Noncausal Vector Autoregression[†]

Markku Lanne[*] University of Helsinki

Pentti Saikkonen[**] University of Helsinki

April 2012

Abstract

In this paper, we propose a new noncausal vector autoregressive (VAR) model for non-Gaussian time series. The assumption of non-Gaussianity is needed for reasons of identifiability. Assuming that the error distribution belongs to a fairly general class of elliptical distributions, we develop an asymptotic theory of maximum likelihood estimation and statistical inference. We argue that allowing for noncausality is of particular importance in economic applications which currently use only conventional causal VAR models. Indeed, if noncausality is incorrectly ignored, the use of a causal VAR model may yield suboptimal forecasts and misleading economic interpretations. Therefore, we propose a procedure for discriminating between causality and noncausality. The methods are illustrated with an application to interest rate data.

[†] We thank Martin Ellison, Juha Kilponen, Mika Meitz, Antti Ripatti, three anonymous referees, and the Coeditor, Robert Taylor, for useful comments. Financial support from the Academy of Finland and the OP-Pohjola Group Research Foundation is gratefully acknowledged. The first version of this paper was completed in May, 2009. It was written while the second author worked at the Bank of Finland whose hospitality is gratefully acknowledged.

* Department of Political and Economic Studies, University of Helsinki, P.O.Box 17 (Arkadiankatu 7), FIN–00014 University of Helsinki, Finland, e-mail: markku.lanne@helsinki.fi

** Department of Mathematics and Statistics, University of Helsinki, P.O.Box 68 (Gustaf Hällströmin katu 2b), FIN–00014 University of Helsinki, Finland, e-mail: pentti.saikkonen@helsinki.fi

1 Introduction

The vector autoregressive (VAR) model is widely applied in various fields of application to summarize the joint dynamics of a number of time series and to obtain forecasts. Especially in economics and finance the model is also employed in structural analyses, and it often provides a suitable framework for conducting tests of theoretical interest. Typically, the error term of a VAR model is interpreted as a forecast error that should be an independent white noise process in order for the model to capture all relevant dynamic dependencies. For the forecast error property of the error term to hold it is necessary that the errors are not serially correlated. However, unless the errors are Gaussian, this is not sufficient to guarantee independence and, even in the absence of serial correlation, it may be possible to predict the error term by lagged values of the considered variables. This is a relevant point because diagnostic checks in empirical analyses often suggest non-Gaussian residuals, and the use of a conventional (causal) VAR model with Gaussian likelihood has typically been justified by properties of quasi maximum likelihood (ML) estimation. Indeed, instead of its conventional causal counterpart a noncausal VAR model, which explicitly allows for the aforementioned predictability of the error term, might provide a correct specification (for noncausal (univariate) autoregressions, see, e.g., Brockwell and Davis (1987, Chapter 3) or Rosenblatt (2000)). These two issues are actually connected, as distinguishing between causality and noncausality is not possible under Gaussianity. Hence, in order to assess the nature of causality, allowance must be made for deviations from Gaussianity when they are backed up by the data. If noncausality indeed is present, confining to (misspecified) causal VAR models may lead to suboptimal forecasts and false conclusions.

The statistical literature on noncausal univariate time series models is relatively small, and, to our knowledge, noncausal VAR models were unexplored prior to our work (for available work on noncausal autoregressions and their applications, see Rosenblatt (2000), Andrews, Davis, and Breidt (2006), Lanne and Saikkonen (2011), and the references therein).[1] In this paper, the previous statistical theory of univariate noncausal autoregressive models is extended to the vector case. Our formulation of the noncausal VAR model is a direct extension of that used by Lanne and Saikkonen (2011) in the univariate case. To obtain a feasible approximation for the non-Gaussian likelihood function, the distribution of the error term is assumed to belong to a fairly general class of elliptical distributions. With this assumption we show the consistency and asymptotic normality of an approximate (local) ML estimator, and justify the applicability of usual likelihood

[1] While revising this paper we learned about the related work of Davis and Song (2010). The formulation of the noncausal VAR model considered by these authors is different from ours but the theoretical results are based on assumptions that are virtually the same as used in this paper.

based tests.

As already indicated, the noncausal VAR model can be used to check the validity of statistical analyses based on a causal VAR model. This is important, for instance, in economic applications where VAR models are commonly applied to test for economic theories. Typically such tests assume the existence of a causal VAR representation whose errors are not predictable by lagged values of the considered time series. If this is not the case, the employed tests based on a causal VAR model are not valid and the resulting conclusions may be misleading. We provide an illustration of this with interest rate data.

The remainder of the paper is structured as follows. Section 2 introduces the noncausal VAR model and discusses issues of identifiability along with other features of the model. Section 3 derives an approximation for the likelihood function and properties of the related approximate ML estimator. Section 4 provides our empirical illustration. Section 5 concludes. An appendix contains proofs and some technical derivations. Further technicalities are provided online at Cambridge Journals Online in supplementary material to this article. Readers may refer to the supplementary material associated with this article, available at Cambridge Journals Online (journals.cambridge.org/ect).

The following notation is used throughout. The expectation operator and the covariance operator are denoted by $\mathbb{E}(\cdot)$ and $\mathbb{C}(\cdot)$ or $\mathbb{C}(\cdot,\cdot)$, respectively, whereas $x \overset{d}{=} y$ means that the random quantities x and y have the same distribution. To simplify notation, we shall write $z = (z_1, \ldots, z_m)$ for the (column) vector z where the components z_i may be either scalars or vectors (or both). By $\text{vec}(A)$ we denote a column vector obtained by stacking the columns of the matrix A one below another. If A is a square matrix then $\text{vech}(A)$ is a column vector obtained by stacking the columns of A from the principal diagonal downwards (including elements on the diagonal). The usual notation $A \otimes B$ is used for the Kronecker product of the matrices A and B. The $mn \times mn$ commutation matrix and the $n^2 \times n(n+1)/2$ duplication matrix are denoted by K_{mn} and D_n, respectively. Both of them are of full column rank. The former is defined by the relation $K_{mn}\text{vec}(A) = \text{vec}(A')$, where A is any $m \times n$ matrix, and the latter by the relation $\text{vec}(B) = D_n \text{vech}(B)$, where B is any symmetric $n \times n$ matrix.

2 Model

2.1 Definition and basic properties

Consider the n-dimensional stochastic process y_t $(t = 0, \pm 1, \pm 2, ...)$ generated by

$$\Pi(B)\,\Phi\left(B^{-1}\right) y_t = \epsilon_t, \tag{1}$$

where $\Pi(B) = I_n - \Pi_1 B - \cdots - \Pi_r B^r$ $(n \times n)$ and $\Phi(B^{-1}) = I_n - \Phi_1 B^{-1} - \cdots - \Phi_s B^{-s}$ $(n \times n)$ are matrix polynomials in the backward shift operator B, and ϵ_t $(n \times 1)$ is a sequence of independent, identically distributed (continuous) random vectors with zero mean and finite positive definite covariance matrix. Moreover, the matrix polynomials $\Pi(z)$ and $\Phi(z)$ $(z \in \mathbb{C})$ have their zeros outside the unit disc so that

$$\det \Pi(z) \neq 0, \quad |z| \leq 1, \quad \text{and} \quad \det \Phi(z) \neq 0, \quad |z| \leq 1. \tag{2}$$

If $\Phi_j \neq 0$ for some $j \in \{1, ..., s\}$, equation (1) defines a noncausal vector autoregression referred to as purely noncausal when $\Pi_1 = \cdots = \Pi_r = 0$. The corresponding conventional causal model is obtained when $\Phi_1 = \cdots = \Phi_s = 0$. Then the former condition in (2) guarantees the stationarity of the model. In the general set up of equation (1) the same is true for the process

$$u_t = \Phi\left(B^{-1}\right) y_t.$$

Specifically, there exists a $\delta_1 > 0$ such that $\Pi(z)^{-1}$ has a well defined power series representation $\Pi(z)^{-1} = \sum_{j=0}^{\infty} M_j z^j = M(z)$ for $|z| < 1 + \delta_1$. Consequently, the process u_t has the causal moving average representation

$$u_t = M(B)\,\epsilon_t = \sum_{j=0}^{\infty} M_j \epsilon_{t-j}. \tag{3}$$

Notice that $M_0 = I_n$ and that (the elements of) the coefficient matrices M_j decay to zero at a geometric rate as $j \to \infty$ (cf. Lemma 3 in Kohn (1979)). When convenient, $M_j = 0$, $j < 0$, will be assumed.

Write $\Pi(z)^{-1} = (\det \Pi(z))^{-1} \Xi(z) = M(z)$, where $\Xi(z)$ is the adjoint polynomial matrix of $\Pi(z)$ with degree at most $(n-1)r$. Then, $\det \Pi(B)\, u_t = \Xi(B)\, \epsilon_t$ and, by the definition of u_t,

$$\Phi\left(B^{-1}\right) w_t = \Xi(B)\, \epsilon_t,$$

where $w_t = (\det \Pi(B)) y_t$. By the latter condition in (2) one can find a $0 < \delta_2 < 1$ such that $\Phi(z^{-1})^{-1} \Xi(z)$ has a well defined power series representation $\Phi(z^{-1})^{-1} \Xi(z) = \sum_{j=-(n-1)r}^{\infty} N_j z^{-j} = N(z^{-1})$ for $|z| > 1 - \delta_2$. Thus, the process w_t has the representation

$$w_t = \sum_{j=-(n-1)r}^{\infty} N_j \epsilon_{t+j}, \tag{4}$$

where the coefficient matrices N_j decay to zero at a geometric rate as $j \to \infty$ and, when convenient, $N_j = 0$, $j < -(n-1)r$, will be assumed.

From (2) it follows that the process y_t itself has the representation

$$y_t = \sum_{j=-\infty}^{\infty} \Psi_j \epsilon_{t-j}, \tag{5}$$

where Ψ_j $(n \times n)$ is the coefficient matrix of z^j in the Laurent series expansion of $\Psi(z) \overset{def}{=} \Phi(z^{-1})^{-1} \Pi(z)^{-1}$ which exists for $1 - \delta_2 < |z| < 1 + \delta_1$ with Ψ_j decaying to zero at a geometric rate as $|j| \to \infty$. The representation (5) implies that y_t is a stationary and ergodic process with finite second moments. We use the abbreviation VAR(r, s) for the model defined by (1). In the causal case $s = 0$, the conventional abbreviation VAR(r) is also used.

Denote by $\mathbb{E}_t(\cdot)$ the conditional expectation operator with respect to the information set $\{y_t, y_{t-1}, ...\}$ and conclude from (1) and (5) that

$$y_t = \sum_{j=-\infty}^{s-1} \Psi_j \mathbb{E}_t(\epsilon_{t-j}) + \sum_{j=s}^{\infty} \Psi_j \epsilon_{t-j}.$$

In the conventional causal case, $s = 0$ and $\mathbb{E}_t(\epsilon_{t-j}) = 0$, $j \leq -1$, so that the right hand side reduces to the moving average representation (3). However, in the noncausal case this does not happen. Then $\Psi_j \neq 0$ for some $j < 0$, which in conjunction with the representation (5) shows that y_t and ϵ_{t-j} are correlated. Consequently, $\mathbb{E}_t(\epsilon_{t-j}) \neq 0$ for some $j < 0$, implying that future errors can be predicted by past values of the process y_t. A possible interpretation of this predictability is that the errors contain factors which are not included in the model and can be predicted by the time series selected in the model. This seems quite plausible, especially, in economic applications where time series are typically interrelated and only a few time series out of a larger selection are used in the analysis. The reason why some variables are excluded may be that data are not available or the underlying economic model only contains the variables for which hypotheses of interest are formulated.

To elaborate the preceding point further, consider a high dimensional time series vector X_t that can be modeled by a finite order causal VAR process whose errors $\xi_t^{(x)}$, say, are independent and identically distributed. Suppose that all components of X_t are not used in the analysis, and let Y_t be the vector containing the selected components. Then Y_t generally does not have a finite order causal VAR representation with errors independent and identically distributed. Instead, Y_t has an infinite order VAR representation whose errors $\xi_t^{(y)}$, say, are uncorrelated and have a linear representation in terms of $\xi_t^{(x)}$, the errors of X_t (see Lemma 2 in Johansen and Juselius (2010)).[2] In the Gaussian case, the errors $\xi_t^{(y)}$ are independent and cannot be predicted by past values of Y_t. The latter fact follows because Y_t has a linear representation in terms of present and past values

[2]This result is formulated in a cointegrated VAR context but can be specialized to the stationary case. Alternatively, Y_t can be represented as a causal vector ARMA process with errors that are uncorrelated and have a linear representation in terms of $\xi_t^{(x)}$ (see Corollaries 11.1.1 and 11.1.2 in Lütkepohl (2005)). The conclusions obtained here for the infinite order causal VAR process apply, in essence, to the causal vector ARMA process.

of $\xi_t^{(y)}$ and, consequently, $\xi_t^{(y)}$ and Y_{t-j} $(j \geq 1)$ are independent. In the non-Gaussian case the situation is different, however. Then $\xi_t^{(y)}$ and Y_{t-j} $(j \geq 1)$ can only be shown to be uncorrelated but, as both of them depend on lagged values of $\xi_t^{(x)}$, they need not be independent, and the possibility that $\xi_t^{(y)}$ can (nonlinearly) be predicted by Y_{t-j} $(j \geq 1)$ appears plausible. Thus, as errors of a noncausal VAR model can be predicted by past values of the observed process, one may expect to observe noncausality when small dimensional VAR models are applied. That this can indeed happen is illustrated by a simulation experiment of Lof (2012) in the context of a bivariate VAR model.

Note that the purpose of the preceding discussion is to demonstrate how a VAR process with an error term potentially predictable by lagged values of the process may arise and lead to observing noncausality in applications. We are not claiming that in the described situation, our noncausal VAR model would be superior to a causal alternative. Making such claims is, in fact, difficult because it is not known whether the selected series have a noncausal VAR representation with errors independent and identically distributed, as assumed in our model. What can be said, however, is that in the non-Gaussian case a causal VAR model is misspecified in the sense that its errors are dependent.

A practical complication with noncausal autoregressive models is that they cannot be identified by second order properties or Gaussian likelihood. In the univariate case this is explained in Brockwell and Davis (1987, p. 124–125). A similar explanation based on the factorization of the spectral density matrix can be obtained from Hannan (1970, p. 64–67). Specifically, one can conclude that the spectral density matrix and, hence, autocovariance function of a noncausal VAR(r, s) process cannot be distinguished from those of a causal VAR$(r + s)$ process (further details on this issue are available in the Supplementary Appendix). Thus, if y_t or, equivalently, the error term ϵ_t is Gaussian, causal and noncausal representations of (1) are statistically indistinguishable and nothing is lost by using a conventional causal representation. However, if the errors are non-Gaussian, using a causal representation of a true noncausal process means using a misspecified VAR model whose errors are only uncorrelated but not independent and can be predicted by past values of the considered series. Thus, potentially better fit and forecasts could be obtained by using the correctly specified noncausal model.

Identification of the noncausal VAR model (1) will be discussed in Section 2.3 after presenting assumptions employed for the error term ϵ_t. Here we only note that finding a correct noncausal VAR model is a larger issue than finding the correct orders r and s in (1). The reason is that equation (1) is not the only possibility to formulate a noncausal VAR model. For instance, as the matrix product does not commute, a different specification is obtained by changing the order of the operators $\Pi(B)$ and $\Phi(B^{-1})$ in (1). We have no strong arguments in favor of the employed order although we believe that it may be the

more convenient choice from the viewpoint of economic applications. The reason is that the chosen order naturally gives rise to a representation of y_t as a function of its future expected values, as is common in economic models involving expectations.[3]

The fact that a different specification results when one changes the order of the operators $\Pi(B)$ and $\Phi(B^{-1})$ in (1) also means that our noncausal VAR model does not include all possible forms of noncausality. A potentially viable alternative might be based on the equation

$$y_t = \Pi_1^* y_{t-1} + \cdots + \Pi_p^* y_{t-p} + \epsilon_t^*. \tag{6}$$

Here ϵ_t^* $(n \times 1)$ is as in (1), that is, a sequence of independent, identically distributed random vectors with zero mean and finite positive definite covariance matrix. Furthermore, the autoregressive polynomial satisfies $\det(I_n - \Pi_1^* B - \cdots - \Pi_p^* B^p) \neq 0$, $|z| = 1$, so that zeros both outside and inside the unit circle are allowed. In the univariate case this yields the formulation used by several previous authors (see, e.g., Breidt et al. (1991) and Rosenblatt (2000)). In the vector case it has recently been considered by Davis and Song (2010).

Unlike (1), the specification (6) is not based on a multiplicative structure and is, in that sense, more general of the two. However, from the viewpoint of interpretation, we find the specification (1) more straightforward as it naturally allows for separating the 'causal' and 'noncausal' parts of the process. Moreover, as the following example demonstrates, there are cases where the specification (1) cannot be imbedded in (6), at least in a straightforward and practically convenient manner.

Consider a bivariate special case of (1) with $r = s = 1$ and suppose that the components of y_t satisfy $y_{1t} = \pi_{11} y_{t-1} + \epsilon_{1t}$ and $y_{2t} = \phi_{22} y_{2,t+1} + \epsilon_{2t}$ where $0 < |\pi_{11}| < 1$ and $0 < |\phi_{22}| < 1$. Thus, as a univariate process y_{1t} is causal and y_{2t} is purely noncausal but assuming that ϵ_{1t} and ϵ_{2t}, the components of ϵ_t, are dependent there are gains in using a bivariate model. Denoting $\pi_{22}^* = 1/\phi_{22}$ we have $y_{2t} = \pi_{22}^* y_{2,t-1} - \pi_{22}^* \epsilon_{2,t-1}$ so that the bivariate process (y_{1t}, y_{2t}) can be written in the form of equation (6) except that the resulting error term $(\epsilon_{1t}, -\pi_{22}^* \epsilon_{2,t-1})$ is not independent in time. It is possible to obtain the specification (6) with an independent error term $\epsilon_t^* = (\epsilon_{1t}, -\pi_{22}^* \epsilon_{2t})$ if one considers the process $(y_{1t}, y_{2,t+1})$. However, from a practical point of view this possibility appears difficult because the structure of the observed process is unknown. In more complicated models this difficulty apparently becomes even more pronounced.[4]

[3] For instance, in the case $s = 1$ equation (1) and the definition of the process u_t imply that $y_t = \Phi_1 y_{t+1} + u_t$ and taking conditional expectations on both sides readily shows how y_t depends on the expected value of y_{t+1}. If the the order of the operators $\Pi(B)$ and $\Phi\left(B^{-1}\right)$ in (1) is changed the situation gets more complicated. For instance, when $r = s = 1$ we have $(I_n + \Phi_1 \Pi_1) y_t = \Phi_1 y_{t+1} + \Pi_1 y_{t-1} + \epsilon_t$ where it is possible that the matrix $I_n + \Phi_1 \Pi_1$ is singular.

[4] The argument used in this example can clearly be reversed by starting from the bivariate special case

In spite of its simplicity the preceding example demonstrates that in the multivariate case noncausal autoregressions appear considerably more complex than in the univariate case where there is no essential difference between the specifications (1) and (6). If a univariate time series can be described by (1) with $\Phi_s \neq 0$ it can be described by (6) and vice verse when $\Pi_p^* \neq 0$ (to get an illustration, consider y_{2t} alone in the preceding example). As seen above, this does not necessarily happen in the multivariate case. Whether a feasible specification covering all or 'most' noncausal VAR processes exists is an interesting question not attempted to solve in this paper.

2.2 Assumptions

In this section, we introduce assumptions that enable us to derive the likelihood function and its derivatives as well as to discuss the identifiability of the model. Further assumptions, needed for the asymptotic analysis of the ML estimator and related tests, will be introduced in subsequent sections.

As already discussed, meaningful application of the noncausal VAR model requires that the distribution of ϵ_t is non-Gaussian. In the following assumption the distribution of ϵ_t is restricted to a general elliptical form. As is well known, the normal distribution belongs to the class of elliptical distributions but we will not rule it out at this point. Other examples of elliptical distributions are discussed in Fang, Kotz, and Ng (1990, Chapter 3). Perhaps the best known non-Gaussian example is the multivariate t-distribution.

Assumption 1. The error process ϵ_t in (1) is independent and identically distributed with zero mean, finite and positive definite covariance matrix, and an elliptical distribution possessing a density.

Results on elliptical distributions we shall need can be found in Fang et al. (1990, Chapter 2) on which the following discussion is based. Let Σ $(n \times n)$ be a symmetric and positive definite parameter matrix. By Assumption 1, we have the representation

$$\epsilon_t \overset{d}{=} \rho_t \Sigma^{1/2} \upsilon_t, \tag{7}$$

where (ρ_t, υ_t) is an independent and identically distributed sequence such that ρ_t (scalar) and υ_t $(n \times 1)$ are independent, ρ_t is nonnegative, and υ_t is uniformly distributed on the unit ball (so that $\upsilon_t'\upsilon_t = 1$).

of (6) given by $y_{it} = \pi_{ii}^* y_{i,t-1} + \epsilon_{it}^*$ $(i = 1, 2)$ with $0 < |\pi_{11}^*| < 1$ and $|\pi_{22}^*| > 1$. This specification can be transformed to the form (1) with an independent error term only if one considers the process $(y_{1t}, y_{2,t-1})$ instead of (y_{1t}, y_{2t}).

The density function of ϵ_t is of the form

$$f_\Sigma(x;\lambda) = \frac{1}{\sqrt{\det(\Sigma)}} f\left(x'\Sigma^{-1}x;\lambda\right) \tag{8}$$

for some nonnegative function $f(\cdot;\lambda)$ of a scalar variable (examples of particular cases with $\Sigma = I_n$ can be found in Fang et al. (1990, p. 69) and for the multivariate t-distribution with general Σ in Section 4 (footnote 4)). In addition to the parameter matrix Σ the distribution of ϵ_t is allowed to depend on the parameter vector λ ($\mathrm{d} \times 1$). The parameter matrix Σ is closely related to the covariance matrix of ϵ_t. Specifically, because $\mathbb{E}(\upsilon_t) = 0$ and $\mathbb{C}(\upsilon_t) = n^{-1}I_n$ (see Fang et al. (1990, Theorem 2.7)) one obtains from (7) that

$$\mathbb{C}(\epsilon_t) = \frac{\mathbb{E}(\rho_t^2)}{n}\Sigma. \tag{9}$$

Note that finiteness of the covariance matrix $\mathbb{C}(\epsilon_t)$ is equivalent to $\mathbb{E}(\rho_t^2) < \infty$. For later purposes we also note that the density of ρ_t^2, denoted by $\varphi_{\rho^2}(\cdot;\lambda)$, is related to the function $f(\cdot;\lambda)$ in (8) via

$$\varphi_{\rho^2}(\zeta;\lambda) = \frac{\pi^{n/2}}{\Gamma(n/2)}\zeta^{n/2-1} f(\zeta;\lambda), \quad \zeta \geq 0, \tag{10}$$

where $\Gamma(\cdot)$ is the gamma function (see Fang et al. (1990, p. 36)).

A convenient feature of elliptical distributions is that we can often work with the scalar random variable ρ_t instead of the random vector ϵ_t. This facilitates the needed mathematical derivations. Elliptical distributions form a fairly large class of multivariate distributions but being symmetric they cannot allow for skewness. Using more general distributional assumptions could be possible, but that might add to the technical complications which are considerable even in the elliptical case.

Assumptions to be imposed on the density of ϵ_t can be expressed by using the function $f(\zeta;\lambda)$ ($\zeta \geq 0$). These assumptions are similar to those previously used by Andrews et al. (2006) and Lanne and Saikkonen (2011) in so-called all-pass models and univariate noncausal autoregressive models, respectively. We denote by Λ the permissible parameter space of λ and use $f'(\zeta;\lambda)$ to signify the partial derivative $\partial f(\zeta,\lambda)/\partial\zeta$ with a similar definition for $f''(\zeta;\lambda)$. Also, we include a subscript (typically λ) in the expectation operator or covariance operator when it seems reasonable to emphasize the parameter value assumed in the calculations. Our second assumption is as follows.

Assumption 2. (i) The parameter space Λ is an open subset of $\mathbb{R}^{\mathrm{d}}$ and that of the parameter matrix Σ is the set of symmetric positive definite $n \times n$ matrices.
(ii) The function $f(\zeta;\lambda)$ is positive and twice continuously differentiable on $(0,\infty) \times \Lambda$. Furthermore, for all $\lambda \in \Lambda$, a finite and positive right limit $\lim_{\zeta\to 0+} f(\zeta;\lambda)$ exists.

(iii) For all $\lambda \in \Lambda$,

$$\int_0^\infty \zeta^{n/2+1} f(\zeta;\lambda)\, d\zeta < \infty \quad \text{and} \quad \int_0^\infty \zeta^{n/2} (1+\zeta) \frac{(f'(\zeta;\lambda))^2}{f(\zeta;\lambda)} d\zeta < \infty.$$

Assuming that the parameter space Λ is open is not restrictive and facilitates exposition. The former part of Assumption 2(ii) is needed to ensure the usual differentiability of the likelihood function. It is similar to condition (A1) in Andrews et al. (2006) and Lanne and Saikkonen (2011) although in these papers the domain of the first argument of the function f is the whole real line. The latter part of Assumption 2(ii) is a mild technical condition that is needed in some proofs. The first condition in Assumption 2(iii) implies that $\mathbb{E}_\lambda(\rho_t^4)$ is finite (see (10)) and altogether this assumption guarantees finiteness of some expectations needed in subsequent developments. In particular, the latter condition implies finiteness of the quantities

$$\boldsymbol{j}(\lambda) = \frac{4\pi^{n/2}}{n\Gamma(n/2)} \int_0^\infty \zeta^{n/2} \frac{(f'(\zeta;\lambda))^2}{f(\zeta;\lambda)} d\zeta = \frac{4}{n} \mathbb{E}_\lambda \left[\rho_t^2 \left(\frac{f'(\rho_t^2;\lambda)}{f(\rho_t^2;\lambda)} \right)^2 \right] \tag{11}$$

and

$$\boldsymbol{i}(\lambda) = \frac{\pi^{n/2}}{\Gamma(n/2)} \int_0^\infty \zeta^{n/2+1} \frac{(f'(\zeta;\lambda))^2}{f(\zeta;\lambda)} d\zeta = \mathbb{E}_\lambda \left[\rho_t^4 \left(\frac{f'(\rho_t^2;\lambda)}{f(\rho_t^2;\lambda)} \right)^2 \right], \tag{12}$$

where the latter equalities are obtained by using the density of ρ_t^2 (see (10)). The quantities $\boldsymbol{j}(\lambda)$ and $\boldsymbol{i}(\lambda)$ can be used to characterize non-Gaussianity of the error term ϵ_t. Specifically we can prove the following (a proof of this lemma as well as other proofs can be found in Appendix B or in the Suplementary Appendix).

Lemma 1. *Suppose that Assumptions 1 and 2 hold. Then,* $\boldsymbol{j}(\lambda) \geq n/\mathbb{E}_\lambda(\rho_t^2)$ *and* $\boldsymbol{i}(\lambda) \geq (n+2)^2 [\mathbb{E}_\lambda(\rho_t^2)]^2 / 4\mathbb{E}_\lambda(\rho_t^4)$ *where equalities hold if and only if* ϵ_t *is Gaussian. If* ϵ_t *is Gaussian,* $\boldsymbol{j}(\lambda) = 1$ *and* $\boldsymbol{i}(\lambda) = n(n+2)/4$.

Lemma 1 shows that assuming $\boldsymbol{j}(\lambda) > n/\mathbb{E}_\lambda(\rho_t^2)$ gives a counterpart of condition (A5) in Andrews et al. (2006) and Lanne and Saikkonen (2011). A difference is, however, that in these papers the lower part of the inequality does not involve a counterpart of the expectation $\mathbb{E}_\lambda(\rho_t^2)$. In subsequent developments we also consider a scaled version of $\boldsymbol{j}(\lambda)$ given by

$$\boldsymbol{\tau}(\lambda) = \boldsymbol{j}(\lambda) \mathbb{E}_\lambda(\rho_t^2) / n. \tag{13}$$

Clearly, $\boldsymbol{\tau}(\lambda) \geq 1$ with equality if and only if ϵ_t is Gaussian.

It appears useful to generalize the model defined in equation (1) by allowing restrictions on the coefficient matrices Π_j $(j = 1, ..., r)$ and Φ_j $(j = 1, ..., s)$. Thus, we make the

following assumption which even allows for general nonlinear restrictions although in practice linear restrictions are presumably the most common ones (see, e.g., Assumption A4 of Kohn (1979) for a similar assumption in (causal) ARMAX models).

Assumption 3. The parameter matrices $\Pi_j = \Pi_j(\vartheta_1)$ $(j = 1, ..., r)$ and $\Phi_j(\vartheta_2)$ $(j = 1, ..., s)$ are twice continuously differentiable functions of the parameter vectors $\vartheta_1 \in \Theta_1 \subseteq \mathbb{R}^{m_1}$ and $\vartheta_2 \in \Theta_2 \subseteq \mathbb{R}^{m_2}$, where the permissible parameter spaces Θ_1 and Θ_2 are open and such that condition (2) holds for all $\vartheta = (\vartheta_1, \vartheta_2) \in \Theta_1 \times \Theta_2$.

Together with Assumption 2(ii) this assumption guarantees the standard requirement that the likelihood function is twice continuously differentiable. The most common example of the restrictions imposed on Π_j and Φ_j restricts some of their elements to zero in which case the parameter vectors ϑ_1 and ϑ_2 contain the unrestricted elements of Π_j and Φ_j, respectively. We will continue to use the notation Π_j and Φ_j when there is no need to make the dependence on the underlying parameter vectors explicit.

2.3 Identifiability

In this section we demonstrate that a correct noncausal model can be distinguished from its causal counterpart or an incorrect noncausal alternative. To this end, we consider the uniqueness of the linear representation (5). This issue has been studied in the univariate case by Rosenblatt (2000, Chapter 1.3) and in the vector case by Chan and Ho (2004) (see also Davis and Song (2010) and Chan, Ho, and Tong (2006) where results of the latter paper are discussed). Chan and Ho (2004) provide conditions under which the process ϵ_t and the coefficient matrices Ψ_j in the linear representation (5) are 'essentially' unique. The linear processes they consider are more general than will be assumed below but, for ease of exposition, we prefer to be more specific here.

Now, suppose that y_t has two linear representations given by

$$y_t = \sum_{j=-\infty}^{\infty} \Psi_j \epsilon_{t-j} = \sum_{j=-\infty}^{\infty} \Psi_j^* \epsilon_{t-j}^*, \tag{14}$$

where former is defined by (5) and the latter is defined analogously. Specifically, ϵ_t^* $(n \times 1)$ is a sequence of independent, identically distributed random vectors with zero mean and finite positive definite covariance matrix, and the coefficient matrices Ψ_j^* decay to zero at a geometric rate as $|j| \to \infty$. As an application of Theorem 7 Chan and Ho (2004) we can obtain the following proposition (see also Theorem 1 of Chan et al. (2006) for a related result). In this proposition we assume that the (excess) kurtosis of the elliptically distributed error term ϵ_t is nonzero. The kurtosis measure we use is the one discussed in Muirhead and Waternaux (1980, p. 33). It only depends on the function f in (8)

and equals the common kurtosis of any component of ϵ_t. Thus, it can be defined by the conventional kurtosis of any component of ϵ_t.

Proposition 1. *Let Assumptions 1 and 2 hold and assume that the (excess) kurtosis of ϵ_t is nonzero and that y_t has the two representations in (14) with the fourth moments of ϵ_t^* finite. Then, there exist an integer l and a nonsingular matrix Q such that $\epsilon_t^* = Q^{-1}\epsilon_{t+l}$ and $\Psi_j^* = \Psi_{j-l}Q$.*

The result of this proposition holds even if the kurtosis of ϵ_t is zero provided a certain condition on higher order cumulants of ϵ_t holds (this condition also implies non-Gaussianity of ϵ_t). Similar results can also be obtained when the distribution of ϵ_t is not elliptical (see Theorem 7 of Chan and Ho (2004) or, for the required assumptions, Conditions 5 and 6 in Chan et al. (2006)). As Theorems 3 and 4 of Chan and Ho (2004) show, different results are obtained if the components of ϵ_t are independent, which cannot happen for non-Gaussian elliptical distributions (see Theorem 4.1.1 of Fang et al. (1990)).

Next we use Proposition 1 to demonstrate that causal and noncausal VAR models can be distinguished. Suppose we have data generated by a noncausal VAR(r, s) process (1) with $s \geq 1$ and the error term ϵ_t satisfying the assumptions stated in Proposition 1. Assume, for simplicity, that $\Phi_s \neq 0$ (the following discussion can readily be modified to the case where $\Phi_s = 0$). Now consider the incorrectly specified causal VAR$(r + s)$ model

$$\mathcal{C}(B)y_t = \xi_t, \quad \mathcal{C}(B) = \sum_{j=0}^{r+s} \mathcal{C}_j B^j, \quad \mathcal{C}_0 = I_n, \tag{15}$$

where $\det \mathcal{C}(z) \neq 0$, $|z| \leq 1$ and the (stationary) error term ξ_t is uncorrelated with zero mean and finite, positive definite covariance matrix (see the discussion in Section 2.1). From this and the linear representation (1) it follows that

$$\xi_t = \mathcal{C}(B)\Phi\left(B^{-1}\right)^{-1}\Pi\left(B\right)^{-1}\epsilon_t. \tag{16}$$

We shall now demonstrate that ξ_t cannot be an independent sequence if the conditions of Proposition 1 hold. Suppose that ξ_t is independent but has the linear representation (16). Then, Proposition 1 (applied to ξ_t instead of y_t) shows that, for some integer l and a nonsingular matrix Q, $\xi_t = Q^{-1}\epsilon_{t+l}$ and $\mathcal{C}(z)\Phi\left(z^{-1}\right)^{-1}\Pi\left(z\right)^{-1}Q$ is of the form $I_n z^{-l}$ (cf. Theorems 8 and 12 of Chan and Ho (2004)). Hence, we have $\mathcal{C}(z) = Q^{-1}\Pi\left(z\right)\Phi\left(z^{-1}\right)z^{-l}$ and, furthermore, $\Pi\left(z\right)^{-1}Q\mathcal{C}(z) = \Phi\left(z^{-1}\right)z^{-l}$. As $\Phi_s \neq 0$ (and $s \geq 1$) is assumed we must have $l = -s$ so that $\Pi\left(z\right)^{-1}Q\mathcal{C}(z) = -(\Phi_s + \Phi_{s-1}z + \cdots + \Phi_1 z^{s-1} - I_n z^s)$. However, as the zeros of $\det\left(\Phi\left(z\right)\right)$ lie outside the unit circle (see (2)) the zeros of $\det\left(\Phi\left(z^{-1}\right)z^s\right)$ lie inside the unit circle and, because $\det\left(\Pi\left(z\right)^{-1}Q^{-1}\mathcal{C}(z)\right) \neq 0$, $|z| \leq 1$, we get a contradiction. Thus, the uncorrelated error term ξ_t of the causal representation (15) is dependent which

makes the causal representation distinguishable from the true noncausal representation (1). The same conclusion is obtained if in place of the causal VAR$(r+s)$ process one considers a misspecified noncausal VAR(r', s') process with $r'+s' = r+s$ and either $r' < r$ or $s' < s$.

The preceding discussion demonstrates that, under the conditions of Proposition 1 (or their extensions), different (causal or noncausal) representations of a correct noncausal VAR(r, s) process can be distinguished because errors of incorrect representations are only uncorrelated but not independent. In particular, even though errors of misspecified representations have a linear structure similar to (16) they can exhibit nonlinear dependence. To demonstrate this, we consider dependence typically related to conditional heteroskedasticity, which is presumably the most common type on nonlinear dependence considered in the context of (causal) VAR models. Again, we concentrate on the case where a causal model is incorrectly specified.

To see whether the error term ξ_t in (15) shows signs of conditional heteroskedasticity, we consider the autocovariances of the squares of its components ξ_{at} $(a = 1, ..., n)$. Denote $\mathbb{C}\left(\xi_{at}, \xi_{b,t+j}\right) = \gamma_{ab}(j)$ and let $\kappa_{abcd}(j,k,l)$ signify the fourth order cumulant of $\left(\xi_{at}, \xi_{b,t+j}, \xi_{c,t+k}, \xi_{d,t+l}\right)$. As is well known, $\mathbb{E}(\xi_{at}^2 \xi_{b,t+j}^2) = \gamma_{aa}(0)\gamma_{bb}(0) + 2\gamma_{ab}(j)^2 + \kappa_{aabb}(0,j,j)$ (cf. Hannan (1970, p. 209)) so that, because the sequence ξ_t is serially uncorrelated,

$$\mathbb{C}(\xi_{at}^2, \xi_{b,t+j}^2) = \kappa_{aabb}(0,j,j), \quad \text{for } j \neq 0.$$

In the non-Gaussian case fourth order cumulants are generally nonzero so that squared residuals of a (misspecified) causal VAR$(r+s)$ model can be expected to exhibit serial correlation. In particular cases the nature of this serial correlation can be studied by expressing the fourth cumulants $\kappa_{aabb}(0,j,j)$ in terms of the fourth cumulants of ϵ_t and the parameters in the series expansion of $\mathcal{C}(z)\Phi\left(z^{-1}\right)^{-1}\Pi(z)$ (see Hannan (1970, p. 211)). The result is rather complicated and therefore not considered here.

The preceding discussion indicates that in non-Gaussian cases residuals of a fitted causal VAR model can appear conditionally heteroskedastic even if the data are generated by a noncausal VAR process with homoskedastic errors. The same can happen when one looks at the residuals of a misspecified noncausal VAR model. Thus, noncausal VAR models can allow for features similar to those typically modelled with GARCH models and, in particular, causal VAR models with GARCH errors. A closer examination of this issue and comparisons of noncausal VAR models and causal VAR models with GARCH errors would be of interest but is beyond the scope of this paper. An empirical illustration of the capability of a univariate noncausal autoregressive model to allow for features typically modelled by a GARCH model is provided by Breidt, Davis, and Trindade (2001). In their illustration the probable alternative to the chosen noncausal model, namely a causal

AR(1) model with GARCH errors, would require at least two more parameters. In our vector case corresponding differences in the number of parameters can easily become considerably larger.

3 Parameter estimation

3.1 Likelihood function

ML estimation of the parameters of a univariate noncausal autoregressive model was studied by Breidt et al. (1991) by using a parametrization different from that in (1). The parametrization (1) was employed by Lanne and Saikkonen (2011) whose results are extended here. Unless otherwise stated, Assumptions 1-3 are supposed to hold. The derivations also assume that $s > 0$ but can readily be specialized to the causal case $s = 0$.

Suppose we have an observed time series $y_1, ..., y_T$ $(T > s + nr)$. As in the univariate case we derive the likelihood function by transforming the vector of observed time series. Denote

$$\det \Pi(z) = a(z) = 1 - a_1 z - \cdots - a_{nr} z^{nr}.$$

Then, $w_t = a(B)\, y_t$ which in conjunction with the definition $u_t = \Phi(B^{-1})\, y_t$ shows that

$$\begin{bmatrix} u_1 \\ \vdots \\ u_{T-s} \\ w_{T-s+1} \\ \vdots \\ w_T \end{bmatrix} = \begin{bmatrix} y_1 - \Phi_1 y_2 - \cdots - \Phi_s y_{s+1} \\ \vdots \\ y_{T-s} - \Phi_1 y_{T-s+1} - \cdots - \Phi_s y_T \\ y_{T-s+1} - a_1 y_{T-s} - \cdots - a_{nr} y_{T-s-nr+1} \\ \vdots \\ y_T - a_1 y_{T-1} - \cdots - a_{nr} y_{T-nr} \end{bmatrix} = \boldsymbol{H}_1 \begin{bmatrix} y_1 \\ \vdots \\ y_{T-s} \\ y_{T-s+1} \\ \vdots \\ y_T \end{bmatrix}$$

or briefly

$$\boldsymbol{x}_1 = \boldsymbol{H}_1 \boldsymbol{y}.$$

The definition of u_t and (1) yield $\Pi(B)\, u_t = \epsilon_t$ so that, by the preceding equality,

$$\begin{bmatrix} u_1 \\ \vdots \\ u_r \\ \epsilon_{r+1} \\ \vdots \\ \epsilon_{T-s} \\ w_{T-s+1} \\ \vdots \\ w_T \end{bmatrix} = \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ u_{r+1} - \Pi_1 u_r - \cdots - \Pi_r u_1 \\ \vdots \\ u_{T-s} - \Pi_1 u_{T-s-1} - \cdots - \Pi_r u_{T-s-r} \\ w_{T-s+1} \\ \vdots \\ w_T \end{bmatrix} = \boldsymbol{H}_2 \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ u_{r+1} \\ \vdots \\ u_{T-s} \\ w_{T-s+1} \\ \vdots \\ w_T \end{bmatrix}$$

or

$$\boldsymbol{x}_2 = \boldsymbol{H}_2 \boldsymbol{x}_1.$$

We also perform a third transformation which transforms the variables $w_{T-s+1}, ..., w_T$ in $\boldsymbol{x}_2$. To this end, define

$$v_{k,T-s+k} = w_{T-s+k} - \sum_{j=-(n-1)r}^{-k} N_j \epsilon_{T-s+k+j}, \quad k = 1, ..., s,$$

where the sum is interpreted as zero when $k > (n-1)\, r$, that is, when the lower bound exceeds the upper bound. Note also that, by (1) and (4), $v_{k,T-s+k}$ can be expressed as a function of the observed data $y_1, ..., y_T$ and that the representation $v_{k,T-s+k} = \sum_{j=-k+1}^{\infty} N_j \epsilon_{T-s+k+j}$ holds, showing that $v_{k,T-s+k}$ $(k = 1, ..., s)$ are independent of ϵ_t, $t \leq T - s$. Now we can introduce the transformation

$$\begin{bmatrix} u_1 \\ \vdots \\ u_r \\ \epsilon_{r+1} \\ \vdots \\ \epsilon_{T-s} \\ v_{1,T-s+1} \\ \vdots \\ v_{s,T} \end{bmatrix} = \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ \epsilon_{r+1} \\ \vdots \\ \epsilon_{T-s} \\ w_{T-s+1} - N_{-(n-1)r}\epsilon_{T-s+1-(n-1)r} - \cdots - N_{-1}\epsilon_{T-s} \\ \vdots \\ w_T - N_{-(n-1)r}\epsilon_{T-(n-1)r} - \cdots - N_{-s}\epsilon_{T-s} \end{bmatrix} = \boldsymbol{H}_3 \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ \epsilon_{r+1} \\ \vdots \\ \epsilon_{T-s} \\ w_{T-s+1} \\ \vdots \\ w_T \end{bmatrix}$$

or

$$\boldsymbol{z} = \boldsymbol{H}_3 \boldsymbol{x}_2.$$

Combining the preceding three transformations yields the equation

$$\boldsymbol{z} = \boldsymbol{H}_3 \boldsymbol{H}_2 \boldsymbol{H}_1 \boldsymbol{y},$$

where the (nonstochastic) matrices $\boldsymbol{H}_1$, $\boldsymbol{H}_2$, and $\boldsymbol{H}_3$ are nonsingular. The nonsingularity of $\boldsymbol{H}_2$ and $\boldsymbol{H}_3$ follows from the fact that $\det(\boldsymbol{H}_2) = \det(\boldsymbol{H}_3) = 1$, as can be easily checked. Justifying the nonsingularity of $\boldsymbol{H}_1$ is somewhat more complicated. Details are available in the Supplementary Appendix.

From (3) and (4) it is seen that the component vectors of $\boldsymbol{z}$ given by $\boldsymbol{z}_1 = (u_1, ..., u_r)$, $\boldsymbol{z}_2 = (\epsilon_{r+1}, ..., \epsilon_{T-s})$, and $\boldsymbol{z}_3 = (v_{1,T-s+1}, ..., v_{s,T})$ are independent. Thus, (under true parameter values) the joint density function of $\boldsymbol{z}$ can be expressed as

$$h_{\boldsymbol{z}_1}(\boldsymbol{z}_1) \left(\prod_{t=r+1}^{T-s} f_\Sigma(\epsilon_t; \lambda) \right) h_{\boldsymbol{z}_3}(\boldsymbol{z}_3),$$

where $h_{\boldsymbol{z}_1}(\cdot)$ and $h_{\boldsymbol{z}_3}(\cdot)$ signify the joint density functions of $\boldsymbol{z}_1$ and $\boldsymbol{z}_3$, respectively. Using (1) and the fact that the determinants of $\boldsymbol{H}_2$ and $\boldsymbol{H}_3$ are unity we can write the joint density function of the data vector $\boldsymbol{y}$ as

$$h_{\boldsymbol{z}_1}(\boldsymbol{z}_1(\vartheta))\left(\prod_{t=r+1}^{T-s} f_\Sigma\left(\Pi(B)\,\Phi\left(B^{-1}\right)y_t;\lambda\right)\right)h_{\boldsymbol{z}_3}(\boldsymbol{z}_3(\vartheta))\left|\det(\boldsymbol{H}_1)\right|.$$

Here the argument $\boldsymbol{z}_1(\vartheta)$ is defined by replacing u_t in the definition of $\boldsymbol{z}_1$ by $\Phi(B^{-1})\,y_t$ $(t=1,...,r)$ and $\boldsymbol{z}_3(\vartheta)$ is defined similarly by replacing $v_{k,T-s+k}$ in the definition of $\boldsymbol{z}_3$ by an analog with $a(B)\,y_{T-s+k}$ and $\Pi(B)\,\Phi(B^{-1})\,y_{T-s+k+j}$ used in place of w_{T-s+k} and $\epsilon_{T-s+k+j}$, respectively $(j=-(n-1)\,r,....,-k,\ k=1,...,s)$.

It is easy to check that the determinant of the $(T-s)\,n\times(T-s)\,n$ block in the upper left hand corner of $\boldsymbol{H}_1$ is unity and, using the well-known formula for the determinant of a partitioned matrix, it can further be seen that the determinant of $\boldsymbol{H}_1$ is independent of the sample size T. This suggests approximating the joint density of $\boldsymbol{y}$ by the second factor in the preceding expression, giving rise to the approximate log-likelihood function

$$l_T(\theta)=\sum_{t=r+1}^{T-s} g_t(\theta), \tag{17}$$

where the parameter vector θ contains the unknown parameters and (cf. (8))

$$g_t(\theta)=\log f\left(\epsilon_t(\vartheta)'\,\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)-\frac{1}{2}\log\det(\Sigma), \tag{18}$$

with

$$\epsilon_t(\vartheta)=u_t(\vartheta_2)-\sum_{j=1}^{r}\Pi_j(\vartheta_1)\,u_{t-j}(\vartheta_2) \tag{19}$$

and $u_t(\vartheta_2)=y_t-\Phi_1(\vartheta_2)\,y_{t+1}-\cdots-\Phi_s(\vartheta_2)\,y_{t+s}$. In addition to ϑ and λ the parameter vector θ also contains the different elements of the matrix Σ, that is, the vector $\sigma=\mathrm{vech}(\Sigma)$. For simplicity, we shall usually drop the word 'approximate' and speak about likelihood function. The same convention is used for related quantities such as the ML estimator of the parameter θ or its score and Hessian.

Maximizing $l_T(\theta)$ over permissible values of θ (see Assumptions 2(i) and 3) gives an approximate ML estimator of θ. Note that here, as well as in the next section, the orders r and s are assumed known. In our empirical example (see Section 4) we present one way to specify these quantities.

3.2 Score vector

At this point we introduce the notation θ_0 for the true value of the parameter θ and similarly for its components. Note that our assumptions imply that θ_0 is an interior point

of the parameter space of θ. To simplify notation we write $\epsilon_t(\vartheta_0) = \epsilon_t$ and $u_t(\vartheta_{20}) = u_{0t}$ when convenient. The subscript '0' will similarly be included in the coefficient matrices of the infinite moving average representations (3), (4), and (5) to emphasize that they are related to the data generation process (i.e. M_{j0}, N_{j0}, and Ψ_{j0}). We also denote $\pi_j(\vartheta_1) = \text{vec}(\Pi_j(\vartheta_1))$ $(j = 1, ..., r)$ and $\phi_j(\vartheta_2) = \text{vec}(\Phi_j(\vartheta_2))$ $(j = 1, ..., s)$, and set

$$\nabla_1(\vartheta_1) = \left[\frac{\partial}{\partial\vartheta_1}\pi_1'(\vartheta_1) : \cdots : \frac{\partial}{\partial\vartheta_1}\pi_r'(\vartheta_1)\right]' \quad \left(n^2 r \times m_1\right)$$

and

$$\nabla_2(\vartheta_2) = \left[\frac{\partial}{\partial\vartheta_2}\phi_1'(\vartheta_2) : \cdots : \frac{\partial}{\partial\vartheta_2}\phi_s'(\vartheta_2)\right]' \quad \left(n^2 s \times m_2\right).$$

In this section, we consider $\partial l_T(\theta_0)/\partial\theta$, the score of θ evaluated at the true parameter value θ_0. Explicit expressions of the components of the score vector are given in Appendix A. Here we first present the expression of the limit $\lim_{T\to\infty} T^{-1}\mathbb{C}(\partial l_T(\theta_0)/\partial\theta)$ and then at the end of the section the asymptotic distribution of the score is presented. To this end, additional assumptions and notation are needed. Some of the assumptions introduced here, and also in the next section, are rather technical, imposing conditions on the distribution of the error term. It may be worth noting that these conditions are not special in that they have been used in one form or another for years in likelihood based statistical inference. They typically hold for distributions usually used in applications. For instance, the multivariate t-distribution, which will be used in our empirical application, satisfies all the assumptions we impose. For the treatment of the score of λ we first make the following assumption.

Assumption 4. (i) There exists a function $f_1(\zeta)$ such that $\int_0^\infty \zeta^{n/2-1} f_1(\zeta)\, d\zeta < \infty$ and, in some neighborhood of λ_0, $|\partial f(\zeta;\lambda)/\partial\lambda_i| \leq f_1(\zeta)$ for all $\zeta \geq 0$ and $i = 1, ..., d$.

(ii) $$\left|\int_0^\infty \frac{\zeta^{n/2-1}}{f(\zeta;\lambda_0)}\frac{\partial}{\partial\lambda_i} f(\zeta;\lambda_0)\frac{\partial}{\partial\lambda_j}\partial f(\zeta;\lambda_0)\, d\zeta\right| < \infty, \quad i, j = 1, ..., d.$$

The first condition is a standard dominance condition which ensures that the score of λ (evaluated at θ_0) has zero mean whereas the second condition ensures that the covariance matrix of the score of λ (evaluated at θ_0) is finite. For other scores the corresponding properties are obtained from the assumptions made in the previous section.

Recall the definition $\boldsymbol{\tau}(\lambda) = \boldsymbol{j}(\lambda)\mathbb{E}_\lambda(\rho_t^2)/n$ where $\boldsymbol{j}(\lambda)$ is defined in (11). In what follows, we denote $\boldsymbol{j}_0 = \boldsymbol{j}(\lambda_0)$ and $\boldsymbol{\tau}_0 = \boldsymbol{j}_0\mathbb{E}_{\lambda_0}(\rho_t^2)/n$. Define the $n \times n$ matrix

$$C_{11}(a, b) = \boldsymbol{\tau}_0 \sum_{k=0}^{\infty} M_{k-a,0}\Sigma_0 M_{k-b,0}'$$

and set $C_{11}(\theta_0) = \left[C_{11}(a,b) \otimes \Sigma_0^{-1}\right]_{a,b=1}^{r}$ $(n^2 r \times n^2 r)$ and, furthermore,

$$\mathcal{I}_{\vartheta_1\vartheta_1}(\theta_0) = \nabla_1(\vartheta_{10})' C_{11}(\theta_0) \nabla_1(\vartheta_{10}).$$

It is straightforward to check that $\mathcal{I}_{\vartheta_1\vartheta_1}(\theta_0)$ is the standardized covariance matrix of the score of ϑ_1 or the (Fisher) information matrix of ϑ_1 evaluated at θ_0 (details can be found in the proof of Proposition 2 in the Supplementary Appendix). In what follows, the term information matrix will be used to refer to the covariance matrix of the asymptotic distribution of the score vector $\partial l_T(\theta_0)/\partial\theta$.

Presenting the information matrix of ϑ_2 is somewhat complicated. Denoting $\boldsymbol{i}_0 = \boldsymbol{i}(\lambda_0)$ we first define

$$J_0 = \boldsymbol{i}_0 \mathbb{E}\left[(\text{vech}(\upsilon_t\upsilon_t'))(\text{vech}(\upsilon_t\upsilon_t'))'\right] - \tfrac{1}{4}\text{vech}(I_n)\,\text{vech}(I_n)',$$

a square matrix of order $n(n+1)/2$ (for the definition of υ_t, see (7)). An explicit expression of the expectation on the right hand side can be obtained from Wong and Wang (1992, p. 274). We also denote $\Pi_{i0} = \Pi_i(\vartheta_{10})$, $i = 1, ..., r$, and $\Pi_{00} = -I_n$, and define the partitioned matrix $C_{22}(\theta_0) = [C_{22}(a,b;\theta_0)]_{a,b=1}^{s}$ $(n^2 s \times n^2 s)$ where the $n^2 \times n^2$ matrix $C_{22}(a,b;\theta_0)$ is

$$\begin{aligned} C_{22}(a,b;\theta_0) &= \boldsymbol{\tau}_0 \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \sum_{i,j=0}^{r} \left(\Psi_{k+a-i,0}\Sigma_0\Psi_{k+b-j,0}' \otimes \Pi_{i0}'\Sigma_0^{-1}\Pi_{j0}\right) \\ &\quad + \sum_{i,j=0}^{r} \left(\Psi_{a-i,0}\Sigma_0^{1/2} \otimes \Pi_{i0}'\Sigma_0^{-1/2}\right)(4D_n J_0 D_n' - K_{nn})\left(\Sigma_0^{1/2}\Psi_{b-j,0}' \otimes \Sigma_0^{-1/2}\Pi_{j0}\right) \end{aligned}$$

(see the end of the introduction for the definitions of the commutation matrix K_{nn} and the duplication matrix D_n). Now set

$$\mathcal{I}_{\vartheta_2\vartheta_2}(\theta_0) = \nabla_2(\vartheta_{20})' C_{22}(\theta_0) \nabla_2(\vartheta_{20}),$$

which is the (limiting) information matrix of ϑ_2 (see Appendix B).

To be able to present the information matrix of the whole parameter vector ϑ we define the $n^2 \times n^2$ matrix

$$C_{12}(a,b;\theta_0) = -\boldsymbol{\tau}_0 \sum_{k=a}^{\infty} \sum_{i=0}^{r} \left(M_{k-a,0}\Sigma_0\Psi_{k+b-i,0}' \otimes \Sigma_0^{-1}\Pi_{i0}\right) + K_{nn}\left(\Psi_{b-a,0}' \otimes I_n\right)$$

and the $n^2 r \times n^2 s$ matrix $C_{12}(\theta_0) = [C_{12}(a,b;\theta_0)] = C_{21}(\theta_0)'$ $(a = 1, ..., r,\ b = 1, ..., s)$. Then the off-diagonal blocks of the (limiting) information matrix of ϑ are given by

$$\mathcal{I}_{\vartheta_1\vartheta_2}(\theta_0) = \nabla_1(\vartheta_{10})' C_{12}(\theta_0) \nabla_2(\vartheta_{20}) = \mathcal{I}_{\vartheta_2\vartheta_1}(\theta_0)'.$$

Combining the preceding definitions we now define the matrix

$$\mathcal{I}_{\vartheta\vartheta}(\theta) = \left[\mathcal{I}_{\vartheta_i\vartheta_j}(\theta)\right]_{i,j=1,2}.$$

For the remaining blocks of the information matrix of θ, we first define

$$\mathcal{I}_{\sigma\sigma}(\theta_0) = D_n'\left(\Sigma_0^{-1/2}\otimes\Sigma_0^{-1/2}\right)D_nJ_0D_n'\left(\Sigma_0^{-1/2}\otimes\Sigma_0^{-1/2}\right)D_n$$

and

$$\mathcal{I}_{\vartheta_2\sigma}(\theta_0) = -2\sum_{j=1}^{s}\frac{\partial}{\partial\vartheta_2}\phi_j'(\vartheta_2)\sum_{i=0}^{r}\left(\Psi_{j-i,0}\Sigma_0^{1/2}\otimes\Pi_{i0}'\Sigma_0^{-1/2}\right)D_nJ_0D_n'\left(\Sigma_0^{-1/2}\otimes\Sigma_0^{-1/2}\right)D_n$$

with $\mathcal{I}_{\vartheta_2\sigma}(\theta)' = \mathcal{I}_{\sigma\vartheta_2}(\theta)$. Finally, define

$$\mathcal{I}_{\lambda\lambda}(\theta_0) = \frac{\pi^{n/2}}{\Gamma(n/2)}\int_0^\infty\frac{\zeta^{n/2-1}}{f(\zeta;\lambda_0)}\left(\frac{\partial}{\partial\lambda}f(\zeta;\lambda_0)\right)\left(\frac{\partial}{\partial\lambda}f(\zeta;\lambda_0)\right)'d\zeta$$

and

$$\mathcal{I}_{\sigma\lambda}(\theta_0) = -D_n'\left(\Sigma_0^{-1/2}\otimes\Sigma_0^{-1/2}\right)D_n\mathrm{vech}(I_n)\frac{\pi^{n/2}}{n\Gamma(n/2)}\int_0^\infty\zeta^{n/2}\frac{f'(\zeta;\lambda_0)}{f(\zeta;\lambda_0)}\frac{\partial}{\partial\lambda'}f(\zeta;\lambda_0)\,d\zeta$$

with $\mathcal{I}_{\sigma\lambda}(\theta_0)' = \mathcal{I}_{\lambda\sigma}(\theta_0)$. Here the integrals are finite by Assumptions 2(iii) and 4(ii), and the Cauchy-Schwarz inequality.

The information matrix of the whole parameter vector θ is given by

$$\mathcal{I}_{\theta\theta}(\theta_0) = \begin{bmatrix} \mathcal{I}_{\vartheta_1\vartheta_1}(\theta_0) & \mathcal{I}_{\vartheta_1\vartheta_2}(\theta_0) & 0 & 0 \\ \mathcal{I}_{\vartheta_2\vartheta_1}(\theta_0) & \mathcal{I}_{\vartheta_2\vartheta_2}(\theta_0) & \mathcal{I}_{\vartheta_2\sigma}(\theta_0) & 0 \\ 0 & \mathcal{I}_{\sigma\vartheta_2}(\theta_0) & \mathcal{I}_{\sigma\sigma}(\theta_0) & \mathcal{I}_{\sigma\lambda}(\theta_0) \\ 0 & 0 & \mathcal{I}_{\lambda\sigma}(\theta_0) & \mathcal{I}_{\lambda\lambda}(\theta_0) \end{bmatrix}.$$

Note that in the scalar case $n = 1$ and in the purely noncausal case $r = 0$ the expressions of $\mathcal{I}_{\vartheta_2\vartheta_2}(\theta_0)$ and $\mathcal{I}_{\vartheta_1\vartheta_2}(\theta_0)$ simplify and $\mathcal{I}_{\vartheta_2\sigma}(\theta_0)$ becomes zero (see equality (B.6) in Appendix B). The latter fact means that in these special cases the parameters ϑ and (σ,λ) are orthogonal so that their ML estimators are asymptotically independent.

Before presenting the limiting distribution of the score of θ we introduce conditions needed to guarantee the positive definiteness of its covariance matrix. Specifically, we assume the following.

Assumption 5. (i) The matrices $\nabla_1(\vartheta_{10})$ $(rn^2\times m_1)$ and $\nabla_2(\vartheta_{20})$ $(sn^2\times m_2)$ are of full column rank.

(ii) The matrix $\begin{bmatrix} \mathcal{I}_{\sigma\sigma}(\theta_0) & \mathcal{I}_{\sigma\lambda}(\theta_0) \\ \mathcal{I}_{\lambda\sigma}(\theta_0) & \mathcal{I}_{\lambda\lambda}(\theta_0) \end{bmatrix}$ is positive definite.

Assumption 5(i) imposes conventional rank conditions on the first derivatives of the functions in Assumption 3. Assumption 5(ii) is analogous to what has been assumed in previous univariate models (see Andrews et al. (2006) and Lanne and Saikkonen (2011)). Note, however, that unlike in the univariate case it is here less obvious that this assumption is sufficient for the positive definiteness of the whole information matrix $\mathcal{I}_{\theta\theta}(\theta_0)$. The reason is that in the univariate case the situation is simpler in that the parameters λ and σ are orthogonal to the autoregressive parameters (here ϑ_1 and ϑ_2). In the present case the orthogonality of σ with respect to ϑ_2 generally fails but it is still possible to do without assuming more than assumed in the univariate case. Note also that, similarly to the aforementioned univariate cases, Assumption 5(ii) is not needed to guarantee the positive definiteness of $\mathcal{I}_{\sigma\sigma}(\theta_0)$. This follows from the definition of $\mathcal{I}_{\sigma\sigma}(\theta_0)$ and the facts that duplication matrices are of full column rank and the matrix J_0 is positive definite even in the Gaussian case (see Lemma 4 in Appendix B).

Now we can present the limiting distribution of the score.

Proposition 2. *Suppose that Assumptions 1–5 hold and that ϵ_t is non-Gaussian. Then,*

$$(T-s-r)^{-1/2}\sum_{t=r+1}^{T-s}\frac{\partial}{\partial\theta}g_t(\theta_0)\xrightarrow{d}N\left(0,\mathcal{I}_{\theta\theta}(\theta_0)\right),$$

where the matrix $\mathcal{I}_{\theta\theta}(\theta_0)$ is positive definite.

This result generalizes the corresponding univariate result given in Breidt et al. (1991) and Lanne and Saikkonen (2011). In the following section we generalize the work of these authors further by deriving the limiting distribution of the (approximate) ML estimator of θ_0. Note that for this result it is crucial that ϵ_t is assumed to be non-Gaussian because in the Gaussian case the information matrix $\mathcal{I}_{\theta\theta}(\theta_0)$ is singular (see the proof of Proposition 2, Step 2). As is well known (see Theorem 1 of Rothenberg (1971)), positive definiteness of the information matrix $\mathcal{I}_{\theta\theta}(\theta_0)$ can be viewed as a local identifiability condition for the parameter θ_0.

3.3 Limiting distribution of the approximate ML estimator

The expressions of the second partial derivatives of the log-likelihood function can be found in Appendix A. The following lemma shows that the expectations of these derivatives evaluated at the true parameter value agree with the corresponding elements of $-\mathcal{I}_{\theta\theta}(\theta_0)$. For this lemma we need the following assumption.

Assumption 6.(i) The integral $\int_0^\infty \zeta^{n/2-1}f'(\zeta;\lambda_0)\,d\zeta$ is finite, $\lim_{\zeta\to\infty}\zeta^{n/2+1}f'(\zeta;\lambda_0)=0$, and a finite right limit $\lim_{\zeta\to 0+}f'(\zeta;\lambda_0)$ exists.

(ii) There exists a function $f_2(\zeta)$ such that $\int_0^\infty \zeta^{n/2-1} f_2(\zeta)\, d\zeta < \infty$ and, in some neighborhood of λ_0, $\zeta \left|\partial f'(\zeta;\lambda)/\partial\lambda_i\right| \leq f_2(\zeta)$ and $\left|\partial^2 f(\zeta;\lambda)/\partial\lambda_i\partial\lambda_j\right| \leq f_2(\zeta)$ for all $\zeta \geq 0$ and $i, j = 1, ..., \mathrm{d}$.

Assumption 6(i) is analogous to Assumptions 2(ii) and (iii) except that it is formulated for the derivative $f'(\zeta;\lambda_0)$. Assumption 6(ii) imposes a standard dominance condition which guarantees that the expectation of $\partial^2 g_t(\theta_0)/\partial\lambda\partial\lambda'$ behaves in the desired fashion. It complements Assumption 4(i) which is formulated similarly to deal with the expectation of $\partial g_t(\theta_0)/\partial\lambda$. Now we can formulate the following lemma.

Lemma 2. *If Assumptions 1-6 hold then* $-(T-s-r)^{-1}\mathbb{E}_{\theta_0}\left[\partial^2 l_T(\theta_0)/\partial\theta\partial\theta'\right] = \mathcal{I}_{\theta\theta}(\theta_0)$.

Lemma 2 shows that the Hessian of the log-likelihood function evaluated at the true parameter value is related to the information matrix in the standard way, implying that $\partial^2 g_t(\theta_0)/\partial\theta\partial\theta'$ obeys a desired law of large numbers. However, to establish the asymptotic normality of the ML estimator more is needed, namely that $\partial^2 g_t(\theta)/\partial\theta\partial\theta'$ obeys a uniform law of large numbers in some neighborhood of θ_0. For that additional assumptions are required. As usual, it suffices to impose appropriate dominance conditions such as those given in the following assumption.

Assumption 7. For all $\zeta \geq 0$ and all λ in some neighborhood of λ_0, the functions

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

$$\frac{1}{f(\zeta;\lambda)}\left|\frac{\partial}{\partial\lambda_j} f'(\zeta;\lambda)\right|, \quad \frac{1}{f(\zeta;\lambda)}\left|\frac{\partial^2}{\partial\lambda_j\partial\lambda_k} f(\zeta;\lambda)\right|, \quad j,k = 1, ..., \mathrm{d},$$

are dominated by $a_1 + a_2\zeta^{a_3}$ with a_1, a_2, and a_3 nonnegative constants and $\int_0^\infty \zeta^{n/2+1+a_3} f(\zeta;\lambda_0)\, d\zeta < \infty$.

The dominance means that, for example, $\left(f'(\zeta;\lambda)/f(\zeta;\lambda)\right)^2 \leq a_1 + a_2\zeta^{a_3}$ for ζ and λ as specified. These dominance conditions are very similar to those assumed in condition (A7) of Andrews et al. (2006) and Lanne and Saikkonen (2011).

Now we can state the main result of this section.

Theorem 1. *Suppose that Assumptions 1–7 hold and that ϵ_t is non-Gaussian. Then there exists a sequence of (local) maximizers $\hat{\theta}$ of $l_T(\theta)$ in (17) such that*

$$(T-s-r)^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \mathcal{I}_{\theta\theta}(\theta_0)^{-1}\right).$$

Furthermore, $\mathcal{I}_{\theta\theta}(\theta_0)$ can consistently be estimated by $-(T-s-r)^{-1}\partial^2 l_T(\hat{\theta})/\partial\theta\partial\theta'$.

Theorem 1 shows that the usual result on asymptotic normality holds for a local maximizer of the likelihood function and that the limiting covariance matrix can consistently be estimated with the Hessian of the log-likelihood function. Based on these results and arguments used in their proof, conventional likelihood based tests with limiting chi-square distribution can be obtained. It is worth noting, however, that consistent estimation of the limiting covariance matrix cannot be based on the outer product of the first derivatives of the log-likelihood function. Specifically, $(T-s-r)^{-1}\sum_{t=r+1}^{T-s}(\partial g_t(\hat{\theta})/\partial\theta)(\partial g_t(\hat{\theta})/\partial\theta')$ is, in general, not a consistent estimator of $\mathcal{I}_{\theta\theta}(\theta_0)$. The reason is that the components of the score vector are serially correlated so that what is needed is an estimator of the long-run covariance matrix of the (stationary) process $\partial g_t(\theta_0)/\partial\theta$. This is not obtained by the aforementioned estimator which does not take nonzero covariances between $\partial g_t(\theta_0)/\partial\theta$ and $\partial g_k(\theta_0)/\partial\theta$, $k \neq t$, into account. Such covariances are responsible, for example, for the term $K_{nn}\left(\Psi'_{b-a}\otimes I_n\right)$ in $\mathcal{I}_{\vartheta_1\vartheta_2}(\theta_0)$ (for details we refer to the definition of $C_{12}(a,b;\theta_0)$ and the related proof in the Supplementary Appendix). However, being based on the Hessian of the log-likelihood function the estimator given in Theorem 1 works as usual estimating $-E_{\theta_0}\left[\partial^2 g_t(\theta_0)/\partial\theta\partial\theta'\right] = \mathcal{I}_{\theta\theta}(\theta_0)$ consistently (see Lemma 2). This, in turn, is due to the fact that $\partial^2 g_t(\theta)/\partial\theta\partial\theta'$ is a stationary and ergodic process obeying a uniform law of large numbers (see the proof of Theorem 1).

4 Empirical application

We illustrate the use of the noncausal VAR model with an application to U.S. interest rate data. Specifically, we consider the expectations hypothesis of the term structure of interest rates, according to which the long-term interest rate is a weighted sum of present and expected future short-term interest rates. Campbell and Shiller (1991) suggested testing the expectations hypothesis by testing the restrictions it imposes on the parameters of a bivariate VAR model for the change in the short-term interest rate and the spread between the long-term and short-term interest rates. The general idea is that a causal VAR model captures the dynamics of interest rates, and therefore, its forecasts can be considered as investors' expectations. If these expectations are rational, i.e., they do not systematically deviate from the observed values, this together with the expectations hypothesis imposes testable restrictions on the parameters of the VAR model. This method, already proposed by Sargent (1979), is straightforward to implement and widely applied in economics besides this particular application. However, it crucially depends on the causality of the employed VAR model, suggesting that the validity of this assumption should be checked to avoid potentially misleading conclusions. If the selected VAR model turns out to be noncausal, the estimates may yield evidence in favor of or against the

expectations hypothesis. In particular, according to the expectations hypothesis, the expected changes in the short rate drive the term structure, and therefore, their coefficients in the Φ matrices should be significant in the equation of the spread.

The specification of a potentially noncausal VAR model is carried out along the same lines as in the univariate case in Breidt et al. (1991) and Lanne and Saikkonen (2011). The first step is to fit a conventional causal VAR model by least squares or Gaussian ML and determine its order by using conventional procedures such as diagnostic checks and model selection criteria. We deem a causal model adequate when its residuals show no signs of serial correlation, and, once such an adequate causal model is found, we check its residuals for Gaussianity. As already discussed, it makes sense to proceed to noncausal models only if deviations from Gaussianity are detected. If this happens, a non-Gaussian error distribution is adopted and all causal and noncausal models of the selected order are estimated. Of these models the one that maximizes the log-likelihood function is selected and its adequacy is checked by diagnostic tests.

We use the Ljung-Box and McLeod-Li tests to check for error autocorrelation and conditional heteroskedasticity, respectively. Note, however, that when the orders of the model are misspecified, these tests are not exactly valid as they do not take estimation errors correctly into account. For instance, as discussed in Section 2.3, squared errors of misspecified noncausal VAR models are, in general, serially correlated, implying that the conventional limiting distribution of the Ljung-Box test does not apply (cf. Francq, Roy, and Zakoïan (2005)). The reason is that misspecification of the model orders makes the errors dependent. Nevertheless, p-values of these tests can be seen as convenient summary measures of the autocorrelation remaining in the residuals and their squares. A similar remark applies to the Shapiro-Wilk test we use to check the normality of the errors.

Our data set comprises the (demeaned) change in the six-month interest rate (Δr_t) and the spread between the five-year and six-month interest rates (S_t) (quarter-end yields on U.S. zero-coupon bonds) from the thirty-year period 1967:1–1996:4 (120 observations) previously used in Duffee (2002). The two series are depicted in Figure 1. The AIC and BIC select Gaussian VAR(3) and VAR(1) models, respectively, but only the third-order model produces serially uncorrelated errors. However, the results in Table 1 show that its squared residuals are autocorrelated, and the Q-Q plots in the upper panel of Figure 2 indicate considerable deviations from normality. The p-values of the Shapiro-Wilk test for the residuals of the equations of Δr_t and S_t equal 5.06e–9 and 7.23e–7, respectively. Because the most severe violations of normality occur at the tails, a more leptokurtic distribution, such as the multivariate t-distribution, might prove suitable for these data.

Results of diagnostic checks of all four third-order VAR models with t-distributed

errors are summarized in Table 1.[5] By a wide margin, the specification maximizing the log-likelihood function is the VAR(2,1)-t model. It also turns out to be the only one of the estimated models that shows no signs of remaining autocorrelation in the residuals or their squares. Given this, it is interesting to recall from Section 2.3 that, when the (non-Gaussian) data generation process is noncausal, squared residuals of a causal VAR model or a misspecified noncausal VAR model tend to be autocorrelated. The Q-Q plots of the residuals in the lower panel of Figure 2 lend support to the adequacy of the multivariate t-distribution of the errors. In particular, the t-distribution seems to capture the tails reasonably well. Moreover, the estimate of the degrees-of-freedom parameter λ is small (4.085), suggesting inadequacy of the Gaussian error distribution. Thus, there is clear evidence of noncausality.

The estimates of the preferred model are presented in Table 2. The estimated Φ_1 matrix seems to have an interpretation that goes contrary to the implications of the expectations hypothesis discussed above: the estimate of $\Phi_{1,21}$ is insignificant at conventional significance levels, indicating that an expected increase of the short-term rate has no significant effect on the spread. Furthermore, an expected future increase of the spread tends to decrease the short-term rate and increase the spread, i.e., the estimates of $\Phi_{1,12}$ and $\Phi_{1,22}$ are both significant at 1% level, with the former being negative and the latter positive. This may be interpreted in favor of (expected) time-varying term premia driving the term structure instead of expectations of future short-term rates as implied by the expectations hypothesis.

In sum, the presence of a noncausal VAR representation of Δr_t and S_t invalidates the test of the expectations hypothesis suggested by Campbell and Shiller (1991). Furthermore, the estimation results of the noncausal VAR model lend little support to the expectations hypothesis. If noncausality prevails more generally in interest rates, this might also explain the common rejections of the expectations hypothesis when testing is based on the assumption of a causal VAR model.

5 Conclusion

In this paper, we have proposed a new noncausal VAR model that contains the commonly used causal VAR model as a special case. Under Gaussianity, causal and noncausal VAR

[5] The density function of the multivariate t-distribution for an n-dimensional random vector x with mean zero, λ degerees of freedom, and covariance matrix $\frac{\lambda}{\lambda-2}\Sigma$ is given by

$$f_\Sigma(x;\lambda) = \frac{\Gamma\left[(\lambda+n)/2\right]}{(\lambda\pi)^{n/2}\,\Gamma(\lambda/2)\,\sqrt{\det(\Sigma)}}\left(1+\frac{1}{\lambda}x'\Sigma^{-1}x\right)^{-(\lambda+n)/2}$$

where $\Gamma(\cdot)$ is the gamma function and $\lambda > 2$ is assumed.

models cannot be distinguished which underlines the importance of careful specification of the error distribution of the model. Assuming that the error distribution belongs to a fairly general class of elliptical distributions we derived asymptotic properties of an approximate (local) ML estimator in the noncausal VAR model. The potential usefulness of the noncausal VAR model was illustrated by means of an empirical application to the U.S. term structure of interest rates. In that case we successfully employed an extension of the model selection procedure presented by Breidt et al. (1991) and Lanne and Saikkonen (2011) in the corresponding univariate case and found evidence of noncausality. This finding invalidates the previously employed test of the expectations hypothesis of the term structure of interest rates explicitly based on a causal VAR model.

While the new model appears useful in providing a more accurate description of time series dynamics and checking for the validity of a causal VAR representation, it may also have other uses. For instance, in economic applications, we expect noncausal VAR models to be valuable in checking for so-called nonfundamentalness. In economics, a model is said to exhibit nonfundamentalness if its solution explicitly depends on the future so that it does not have a causal VAR representation (for a recent survey of the relevant literature, see Alessi, Barigozzi, and Capasso (2011)). Hence, nonfundamentalness is closely related to noncausality, and checking for noncausality can be seen as one way of testing for nonfundamentalness. Because nonfundamentalness often invalidates the use of conventional econometric methods, being able to detect it in advance is important. However, the test procedures suggested in the previous literature are not very convenient and have not been much applied in practice.

Checking for causality (or fundamentalness) is an important application of our methods, but it can only be considered as the first step in the empirical analysis of time series data. Once noncausality has been detected, it would be natural to use the noncausal VAR model for forecasting and structural analysis. These, however, require methods that are not readily available. Because the prediction problem in noncausal VAR models is generally nonlinear (cf. Rosenblatt (2000, Chapter 5)) methods used in the causal case are not applicable and, due to the explicit dependence on the future, the same is true for conventional simulation-based methods. In the univariate case, Lanne, Luoto, and Saikkonen (2012) have proposed a forecasting method that could plausibly be extended to the noncausal VAR model.

Regarding statistical aspects, the theory presented in this paper is confined to the class of elliptical distributions. Even though the multivariate t-distribution belonging to this class seemed adequate in our empirical applications, it would be desirable to make extensions to other relevant classes of distributions. Also, the finite-sample properties of the employed model selection procedure could be examined by means of simulation

experiments. We leave all of these issues for future research.

Mathematical Appendix

Some of the mathematical derivations require rather long and tedious calculations. In what follows, we shall therefore omit several details which can be found in the Supplementary Appendix.

A Derivatives of the log-likelihood function

It will be sufficient to consider the derivatives of $g_t(\theta)$ which can be obtained by straightforward differentiation. To simplify notation we set $h(\zeta;\lambda) = f'(\zeta;\lambda)/f(\zeta;\lambda)$,

$$e_t(\theta) = h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\Sigma^{-1/2}\epsilon_t(\vartheta) \quad \text{and} \quad e_{0t} = e_t(\theta_0). \tag{A.1}$$

Then,

$$h'\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right) = \frac{f''\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)}{f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)} - \left(\frac{f'\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)}{f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)}\right)^2 \tag{A.2}$$

and (see (7))

$$e_{0t} \overset{d}{=} \rho_t h\left(\rho_t^2;\lambda_0\right)\upsilon_t = \rho_t h_0\left(\rho_t^2\right)\upsilon_t, \tag{A.3}$$

where the latter equality defines the notation $h_0(\cdot) = h(\cdot;\lambda_0)$.

First derivatives of $l_T(\theta)$. First, conclude from (19) that

$$\tfrac{\partial}{\partial\vartheta_1}\epsilon_t'(\vartheta) = -\sum_{i=1}^{r}\tfrac{\partial}{\partial\vartheta_1}\pi_i'(\vartheta_1)\left(u_{t-i}(\vartheta_2)\otimes I_n\right) \tag{A.4}$$

and

$$\tfrac{\partial}{\partial\vartheta_2}\epsilon_t'(\vartheta) = \sum_{i=0}^{r}\sum_{j=1}^{s}\tfrac{\partial}{\partial\vartheta_2}\phi_j'(\vartheta_2)\left(y_{t+j-i}\otimes\Pi_i'\right), \tag{A.5}$$

with $\Pi_0 = -I_n = \Pi_{00}$. With this notation and $\sigma = \text{vech}(\Sigma)$ one obtains from (18) that

$$\begin{aligned}
\tfrac{\partial}{\partial\vartheta_i}g_t(\theta) &= 2h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\tfrac{\partial}{\partial\vartheta_i}\epsilon_t'(\vartheta)\Sigma^{-1}\epsilon_t(\vartheta), \quad i=1,2\\
\tfrac{\partial}{\partial\sigma}g_t(\theta) &= -h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\mathcal{G}\left(\Sigma^{-1}\right)'\left(\epsilon_t(\vartheta)\otimes\epsilon_t(\vartheta)\right) - \tfrac{1}{2}D_n'\text{vec}\left(\Sigma^{-1}\right)\\
\tfrac{\partial}{\partial\lambda}g_t(\theta) &= \tfrac{1}{f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)}\tfrac{\partial}{\partial\lambda}f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right),
\end{aligned}$$

where, for brevity, $\mathcal{G}\left(\Sigma^{-1}\right)' = D_n'\left(\Sigma^{-1}\otimes\Sigma^{-1}\right)$.

Second derivatives of $l_T(\theta)$. Using the fact that

$$\begin{aligned}
\tfrac{\partial}{\partial\vartheta_i'}e_t(\theta) &= h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\Sigma^{-1/2}\tfrac{\partial}{\partial\vartheta_i'}\epsilon_t(\vartheta) \qquad (A.6)\\
&\quad +2h'\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\Sigma^{-1/2}\epsilon_t(\vartheta)\,\epsilon_t(\vartheta)'\Sigma^{-1}\tfrac{\partial}{\partial\vartheta_i'}\epsilon_t(\vartheta), \quad i=1,2,
\end{aligned}$$

we first have

$$\begin{aligned}\frac{\partial^2}{\partial\vartheta_1\partial\vartheta_1'}g_t(\theta) &= -2\sum_{i=1}^r\left(u_{t-i}(\vartheta_2)'\otimes e_t(\theta)'\Sigma^{-1/2}\otimes I_{m_1}\right)\frac{\partial}{\partial\vartheta_1'}\mathrm{vec}(\frac{\partial}{\partial\vartheta_1}\pi_i'(\vartheta_1))\\ &\quad -2\sum_{i=1}^r\frac{\partial}{\partial\vartheta_1}\pi_i'(\vartheta_1)(u_{t-i}(\vartheta_2)\otimes I_n)\Sigma^{-1/2}\frac{\partial}{\partial\vartheta_1'}e_t(\theta)\\ \frac{\partial^2}{\partial\vartheta_2\partial\vartheta_2'}g_t(\theta) &= 2\sum_{j=1}^s\sum_{i=0}^r\left(y_{t+j-i}'\otimes e_t(\theta)'\Sigma^{-1/2}\Pi_i\otimes I_{m_2}\right)\frac{\partial}{\partial\vartheta_2'}\mathrm{vec}(\frac{\partial}{\partial\vartheta_2}\phi_j'(\vartheta_2))\\ &\quad +2\sum_{j=1}^s\frac{\partial}{\partial\vartheta_2}\phi_j'(\vartheta_2)\sum_{i=0}^r(y_{t+j-i}\otimes\Pi_i')\Sigma^{-1/2}\frac{\partial}{\partial\vartheta_2'}e_t(\theta)\\ \frac{\partial^2}{\partial\vartheta_1\partial\vartheta_2'}g_t(\theta) &= -2\sum_{i=1}^r\frac{\partial}{\partial\vartheta_1}\pi_i'(\vartheta_1)\left(I_n\otimes\Sigma^{-1/2}e_t(\theta)\right)\frac{\partial}{\partial\vartheta_2'}u_{t-i}(\vartheta_2)\\ &\quad -2\sum_{i=1}^r\frac{\partial}{\partial\vartheta_1}\pi_i'(\vartheta_1)(u_{t-i}(\vartheta_2)\otimes I_n)\Sigma^{-1/2}\frac{\partial}{\partial\vartheta_2'}e_t(\theta),\end{aligned}$$

where $\partial u_{t-i}(\vartheta_2)/\partial\vartheta_2' = -\sum_{j=1}^s(y_{t+j-i}'\otimes I_n)\partial\phi_j(\vartheta_2)/\partial\vartheta_2'$ (see below (19)).

Next,

$$\begin{aligned}\frac{\partial^2}{\partial\sigma\partial\sigma'}g_t(\theta) &= 2h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)D_n'\left(\Sigma^{-1}\epsilon_t(\vartheta)\epsilon_t(\vartheta)'\Sigma^{-1}\otimes\Sigma^{-1}\right)D_n+\tfrac{1}{2}D_n'\mathcal{G}(\Sigma^{-1})\\ &\quad +h'\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\mathcal{G}(\Sigma^{-1})'\left(\epsilon_t(\vartheta)\epsilon_t(\vartheta)'\otimes\epsilon_t(\vartheta)\epsilon_t(\vartheta)'\right)\mathcal{G}(\Sigma^{-1})\\ \frac{\partial^2}{\partial\vartheta_i\partial\sigma'}g_t(\theta) &= -2h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\left(\epsilon_t(\vartheta)'\otimes\frac{\partial}{\partial\vartheta_i}\epsilon_t'(\vartheta)\right)\mathcal{G}(\Sigma^{-1})\\ &\quad -2h'\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\frac{\partial}{\partial\vartheta_i}\epsilon_t'(\vartheta)\Sigma^{-1}\epsilon_t(\vartheta)\left(\epsilon_t(\vartheta)'\otimes\epsilon_t(\vartheta)'\right)\mathcal{G}(\Sigma^{-1}),\quad i=1,2\\ \frac{\partial^2}{\partial\sigma\partial\lambda'}g_t(\theta) &= -\mathcal{G}(\Sigma^{-1})'(\epsilon_t(\vartheta)\otimes\epsilon_t(\vartheta))\frac{\partial}{\partial\lambda'}h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right).\end{aligned}$$

Finally,

$$\begin{aligned}\frac{\partial^2}{\partial\lambda\partial\lambda'}g_t(\theta) &= -\frac{1}{\left(f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\right)^2}\frac{\partial}{\partial\lambda}f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\frac{\partial}{\partial\lambda'}f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\\ &\quad +\frac{1}{f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)}\frac{\partial^2}{\partial\lambda\partial\lambda'}f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\\ \frac{\partial^2}{\partial\vartheta_i\partial\lambda'}g_t(\theta) &= 2\frac{\partial}{\partial\vartheta_i}\epsilon_t'(\vartheta)\Sigma^{-1}\epsilon_t(\vartheta)\frac{\partial}{\partial\lambda'}h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right),\quad i=1,2,\end{aligned}$$

where

$$\begin{aligned}\frac{\partial}{\partial\lambda'}h\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right) &= \frac{1}{f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)}\frac{\partial}{\partial\lambda'}f'\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\\ &\quad -\frac{f'\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)}{\left(f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right)\right)^2}\frac{\partial}{\partial\lambda'}f\left(\epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta);\lambda\right).\end{aligned}$$

B Proofs for Sections 2 and 3

Proof of Lemma 1. For the former inequality, first conclude from the definition of the function h (see the beginning of Appendix A) and the density of ρ_t^2 (see (10)) that

$$\mathbb{E}_\lambda\left[\rho_t^2h\left(\rho_t^2;\lambda\right)\right] = \frac{\pi^{n/2}}{\Gamma(n/2)}\int_0^\infty\zeta^{n/2}f'(\zeta;\lambda)\,d\zeta = -\frac{n}{2}. \tag{B.1}$$

Here the latter equality follows because

$$\int_0^\infty\zeta^{n/2}f'(\zeta;\lambda)\,d\zeta = \zeta^{n/2}f(\zeta;\lambda)\,|_0^\infty - \frac{n}{2}\int_0^\infty\zeta^{n/2-1}f(\zeta;\lambda)\,d\zeta = -\frac{n\Gamma(n/2)}{2\pi^{n/2}}$$

by Assumptions 2(ii) and (iii) (cf. Fang et al. (1990, p. 35)). Equation (B.1), the Cauchy-Schwarz inequality, and the definition of $\boldsymbol{j}(\lambda)$ (see (11)) yield

$$\begin{aligned} 1 &= \left\{ \frac{2\pi^{n/2}}{n\Gamma(n/2)} \int_0^\infty \zeta^{n/4} \frac{f'(\zeta;\lambda)}{\sqrt{f(\zeta;\lambda)}} \zeta^{n/4} \sqrt{f(\zeta;\lambda)} d\zeta \right\}^2 \\ &\leq \frac{4\pi^{n/2}}{n\Gamma(n/2)} \int_0^\infty \zeta^{n/2} \frac{(f'(\zeta;\lambda))^2}{f(\zeta;\lambda)} d\zeta \cdot \frac{\pi^{n/2}}{n\Gamma(n/2)} \int_0^\infty \zeta^{n/2} f(\zeta;\lambda)\, d\zeta \qquad \text{(B.2)} \\ &= \boldsymbol{j}(\lambda) \cdot \mathbb{E}_\lambda\left(\rho_t^2\right)/n. \end{aligned}$$

Thus, we have shown the claimed inequality.

From the preceding proof it is seen that equality holds if and only if there is equality in (B.2). As is well known, this happens if and only if $\zeta^{n/4} f'(\zeta;\lambda)/\sqrt{f(\zeta;\lambda)}$ is proportional to $\zeta^{n/4}\sqrt{f(\zeta;\lambda)}$ or if and only if

$$\frac{f'(\zeta;\lambda)}{f(\zeta;\lambda)} = \frac{\partial}{\partial\zeta} \log f(\zeta;\lambda) = c \quad \text{for some constant } c.$$

This implies $f(\zeta;\lambda) = b\exp(-a\zeta)$ with $a > 0$ and $b > 0$. From the fact that $f(x'x;\lambda)$, $x \in \mathbb{R}^n$, is the density function of $\rho_t \upsilon_t$ (see (7) and (8)) it further follows that $b = (a/\pi)^{n/2}$ and that $\rho_t \upsilon_t$ has the normal density $(2\pi)^{-n/2}\exp(-x'x/2)$. Here the identity covariance matrix is obtained because $\rho_t^2 \sim \chi_n^2$, and hence from (9), $\mathbb{C}(\rho_t^2 \upsilon_t) = I_n$ (cf. the corollary to Lemma 1.4 and Example 1.3 of Fang et al. (1990, p. 23). Thus, ϵ_t is Gaussian as a linear transformation of $\rho_t \upsilon_t$. On the other hand, if ϵ_t is Gaussian the equality $f'(\zeta;\lambda)/f(\zeta;\lambda) = c$ clearly holds with $c = -1/2$ and, because then $\rho_t^2 \sim \chi_n^2$, $\mathbb{E}_\lambda(\rho_t^2) = n$ and $\boldsymbol{j}(\lambda) = 1$. This completes the proof for $\boldsymbol{j}(\lambda)$. The proof for $i(\lambda)$ makes use of similar arguments. Details can be found in the Supplementary Appendix. □

Proof of Proposition 1. The proof is obtained as an application of Theorem 7 of Chan and Ho (2004). First note that, by condition (2), the coefficient matrices Ψ_j are square summable and the matrix $\sum_{j=-\infty}^{\infty} \Psi_j e^{-ij\omega}$ is nonsingular for all $\omega \in (-\pi, \pi]$. Thus, conditions (i)-(iii) of Lemma 2 of Chan and Ho (2004) hold and we only need to verify conditions (D1) and (D2) of their Theorem 7 (or Conditions 5 and 6 in Chan et al. (2006)). The latter requires that any two linear combinations of ϵ_t with nonzero coefficient vectors must be dependent. Thus, let $\alpha_1'\epsilon_t$ and $\alpha_2'\epsilon_t$ be such linear combinations ($\alpha_1 \neq 0 \neq \alpha_2$). By Theorem 2.16 of Fang et al. (1990) the bivariate random vector $(\alpha_1'\epsilon_t, \alpha_2'\epsilon_t)$ is elliptically distributed and also non-Gaussian because ϵ_t is non-Gaussian by assumption. Now, if $\alpha_1'\epsilon_t$ and $\alpha_2'\epsilon_t$ are independent they are uncorrelated, and by Theorem 4.11 of Fang et al. (1990), $(\alpha_1'\epsilon_t, \alpha_2'\epsilon_t)$ is necessarily Gaussian. As this is a contradiction, condition (D2) of Chan and Ho (2004) holds.

To verify condition (D1) of Chan and Ho (2004), let $\mathsf{cum}(\cdot, ..., \cdot)$ signify the cumulant of the indicated random variables. As in the example of Chan and Ho (2004) after their

Theorem 7, it suffices to show that the symmetric matrix $\mathsf{cum}\left(\epsilon_t, \epsilon_t, \epsilon_{1t}, \epsilon_{1t}\right)$ is nonsingular (here ϵ_{1t} is the first component of ϵ_t). To this end, we first show that this matrix is positive definite when the kurtosis of ϵ_t, denoted by $\mathcal{K}$, is positive (note that $|\mathcal{K}| < \infty$ because ρ_t, and hence the components of ϵ_t have finite fourth moments, as discussed after Assumption 2). Denote $\iota = (1, 0, ..., 0)$ $(n \times 1)$ and let α $(n \times 1)$ be an arbitrary nonzero constant vector. Then,

$$\begin{aligned} \alpha' \mathsf{cum}\left(\epsilon_t, \epsilon_t, \epsilon_{1t}, \epsilon_{1t}\right) \alpha &= \mathsf{cum}\left(\alpha'\epsilon_t, \alpha'\epsilon_t, \iota'\epsilon_t, \iota'\epsilon_t\right) \\ &= \mathcal{K}\left\{\alpha'\mathbb{E}\left(\epsilon_t\epsilon_t'\right)\alpha \cdot \iota'\mathbb{E}\left(\epsilon_t\epsilon_t'\right)\iota + 2\left[\alpha'\mathbb{E}\left(\epsilon_t\epsilon_t'\right)\iota\right]^2\right\} \end{aligned}$$

(cf. Muirhead and Waternaux (1980, p. 33)). The last expression is positive, showing the desired result. If $\mathcal{K} < 0$ it can similarly be seen that the matrix $\mathsf{cum}\left(\epsilon_t, \epsilon_t, \epsilon_{1t}, \epsilon_{1t}\right)$ is negative definite. Thus, we have verified condition (D1) of Chan and Ho (2004). □

Next we present some auxiliary results needed to prove Proposition 2. Here the true parameter value is assumed, so the notation $\mathbb{E}\left(\cdot\right)$ will be used instead of $\mathbb{E}_{\lambda_0}\left(\cdot\right)$ and similarly for $\mathbb{C}\left(\cdot\right)$. In these proofs frequent use will be made of well-known properties of the Kronecker product and the vec operator, especially the result $\mathrm{vec}(ABC) = (C' \otimes A)\mathrm{vec}(B)$ which holds for any conformable matrices A, B, and C. This and other results of matrix algebra to be employed can be found in Lütkepohl (1996). To simplify notation, we define $\varepsilon_t = \Sigma_0^{-1/2}\epsilon_t$ and note that (see (7))

$$\varepsilon_t \overset{d}{=} \rho_t \upsilon_t. \tag{B.3}$$

We will also frequently write $f\left(\cdot; \lambda_0\right) = f_0\left(\cdot\right)$ and similarly for $f_0'\left(\cdot\right)$ and $f_0''\left(\cdot\right)$.

Lemma 3. *Under the conditions of Proposition 2,*

$$\mathbb{E}\left(e_{0t}\right) = 0 \quad \textit{and} \quad \mathbb{C}\left(e_{0t}\right) = \frac{\boldsymbol{j}_0}{4} I_n, \tag{B.4}$$

and

$$\mathbb{C}\left(\varepsilon_t, e_{0k}\right) = \begin{cases} 0, & \textit{if } \ t \neq k \\ -\frac{1}{2} I_n, & \textit{if } \ t = k \end{cases} \tag{B.5}$$

Proof. By the definition of the function $h_0\left(\cdot\right)$ (see (A.3)) and the density of ρ_t^2 (see (10)),

$$\mathbb{E}\left[\rho_t^2\left(h_0\left(\rho_t^2\right)\right)^2\right] = \frac{\pi^{n/2}}{\Gamma\left(n/2\right)} \int_0^\infty \zeta^{n/2} \frac{\left(f_0'\left(\zeta\right)\right)^2}{f_0\left(\zeta\right)} d\zeta = \frac{n}{4}\boldsymbol{j}_0,$$

where the latter equality is due to (11). Thus, because $\mathbb{E}\left(\upsilon_t\right) = 0$ and $\mathbb{C}\left(\upsilon_t\right) = n^{-1} I_n$, independence of the processes ρ_t and υ_t in conjunction with (A.3) proves (B.4). The same arguments and (B.3) yield $\mathbb{E}\left(\varepsilon_t e_{0k}'\right) = \mathbb{E}\left[\rho_t \rho_k h_0\left(\rho_k^2\right)\right] \mathbb{E}\left(\upsilon_t \upsilon_k'\right)$, where $\mathbb{E}\left(\upsilon_t \upsilon_k'\right) = 0$ for $t \neq k$. Thus, one obtains (B.5) from this and (B.1). □

Lemma 4. . *Under the conditions of Proposition 2,*

$$\mathbb{C}\left(\varepsilon_{t-i}\otimes e_{0t},\varepsilon_{k-j}\otimes e_{0k}\right)=\begin{cases} D_nJ_0D_n', & \text{if } t=k,\ i=j=0\\ \frac{\tau_0}{4}I_{n^2}, & \text{if } t=k,\ i=j\neq 0\\ \frac{1}{4}K_{nn}, & \text{if } t\neq k,\ i=t-k,\ j=k-t\\ 0, & \text{otherwise.}\end{cases}$$

Moreover, the matrix J_0 is positive definite even when ϵ_t is Gaussian.

Proof. We only prove the case $t=k$ and $i=j=0$. The other cases can be established with similar arguments (details are available in the Supplementary Appendix). First note that $\varepsilon_{t-i}\otimes e_{0t}\overset{d}{=}\rho_{t-i}\rho_t h_0\left(\rho_t^2\right)\left(\upsilon_{t-i}\otimes\upsilon_t\right)$ (see (B.3) and (A.3)). This and independence of ρ_t and υ_t yields

$$\mathbb{E}\left(\varepsilon_t\otimes e_{0t}\right)=\mathbb{E}\left[\rho_t^2h_0\left(\rho_t^2\right)\right]\mathbb{E}\left(\upsilon_t\otimes\upsilon_t\right)=-\frac{1}{2}D_n\text{vech}\left(I_n\right),$$

where the latter equality is due to (B.1) and the fact $\mathbb{E}\left(\upsilon_t\otimes\upsilon_t\right)=n^{-1}\text{vec}(I_n)$. By the same arguments we also find that

$$\mathbb{E}\left[(\varepsilon_t\otimes e_{0t})(\varepsilon_t\otimes e_{0t})'\right]=\mathbb{E}\left[\rho_t^4\left(h_0\left(\rho_t^2\right)\right)^2\right]\mathbb{E}\left(\upsilon_t\upsilon_t'\otimes\upsilon_t\upsilon_t'\right)=\boldsymbol{i}_0\mathbb{E}\left(\upsilon_t\upsilon_t'\otimes\upsilon_t\upsilon_t'\right),$$

where the latter equality follows from the definition of $\boldsymbol{i}_0$ (see (12)). Because

$$\mathbb{E}\left(\upsilon_t\upsilon_t'\otimes\upsilon_t\upsilon_t'\right)=\mathbb{E}\left[\left(\upsilon_t\otimes\upsilon_t\right)\left(\upsilon_t'\otimes\upsilon_t'\right)\right]=D_n\mathbb{E}\left[\left(\text{vech}(\upsilon_t\upsilon_t')\right)\left(\text{vech}(\upsilon_t\upsilon_t')\right)'\right]D_n',$$

the stated result is obtained from the definition of the matrix J_0.

The matrix J_0 is clearly symmetric and from the definition of $\boldsymbol{i}_0$ and (B.1) it follows that, even when ϵ_t is Gaussian, $\boldsymbol{i}_0>\left\{\mathbb{E}\left[\rho_t^2h_0\left(\rho_t^2\right)\right]\right\}^2=n^2/4$, where the inequality is strict because ρ_t^2 has positive density. Now, let x be a nonzero $n\times 1$ vector and conclude from the preceding inequality and the definition of J_0 that

$$4x'J_0x>n^2x'\mathbb{E}\left[\left(\text{vech}(\upsilon_t\upsilon_t')\right)\left(\text{vech}(\upsilon_t\upsilon_t')\right)'\right]x-x'\text{vech}\left(I_n\right)\text{vech}\left(I_n\right)'x.$$

As $\mathbb{E}\left[\text{vech}(\upsilon_t\upsilon_t')\right]=n^{-1}\text{vech}(I_n)$, the right hand side equals $n^2x'\mathbb{C}\left(\text{vech}(\upsilon_t\upsilon_t')\right)x$, which is clearly nonnegative and, consequently, J_0 is positive definite. □

Proof of Proposition 2. The proof consists of three steps. The first one shows that the expectation of the score of θ at the true parameter value is zero and its limiting covariance matrix is $\mathcal{I}_{\theta\theta}\left(\theta_0\right)$. The positive definiteness of $\mathcal{I}_{\theta\theta}\left(\theta_0\right)$ is established in the second step and the third step proves the asymptotic normality of the score.

Step 1. We only give a proof for the score of ϑ_2 which differs most from standard cases (proofs of the other cases are available in the Supplementary Appendix). First

we demonstrate that the expectation of this score is zero. Denote $\Phi_0(z)^{-1} = L_0(z) = \sum_{j=0}^{\infty} L_{j0} z^j$ where $L_{00} = I_n$ and the subscript zero again refers to true parameter values. We also define $L_{j0} = 0$ for $j < 0$. Using the identity $L_0(z^{-1}) = \Psi_0(z)\Pi_0(z)$ it can be seen that

$$-\sum_{i=0}^{r} \Psi_{j-i,0}\Pi_{i0} = \begin{cases} 0, & j > 0 \\ I_n, & j = 0 \\ L_{-j0}, & j < 0, \end{cases} \tag{B.6}$$

where $\Pi_{00} = -I_n$. To simplify notation we denote

$$A_0(k,i) = \Psi_{k0}\Sigma_0^{1/2} \otimes \Pi_{i0}'\Sigma_0^{-1/2} \quad \text{and} \quad B_0(d) = M_{d0}\Sigma_0^{1/2} \otimes \Sigma_0^{-1/2}$$

and note that, by (B.6),

$$\sum_{i=0}^{r} A_0(a-i,i)\,\mathrm{vec}(I_n) = \mathrm{vec}\left(\sum_{i=0}^{r} \Pi_{i0}'\Psi_{a-i,0}'\right) = 0, \quad a \in \{1,...,s\}. \tag{B.7}$$

Next, observe that

$$\frac{\partial}{\partial \vartheta_2} g_t(\theta_0) = 2\sum_{j=1}^{s} \frac{\partial}{\partial \vartheta_2}\phi_j'(\vartheta_{20}) \sum_{i=0}^{r} (y_{t+j-i} \otimes \Pi_{i0}')\,\Sigma_0^{-1/2} e_{0t} \tag{B.8}$$

(see Appendix A) and consider the expectation

$$\mathbb{E}\left(\sum_{i=0}^{r} (y_{t+a-i} \otimes \Pi_{i0}')\,\Sigma_0^{-1/2} e_{0t}\right) = \sum_{i=0}^{r}\sum_{k=-\infty}^{\infty} A_0(k,i)\,\mathbb{E}\left(\varepsilon_{t+a-i-k} \otimes e_{0t}\right),$$

which is obtained by using equation (5), the definition of $A_0(k,i)$, and the definition $\varepsilon_t = \Sigma_0^{-1/2}\epsilon_t$. By Lemma 3, the expectation in the last expression equals zero if $k \neq a-i$ and $-\frac{1}{2}\mathrm{vec}(I_n)$ if $k = a-i$. From this and (B.7) we find that

$$\mathbb{E}\left(\sum_{i=0}^{r} (y_{t+a-i} \otimes \Pi_{i0}')\,\Sigma_0^{-1/2} e_{0t}\right) = -\frac{1}{2}\sum_{i=0}^{r} A_0(a-i,i)\,\mathrm{vec}(I_n) = 0.$$

This in conjunction with (17) and (B.8) implies that $\mathbb{E}\left(\partial l_T(\theta_0)/\partial \vartheta_2\right) = 0$.

As for the covariance matrix of the score of ϑ_2, let $\mathbf{1}(\cdot)$ stand for the indicator function

and, for $a, b \in \{1, ..., s\}$, consider the covariance matrix

$$\begin{aligned}
& \mathbb{C}\left(\sum_{i=0}^{r}\left(y_{t+a-i}\otimes\Pi_{i0}'\right)\Sigma_0^{-1/2}e_{0t}, \sum_{j=0}^{r}\left(y_{k+b-j}\otimes\Pi_{i0}'\right)\Sigma_0^{-1/2}e_{0k}\right) \\
= & \sum_{c,d=-\infty}^{\infty}\sum_{i,j=0}^{r} A_0(c,i)\,\mathbb{C}\left(\left(\varepsilon_{t+a-i-c}\otimes e_{0t}\right),\left(\varepsilon_{k+b-j-d}\otimes e_{0k}\right)\right)A_0(d,j)' \\
= & \frac{\boldsymbol{\tau}_0}{4}\sum_{\substack{c=-\infty\\ c\neq 0}}^{\infty}\sum_{i,j=0}^{r} A_0(c+a-i,i)\,A_0(c+b-j,j)'\,\mathbf{1}(t=k) \\
& +\frac{1}{4}\sum_{i,j=0}^{r} A_0(t-k+a-i,i)\,K_{nn}A_0(k-t+b-j,j)'\,\mathbf{1}(t\neq k) \\
& +\sum_{i,j=0}^{r} A_0(a-i,i)\,D_nJ_0D_n'A_0(b-j,j)'\,\mathbf{1}(t=k).
\end{aligned}$$

Here the former equality is again obtained by using (5) and the definition of $A_0(k,i)$ whereas the latter is justified by Lemma 4. Summing the last expression over $t, k = r+1, ..., T-s$, multiplying by $4/(T-s-r)$, and letting T tend to infinity yields the matrix $C_{22}(a,b;\theta_0)$ (see (B.8) and the definition of $\mathcal{I}_{\vartheta_2\vartheta_2}(\theta_0)$). Thus,

$$\begin{aligned}
C_{22}(a,b;\theta_0) = & \boldsymbol{\tau}_0\sum_{\substack{k=-\infty\\ k\neq 0}}^{\infty}\sum_{i,j=0}^{r} A_0(k+a-i,i)\,A_0(k+b-j,j)' \\
& +\sum_{\substack{k=-\infty\\ k\neq 0}}^{\infty}\sum_{i,j=0}^{r} A_0(k+a-i,i)\,K_{nn}A_0(-k+b-j,j)' \\
& +4\sum_{i,j=0}^{r} A_0(a-i,i)\,D_nJ_0D_n'A_0(b-j,j)'. \qquad \text{(B.9)}
\end{aligned}$$

To see that the right hand side really equals the expression given in the main text, we have to show that the second term vanishes when the range of summation is changed to $k = 0, \pm 1, \pm 2, ...$, or that

$$\sum_{k=-\infty}^{\infty}\sum_{i,j=0}^{r}\left(\Psi_{k+a-i,0}\Sigma_0^{1/2}\otimes\Pi_{i0}'\Sigma_0^{-1/2}\right)K_{nn}\left(\Sigma_0^{1/2}\Psi_{-k+b-j,0}'\otimes\Sigma_0^{-1/2}\Pi_{j0}\right)=0.$$

One can show this by using the identity $(\Psi_{k+a-i,0}\Sigma_0^{1/2}\otimes\Pi_{i0}'\Sigma_0^{-1/2})K_{nn} = K_{nn}(\Pi_{i0}'\Sigma_0^{-1/2}\otimes\Psi_{k+a-i,0}\Sigma_0^{1/2})$ (see Lütkepohl (1996), Result 9.2.2 (5)(a)), (B.6) and straightforward calculation (further details can be found in the Supplementary Appendix).

From (B.8), the definition of $A_0(c,i)$, and the preceding derivations it follows that the covariance matrix of the score of ϑ_2 divided by $T-s-r$ converges to $\mathcal{I}_{\vartheta_2\vartheta_2}(\theta_0)$.

Step 2. In view of Assumption 5(i) it suffices to prove the positive definiteness of $\mathcal{I}_{\theta\theta}(\theta_0)$ when $\nabla_1(\vartheta_{10}) = I_{rn^2}$ and $\nabla_2(\vartheta_{20}) = I_{sn^2}$. Using the matrices $A_0(k,i)$ and $B_0(d)$ introduced in the preceding step we define the $sn^2 \times n^2$ and $rn^2 \times n^2$ matrices

$$\underline{A}_0(k) = \left[\sum_{i=0}^{r} A_0(k+j-i,i)\right]_{j=1}^{s} \quad \text{and} \quad \underline{B}_0(k) = [B_0(k-i)]_{i=1}^{r},$$

We also set

$$F_0 = \frac{\pi^{n/2}}{\Gamma(n/2)} \int_0^\infty \zeta^{n/2} \frac{f'(\zeta;\lambda_0)}{f(\zeta;\lambda_0)} \frac{\partial}{\partial\lambda} f(\zeta;\lambda_0)\, d\zeta \cdot \operatorname{vech}(I_n)' J_0^{-1} \quad \left(\mathrm{d} \times \tfrac{1}{2}n(n+1)\right).$$

Let $\eta_t = (\eta_{1t}, \eta_{2t}, \eta_{3t}, \eta_{4t})$ be a sequence of independent and identically distributed random vectors with zero mean. The covariance matrix of η_t as well as the dimensions of its components will be specified shortly. We consider the linear process

$$x_t = \sum_{k=1}^{\infty} \underline{G}_0(k)\, \eta_t,$$

where $x_t = (x_{1t}, x_{2t}, x_{3t}, x_{4t})$ and

$$\underline{G}_0(k) = \begin{bmatrix} -\underline{B}_0(k) & 0 & 0 & 0 \\ \underline{A}_0(k) & \underline{A}_0(-k) & 2\mathbf{1}(k=1)\,\underline{A}_0(k-1)\,D_n & 0 \\ 0 & 0 & -\mathbf{1}(k=1)\,D_n'(\Sigma_0^{-1/2} \otimes \Sigma_0^{-1/2})D_n & 0 \\ 0 & 0 & \mathbf{1}(k=1)\,F_0 & \mathbf{1}(k=1)\,I_{\mathrm{d}} \end{bmatrix}$$

With an appropriate definition of the covariance matrix of η_t we have $\mathbb{C}(x_t) = \mathcal{I}_{\theta\theta}(\theta_0)$. This is achieved by assuming

$$\mathbb{C}(\eta_t) = \operatorname{diag}\left(\begin{bmatrix} \boldsymbol{\tau}_0 I_{n^2} & K_{nn} \\ K_{nn}' & \boldsymbol{\tau}_0 I_{n^2} \end{bmatrix} \ : \ J_0 \ : \ \mathcal{I}_{\lambda\lambda}(\theta_0) - F_0 J_0 F_0' \right),$$

where the first block defines the covariance matrix of (η_{1t}, η_{2t}). Thus, (η_{1t}, η_{2t}), η_{3t}, and η_{4t} are uncorrelated and the dimensions of x_{it} agree with those of η_{it} $(i = 1, ..., 4)$. By straightforward calculations one can check that the equality $\mathbb{C}(x_t) = \mathcal{I}_{\theta\theta}(\theta_0)$ really holds (with $\nabla_1(\vartheta_{10}) = I_{rn^2}$ and $\nabla_2(\vartheta_{20}) = I_{sn^2}$).

From Lemma 1 and the fact that K_{nn} is a permutation matrix it follows that the first block of $\mathbb{C}(\eta_t)$ is positive definite. Indeed, this is implied by the positive definiteness of $\boldsymbol{\tau}_0 I_{n^2} - \boldsymbol{\tau}_0^{-1} K_{nn}' K_{nn} = \boldsymbol{\tau}_0 I_{n^2} - \boldsymbol{\tau}_0^{-1} I_{n^2}$, which holds because $\boldsymbol{\tau}_0 > 1$. That J_0 is positive definite follows from Lemma 4 whereas the positive definiteness of the third block of $\mathbb{C}(\eta_t)$ is due to Assumption 5(ii) and the identity $\mathcal{I}_{\lambda\lambda}(\theta_0) - F_0 J_0 F_0' = \mathcal{I}_{\lambda\lambda}(\theta_0) - \mathcal{I}_{\lambda\sigma}(\theta_0)\mathcal{I}_{\sigma\sigma}(\theta_0)^{-1}\mathcal{I}_{\sigma\lambda}(\theta_0)$, which can be checked by direct calculation. Thus, the covariance matrix $\mathbb{C}(\eta_t)$ is positive definite.

The preceding discussion implies that the matrix $\mathcal{I}_{\theta\theta}(\theta_0)$ is positive definite if the covariance matrix $\mathbb{C}(x_t)$ is positive definite. This, in turn, holds if the infinite dimensional matrix $[\underline{G}_0(1) : \underline{G}_0(2) : \cdots]$ is of full row rank. Proving this last fact is somewhat tedious, so we omit details which are available in the Supplementary Appendix.

Step 3. The asymptotic normality can be proved in the same way as in previous univariate models (see Proposition 2 of Breidt et al. (1991)). The idea is to use (3) and (5) to approximate the processes $u_{t-i}(\vartheta_{10})$ and y_{t+j-i} $(i = 1, ..., r,\ j = 1, ..., s)$ in $\partial g_t(\theta_0)/\partial\vartheta_1$ and $\partial g_t(\theta_0)/\partial\vartheta_1$, respectively, by long moving averages. After this, a central limit theorem for finitely dependent stationary and ergodic processes in conjunction with a standard approximation technique completes the proof. □

Proof of Lemma 2. The arguments used in the proof are analogous to those used in the proof of Proposition 2. A detailed proof is available in the Supplementary Appendix. □

Proof of Theorem 1. First note that our Proposition 2 and Lemma 2 are analogous to Lemmas 1 and 2 of Andrews et al. (2006). Thus, as in the proof of Theorem 1 of that paper we can use a standard Taylor expansion and conclude that it suffices to show that the Hessian of the log-likelihood function satisfies

$$\sup_{\theta\in\Theta_0}\left\| N^{-1}\sum_{t=r+1}^{T-s}\left(\frac{\partial^2}{\partial\theta\partial\theta'}g_t(\theta) - \frac{\partial^2}{\partial\theta\partial\theta'}g_t(\theta_0)\right)\right\| \xrightarrow{p} 0, \tag{B.10}$$

where Θ_0 is a small compact neighborhood of θ_0 with non-empty interior (cf. Lanne and Saikkonen (2011)). It can readily be checked that $\partial^2 g_t(\theta)/\partial\theta\partial\theta'$ is stationary and ergodic, and, as a function of θ, continuous. Hence, a sufficient condition for (B.10) to hold is that $\partial^2 g_t(\theta)/\partial\theta\partial\theta'$ obeys a uniform law of large numbers over Θ_0. This in turn is implied by

$$\mathbb{E}_{\theta_0}\left(\sup_{\theta\in\Theta_0}\left\|\frac{\partial^2}{\partial\theta\partial\theta'}g_t(\theta)\right\|\right) < \infty \tag{B.11}$$

(see Theorem A.2.2 in White (1994)). Proving (B.11) is straightforward and, therefore, omitted (details can be found in the Supplementary Appendix). □

References

[1] Alessi, L., M. Barigozzi, and M. Capasso (2008). Non-Fundamentalness in Structural Econometric Models: A Review. International Statistical Review 79, 16–47.

[2] Andrews, B. R.A. Davis, and F.J. Breidt (2006). Maximum Likelihood Estimation for All-Pass Time Series Models. Journal of Multivariate Analysis 97, 1638-1659.

[3] Breidt, J., R.A. Davis, K.S. Lii, and M. Rosenblatt (1991). Maximum Likelihood Estimation for Noncausal Autoregressive Processes. Journal of Multivariate Analysis 36, 175–198.

[4] Breidt, J., R.A. Davis, and A.A. Trindade (2001). Least Absolute Deviation Estimation for All-Pass Time Series Models. The Annals of Statistics 29, 919-946.

[5] Brockwell , P.J. and R.A. Davis (1987). Time Series: Theory and Methods. Springer-Verlag. New York.

[6] Campbell, J.Y., and R.J. Shiller (1991). Yield Spreads and Interest Rate Movements: A Bird's Eye View. Review of Economic Studies 58, 495–514.

[7] Chan, K.S. and L. Ho (2004). On the Unique Representation of Non-Gaussian Multivariate Linear Processes. Technical Report #341, University of Iowa. http://www.stat.uiowa.edu/techrep/

[8] Chan, K.S., L. Ho, and H. Tong (2006). A Note on Time-irreversibility of Multivariate Linear Processes. Biometrika 93, 221–227.

[9] Davis, R.A., and L. Song (2010). Noncausal Vector AR Processes with Application to Financial Time Series. Technical Report. Columbia University.

[10] Duffee, G. (2002). Term Premia and Interest Rate Forecasts in Affine Models. Journal of Finance 57, 405–443.

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

[12] Francq, C., R. Roy, and J.-M. Zakoïan (2005). Diagnostic checking in ARMA models with uncorrelated errors. Journal of the American Statistical Association 100, 532–544.

[13] Hannan, E.J. (1970). Multiple Time Series. John Wiley and Sons. New York.

[14] Johansen. S. and K. Juselius (2010). An Invariance Property of the Common Trends under Linear Transformations of the Data. CREATES Research Papers 2010-72, School of Economics and Management, University of Aarhus.

[15] Kohn, R. Asymptotic Estimation and Hypothesis Testing Results for Vector Linear Time Series Models. Econometrica, 47, 1005–1029.

[16] Lanne, M., J. Luoto, and P. Saikkonen (2012). Optimal Forecasting of Noncausal Autoregressive Time Series. International Journal of Forecasting (forthcoming).

[17] Lanne, M., and P. Saikkonen (2011). Noncausal Autoregressions for Economic Time Series. Journal of Time Series Econometrics 3 (3), Article 3.

[18] Lof, M. (2012). Noncausality and Asset Pricing. Studies in Nonlinear Dynamics and Econometrics (forthcoming).

[19] Lütkepohl, H. (1996). Handbook of Matrices. John Wiley & Sons,New York.

[20] Lütkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin.

[21] Rosenblatt, M. (2000). Gaussian and Non-Gaussian Linear Time Series and Random Fields. Springer-Verlag, New York.

[22] Rothenberg, T. J. (1971). Identification in Parametric Models. Econometrica 39, 577–591.

[23] Sargent, T.J. (1979). A Note on Maximum Likelihood Estimation of the Rational Expectations Model of the Term Structure. Journal of Monetary Economics 5, 133–143.

[24] White, H. (1994). Estimation, Inference and Specification Analysis. Cambridge University Press. New York.

[25] Wong, C.H. and T. Wang (1992). Moments for Elliptically Countered Random Matrices. Sankhyā 54, 265–277.

Figure 1: The quarterly change in the six-month U.S. interest rate (solid line) and the spread between the five-year and six-month U.S. interest rates (dasdhes).

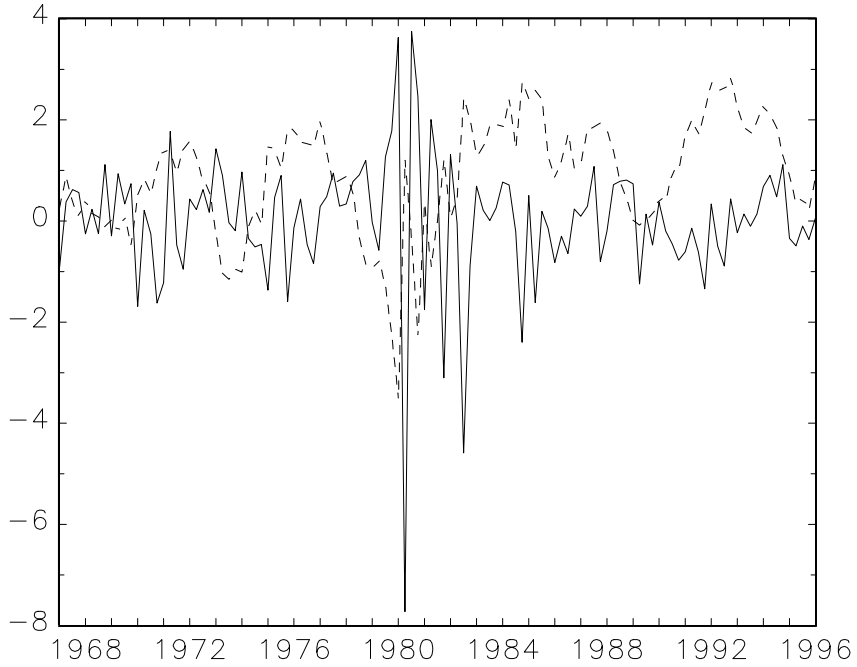

Figure 2: Quantile-quantile plots of the residuals of the VAR(3,0)-N (upper panel) and VAR(2,1)-t (lower panel) models for the U.S. term structure data.

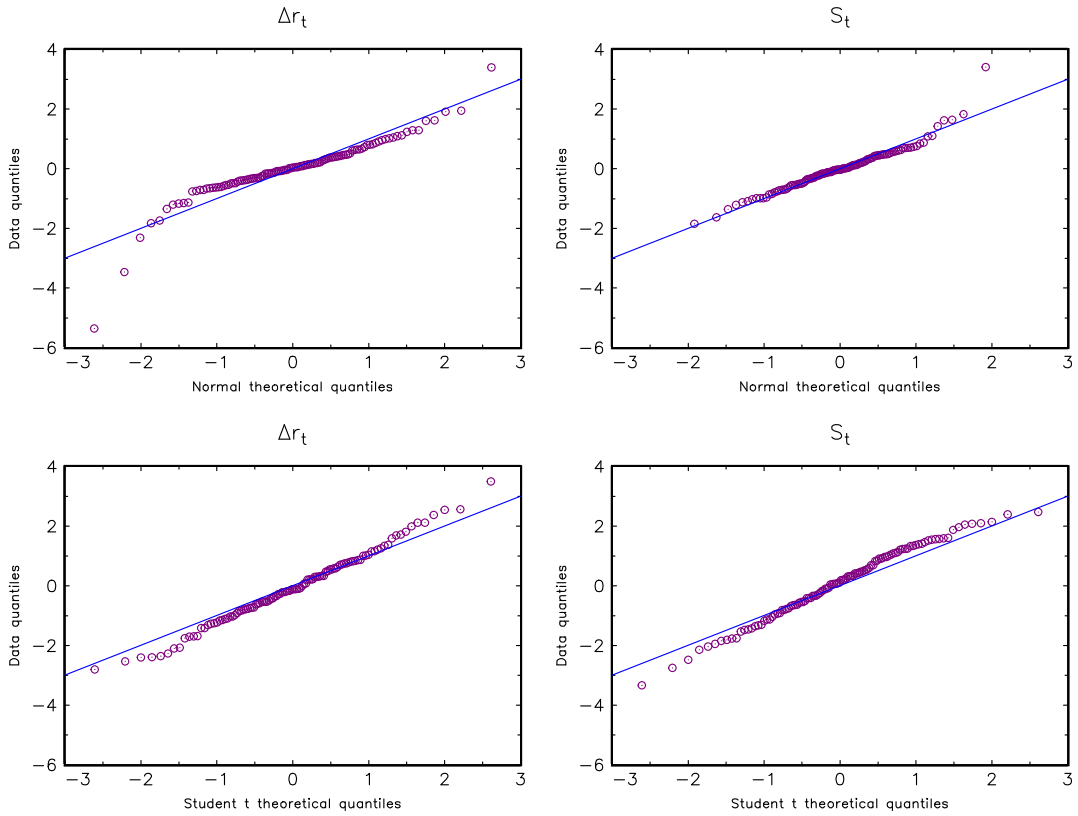

Table 1: Results of diagnostic checks of the third-order VAR models for the term structure.

	Model				
	VAR(3,0)-N	VAR(3,0)-t	VAR(2,1)-t	VAR(1,2)-t	VAR(0,3)-t
Ljung-Box (4)	0.172	0.014	0.094	9.4e–5	0.003
	0.118	0.069	0.063	3.2e–5	0.027
McLeod-Li (4)	0.4.2e–4	0.023	0.896	5.2e–5	0.101
	0.002	0.183	0.930	0.018	0.003
Log-likelihood	–258.510	–229.985	–222.953	-227.454	–231.252

VAR(r, s) denotes the vector autoregressive model for $(\Delta r_t, S_t)'$ with the rth and sth order polynomials $\Pi(B)$ and $\Phi(B^{-1})$, respectively. N and t refer to Gaussian and t-distributed errors, respectively. Marginal significance levels of the Ljung-Box and McLeod-Li tests with 4 lags are reported for each equation.

Table 2: Estimation results of the VAR(2,1)-t model for $(\Delta r_t, S_t)'$.

Π_1	–0.458	0.782	Π_2	–0.241	0.298
	(0.156)	(0.189)		(0.090)	(0.184)
	0.138	0.075		0.320	–0.006
	(0.143)	(0.183)		(0.097)	(0.164)
Φ_1	0.399	–0.210			
	(0.126)	(0.067)			
	–0.240	0.673			
	(0.260)	(0.144)			
Σ	0.296	–0.167			
	(0.096)	(0.106)			
	–0.167	0.312			
	(0.106)	(0.189)			
λ	4.085				
	(1.210)				

The figures in parentheses are standard errors based on the Hessian of the log-likelihood function.