Noncausal Vector AR Processes with Application to Economic Time Series

Richard A. Davis and Li Song
Columbia University

October 5, 2012

Abstract

Inference procedures for noncausal autoregressive (AR) models have been well studied and applied in a variety of applications from environmental to financial. For such processes, the observations of time t may depend on both past and future shocks in the system. In this paper, we consider extension of the univariate noncausal AR models to the vector AR (VAR) case. The extension presents several interesting challenges since even a first-order VAR can possess both causal and noncausal components. Assuming a non-Gaussian distribution for the noise, we show how to compute an approximation to the likelihood function. Under suitable conditions, it is shown that the maximum likelihood estimator (MLE) of the vector of AR parameters is asymptotically normal. The estimation procedure is illustrated with a simulation study for a VAR(1) process and with two real data examples.

KEYWORDS: Autoregressive, noncausal, non-Gaussian

1 Introduction

An autoregressive process of order p (or AR(p)) is perhaps the best known and most studied process in time series analysis. In the univariate case, such a process is defined through the recursions

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t, \tag{1}$$

where $\{Z_t\}$ is an independent and identically distributed (iid) sequence of random variables with mean zero, variance σ^2. It is typically assumed that the AR(p) process is causal which is equivalent to

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \neq 0, \tag{2}$$

for all $z \in \mathbb{C}$ such that $|z| \leq 1$. In such cases, the solution $\{X_t\}$ to (1) exists and X_t can be expressed as a function of only the present and past of the noise process, Z_t, Z_{t-1}, ..., i.e., $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ for some sequence of constants $\{\psi_j\}$. The solution is said to be noncausal if $\phi(z)$ has any roots inside the unit circle. In particular, X_t is purely noncausal if all the roots of $\phi(z)$ are inside the unit circle; in this case, X_t is a function of only the future of the noise process, Z_{t+1}, Z_{t+2}, ... Finally, if $\phi(z)$ has roots both inside and outside the unit circle, we will say that X_t is mixed in the sense that X_t can be expressed as a two-sided infinite moving average of the noise process.

It should be noted that if $\{X_t\}$ is noncausal, then there exists a causal version of $\{X_t\}$, i.e., $\tilde{\phi}(B)X_t = \widetilde{Z}_t$ where B is the backward shift operator, $\tilde{\phi}(z)$ satisfies the causal condition 2, and $\{Z_t\}$ uncorrelated with mean 0 and variance $\tilde{\sigma}^2$. However, $\{\widetilde{Z}_t\}$ is only independent in the Gaussian case. Thus in estimating the AR parameters using the Gaussian likelihood, one needs to assume causality to ensure identifiability of the parameters. Hence, Gaussian or any estimation method based on second order properties of the process cannot be used for noncausal modeling. In the non-Gaussian case, causal and noncausal models are identifiable (see [3]). [4] studied the inference of MLE's for noncausal AR processes using the Laplace likelihood. They provided a general method for analyzing univariate noncausal AR processes. They started by reparameterizing the original parameters. They expressed the AR polynomial $\phi(z)$ as a product of a causal component and a purely noncausal component. Namely,

$$\phi(z) = \phi_c(z)\phi_n(z) = (1 - \theta_1 z - \cdots - \theta_r z^r)(1 - \theta_{r+1} z - \cdots - \theta_p z^s), \tag{3}$$

where $r + s = p$, $\phi_c(z) \neq 0$ for $|z| \leq 1$, and $\phi_n(z) \neq 0$ for $|z| \geq 1$. Then they showed that the score function of the new set of parameters $\{\theta_1, \ldots, \theta_p\}$ is asymptotically normal and provided the explicit form of the Fisher information of the score. Finally the asymptotic distribution of the original parameters $\{\phi_1, \ldots, \phi_p\}$ were derived based on the distribution of the score. [10] analyzed the noncausal AR in a similar way as in [4], although they had a slightly different model specification. They assumed $\phi(z)$ has the form

$$\phi(z) = \varphi(z^{-1})\phi_c(z) = (1 - \varphi_1 z^{-1} - \cdots - \varphi_s z^{-s})(1 - \psi_1 z - \cdots - \psi_r z^r), \tag{4}$$

where $\varphi(z) \neq 0$ and $\phi_c(z) \neq 0$ for $|z| \leq 1$. Furthermore, they considered a general class of distributions for the noise process, which includes the univariate Laplace and t distributions.

When it comes to noncausal multivariate AR models, only a few references can be found in the statistical literature. This may be due in part to the complicated nature of causality constraints in the multivariate case. For example, a VAR(1) model can have both causal and noncausal components. [11] tried to extend their ideas (4) to the multivariate case by assuming the AR process has a representation given by

$$\mathbf{\Pi}(B)\mathbf{\Phi}(B^{-1})\mathbf{X}_t = \mathbf{Z}_t, \tag{5}$$

where $\mathbf{X}_t := (X_{t,1}, \ldots, X_{t,m})^T$ is an m-dimensional stochastic process, $\mathbf{Z}_t := (Z_{t,1}, \ldots, Z_{t,m})^T$ is an iid sequence of continuous random vectors with mean $\mathbf{0}$ and covariance matrix Σ_0^*, and matrix polynomials $\mathbf{\Pi}(z) = I_n - \Pi_1 z - \cdots - \Pi_r z^r$ and $\mathbf{\Phi}(z) = I_n - \Phi_1 z - \cdots - \Phi_s z^s$, satisfy conditions $\det \mathbf{\Pi}(z) \neq 0$ and $\det \mathbf{\Phi}(z) \neq 0$ for $|z| \leq 1$. This is a natural extension of the model Lanne and Saikkonen used in [10]. However, there are two major issues associated with this model specification. First, in the vector case, all coefficients are matrices and hence $\mathbf{\Pi}(B)$ and $\mathbf{\Phi}(B^{-1})$ may not be commutative. In other words, one can also fit a model given by $\widetilde{\mathbf{\Phi}}(B^{-1})\widetilde{\mathbf{\Pi}}(B)\mathbf{X}_t = \mathbf{Z}_t$ to the same data and obtain completely different coefficients. Second, this model specification only covers a subset of VAR models. For example, a VAR(p) model in its conventional representation given by

$$\mathbf{\Phi}(B)\mathbf{X}_t = \mathbf{X}_t - \Phi_1 \mathbf{X}_{t-1} - \cdots - \Phi_p \mathbf{X}_{t-p} = \mathbf{Z}_t, \tag{6}$$

where $\{\Phi_1, \ldots, \Phi_p\}$ can be any arbitrary real matrices as long as $\det \mathbf{\Phi}(z) \neq 0$ for $|z| = 1$, does not always have a representation like (5). For instance, the simple AR(1) model

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} 0.8 & 0.6 \\ 0.6 & 1.7 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \end{pmatrix}, \tag{7}$$

whose AR polynomial has two roots 0.5 and 2, is not included in model (5).

In this paper, we consider the general representation of VAR(p) given in (6) and derive the asymptotic distributions for the estimates of the original parameters Φ_i's directly. We also allow our VAR models to be either causal, purely noncausal or mixed. In this way, we avoid the noncommutative issue of the matrix multiplication mentioned earlier and are able to deal with situations like (27).

In the univariate case, noncausal AR models have proven to be useful in a number of applications. For example, the Wal-Mart stock volume data in [1], the Microsoft stock volume data in [5] and the U.S. inflation data in [10]. In the multivariate case, [11] fitted noncausal VAR models to two economic data (US GDP data and US interest rate data) and argued that the noncausal fitting is more appropriate. While a noncausal AR model may provide a better fit to the data, the lack of causality could indicate that one needs to consider other exogenous information, such as news, which may contain "predictive information" that is external from the past observations.

The rest of the paper is organized as follows. Section 2 provides details about mixed VAR(p) models and includes a state-space representation for such models. This representation is useful not only in generating realizations of the process but also in deriving the likelihood. We also show that the nature of noncausality in model fitting can be examined through the impulse response coefficients. Section 3 establishes the uniqueness of the representation (6) in the non-Gaussian case using a key theorem from [6]. Section 4 develops the asymptotic theory for the MLE. The results are similar to those provided in [11]. Section 5 provides methods to do model selection and model fitting under the non-causality framework. Section 6 gives one simulation example and two real data examples to further illustrate the results in Section 4. The impulse response analysis is applied to the examples as well.

2 Model Specification

Consider the m-dimensional stationary process $\mathbf{X}_t := (X_{t,1}, \ldots, X_{t,m})^T$, $t = 1, 2, \ldots, n$, generated by (6). The conventional causal VAR models require that $\det \mathbf{\Phi}(z) \neq 0$, for all $z \in \mathbb{C}$, such that $|z| \leq 1$. Then the $\mathbf{X}_t$ can be written as

$$\mathbf{X}_t = \sum_{j=0}^{\infty} \Psi_j \mathbf{Z}_{t-j},$$

where the matrices Ψ_j are determined uniquely by

$$\mathbf{\Psi}(z) := \sum_{j=0}^{\infty} \Psi_j z^j = \mathbf{\Phi}^{-1}(z).$$

Here we drop the causality assumption and allow $\det \mathbf{\Phi}(z)$ to have roots inside the unit circle. Without loss of generality, assume that $\det \mathbf{\Phi}(z)$ has l roots outside the unit circle (the causal roots) and $pm - l$ roots inside the unit circle (the noncausal roots). However, it is not clear how to generate such a stationary process if we allow some of the roots to be noncausal. One possibility is to use the state-space representation to first transform a VAR(p) process to a VAR(1) process. Define two new processes $\mathbf{Y}_t$ and $\mathbf{Z}_t^*$ given by

$$\mathbf{Y}_t = \begin{pmatrix} \mathbf{X}_t \\ \mathbf{X}_{t-1} \\ \vdots \\ \mathbf{X}_{t-p+1} \end{pmatrix}_{pm \times 1}$$

and

$$\mathbf{Z}_t^* = \begin{pmatrix} \mathbf{Z}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}_{pm \times 1}.$$

It is well-known that $\mathbf{Y}_t$ has an VAR(1) representation given by

$$\mathbf{Y}_t = \mathbf{\Phi}_Y \mathbf{Y}_{t-1} + \mathbf{Z}_t^*, \tag{8}$$

where

$$\mathbf{\Phi}_Y = \begin{pmatrix} \Phi_1 & \Phi_2 & \cdots & \cdots & \Phi_p \\ I_m & O_m & \cdots & \cdots & O_m \\ O_m & I_m & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ O_m & O_m & \cdots & I_m & O_m \end{pmatrix}_{pm \times pm},$$

and O_m is the $m \times m$ matrix of zeros. An immediate advantage of using (8) is that the roots of $\det \mathbf{\Phi}(z)$ are now reciprocals of the eigenvalues of the matrix $\mathbf{\Phi}_Y$ (see also [13]). Hence, the causal roots of the AR polynomial correspond to the eigenvalues of $\mathbf{\Phi}_Y$ that are inside the unit circle and the noncausal roots correspond to those that are outside the unit circle. Notice that the matrix $\mathbf{\Phi}_Y$ can be further decomposed into its Jordan canonical form, i.e., there exists an invertible $pm \times pm$ matrix $\mathbf{A}$ such that

$$\mathbf{\Phi}_Y \mathbf{A} \;=\; \mathbf{AJ} = \mathbf{A} \begin{pmatrix} \mathbf{J}_1 & \\ & \mathbf{J}_2 \end{pmatrix},$$

where $\mathbf{J}$ is given by

$$\begin{pmatrix} \lambda_1 & & & & & & & \\ s_1 & \ddots & & & & & & \\ & \ddots & \ddots & & & & & \\ & & s_{l-1} & \lambda_l & & & & \\ & & & 0 & \lambda_{l+1} & & & \\ & & & & s_{l+1} & \ddots & & \\ & & & & & \ddots & \ddots & \\ & & & & & & s_{pm-1} & \lambda_{pm} \end{pmatrix}$$

and $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_l| < 1 < |\lambda_{l+1}| \leq \cdots \leq |\lambda_{pm}|$. The sub diagonal elements $\{s_i\}$ are either 0 or 1. Then from (8), we get

$$\begin{aligned} (\mathbf{I} - \mathbf{J}B)\mathbf{A}^{-1}\mathbf{Y}_t &:= (\mathbf{I} - \mathbf{J}B)\widetilde{\mathbf{Y}}_t \\ &= \mathbf{A}^{-1}\mathbf{Z}_t^* := \widetilde{\mathbf{Z}}_t. \end{aligned} \tag{9}$$

Notice that $\widetilde{\mathbf{Z}}_t$ is still a sequence of iid noise. The factors $\widetilde{\mathbf{Y}}_t$ and $\widetilde{\mathbf{Z}}_t$ can also be further partitioned according to $\mathbf{J}$. That is $\widetilde{\mathbf{Y}}_t = (\widetilde{\mathbf{Y}}_{t,1}^T, \widetilde{\mathbf{Y}}_{t,2}^T)^T$ and $\widetilde{\mathbf{Z}}_t = (\widetilde{\mathbf{Z}}_{t,1}^T, \widetilde{\mathbf{Z}}_{t,2}^T)^T$, where $\widetilde{\mathbf{Y}}_{t,1} = (\widetilde{Y}_{t,1}, \ldots, \widetilde{Y}_{t,l})^T$, $\widetilde{\mathbf{Y}}_{t,2} = (\widetilde{Y}_{t,l+1}, \ldots, \widetilde{Y}_{t,pm})^T$, $\widetilde{\mathbf{Z}}_{t,1} = (\widetilde{Z}_{t,1}, \ldots, \widetilde{Z}_{t,l})^T$ and $\widetilde{\mathbf{Z}}_{t,2} = (\widetilde{Z}_{t,l+1}, \ldots, \widetilde{Z}_{t,pm})^T$. From (10), we obtain,

$$\begin{cases} \widetilde{\mathbf{Y}}_{t,1} - \mathbf{J}_1\widetilde{\mathbf{Y}}_{t-1,1} = \widetilde{\mathbf{Z}}_{t,1} \\ \widetilde{\mathbf{Y}}_{t,2} - \mathbf{J}_2\widetilde{\mathbf{Y}}_{t-1,2} = \widetilde{\mathbf{Z}}_{t,2} \end{cases}, \tag{10}$$

which implies that $\widetilde{\mathbf{Y}}_{t,1}$ is a purely causal process and $\widetilde{\mathbf{Y}}_{t,2}$ is a purely non-causal process. Therefore, in order to generate the $\mathbf{X}_t$ process, one can first generate $\mathbf{Z}_t$ and transform $\mathbf{Z}_t$ to $\widetilde{\mathbf{Z}}_t$; use the standard forward (backward) method to generate $\widetilde{\mathbf{Y}}_{t,1}$ ($\widetilde{\mathbf{Y}}_{t,2}$); and finally, transform $\widetilde{\mathbf{Y}}_t$ back to $\mathbf{X}_t$. From (10), $\widetilde{\mathbf{Y}}_{t,1}$ and $\widetilde{\mathbf{Y}}_{t,2}$ can also be represented as infinite moving averages, i.e.,

$$\begin{cases} \widetilde{\mathbf{Y}}_{t,1} = \sum_{i=0}^{\infty} \mathbf{J}_1^i \widetilde{\mathbf{Z}}_{t-i,1} \\ \widetilde{\mathbf{Y}}_{t,2} = -\sum_{i=1}^{\infty} \mathbf{J}_2^{-i} \widetilde{\mathbf{Z}}_{t+i,2} \end{cases}.$$

This representation also leads to the representation of $\mathbf{Y}_t$ which is given by

$$\mathbf{Y}_t = \sum_{i=-\infty}^{\infty} \mathbf{A}\mathbf{F}_i\mathbf{A}^{-1}\mathbf{Z}_{t-i}^*, \tag{11}$$

where

$$\mathbf{F}_i = \begin{cases} \begin{pmatrix} \mathbf{J}_1^i & \\ & O_{pm-l} \end{pmatrix}, & i \geq 0, \\ \begin{pmatrix} O_l & \\ & -\mathbf{J}_2^i \end{pmatrix}, & i \leq -1. \end{cases}$$

Denote the upper-left $m \times m$ sub-matrix of $\mathbf{A}\mathbf{F}_i\mathbf{A}^{-1}$ by M_i, then the original time series $\mathbf{X}_t$ can be written as

$$\mathbf{X}_t = \sum_{i=-\infty}^{\infty} M_i\mathbf{Z}_{t-i}. \tag{12}$$

Because of the special structure of $\mathbf{F}_i$, the coefficient matrices M_i decay to zero at a geometric rate as $i \to \pm\infty$. It can also be shown that (12) is the unique stationary solution of $\mathbf{X}_t$ specified by model (6).

It should be noted that these matrix coefficients M_i's in (12) are often referred to as the *impulse response coefficients* of this linear filter. If the covariance matrix Σ_0^* of $\mathbf{Z}_t$ is not diagonal, then it is common to apply a transformation, so that the noise becomes uncorrelated. A popular choice of such transformations comes from the Cholesky decomposition. Let P_L be a lower triangular matrix such that $\Sigma_0^* = P_L P_L'$. Then $\{M_i P_L;\ i = \cdots, -1, 0, 1, \cdots\}$ are the new impulse response coefficients for a new sequence of noise $\{P_L^{-1}\mathbf{Z}_t\}$. For instance, assuming we have a model given in (27) and the covariance matrix of $\mathbf{Z}_t$ is diagonal, the impulse response coefficients of model (27) are plotted in Figure 1. If the linear filter were causal, the impulse response coefficients of the filter would be zero for all negative lags. But, as shown in Figure 1, the four curves are non-zero for $t < 0$ and this implies noncausality. It is also interesting to see that for this particular model, if there is a shock $\mathbf{Z}_0 = (-1, 1/2)'$ at time 0 and other noise terms are 0, the response $\mathbf{X}_t$ will be zero for $t < 0$. In other words, we would not anticipate this special kind of shocks before time 0. On the other hand, if a large shock is not proportional to $(-1, 1/2)'$ at time 0, the response is essentially non-zero for $t < 0$ and thus we would anticipate the shock before time 0.

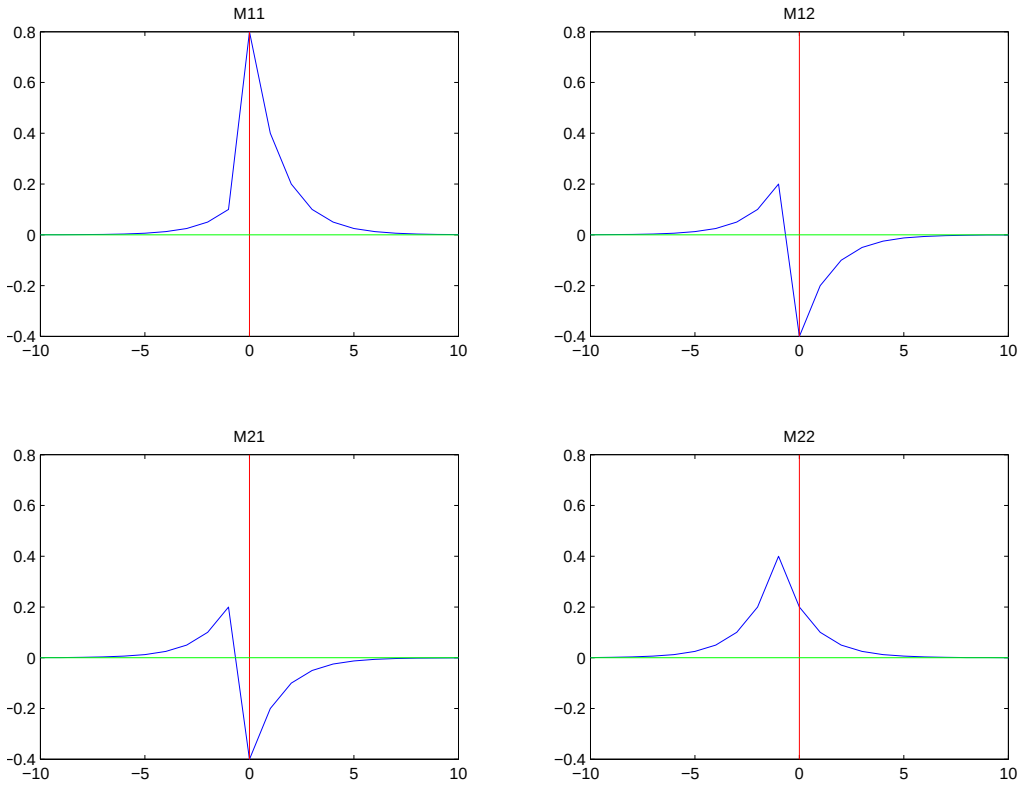

Figure 1: Impulse response coefficients of model (27).

3 Identifiability Issues

A practical complication with noncausal AR models is the non-identifiability issue under the Gaussian likelihood. To see this, note that the spectral density matrix of the process $\mathbf{X}_t$ defined by (6) is given by

$$\begin{aligned} f(\omega) &= \frac{1}{2\pi}\mathbf{\Phi}\left(e^{i\omega}\right)^{-1}\mathbf{Var}(\mathbf{Z}_t)\left(\mathbf{\Phi}\left(e^{-i\omega}\right)^{-1}\right)^T \\ &= \frac{1}{2\pi}\left[\mathbf{\Phi}\left(e^{-i\omega}\right)^T\Sigma_0^{*-1}\mathbf{\Phi}\left(e^{i\omega}\right)\right]^{-1}. \end{aligned} \tag{13}$$

In the expression (13) above, the matrix in the square brackets is 2π times the spectral density matrix of a vector MA(p) process. Furthermore, a slight modification of Theorem 10' of [9] leads to the representation

$$\begin{aligned} &\mathbf{\Phi}\left(e^{-i\omega}\right)^T\Sigma_0^{*-1}\mathbf{\Phi}\left(e^{i\omega}\right) \\ &= \left(\sum_{j=0}^{p}\mathcal{A}_j e^{-ij\omega}\right)^T\left(\sum_{j=0}^{p}\mathcal{A}_j e^{ij\omega}\right), \end{aligned}$$

where the $m\times m$ matrices $\mathcal{A}_0,\ldots,\mathcal{A}_p$ are real with $\mathcal{A}_0$ positive definite, and the zeros of $\det\left(\sum_{j=0}^{p}\mathcal{A}_j z^j\right)$ lie outside the unique disc. This implies the spectral density matrix of $\mathbf{X}_t$ has the representation

$$f(\omega)=\frac{1}{2\pi}\left(\sum_{j=0}^{p}\mathcal{A}_j e^{ij\omega}\right)^{-1}\left(\sum_{j=0}^{p}\mathcal{A}_j^T e^{-ij\omega}\right)^{-1},$$

which is the spectral density matrix of a causal vector AR(p) process.

The preceding discussion means that, under the Gaussian niose assumption where the distribution of $\mathbf{X}_t$ is purely determined by the second order properties, one cannot distinguish between a noncausal VAR process and a causal VAR process. Therefore, under the noncausal vector AR framework, one has to consider non-Gaussianity for the innovation. The identifiability under non-Gaussian errors can be established by Theorem 1 from [6]. We modify their theorem slightly to adapt to our setup and state it as a lemma below.

Lemma 3.1. *Let* $\mathbf{X}$ *and* $\widetilde{\mathbf{X}}$ *be two non-Gaussian m-dimensional linear processes defined by*

$$\begin{aligned} \mathbf{X}_t &= \sum_{j=-\infty}^{\infty}C_j\mathbf{Z}_{t-j}, \quad t=\cdots,-1,0,1,\cdots, \\ \widetilde{\mathbf{X}}_t &= \sum_{j=-\infty}^{\infty}\widetilde{C}_j\widetilde{\mathbf{Z}}_{t-j}, \quad t=\cdots,-1,0,1,\cdots, \end{aligned}$$

where the coefficients C_j *and* $\widetilde{C}_j$ *are assumed to be square-summable, and the innovations* $\{\mathbf{Z}_t\}$ *and* $\{\widetilde{\mathbf{Z}}_t\}$ *are m-dimensional iid sequences. Then the processes* $\{\mathbf{X}_t\}$ *and* $\{\widetilde{\mathbf{X}}_t\}$ *are equivalent in distribution if and only if there exist an integer q and a matrix H such that, for all t,*

$$\widetilde{\mathbf{Z}}_{t-q}\overset{\mathrm{d}}{=}H\mathbf{Z}_t \quad \textit{and} \quad C_{t-q}=\widetilde{C}_t H, \tag{14}$$

under the following two conditions:

(1) The innovation sequence $\mathbf{Z}_t$ *admits an invertible* B_K *for some* K *with an* $r \geq 3$*, where* $K = (k_3, k_4, \ldots, k_r)$ *is a multi-index,* $1 \leq k_i \leq m$ *for all* $3 \leq i \leq r$*, and the matrix* B_K *is the* $m \times m$ *matrix where the* (i,j)*th entry of* B_K *is the cumulant* $\alpha_{ijK} = cum(Z_{t,i}, Z_{t,j}, Z_{t,k_3}, \ldots, Z_{t,k_r})$*, where* $Z_{t,l}$ *means the* l*-th component of* $\mathbf{Z}_t$*.*

(2) Any two linear combinations of $\mathbf{Z}_t$ *with non-zero coefficients must be stochastically dependent.*

Remark 3.2. For the definition of cumulants, see Jammalamadaka, Rao and Terdik (2004).

Remark 3.3. The above lemma essentially points out an important fact that, under some regularity conditions on the distribution of the innovation, then the linear representation of the non-Gaussian process is unique aside from changes in scale and shifts in the time origin of the innovation series.

Remark 3.4. Chan and Ho (2004) argued that the two conditions of the innovation process are rather mild and are satisfied by many distributions, including the multivariate t-distribution.

4 Asymptotic Theory

4.1 Approximate Likelihood

The discussion from Section 3 demonstrates that the only meaningful application of non-causal VAR models requires that the distribution of $\mathbf{Z}_t$ to be non-Gaussian. On the other hand, when the innovation process satisfies the conditions of Lemma 3.1, then the coefficient matrices in the VAR model are identifiable everywhere. Here, we adopt the assumptions made by Lanne and Saikkonen in [11]. Suppose $\mathbf{Z}_t$ follows some elliptical distribution, which has a density of the form

$$f_\Sigma(\mathbf{z}; \boldsymbol{\nu}) = \det(\Sigma)^{-1/2} f(\mathbf{z}^T \Sigma^{-1} \mathbf{z}; \boldsymbol{\nu}). \tag{15}$$

As shown in [8], $\mathbf{Z}_t$ also has the representation,

$$\mathbf{Z}_t \overset{\mathrm{d}}{=} \Sigma^{1/2} \boldsymbol{\epsilon}_t \overset{\mathrm{d}}{=} \rho_t \Sigma^{1/2} \mathbf{v}_t, \tag{16}$$

where $(\rho_t, \mathbf{v}_t)$ is an iid sequence such that ρ_t (scalar) and $\mathbf{v}_t$ $(m \times 1)$ are independent, ρ_t is non-negative, and $\mathbf{v}_t$ is uniformly distributed on the unit ball (i.e., $\mathbf{v}_t^T \mathbf{v}_t = 1$). The vector $\boldsymbol{\nu}$ $(d \times 1)$ and the matrix Σ $(m \times m)$, which is assumed to be symmetric and positive definite, are parameters associated with the distribution of $\mathbf{Z}_t$. Specifically, since $\mathbf{E}(\mathbf{v}_t) = 0$ and $\mathbf{Var}(\mathbf{v}_t) = m^{-1} I_m$, one obtains from (16) that

$$\mathbf{Var}(\mathbf{Z}_t) = \frac{\mathbf{E}(\rho_t^2)}{m} \Sigma. \tag{17}$$

A convenient feature of elliptical distributions is that we can often work with the scalar random variable ρ_t instead of the random vector $\mathbf{Z}_t$. As stated in [8] and [11], the density function of ρ_t^2 denoted by $\phi_{\rho^2}(\cdot; \boldsymbol{\nu})$ is related to the function $f(\cdot; \boldsymbol{\nu})$ in (15) via

$$\varphi_{\rho^2}(\zeta; \boldsymbol{\nu}) = \frac{\pi^{m/2}}{\Gamma(m/2)} \zeta^{m/2-1} f(\zeta; \boldsymbol{\nu}), \quad \zeta \geq 0,$$

where $\Gamma(\cdot)$ is the gamma function. Similar notation as in [11] is used in this paper. For example $f'(\zeta; \boldsymbol{\nu})$ is used to signify the partial derivative $\partial f(\zeta; \boldsymbol{\nu}) / \partial \zeta$ and $f''(\zeta; \boldsymbol{\nu})$ is the second derivative of f with respect to ζ. The following assumptions are imposed (see also [11], [4]) for a function $f(\zeta; \boldsymbol{\nu})$ in order to ensure the validity of the asymptotic theory.

Assumptions:

A.1. The parameter space of $\boldsymbol{\nu}$, denoted by Λ, is an open subset of $\mathbb{R}^d$ and that of the parameter matrix Σ is the set of symmetric positive definite $m \times m$ matrices.

A.2. The function $f(\zeta;\boldsymbol{\nu})$ is positive and twice continuously differentiable on $(0,\infty)\times\Lambda$. Furthermore, for all $\boldsymbol{\nu} \in \Lambda$, $\lim_{\zeta\to\infty} \zeta^{m/2} f(\zeta;\boldsymbol{\nu}) = 0$, and a finite and non-negative right limit $\lim_{\zeta\to 0+} f(\zeta;\boldsymbol{\nu})$ exists.

A.3. $\int_0^\infty \zeta^{m/2-1} f'(\zeta;\boldsymbol{\nu}_0)d\zeta < \infty$, $\lim_{\zeta\to\infty} \zeta^{m/2+1} f'(\zeta;\boldsymbol{\nu}_0) = 0$, and a finite right limit $\lim_{\zeta\to 0+} f'(\zeta;\boldsymbol{\nu}_0)$ exists.

A.4. For all $\boldsymbol{\nu} \in \Lambda$,

$$\int_0^\infty \zeta^{m/2+1} f(\zeta;\boldsymbol{\nu})d\zeta < \infty$$

and

$$\int_0^\infty \zeta^{m/2}(1+\zeta)\frac{(f'(\zeta;\boldsymbol{\nu}))^2}{f(\zeta;\boldsymbol{\nu})}d\zeta < \infty.$$

A.5. There exists a function $f_1(\zeta)$ such that $\int_0^\infty \zeta^{m/2-1} f_1(\zeta)d\zeta < \infty$ and in some neighborhood of λ_0, $||\partial f(\zeta;\boldsymbol{\nu})/\partial\boldsymbol{\nu}|| \le f_1(\zeta)$ for all $\zeta \ge 0$, where $||\cdot||$ is the maximum normal of a vector or a matrix. Moreover,

$$\left|\left|\int_0^\infty \frac{\zeta^{m/2-1}}{f(\zeta;\boldsymbol{\nu}_0)}\frac{\partial^2 f(\zeta;\boldsymbol{\nu}_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^T}d\zeta\right|\right| < \infty,$$

for all $\zeta \ge 0$.

A.6. There exists a function $f_2(\zeta)$ such that $\int_0^\infty \zeta^{m/2-1} f_1(\zeta)d\zeta < \infty$ and in some neighborhood of $\boldsymbol{\nu}_0$, $\zeta||\partial f'(\zeta;\boldsymbol{\nu})/\partial\boldsymbol{\nu}|| \le f_2(\zeta)$ and $||\partial^2 f(\zeta;\boldsymbol{\nu})/\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^T|| \le f_2(\zeta)$ for all $\zeta \ge 0$.

A.7. For all $\zeta \ge 0$ and all $\boldsymbol{\nu}$ in some neighborhood of $\boldsymbol{\nu}_0$, the functions

$$\left(\frac{f'(\zeta;\boldsymbol{\nu})}{f(\zeta;\boldsymbol{\nu})}\right)^2, \quad \left|\frac{f''(\zeta;\boldsymbol{\nu})}{f(\zeta;\boldsymbol{\nu})}\right|, \quad \frac{1}{f^2(\zeta;\boldsymbol{\nu})}\left|\left|\frac{\partial}{\partial\boldsymbol{\nu}}f(\zeta;\boldsymbol{\nu})\right|\right|^2, \quad \frac{1}{f(\zeta;\boldsymbol{\nu})}\left|\left|\frac{\partial}{\partial\boldsymbol{\nu}}f'(\zeta;\boldsymbol{\nu})\right|\right|$$

and

$$\frac{1}{f(\zeta;\boldsymbol{\nu})}\left|\left|\frac{\partial^2}{\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^T}f(\zeta;\boldsymbol{\nu})\right|\right|$$

are dominated by $a_1 + a_2\zeta^{a_3}$ with a_1, a_2 and a_3 being nonnegative constants and $\int_0^\infty \zeta^{m/2+1+a_3} f(\zeta;\boldsymbol{\nu})d\zeta < \infty$.

Remark 4.1. All the assumptions above are given in [11]. As they pointed out, these assumptions also have a connection to the assumptions used in the analysis of univariate non-Gaussian noncausal AR processes, see also [2] and [4].

Remark 4.2. The first condition in A.4 implies $\mathbf{E}(\rho_t^4)$ is finite, and the second condition guarantees finiteness of some expectations needed in the subsequent developments (the information matrix of the score). More specifically, the second condition implies

$$\begin{aligned}
\mathbf{E}\left[\rho_t^2\left(\frac{f'(\rho_t^2;\boldsymbol{\nu})}{f(\rho_t^2;\boldsymbol{\nu})}\right)^2\right] &= \frac{\pi^{m/2}}{\Gamma(m/2)}\int_0^\infty \zeta^{m/2}\frac{(f'(\zeta;\boldsymbol{\nu}))^2}{f(\zeta;\boldsymbol{\nu})}d\zeta < \infty; \\
\mathbf{E}\left[\rho_t^4\left(\frac{f'(\rho_t^2;\boldsymbol{\nu})}{f(\rho_t^2;\boldsymbol{\nu})}\right)^2\right] &= \frac{\pi^{m/2}}{\Gamma(m/2)}\int_0^\infty \zeta^{m/2+1}\frac{(f'(\zeta;\boldsymbol{\nu}))^2}{f(\zeta;\boldsymbol{\nu})}d\zeta < \infty.
\end{aligned}$$

Remark 4.3. Assumptions A.5 and A.6 impose standard dominance conditions for $f(\cdot)$ and $f'(\cdot)$ which guarantee the interchangeability of limit operations and ensures the score vector and the information matrix corresponding to $\boldsymbol{\nu}$ behave in a desired fashion.

Remark 4.4. Assumption A.7 gives dominance conditions needed to establish the asymptotic normality of the MLE's of the parameters.

From (10), it's easy to see that

$$\begin{pmatrix} \widetilde{\mathbf{Y}}_{p,1} \\ \widetilde{\mathbf{Z}}_{p+1,1} \\ \vdots \\ \widetilde{\mathbf{Z}}_{n,1} \end{pmatrix} = \begin{pmatrix} I_{pm} & & & \\ -\mathbf{J}_1 & I_{pm} & & \\ & \ddots & \ddots & \\ & & -\mathbf{J}_1 & I_{pm} \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{Y}}_{p,1} \\ \widetilde{\mathbf{Y}}_{p+1,1} \\ \vdots \\ \widetilde{\mathbf{Y}}_{n,1} \end{pmatrix} = T_1 \begin{pmatrix} \widetilde{\mathbf{Y}}_{p,1} \\ \widetilde{\mathbf{Y}}_{p+1,1} \\ \vdots \\ \widetilde{\mathbf{Y}}_{n,1} \end{pmatrix},$$

and

$$\begin{pmatrix} \widetilde{\mathbf{Y}}_{p,2} \\ \widetilde{\mathbf{Y}}_{p+1,2} \\ \vdots \\ \widetilde{\mathbf{Y}}_{n,2} \end{pmatrix} = \begin{pmatrix} -\mathbf{J}_2^{-1} & \cdots & \cdots & -\mathbf{J}_2^{-(n-1)} \\ & \ddots & & \vdots \\ & & -\mathbf{J}_2^{-1} & -\mathbf{J}_2^{-1} \\ & & & I_{pm} \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{Z}}_{p+1,2} \\ \vdots \\ \widetilde{\mathbf{Z}}_{n,2} \\ \widetilde{\mathbf{Y}}_{n,2} \end{pmatrix} = T_2^{-1} \begin{pmatrix} \widetilde{\mathbf{Z}}_{p+1,2} \\ \vdots \\ \widetilde{\mathbf{Z}}_{n,2} \\ \widetilde{\mathbf{Y}}_{n,2} \end{pmatrix}.$$

Therefore, we have the relationship

$$\begin{aligned} \mathbf{P}_1 \begin{pmatrix} \widetilde{\mathbf{Y}}_{p,1} \\ \widetilde{\mathbf{Z}}_{p+1} \\ \vdots \\ \widetilde{\mathbf{Z}}_{n} \\ \widetilde{\mathbf{Y}}_{n,2} \end{pmatrix} &= \begin{pmatrix} \widetilde{\mathbf{Y}}_{p,1} \\ \widetilde{\mathbf{Z}}_{p+1,1} \\ \vdots \\ \widetilde{\mathbf{Z}}_{n,1} \\ \widetilde{\mathbf{Z}}_{p+1,2} \\ \vdots \\ \widetilde{\mathbf{Z}}_{n,2} \\ \widetilde{\mathbf{Y}}_{n,2} \end{pmatrix} = \begin{pmatrix} T_1 & \\ & T_2 \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{Y}}_{p,1} \\ \widetilde{\mathbf{Y}}_{p+1,1} \\ \vdots \\ \widetilde{\mathbf{Y}}_{n,1} \\ \widetilde{\mathbf{Y}}_{p,2} \\ \vdots \\ \widetilde{\mathbf{Y}}_{n-1,2} \\ \widetilde{\mathbf{Y}}_{n,2} \end{pmatrix} \\ &= \begin{pmatrix} T_1 & \\ & T_2 \end{pmatrix} \mathbf{P}_2 \begin{pmatrix} \widetilde{\mathbf{Y}}_{p} \\ \widetilde{\mathbf{Y}}_{p+1} \\ \vdots \\ \widetilde{\mathbf{Y}}_{n} \end{pmatrix}, \end{aligned}$$

where $\mathbf{P}_1$ and $\mathbf{P}_2$ are permutation matrices with determinant 1. From (10), the likelihood of the observed time series $\{\mathbf{X}_t\}$ can be written as

$$
\begin{aligned}
& L(\mathbf{X}_1, \ldots, \mathbf{X}_n) \\
= \; & L(\mathbf{Y}_p, \mathbf{Y}_{p+1}, \ldots, \mathbf{Y}_n) \\
= \; & L(\widetilde{\mathbf{Y}}_p, \widetilde{\mathbf{Y}}_{p+1}, \ldots, \widetilde{\mathbf{Y}}_n) |\det(\mathbf{A})|^{-(n-p+1)} \\
= \; & p_1(\widetilde{\mathbf{Y}}_{p,1}) p_2(\widetilde{\mathbf{Y}}_{n,2}) |\det(\mathbf{A})|^{-(n-p+1)} \prod_{i=p+1}^{n} p(\widetilde{\mathbf{Z}}_i) \cdot |\det(T_1)| \cdot |\det(T_2)| \\
= \; & p_1(\widetilde{\mathbf{Y}}_{p,1}) p_2(\widetilde{\mathbf{Y}}_{n,2}) |\det(\mathbf{A})|^{-(n-p+1)} \prod_{i=p+1}^{n} p(\widetilde{\mathbf{Z}}_i) \cdot |\det(\mathbf{J}_2)|^{n-p} \\
= \; & p_1(\widetilde{\mathbf{Y}}_{p,1}) p_2(\widetilde{\mathbf{Y}}_{n,2}) \cdot |\det(\mathbf{A})|^{-(n-p+1)} \cdot \prod_{i=p+1}^{n} (f_\Sigma(\mathbf{Z}_i; \lambda) \cdot |\det(\mathbf{A})|) \cdot |\det(\mathbf{J}_2)|^{n-p} \\
= \; & p_1(\widetilde{\mathbf{Y}}_{p,1}) p_2(\widetilde{\mathbf{Y}}_{n,2}) \cdot |\det(\mathbf{A})|^{-1} \prod_{i=p+1}^{n} (f_\Sigma(\mathbf{Z}_i; \lambda) \cdot |\det(\mathbf{J_2})|) \,, \qquad (18)
\end{aligned}
$$

where $p_1(\cdot)$, $p_2(\cdot)$, and $p(\cdot)$ are density functions of $\widetilde{\mathbf{Y}}_{p,1}$, $\widetilde{\mathbf{Y}}_{n,2}$ and $\{\widetilde{\mathbf{Z}}_i\}$, respectively. The third equality follows from the facts that the $\widetilde{\mathbf{Z}}_i$'s are independent, $\widetilde{\mathbf{Y}}_{p,1}$ only depends on $\{\widetilde{\mathbf{Z}}_{-\infty}, \ldots, \widetilde{\mathbf{Z}}_p\}$ and $\widetilde{\mathbf{Y}}_{n,2}$ only depends on $\{\widetilde{\mathbf{Z}}_{n+1}, \ldots, \widetilde{\mathbf{Z}}_\infty\}$. It is easy to see that the first part of the above formula, $p_1(\widetilde{\mathbf{Y}}_{p,1}) p_2(\widetilde{\mathbf{Y}}_{n,2}) \cdot |\det(\mathbf{A})|^{-1}$, remains bounded in probability and hence plays little role in the asymptotics. This suggests approximating the joint density of $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ by the second term in (18), namely $\prod_{i=p+1}^{n} (f_\Sigma(\mathbf{Z}_i; \lambda) \cdot |\det(\mathbf{J_2})|)$.

Some matrix notation is adopted from [11] in the following discussion. By $\mathrm{vec}(A)$ we denote a column vector obtained by stacking the columns of the matrix A one below another. If A is a square matrix then $\mathrm{vech}(A)$ is a column vector obtained by stacking the columns of A from the principal diagonal downwards (including elements on the diagonal), and $A \otimes B$ is used for the Kronecker product of the matrices A and B. The $ml \times ml$ commutation matrix and the $m^2 \times m(m+1)/2$ duplication matrix are denoted by K_{ml} and D_m, respectively. Both of them are of full column rank. The former is defined by the relation $K_{ml}\mathrm{vec}(A) = \mathrm{vec}(A^T)$, where A is any $m \times l$ matrix, and the latter by the relation $\mathrm{vec}(B) = D_m \mathrm{vech}(B)$, where B is any symmetric $m \times m$ matrix.

In stacking the parameters into a vector, define

$$
\boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{pm^2} \end{pmatrix} = \begin{pmatrix} \mathrm{vec}(\Phi_1) \\ \mathrm{vec}(\Phi_2) \\ \vdots \\ \mathrm{vec}(\Phi_p) \end{pmatrix}
$$

and

$$
\boldsymbol{\sigma} = \mathrm{vech}(\Sigma).
$$

Let $\boldsymbol{\theta}$ be the parameter vector that contains all the unknown parameters, namely

$$
\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\phi} \\ \boldsymbol{\sigma} \\ \boldsymbol{\nu} \end{pmatrix}. \qquad (19)
$$

It is also clear that $|\det(\mathbf{J}_2)|$ is a function of $\boldsymbol{\phi}$ only. In fact, $|\det(\mathbf{J}_2)| = |\lambda_{l+1}\cdots\lambda_{pm}|$, which is a product of all eigenvalues of $\boldsymbol{\Phi}_Y$ that are outside the unit circle. Thus, $\kappa(\boldsymbol{\phi})$ is used in the following discussion to represent $\log|\det(\mathbf{J}_2)|$.

4.2 Limiting Distribution of MLE

In this section, we discuss the asymptotic properties of the MLE's of the parameters. First let $\tilde{l}_n(\boldsymbol{\theta})$ and $l_n(\boldsymbol{\theta})$ be the complete and approximate log-likelihood of the time series $\{\mathbf{X}_t\}$, i.e.,

$$
\begin{aligned}
(20)\qquad \tilde{l}_n(\boldsymbol{\theta}) &= \log(L(\mathbf{X}_1,\ldots,\mathbf{X}_n)) \\
&= \log(p_1(\widetilde{\mathbf{Y}}_{p,1})p_2(\widetilde{\mathbf{Y}}_{n,2})\cdot|\det(\mathbf{A})|^{-1}) + \sum_{i=p+1}^{n}(\log f_\Sigma(\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})+\kappa(\boldsymbol{\phi})) \\
&= \log(p_1(\widetilde{\mathbf{Y}}_{p,1})p_2(\widetilde{\mathbf{Y}}_{n,2})\cdot|\det(\mathbf{A})|^{-1}) \\
&\quad + \sum_{i=p+1}^{n}\left(-1/2\log\det(\Sigma)+\log f(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})+\kappa(\boldsymbol{\phi})\right), \\
(21)\qquad l_n(\boldsymbol{\theta}) &= \sum_{i=p+1}^{n}\left(-1/2\log\det(\Sigma)+\log f(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})+\kappa(\boldsymbol{\phi})\right) \\
&:= \sum_{i=p+1}^{n} g_i(\boldsymbol{\theta}),
\end{aligned}
$$

where $\mathbf{Z}_t(\boldsymbol{\phi})$ is defined through recursions $\mathbf{Z}_t(\boldsymbol{\phi}) = \mathbf{X}_t - \Phi_1\mathbf{X}_{t-1} - \cdots - \Phi_p\mathbf{X}_{t-p}$. Let $\boldsymbol{\theta}_0$, $\boldsymbol{\phi}_0$, $\boldsymbol{\sigma}_0$ and $\boldsymbol{\nu}_0$ be the true values of the parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\nu}$. Therefore, $\mathbf{Z}_i(\boldsymbol{\phi}_0) = \mathbf{Z}_i$. We first consider the distribution of the score vector evaluated at the true value, i.e., $\partial l_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$. For convenience, also define

$$h(\zeta;\nu) = \frac{f'(\zeta;\boldsymbol{\nu})}{f(\zeta;\boldsymbol{\nu})}$$

and

$$(22)\qquad \mathbf{e}_i(\boldsymbol{\theta}) = h(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})\Sigma^{-1/2}\mathbf{Z}_i(\boldsymbol{\phi}).$$

The first and second derivatives of the function $g_i(\boldsymbol{\theta})$ can be computed analytically and are presented in the appendix. We will then show in the following theorem that the score vector of the approximate likelihood evaluated at the true value, i.e.,

$$(23)\qquad \frac{\partial l_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} = \sum_{i=p+1}^{n}\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} = \begin{pmatrix} \sum_{i=p+1}^{n}\partial g_i(\boldsymbol{\theta}_0)/\partial\boldsymbol{\phi} \\ \sum_{i=p+1}^{n}\partial g_i(\boldsymbol{\theta}_0)/\partial\boldsymbol{\sigma} \\ \sum_{i=p+1}^{n}\partial g_i(\boldsymbol{\theta}_0)/\partial\boldsymbol{\nu} \end{pmatrix},$$

has a normal limit.

Theorem 4.5. *Suppose that Assumptions A.1-A.6 hold and that $\mathbf{Z}_t$ is non-Gaussian. Then,*

$$
\begin{aligned}
\frac{1}{\sqrt{n-p}}\frac{\partial l_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} &= \frac{1}{\sqrt{n-p}}\sum_{i=p+1}^{n}\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} \\
(24)\qquad &\xrightarrow{\mathrm{d}} N(\mathbf{0},\mathcal{I}_{\theta\theta}(\boldsymbol{\theta}_0)),
\end{aligned}
$$

where the matrix $\mathcal{I}_{\theta\theta}(\boldsymbol{\theta}_0)$ *is given by*

$$-\mathbf{E}\begin{pmatrix} \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T} & \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\phi}\partial\boldsymbol{\sigma}^T} & \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\phi}\partial\boldsymbol{\nu}^T} \\ \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\phi}^T} & \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^T} & \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\nu}^T} \\ \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\phi}^T} & \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\sigma}^T} & \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^T} \end{pmatrix}. \tag{25}$$

Further assume,

$$-\mathbf{E}\begin{pmatrix} \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^T} & \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\nu}^T} \\ \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\sigma}^T} & \frac{\partial^2 g_{p+1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^T} \end{pmatrix}$$

is positive definite, then $\mathcal{I}_{\theta\theta}(\boldsymbol{\theta}_0)$ *is positive definite.*

The proof of the theorem is given in Appendix B.

Remark 4.6. Most previous studies (for example, [4] and [11]) about noncausal AR processes proved (33) and (34) in a more indirect fashion. Our proof showed that these properties of the approximate likelihood should follow naturally from those of the complete likelihood.

Next we argue that there there exists a sequence of solutions, $\hat{\boldsymbol{\theta}}_n$, to the likelihood equations,

$$\frac{\partial l_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = 0,$$

with $l_n(\boldsymbol{\theta})$ given in (21), which is consistent and asymptotically efficient in the sense that

$$\sqrt{n-p}\cdot(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \mathcal{I}_{\theta\theta}^{-1}(\boldsymbol{\theta}_0)),$$

where $\mathcal{I}_{\theta\theta}(\boldsymbol{\theta}_0)$ is the Fisher information matrix given in (25). In order to establish the asymptotic normality of the estimator of the parameter, one needs to prove the local consistency, which usually requires that the Hessian of the log-likelihood function satisfies

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}\left\|\frac{1}{n-p}\sum_{i=p+1}^{n}\left(\frac{\partial^2 g_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} - \frac{\partial^2 g_i(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)\right\| \xrightarrow{\mathrm{p}} 0, \tag{26}$$

where $\boldsymbol{\Theta}_0$ is some small enough compact neighborhood of $\boldsymbol{\theta}_0$. With one additional assumption A.7 and following the same argument as in [11] (proof of Theorem 4), (26) and hence the following theorem can be established.

Theorem 4.7. *Suppose that Assumptions A.1-A.7 hold and that* $\mathbf{Z}_t$ *is non-Gaussian and satisfies the conditions given by Lemma 3.1. Then there exists a sequence of local maximizers* $\hat{\boldsymbol{\theta}}$ *of* $l_n(\boldsymbol{\theta})$ *in (21) such that*

$$\sqrt{n-p}\cdot(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \mathcal{I}_{\theta\theta}^{-1}(\boldsymbol{\theta}_0)).$$

Furthermore, $\mathcal{I}_{\theta\theta}(\boldsymbol{\theta}_0)$ *can be consistently estimated by* $-(n-p)^{-1}\partial^2 l_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$.

5 Model Selection and Model Fitting

In fitting a VAR model, causal or noncausal, to data. The first challenge is to identify the order p. For causal models, one can use standard information criteria such as AIC or BIC. Since AIC and BIC essentially rely only on second order properties of the model, which are indistinguishable

between causal and noncausal models, this suggests that AIC and BIC will also be effective for estimating p in the noncausal case. And if the underlying noise process does exhibit a heavier tail than Gaussian, then BIC is expected to perform better than AIC, as the BIC generally penalizes free parameters more strongly than does the AIC. This is supported by the following empirical study.

Consider the process

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} -7.64 & 12.62 \\ -5.88 & 10.04 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} 3.6 & -5.64 \\ 4.2 & -6.63 \end{pmatrix} \begin{pmatrix} X_{t-2,1} \\ X_{t-2,2} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \end{pmatrix},$$

where $Z_{t,1}$ and $Z_{t,2}$ are two independent t-distributed noise sequences. In the simulation study, a standard t-distribution with 4 degrees of freedom is used. This VAR model's characteristic function has 4 roots; three outside the unit circle (1.667, 2, 3.333) and one inside the unit circle (0.5).

In this simulation, we generated 10,000 realizations of the time series $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ with the length $n = 300$. For each simulated time series, a causal VAR model was selected using either AIC or BIC criterion. The following table summarizes the counts of different selected. As seen in this table, BIC is extremely effective in identifying the order p for a noncausal model.

Model order	2	3	4	5	6	7	8	9	10	11	12
AIC	8726	828	270	92	42	25	6	9	1	0	1
BIC	9998	2	0	0	0	0	0	0	0	0	0

Table 1: Simulation results: order selection based by AIC and BIC criteria.

To facilitate the maximization of the potentially noncausal AR model, we can select the order p using BIC and fit a causal Gaussian model to the data. The parameter estimates and its maximized likelihood will serve as a benchmark. It is easy to envision that the non-Gaussian likelihood of the non-causal model will have many local maxima, which makes finding the global maximum difficult. To avoid getting trapped in finding only local maxima and not the global maxima, one should start the optimization from various starting points in the parameter set. The choice of starting points can be guided by the estimates obtained from the fitting causal Gaussian model, allowing for noncausal coefficients.

6 Examples

6.1 Simulation studies

In order to illustrate the theory from the previous section, first consider a bivariate ($m = 2$) vector AR(1) process generated from bivariate t-distributed innovations, i.e.,

$$\mathbf{X}_t = \Phi_1 \mathbf{X}_{t-1} + \mathbf{Z}_t,$$

where $\mathbf{Z}_t$ has the density function given by

$$\begin{aligned} f_{\Sigma_0}(\mathbf{z}) &= \det(\Sigma_0)^{-1} f(\mathbf{z}^T \Sigma_0^{-1} \mathbf{z}; \nu) \\ &= \det(\Sigma_0)^{-1} \frac{\nu^{\nu/2} \Gamma\left((\nu+m)/2\right)}{\pi^{m/2} \Gamma\left(\nu/2\right)} \cdot \frac{1}{\left(\nu + \mathbf{z}^T \Sigma_0^{-1} \mathbf{z}\right)^{(\nu+m)/2}}. \end{aligned}$$

Therefore, the $h(\cdot)$ function defined in (22) and its derivative have a simple form given by

$$h(\zeta;\nu) = -\frac{\nu+m}{2(\nu+\zeta)} \quad \text{and} \quad h'(\zeta;\nu) = \frac{\nu+m}{2(\nu+\zeta)^2}.$$

Assume Φ_1 has one eigenvalue inside the unit disc $|\lambda_1| < 1$ and the other outside $\lambda_2 > 1$ (the case $\lambda_2 < -1$ can be argued in a similar way) and has a decomposition

$$\Phi_1 = AJA^{-1} = \begin{pmatrix} 1 & 1 \\ \frac{\lambda_1-\phi_1}{\phi_3} & \frac{\lambda_2-\phi_1}{\phi_3} \end{pmatrix} \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \frac{\lambda_1-\phi_1}{\phi_3} & \frac{\lambda_2-\phi_1}{\phi_3} \end{pmatrix}^{-1}.$$

Notice that in this case λ_2 is the bigger root, thus

$$\begin{aligned} \lambda_1 &= \frac{\phi_1+\phi_2-\sqrt{(\phi_1+\phi_4)^2-4(\phi_1\phi_4-\phi_2\phi_3)}}{2}, \\ \lambda_2 &= \frac{\phi_1+\phi_2+\sqrt{(\phi_1+\phi_4)^2-4(\phi_1\phi_4-\phi_2\phi_3)}}{2}. \end{aligned}$$

Therefore, $\kappa(\boldsymbol{\phi}) = \log(\lambda_2)$. From (11), one can write down the explicit infinite moving average representation of $\mathbf{X}_t$, i.e.,

$$\mathbf{X}_t = C_1 \sum_{i=0}^{\infty} \lambda_1^i \mathbf{Z}_{t-i} + C_2 \sum_{j=1}^{\infty} \lambda_2^{-j} \mathbf{Z}_{t+j}, \tag{27}$$

where the constant matrix is given by

$$C_1 = \begin{pmatrix} \frac{\lambda_2-\phi_1}{\lambda_2-\lambda_1} & \frac{-\phi_3}{\lambda_2-\lambda_1} \\ \frac{(\lambda_1-\phi_1)(\lambda_2-\phi_1)}{\phi_3(\lambda_2-\lambda_1)} & \frac{-(\lambda_1-\phi_1)}{\lambda_2-\lambda_1} \end{pmatrix},$$

and

$$C_2 = \begin{pmatrix} \frac{-(\lambda_1-\phi_1)}{\lambda_2-\lambda_1} & \frac{\phi_3}{\lambda_2-\lambda_1} \\ \frac{-(\lambda_1-\phi_1)(\lambda_2-\phi_1)}{\phi_3(\lambda_2-\lambda_1)} & \frac{\lambda_2-\phi_1}{\lambda_2-\lambda_1} \end{pmatrix}.$$

Note that the representations of C_1, C_2 and A are not unique. The expression in (27) essentially characterizes the mixture series $\mathbf{X}_t$ as a sum of a purely causal AR(1) process (the first component) and a purely noncausal AR(1) process (the second component).

In our simulation example, we generate the process $\mathbf{X}_t$ from the model (27) mentioned in Section 1. That is, we set

$$\begin{pmatrix} \phi_1 & \phi_3 \\ \phi_2 & \phi_4 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.6 \\ 0.6 & 1.7 \end{pmatrix},$$

which gives $\lambda_1 = 0.5$, $\lambda_2 = 2$,

$$A = \begin{pmatrix} 1 & 1 \\ -0.5 & 2 \end{pmatrix}$$

and

$$A^{-1} = \begin{pmatrix} 0.8 & -0.4 \\ 0.2 & 0.4 \end{pmatrix}.$$

For simplicity, set $\Sigma_0 = I_2$, $\nu =$ degrees of freedom $= 6$. By Theorem 4.7, the asymptotic variance of the MLE's can be calculated using numerical methods. The simulation is conducted in two cases. First, assume the covariance matrix and the degrees of freedom are given, and hence only $\boldsymbol{\phi}$ is to

be estimated. Second, consider a general case in which all parameters are to be estimated. In both cases, the noncausal vector AR process is generated using the method stated in Section 2. Time series of sizes 100, 200, 500 and 1000 are simulated respectively. For each time series, a vector AR model is fitted with MLE of the parameter, and the resulting estimates and their associated standard errors were computed. For each sample size, this procedure is replicated 15,000 times. The results from this experiment are summarized in the following tables.

6.1.1 Case I: estimating only the AR coefficients

Table 2 compares the empirical mean of the MLE's of $\boldsymbol{\phi}$ with its true value. The empirical standard errors are also recorded in the parentheses. Table 3 compares the empirical variance of $\sqrt{n-p}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)$ with its true value. The empirical variance is recorded in $(\cdot)$ and the subscripts of the parentheses stand for the sample size (1 for 100, 2 for 200, etc.). From Theorem 4.7, we know that the standard errors can also be calculated from the second derivative of the approximate likelihood function. Thus, for each simulated time series, standard errors of the MLE's can be estimated. The average of such estimates are recorded in $[\cdot]$. From the results, we see that it is often more accurate to use these estimates of the standard errors to describe the finite sample behavior of the MLE's than to use the asymptotic ones. It is also worth noticing that although Φ and Σ are both symmetric, $\hat{\phi}_2$ and $\hat{\phi}_3$ have different asymptotic distributions.

	True value	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
ϕ_1	0.8	0.8002 (0.1373)	0.8000 (0.0932)	0.7995 (0.0568)	0.8001 (0.0400)
ϕ_2	0.6	0.6084 (0.0986)	0.6047 (0.0669)	0.6025 (0.0423)	0.6010 (0.0293)
ϕ_3	0.6	0.6113 (0.1818)	0.6065 (0.1260)	0.6014 (0.0775)	0.6013 (0.0545)
ϕ_4	1.7	1.7161 (0.1707)	1.7072 (0.1150)	1.7041 (0.0729)	1.7019 (0.0509)

Table 2: The true and empirical mean (standard error) of the MLE's of ϕ_1, ϕ_2, ϕ_3 and ϕ_4.

6.1.2 Case II: estimating all parameters

In this case, neither the covariance matrix nor the degrees of freedom is given. A complete model is fitted to the simulated data. Table 3 is similar to Table 1 which compares the empirical averages of the MLE's with their corresponding true values. Since more parameters are included, the asymptotic variance of $\hat{\boldsymbol{\phi}}$ also increases. The theoretical standard errors of $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ are given in $(\cdot)_T$. The empirical standard errors of $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ are given in $(\cdot)$. From Theorem 4.7, we can also estimate the covariance matrix of $\sqrt{n-p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ for any particular time series, and thus the standard errors of parameters can be estimated as well. The empirical averages of such estimated standard errors are given in $[\cdot]$. Notice that the degree of freedom parameter ν has a very large standard error compared to other parameters. This also implies that ν in some sense is the most difficult to estimate and one should always be cautious about the estimation of ν. When the sample size is small, $\hat{\nu}$ could potentially be very inaccurate (see also Figure 2).

There are several issues when more parameters are estimated. First, when $-(n-p)^{-1}\partial^2 l_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ is used to estimate the covariance matrix of the MLE's, there is no guarantee that the estimated matrix is positive definite when the sample size is small. Second, there is no guarantee that the $\hat{\boldsymbol{\phi}}$ are in the domain of mixed models. Meaning, when the sample size is not big enough, the MLE's may appear in the causal region or in the purely noncausal region. For example, in our simulation, when the sample size is 100, we found 1078 of the 15,000 replications were in the causal region

	ϕ_1		ϕ_2		ϕ_3		ϕ_4	
ϕ_1	$(1.613)_T$		$(0.109)_T$		$(1.530)_T$		$(-0.664)_T$	
	$(1.865)_1$	$[1.704]_1$	$(0.086)_1$	$[0.114]_1$	$(1.642)_1$	$[1.603]_1$	$(-0.750)_1$	$[-0.701]_1$
	$(1.728)_2$	$[1.657]_2$	$(0.136)_2$	$[0.112]_2$	$(1.645)_2$	$[1.562]_2$	$(-0.729)_2$	$[-0.683]_2$
	$(1.609)_5$	$[1.625]_5$	$(0.095)_5$	$[0.110]_5$	$(1.531)_5$	$[1.537]_5$	$(-0.685)_5$	$[-0.667]_5$
	$(1.599)_{10}$	$[1.613]_{10}$	$(0.117)_{10}$	$[0.109]_{10}$	$(1.508)_{10}$	$[1.527]_{10}$	$(-0.661)_{10}$	$[-0.663]_{10}$
ϕ_2			$(0.860)_T$		$(0.247)_T$		$(0.529)_T$	
			$(0.963)_1$	$[0.935]_1$	$(0.287)_1$	$[0.272]_1$	$(0.540)_1$	$[0.570]_1$
			$(0.890)_2$	$[0.896]_2$	$(0.265)_2$	$[0.260]_2$	$(0.524)_2$	$[0.548]_2$
			$(0.894)_5$	$[0.875]_5$	$(0.239)_5$	$[0.252]_5$	$(0.538)_5$	$[0.539]_5$
			$(0.856)_{10}$	$[0.867]_{10}$	$(0.274)_{10}$	$[0.248]_{10}$	$(0.522)_{10}$	$[0.535]_{10}$
ϕ_3					$(2.992)_T$		$(-0.887)_T$	
					$(3.273)_1$	$[3.180]_1$	$(-1.010)_1$	$[-0.917]_1$
					$(3.158)_2$	$[3.074]_2$	$(-0.953)_2$	$[-0.902]_2$
					$(3.000)_5$	$[3.017]_5$	$(-0.929)_5$	$[-0.888]_5$
					$(2.972)_{10}$	$[2.996]_{10}$	$(-0.878)_{10}$	$[-0.883]_{10}$
ϕ_4							$(2.561)_T$	
							$(2.884)_1$	$[2.737]_1$
							$(2.634)_2$	$[2.651]_2$
							$(2.655)_5$	$[2.596]_5$
							$(2.591)_{10}$	$[2.579]_{10}$

Table 3: The true and empirical variance of the MLE's of ϕ_1, ϕ_2, ϕ_3 and ϕ_4. Results are already normalized by their corresponding sample sizes.

($\approx 7.2\%$) and 148 in the purely noncausal region ($\approx 0.99\%$). When the sample size is 200, 179 out of 15,000 replications are in the causal region ($\approx 1.2\%$) and 4 are in the purely noncausal region ($\approx 0.03\%$); when the sample size is 500 or 1000, the MLE's behaved well and are all in the domain of mixed models.

Finally, Figure 2 shows the convergence to normality. We take the parameter ν as an example. The histograms of the estimators for different sample sizes are plotted together with the corresponding asymptotic normal density curves as reference. As shown in Figure 2, when sample size is small, the empirical distribution of ν is highly skewed to the right. This skewness is corrected when sample size increases. When the sample size is 1000, one can see that the empirical distribution matches very well with the reference normal curve.

	True value	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
ϕ_1	0.8	0.8435 (0.3851) $(0.2985)_T$ [0.3380]	0.8220 (0.2394) $(0.2106)_T$ [0.2390]	0.8068 (0.1345) $(0.1330)_T$ [0.1399]	0.8026 (0.0932) $(0.0940)_T$ [0.0958]
ϕ_2	0.6	0.5274 (0.2612) $(0.1487)_T$ [0.2080]	0.5883 (0.1338) $(0.1049)_T$ [0.1263]	0.5992 (0.0665) $(0.0662)_T$ [0.0698]	0.5989 (0.0456) $(0.0468)_T$ [0.0476]
ϕ_3	0.6	0.5359 (0.5491) $(0.4614)_T$ [0.5126]	0.5938 (0.3695) $(0.3254)_T$ [0.3658]	0.5988 (0.2102) $(0.2055)_T$ [0.2167]	0.5989 (0.1444) $(0.1452)_T$ [0.1484]
ϕ_4	1.7	1.6297 (0.5576) $(0.3839)_T$ [0.4377]	1.6987 (0.3228) $(0.2708)_T$ [0.3034]	1.7067 (0.1693) $(0.1710)_T$ [0.1783]	1.7021 (0.1186) $(0.1029)_T$ [0.1226]
σ_1	1	1.1382 (0.5579) $(0.3115)_T$ [0.5939]	1.0561 (0.2693) $(0.2197)_T$ [0.3054]	1.0159 (0.1421) $(0.1387)_T$ [0.1552]	1.0075 (0.0987) $(0.0980)_T$ [0.1034]
σ_2	0	-0.0663 (0.3529) $(0.2444)_T$ [0.3866]	-0.0115 (0.2060) $(0.1724)_T$ [0.2210]	-0.0024 (0.1130) $(0.1089)_T$ [0.1188]	-0.0024 (0.0770) $(0.0769)_T$ [0.0798]
σ_3	1	1.0226 (0.5093) $(0.4063)_T$ [0.5600]	1.0264 (0.3137) $(0.2866)_T$ [0.3410]	1.0108 (0.1760) $(0.1810)_T$ [0.1906]	1.0040 (0.1233) $(0.1279)_T$ [0.1295]
ν	6	6.7925 (3.5194) $(2.4436)_T$ [6.5337]	6.4947 (2.2386) $(1.7235)_T$ [2.7905]	6.1386 (1.1061) $(1.0884)_T$ [1.2222]	6.0816 (0.7712) $(0.7692)_T$ [0.8119]

Table 4: The true and empirical mean of the MLE's of ϕ_1, ϕ_2, ϕ_3, ϕ_4, σ_1, σ_2, σ_3 and ν. The empirical standard errors are given in $(\cdots)$, the theoretical standard errors are given in $(\cdots)_T$ and the averages of the estimated standard errors are given in $[\cdots]$.

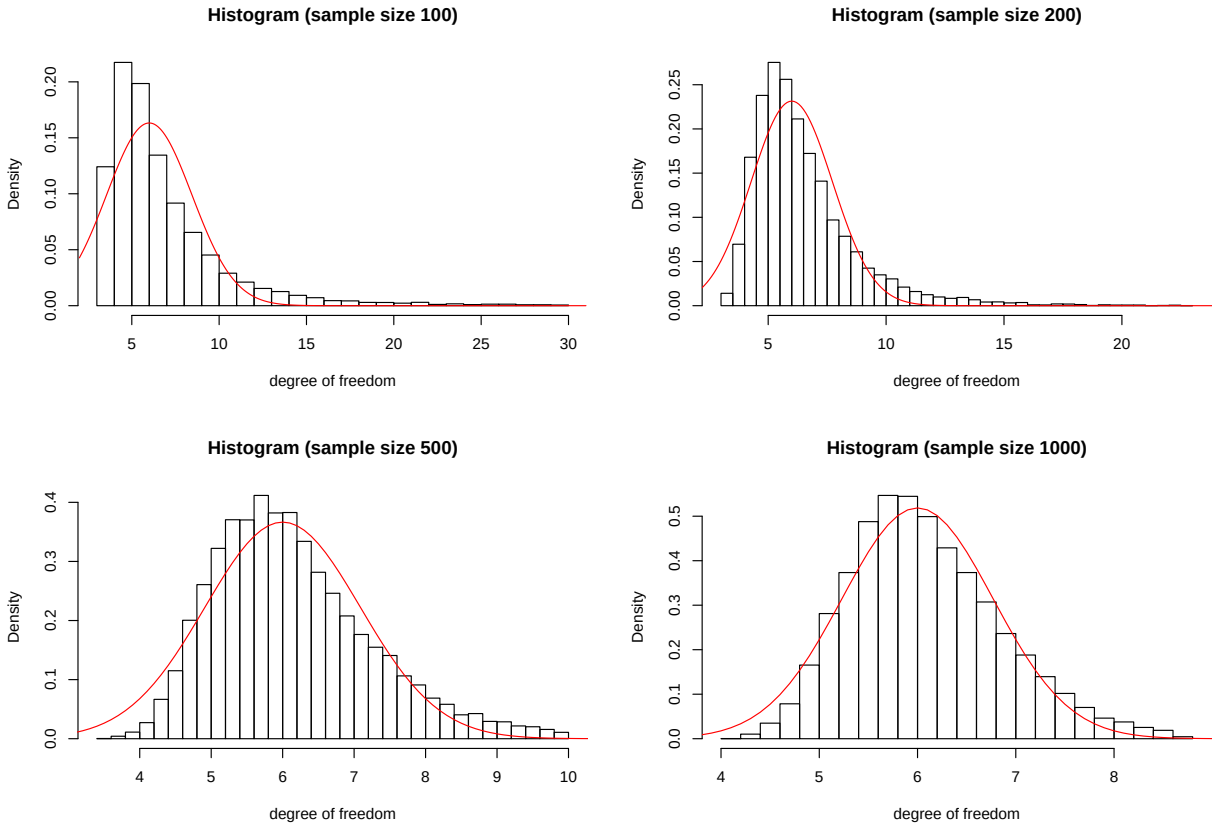

Figure 2: Comparison of empirical histogram of $\hat{\nu}$ and its asymptotic normal distribution.

6.2 Real data examples

In this subsection, we consider two datasets that were previously studied by [11]. Conventionally, in many economic applications, VAR models are fitted under causality assumptions. Lanne and Saikkonen argued that, it would be incorrect to base a test of theory on the assumption of causality in the presence of noncausality. For the following two datasets, conventional causal VAR models are first fitted to the data. Standard procedures for testing Gaussianity are then carried out. It is shown that the residuals from these causal models present some non-Gaussian (heavy tail) behavior (see also [11]). Then it makes sense to include both causal and noncausal models in our fitting procedure, the model that maximizes the likelihood is selected as the best fit. For both examples, multivariate t-distribution is used for the noise.

6.2.1 Fiscal foresight

First, a simple trivariate VAR model for the demeaned differences of U.S. GDP, total government expenditure, and total government revenue (all in real per capita terms). The quarterly data from 1955:1 to 2000:4 (184 observations) were previously used by [12], who also provide a detailed description of the construction of the variables. Figure 3 contains time series plots of the three series. [11] suggested using a VAR model of order 2 for this data based on AIC and BIC. Table 5 compares the log-likelihood function values of the best fit of casual Gaussian (CG) VAR(2), causal non-Gaussian (CN) VAR(2), purely noncausal (PN) VAR(2) and the mixed (MX) VAR(2). The estimators and their standard errors from the mixed model fit are summarized in Table 6.

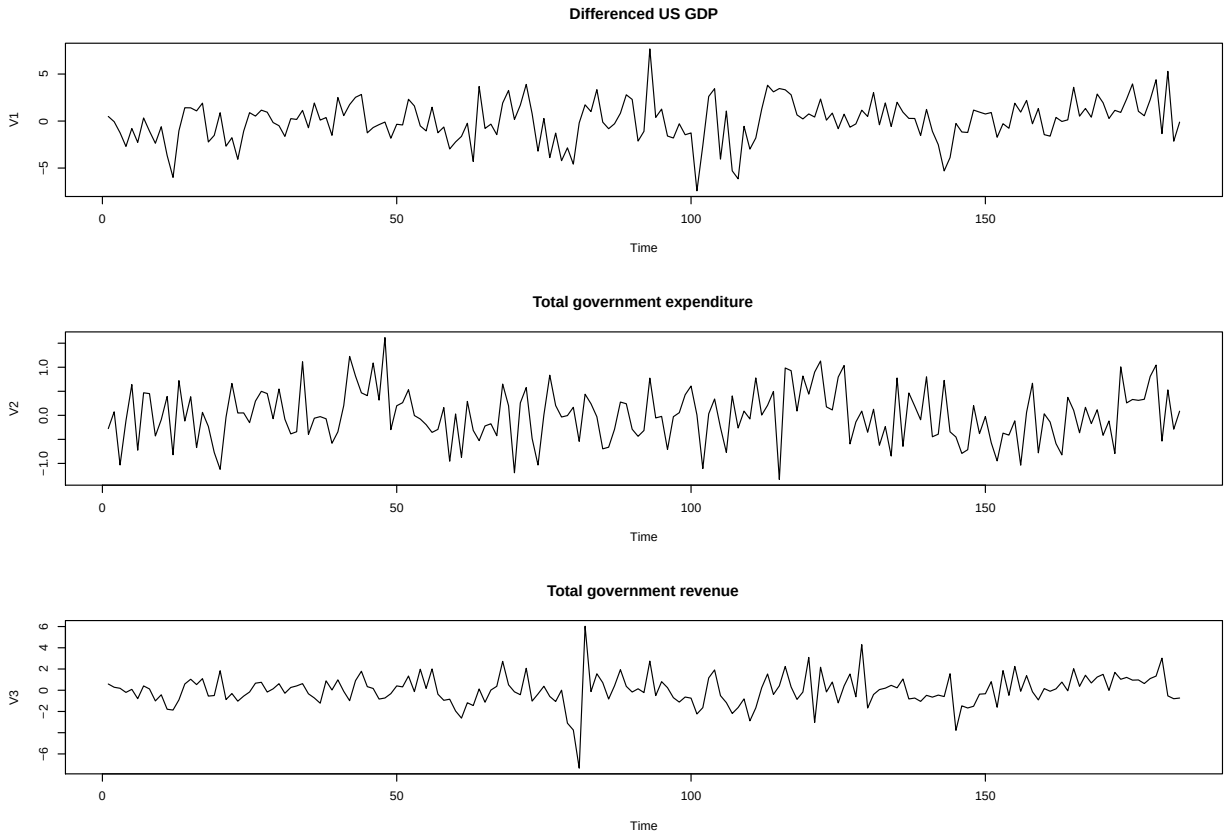

Figure 3: Time series plot of the GDP data.

Model assumption	CG	CN	PN	MX
Log-likelihood	-819.230	-802.865	-800.355	-791.270

Table 5: Comparison of log-likelihood. (CG: Causal Gaussian; CN: Causal Non-Gaussian; PN: Purely Noncausal; MX: Mixed)

Φ_1			Φ_2			Σ			ν
3.516	3.710	-3.799	1.789	-6.268	-1.162	49.108	-12.746	-14.666	
(1.226)	(3.631)	(1.527)	(0.924)	(2.946)	(0.673)	(9.654)	(5.728)	(7.215)	
-0.512	-3.104	0.679	-0.586	2.658	0.295	-12.746	5.135	4.019	5.751
(0.456)	(1.267)	(0.625)	(0.311)	(1.081)	(0.312)	(5.728)	(3.373)	(3.494)	(1.399)
-1.099	-1.811	2.918	-0.505	1.149	1.257	-14.666	4.019	9.376	
(0.475)	(1.889)	(0.762)	(0.435)	(1.534)	(0.314)	(7.215)	(3.494)	(5.203)	

Table 6: The MLE's of the parameters and their associated standard errors for the GDP data.

Notice that the degrees of freedom parameter ν is estimated as 5.75 which confirms the heavy tail behavior of the data. The AR polynomial $\det\left(I - \Phi_1 z - \Phi_2 z^2\right)$, where Φ_1 and Φ_2 are given in

Table 5, has three roots inside the unit circle and three roots outside the unit circle. In theory, one could solve for Ψ_1 and Ψ_2 which satisfy

$$I - \Phi_1 z - \Phi_2 z^2 = (I - \Psi_1 z)(I - \Psi_2 z),$$

or $\Phi_1 = \Psi_1 + \Psi_2$ and $\Phi_2 = -\Psi_1 * \Psi_2$. Unlike the univariate case, where there is a unique solution for this decomposition, there are usually more than one solution in the multivariate case. For example, in our case, one possible solution is

$$\Psi_1 = \begin{pmatrix} 3.994 & 1.706 & -3.719 \\ -0.390 & -3.740 & 0.661 \\ -1.010 & -2.569 & 3.309 \end{pmatrix},$$

$$\Psi_2 = \begin{pmatrix} -0.478 & 2.003 & -0.080 \\ -0.122 & 0.636 & 0.018 \\ -0.089 & 0.758 & -0.390 \end{pmatrix}.$$

It is also easy to check that Ψ_1 has all of its eigenvalues outside the unit circle and Ψ_2 has all of its eigenvalues inside the unit circle. Thus this is similar to the setup of Lanne and Saikkonen's model (5), where one can perfectly separate the causal component and the purely noncausal component. However, as mentioned above, we can also obtain other decompositions, for instance, another possible solution would be

$$\Psi_1 = \begin{pmatrix} 4.011 & 4.130 & -7.087 \\ -0.390 & -3.730 & 0.646 \\ -0.999 & -1.053 & 1.202 \end{pmatrix},$$

$$\Psi_2 = \begin{pmatrix} -0.495 & -0.420 & 3.288 \\ -0.123 & 0.625 & 0.033 \\ -0.099 & -0.758 & 1.716 \end{pmatrix}.$$

In this case, Ψ_1 has two eigenvalues outside the unit circle and one inside the unit circle, whereas Ψ_2 has two eigenvalues inside the unit circle and one outside. This means that very different product representations may correspond to the same VAR model. Therefore, any theory or interpretation obtained from the product representation may be misleading. This also suggests that our model setup is more general than the one Lanne and Saikkonen used.

Finally, the ACF's of the residuals can be used to show that the noncausal model is more appropriate than the causal model. Figure 4 contains the ACF's (including the cross ACF's) of the residuals from a noncausal model. It shows that the model effectively removes the serial dependence of the original data. Although not included here, the ACF's of the residuals from a causal Gaussian model look very similar to Figure 4, which means the causal Gaussian model also does a good job in removing the serial dependence structure of the data. However, the ACF's of the squares of the residuals (Figure 5) from the causal Gaussian model show serial correlation again. This is evidence that the fitted residuals are uncorrelated but not independent. In comparison, Figure 6 contains the ACF's of the squares of the residuals from a noncausal model. It can be seen that the squares of the residuals from this noncausal model are also uncorrelated. This indicates that the noncausal model is a better fit than the causal Gaussian model in the sense that the residuals are more independent.

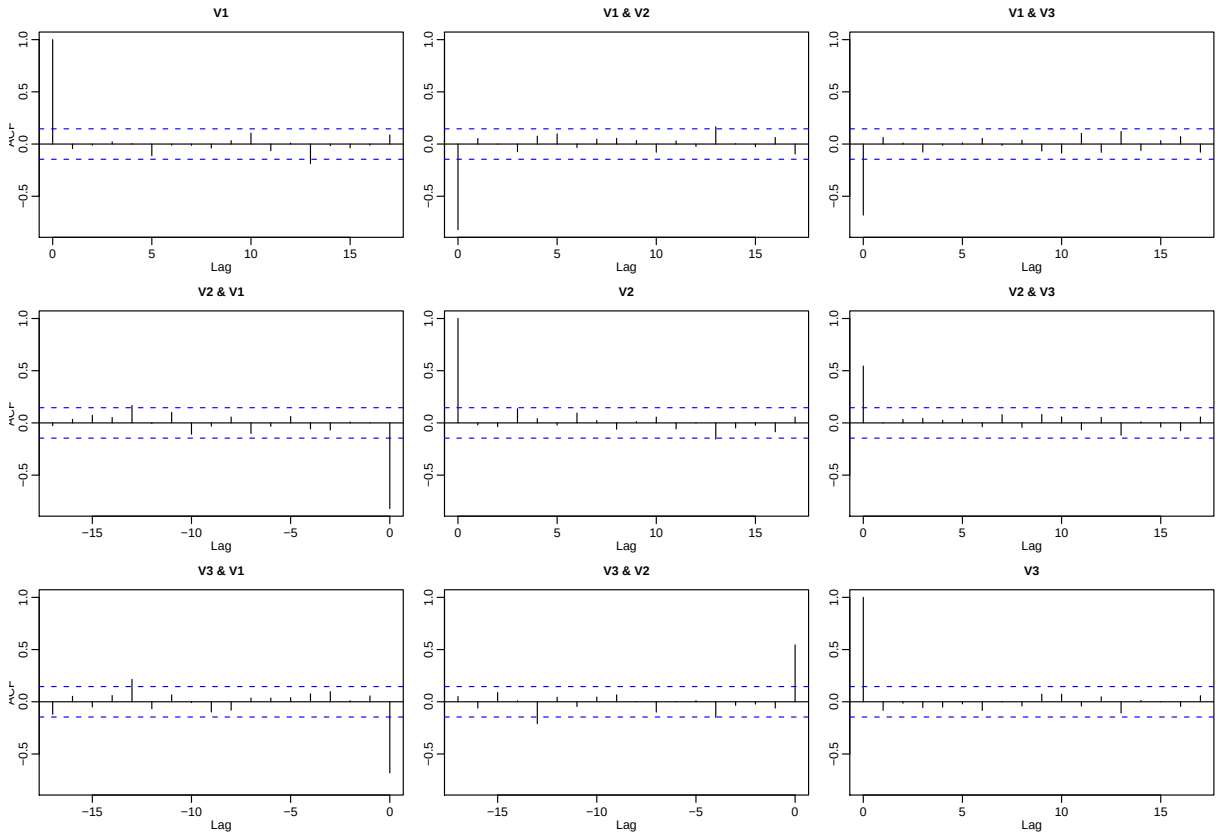

Figure 4: Cross correlation of the residuals from the mixed non-Gaussian model.

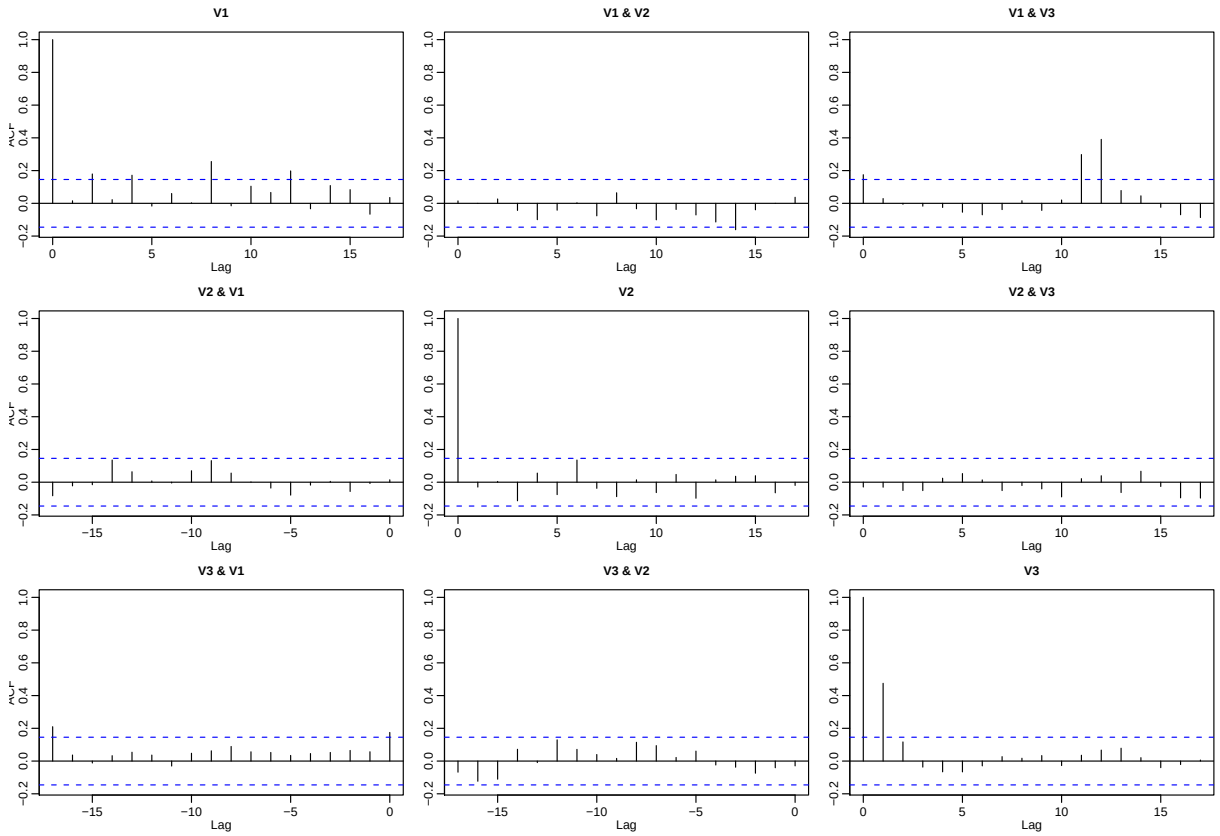

Figure 5: Cross correlation of the squares of the residuals from the causal Gaussian model.

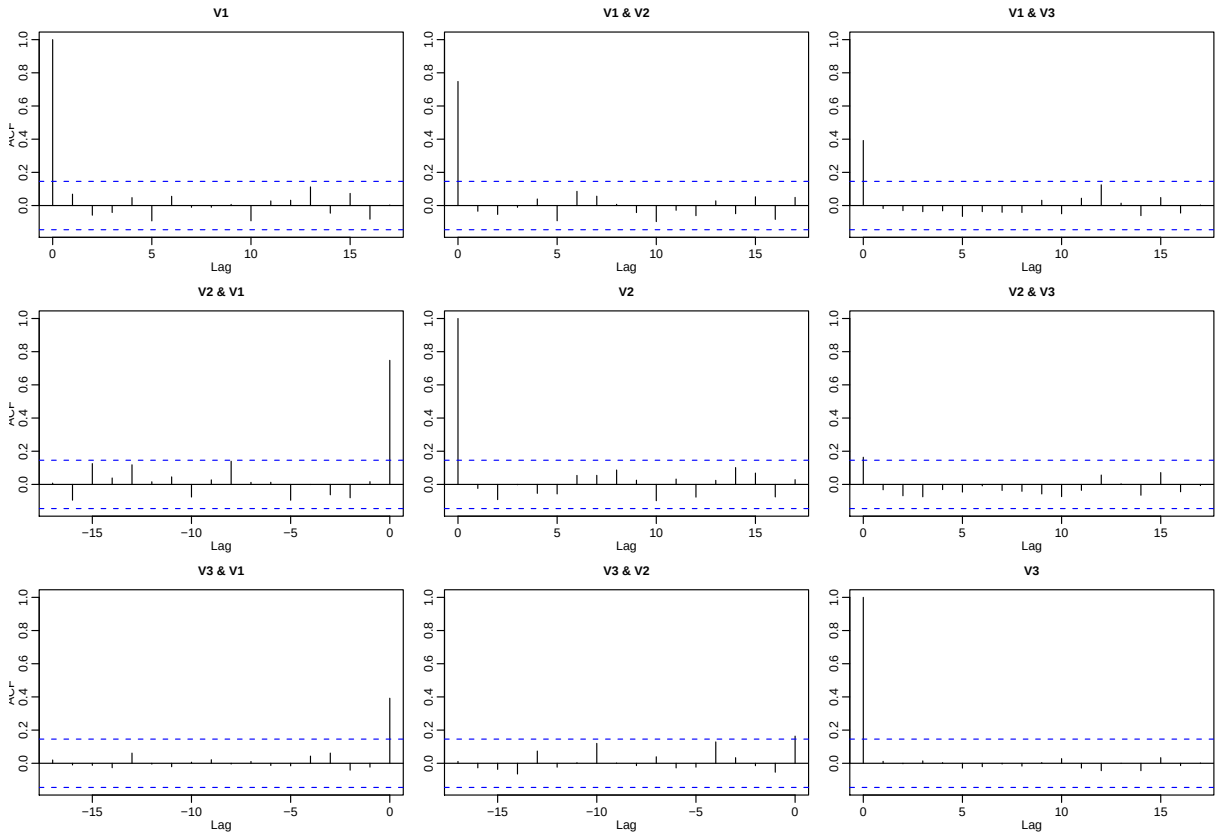

Figure 6: Cross correlation of the squares of the residuals from the mixed non-Gaussian model.

Figure 7 shows the impulse response coefficients of the noncausal model. Since the estimator of Σ is not diagonal, the Cholesky decomposition, as discussed earlier in Section 2, is applied to make the three components of the noise uncorrelated. From Figure 7, we can see that a large positive shock in the first component of the transformed noise would contribute a big negative bump to the noncausal element in the change of GDP and little to the noncausal components for changes of total expenditure and total revenue. A large shock in the second component of the transformed noise would have the most impact on the change of GDP in the noncausal part. It would have less impact on the change of total expenditure and minimum impact on the change of total revenue. Finally, a large shock in the third component of the transformed noise would mainly influence the noncausal pieces of the GDP and revenue components of the response. As for the causal part of the response, large shocks in the second and third components would contribute some reverberations in the causal element of the change of GDP and the change of total revenue.

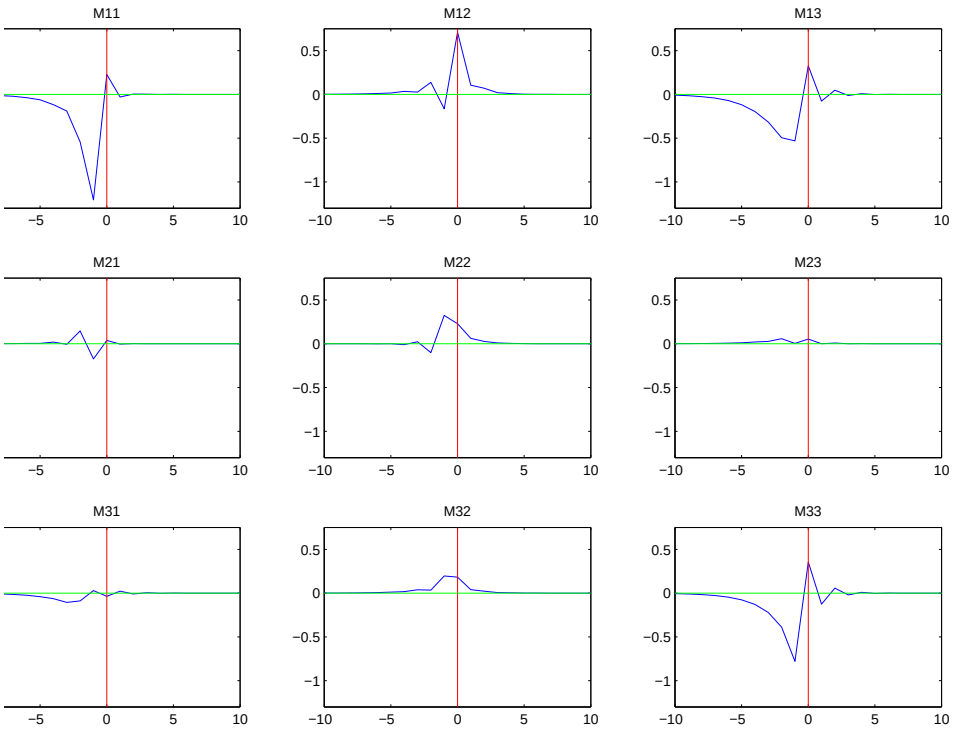

Figure 7: Impulse response coefficients of the noncausal model fitted to the GDP data.

6.2.2 Term structure of interest rates

For the second example, we concentrate on a bivariate VAR model for the demeaned change in the three-month interest rate and the spread between the ten-year and three-month interest rates (quarter-end yields on U.S. zero-coupon bonds) from 1970:1 to 1998:4 (116 observations). Figure 8 contains time series plots of the two series. The data was previously used by [7] and are provided on his website. Lanne and Saikkonen suggested using a VAR model of order 3 for this data. Table 7 compares the log-likelihood function values of the best fit of casual Gaussian (CG) VAR(3), causal non-Gaussian (CN) VAR(3), purely noncausal non-Gaussian (PN) VAR(3) and mixed (MX) VAR(3). The estimators and their standard errors from the mixed model we found are summarized in Table 8.

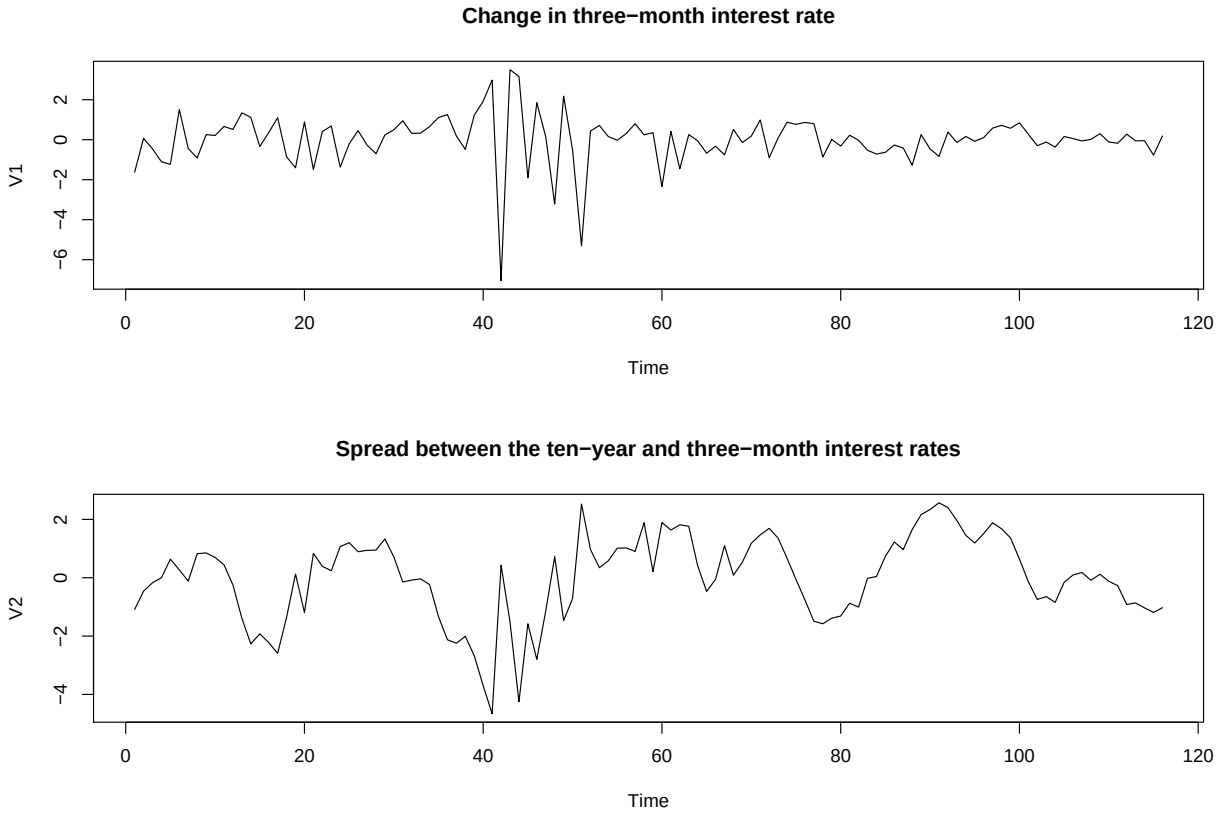

Figure 8: Time series plot of the interest rate data.

Model assumption	CG	CN	PN	MX
Log-likelihood	-256.624	-237.345	-235.481	-229.054

Table 7: Comparison of log-likelihood. (CG: Causal Gaussian; CN: Causal Non-Gaussian; PN: Purely Noncausal; MX: Mixed)

Φ_1		Φ_2		Φ_3		Σ		ν
0.789	0.009	0.434	0.110	0.728	-0.268	0.774	-0.448	2.806
(0.289)	(0.255)	(0.202)	(0.327)	(0.128)	(0.232)	(0.245)	(0.145)	(0.715)
-0.548	0.818	-0.209	0.014	-0.552	0.136	-0.448	0.393	
(0.170)	(0.174)	(0.128)	(0.240)	(0.079)	(0.175)	(0.145)	(0.106)	

Table 8: The MLE's of the parameters and their associated standard errors for the interest rate data.

Notice that the degrees of freedom ν is estimated to be 2.806, which is corresponding to a very heavy tail distribution. In this case, the noise process $\mathbf{Z}_t$ has not have a third moment, which leads to the violation of some of the regularity conditions (for example, A.4.). Thus,

one should be cautious in using these estimators. Despite the heavy-tailness, the AR polynomial $\det\left(I - \Phi_1 z - \Phi_2 z^2 - \Phi_3 z^3\right)$ has only one root inside the unit circle and the other five roots are all outside the unit circle. Thus it is impossible to have a representation like Lanne and Saikkonen's model, i.e., there is no way to separate the polynomial into a causal component and a purely noncausal component.

Similar analyses as in the previous example can also be used to show that the noncausal model is more appropriate than the causal model. By looking at the ACF's of the residuals from the models, we would see similar results as in the previous example. That is, both the causal Gaussian model and the noncausal model do well in terms of removing the serial dependence structure of the original data. However, the ACF's of the squares of the residuals (Figure 9) from the causal Gaussian model present serial correlation. On the other hand, the ACF's of the squares of the residuals from the noncausal model are also proved to be uncorrelated, see Figure 10. This indicates that the noncausal model does a better job than the causal Gaussian model in terms of producing more independent residuals.

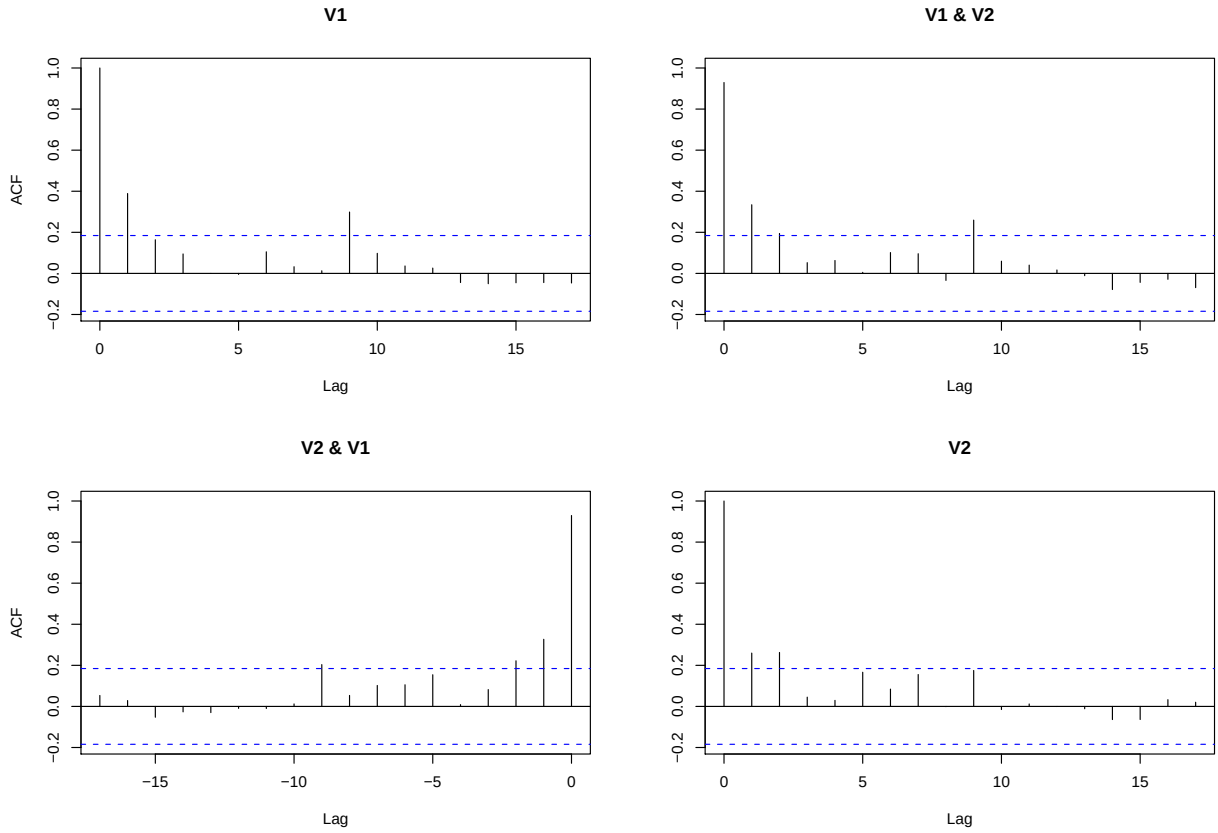

Figure 9: Cross correlation of the squares of the residuals from the causal Gaussian model.

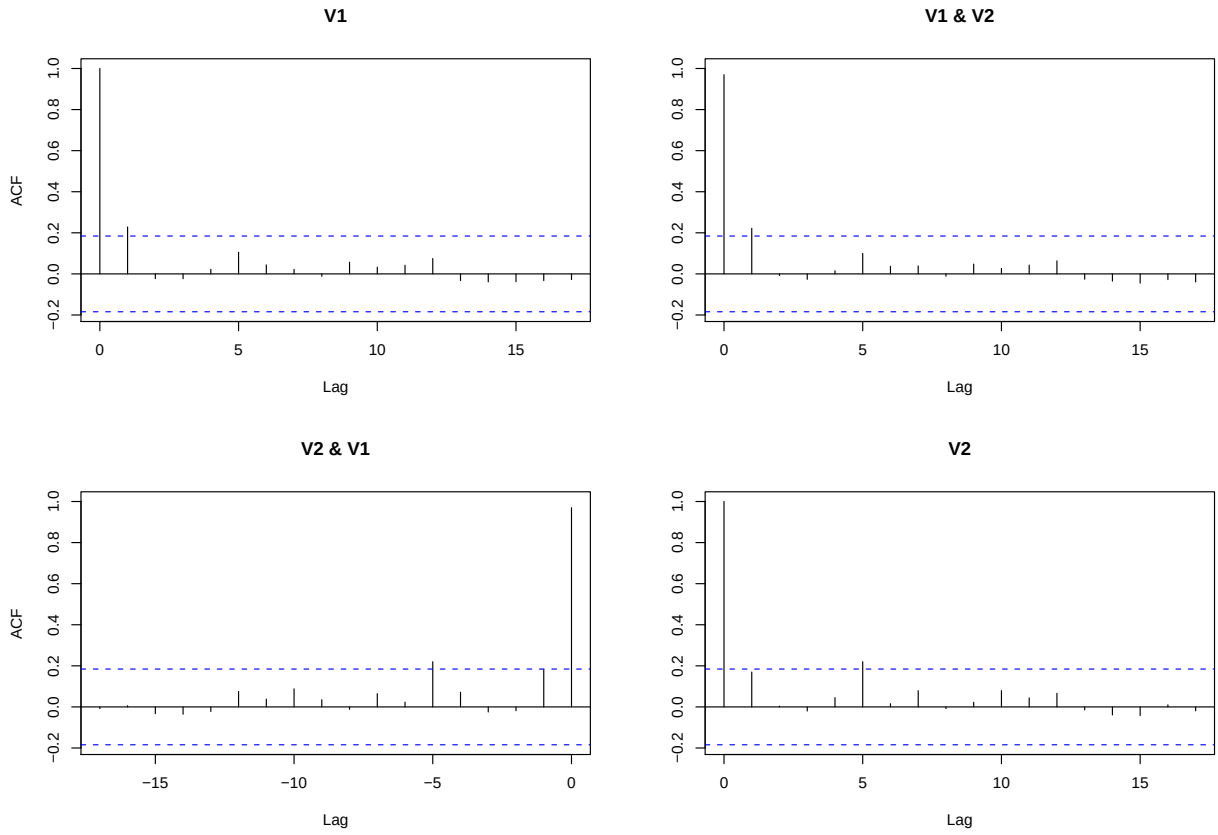

Figure 10: Cross correlation of the squares of the residuals from the mixed non-Gaussian model.

As before, we plot the impulse response coefficients for this noncausal model where the Cholesky decomposition is used to de-correlate the noise components. From Figure 11, we can see the non-causality mainly comes from the first component of the transformed noise. The second component of the transformed noise barely contributes anything to the noncausal element. The first component of the transformed noise also contributes some reverberation to the causal components of both differenced short-term rates and rates spread while the second component only impacts the causal component of the rates spread.

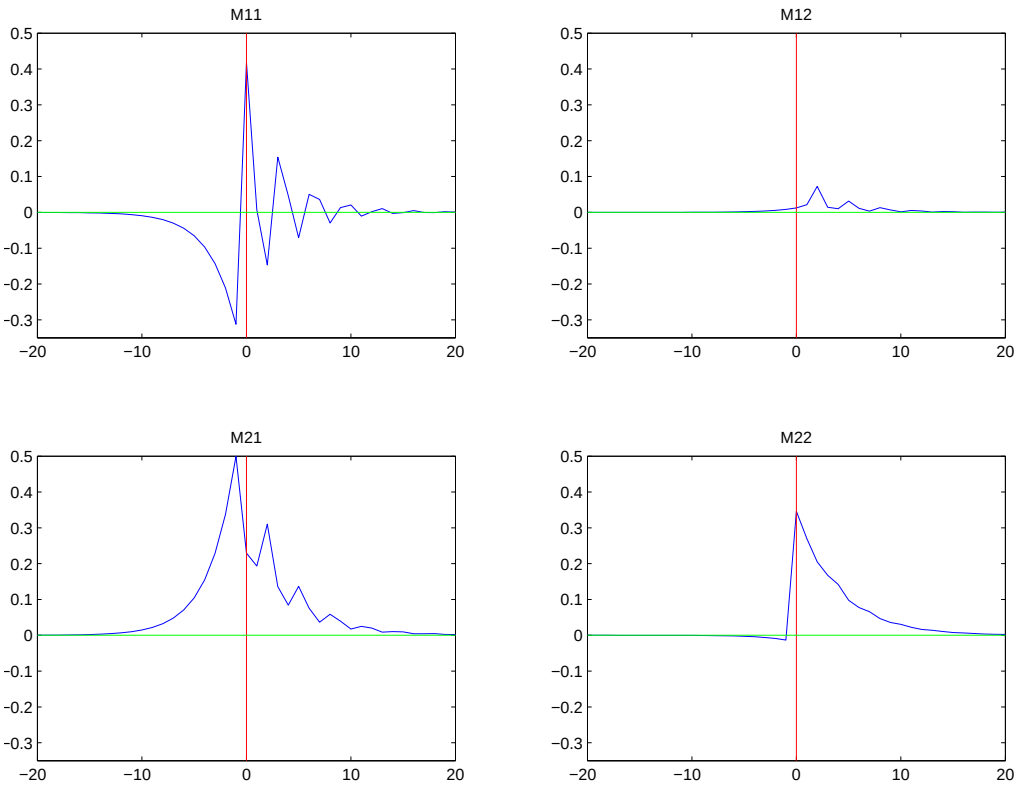

Figure 11: Impulse response coefficients of the noncausal model fitted to the interest rate data.

7 Conclusion

A general noncausal non-Gaussian VAR model has been proposed in this paper. This process has a state-space representation that neatly separate the causal and noncausal components. Model simulation and model selection techniques are introduced. This representation is the backbone for simulating realizations from a noncausal VAR model as well as providing a mechanism for the likelihood function. In particular, maximum likelihoods can be computed. Allowing noncausality in the VAR model opens doors to a much larger pool of models, and in many cases these noncausal models provide a better fit to the data, in the sense that the residuals appear more "white". The models were fitted to two macro time series and gave an improved fit as compared to the traditional causal VAR model.

APPENDIX

A Derivatives of log-likelihood

This section contains the analytical form of the first and second derivatives of function $g_i(\boldsymbol{\theta})$ from Section (4.2). Straightforward calculations show that

$$\begin{aligned}
\frac{\partial g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}} &= 2h(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu}) \cdot \frac{\partial}{\partial \boldsymbol{\phi}}\mathbf{Z}_i(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi}) + \frac{\partial \kappa(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \\
&= -2\begin{pmatrix} \mathbf{X}_{i-1} \\ \vdots \\ \mathbf{X}_{i-p} \end{pmatrix} \otimes \left(\Sigma^{-1/2}\mathbf{e}_t(\boldsymbol{\theta})\right) + \frac{\partial \kappa(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \\
&= -2\mathbf{Y}_{i-1} \otimes \left(\Sigma^{-1/2}\mathbf{e}_t(\boldsymbol{\theta})\right) + \frac{\partial \kappa(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}; \\
\frac{\partial g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}} &= -h(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})D_m^T \cdot (\Sigma^{-1}\otimes\Sigma^{-1})(\mathbf{Z}_i(\boldsymbol{\phi})\otimes\mathbf{Z}_i(\boldsymbol{\phi})) - \frac{1}{2}D_m^T\mathrm{vec}(\Sigma^{-1}) \\
&= -D_m^T(\Sigma^{-1}\otimes\Sigma^{-1}) \cdot \left(\mathbf{Z}_i(\boldsymbol{\theta})\otimes\Sigma^{1/2}\mathbf{e}_i(\boldsymbol{\theta}) + \frac{1}{2}\mathrm{vec}(\Sigma)\right); \\
\frac{\partial g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\nu}} &= \frac{1}{f(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})} \cdot \frac{\partial}{\partial \boldsymbol{\nu}} f(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu}).
\end{aligned}$$

Also notice that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\phi}^T}\mathbf{e}_t(\boldsymbol{\theta}) &= -h(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})\Sigma^{-1/2}(\mathbf{Y}_{i-1}^T\otimes I_m) \\
&\quad -2h'(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})\Sigma^{-1/2}\cdot\mathbf{Z}_i(\boldsymbol{\phi})\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}(\mathbf{Y}_{i-1}^T\otimes I_m).
\end{aligned}$$

The second derivatives can then be calculated as

$$\begin{aligned}
\frac{\partial^2 g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T} &= -2\mathbf{Y}_{i-1}\otimes I_m\Sigma^{-1/2}\frac{\partial}{\partial \boldsymbol{\phi}^T}\mathbf{e}_t(\boldsymbol{\theta}) + \frac{\partial^2\kappa(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}; \\
\frac{\partial^2 g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\sigma}^T} &= 2h(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})\left(\mathbf{Z}_i^T(\boldsymbol{\phi})\otimes(\mathbf{Y}_{i-1}\otimes I_m)\right)(\Sigma^{-1}\otimes\Sigma^{-1})D_m \\
&\quad +2h'(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})\cdot(\mathbf{Y}_{i-1}\otimes I_m)\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi})\cdot\left(\mathbf{Z}_i^T(\boldsymbol{\phi})\otimes\mathbf{Z}_i^T(\boldsymbol{\phi})\right)(\Sigma^{-1}\otimes\Sigma^{-1})D_m; \\
\frac{\partial^2 g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\nu}^T} &= -2\mathbf{Y}_{i-1}\otimes\left(\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi})\right)\cdot\frac{\partial}{\partial\boldsymbol{\nu}^T}h(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu}); \\
\frac{\partial^2 g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}\partial \boldsymbol{\sigma}^T} &= h(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})\cdot(\mathbf{Z}_i^T(\boldsymbol{\phi})\otimes\mathbf{Z}_i^T(\boldsymbol{\phi})\otimes D_m^T)\cdot(I_m\otimes K_{mm}\otimes I_m) \\
&\quad \cdot\left[\Sigma^{-1}\otimes\Sigma^{-1}\otimes\mathrm{vec}(\Sigma^{-1}) + \mathrm{vec}(\Sigma^{-1})\otimes\Sigma^{-1}\otimes\Sigma^{-1}\right]D_m \\
&\quad +h'(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})D_m^T\cdot(\Sigma^{-1}\otimes\Sigma^{-1})\left(\mathbf{Z}_i(\boldsymbol{\phi})\mathbf{Z}_i^T(\boldsymbol{\phi})\otimes\mathbf{Z}_i(\boldsymbol{\phi})\mathbf{Z}_i^T(\boldsymbol{\phi})\right)(\Sigma^{-1}\otimes\Sigma^{-1})D_m \\
&\quad +\frac{1}{2}D_m^T(\Sigma^{-1}\otimes\Sigma^{-1})D_m; \\
\frac{\partial^2 g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}\partial \boldsymbol{\nu}^T} &= -D_m^T(\Sigma^{-1}\otimes\Sigma^{-1})(\mathbf{Z}_i(\boldsymbol{\phi})\otimes\mathbf{Z}_i(\boldsymbol{\phi}))\cdot\frac{\partial}{\partial\boldsymbol{\nu}^T}h(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu}); \\
\frac{\partial^2 g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\nu}\partial \boldsymbol{\nu}^T} &= -\frac{\partial f(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})/\partial\boldsymbol{\nu}}{f^2(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})}\cdot\frac{\partial f(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})}{\partial\boldsymbol{\nu}^T} \\
&\quad +\frac{\partial^2 f(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})/\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^T}{f(\mathbf{Z}_i^T(\boldsymbol{\phi})\Sigma^{-1}\mathbf{Z}_i(\boldsymbol{\phi});\boldsymbol{\nu})};
\end{aligned}$$

B Proof for Theorem 4.5.

Proof. It is worth noting that, by the independence of $\widetilde{\mathbf{Y}}_{p,1}$ and $\widetilde{\mathbf{Y}}_{n,2}$ from $\mathbf{Z}_{p+1}(\boldsymbol{\phi}_0), \ldots, \mathbf{Z}_n(\boldsymbol{\phi}_0)$, $\tilde{l}_n(\boldsymbol{\theta})$ in (20) is also a complete likelihood for any choice of initial distributions $p_1(\cdot)$ and $p_2(\cdot)$ for $\widetilde{\mathbf{Y}}_{p,1}$ and $\widetilde{\mathbf{Y}}_{n,2}$. In fact, one can choose $p_1(\cdot)$ and $p_2(\cdot)$ independent of $\boldsymbol{\theta}$, then such a complete likelihood will have the following properties:

$$\widetilde{\mathbf{E}}\left(\frac{\partial \tilde{l}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) = \mathbf{0}, \tag{28}$$

and

$$\widetilde{\mathbf{Var}}\left(\frac{\partial \tilde{l}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) = \widetilde{\mathbf{E}}\left(\frac{\partial \tilde{l}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)^2 = -\widetilde{\mathbf{E}}\left(\frac{\partial^2 \tilde{l}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right). \tag{29}$$

Here, $\widetilde{\mathbf{E}}(\cdot)$ denotes the expectation under the measure with the new choices of $p_1(\cdot)$ and $p_2(\cdot)$. For convenience, we will call this the star measure in the following context, and we will continue to use $\mathbf{E}(\cdot)$ to denote the expectation under the measure when $p_1(\cdot)$ and $p_2(\cdot)$ are chosen to be the stationary distribution. This will be referred to as the stationary measure. From (28), we have

$$\begin{aligned} 0 &= \frac{1}{n-p}\widetilde{\mathbf{E}}\left(\frac{\partial \tilde{l}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) \\ &= \frac{\widetilde{\mathbf{E}}\left(\partial \log(p_1(\widetilde{\mathbf{Y}}_{p,1})p_2(\widetilde{\mathbf{Y}}_{n,2}) \cdot |\det(\mathbf{A})|^{-1})/\partial \boldsymbol{\theta}\right)}{n-p} + \frac{1}{n-p}\widetilde{\mathbf{E}}\left(\frac{\partial l_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) \\ &= o(1) + \frac{1}{n-p}\sum_{i=p+1}^{n} \widetilde{\mathbf{E}}\left(\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right). \end{aligned}$$

Next we show that the average of the expectations under the new measure would converge to a constant which is $\mathbf{E}\left(\partial g_i(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}\right)$. From the calculation in Appendix A, we know that $\{\partial g_i(\boldsymbol{\theta}_0)/\partial \boldsymbol{\sigma}\}$ and $\{\partial g_i(\boldsymbol{\theta}_0)/\partial \lambda\}$ are two iid sequences only depending on the true values of $\{\mathbf{Z}_i\}$. Since the $\mathbf{Z}_i$ have the same distribution under either the stationary measure or the star measure, the identities

$$\begin{aligned} \widetilde{\mathbf{E}}\left(\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\sigma}}\right) &= \mathbf{E}\left(\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\sigma}}\right), \\ \widetilde{\mathbf{E}}\left(\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial \lambda}\right) &= \mathbf{E}\left(\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial \lambda}\right), \end{aligned} \tag{30}$$

follow easily. Unfortunately, $\{\partial g_i(\boldsymbol{\theta}_0)/\partial \boldsymbol{\phi}\}$ is not an iid sequence because of the $\mathbf{Y}_{i-1}$ term in it. However, we can use a similar argument as in Section 2 equation (11) to derive a truncated representation for $\mathbf{Y}_i$, i.e., for any $p < i < n$,

$$\mathbf{Y}_i = \mathbf{A}\mathbf{F}_{i-p}\mathbf{A}^{-1}\mathbf{Y}_p + \sum_{j=i-n}^{i-p-1} \mathbf{A}\mathbf{F}_j\mathbf{A}^{-1}\mathbf{Z}^*_{i-j} + \mathbf{A}\mathbf{F}_{i-n}\mathbf{A}^{-1}\mathbf{Y}_n. \tag{31}$$

Now let $\mathbf{Y}^*_p$ and $\mathbf{Y}^*_n$ be the initial values generated under the new choices of p_1 and p_2 and let $\mathbf{Y}^*_i$ $(p < i < n)$ be the random variables under the star measure. When $i = \lfloor n/2 \rfloor$, we have

$$\mathbf{Y}^*_{\lfloor n/2 \rfloor} = \mathbf{A}\mathbf{F}_{\lfloor n/2 \rfloor - p}\mathbf{A}^{-1}\mathbf{Y}^*_p + \sum_{j=\lfloor n/2 \rfloor - n}^{\lfloor n/2 \rfloor - p-1} \mathbf{A}\mathbf{F}_j\mathbf{A}^{-1}\mathbf{Z}^*_{\lfloor n/2 \rfloor - j} + \mathbf{A}\mathbf{F}_{\lfloor n/2 \rfloor - n}\mathbf{A}^{-1}\mathbf{Y}^*_n. \tag{32}$$

By the definition of $\mathbf{F}_j \to 0$ as $j \to \pm\infty$ at a geometric rate. Therefore, it is easy to show that, regardless of the choice of $\mathbf{Y}_p^*$ and $\mathbf{Y}_n^*$, $\mathbf{Y}_{\lfloor n/2 \rfloor}^*$ and $\mathbf{Y}_{\lfloor n/2 \rfloor} \stackrel{\mathrm{d}}{=} \mathbf{Y}_{p+1}$ have the same distribution as $n \to \infty$. Using stationarity of $\{\partial g_i(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}\}$ under the stationary measure, we obtain the following relationship,

$$\begin{aligned} \mathbf{E}\left(\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) &= \lim_{n\to\infty} \mathbf{E}\left(\frac{\partial g_{\lfloor n/2 \rfloor}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) = \lim_{n\to\infty} \widetilde{\mathbf{E}}\left(\frac{\partial g_{\lfloor n/2 \rfloor}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) \\ &= \lim_{n\to\infty} \frac{1}{n-p} \sum_{i=p+1}^{n} \widetilde{\mathbf{E}}\left(\frac{\partial g_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) = 0 \end{aligned} \tag{33}$$

Similarly, for the variance of the score, we have

$$\begin{aligned} & \widetilde{\mathbf{E}}\left(\frac{1}{\sqrt{n-p}} \frac{\partial \log(p_1(\widetilde{\mathbf{Y}}_{p,1}) p_2(\widetilde{\mathbf{Y}}_{n,2}) \cdot |\det(\mathbf{A})|^{-1})}{\partial \boldsymbol{\theta}} + \frac{1}{\sqrt{n-p}} \frac{\partial l_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)^2 \\ = & \frac{1}{n-p} \widetilde{\mathbf{Var}}\left(\frac{\partial \tilde{l}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) \\ = & -\frac{1}{n-p} \widetilde{\mathbf{E}}\left(\frac{\partial^2 \tilde{l}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) \\ = & -\frac{1}{n-p} \widetilde{\mathbf{E}}\left(\frac{\partial^2 \log(p_1(\widetilde{\mathbf{Y}}_{p,1}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) - \frac{1}{n-p} \widetilde{\mathbf{E}}\left(\frac{\partial^2 \log(p_2(\widetilde{\mathbf{Y}}_{n,2}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) \\ & -\frac{1}{n-p} \widetilde{\mathbf{E}}\left(\frac{\partial^2 \log(|\det(\mathbf{A})|^{-1})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) - \frac{1}{n-p} \widetilde{\mathbf{E}}\left(\frac{\partial^2 l_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) \\ = & o(1) - \frac{1}{n-p} \sum_{i=p+1}^{n} \widetilde{\mathbf{E}}\left(\frac{\partial^2 g_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right). \end{aligned}$$

Using a similar argument as for the expectation of the score, we can establish

$$\begin{aligned} \mathcal{I}_{\theta\theta} &= \lim_{n\to\infty} \mathbf{E}\left(\frac{1}{\sqrt{n-p}} \frac{\partial l_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)^2 = \lim_{n\to\infty} \widetilde{\mathbf{E}}\left(\frac{1}{\sqrt{n-p}} \frac{\partial l_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)^2 \\ &= -\lim_{n\to\infty} \frac{1}{n-p} \sum_{i=p+1}^{n} \widetilde{\mathbf{E}}\left(\frac{\partial^2 g_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) = -\lim_{n\to\infty} \widetilde{\mathbf{E}}\left(\frac{\partial^2 g_{\lfloor n/2 \rfloor}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) \\ &= -\lim_{n\to\infty} \mathbf{E}\left(\frac{\partial^2 g_{\lfloor n/2 \rfloor}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) = -\mathbf{E}\left(\frac{\partial^2 g_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right). \end{aligned} \tag{34}$$

Therefore, (24) is proved. A similar argument as in Lanne and Saikkonen can be applied here to establish the positive definiteness of the matrix $\mathcal{I}_{\theta\theta}$.

The asymptotic normality can be shown in a standard manner as in the previous papers, see [4] and [11]. The idea is to truncate the infinite moving average representations of $\mathbf{Y}_t$ and $\mathbf{X}_t$ in (11) and (12) at some large number. Then a standard central limit theorem for finite dependent stationary sequences can be applied to establish the asymptotic normality of the truncated score vector. Making use of the coefficient matrices in (11) and (12) decaying to zero at a geometric rate, a standard result can be used to deal with the approximation error made by truncation and thus prove the stated asymptotic normality. Details are omitted here. □

References

[1] B. Andrews, M. Calder, and R. A. Davis. Maximum likelihood estimation for α-stable autoregressive processes. *Annals of Statistics*, 37:1946–1982, 2007.

[2] B. Andrews, R. A. Davis, and F. J. Breidt. Maximum likelihood estimation for all-pass time series models. *Journal of Multivariate Analysis*, 97:1638–1659, 2006.

[3] F. J. Breidt and R. A. Davis. Time-reversibility, identifiability, and independence of innovations for stationary time series. *Journal of Time Series Analysis*, 13:377–390, 1992.

[4] F. J. Breidt, R. A. Davis, K. S. Lii, and M. Rosenblatt. Maximum likelihood estimation for noncausal autoregressive processes. *Journal of Multivariate Analysis*, 36:175–198, 1991.

[5] F. J. Breidt, R. A. Davis, and A. Trindade. Least absolute deviation estimation for all-pass time series models. *Annals of Statistics*, 29:919–946, 2001.

[6] K. Chan, L. Ho, and H. Tong. A note on time-reversibility of multivariate linear processes. *Biometrika*, 93:221–227, 2006.

[7] G. Duffee. Term premia and interest rate forecasts in affine models. *Journal of Finance*, 57:405–443, 2002.

[8] K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London, 1990.

[9] E. J. Hannan. *Multiple Time Series*. John Wiley and sons, New York, 1970.

[10] M. Lanne and P. Saikkonen. Modeling expectations with noncausal autoregressions. *HECER Discussion Paper*, 212, 2008.

[11] M. Lanne and P. Saikkonen. Noncausal vector autoregression. Research Discussion Papers, Bank of Finland, 2009.

[12] A. Mountford and H. Uhlig. What are the effects of fiscal policy shocks? *Journal of Applied Econometrics*, 24:960–992, 2009.

[13] G. C. Reinsel. *Elements of Multivariate Time Series Analysis*. Springer, New York, 1993.