HHS Public Access

Author manuscript

JMLR Workshop Conf Proc. Author manuscript; available in PMC 2017 January 25.

Published in final edited form as:

JMLR Workshop Conf Proc. 2015 July ; 37: 1843–1851.

Robust Estimation of Transition Matrices in High Dimensional Heavy-tailed Vector Autoregressive Processes

Huitong Qiu,
Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21210 USA

Sheng Xu,
Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21210 USA

Fang Han,
Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21210 USA

Han Liu,
Princeton University, 98 Charlton Street, Princeton, NJ 08544 USA

Brian Caffo
Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21210 USA

Huitong Qiu: HQIU7@JHU.EDU; Sheng Xu: SHXU@JHU.EDU; Fang Han: FHAN@JHU.EDU; Han Liu: HANLIU@PRINCETON.EDU; Brian Caffo: BCAFFO@JHU.EDU

Abstract

Gaussian vector autoregressive (VAR) processes have been extensively studied in the literature. However, Gaussian assumptions are stringent for heavy-tailed time series that frequently arises in finance and economics. In this paper, we develop a unified framework for modeling and estimating heavy-tailed VAR processes. In particular, we generalize the Gaussian VAR model by an elliptical VAR model that naturally accommodates heavy-tailed time series. Under this model, we develop a quantile-based robust estimator for the transition matrix of the VAR process. We show that the proposed estimator achieves parametric rates of convergence in high dimensions. This is the first work in analyzing heavy-tailed high dimensional VAR processes. As an application of the proposed framework, we investigate Granger causality in the elliptical VAR process, and show that the robust transition matrix estimator induces sign-consistent estimators of Granger causality. The empirical performance of the proposed methodology is demonstrated by both synthetic and real data. We show that the proposed estimator is robust to heavy tails, and exhibit superior performance in stock price prediction.

1. Introduction

Vector autoregressive models are widely used in analyzing multivariate time series. Examples include financial time series (Tsay, 2005), macroeconomic time series (Sims, 1980), gene expression series (Fujita et al., 2007; Opgen-Rhein & Strimmer, 2007), and functional magnetic resonance images (Qiu et al., 2015).

Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T \in \mathbb{R}^d$ be a stationary multivariate time series. We consider VAR models[1] such that

$$\boldsymbol{X}_t = \mathbf{A}\boldsymbol{X}_{t-1} + \boldsymbol{E}_t \text{ for } t = 2, \ldots, T,$$

where $\mathbf{A}$ is the transition matrix, and $\boldsymbol{E}_2, \ldots, \boldsymbol{E}_T$ are latent innovations. The transition matrix characterizes the dependence structure of the VAR process, and plays a fundamental role in forecasting. Moreover, the sparsity pattern of the transition matrix is often closely related to Granger causality. In this paper, we focus on estimating the transition matrix in high dimensional VAR processes.

VAR models have been extensively studied under the Gaussian assumption. The Gaussian VAR model assumes that the latent innovations are i.i.d. Gaussian random vectors, and are independent from past observations (Lütkepohl, 2007). Under this model, there is vast literature on estimating the transition matrix under high dimensional settings. These estimators can be categorized into regularized estimators and Dantzig-selector-type estimators. The former can be formulated by

$$\hat{\mathbf{A}}^{\text{reg}} := \underset{\mathbf{M} \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \; l(\mathbf{Y} - \mathbf{M}\mathbf{X}) + P_\rho(\mathbf{M}), \tag{1.1}$$

where $\mathbf{Y} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{T-1}) \in \mathbb{R}^{d \times (T-1)}$, $\mathbf{X} := (\boldsymbol{X}_2, \ldots, \boldsymbol{X}_T) \in \mathbb{R}^{d \times (T-1)}$, $l(\cdot)$ is a loss function, and $P_\rho(\cdot)$ is a penalty function with penalty parameter ρ. Common choices of the loss function include least squares loss and negative log-likelihood (Hamilton, 1994). For the penalty function, various ℓ_1 penalties (Wang et al., 2007; Hsu et al., 2008; Shojaie & Michailidis, 2010) and ridge penalty (Hamilton, 1994) are widely used. Theoretical properties of ℓ_1 penalized estimators are studied in Narki & Rinaldo (2011), Song & Bickel (2011), and Basu & Michailidis (2013).

In parallel to the penalized minimum loss estimators, Han & Liu (2013) proposed a Dantzig-selector-type estimator, which is formulated as the solution to a linear programming problem. In contrast to the ℓ_1 regularized estimators, consistency of the Dantzig-selector-type estimator do not rely on restricted eigenvalue conditions. These conditions do not explicitly account for the effect of serial dependence. Moreover, the Dantzig-selector-type estimator weakens the sparsity assumptions required by the ℓ_1 regularized estimators.

Although extensively studied in the literature, Gaussian VAR models are restrictive in their implications of light tails. Heavy-tailed time series frequently arise in finance, macroeconomics, signal detection, and statistical physics, to name just a few (Feldman & Taqqu, 1998). For analyzing these data, more flexible models and robust estimators are desired.

In this paper, we develop a unified framework for modeling and estimating heavy-tailed VAR processes. In particular, we propose an elliptical VAR model that allows for heavy-

[1]For simplicity, we only consider order one VAR models in this paper. Extensions to higher orders can be obtained using the same technique as in Chapter 2.1 of Lütkepohl (2007).

tailed processes. The elliptical VAR model covers the Gaussian VAR model as a special case. Under this model, we show that the transition matrix is closely related to quantile-based scatter matrices. The relation serves as a quantile-based counterpart of the Yule-Walker equation[2] (Lütkepohl, 2007). Motivated by this relation, we propose a quantile-based robust estimator of the transition matrix. The estimator falls into the category of Dantzig-selector-type estimators, and enjoys similar favorable properties as the estimator in Han & Liu (2013). We investigate the asymptotic behavior of the estimator in high dimensions, and show that although set in a more general model, it achieves the same rates of convergence as the Gaussian-based estimators. The effect of serial dependence is also explicitly characterized in the rates of convergence.

As an application of the framework developed in this paper, we investigate Granger causality estimation under the elliptical VAR process. We show that just as in Gaussian VAR models, Granger causality relations are also captured by the sparsity patterns of the transition matrix. The robust transition matrix estimator developed in this paper induces sign-consistent estimators of these relations.

2. Background

In this section, we introduce the notation employed in this paper, and provide a review on elliptical distributions and robust scales. Elliptical distributions provide a basis for our model, while robust scales motivate our methodology.

2.1. Notation

Let $\mathbf{v} = (\upsilon_1, \ldots, \upsilon_d)^\top$ be a d-dimensional real vector, and $\mathbf{M} = [\mathbf{M}_{jk}] \in \mathbb{R}^{d_1 \times d_2}$ be a $d_1 \times d_2$ matrix with $\mathbf{M}_{jk}$ as the (j, k) entry. Denote by $\mathbf{v}_I$ the subvector of $\mathbf{v}$ whose entries are index by a set $I \subset \{1, \ldots, d\}$. Similarly, denote by $\mathbf{M}_{U,V}$ the submatrix of $\mathbf{M}$ whose entries are indexed by $U \subset \{1, \ldots, d_1\}$ and $V \subset \{1, \ldots, d_2\}$. Let $\mathbf{M}_{U,*} = \mathbf{M}_{U,\{1, \ldots, d_2\}}$. For $0 < q < \infty$, we define the vector ℓ_q norm of $\mathbf{v}$ as $\|\mathbf{v}\|_q := (\sum_{j=1}^{d} |\upsilon_j|)^{1/q}$, and the vector ℓ_∞ norm of $\mathbf{v}$ as $\|\mathbf{v}\|_\infty := \max_{j=1}^{d} |\upsilon_j|$. Let the matrix $\ell_{\max}$ norm of $\mathbf{M}$ be $\|\mathbf{M}\|_{\max} := \max_{jk} |M_{jk}|$, the matrix ℓ_∞ norm be $\|\mathbf{M}\|_\infty := \max_j \sum_{k=1}^{d} |\mathbf{M}_{jk}|$, and the Frobenius norm be $\|\mathbf{M}\|_F := \sqrt{\sum_{jk} M_{jk}^2}$. Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\top$ be two random vectors. We write $\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{Y}$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed. We use $\mathbf{0}, \mathbf{1}, \ldots$ to denote vectors with 0, 1, … at every entry.

2.2. Elliptical Distribution

Definition 2.1 (Fang et al. (1990))—*A random vector $\boldsymbol{X} \in \mathbb{R}^d$ follows an elliptical distribution with location $\boldsymbol{\mu} \in \mathbb{R}^d$ and scatter $\mathbf{S} \in \mathbb{R}^{d \times d}$ if and only if there exists a nonnegative random variable $\xi \in \mathbb{R}$, a rank k matrix $\mathbf{R} \in \mathbb{R}^{d \times k}$ with $\mathbf{S} = \mathbf{R}\mathbf{R}^\top$, a random vector $\mathbf{U} \in \mathbb{R}^k$ independent of ξ and uniformly distributed in the k dimensional sphere, $\mathbb{S}^{k-1}$, such that*

[2]The Yule-Walker equation connects the transition matrix with the covariance matrix and the lag-one autocovariance matrix of the process.

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \xi \mathbf{R}\boldsymbol{U}\,. \tag{2.1}$$

In this case, we denote $\boldsymbol{X} \sim \mathrm{EC}_d(\boldsymbol{\mu}, \mathbf{S}, \xi)$. $\mathbf{S}$ *is called the scatter matrix, and* ξ *is called the generating variate*.

Remark 2.2—(2.1) is often referred to as the stochastic representation of the elliptical random vector $\boldsymbol{X}$. Of note, by Theorem 2.3 in Fang et al. (1990) and the proof of Theorem 1 in Cambanis et al. (1981), Definition 2.1 is equivalent if we replace "$\overset{\mathrm{d}}{=}$" with simply "=".

Proposition 2.3 (Theorems 2.15 and 2.16 in Fang et al. (1990))—*Suppose* $\boldsymbol{X} \sim \mathrm{EC}_d(\boldsymbol{\mu}, \mathbf{S}, \xi)$ *and* $\mathrm{rank}(\mathbf{S}) = k$. *Let* $\mathbf{B} \in \mathbb{R}^{p \times d}$ *be a matrix and* $\boldsymbol{\nu} \in \mathbb{R}^p$ *be a vector. Denote* $l = \mathrm{rank}(\mathbf{BSB}^\top)$. *Then, we have*

$$\boldsymbol{\nu} + \mathbf{B}\boldsymbol{X} \sim \mathrm{EC}_p(\boldsymbol{\nu} + \mathbf{B}\boldsymbol{\mu}, \mathbf{BSB}^\top, \xi\sqrt{B}),$$

where $B \sim \mathrm{Beta}(l/2, (k-l)/2)$ *follows a Beta distribution if* $k > l$, *and* $B = 1$ *if* $k = l$.

2.3. Robust Scales

Let $X \in \mathbb{R}$ be a random variable with a sequence of observations $X_1, \ldots, X_T$. Denote F as the distribution function of X. For a constant $q \in [0, 1]$, we define the q-quantiles of X and $\{X_t\}_{t=1}^T$ to be

$$Q(X; q) = Q(F; q) \coloneqq \inf\{x : \mathbb{P}(X \le x) \ge q\},$$

$$\widehat{Q}(\{X_t\}_{t=1}^T; q) \coloneqq X^{(k)} \text{ where } k = \min\left\{t : \frac{t}{T} \ge q\right\}.$$

Here $X^{(1)} \le \cdots \le X^{(T)}$ are the order statistics of the sample $\{X_t\}_{t=1}^T$. We say $Q(X; q)$ is unique if there exists a unique x such that $\mathbb{P}(X \le x) = q$. We say $\widehat{Q}(\{X_t\}_{t=1}^T; q)$ is unique if there exists a unique $X \in \{X_t\}_{t=1}^T$ such that $X = X^{(k)}$. Following Rousseeuw & Croux (1993), we define the population and sample quantile-based scales as

$$\sigma^{\mathrm{Q}}(X) \coloneqq Q(|X - \widetilde{X}|; 1/4), \tag{2.2}$$

$$\hat{\sigma}^{\mathrm{Q}}(\{X_t\}_{t=1}^T) \coloneqq \widehat{Q}(\{|X_s - X_t|\}_{1 \le s \le t \le T}; 1/4),$$

where $\widetilde{X}$ is an independent copy of X. $\hat{\sigma}^{\mathrm{Q}}(\{X_t\}_{t=1}^T)$ can be computed using $O(T \log T)$ time and $O(T)$ storage (Rousseeuw & Croux, 1993).

3. Model

In this paper, we model the time series of interest by an elliptical VAR process.

Definition 3.1

A sequence of observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T \in \mathbb{R}^d$ *is an elliptical VAR process if and only if the following conditions are satisfied*:

1. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ *follow a lag-one VAR process*

$$\boldsymbol{X}_t = \mathbf{A}\boldsymbol{X}_{t-1} + \boldsymbol{E}_t, \quad for\ t = 2, \ldots, T, \tag{3.1}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ *is the transition matrix, and* $\boldsymbol{E}_2, \ldots, \boldsymbol{E}^T \in \mathbb{R}^d$ *are latent innovations*.

2. $\{(\boldsymbol{X}_t^\top, \boldsymbol{E}_{t+1}^\top)^\top\}_{t=1}^{T-1}$ *are stationary and absolutely continuous elliptical random vectors*:

$$\begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{E}_{t+1} \end{pmatrix} \sim \mathrm{EC}_{2d}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{pmatrix}, \xi\right), \tag{3.2}$$

where $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\Psi}$ *are positive definite matrices, and* $\xi > 0$ *with probability 1.*

Remark 3.2

The elliptical VAR process in Definition 3.1 can be generated by an iterative algorithm following Rémillard et al. (2012). In detail, by the property of elliptical distributions, the density function of $(\boldsymbol{X}_t^\top, \boldsymbol{E}_{t+1}^\top)^\top$ can be written by $h(\mathbf{x}, \mathbf{e}) = 1/\sqrt{|\boldsymbol{\Sigma}||\boldsymbol{\Psi}|}\, g(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{e}^\top \boldsymbol{\Psi}^{-1}\mathbf{e})$ for some function g, and the density function of $\boldsymbol{X}_t$ and the conditional density function of $\boldsymbol{E}_{t+1}$ given $\boldsymbol{X}_t$ can be written by

$$h_1(\mathbf{x}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} g_1(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x})$$

$$\text{and } h_2(\mathbf{e}|\mathbf{x}) = \frac{1}{\sqrt{|\boldsymbol{\Psi}|}} g_2(\mathbf{e}^\top \boldsymbol{\Psi}^{-1}\mathbf{e}),$$

where g_1 and g_2 are defined by

$$g_1(r) = \int_{\mathbb{R}^d} g(\|\mathbf{z}\|_2^2 + r)d\mathbf{z} \text{ and } g_2(r) = \frac{g(r + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x})}{g_1(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x})}.$$

The elliptical VAR process $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ can be generated by the following algorithm:

1. Generate $\boldsymbol{X}_1$ from $h_1(\mathbf{x})$.
2. For $t = 2, \ldots, T$,

a. generate $\boldsymbol{E}_t$ from $h_2(\mathbf{e} \mid \boldsymbol{X}_{t-1})$;

b. set $\boldsymbol{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{E}_t$.

Remark 3.3

By definition, it follows that an elliptical VAR process is a stationary process. A special case of the elliptical VAR process is the Gaussian VAR process. An elliptical VAR process is Gaussian VAR if (3.2) is replaced by

$$\begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{E}_{t+1} \end{pmatrix} \sim N_{2d}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{pmatrix}\right).$$

The elliptical VAR process generalizes the Gaussian VAR process in two aspects. First, the elliptical model generalizes the Gaussian model by allowing heavy tails. This makes robust methodologies necessary for estimating the process. Secondly, the elliptical VAR model does not require that the observations are independent from future latent innovations.

Next, we show that there exists an elliptical random vector $\boldsymbol{L} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top, \boldsymbol{E}_2^\top, \ldots, \boldsymbol{E}_T^\top)^\top$ such that the two conditions in Definition 3.1 are satisfied. To this end, let $\boldsymbol{L}_0 := (\boldsymbol{X}_1^\top, \boldsymbol{E}_2^\top, \ldots, \boldsymbol{E}_T^\top)^\top \sim \mathrm{EC}_{Td}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}, \ldots, \boldsymbol{\Psi}), \zeta)$ and define

$$\boldsymbol{L} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top, \boldsymbol{E}_2^\top, \ldots, \boldsymbol{E}_T^\top)^\top := \mathbf{B}\boldsymbol{L}_0, \tag{3.3}$$

where

$$\mathbf{B} := \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{A} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{A}^2 & \mathbf{A} & \mathbf{I} & \cdots & \mathbf{0} \\ & & \cdots & & \\ \mathbf{A}^{T-1} & \mathbf{A}^{T-2} & \mathbf{A}^{T-3} & \cdots & \mathbf{I} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ & & \cdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{pmatrix} \in \mathbb{R}^{(2T-1)d \times Td}.$$

By Proposition 2.3, $\boldsymbol{L}$ is an elliptical random vector. The next Lemma gives sufficient and necessary conditions for $\boldsymbol{L}$ to satisfy the two conditions in Definition 3.1.

Lemma 3.4

1. $\boldsymbol{L} \sim \mathrm{EC}_{(2T-1)d}(\mathbf{0}, \boldsymbol{\Omega}, \zeta)$ *satisfies Condition 1. Partition the scatter* $\boldsymbol{\Omega}$ *according to the dimensions of* $\{\boldsymbol{X}_t\}_{t=1}^T$ *and* $\{\boldsymbol{E}_t\}_{t=2}^T$:

$$\boldsymbol{\Omega} := \begin{pmatrix} \boldsymbol{\Omega}_X & \boldsymbol{\Omega}_{XE} \\ \boldsymbol{\Omega}_{XE}^\top & \boldsymbol{\Omega}_E \end{pmatrix}. \tag{3.4}$$

We have

$$\mathbf{\Omega}_E = \begin{pmatrix} \mathbf{\Psi} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{\Psi} \end{pmatrix} \in \mathbb{R}^{(T-1)d \times (T-1)d}. \tag{3.5}$$

2. ***L*** *satisfies Condition 2 if and only if the following equations hold*:

$$\mathbf{\Omega}_X = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1T} \\ \mathbf{\Sigma}_{12}^\top & \mathbf{\Sigma} & \cdots & \mathbf{\Sigma}_{2T} \\ & \cdots & & \\ \mathbf{\Sigma}_{1T}^\top & \mathbf{\Sigma}_{2T}^\top & \cdots & \mathbf{\Sigma} \end{pmatrix} \in \mathbb{R}^{Td \times Td}, \tag{3.6}$$

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top + \mathbf{\Psi}, \tag{3.7}$$

$$\mathbf{\Sigma}_{t,t+u} = \mathbf{\Sigma}(\mathbf{A}^\top)^u, \tag{3.8}$$

$$\text{and } (\mathbf{\Omega}_{XE})_{I_j I_k} = \begin{cases} \mathbf{0}, & if\ j \le k; \\ \mathbf{A}^{j-k-1}\mathbf{\Psi}, & if\ j > k, \end{cases} \tag{3.9}$$

for $t = 1, \ldots, T-1$, $u = 1, \ldots, T-t$, $j = 1, \ldots, T$, *and* $k = 2, \ldots, T-1$. *Here* $\mathbf{\Omega}_{XE} = [(\mathbf{\Omega}_{XE})_{I_j I_k}]$ *is a partition of* $\mathbf{\Omega}_{XE}$ *into* $d \times d$ *matrices, where* $I_l := \{(l-1)d + 1, \ldots, ld\}$ *for* $l = 1, \ldots, T$.

Lemma 3.4 is a consequence of Proposition 2.3. Detailed proof is collected in the supplementary material. Lemma 3.4 shows that there exists an elliptical random vector $\boldsymbol{L} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top, \boldsymbol{E}_2^\top, \ldots, \boldsymbol{E}_T^\top)^\top$ that satisfies the two conditions in Definition 3.1. On the other hand, the algorithm in Remark 3.2 generate a unique sequence of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_T$. Therefore, we immediately have the following proposition.

Proposition 3.5

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ *be an elliptical VAR process with latent innovations* $\boldsymbol{E}_2, \ldots, \boldsymbol{E}_T$. *Then* $\boldsymbol{L} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top, \boldsymbol{E}_2^\top, \ldots, \boldsymbol{E}_T^\top)^\top$ *is an absolutely continuous elliptical random vector*.

Denote $\mathbf{\Sigma}_1 := \mathbf{\Sigma}_{t,t+1}$. We call $\mathbf{\Sigma}$ a scatter matrix of the elliptical VAR process, and $\mathbf{\Sigma}_1$ a lag-one scatter matrix. For any $c > 0$, since $\boldsymbol{L} \sim \mathrm{EC}_d(\mathbf{0}, \mathbf{\Omega}, \zeta)$ implies $\boldsymbol{L} \sim \mathrm{EC}_d(\mathbf{0}, c\mathbf{\Omega}, \zeta/\sqrt{c})$, $c\mathbf{\Sigma}$ and $c\mathbf{\Sigma}_1$ are also scatter matrix and lag-one scatter matrix of the elliptical VAR process.

Next, we show that the scatter matrix and lag-one scatter matrix are closely related to the robust scales defined in Section 2.3. In particular, we show that the robust scale σ^Q motivates an alternative definition of the scatter matrix and lag-one scatter matrix.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ be an elliptical VAR process with $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{td})^\top$. We define

$$\mathbf{R}^{\mathrm{Q}} = [\mathbf{R}^{\mathrm{Q}}_{jk}] \text{ and } \mathbf{R}^{\mathrm{Q}}_{1} = [(\mathbf{R}^{\mathrm{Q}}_{1})_{j'k'}], \tag{3.10}$$

where the entries are given by

$$\mathbf{R}^{\mathrm{Q}}_{jj} \coloneqq \sigma^{\mathrm{Q}}(X_{1j})^2, \text{ for } j = 1, \ldots, d,$$

$$\mathbf{R}^{\mathrm{Q}}_{jk} \coloneqq \frac{1}{4}\left[\sigma^{\mathrm{Q}}(X_{1j} + X_{1k})^2 - \sigma^{\mathrm{Q}}(X_{1j} - X_{1k})^2\right], \text{ for } j \neq k,$$

$$(\mathbf{R}^{\mathrm{Q}}_{1})_{j'k'} \coloneqq \frac{1}{4}\left[\sigma^{\mathrm{Q}}(X_{1j'} + X_{2k'})^2 - \sigma^{\mathrm{Q}}(X_{1j'} - X_{2k'})^2\right], \text{ for } j', k' = 1, \ldots, d.$$

The next theorem shows that $\mathbf{R}^{\mathrm{Q}}$ and $\mathbf{R}^{\mathrm{Q}}_{1}$ are scatter matrix and lag-one scatter matrix of the elliptical VAR process.

Theorem 3.6

For the elliptical VAR process in Definition 3.1, we have

$$\mathbf{R}^{Q} = m^{\mathrm{Q}}\boldsymbol{\Sigma} \text{ and } \mathbf{R}^{\mathrm{Q}}_{1} = m^{\mathrm{Q}}\boldsymbol{\Sigma}_{1}, \tag{3.11}$$

where m^{Q} is a constant.

The proof of Theorem 3.6 exploits the summation stability of elliptical distributions and Proposition 2.3. Due to space limit, the detailed proof is collected in the supplementary material. Combining Lemma 3.4 and Theorem 3.6, we obtain the following theorem.

Theorem 3.7

For the elliptical VAR process in Definition 3.1, let $\mathbf{R}^{\mathrm{Q}}$ and $\mathbf{R}^{\mathrm{Q}}_{1}$ be defined as in (3.10). *Then, we have*

$$\mathbf{R}^{\mathrm{Q}}_{1} = \mathbf{R}^{\mathrm{Q}}\mathbf{A}^{\top}. \tag{3.12}$$

(3.12) serves as a quantile-based counterpart as the Yule Walker equation $\mathrm{Var}(\boldsymbol{X}_1) = \mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2)\mathbf{A}^{\top}$. Theorem 3.7 motivates the robust estimator of $\mathbf{A}$ introduced in the next section.

4. Method

In this section, we propose a robust estimator for the transition matrix $\mathbf{A}$. We first introduce robust estimators of $\mathbf{R}^{\mathrm{Q}}$ and $\mathbf{R}^{\mathrm{Q}}_{1}$. Based on these estimators, the transition matrix $\mathbf{A}$ can be estimated by solving an optimization problem.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ be an elliptical VAR process. We define

$$\widehat{\mathbf{R}}^{\mathrm{Q}} \coloneqq [\widehat{\mathbf{R}}^{\mathrm{Q}}_{jk}] \text{ and } \widehat{\mathbf{R}}^{\mathrm{Q}}_{1} \coloneqq [(\widehat{\mathbf{R}}^{\mathrm{Q}}_{1})_{jk}],$$

where the entries are given by

$$\widehat{\mathbf{R}}^{\mathrm{Q}}_{jj} \coloneqq \hat{\sigma}^{\mathrm{Q}}(\{X_{tj}\}_{t=1}^{T})^2, \text{ for } j=1, \ldots, d,$$

$$\widehat{\mathbf{R}}^{\mathrm{Q}}_{jk} \coloneqq \frac{1}{4}\left[\hat{\sigma}^{\mathrm{Q}}(\{X_{tj}+X_{tk}\}_{t=1}^{T})^2 - \hat{\sigma}^{\mathrm{Q}}(\{X_{tj}-X_{tk}\}_{t=1}^{T})^2\right], \text{ for } j \neq k \in \{1, \ldots, d\},$$

$$(\widehat{\mathbf{R}}^{\mathrm{Q}}_{1})_{jk} \coloneqq \frac{1}{4}\left[\hat{\sigma}^{\mathrm{Q}}(\{X_{tj}+X_{t+1,k}\}_{t=1}^{T-1})^2 - \hat{\sigma}^{\mathrm{Q}}(\{X_{tj}-X_{t+1,k}\}_{t=1}^{T-1})^2\right], \text{ for } j, k = 1, \ldots, d.$$

Motivated by Theorem 3.7, we proposed to estimate $\mathbf{A}$ by

$$\hat{\mathbf{A}} = \operatorname{argmin}_{\mathbf{M} \in \mathbb{R}^{d \times d}} \sum_{jk} |\mathbf{M}_{jk}| \tag{4.1}$$

$$\text{s.t.} \|\widehat{\mathbf{R}}^{\mathrm{Q}}\mathbf{M}^{\top} - \widehat{\mathbf{R}}^{\mathrm{Q}}_{1}\|_{\max} \leq \lambda.$$

The optimization problem (4.1) can be further decomposed into d subproblems (Han & Liu, 2013). Specifically, the j-th row of $\hat{\mathbf{A}}$ can be estimated by

$$\hat{\mathbf{A}}_{j*} = \operatorname{argmin}_{\mathbf{v} \in \mathbb{R}^d} \|\mathbf{v}\|_1 \tag{4.2}$$

$$\text{s.t.} \|\widehat{\mathbf{R}}^{\mathrm{Q}}\mathbf{v} - (\widehat{\mathbf{R}}^{\mathrm{Q}}_{1})_{*j}\|_{\infty} \leq \lambda.$$

Thus, the d rows of $\mathbf{A}$ can be estimated in parallel. (4.2) is essentially a linear programming problem, and can be solved efficiently using the simplex algorithm.

Remark 4.1

Since $\hat{\sigma}^{\mathrm{Q}}$ can be computed using $O(T \log T)$ time (Rousseeuw & Croux, 1993), the computational complexity of $\hat{\mathbf{R}}^{\mathrm{Q}}$ and $\widehat{\mathbf{R}}^{\mathrm{Q}}_{1}$ are $O(d^2 T \log T)$. Since $T \ll d$ in practice, $\hat{\mathbf{R}}^{\mathrm{Q}}$ and $\widehat{\mathbf{R}}^{\mathrm{Q}}_{1}$ can be computed almost as efficiently as their moment-based counterparts

$$\hat{\mathbf{S}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{X}_t \boldsymbol{X}_t^{\top} \text{ and } \hat{\mathbf{S}}_1 = \frac{1}{T-1}\sum_{t=1}^{T-1} \boldsymbol{X}_t \boldsymbol{X}_{t+1}^{\top}, \tag{4.3}$$

which have $O(d^2 T)$ complexity and are used in Han & Liu (2013).

5. Theoretical Properties

In this section, we present theoretical analysis of the proposed transition matrix estimator. Due to space limit, the proofs of the results in this section are collected in the supplementary material.

The consistency of the estimator depends on the degree of dependence over the process $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$. We first introduce the ϕ-mixing coefficient for quantifying the degree of dependence.

Definition 5.1

Let $\{X_t\}_{t\in Z}$ *be a stationary process. Define* $\mathcal{F}_{-\infty}^{0} := \sigma(X_t : t \le 0)$ *and* $\mathcal{F}_{n}^{\infty} := \sigma(X_t : t \ge n)$ *to be the* σ*-fileds generated by* $\{X_t\}_{t\le 0}$ *and* $\{X_t\}_{t\ge n}$*, respectively. The* ϕ*-mixing coefficient is defined by*

$$\phi(n) := \sup_{B \in \mathcal{F}_{-\infty}^{0}, A \in \mathcal{F}_{n}^{\infty}, \mathbb{P}(B) > 0} |\mathbb{P}(A \mid B) - \mathbb{P}(A)| .$$

Let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ be an infinite elliptical VAR process in the sense that any contiguous subsequence of $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is an elliptical VAR process. For brevity, we also call $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ an elliptical VAR process. Let $\phi_j(n)$, $\phi_{jk}^{+}(n), \phi_{jk}^{-}(n), \psi_{j'k'}^{+}(n)$, and $\psi_{j'k'}^{-}(n)$ be the ϕ-mixing coefficients of $\{X_{tj}\}_{t\in\mathbb{Z}}$, $\{X_{tj} + X_{tk}\}_{t\in\mathbb{Z}}$, $\{X_{tj} - X_{tk}\}_{t\in\mathbb{Z}}$, $\{X_{tj'} + X_{t+1,\,k'}\}_{t\in\mathbb{Z}}$, and $\{X_{tj'} - X_{t+1,k'}\}_{t\in\mathbb{Z}}$, respectively. Here $j, k, j', k' \in \{1, \ldots, d\}$ but $j \ne k$. Define

$$\Phi(n) = \sup_{j,k,j',k'} \{\phi_j(n), \phi_{jk}^{+}(n), \phi_{jk}^{-}(n), \psi_{j'k'}^{+}(n), \psi_{j'k'}^{-}(n)\},$$

and $\Theta(T) := \sum_{n=1}^{T} \Phi(n)$. Φ and Θ characterize the degree of dependence over the multivariate process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$.

Next, we introduce an identifiability condition on the distribution function of $\boldsymbol{X}_1$.

Condition 1

Let $\tilde{\boldsymbol{X}}_1 = (\tilde{X}_{11}, \ldots, \tilde{X}_{1d})^{\top}$ *and* $\tilde{\boldsymbol{X}}_2 = (\tilde{X}_{21}, \ldots, \tilde{X}_{2d})^{\top}$ *be independent copies of* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$*. Let* F_j, $F_{jk}^{+}, F_{jk}^{-}, G_{j'k'}^{+}$, *and* $G_{j'k'}^{-}$ *be the distribution functions of* $|X_{1j} - \tilde{X}_{1j}|$, $|X_{1j} + X_{1k} - \tilde{X}_{1j} - \tilde{X}_{1k}|$, $|X_{1j} - X_{1k} - \tilde{X}_{1j} + \tilde{X}_{1k}|$, $|X_{1j'} + X_{2k'} - \tilde{X}_{1j'} - \tilde{X}_{2k'}|$, *and* $|X_{1j'} - X_{2k'} - \tilde{X}_{1j'} + \tilde{X}_{2k'}|$. *We assume that there exist constants* $\kappa > 0$ *and* $\eta > 0$ *such that*

$$\inf_{|y - Q(F;\,1/4)| \le \kappa} \frac{d}{dy} F(y) \ge \eta$$

for any $F \in \{F_j, F_{jk}^{+}, F_{jk}^{-}, G_{j'k'}^{+}, G_{j'k'}^{-} : j \ne k \text{ and } j, k, j', k' = 1, \ldots, d\,.\}$.

Then next lemma presents the rates of convergence for $\hat{\mathbf{R}}^{\mathrm{Q}}$ and $\widehat{\mathbf{R}}_1^{\mathrm{Q}}$.

Lemma 5.2

Let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ *be an elliptical VAR process satisfying Condition 1. Let* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ *be a sequence of observations from* $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$*. Suppose that* $\log d/T \to 0$ *as* $T \to \infty$*. Then, for* T *large enough, with probability no smaller than* $1 - 8/d^2$*, we have*

$$\|\widehat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \le r(T), \tag{5.1}$$

$$\|\widehat{\mathbf{R}}_1^{\mathrm{Q}} - \mathbf{R}_1^{\mathrm{Q}}\|_{\max} \leq r_1(T), \tag{5.2}$$

where the rates of convergence are defined by

$$r(T) = \max\left\{\frac{2}{\eta^2}\left[\sqrt{\frac{8(1+2\Theta(T))\log d}{T}} + \frac{4\Theta(T)}{T}\right]^2, \frac{4\sigma_{\max}^{\mathrm{Q}}}{\eta}\left[\sqrt{\frac{8(1+2\Theta(T))\log d}{T}} + \frac{4\Theta(T)}{T}\right]\right\}, \tag{5.3}$$

$$r_1(T) = \max\left\{\frac{1}{\eta^2}\left[\sqrt{\frac{16(1+2\Theta(T))\log d}{T}} + \frac{8\Theta(T)}{T}\right]^2, \frac{2\tau_{\max}^{\mathrm{Q}}}{\eta}\left[\sqrt{\frac{16(1+2\Theta(T))\log d}{T}} + \frac{8\Theta(T)}{T}\right]\right\}. \tag{5.4}$$

Here
$\sigma_{\max}^{\mathrm{Q}} := \max\{\sigma^{\mathrm{Q}}(X_{1j}), \sigma^{\mathrm{Q}}(X_{1j} + X_{1k}), \sigma^{\mathrm{Q}}(X_{1j} - X_{1k}) : j \neq k \in \{1, \ldots, d\}\}$, $\tau_{\max}^{\mathrm{Q}} := \max\{\sigma^{\mathrm{Q}}(X_{1j} + X_{2k}), \sigma^{\mathrm{Q}}(X_{1j} - X_{2k}) : j, k \in \{1, \ldots, d\}\}$

Based on Lemma 5.2, we can further deliver the rates of convergence for $\hat{\mathbf{A}}$ under the matrix $\ell_{\max}$ norm and ℓ_1 norm. We start with some additional notation. For $\alpha \in [0, 1)$, $s > 0$, and $M_T > 0$ that may scale with T, we define the matrix class

$$\mathcal{M}(\alpha, s, M_T) := \left\{\mathrm{M} \in \mathbb{R}^{d \times d} : \max_{1 \leq j \leq d} \sum_{k=1}^{d} |\mathrm{M}_{jk}|^{\alpha} \leq s, \|\mathrm{M}\|_1 \leq M_T\right\}.$$

$\mathcal{M}(0, s, M_T)$ is the set of sparse matrices with at most s non-zero entries in each row and bounded ℓ_1 norm. $\mathcal{M}(\alpha, s, M_T)$ is also investigated in Cai et al. (2011) and Han & Liu (2013).

Theorem 5.3

Let $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ *be an elliptical VAR process satisfying Condition 1, and* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ *be a sequence of observations. Suppose that* $\log d / T \to 0$ *as* $T \to \infty$, *the transition matrix* $\mathbf{A} \in \mathcal{M}(\alpha, s, M_T)$, *and* $\mathbf{R}^{\mathrm{Q}}$ *is non-singular. Define*

$$r_{\max}(T) = \max\left\{\frac{2}{\eta^2}\left[\sqrt{\frac{16(1+2\Theta(T))\log d}{T}} + \frac{8\Theta(T)}{T}\right]^2, \frac{4\max(\sigma_{\max}^{\mathrm{Q}}, \tau_{\max}^{\mathrm{Q}})}{\eta}\left[\sqrt{\frac{16(1+2\Theta(T))\log d}{T}} + \frac{8\Theta(T)}{T}\right]\right\}.$$

If we choose the tuning parameter

$$\lambda = (1 + M_T) r_{\max}(T)$$

in (4.1), *then, for T large enough, with probability no smaller than* $1 - 8/d^2$, *we have*

$$\|\hat{\mathbf{A}} - \mathbf{A}\|_{\max} \leq 2\|(\mathbf{R}^{\mathrm{Q}})^{-1}\|_1(1 + M_T)r_{\max}(T), \tag{5.5}$$

$$\|\hat{\mathbf{A}} - \mathbf{A}\|_{\infty} \leq 4s\left[2\|(\mathbf{R}^{\mathrm{Q}})^{-1}\|_1(1 + M_T)r_{\max}(T)\right]^{1-\alpha}. \tag{5.6}$$

Remark 5.4

If we assume that $\eta \geq C_1$ and $\sigma^{\mathrm{Q}}_{\max}, \tau^{\mathrm{Q}}_{\max}, \|(\mathbf{R}^{\mathrm{Q}})^{-1}\|_1 \leq C_2$ for some absolute constants C_1, $C_2 > 0$, the rates of convergence in Theorem 5.3 reduces to

$$\|\hat{\mathbf{A}} - \mathbf{A}\|_{\max} = O_P\left(M_T\sqrt{\frac{\Theta(T)\log d}{T}}\right),$$

$$\|\hat{\mathbf{A}} - \mathbf{A}\|_{\infty} = O_P\left[s\left(M_T\sqrt{\frac{\Theta(T)\log d}{T}}\right)^{1-\alpha}\right].$$

Here $\Theta(T) = \sum_{n=1}^{T}\Phi(n)$ characterizes the degree of serial dependence in the process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$. If we further assume polynomial decaying ϕ-mixing coefficients

$$\Phi(n) \leq 1/n^{1+\varepsilon} \text{ for some } \varepsilon > 0, \tag{5.7}$$

we have $\Theta(T) \leq \sum_{n=1}^{\infty} 1/n^{1+\varepsilon} < \infty$ and the rate of convergence are further reduced to $\|\hat{\mathbf{A}} - \mathbf{A}\|_{\max} = O_P(M_T\sqrt{\log d/T})$ and $\|\hat{\mathbf{A}} - \mathbf{A}\|_{\infty} = O_P[s(M_T\sqrt{\log d/T})^{1-\alpha}]$, which are the parametric rates obtained in Han & Liu (2013) and Basu & Michailidis (2013). Condition (5.7) has been commonly assumed in the time series literature (Pan & Yao, 2008)

6. Granger Causality

In this section, we demonstrate an application of framework developed in this paper. In particular, we discuss the characterization and estimation of Granger causality under the elliptical VAR model. We start with the definition of Granger causality.

Definition 6.1 (Granger (1980))

Let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ be a stationary process, where $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{td})^{\top}$. For $j \neq k \in \{1, \ldots, d\}$, $\{X_{tk}\}_{t\in\mathbb{Z}}$ Granger causes $\{X_{tj}\}_{t\in\mathbb{Z}}$ if and only if there exists a measurable set A such that

$$\mathbb{P}(X_{t+1,j} \in A \,|\, \{\boldsymbol{X}_s\}_{s\leq t}) \neq \mathbb{P}(X_{t+1,j} \in A \,|\, \{\boldsymbol{X}_{s,\backslash k}\}_{s\leq t}),$$

for all $t \in \mathbb{Z}$, where $\boldsymbol{X}_{s,\backslash k}$ is the subvector obtained by removing X_{sk} from $\boldsymbol{X}_s$.

For a Gaussian VAR process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$, we have that $\{X_{tk}\}_{t\in\mathbb{Z}}$ Granger causes $\{X_{tj}\}_{t\in\mathbb{Z}}$ if and only if the $\{j, k\}$ entry of the transition matrix is non-zero (Lütkepohl, 2007). In the next theorem, we show that a similar property holds for the elliptical VAR process.

Theorem 6.2

Let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ *be an elliptical VAR process with transition matrix* $\mathbf{A}$. *Suppose* $\boldsymbol{X}_t$ *has finite second order moment, and* $\mathrm{Var}(X_{tk} \mid \boldsymbol{X}_{s,\setminus k}\}_{s\le t} \neq 0$ *for any* $k \in \{1, \ldots, d\}$. *Then, for* $j \neq k \in \{1, \ldots, d\}$, *we have*

1. *If* $\mathbf{A}_{jk} \neq 0$, *then* $\{X_{tk}\}_{t\in\mathbb{Z}}$ *Granger causes* $\{X_{tj}\}_{t\in\mathbb{Z}}$.
2. *If we further assume that* $\boldsymbol{E}_{t+1}$ *is independent of* $\{\boldsymbol{X}_s\}_{s\le t}$ *for any* $t \in \mathbb{Z}$, *we have that* $\{X_{tk}\}_{t\in\mathbb{Z}}$ *Granger causes* $\{X_{tj}\}_{t\in\mathbb{Z}}$ *if and only if* $\mathbf{A}_{jk} \neq 0$.

The proof of Theorem 6.2 exploits the autoregressive structure of the process $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$, and the properties on conditional distributions of elliptical random vectors. We refer to the supplementary material for the detailed proof.

Remark 6.3

The assumption that $\mathrm{Var}(X_{tk} \mid \boldsymbol{X}_{s,\setminus k}\}_{s\le t} \neq 0$ requires that X_{tk} cannot be perfectly predictable from the past or from the other observed random variables at time t. Otherwise, we can simply remove $\{X_{tk}\}_{t\in\mathbb{Z}}$ from the process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$, since predicting $\{X_{tk}\}_{t\in\mathbb{Z}}$ is trivial.

Assuming that $\boldsymbol{E}_{t+1}$ is independent of $\{\boldsymbol{X}_s\}_{s\le t}$ for any $t \in \mathbb{Z}$, the Granger causality relations among the processes $\{\{X_{jt}\}_{t\in\mathbb{Z}} : j = 1, \ldots, d\}$ is characterized by the non-zero entries of $\mathbf{A}$. To estimate the Granger causality relations, we define $\tilde{\mathbf{A}} = [\tilde{\mathbf{A}}_{jk}]$, where

$$\tilde{\mathbf{A}}_{jk} \coloneqq \hat{\mathbf{A}}_{jk} I(|\hat{\mathbf{A}}_{jk}| \ge \gamma),$$

for some threshold parameter $\boldsymbol{\gamma}$. To evaluate the consistency between $\tilde{\mathbf{A}}$ and $\mathbf{A}$ regarding sparsity pattern, we define function $\mathrm{sign}(x) \coloneqq I(x > 0) - I(x < 0)$. For a matrix $\mathbf{M}$, define $\mathrm{sign}(\mathbf{M}) \coloneqq [\mathrm{sign}(\mathbf{M}_{jk})]$.

The next theorem gives the rate of $\boldsymbol{\gamma}$ such that $\tilde{\mathbf{A}}$ recovers the sparsity pattern of $\mathbf{A}$ with high probability.

Theorem 6.4

Assume that the conditions in Theorem 5.3 holds, and $\mathbf{A} \in \mathcal{M}(0, s, M_T)$. *If we set*

$$\gamma = 2\|(\mathbf{R}^{\mathrm{Q}})^{-1}\|_1(1 + M_T)r_{\max}(T),$$

then, with probability no smaller than $1 - 8/d^2$, *we have* $\mathrm{sign}(\tilde{\mathbf{A}}) = \mathrm{sign}(\mathbf{A})$, *provided that*

$$\min_{\{(j,k):\mathbf{A}_{jk}>0\}} |\mathbf{A}_{jk}| \ge 2\gamma\,. \tag{6.1}$$

Theorem 6.4 is a direct consequence of Theorem 5.3. We refer to the supplementary material for a detailed proof.

7. Experiments

In this section, we demonstrate the empirical performance of the proposed transition matrix estimator using both synthetic and real data. In addition to the proposed robust Dantzig-selector-type estimator (`R-Dantzig`), we consider the following two competitors for comparison:

1. `Lasso`: an ℓ_1 regularized estimator defined in (1.1) with $l(\mathbf{Y} - \mathbf{MX}) = \|\mathbf{Y} - \mathbf{MX}\|_F^2$ and $P_\rho(\mathbf{M}) = \rho \sum_{jk} \mathbf{M}_{jk}$.
2. `Dantzig`: the estimator proposed in Han & Liu (Han & Liu, 2013), which solves (4.1) with $\hat{\mathbf{R}}^{\mathrm{Q}}$ and $\widehat{\mathbf{R}}_1^{\mathrm{Q}}$ replaced by $\hat{\mathbf{S}}$ and $\hat{\mathbf{S}}_1$ defined in (4.3).

`Lasso` is solved using R package *glmnet*. `Dantzig` and `R-Dantzig` are solved by the simplex algorithm.

7.1. Synthetic Data

In this section, we demonstrate the effectiveness of `R-Dantzig` under synthetic data. To generate the time series, we start with an initial observation $\boldsymbol{X}_1$ and innovations $\boldsymbol{E}_2, \ldots, \boldsymbol{E}_T$. Specifically, we consider three distributions for $(\boldsymbol{X}_1^\top, \boldsymbol{E}_2^\top, \ldots, \boldsymbol{E}_T^\top)^\top$:

`Setting 1`: a multivariate Gaussian distribution: $N(\mathbf{0}, \mathbf{\Phi})$;

`Setting 2`: a multivariate t distribution with degree of freedom 3, and covariance matrix $\mathbf{\Phi}$;

`Setting 3`: an elliptical distribution with log-normal generating variate, log $N(0, 2)$, and covariance matrix $\mathbf{\Phi}$.

Here the covariance matrix $\mathbf{\Phi}$ is block diagonal: $\mathbf{\Phi} = \mathrm{diag}(\mathbf{\Sigma}, \mathbf{\Psi}, \ldots, \mathbf{\Psi}) \in \mathbb{R}^{Td \times Td}$. We set $d = 50$ and $T = 25$. Using $(\boldsymbol{X}_1^\top, \boldsymbol{E}_2^\top, \ldots, \boldsymbol{E}_T^\top)^\top$, we can generate $(\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top)^\top$ by

$$(\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top)^\top = \mathbf{G}(\boldsymbol{X}_1^\top, \boldsymbol{E}_2^\top, \ldots, \boldsymbol{E}_T^\top)^\top,$$

where $\mathbf{G}$ is given by

$$\mathbf{G} := \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{A} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{A}^2 & \mathbf{A} & \mathbf{I} & \cdots & \mathbf{0} \\ & & \cdots & & \\ \mathbf{A}^{T-1} & \mathbf{A}^{T-2} & \mathbf{A}^{T-3} & \cdots & \mathbf{I} \end{pmatrix} \in \mathbb{R}^{Td \times Td}.$$

By Proposition 2.3, $(\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top)^\top$ follows a multivariate Gaussian distribution in `Setting 1`, a multivariate t distribution in `Setting 2`, and an elliptical distribution in `Setting 3` with the same log-normal generating variate.

We generate the parameters $\mathbf{A}$ and $\mathbf{\Sigma}$ following Han & Liu (2013). Specifically, we generate the transition matrix $\mathbf{A}$ using the *huge* R package, with patterns *band, cluster, hub*, and *random*. We refer to Han & Liu (2013) for a graphical illustration of the patterns. Then we rescale $\mathbf{A}$ so that $\|\mathbf{A}\|_2 = 0.8$. Given $\mathbf{A}$, we generate $\mathbf{\Sigma}$ such that $\|\mathbf{\Sigma}\|_2 = 2\|\mathbf{A}\|_2$. Using (3.7), we set $\mathbf{\Psi} = \mathbf{\Sigma} - \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$.

Table 1 presents the errors in estimating the transition matrix and their standard deviations. The tuning parameters λ and ρ are chosen by cross validation. The results are based on 1,000 replicated simulations. We note two observations: (i) Under the Gaussian model (`Setting 1`), `R-Dantzig` has comparable performance as Dantzig, and out-performs `Lasso`. (ii) In `Settings 2–3`, `R-Dantzig` produces significantly smaller estimation errors than `Lasso` and `Dantzig`. Thus, we conclude that `R-Dantzig` is robust to heavy tails.

Figure 1 plots the prediction errors ε_s against sparsity s for the three transition matrix estimators. We observe that `R-Dantzig` achieves smaller prediction errors compared to `Lasso` and `Dantzig`.

7.2. Real Data

In this section, we exploit the VAR model in stock price prediction. We collect adjusted daily closing prices[3] of 435 stocks in the S&P 500 index from January 1, 2003 to December 31, 2007. This gives us $T = 1,258$ closing prices of the 435 stocks. Let $\boldsymbol{X}_t$ be a vector of the 435 closing prices on day t, for $t = 1, \ldots, T$. We model $\{\boldsymbol{X}_t\}_{t=1}^T$ by a VAR process, and estimate the transition matrix using `Lasso, Dantzig`, and `R-Dantzig`. Let $\hat{\mathbf{A}}_s$ be an estimate of the transition matrix with sparsity s[4]. We define the prediction error associated with $\hat{\mathbf{A}}_s$ to be

$$\varepsilon_s := \frac{1}{T-1}\sum_{t=2}^{T} \|\boldsymbol{X}_t - \hat{\mathbf{A}}_s \boldsymbol{X}_{t-1}\|_2 .$$

8. Conclusion

In this paper, we developed a unified framework for modeling and estimating heavy-tailed VAR processes in high dimensions. Our contributions are three-fold. (i) In model level, we generalized the Gaussian VAR model by an elliptical VAR model to accommodate heavy-tailed time series. The model naturally couples with quantile-based scatter matrices and Granger causality. (ii) Methodologically, we proposed a quantile-based estimator of the transition matrix, which induces an estimator of Granger causality. Experimental results

[3]The adjusted closing prices account for all corporate actions such as stock splits, dividends, and rights offerings.
[4]$s \in [0, 1]$ is defined to be the fraction of non-zero entries in $\hat{\mathbf{A}}_s$, and can be controlled by the tuning parameters λ and ρ.

demonstrate that the proposed estimator is robust to heavy tails. (iii) Theoretically, we showed that the proposed estimator achieves parametric rates of convergence in matrix $\ell_{\max}$ norm and ℓ_∞ norm. The theory explicitly captures the effect of serial dependence, and implies sign-consistency of the induced Granger causality estimator. To our knowledge, this is the first work on modeling and estimating heavy-tailed VAR processes in high dimensions. The methodology and theory proposed in this paper have broad impact in analyzing non-Gaussian time series. The techniques developed in the proofs have independent interest in understanding robust estimators under high dimensional dependent data.

References

Basu, Sumanta; Michailidis, George. Estimation in high-dimensional vector autoregressive models. arXiv preprint arXiv:1311.4175. 2013

Cai, Tony; Liu, Weidong; Luo, Xi. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association. 2011; 106(494):594–607.

Cambanis, Stamatis; Huang, Steel; Simons, Gordon. On the theory of elliptically contoured distributions. Journal of Multivariate Analysis. 1981; 11(3):368–385.

Fang, Kai-Tai; Kotz, Samuel; Ng, Kai Wang. Symmetric Multivariate and Related Distributions. Chapman and Hall; 1990.

Feldman, Raya; Taqqu, Murad. A Practical Guide to Heavy Tails: Statistical Techniques and Applications. Springer; 1998.

Fujita, André; Sato, João R; Garay-Malpartida, Humberto M; Yamaguchi, Rui; Miyano, Satoru; Sogayar, Mari C; Ferreira, Carlos E. Modeling gene expression regulatory networks with the sparse vector autoregressive model. BMC Systems Biology. 2007; 1(1):1–39. [PubMed: 17408505]

Granger, Clive WJ. Testing for causality: a personal viewpoint. Journal of Economic Dynamics and Control. 1980; 2:329–352.

Hamilton, James Douglas. Time series analysis. Vol. 2. Princeton University Press; 1994.

Han, Fang; Liu, Han. Transition matrix estimation in high dimensional time series; Proceedings of the 30th International Conference on Machine Learning; 2013. 172–180.

Hsu, Nan-Jung; Hung, Hung-Lin; Chang, Ya-Mei. Subset selection for vector autoregressive processes using lasso. Computational Statistics & Data Analysis. 2008; 52(7):3645–3657.

Lütkepohl, Helmut. New Introduction to Multiple Time Series Analysis. Springer; 2007.

Nardi, Yuval; Rinaldo, Alessandro. Autoregressive process modeling via the lasso procedure. Journal of Multivariate Analysis. 2011; 102(3):528–549.

Opgen-Rhein, Rainer; Strimmer, Korbinian. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. BMC bioinformatics. 2007; 8(Suppl 2):S3.

Pan, Jiazhu; Yao, Qiwei. Modeling multiple time series via common factors. Biometrika. 2008; 95(2):365–379.

Qiu, Huitong; Han, Fang; Liu, Han; Caffo, Brian. Joint estimation of multiple graphical models from high dimensional time series. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2015

Rémillard, Bruno; Papageorgiou, Nicolas; Soustra, Frédéric. Copula-based semiparametric models for multivariate time series. Journal of Multivariate Analysis. 2012; 110:30–42.

Rousseeuw, Peter J; Croux, Christophe. Alternatives to the median absolute deviation. Journal of the American Statistical Association. 1993; 88(424):1273–1283.

Shojaie, Ali; Michailidis, George. Discovering graphical granger causality using the truncating lasso penalty. Bioinformatics. 2010; 26(18):i517–i523. [PubMed: 20823316]

Sims, Christopher A. Macroeconomics and reality. Econometrica. 1980:1–48.

Song, Song; Bickel, Peter J. Large vector autoregressions. arXiv preprint arXiv:1106.3915. 2011

Tsay, Ruey S. Analysis of financial time series. Vol. 543. John Wiley & Sons; 2005.

Wang, Hansheng; Li, Guodong; Tsai, Chih-Ling. Regression coefficient and autoregressive order shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007; 69(1):63–78.

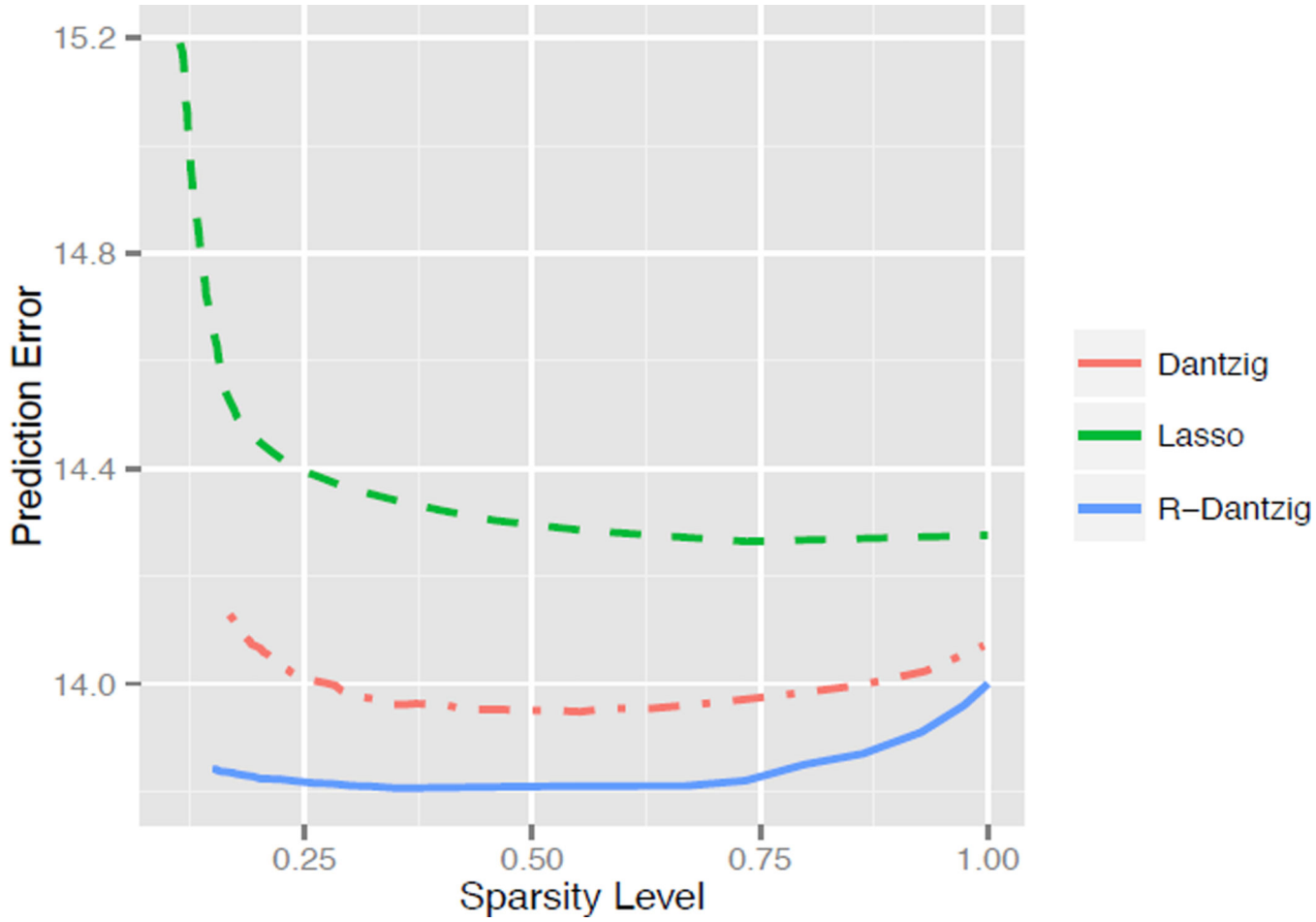

Figure 1.
Prediction errors in stock prices plotted against the sparsity of the estimated transition matrix.

Table 1

Averaged errors and standard deviations in estimating the transition matrix under the matrix Frobenius norm (ℓ_F), $\ell_{\max}$ norm, and ℓ_∞ norm. The results are based on 1,000 replications.

		Lasso			Dantzig			R-Danzig		
		ℓ_F	$\ell_{\max}$	ℓ_∞	ℓ_F	$\ell_{\max}$	ℓ_∞	ℓ_F	$\ell_{\max}$	ℓ_∞
Setting 1	*band*	4.26(0.74)	0.60(0.25)	2.15(0.38)	**3.24(1.07)**	**0.49(0.22)**	1.10(0.07)	3.65(0.01)	0.50(0.03)	**1.06(0.05)**
	cluster	3.04(0.65)	0.52(0.19)	1.82(0.35)	**2.25(0.39)**	**0.41(0.11)**	**1.00(0.58)**	2.47(0.01)	0.44(0.01)	1.13(0.04)
	hub	2.77(0.61)	0.66(0.05)	2.53(0.22)	**1.87(0.01)**	**0.64(0.02)**	**1.87(0.02)**	1.90(0.01)	0.65(0.01)	1.90(0.06)
	random	2.71(0.01)	0.47(0.01)	1.08(0.02)	**2.58(0.36)**	0.48(0.19)	1.21(0.83)	2.74(0.01)	**0.47(0.01)**	**1.19(0.08)**
Setting 2	*band*	9.53(0.58)	1.11(0.22)	10.45(1.29)	3.72(0.19)	0.52(0.10)	1.18(0.66)	**3.62(0.01)**	**0.47(0.02)**	**0.84(0.08)**
	cluster	8.52(0.38)	1.00(0.13)	9.24(1.16)	2.58(0.22)	0.46(0.03)	1.24(0.27)	**2.57(0.27)**	**0.44(0.01)**	**1.09(0.50)**
	hub	8.20(0.28)	0.97(0.09)	8.53(1.02)	3.87(0.01)	0.78(0.03)	3.30(0.02)	**1.88(0.01)**	**0.64(0.01)**	**1.90(0.06)**
	random	8.65(0.19)	0.98(0.10)	9.55(1.42)	2.79(0.07)	0.56(0.01)	1.35(0.15)	**2.72(0.02)**	**0.48(0.01)**	**1.12(0.10)**
Setting 3	*band*	9.43(0.25)	1.07(0.16)	10.83(1.16)	3.79(0.18)	0.52(0.02)	1.16(0.01)	**3.69(0.11)**	**0.49(0.04)**	**1.14(0.45)**
	cluster	8.59(0.34)	0.94(0.10)	9.70(0.98)	2.66(0.10)	0.44(0.02)	1.51(0.22)	**2.55(0.11)**	**0.43(0.01)**	**1.32(0.26)**
	hub	8.16(0.35)	0.95(0.10)	8.79(0.88)	2.51(0.11)	0.66(0.03)	2.34(0.15)	**2.01(0.23)**	**0.64(0.01)**	**2.07(0.30)**
	random	8.81(0.43)	1.04(0.12)	9.31(1.25)	2.71(0.13)	0.47(0.01)	1.28(0.16)	**2.55(0.10)**	**0.46(0.01)**	**1.04(0.29)**