

**Statistica Sinica Preprint No: SS-2022-0249**

|                                 |   |
|---------------------------------|---|
| <b>Title</b>                    | A Bernstein-type Inequality for High Dimensional Linear Processes with Applications to Robust Estimation of Time Series Regressions |
| <b>Manuscript ID</b>            | SS-2022-0249  |
| <b>URL</b>                      | <a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a>                                   |
| <b>DOI</b>                      | 10.5705/ss.202022.0249  |
| <b>Complete List of Authors</b> | Linbo Liu and<br>Danna Zhang  |
| <b>Corresponding Authors</b>    | Danna Zhang   |
| <b>E-mails</b>                  | daz076@ucsd.edu   |

# **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 applications, data are often serially correlated and have an ultrahigh dimension and fat tails, making it difficult to develop new time series analysis tools. 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) a time series regression with fat-tailed and correlated covariates and errors, and (2) a fat-tailed vector autoregression. Our proposed approach allows for exponential increases in the dimension with the 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

The growing prevalence of massive data sets has increased the importance of high-dimensional data analysis, and particularly, high-dimensional linear regression. Specifically, consider the linear regression models

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

where  $Y_i$ ,  $X_i$ , and  $\xi_i$  are the response, covariate, and error variables, respectively. Various regularization methods have been used to estimate the  $p$ -dimensional regression parameter vector, including those of 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  $\xi_i$  are independent and identically distributed (i.i.d.) Gaussian or sub-Gaussian random variables, which can be too restrictive in practice.

On the one hand, serial correlation might occur when data are collected over time, requiring, for example, a linear regression with time series regressors and autoregressive errors (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, including those in quantitative finance (Cont (2001)), portfolio allocation (Kim et al. (2012)), risk management (Koopman and Lucas (2008)), brain networks (Friston (2011)), and geophysical dynamic studies (Kondrashov et al. (2005)).

Previous works have examined linear regression with correlated errors. Specifically, the Lasso estimator is 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 studies focus on cases in which the dimension  $p$  is smaller than the sample size  $n$ , or 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 required for the dimension in the presence of non-Gaussian and dependent errors. To address the ultrahigh-dimensional cases, where  $p$  can grow exponentially with  $n$ , various robust methods have been proposed for linear regression with i.i.d. fat-tailed errors, including the

penalized Huber  $M$ -estimation (Fan et al. (2017), Loh (2017, 2021)), sparse least trimmed squares (Alfons et al. (2013)), and ESL-Lasso (Wang et al. (2013)), among others. In this study, we consider a robust estimation of a time series regression, allowing for ultrahigh dimensions and fat-tailed and correlated errors.

Vector autoregression (VAR) is another popular linear model for describing the evolution of a set of variables over time, and there has been significant progress in estimating high-dimensional VAR models. Inspired by its development in high-dimensional linear regression, Hsu et al. (2008), Nardi and Rinaldo (2011), and Basu and Michailidis (2015) considered the Lasso estimator with an  $\ell_1$ -penalty. Kock and Callot (2015) established oracle inequalities for high-dimensional VAR models. Han et al. (2015) adopted a Dantzig-type penalization. Guo et al. (2016) proposed a Bayesian information criterion based on residual sums of the least squares estimator to estimate a high-dimensional banded autoregression. However, most of these studies require the Gaussian assumption or the existence of a finite exponential moment. In terms of 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 because of the estimation of a possibly large number of parameters. Therefore, there is a need to

investigate robust estimation methods for high-dimensional fat-tailed VAR models.

In summary, we focus on tackling the challenges posed by high-dimensional time series analysis with time series covariates, possibly correlated errors, fat tails, and an ultrahigh dimension. This requires new statistical tools tailored to the characteristics of these data sets. 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 is instrumental in obtaining consistent estimators under mild conditions, such as  $\log p = o(n^c)$ , for some  $c > 0$ .

The remainder of the paper is organized as follows. In Section 2, we introduce the framework of high-dimensional linear processes and the important quantities that 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) a time series linear regression with correlated and fat-tailed covariates and errors, and (2) autoregressive models with fat-tailed errors. We provide simulation results in Section 4 to assess the empirical performance of the robust estimators. All proofs are relegated to the Supplementary Material.

We first introduce some notation. For a vector  $\beta = (\beta_1, \dots, \beta_p)^\top$ , let

A Bernstein-type Inequality for High Dimensional Linear Processes

---

$|\beta|_1 = \sum_i |\beta_i|$ ,  $|\beta|_2 = (\sum_i \beta_i^2)^{1/2}$ ,  $|\beta|_0 = |\{i : \beta_i \neq 0\}|$ , and  $|\beta|_\infty = \max_i |\beta_i|$ . Let  $\text{Supp}(\beta)$  be the support of  $\beta$ . For a matrix  $A = (a_{ij})_{1 \leq i, j \leq p} \in \mathbb{R}^{p \times p}$ , let  $\lambda_i$ , for  $i = 1, \dots, p$ , 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  $|A|_1 = \sum_{i,j} |a_{ij}|$ ,  $\|A\|_1 = \max_j \sum_i |a_{ij}|$ ,  $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ , the spectral norm  $\|A\| = \|A\|_2 = \sup_{|x|_2 \neq 0} |Ax|_2 / |x|_2$ , and the Frobenius norm  $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$ . Moreover, let  $\text{tr}(A)$  be the trace of  $A$ ,  $\|A\|_{\max} = \max_{i,j} |a_{ij}|$  be the entry-wise maximum norm, and  $|A|$  be a matrix after taking the absolute value of  $A$ , that is,  $|A| = (|a_{ij}|)_{i,j}$ . 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  $a_n/b_n \leq C$  as  $n \rightarrow \infty$ , and write  $a_n \asymp b_n$  if  $a_n \lesssim b_n$  and  $b_n \lesssim a_n$ . We use  $c_0, c_1, \dots$  and  $C_0, C_1, \dots$  to denote universal positive constants, the values of which may vary in different contexts. Throughout the paper, we consider the high-dimensional regime, allowing the dimension  $p$  to grow with the sample size  $n$ , that is, we assume  $p = p_n \rightarrow \infty$  as  $n \rightarrow \infty$ .

---

## 2. Bernstein-type Inequality for High-dimensional Linear Processes

We consider a general framework of  $p$ -dimensional stationary linear processes

$$X_i = (X_{i1}, \dots, X_{ip})^\top = \mu + \sum_{k=0}^{\infty} A_k \varepsilon_{i-k}, \quad (2.1)$$

where  $\mu \in \mathbb{R}^p$  is the mean vector,  $A_0 = I_p$ ,  $A_k$ , for  $k \geq 1$ , are  $p \times p$  coefficient matrices with real entries such that  $\sum_{k=0}^{\infty} \text{tr}(A_k^\top A_k) < \infty$ ,  $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{ip})^\top$ , and  $\varepsilon_{ij}$ , for  $i \in \mathbb{Z}$ ,  $1 \leq j \leq p$ , are i.i.d. random variables with a zero mean and finite variance. Kolmogorov's three-series theorem ensures that the linear process (2.1) is well defined. Many researchers have 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 VAR model, which is 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  $\lim p^{-1}\text{tr}(\Pi)$  exists and is finite for any polynomial  $\Pi$  in  $\{A_k, A_k^\top\}$ , a joint convergence assumption that is difficult to verify. In this work, we 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

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

for all  $k \geq 0$ . This implies short-range dependence, in the sense that the autocovariance matrices  $\text{Cov}(X_0, X_j) = \sum_{k=0}^{\infty} A_k A_{k+j}^\top$  are absolutely summable. The proposed quantities  $\rho_p$  and  $\gamma_p$  can capture temporal and spatial dependence in the underlying high-dimensional process. In particular,  $\rho_p$  represents the strength of the temporal dependence, with smaller values indicating faster decay rates and weaker temporal dependence. The magnitude of  $\gamma_p$  naturally quantifies the spatial dependence. A notable feature is that both  $\gamma_p$  and  $\rho_p$  may depend on  $p$  in the high-dimensional regime. For example, when  $p$  is large,  $\rho_p$  may be a relatively large rate,

---

close to one, indicating a slow decay speed. In fact, there exists an absolute constant, independent of  $p$  and strictly smaller than one, such that (2.2) can be rewritten as

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

In particular, we define  $\tau_p \equiv 1$  if there exists  $\rho_0$  such that  $\rho_p \leq \rho_0 < 1$ , and  $\tau_p = \log \rho_0 / \log \rho_p$ , for  $\rho_0$  satisfying  $0 < \rho_0 \leq \rho_p$ , if  $\rho_p$  is large and increases with  $p$ . In the latter case, it could happen that  $\tau := \tau_p$  is an unbounded function in terms of the dimension  $p$ . Note that few studies have examined measures of dependence quantified by the dimension  $p$ , despite their relevance in analyzing high-dimensional time series. This feature is illustrated by the high-dimensional VAR model in Example 1. Henceforth, for notational simplicity, we omit the dimension subscript in  $\gamma_p, \tau_p$ , and refer to them as  $\gamma, \tau$ . In addition, 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 1* (High-dimensional VAR Models). Consider the VAR(1) model

$$X_i = AX_{i-1} + \varepsilon_i, \quad (2.4)$$

where  $A \in \mathbb{R}^{p \times p}$  is the transition matrix, and  $\varepsilon_i$ , for  $i \in \mathbb{Z}$ , are i.i.d. error vectors with mean zero and covariance matrix  $I_p$ . Equivalently, the mod-

---

el 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 consider condition (2.3), and it could happen that  $\tau$  may increase with the dimension  $p$ . Consider the design  $A = (a_{ij})_{i,j=1}^p$ , with  $a_{ij} = \lambda^{j-i+1} \mathbf{1}\{0 \leq j - i \leq B - 1\}$ , for some  $0 < \lambda < 1$  and  $1 \leq B \leq p$ . Here,  $B$  depicts how many variables, at most, in  $X_{i-1}$  have a spatial effect on  $X_{ij}$ . Figure 1 shows a plot of  $\|A^k\|$  under the numerical setup  $\lambda = 0.55$ ,  $B = 3, 4$ , and  $p = 20, 25, 30$ . As shown,  $\|A^k\|$  decays after a certain lag that moves forward as  $p$  increases. This lag can be defined as  $\tau$  in condition (2.3), so  $\tau$  increases with  $p$  in this design. Additionally, the geometric decay (its existence is shown later) occurs at a slow speed, which is further evidence of large  $\rho_p$  (or large  $\tau$ , equivalently). For example, when  $B = 3$  and  $p = 30$ ,  $\|A^k\|$  decreases 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  $\gamma$ , indicating the strength of spatial dependence. Comparing the two plots, we can see that stronger spatial dependence with a larger  $B$  results in a larger  $\gamma$ .

Concentration inequalities play an important role in the study of sums

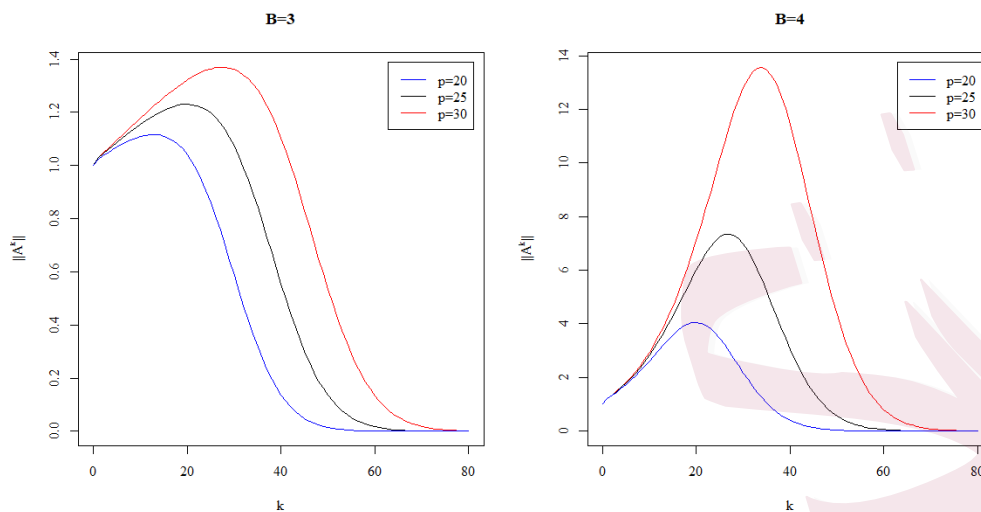


Figure 1: The graph of  $\|A^k\|$  for  $B = 3, 4$ , and  $p = 20, 25, 30$ .

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 a powerful tool for analyzing concentration behavior that provides an exponential inequality for sums of independent random variables that are uniformly bounded. For example, let  $Y_1, \dots, Y_n$  be i.i.d. random variables such that  $\mathbb{E}Y_i = 0$ ,  $\text{Var}(Y_i) = \sigma^2 < \infty$ , and  $|Y_i| \leq M$ , for all  $i$ . Then, for any  $x > 0$ , we have

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

which suggests two types of bound for the tail probability: a sub-Gaussian-type tail  $\exp\{-x^2/(Cn\sigma^2)\}$  in terms of the variance of  $\sum_{i=1}^n Y_i$ , and a sub-exponential-type tail  $\exp\{-x/(CM)\}$  in terms of the uniform bound  $M$ .

---

Bernstein-type inequalities have been developed for Markov chains and 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 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 let condition (2.3) be satisfied. Let  $G : \mathbb{R}^p \rightarrow \mathbb{R}$  be a function with  $|G(u)| \leq M$ , for all  $u \in \mathbb{R}^p$ . Suppose there exists a vector  $g = (g_1, \dots, g_p)^\top$  with  $g_i \geq 0$  and  $\sum_{i=1}^p g_i = 1$  such that the following Lipschitz condition holds: for all  $u = (u_1, \dots, u_p)^\top$  and  $v = (v_1, \dots, v_p)^\top$ ,*

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

Then, for any  $x > 0$ , we have

$$\mathbb{P}\left(\sum_{i=1}^n G(X_i) - \mathbb{E}G(X_i) \geq x\right) \leq 2 \exp\left\{-\frac{x^2}{C_1 n \sigma^2 \tau^2 \gamma^2 + C_2 \tau M x}\right\}, \quad (2.7)$$

where the constants  $C_1$  and  $C_2$  are given by

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

---

*Remark 2.* Equipped with our new inequality (2.7), we 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  $\tau$  and  $\gamma$ , respectively. In the special case that the processes are one-dimensional, denoted by  $X_i \in \mathbb{R}$ , and  $\tau = 1$  and  $\gamma$  is of a constant order that satisfies condition (2.2), our probability inequality (2.7) is just as sharp as the classical Bernstein inequality (2.5). Note that our inequality is strictly sharper than the Bernstein-type inequalities for univariate time series established by Merlevède et al. (2009) and Zhang (2021). Recall that Merlevède et al. (2009) derived a concentration inequality for a univariate strong mixing process  $(X_i)$  with mean zero and upper bounded by  $M$  in magnitude:

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

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 made 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 transfor-

---

mations is linear combinations of transformed component processes, that is,  $G(X_i) = \sum_{j=1}^p a_j h_j(X_{ij})$ , where  $\sum_{j=1}^p |a_j| = 1$ ,  $h_j : \mathbb{R} \rightarrow \mathbb{R}$  are univariate functions satisfying  $|h_j(x)| \leq M$  and  $|h_j(x) - h_j(y)| \leq 1$ , for any  $x, y \in \mathbb{R}$ , and 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 when 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  $\ell_\infty$ -norm convergence for the robust mean estimator, which is as sharp as the optimal rate for i.i.d. processes.

Condition (2.3) requires that  $\|A_k\|$  decays geometrically up to the quantity  $\gamma$ , and that the decay speed is controlled by  $\tau$ . 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, an exponential-type probability inequality does not hold, in general, even if it is a one-dimensional process with a uniform bound. That is, if we relax 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 1, we assume the existence of a finite variance of  $\varepsilon_{ij}$ . If this is relaxed to the existence of a finite exponential moment, a similar bound can be achieved with  $G$  not necessarily bounded; see Theorem 2.

**Theorem 2.** *In 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*

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

where the constants  $C_3$  and  $C_4$  depend on  $\rho_0$  and  $c_0$ .

One immediate application of Theorem 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  $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_p)^\top$ , 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 3.** *Let  $X_i$  be generated from model (2.1), with  $\mathbb{E}\varepsilon_{ij} = 0$ ,  $\text{Var}(\varepsilon_{ij}) = 1$ ,  $\mu = \mathbb{E}X_i$ , and  $\max_{1 \leq j \leq p} \text{Var}(X_{ij}) = \mu_2^2 < \infty$ . Choose  $\nu \asymp \mu_2 \sqrt{n/\log p}$ .*

*With probability at least  $1 - 4p^{-c}$ , for some  $c > 0$ , it holds that*

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



---

under the scaling condition  $(\gamma + \mu_2)\tau\sqrt{\log p/n} \rightarrow 0$ , where  $C$  is a positive constant depending on  $c$  and the constants  $C_1, C_2$  in Theorem 1.

*Remark 3.* Theorem 3 delivers a rate of  $\ell_\infty$ -norm convergence for the robust mean estimator  $\hat{\mu}$ , and involves a delicate interplay between the cross-sectional dependence, temporal dependence, and the variance of the process.

If  $\gamma, \mu_2$ , and  $\tau$  are all of a constant order, it follows that

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

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

### 3. Robust Estimation of Time Series Regression

In this section, we investigate a robust estimation of a high-dimensional time series linear regression and autoregression with fat-tailed covariates and errors. However, we expect our framework of high-dimensional linear processes and Bernstein-type inequalities to be useful in other high-

### 3.1 Estimating Time Series Regression with Correlated Errors

---

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 a random design that involve time dependent covariates and errors:

$$Y_i = X_i^\top \beta^* + \xi_i, \quad (3.1)$$

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  $\text{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2 < \infty$ . Condition (2.3) is satisfied with  $\gamma$  and  $\tau$ ,

which may depend on  $p$ .

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

with  $\mathbb{E}(\eta_i) = 0$  and  $\text{Var}(\eta_i) = \sigma_\eta^2 < \infty$ , 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  $\xi_i$ , respectively, as

defined in (A1) and (A2).

### 3.1 Estimating Time Series Regression with Correlated Errors

---

The framework (3.1) is quite general, because the linear process includes a wide range of commonly used time series models. For linear regression models with dependent errors, early works focused on a fixed design or i.i.d. covariates. Wang et al. (2007) and Yoon et al. (2013) considered the case where  $\xi_i$  follows an autoregressive process, which is one type of linear process. Gupta (2012) examined the weakly dependent  $\xi_i$  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  $X_i = g(\dots, \varepsilon_{i-1}, \varepsilon_i)$  and  $\xi_i = h(\dots, \eta_{i-1}, \eta_i)$ .

We form a modified  $\ell_1$ -regularized Huber estimator of  $\beta$ , 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  $\Phi_{\nu}$  is Huber loss function (Huber (1964))

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

defined with respect to the robustification parameter  $\nu > 0$ . For additional properties of the Huber regression, refer to Huber (1973), Yohai and Maronna (1979), Mammen (1989), Sun et al. (2020), and Fan et al. (2017), among

### 3.1 Estimating Time Series Regression with Correlated Errors

others. Motivated by Loh (2021),  $w(x) : \mathbb{R}^p \rightarrow \mathbb{R}$  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  $|w(x)x|_2 \leq b/\lambda_{\min}(B) =: b_0$ . In contrast to the regular Huber regression for well-behaved  $X_i$  (e.g., Gaussian or sub-Gaussian), we incorporate an additional weight function 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 examine the effect of  $w(x)$ .

As a popular convention,  $\beta^*$  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 4 describes the estimation consistency of  $\hat{\beta}$ .

**Theorem 4.** *Let Assumptions (A1), (A2), and (A3) be satisfied. Assume*

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

*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*

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

### 3.1 Estimating Time Series Regression with Correlated Errors

---

The scaling condition (3.2) to ensure consistency indicates a subtle interplay between the dimensionality parameters  $(s, p, n)$ , internal parameters  $(\tau, \gamma, \sigma_\varepsilon)$ , and 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 suggests that we cannot shrink the covariates too aggressively. If  $X_i$  is well behaved, with the existence of a finite exponential moment, one may eliminate the weight function and replace the factor with 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) provide an optimal convergence rate of  $|\hat{\beta} - \beta|_2$  for a weakly sparse model under 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 that the errors be drawn i.i.d. from a symmetric distribution, and thus selected  $\nu$  at a fixed constant order (cf. Theorem 1). In contrast, Fan et al. (2017) waived the symmetry requirement by allowing  $\nu$  to diverge in order to reduce the bias induced by the Huber loss when the distribution of  $\xi_i$  is asymmetric. We borrow ideas from both, and further account for

### 3.1 Estimating Time Series Regression with Correlated Errors

---

time-dependent covariates and errors. Compared with 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  $\gamma, \tau$  and a larger power of  $\log p$  to handle both dependent covariates and dependent errors.

Applying  $\ell_1$ -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 the finite moments for  $\xi_i$ . 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 considering the performance bound with respect to the prediction norm (cf. Corollary 5.4). Compared with the two aforementioned works, we allow a much wider range for the dimension  $p$  under mild conditions.

The tuning parameter  $\nu$  plays a key role by adapting to errors with fat tails. In practical applications, the optimal values of the tuning parameters  $\nu$  and  $\lambda$  can be chosen using a two-dimensional grid search and

---

### 3.2 Estimating Transition Matrix in VAR Models

cross-validation or an information-based criterion such as the AIC or BIC. We leave the theoretical investigation of selecting the tuning parameters as important future work.

### 3.2 Estimating Transition Matrix in VAR Models

VAR models are popular for studying the evolution of a set of endogenous variables over time. Interpretations of large VAR 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)).

Because a general VAR model of order  $d$  can be reformulated as a VAR(1) model by appropriately redefining the random vectors, many works (Han et al. (2015), Guo et al. (2016)) consider a model with lag 1, as given in (2.4). Most works on high-dimensional VAR models 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 1 is one such), it could

### 3.2 Estimating Transition Matrix in VAR Models

---

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 fill the gap between the spectral radius of a matrix and its spectral norm. The following proposition provides a sufficient and necessary condition for  $\lambda_{\max}(A) < 1$  by relating to the spectral norm.

**Proposition 1.** *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 model (2.4) without extra requirements. We now introduce the notation. Let  $\mathbf{a}_j^\top$  be the  $j$ th row of  $A$  and  $s_j$  be the cardinality of the support set of  $\mathbf{a}_j$ , that is,  $s_j = |\text{supp}(\mathbf{a}_j)| = |\{i : a_{ij} \neq 0\}|$ . Denote  $s = \max_{1 \leq j \leq p} s_j$  and  $\mathcal{S} = \sum_{i=1}^p s_j$ . For robustness, we first truncate the data by obtaining  $\tilde{X}_i = \phi_\nu(X_i)$ , where  $\nu$  is the truncation parameter, determined later. For notational convenience, we assume  $X_0$  is also observed. Based on the truncated sample  $\tilde{X}_i$  and the tuning parameter  $\lambda > 0$ , we propose



### 3.2 Estimating Transition Matrix in VAR Models

estimating  $A$  by solving the following Lasso problem:

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

which is equivalent to solving the  $p$  sub-problems:

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

Before proceeding, we state the key assumptions on the process (2.4) and some scaling conditions that guarantee the 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)  $\mu_q \gamma \tau s^2 [(\log p)/n]^{(q-2)/(2q-2)} \rightarrow 0$ .

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

Assumption (B1) imposes polynomial moment conditions on the underlying VAR process. Assumption (B2) (or (B2')) assumes a vanishing scaling property. If  $\mu_q$ ,  $\tau$ , and  $\gamma$  are of a constant order, (B2) reduces to the scaling condition that involves  $s$  (or  $\mathcal{S}$ ),  $n$ , and  $p$  only.

**Theorem 5.** *Let Assumptions (B1) and (B2) be satisfied. Choose the truncation parameter  $\nu \asymp \mu_q (n/\log p)^{1/(2q-2)}$ . Let  $\hat{A}$  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*

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

### 3.2 Estimating Transition Matrix in VAR Models

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

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

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 sample size  $n$ ; (ii) internal parameters: the moment  $\mu_q$ , dependence quantities  $\tau$  and  $\gamma$ , and maximum absolute row sum  $\|A\|_\infty$ . If we assume that the internal parameters are of a constant order, we have

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

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

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

In the presence of fat tails and with the existence of a finite  $q$ th moment,

### 3.2 Estimating Transition Matrix in VAR Models

our result yields a slightly slower convergence rate, characterized by the moment order  $q$ , and it becomes closer to their bound as  $q$  increases.

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

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

The Dantzig-type estimator is then modified to solve the following convex programming problem:

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

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

$$\hat{\mathbf{a}}_{\cdot j} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} |\mathbf{b}|_1 \text{ s.t. } |\hat{\Sigma}_0 \mathbf{b} - \hat{\Sigma}_1 u_j|_\infty \leq \lambda, \quad j = 1, \dots, p, \quad (3.9)$$

### 3.2 Estimating Transition Matrix in VAR Models

for any unit vector  $u_j$ . Let  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p$  be columns of  $A$ , and denote  $s^* = \max_{1 \leq j \leq p} |\text{supp}(\mathbf{a}_j)|$ . We can obtain  $\widehat{A}$  by simply concatenating all the columns  $\widehat{\mathbf{a}}_j$ , that is,  $\widehat{A} = (\widehat{\mathbf{a}}_1, \widehat{\mathbf{a}}_2, \dots, \widehat{\mathbf{a}}_p)$ . The next theorem delivers an upper bound on the statistical accuracy.

**Theorem 6.** *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*

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

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

Interestingly, the convergence rate of the modified Dantzig-type estimator has a similar form to that of the robust Lasso estimator developed in Theorem 5, if the included internal parameters for the process are of a constant order. Both methods involve  $p$  parallel programming problems, with the lasso-based method performing a row-by-row estimation, and the Dantzig method performing a column-by-column estimation. The case of  $\|A\| < 1$  studied by Han et al. (2015) is the special case where  $\gamma = 1$  and  $\tau = 1$  in our framework. The latter work imposes a more flexible sparse condition, namely, that the transition matrix  $A$  belongs to a class of weakly sparse matrices defined in terms of a strong  $\ell^r$ -ball ( $0 \leq r < 1$ ). This condi-

---

tion is also considered by Bickel and Levina (2008), Rothman et al. (2009), Cai et al. (2011), and Cai and Zhou (2012) when 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

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

which is a bit sharper than our result (3.11). There is an additional cost for fat-tailed processes with robustness absorbed. Note that we are also able to derive the bound of  $\|\hat{A} - A\|_1$  for weakly sparse  $A$  based on the result (3.10).

#### 4. Simulation Study

In this section, we evaluate the finite-sample performance of the robust Lasso and Dantzig estimators proposed in Section 3.2, and compare it with that of the traditional Lasso and Dantzig methods. A simulation on time series linear regression is presented in the Supplementary Material. We consider the model (2.4), where  $\varepsilon_{ij}$  are i.i.d. standardized Student's  $t$ -distributions with  $\text{df} = 5$ . We adopt 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 1 with  $B = 2$  and  $\lambda_i \sim \text{Unif}(-0.8, 0.8)$ .
  - (3) Toeplitz:  $A = (\lambda^{|i-j|})$  and  $\lambda = 0.5$ .
  - (4) Random Sparse:  $a_{ii} \sim \text{Unif}(-0.8, 0.8)$  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) are further scaled by a factor of  $2\lambda_{\max}(A)$  to ensure that the spectral radius of the transition matrix is less than one. Figure 2 shows the plot of  $\|A^k\|$  under the four designs, with  $p = 100, 500$ . These patterns of matrix  $A$  were studied previously 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, 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 (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  $\tau$  and  $\gamma$  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  $\{X_1, \dots, X_n\}$  using a two-

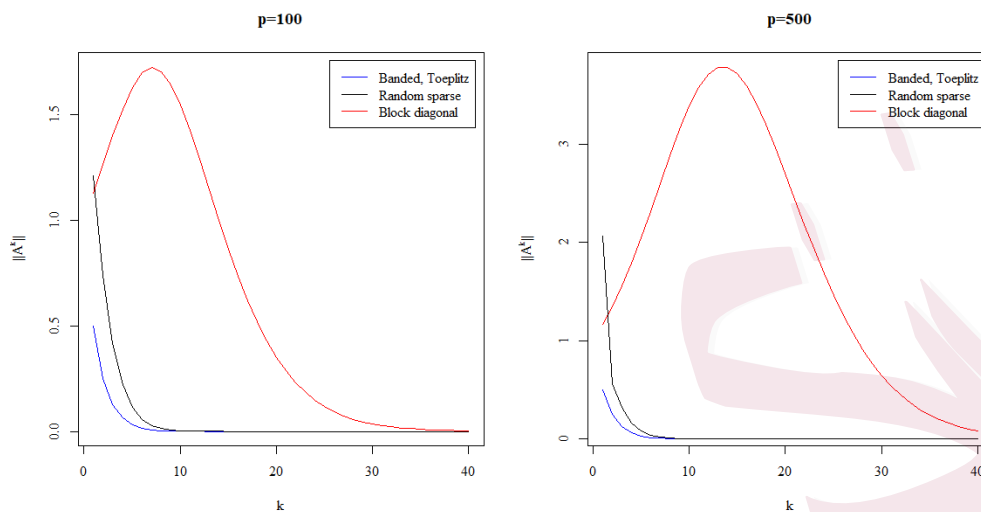


Figure 2: The graph of  $\|A^k\|$  for the four designs of  $A$ , with  $p = 100, 500$  dimensional grid search for the tuning parameters  $\nu$  and  $\lambda$ . For each  $(\nu, \lambda)$  in the grid, denote the estimator by  $\hat{A}(\nu, \lambda)$ . Then,  $(\nu, \lambda)$  is chosen to minimize  $n^{-1} \sum_{t=n+1}^{2n} |X_t - \hat{A}(\nu, \lambda)X_{t-1}|_2^2$ , the average prediction error on  $\{X_{n+1}, \dots, X_{2n}\}$ . The following tables report the average and standard deviation (in parentheses) of the estimation error based on 1000 repetitions in different matrix norms consistent with Theorem 5 and Theorem 6. As comparisons, we obtain the results for the robust methods and the traditional versions (Lasso estimator in Tibshirani (1996) and Dantzig-based estimator in Han et al. (2015)) in different designs.

From a statistical perspective, the tables indicate that both robust estimation methods outperform the regular Lasso and Dantzig when the inno-

| $p = 50, n = 100$       | Banded       | Block        | Toeplitz     | Random       |
|-------------------------|--------------|--------------|--------------|--------------|
| Lasso $L_\infty$        | 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 $L_\infty$ | 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) |

| $p = 100, n = 50$       | Banded       | Block        | Toeplitz     | Random       |
|-------------------------|--------------|--------------|--------------|--------------|
| Lasso $L_\infty$        | 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 $L_\infty$ | 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 (0.065) | 2.45 (0.090) |



| $p = 500, n = 100$      | Banded       | Block        | Toeplitz     | Random       |
|-------------------------|--------------|--------------|--------------|--------------|
| Lasso $L_\infty$        | 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 $L_\infty$ | 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) |

vation vectors have a fat tail and the transition matrix exhibits a sparsity pattern. In summary, our robust methods work particularly well for non-Gaussian time series.

## 5. Conclusion

Conventional time series regression tools are inadequate when analyzing high-dimensional temporal-dependent and fat-tailed data. In this paper, we have proposed a novel Bernstein inequality for high-dimensional linear processes, thus contributing to the robust estimation theory of high-dimensional time series regression in the presence of fat tails. The convergence rate depends on the strength of the temporal and cross-sectional

## REFERENCES

---

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. A statistical inference of the estimates, such as hypothesis testing and constructing confidence intervals, requires additional research in terms of asymptotic distributional theory. This is left to future work.

### Supplementary Material

The online Supplementary Material contains a simulation on time series regression and the proofs of all the results presented in the paper.

### Acknowledgements

We thank two anonymous referees, an Associate Editor and the Editor for their helpful comments that have improved the paper. The work of Danna Zhang was supported by the NSF grant DMS-1916290.

### 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.

## REFERENCES

---

- 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  $\ell_1$  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.

## REFERENCES

---

- 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*

## REFERENCES

---

- 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. *Ann. Stat.* 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*

## REFERENCES

---

- 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*

---

## REFERENCES

- 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  $\ell_q$ -balls. *IEEE transactions on information theory* 57(10), 6976–6994.
- Reinsel, G. C. (2003). *Elements of multivariate time series analysis*. Springer.
- 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*

## REFERENCES

---

- 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



## REFERENCES

---

- 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. *Electron. J. Stat.* 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