}$, i.e. $\widehat{A} = (\widehat{\boldsymbol{a}}_{\cdot 1}, \widehat{\boldsymbol{a}}_{\cdot 2}, \ldots, \widehat{\boldsymbol{a}}_{\cdot p})$. The next theorem delivers an upper bound on the statistical accuracy.

Theorem 3.4. *Let Assumption (B1) be satisfied. Let $\widehat{A}$ be the solution of (3.8) with $\nu \asymp \mu_q (n/\log p)^{1/(2q-2)}$ and $\lambda \asymp \mu_q \gamma\tau(\|A\|_1+1)[(\log p)/n]^{(q-2)/(2q-2)}$. With probability at least $1 - 8p^{-c'}$ for some constant $c' > 0$, it holds that*

$$\|\widehat{A} - A\|_{\max} \le C\mu_q\gamma\tau\|\Sigma_0^{-1}\|_1(\|A\|_1 + 1)\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{1}{2q-2}}, \tag{3.10}$$

$$\|\widehat{A} - A\|_1 \le C'\mu_q\gamma\tau\|\Sigma_0^{-1}\|_1(\|A\|_1 + 1)s^*\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{1}{2q-2}}. \tag{3.11}$$

It is interesting to see that the convergence rate of the modified Dantzig-type estimator has a similar form to that of the robust Lasso estimator developed in Theorem 3.3, if the included internal parameters for the process are of a constant order. Both methods involve p parallel programming

problems with the lasso-based one concerning row-by-row estimation while the Dantzig method concerning column-by-column instead. The situation $\|A\| < 1$ studied by Han et al. (2015) is the special case where $\gamma = 1$ and $\tau = 1$ in our framework. In their paper, a more flexible sparse condition was imposed: the transition matrix A belongs to a class of weakly sparse matrices defined in terms of strong ℓ^r-ball ($0 \leq r < 1$), which was also considered by Bickel and Levina (2008), Rothman et al. (2009), Cai et al. (2011), Cai and Zhou (2012) in estimating covariance and precision matrices. For $r = 0$, it is the exact sparse case and Theorem 1 in Han et al. (2015) implies the dimension parameter order

$$\|\widehat{A} - A\|_1 = O_{\mathbb{P}}\Big(s^* \sqrt{\frac{\log p}{n}}\Big),$$

a bit sharper than our result (3.11). There is additional cost for fat-tailed processes with robustness absorbed. We shall remark that we can also derive the bound of $\|\widehat{A} - A\|_1$ accordingly for weakly sparse A based on the result (3.10) without any technical difficulty.

4. Simulation Study

In this section, we evaluate the finite sample performance of both robust Lasso and Dantzig estimators that are proposed in Section 3.2 and compare with the traditional Lasso and Dantzig methods. Simulation on time

series linear regression is presented in the supplementary material. We consider the model (2.4), where ε_{ij}'s are i.i.d. standardized Student's t-distributions with df $= 5$. We take the numerical setup of $n = 50, 100$ and $p = 50, 100, 500$ and set $s = \lfloor \log p \rfloor$. For the true transition matrix $A = (a_{ij})$, we consider the following designs.

(1) Banded: $A = (\lambda^{|i-j|}\mathbf{1}\{|i-j| \leq s\})$ and $\lambda = 0.5$.

(2) Block diagonal: $A = \text{diag}\{A_i\}$, where each $A_i \in \mathbb{R}^{s\times s}$ follows the structure in Example 2.1 with $B = 2$ and $\lambda_i \sim Unif(-0.8, 0.8)$.

(3) Toeplitz: $A = (\lambda^{|i-j|})$ and $\lambda = 0.5$.

(4) Random Sparse: $a_{ii} \sim Unif(-0.8, 0.8)$ and $a_{ij} \sim N(0, 1)$ for $(i, j) \in C \subset \{(i, j) : i \neq j\}$ where C is randomly selected and $|C| = s^2$.

To ensure stationarity of the VAR model, the designs in (1), (3), and (4) were further scaled by a factor of $2\lambda_{\max}(A)$ to ensure that the spectral radius of the transition matrix is less than 1. Figure 2 shows the plot of $\|A^k\|$ under the four designs with $p = 100, 500$. These patterns of matrix A were previously studied in Han et al. (2015), where the assumption $\|A\| < 1$ was necessary. In this study, we keep the designs of symmetric sparse and weakly sparse matrices, which are presented in cases (1) and (3), respectively. For these two cases, it holds that $\|A^k\| = (\lambda_{\max}(A))^k = (0.5)^k$, and condition

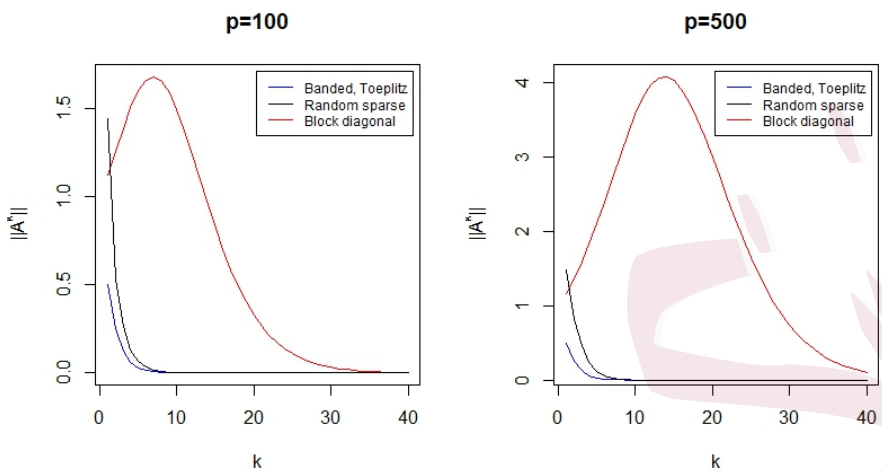

Figure 2: The graph of $\|A^k\|$ for the four designs of A with $p = 100, 500$

(2.3) is satisfied with $\tau = 1$, $\gamma = 1$ and $\rho_0 = 0.5$. However, for the designs using asymmetric coefficient matrices (cases (2) and (4)), we allow $\|A\| > 1$, and τ and γ in condition (2.3) may depend on the value of p.

In each repetition, we generate a process of length $2n$. We run the estimation procedure in (3.4) or (3.8) based on $\{X_1, \ldots, X_n\}$ by a two-dimensional grid search for the tuning parameters ν and λ. For each (ν, λ) in the grid, denote the estimator by $\hat{A}(\nu, \lambda)$. Then (ν, λ) is chosen such that $n^{-1} \sum_{t=n+1}^{2n} |X_t - \hat{A}(\nu, \lambda) X_{t-1}|_2^2$, the average prediction error on $\{X_{n+1}, \ldots, X_{2n}\}$, is minimized. The following tables report the average (as an entry) and standard deviation (in parentheses) of estimation error based on 1000 repetitions in different matrix norms consistent with Theorem 3.3 and Theorem 3.4. As comparisons, we obtain the results for robust

$p = 50, n = 100$	Banded	Block	Toeplitz	Random
Lasso L_∞	1.49 (0.060)	0.96 (0.072)	1.46 (0.143)	1.28 (0.124)
Lasso L_F	1.56 (0.112)	1.22 (0.112)	1.55 (0.121)	1.30 (0.084)
Robust-Lasso L_∞	1.35 (0.049)	0.80 (0.078)	1.30 (0.072)	1.17 (0.065)
Robust-Lasso L_F	1.37 (0.076)	1.05 (0.041)	1.36 (0.090)	1.23 (0.038)
Dantzig L_1	2.01 (0.121)	1.91 (0.087)	2.02 (0.140)	2.40 (0.159)
Dantzig L_F	2.10 (0.095)	1.92 (0.074)	2.04 (0.125)	2.69 (0.078)
Robust-Dantzig L_1	1.86 (0.050)	1.08 (0.058)	1.86 (0.043)	1.47 (0.077)
Robust-Dantzig L_F	1.90 (0.049)	1.41 (0.044)	1.89 (0.033)	2.02 (0.073)

methods and the traditional versions (Lasso estimator in Tibshirani (1996) and Dantzig-based estimator in Han et al. (2015)) in different designs.

From statistical perspective, the tables indicate that both of our robust estimation methods outperform the regular Lasso and Dantzig, when the innovation vectors have fat tail and the transition matrix enjoys a sparsity pattern. In a nutshell, our robust methods is more advantageous in tackling non-Gaussian time series.

5. Concluding Remarks

Time series regression arises in a wide range of disciplines. Conventional tools are inadequate when it involves high dimensional temporal dependent

$p = 100, n = 50$	Banded	Block	Toeplitz	Random
Lasso L_∞	2.64 (0.205)	2.31 (0.093)	2.49 (0.308)	2.40 (0.114)
Lasso L_F	2.73 (0.168)	2.44 (0.141)	2.74 (0.125)	2.48 (0.119)
Robust-Lasso L_∞	2.65 (0.073)	2.26 (0.101)	2.67 (0.039)	2.18 (0.084)
Robust-Lasso L_F	2.67 (0.080)	2.38 (0.139)	2.69 (0.052)	2.32 (0.131)
Dantzig L_1	3.13 (0.177)	2.70 (0.146)	3.15 (0.140)	3.21 (0.136)
Dantzig L_F	3.16 (0.073)	3.06 (0.172)	3.58 (0.116)	3.75 (0.173)
Robust-Dantzig L_1	1.80 (0.069)	1.82 (0.051)	1.72 (0.047)	1.51 (0.073)
Robust-Dantzig L_F	2.78 (0.071)	2.01 (0.104)	2.77 (065)	2.45 (0.090)

$p = 500, n = 100$	Banded	Block	Toeplitz	Random
Lasso L_∞	4.99 (0.091)	4.12 (0.043)	4.27 (0.052)	4.49 (0.019)
Lasso L_F	8.16 (0.070)	7.98 (0.004)	8.05 (0.021)	7.82 (0.052)
Robust-Lasso L_∞	4.80 (0.012)	3.31 (0.015)	3.55 (0.051)	3.40 (0.017)
Robust-Lasso L_F	7.51 (0.120)	7.50(0.177)	7.69 (0.158)	6.69 (0.220)
Dantzig L_1	5.03 (0.070)	5.64 (0.034)	5.18 (0.055)	5.43 (0.050)
Dantzig L_F	8.64 (0.169)	9.03 (0.199)	9.18 (0.222)	8.43 (0.192)
Robust-Dantzig L_1	4.51 (0.030)	4.50 (0.017)	4.69 (0.037)	4.69 (0.034)
Robust-Dantzig L_F	7.11 (0.123)	7.05 (0.102)	7.09 (0.099)	6.76 (0.122)

and fat-tailed data. In this paper, we develop a novel Bernstein inequality for high dimensional linear processes, with the help of which, we have made contributions towards the robust estimation theory of high dimensional time series regression in the presence of fat tails. The convergence rate depends on the strength of temporal and cross-sectional dependence, the moment condition, the dimension and the sample size. We allow the dimension to increase exponentially with the sample size as a natural requirement of consistency. To perform statistical inference of the estimates such as hypothesis testing and construction of confidence intervals, one needs to develop the deeper result in terms of asymptotic distributional theory. The latter is more challenging and we leave it as future work.

References

Adamczak, R. et al. (2008). A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability 13*, 1000–1034.

Alfons, A., C. Croux, and S. Gelper (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics 7*(1), 226–248.

Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics 43*(4), 1535–1567.

Bernstein, S. (1946). *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow.

Bhattacharjee, M. and A. Bose (2014). Estimation of autocovariance matrices for infinite dimensional vector linear process. *Journal of Time Series Analysis 35*(3), 262–281.

Bhattacharjee, M. and A. Bose (2016). Large sample behaviour of high dimensional autocovariance matrices. *The Annals of Statistics 44*(2), 598–628.

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *The Annals of statistics 36*(6), 2577–2604.

Bickel, P. J., Y. Ritov, A. B. Tsybakov, et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Brockwell, P. J. and R. A. Davis (2009). *Time series: theory and methods*. Springer science & business media.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T., W. Liu, and X. Luo (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Cai, T. T. and H. H. Zhou (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics 40*(5), 2389–2420.

Candes, E., T. Tao, et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics 35*(6), 2313–2351.

Chen, L. and W. Wu (2018). Concentration inequalities for empirical processes of linear time

series. *Journal of Machine Learning Research 18*(231), 1–46.

Chen, X., M. Xu, and W. B. Wu (2016). Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Transactions on Signal Processing 64*(24), 6459–6470.

Chernozhukov, V., W. K. Härdle, C. Huang, and W. Wang (2021). Lasso-driven inference in time and space. *The Annals of Statistics 49*(3), 1702–1735.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance 1*(2), 223–236.

Doukhan, P. and S. Louhichi (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic processes and their applications 84*(2), 313–342.

Fan, J., Q. Li, and Y. Wang (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 247–265.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fan, J., J. Lv, and L. Qi (2011). Sparse high dimensional models in economics. *Annual review of economics 3*, 291–317.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity 1*(1), 13–36.

Gourieroux, C. and J. Jasiak (2011). *Financial Econometrics: Problems, Models, and Methods.* Princeton University Press.

Guo, S., Y. Wang, and Q. Yao (2016). High-dimensional and banded vector autoregressions. *Biometrika 103*(4), 889–903.

Gupta, S. (2012). A note on the asymptotic distribution of lasso estimator for correlated data. *Sankhya A 74*(1), 10–28.

Han, F., H. Lu, and H. Liu (2015). A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research 16*, 3115–3150.

Hang, H. and I. Steinwart (2017). A bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics 45*(2), 708–743.

Harvey, A. C. (1990). *The econometric analysis of time series.* Mit Press.

Hsu, N.-J., H.-L. Hung, and Y.-M. Chang (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis 52*(7), 3645–3657.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics 35*(1), 73–101.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics 1*(5), 799–821.

Kaul, A. (2014). Lasso with long memory regression errors. *Journal of Statistical Planning and*

Inference 153, 11–26.

Kim, Y., R. Giacometti, S. Rachev, F. Fabozzi, and D. Mignacca (2012). Measuring financial risk and portfolio optimization with a non-gaussian multivariate model. *Annals of operations research 201*(1), 325–343.

Kock, A. B. and L. Callot (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics 186*(2), 325–344.

Kondrashov, D., S. Kravtsov, A. W. Robertson, and M. Ghil (2005). A hierarchy of data-based enso models. *Journal of climate 18*(21), 4425–4444.

Koopman, S. J. and A. Lucas (2008). A non-gaussian panel time series model for estimating and decomposing default risk. *Journal of Business & Economic Statistics 26*(4), 510–525.

Ledoit, O. and M. Wolf (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance 10*(5), 603–621.

Liu, H., A. Aue, and D. Paul (2015). On the marčenko–pastur law for linear time series. *The Annals of Statistics 43*(2), 675–712.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics 45*(2), 866–896.

Loh, P.-L. (2021). Scale calibration for high-dimensional robust regression. *Electronic Journal of Statistics 15*(2), 5933–5994.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science &

Business Media.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics 17*(1), 382–400.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics 37*(1), 246–270.

Merlevède, F., M. Peligrad, E. Rio, et al. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pp. 273–292. Institute of Mathematical Statistics.

Nardi, Y. and A. Rinaldo (2011). Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis 102*(3), 528–549.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory 57*(10), 6976–6994.

Reinsel, G. C. (2003). *Elements of multivariate time series analysis*. Springer Science & Business Media.

Rothman, A. J., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association 104*(485), 177–186.

Sameshima, K. and L. A. Baccala (2014). *Methods in Brain Connectivity Inference Through Multivariate Time Series Analysis*. CRC press.

Shojaie, A., S. Basu, and G. Michailidis (2012). Adaptive thresholding for reconstructing regulatory networks from time-course gene expression data. *Statistics in Biosciences 4*(1), 66–83.

Shumway, R. H., D. S. Stoffer, and D. S. Stoffer (2000). *Time series analysis and its applications*, Volume 3. Springer.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica 48*(1), 1–48.

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review 36*(5), 975–1000.

Stock, J. H. and M. W. Watson (2001). Vector autoregressions. *Journal of Economic perspectives 15*(4), 101–115.

Sun, Q., W.-X. Zhou, and J. Fan (2020). Adaptive huber regression. *Journal of the American Statistical Association 115*(529), 254–265.

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

Tsay, R. S. (1984). Regression models with time series errors. *Journal of the American Statistical Association 79*(385), 118–124.

Tsay, R. S. (2005). *Analysis of financial time series*. John wiley & sons.

Wang, H., G. Li, and C.-L. Tsai (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical*

Methodology) 69(1), 63–78.

Wang, X., Y. Jiang, M. Huang, and H. Zhang (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association 108*(502), 632–643.

Wu, W.-B. and Y. N. Wu (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics 10*(1), 352–379.

Yohai, V. J. and R. A. Maronna (1979). Asymptotic behavior of m-estimators for the linear model. *The Annals of Statistics 7*(2), 258–268.

Yoon, Y. J., C. Park, and T. Lee (2013). Penalized regression models with autoregressive error terms. *Journal of Statistical Computation and Simulation 83*(9), 1756–1772.

Zhang, D. (2021). Robust estimation of the mean and covariance matrix for high dimensional time series. *Statistica Sinica 31*(2), 797–820.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

Department of Mathematics,

University of California, San Diego,

La Jolla, CA 92093

E-mail: linbo@ucsd.edu

E-mail: daz076@ucsd.edu