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).

Let \( X_1, \ldots, X_T \in \mathbb{R}^d \) be a stationary multivariate time series. We consider VAR models\(^1\) such that

\[
X_t = A X_{t-1} + E_t \quad \text{for} \quad t = 2, \ldots, T,
\]

where \( A \) is the transition matrix, and \( E_2, \ldots, E_T \) 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

\(^1\)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).
can be formulated by
\[
\hat{\mathbf{A}}_{\text{reg}} := \arg\min_{\mathbf{M} \in \mathbb{R}^{d \times d}} l(Y - \mathbf{MX}) + P_{\rho}(\mathbf{M}),
\]
where \( Y := (X_1, \ldots, X_{T-1}) \in \mathbb{R}^{d \times (T-1)} \), \( \mathbf{X} := (X_2, \ldots, X_T) \in \mathbb{R}^{d \times (T-1)} \), \( l(\cdot) \) is a loss function, and \( P_{\rho}(\cdot) \) is a penalty function with penalty parameter \( \rho \). Common choices of the loss function include least squares loss and negative log-likelihood (Hamilton, 1994). For the penalty function, various \( \ell_1 \) penalties (Wang et al., 2007; Hsu et al., 2008; Shojaie & Michailidis, 2010) and ridge penalty (Hamilton, 1994) are widely used. Theoretical properties of \( \ell_1 \) 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 \( \ell_1 \) 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 \( \ell_1 \) 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-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 \( \mathbf{v} = (v_1, \ldots, v_d)^T \) 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 \( \ell_q \) norm of \( \mathbf{v} \) as \( ||\mathbf{v}||_q := (\sum_{j=1}^{d} |v_j|^q)^{1/q} \), and the vector \( \ell_\infty \) norm of \( \mathbf{v} \) as \( ||\mathbf{v}||_\infty := \max_{j=1}^{d} |v_j| \). Let the matrix \( \ell_{\max} \) norm of \( \mathbf{M} \) be \( ||\mathbf{M}||_{\max} := \max_{j,k} |\mathbf{M}_{jk}| \), the matrix \( \ell_{\infty} \) norm be \( ||\mathbf{M}||_{\infty} := \max_{j} \sum_{k=1}^{d_2} |\mathbf{M}_{jk}| \), and the Frobenius norm be \( ||\mathbf{M}||_F := \sqrt{\sum_{jk} M_{jk}^2} \). Let \( \mathbf{X} = (X_1, \ldots, X_d)^T \) and \( \mathbf{Y} = (Y_1, \ldots, Y_d)^T \) be two random vectors. We write \( \mathbf{X} \overset{\text{d}}{=} \mathbf{Y} \) if \( \mathbf{X} \) and \( \mathbf{Y} \) are identically distributed. We use \( 0, 1, \ldots \) to denote vectors with 0, 1, \ldots at every entry.

2.2. Elliptical Distribution

Definition 2.1 (Fang et al. (1990)). A random vector \( \mathbf{X} \in \mathbb{R}^d \) follows an elliptical distribution with location \( \mu \in \mathbb{R}^d \) 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 \( \mathbf{S} = \mathbf{R} \mathbf{R}^T \), a random vector \( \mathbf{U} \in \mathbb{R}^k \) independent of \( \xi \) and uniformly distributed in the \( k \) dimensional sphere, \( \mathbb{S}^{k-1} \), such that
\[
\mathbf{X} \overset{\text{d}}{=} \mu + \xi \mathbf{R} \mathbf{U}.
\]
In this case, we denote \( \mathbf{X} \sim EC_d(\mu, \mathbf{S}, \xi) \). \( \mathbf{S} \) is called the scatter matrix, and \( \xi \) is called the generating variate.

Remark 2.2. (2.1) is often referred to as the stochastic representation of the elliptical random vector \( \mathbf{X} \). 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 “\( \overset{\text{d}}{=} \)” with simply “\( = \)”.

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

$$\nu + BX \sim EC_p(\nu + B\mu, BSB^T, \xi\sqrt{B}),$$

where $B \sim \text{Beta}(l/2, (k-l)/2)$ 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

$$Q(X; q) = Q(F; q) := \inf\{x : \mathbb{P}(X \leq x) \geq q\},$$

$$\bar{Q}(\{X_t\}_{t=1}^T; q) := X^{(k)}$$  

where $k = \min\{t : \frac{t}{T} \geq q\}$.

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

$$\sigma^Q(X) := Q(\{|X - \bar{X}|; 1/4\},$$

$$\bar{\sigma}^Q((X_t)_{t=1}^T) := \bar{Q}(\{|X_t - X_t|_{t=1}^T; 1/4\},$$

where $\bar{X}$ is an independent copy of $X$. $\bar{\sigma}^Q((X_t)_{t=1}^T)$ 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 $X_1, \ldots, X_T \in \mathbb{R}^d$ is an elliptical VAR process if and only if the following conditions are satisfied:

1. $X_1, \ldots, X_T$ follow a lag-one VAR process

   $$X_t = AX_{t-1} + E_t, \text{ for } t = 2, \ldots, T,$$

   where $A \in \mathbb{R}^{d \times d}$ is the transition matrix, and $E_2, \ldots, E_T \in \mathbb{R}^d$ are latent innovations.

2. $(X_t^T, E_{t+1}^T)_{t=1}^{T-1}$ are stationary and absolutely continuous elliptical random vectors:

   $$\begin{pmatrix} X_t \\ E_{t+1} \end{pmatrix} \sim EC_{2d}(0, \begin{pmatrix} \Sigma & 0 \\ 0 & \Psi \end{pmatrix}, \zeta),$$

   where $\Sigma$ and $\Psi$ are positive definite matrices, and $\zeta > 0$ 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 $(X_t^T, E_{t+1}^T)$ can be written by $h(x, e) = 1/\sqrt{\det(\Psi)} g(x^T \Sigma^{-1} x + e^T \Psi^{-1} e)$ for some function $g$, and the density function of $X_t$ and the conditional density function of $E_{t+1}$ given $X_t$ can be written by

$$h_1(x) = \frac{1}{\sqrt{\det(\Sigma)}} g_1(x^T \Sigma^{-1} x)$$

and

$$h_2(e \mid x) = \frac{1}{\sqrt{\det(\Psi)}} g_2(e^T \Psi^{-1} e),$$

where $g_1$ and $g_2$ are defined by

$$g_1(r) = \int_{\mathbb{R}^d} g(\|z\|^2 + r) dz \text{ and } g_2(r) = g(r + x^T \Sigma^{-1} x).$$

The elliptical VAR process $X_1, \ldots, X_T$ can be generated by the following algorithm:

1. Generate $X_1$ from $h_1(x)$.

2. For $t = 2, \ldots, T$,
   
   (a) generate $E_t$ from $h_2(e \mid X_{t-1})$;
   (b) set $X_t = AX_{t-1} + E_t$.

**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

$$\begin{pmatrix} X_t \\ E_{t+1} \end{pmatrix} \sim N_{2d}(0, \begin{pmatrix} \Sigma & 0 \\ 0 & \Psi \end{pmatrix}).$$

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 $L = (X_1^T, X_2^T, \ldots, X_T^T, E_2^T, \ldots, E_T^T)^T$ such that the two conditions in Definition 3.1 are satisfied. To this end, let $L_0 := (X_1^T, E_2^T, \ldots, E_T^T)^T \sim EC_{2d}(0, \text{diag}(\Sigma, \Psi, \ldots, \Psi), \zeta)$ and define

$$L = (X_1^T, \ldots, X_T^T, E_2^T, \ldots, E_T^T)^T := BL_0,$$

where $B$ is a matrix.
where

\[
\begin{pmatrix}
1 & 0 & 0 & \ldots & 0 \\
A & 1 & 0 & \ldots & 0 \\
A^2 & A & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 1 \\
0 & 0 & 0 & \ldots & 1
\end{pmatrix} \in \mathbb{R}^{(2T-1)d \times Td}.
\]

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

**Lemma 3.4.**

1. **L** ∼ EC\((2T-1)d(0, \Omega, \zeta)\) satisfies Condition 1. Partition the scatter \( \Omega \) according to the dimensions of \( \{X_t\}_{t=1}^T \) and \( \{E_t\}_{t=2}^T \):

\[
\Omega := \begin{pmatrix}
\Omega_X & \Omega_{XE} \\
\Omega_{XE}^T & \Omega_E
\end{pmatrix},
\]

(3.4)

We have

\[
\Omega_E = \begin{pmatrix}
\Psi & 0 \\
\vdots & \ddots \\
0 & \Psi
\end{pmatrix} \in \mathbb{R}^{(T-1)d \times (T-1)d}.
\]

(3.5)

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

\[
\Omega_X = \begin{pmatrix}
\Sigma & \Sigma_{12} & \ldots & \Sigma_{1T} \\
\Sigma_{12}^T & \Sigma & \ldots & \Sigma_{2T} \\
\vdots & \vdots & \ddots & \vdots \\
\Sigma_{1T}^T & \Sigma_{2T}^T & \ldots & \Sigma
\end{pmatrix} \in \mathbb{R}^{Td \times Td},
\]

(3.6)

\[
\Sigma = A \Sigma A^T + \Psi,
\]

(3.7)

\[
\Sigma_{t,t+u} = \Sigma(A^T)^u,
\]

(3.8)

and

\[
(\Omega_{XE})_{t,t+k} = \begin{cases}
0, & \text{if } j \leq k; \\
A^{j-k-1} \Psi, & \text{if } j > k,
\end{cases}
\]

(3.9)

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

**Proposition 3.5.** Let \( X_1, \ldots, X_T \) be an elliptical VAR process with latent innovations \( E_2, \ldots, E_T \). Then \( L = (X_1^T, \ldots, X_T^T, E_2^T, \ldots, E_T^T)^T \) is an absolutely continuous elliptical random vector.

Denote \( \Sigma_1 := \Sigma_{t,t+1} \). We call \( \Sigma \) a scatter matrix of the elliptical VAR process, and \( \Sigma_1 \) a lag-one scatter matrix. For any \( c > 0 \), since \( L \sim EC_d(0, \Omega, \zeta) \) implies \( L \sim EC_d(0, c\Omega, \zeta / \sqrt{c}) \), \( c \Sigma \) and \( c \Sigma_1 \) 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 \( \sigma^Q \) motivates an alternative definition of the scatter matrix and lag-one scatter matrix.

Let \( X_1, \ldots, X_T \) be an elliptical VAR process with \( X_t = (X_{t1}, \ldots, X_{td})^T \). We define

\[
R_Q = [R_{Qj}^j] \quad \text{and} \quad R_1^Q = [(R_1^Q)_{j'k'}],
\]

(3.10)

where the entries are given by

\[
R_{Qj}^j := \sigma^Q(X_{1j})^2, \quad \text{for } j = 1, \ldots, d,
\]

\[
R_{Qj}^j := \frac{1}{4} \left[ \sigma^Q(X_{1j} + X_{1k})^2 - \sigma^Q(X_{1j} - X_{1k})^2 \right], \quad \text{for } j \neq k,
\]

(3.11)

\[
(R_{1}^Q)_{j'k'} := \frac{1}{4} \left[ \sigma^Q(X_{1j'} + X_{2k'})^2 - \sigma^Q(X_{1j'} - X_{2k'})^2 \right],
\]

(3.12)

for \( j', k' = 1, \ldots, d \).

The next theorem shows that \( R^Q \) and \( R_1^Q \) 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

\[
R^Q = m^Q \Sigma \quad \text{and} \quad R_1^Q = m^Q \Sigma_1,
\]

(3.13)

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 \( R^Q \) and \( R_1^Q \) be defined as in (3.10). Then, we have

\[
R_1^Q = R^Q A^T.
\]

(3.14)

(3.14) serves as a quantile-based counterpart as the Yule Walker equation \( \text{Var}(X_1) = \text{Cov}(X_1, X_2)A^T \). Theorem 3.7 motivates the robust estimator of \( A \) introduced in the next section.
4. Method
In this section, we propose a robust estimator for the transition matrix $\varepsilon$. We first introduce robust estimators of $\varepsilon^2$ and $\varepsilon^3$. Based on these estimators, the transition matrix $\varepsilon$ can be estimated by solving an optimization problem.

Let $X_0, \ldots, X_T$ be an elliptical VAR process. We define

\[
\varepsilon^2 := [\hat{\varepsilon}^2_{jk}], \quad \varepsilon^3 := [\{\hat{\varepsilon}^3_{jk}\}_{jk}],
\]

where the entries are given by

\[
\hat{\varepsilon}^2_{jk} := \frac{1}{T-1} \left[ \sum_{t=1}^{T} (X_{t+j} - X_{t+k})^T (X_{t+j} - X_{t+k}) \right]^{1/2},
\]

\[
\hat{\varepsilon}^3_{jk} := \frac{1}{T-1} \left[ \sum_{t=1}^{T} (X_{t+j} - X_{t+k})^T (X_{t+j} - X_{t+k}) \right].
\]

The consistency of the estimator depends on the degree of dependence over the multivariate process $X_0, \ldots, X_T$. We first introduce the $\phi$-mixing coefficient for quantifying the degree of dependence.

Definition 5.1. Let $\{X_t\}_{t \in \mathbb{Z}}$ be a stationary process. Define $F^0_{-\infty,\infty} := \mathbb{P}(X_t \leq 0)$ and $F^0_{\phi} := \mathbb{P}(X_t > 0)$ to be the $\phi$-filed generated by $\{X_t\}_{t \in \mathbb{Z}}$ and $\{X_t\}_{t \in \mathbb{Z}}$, respectively. The $\phi$-mixing coefficient is defined by

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

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.
where the rates of convergence are defined by

\[ r(T) = \max\left\{ \frac{2}{\eta^2} \left[ \sqrt{\frac{8(1 + 2\Theta(T))}{T} \log \frac{d}{T}} + \frac{4\Theta(T)}{T} \right]^2, \right. \]

\[ \frac{4\sigma^Q_{\max}}{\eta} \left[ \sqrt{\frac{8(1 + 2\Theta(T))}{T} \log \frac{d}{T}} + \frac{4\Theta(T)}{T} \right], \] (5.3)

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

\[ \frac{2\sigma^Q_{\max}}{\eta} \left[ \sqrt{\frac{16(1 + 2\Theta(T))}{T} \log \frac{d}{T}} + \frac{8\Theta(T)}{T} \right]. \] (5.4)

Here \( \sigma^Q_{\max} := \max\{\sigma^Q(X_{1,j}), \sigma^Q(X_{1,j} + X_{1,k}), \sigma^Q(X_{1,j} - X_{1,k}) : j \neq k \in \{1, \ldots, d\}, \}, \tau^Q_{\max} := \max\{\sigma^Q(X_{1,j} + X_{2,k}), \sigma^Q(X_{1,j} - X_{2,k}) : j, k \in \{1, \ldots, d\}\}. \)

Based on Lemma 5.2, we can further deliver the rates of convergence for \( \hat{A} \) under the matrix \( \ell_{\max} \) norm and \( \ell_1 \) 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

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

\( \mathcal{M}(0, s, M_T) \) is the set of sparse matrices with at most \( s \) non-zero entries in each row and bounded \( \ell_1 \) norm. \( \mathcal{M}(\alpha, s, M_T) \) is also investigated in Cai et al. (2011) and Han & Liu (2013).

**Theorem 5.3.** Let \( \{X_t\}_{t \in \mathbb{Z}} \) be an elliptical VAR process satisfying Condition 1, and \( X_1, \ldots, X_T \) be a sequence of observations. Suppose that \( \log d/T \to 0 \) as \( T \to \infty, \) the transition matrix \( A \in \mathcal{M}(\alpha, s, M_T), \) and \( R^Q \) is non-singular. Define

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

\[ \frac{4\max\{\sigma^Q_{\max}, \tau^Q_{\max}\}}{\eta} \left[ \sqrt{\frac{16(1 + 2\Theta(T))}{T} \log \frac{d}{T}} + \frac{8\Theta(T)}{T} \right]. \]

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

\[ \|\hat{A} - A\|_{\max} \leq 2\left\| (R^Q)^{-1}\right\|_1 (1 + M_T)r_{\max}(T), \] (5.5)

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

**Remark 5.4.** If we assume that \( \eta \geq C_1 \) and \( \sigma^Q_{\max}, \tau^Q_{\max}, \left\| (R^Q)^{-1}\right\|_1 \leq C_2 \) for some absolute constants \( C_1, C_2 > 0, \) the rates of convergence in Theorem 5.3 reduces to

\[ \|\hat{A} - A\|_{\max} = O_P \left( M_{T} \sqrt{\frac{\Theta(T) \log d}{T}} \right), \]

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

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

\[ \Phi(n) \leq 1/n^{1+\epsilon} \text{ for some } \epsilon > 0, \] (5.7)

we have \( \Theta(T) \leq \sum_{n=1}^{\infty} 1/n^{1+\epsilon} < \infty \) and the rate of convergence are further reduced to \( \|\hat{A} - A\|_{\max} = O_P(M_{T} \sqrt{\log d/T}) \) and \( \|\hat{A} - A\|_{\infty} = O_P\left[s(M_{T} \sqrt{\log d/T})^{1-\alpha}\right], \) 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 \( \{X_t\}_{t \in \mathbb{Z}} \) be a stationary process, where \( X_t = (X_{t1}, \ldots, X_{td})^T. \) 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

\[ \mathbb{P}(X_{t+1, j} \in A | \{X_s\}_{s \in \mathbb{Z}}) \neq \mathbb{P}(X_{t+1, j} \in A | \{X_{s,k}\}_{s \in \mathbb{Z}}), \]

for all \( t \in \mathbb{Z}, \) where \( X_{s,k} \) is the subvector obtained by removing \( X_{sk} \) from \( X_s. \)

For a Gaussian VAR process \( \{X_t\}_{t \in \mathbb{Z}} \), 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 \( \{X_t\}_{t \in \mathbb{Z}} \) be an elliptical VAR process with transition matrix \( A. \) Suppose \( X_t \) has finite second order moment, and \( \text{Var}(X_{tk} | X_{s,k})_{s \leq t} \neq 0 \) for any \( k \in \{1, \ldots, d\}. \) Then, for \( j \neq k \in \{1, \ldots, d\}, \) we have

1. If \( A_{jk} \neq 0, \) then \( \{X_{tk}\}_{t \in \mathbb{Z}} \) Granger causes \( \{X_{tj}\}_{t \in \mathbb{Z}}. \)
2. If we further assume that \( E_{t+1} \) is independent of \( \{X_s\}_{s \leq t} \) 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 \( A_{jk} \neq 0. \)
The proof of Theorem 6.2 exploits the autoregressive structure of the process $X_1, \ldots, X_T$, 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 $\text{Var}(X_{tk} \mid X_{tk}, \ldots, X_{t1}) \neq 0$ 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 $\{X_t\}_{t \in \mathbb{Z}}$, since predicting $\{X_{tk}\}_{t \in \mathbb{Z}}$ is trivial.

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

$$\hat{A}_{jk} := \hat{A}_{jk}I(|\hat{A}_{jk}| \geq \gamma),$$

for some threshold parameter $\gamma$. To evaluate the consistency between $A$ and $\hat{A}$ regarding sparsity pattern, we define function $\text{sign}(x) := I(x > 0) - I(x < 0)$. For a matrix $M$, define $\text{sign}(M) := [\text{sign}(M_{jk})]$.

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

**Theorem 6.4.** Assume that the conditions in Theorem 5.3 hold, and $A \in \mathcal{M}(0, s, M_T)$. If we set

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

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

$$\min_{(j,k) : A_{jk} > 0} |A_{jk}| \geq 2\gamma. \quad (6.1)$$

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 $\ell_1$ regularized estimator defined in (1.1) with $\|Y - MX\|_F^2$ and $P_p(M) = \rho \sum_{j,k} M_{jk}$.

2. **Dantzig:** the estimator proposed in Han & Liu (Han & Liu, 2013), which solves (4.1) with $\hat{R}^Q$ and $\hat{R}^Q_1$ replaced by $\hat{S}$ and $\hat{S}_1$ 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 $X_1$ and innovations $E_2, \ldots, E_T$. Specifically, we consider three distributions for $(X_1^T, E_2^T, \ldots, E_T^T)^T$:

**Setting 1:** a multivariate Gaussian distribution: $N(0, \Phi)$;

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

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

Here the covariance matrix $\Phi$ is block diagonal: $\Phi = \text{diag}(\Sigma_1, \Psi, \ldots, \Psi) \in \mathbb{R}^{Td \times Td}$. We set $d = 50$ and $T = 25$. Using $(X_1^T, E_2^T, \ldots, E_T^T)^T$, we can generate $(X_1^T, \ldots, X_T^T)^T$ by

$$(X_1^T, \ldots, X_T^T)^T = G(X_1^T, E_2^T, \ldots, E_T^T)^T,$$

where $G$ is given by

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

By Proposition 2.3, $(X_1^T, \ldots, X_T^T)^T$ 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 $A$ and $\Sigma$ following Han & Liu (2013). Specifically, we generate the transition matrix $A$ 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 $A$ so that $\|A\|_2 = 0.8$. Given $A$, we generate $\Sigma$ such that $2\|\Sigma\|_2 = 2\|A\|_2$. Using (3.7), we set $\Psi = \Sigma - A \Sigma A^T$.

Table 1 presents the errors in estimating the transition matrix and their standard deviations. The tuning parameters $\lambda$ and $\rho$ 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 outperforms 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 $\epsilon_s$ against sparsity $s$ for the three transition matrix estimators. We observe that R-Dantzig achieves smaller prediction errors compared to Lasso and Dantzig.
Robust Estimation of VAR Processes

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

<table>
<thead>
<tr>
<th>Setting</th>
<th></th>
<th>Lasso</th>
<th></th>
<th>Dantzig</th>
<th></th>
<th>R-Dantzig</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\ell_F$</td>
<td>$\ell_{\max}$</td>
<td>$\ell_{\infty}$</td>
<td>$\ell_F$</td>
<td>$\ell_{\max}$</td>
<td>$\ell_{\infty}$</td>
<td>$\ell_F$</td>
</tr>
<tr>
<td><strong>Setting 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>band</td>
<td>4.26(0.74)</td>
<td>0.60(0.25)</td>
<td>2.15(0.38)</td>
<td>3.24(1.07)</td>
<td>0.49(0.22)</td>
<td>1.10(0.07)</td>
<td>3.65(0.01)</td>
</tr>
<tr>
<td>cluster</td>
<td>3.04(0.65)</td>
<td>0.52(0.19)</td>
<td>1.82(0.35)</td>
<td>2.25(0.39)</td>
<td>0.41(0.11)</td>
<td>1.00(0.08)</td>
<td>2.47(0.01)</td>
</tr>
<tr>
<td>hub</td>
<td>2.77(0.61)</td>
<td>0.66(0.05)</td>
<td>2.53(0.22)</td>
<td>1.87(0.01)</td>
<td>0.64(0.02)</td>
<td>1.87(0.02)</td>
<td>1.90(0.01)</td>
</tr>
<tr>
<td>random</td>
<td>2.71(0.01)</td>
<td>0.47(0.01)</td>
<td>1.08(0.02)</td>
<td>2.58(0.36)</td>
<td>0.48(0.19)</td>
<td>1.21(0.83)</td>
<td>2.74(0.01)</td>
</tr>
<tr>
<td><strong>Setting 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>band</td>
<td>9.53(0.58)</td>
<td>1.11(0.22)</td>
<td>10.45(1.29)</td>
<td>3.72(0.19)</td>
<td>0.52(0.10)</td>
<td>1.18(0.66)</td>
<td>3.62(0.01)</td>
</tr>
<tr>
<td>cluster</td>
<td>8.52(0.38)</td>
<td>1.00(0.13)</td>
<td>9.24(1.16)</td>
<td>2.58(0.22)</td>
<td>0.46(0.03)</td>
<td>1.24(0.27)</td>
<td>2.57(0.27)</td>
</tr>
<tr>
<td>hub</td>
<td>8.20(0.28)</td>
<td>0.97(0.09)</td>
<td>8.53(1.02)</td>
<td>3.87(0.01)</td>
<td>0.78(0.03)</td>
<td>3.30(0.02)</td>
<td>1.88(0.01)</td>
</tr>
<tr>
<td>random</td>
<td>8.65(0.19)</td>
<td>0.98(0.10)</td>
<td>9.55(1.42)</td>
<td>2.79(0.07)</td>
<td>0.56(0.01)</td>
<td>1.35(0.15)</td>
<td>2.72(0.02)</td>
</tr>
<tr>
<td><strong>Setting 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>band</td>
<td>9.43(0.25)</td>
<td>1.07(0.16)</td>
<td>10.83(1.16)</td>
<td>3.79(0.18)</td>
<td>0.52(0.02)</td>
<td>1.16(0.01)</td>
<td>3.69(0.11)</td>
</tr>
<tr>
<td>cluster</td>
<td>8.59(0.34)</td>
<td>0.94(0.10)</td>
<td>9.70(0.98)</td>
<td>2.66(0.10)</td>
<td>0.44(0.02)</td>
<td>1.51(0.22)</td>
<td>2.55(0.11)</td>
</tr>
<tr>
<td>hub</td>
<td>8.16(0.35)</td>
<td>0.95(0.10)</td>
<td>8.79(0.88)</td>
<td>2.51(0.11)</td>
<td>0.66(0.03)</td>
<td>2.34(0.15)</td>
<td>2.01(0.23)</td>
</tr>
<tr>
<td>random</td>
<td>8.81(0.43)</td>
<td>1.04(0.12)</td>
<td>9.31(1.25)</td>
<td>2.71(0.13)</td>
<td>0.47(0.01)</td>
<td>1.28(0.16)</td>
<td>2.55(0.10)</td>
</tr>
</tbody>
</table>

7.2. Real Data

In this section, we exploit the VAR model in stock price prediction. We collect adjusted daily closing prices 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 $X_t$ be a vector of the 435 closing prices on day $t$, for $t = 1, \ldots, T$. We model $\{X_t\}_{t=1}^T$ by a VAR process, and estimate the transition matrix using Lasso, Dantzig, and R-Dantzig. Let $\hat{A}_s$ be an estimate of the transition matrix with sparsity $s$. We define the prediction error associated with $\hat{A}_s$ to be

$$
\epsilon_s := \frac{1}{T-1} \sum_{t=2}^{T} \|X_t - \hat{A}_s X_{t-1}\|_2.
$$

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 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 $\ell_{\infty}$ 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.

---

3The 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 $\hat{A}_s$, and can be controlled by the tuning parameters $\lambda$ and $\rho$.

---

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


