Testing for Noncausal Vector Autoregressive Representation

Mehdi Hamidi Sahneh

University of Carlos III

5 August 2013

Testing for Noncausal Vector Autoregressive Representation

Mehdi Hamidi Sahneh *

Abstract

We propose a test for noncausal vector autoregressive representation generated by non-Gaussian shocks. We prove that in these models the Wold innovations are martingale difference if and only if the model is correctly specified. We propose a test based on a generalized spectral density to check for martingale difference property of the Wold innovations. Our approach does not require to identify and estimate the noncausal models. No specific estimation method is required, and the test has the appealing nuisance parameter free property. The test statistic uses all lags in the sample and it has a convenient asymptotic standard normal distribution under the null hypothesis. A Monte Carlo study is conducted to examine the finite-sample performance of our test.

Keywords: Explosive Bubble; Identification; Noncausal Process; Vector Autoregressive.

JEL classification: C5, C32, E62.

*Departamento de Economia, Universidad Carlos III de Madrid, Getafe, 28903, Spain. Email address: *mhamidis@eco.uc3m.es*. The author is deeply indebted to Carlos Velasco for guidance and encouragement. I also benefited from Miguel Delgado, Juan Dolado, Jesus Gonzalo, Hernan Seoane, Fabio Canova, Pentti Saikkonen, Markku Lanne, and seminar participants at The 68th European Meeting of the Econometric Society (ESEM2014), The 8th Nordic Econometric Meeting (NEM2015), for comments and discussions. The research was supported by the Spanish Plan Nacional de I+D+I (ECO2012-31748) and (ECO2014-57007).

1 Introduction

Vector Autoregressions (VAR) have been used extensively by economists and statisticians for economic analysis and to obtain forecasts. If the model is misspecified, though, interesting dynamics of the time series process can be ignored and conclusions from the model might be misleading. Since estimation methods based on second-order moment techniques do not identify noncausal processes, most economic applications restrict themselves to causal autoregressive models. Indeed if noncausality is incorrectly ignored, the estimates may yield suboptimal forecasts and misleading economic interpretations. In this paper we propose a test for noncausal VAR models generated by non-Gaussian shocks.

Causality is the standard assumption in the analysis of time series, because without this assumption the model is unidentified using econometrics methods based on second-order moments. However, in the non-Gaussian case, causal and noncausal representations are distinguishable on the basis of higher order cumulants; see, e.g. Rosenblatt (2000). Despite the significant implications for empirical work, little is known about how to empirically detect noncausality. The only proposal that we are aware of is Breidt et al. (1991), which is based on maximizing the likelihood function. Specifically, all combinations of causal and noncausal models of a given order are estimated, and the model yielding the greatest value of the likelihood function is selected. However, this method crucially relies on the choice of non-Gaussian distribution. If the non-Gaussian distribution is misspecified, the correct noncausal model might not be among these representations. Even if the noncausality is correctly identified, this procedure may pick the wrong specification because of the misspecification of the non-Gaussian distribution.

We prove that the Wold innovations from fitting a noncausal VAR are not martingale difference (MD), if the true errors are non-Gaussian. Using our theoretical results, we are able to propose a test for noncausal VAR, which follows the tradi-

tional modeling strategy of imposing causality. Therefore, this approach does not require to estimate noncausal models. Under the null hypothesis, Wold innovations are martingale difference and standard inference applies. Under the alternative hypothesis we face the situation where the econometrician fits a wrong model, and the Wold innovations are not martingale difference.

Portmanteau test proposed by Box and Pierce (1970) and Ljung and Box (1978) are not able to capture the nonlinear dependence structure. There are many proposals to test for the martingale difference property, which to the best of our knowledge, none of them are applicable to the multivariate setting of this paper. To test for the MD property of the Wold innovations, we extend Hong and Lee's (2005) test from univariate to multivariate setting. The proposed test statistic has a convenient asymptotic standard normal distribution under the null hypothesis. No specific estimation method is required, and the test has the appealing nuisance parameter free property. Moreover, our test only require as inputs estimated model residuals, obtained from any $\sqrt{T}$-consistent parameter estimates.

The rest of the paper is organized as follows: Section 2 provides a formal statement of the characterization of noncausal VAR representations and the testing problem. Section 3 introduces formally the test statistic based on the generalized spectral density and Section 4 investigates its asymptotic properties. Section 5 examines the finite-sample performance of the test through some Monte Carlo simulation experiments and an empirical application. Section 6 concludes. An Appendix contains the proofs.

2 Characterization of noncausal VAR representations

Let $\{x_t\}$ be a d-dimensional stationary solution of the VAR model, satisfying the difference equation:

$$\Phi(L)x_t = \xi_t \ , \qquad t = 0, \pm 1, \pm 2, \cdots \tag{1}$$

where $\{\xi_t\}$ are independent non-Gaussian process, and $\Phi(L) := I_d - \Phi_1 L - \cdots - \Phi_p L^p$ is the autoregressive polynomial. Henceforth, I_d is the $d \times d$ identity matrix, $\Phi_p \neq 0$ and L is the lag operator, i.e., $Lx_t = x_{t-1}$. We can factor the autoregressive polynomial as

$$\Phi(z) = \Phi^{\dagger}(z)\Phi^{*}(z)$$

where

$$\Phi^{\dagger}(z) = \prod_{1 \leq i \leq r} (1 - b_i^{-1} z), \quad |b_i| > 1$$
$$\Phi^{*}(z) = \prod_{r < i \leq p} (1 - b_i^{-1} z), \quad |b_i| < 1$$

and where $\Phi^{*}(z) = 1$ if $r = p$.

A VAR process defined by (1) is said to be causal if and only if all the roots of $\Phi(z)$ lie outside the unit circle in the complex plane (i.e. $r = p$). If some of the roots of $\Phi(z)$ lie inside the unit circle, then we say the VAR model is noncausal (see Brockwell and Davis, 1991, ch 3). We use the abbreviation VAR(r,s), where $s = p - r$, for the noncausal VAR model specified by (1), where r is the number of roots outside the unit disk and s is the number of roots inside the unit disk. In the causal case, i.e. $s = 0$, we use the conventional VAR(p) abbreviation.

Despite the evidence pointing out to noncausal representations in econometrics and statistics models, little is known about how to empirically detect noncausality. The only proposal in the literature that we are aware of is that of Breidt et al. (1991) and Lanne and Saikkonen (2011). These authors propose to fit a conventional causal VAR model by least squares or Gaussian ML, using conventional model selection criteria to specify the lag order p. Assuming a non-Gaussian error distribution, all causal and noncausal models of order p are estimated and of these models the one that maximizes the log-likelihood function is selected. However, if the non-Gaussian distribution is misspecified, this procedure may pick the wrong specification because of the misspecification of the non-Gaussian distribution.

A natural way of testing the specification of a causal VAR(p) model, is to check if the residuals are uncorrelated. In practice, the order p is often selected so that the residuals are white noise. However, one can show that if noncausality is excluded incorrectly, the Wold innovations are still uncorrelated. Therefore, estimation methods based on second-order moment techniques do not identify noncausality.

In the non-Gaussian case, however, causal and noncausal models are distinguishable using higher order cumulants (Lii and Rosenblatt, 1982). Using time-reversibility argument, Breidt and Davis (1992) proved that the Wold innovations from fitting a causal model to a noncausal one are *iid*, if and only if the error is non-Gaussian. Unfortunately, this result does not extend to the multivariate case (Chan et al., 2006). Moreover, testing for serial dependence of the Wold innovations is restrictive and may lead to rejection of the null of causality by mistake. To see this, consider the case where the true unobserved errors are martingale difference process, for example GARCH. If the model is causal, then Wold innovations have the same structure as the true unobserved errors. Therefore, if we test for serial dependence, we reject the null of causality, although the model is causal.

In this paper, I use the information structure available in the *Blaschke* matrix

to propose a new test to empirically detect noncausality[1]. A standard result for ARMA processes is that any VAR(r,s) process $\{x_t\}$ which is noncausal with respect to the noise sequence $\{\xi_t\}$ can also be modeled as a causal VAR(r,s) with respect to a new noise sequence $\{\epsilon_t\}$. One can show that the true unobserved shocks, $\{\xi_t\}$, will be related to the Wold innovations, $\{\epsilon_t\}$, through *Blaschke* matrices. Under some mild conditions stated in Assumption 1, I prove that if the model is noncausal, the Wold innovations are not MD, i.e., they are non-linearly predictable, despite being white noise.

Assumption 1. ξ_t is an independent process that is continuously distributed with a non-Gaussian distribution such that $(a+1)th$ moment finite for some $a \geq 2$ and $\text{Var}(\xi_t) > 0$.

Proposition 2.1: Let Assumption 1 hold. The non-Gaussian VAR model (1) is causal if and only if the Wold innovations $\{\epsilon_t\}$ are MD.

For the proof see appendix. Assumption 1. is commonly used in the empirical studies. It can be further relaxed to allow for the true unobserved shocks to be dependent. The proof holds under sub-independence assumption[2]. This is a generalization of the concept of independence of random variables, i.e., if two random variables are independent then they are sub-independent, but not conversely, see Hamedani (2013). Unfortunately, the connection between sub-independence and MD is not clear in the literature, and we do not attempt to justify it here.

Non-Gausianity is needed to achieve identification. In fact, there are many studies that emphasize considering non-Gaussian distributions and other higher order time-varying moments (see e.g., Harvey and Siddique, 1999, 2000; Jondeau and Rockinger, 2003). Note that, what is needed is the existence of some moments

[1] *Blaschke* matrices are complex-valued filters which take the roots from inside to outside the unit disc (Lippi and Reichlin, 1994).

[2] Two random variables are said to be sub-independent if the characteristic function of their sum is equal to the product of their marginal characteristic functions, i.e., $\phi_{x+y}(t) = \phi_x(t)\phi_y(t)$.

higher than the third for at least one of the shocks, and no specific distributional assumption is needed. The continuity assumption is also mild and could be dropped in the univariate case or if there is only one root of the $\det \Theta(L)$ that is inside the unit circle. This is stated in the following corollary.

3 Testing for noncausal representations

Under the null of causality $\xi_t(\theta_0) = \epsilon_t(\theta_0)$, which following Proposition 2.1 can be restated as

$$\mathbb{H}_0 : \epsilon_t(\theta_0) \text{ is MD for some } \theta_0 \in \Xi \tag{2}$$

where $\theta_0 = vec\{\Phi_1, \cdots, \Phi_p, \Theta_1, \cdots, \Theta_q, \Sigma_\epsilon\}$, and $vec(.)$ denote an operator on a matrix which cascades the columns of the matrix from the left to the right and forms a column vector.

Testing (2) is not an easy task. There are many proposals to test for the martingale difference property see Hong (1999), Domínguez and Lobato (2003), Hong and Lee (2005), among others. To the best of our knowledge, none of these tests are applicable to the multivariate setting of this paper. Alternatively, it is possible to apply a sequence of univariate test to each series. However, using a multivariate procedure will avoid the multiple testing problem and is more powerful, since it is possible that a single series is not MD, but the collection of several series is MD. Moreover, $\{\epsilon_t\}$ is unobserved and residuals depend on a $\sqrt{T}$-consistent estimator for θ_0, which may cause the loss of the nuisance parameter-free property of the asymptotic distribution of the test statistics.

To overcome these problems and checking for non-linear predictability at all lags in the sample, I extend the generalized spectral test of Hong and Lee (2005) to the multivariate setting. Compared with the existing tests in the literature, this test has some advantages: first, with the frequency domain approach, one can allow

infinite number of lags as the sample size increases; second, the test has a standard normal limiting distribution and parameter estimation uncertainty has no impact on the asymptotic distribution of the test statistics.[3] The proposed test can also be used to test the martingale hypothesis in the multivariate setting for observed raw data without any modification.

My proposal for testing the MD property of the Wold innovations is based upon the generalized spectrum of Hong (1999):

$$f(\omega, u, v) \equiv \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_j(u, v) e^{-ij\omega}, \tag{3}$$

where $\omega \in [-\pi, \pi]$ is the frequency, $i \equiv \sqrt{-1}$, $(u, v) \in \mathbb{R}^d \times \mathbb{R}^d$, and

$$\sigma_j(u, v) = \text{cov}(e^{iu'\epsilon_t}, e^{iv'\epsilon_{t-|j|}}), \quad j = 0, \pm 1, ...$$

where $\epsilon_t \equiv \epsilon_t(\theta)$. Note that $f(\omega, u, v)$ is a complex-valued scalar function, although ϵ_t is a $d \times 1$ vector. The function $f(\omega, u, v)$ captures any type of pairwise serial dependence in $\{\epsilon_t\}$, including that with zero autocorrelation function.

The generalized spectrum $f(\omega, u, v)$ is not suitable for testing (2), because it also captures the serial dependence in higher order moments. For example $f(\omega, u, v)$ captures GARCH dependence, although the process could be a MDS. However, just as the characteristic function can be differentiated to generate various moments of ϵ_t, $f(\omega, u, v)$ can be differentiated to capture the serial dependence in various moments. To capture (and only capture) the serial dependence in the conditional mean, one can use

$$f^{(0,1,0)}(\omega, u, v) \equiv \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_j^{(1,0)}(0, v) e^{-ij\omega}, \quad \omega \in [-\pi, \pi]$$

where

$$\sigma_j^{(1,0)}(0, v) \equiv \frac{\partial}{\partial u} \sigma_j(u, v)\big|_{u=0} = \text{cov}(i\epsilon_t, e^{iv'\epsilon_{t-|j|}})$$

[3] Tests based on bootstrap procedures which take into account the impact of parameter estimation uncertainty may also be considered (see e.g., Gonçalves and Kilian, 2004).

is a $d\times 1$ vector. The measure $\sigma_j^{(1,0)}(0,v)$ checks whether the autoregression function $E(\epsilon_t|\epsilon_{t-j})=0$ at lag j is zero.[4]

In the present context, ϵ_t is not observed. Suppose we have T observations $\{x_t\}_{t=1}^T$ which is used to estimate the model and to obtain the estimated model residual

$$\hat{\epsilon}_t \equiv \hat{\Theta}^{-1}(L)\hat{\Phi}(L)x_t \tag{4}$$

where $\hat{\theta}$ is a $\sqrt{T}$-consistent estimator for θ_0. Examples of $\hat{\theta}$ are conditional least squares and quasi-maximum likelihood estimator. We can estimate $f^{(0,1,0)}(\omega,0,v)$ by a smoothed kernel estimator

$$\hat{f}^{(0,1,0)}(\omega,0,v) \equiv \frac{1}{2\pi}\sum_{j=T-1}^{T-1}(1-\frac{|j|}{T})^{1/2}k(j/h)\hat{\sigma}_j^{(1,0)}(0,v)e^{-ij\omega}, \quad \omega\in[-\pi,\pi] \tag{5}$$

where $\hat{\sigma}_j^{(1,0)}(0,v)=\frac{\partial}{\partial u}\hat{\sigma}_j(u,v)\big|_{u=0}$, $\hat{\sigma}_j(u,v)=\hat{\varphi}_j(u,v)-\hat{\varphi}_j(u,0)\hat{\varphi}_j(0,v)$, and

$$\hat{\varphi}_j(u,v)=\frac{1}{T-|j|}\sum_{t=j+1}^{T}e^{iu'\hat{\epsilon}_t+iv'\hat{\epsilon}_{t-|j|}}$$

where $h\equiv h(T)$ is a bandwidth, and $k:\mathbb{R}\to[-1,1]$ is a symmetric kernel. Examples of $k(\cdot)$ include the Bartlett, Daniell, Parzen and Quadratic spectral kernels. The factor $(1-\frac{|j|}{T})^{1/2}$ is a finite-sample correction. The effect of this correction factor is to put less weight on very large lags, for which we have less sample information. It could be replaced by unity.

Under $\mathbb{H}_0$, the generalized spectral derivative $f^{(0,1,0)}(\omega,0,v)$ becomes a flat spectrum:

$$f_0^{(0,1,0)}(\omega,0,v)\equiv\frac{1}{2\pi}\sigma_0^{(1,0)}(0,v), \quad \omega\in[-\pi,\pi]$$

[4]The hypothesis of $E(\epsilon_t|I_{t-j}^{\epsilon})=0$ *a.s.* is not the same as the hypothesis of $E(\epsilon_t|\epsilon_{t-j})=0$ *a.s.* for all $j>0$. The former checks all type of dependencies, whereas the latter one only captures pairwise dependencies. See Hong (1999) for more discussion on this.

which can be consistently estimated by

$$\hat{f}_0^{(0,1,0)}(\omega,0,v) \equiv \frac{1}{2\pi}\hat{\sigma}_0^{(1,0)}(0,v), \quad \omega \in [-\pi,\pi]$$

The estimators $\hat{f}^{(0,1,0)}(\omega,0,v)$ and $\hat{f}_0^{(0,1,0)}(\omega,0,v)$ converge to the same limit under $\mathbb{H}_0$, and generally converge to different limits under $\mathbb{H}_1$. Thus, any significant divergence between them can be interpret as evidence of the violation of the MDS property, and hence, of the non-fundamentalness of the process.

Our test statistic, which is the multivariate version of $\hat{M}$ of Hong and Lee (2005), is given as follows:

$$\hat{M} \equiv \left[\sum_{j=1}^{T-1} k^2(j/h)T_j \int \left\|\hat{\sigma}_j^{(1,0)}(0,v)\right\|^2 d\mathcal{W}(v) - \hat{C}\right] \Big/ \sqrt{\hat{D}} \tag{6}$$

where $T_j = T - j$, $\mathcal{W}(v) = \prod_{c=1}^{d} W(v_c)$, $W : \mathbb{R} \to \mathbb{R}^+$ is a nondecreasing function that weighs sets symmetric about zero equally, and the unspecified integrals are taken over the support of $\mathcal{W}(\cdot)$. Examples of $W(\cdot)$ include the CDF of any symmetric probability distribution, either discrete or continuous. $\hat{C}$ and $\hat{D}$ are estimate of the mean and the variance of $T\int\int_{-\pi}^{\pi} \|\hat{f}^{(0,1,0)}(\omega,0,v) - \hat{f}_0^{(0,1,0)}(\omega,0,v)\|^2 d\omega d\mathcal{W}(v)$,

$$\hat{C}(p) \equiv \sum_{j=1}^{T-1} k^2(j/p)\frac{1}{T-j}\sum_{t=j+1}^{T-1} \|\hat{\epsilon}_t\|^2 \int \left|\hat{\psi}_{t-j}(v)\right|^2 dW(v)$$

$$\hat{D}(p) = 2\sum_{j=1}^{T-2}\sum_{l=1}^{T-2} k^2(j/p)k^2(l/p)\sum_{a=1}^{d}\sum_{b=a}^{d}\int\int \left|\frac{1}{T-\max(j,l)} \times \sum_{t=\max(j,l)+1}^{T} \hat{\epsilon}_{at}\hat{\epsilon}'_{bt}\hat{\psi}_{t-j}(v)\hat{\psi}^*_{t-l}(v')\right|^2 dW(v)dW(v')$$

where $\hat{\psi}_t(v) = e^{iv'\hat{\epsilon}_t} - T^{-1}\sum_{t=1}^{T} e^{iv'\hat{\epsilon}_t}$.

To derive the limit distribution of the test, I need to impose some regularity

conditions. Throughout, I use C to denote a generic bounded constant, $\|.\|$ the Euclidean norm, and A^* the complex conjugate of A.

Assumption A1. $\{x_t\}$ is a $d \times 1$ strictly stationary time series process, and ϵ_t are MDS with $\mathrm{E}\|\epsilon_t^4\| \leq C$, where ϵ_t is Wold innovation from estimating an invertible model.

Assumption A2. For q sufficiently large, there exists a strictly stationary process $\{\epsilon_{q,t}\}$ measurable with respect to the sigma field generated by $\{\epsilon_{t-1}, \epsilon_{t-2}, \cdots, \epsilon_{t-q}\}$ s.t. as $q \to \infty$, $\epsilon_{q,t}$ is independent of $\{\epsilon_{t-q-1}, \epsilon_{t-q-2}, \cdots\}$ for each t, $\mathrm{E}[\epsilon_{q,t}|I_{t-1}] = 0$ a.s., $\mathrm{E}\|\epsilon_t - \epsilon_{q,t}\|^2 \leq Cq^{-\kappa}$ for some constant $\kappa \geq 1$, and $\mathrm{E}\|\epsilon_{q,t}\|^4 \leq C$ for all large q.

Assumption A3. The estimator $\hat{\theta}$ is such that $\sqrt{T}(\hat{\theta} - \theta^*) = O_P(1)$, where $\theta^* \equiv plim_{T\to\infty}\hat{\theta}$. Under $\mathbb{H}_0$, $\theta^* = \theta_0$.

Assumption A4. Let $\bar{x}_0 = (x_0; \cdots; x_{1-p}; \epsilon_0; \cdots; \epsilon_{1-q})$ be some assumed initial values. Then $\mathrm{E}\|\bar{x}_0^2\| < \infty$.

Assumption A5. $k : \mathbb{R} \to [-1, 1]$ is symmetric about 0, and is continuous at 0 and all points except a finite number of points, with $k(0) = 1$ and $|k(z)| \leq C|z|^{-b}$ as $z \to \infty$ for some $b > 1$.

Assumption A6. $W : \mathbb{R} \to \mathbb{R}^+$ is nondecreasing and weights sets symmetric about zero equally, with $\int \|v\|^4 dW(v) \leq C$.

Assumption A7. Define $\psi_t(v) \equiv e^{iv\epsilon_t} - T^{-1}\sum_{t=1}^{T} e^{iv\epsilon_t}$ and $\Sigma \equiv \mathrm{E}(\epsilon_t\epsilon_t')$. Then, $\{\frac{\partial\epsilon_t}{\partial\theta}, \epsilon_t\}$ is a strictly stationary process such that

(a) $\sum_{j=1}^{\infty} \|\mathrm{cov}[\frac{\partial\epsilon_t}{\partial\theta}, \psi_{t-j}(v)]\| \leq C$;

(b) $\sum_{j=1}^{\infty} \sup_{(u,v)\in\mathbb{R}^2} |\sigma_j(u, v)| \leq C$;

(c) $\sum_{j=1}^{\infty}\sum_{l=1}^{\infty} \sup_{(u,v)\in\mathbb{R}^2} \left\|\mathrm{E}[(\epsilon_t\epsilon_t' - \Sigma)\psi_{t-j}(u)\psi_{t-l}(v)]\right\| \leq C$;

(d) $\sum_{j=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\sum_{\tau=-\infty}^{\infty}\sup_{v\in\mathbb{R}}\|\kappa_{j,l,\tau}(v)\|\leq C$, where $\kappa_{j,l,\tau}(v)$ is the fourth order cumulant of the joint distribution of the process $\{\frac{\partial\epsilon_t}{\partial\theta},\psi_{t-j}(v),\frac{\partial\epsilon_{t-l}}{\partial\theta},\psi^*_{t-\tau}(v)\}$.

Assumption A8. $\sum_{j=1}^{\infty}\sup_{v\in\mathbb{R}}\|\sigma_j^{(1,0)}(0,v)\|\leq C$.

Assumption A1 is a regularity condition on the data generating process (DGP) $\{x_t\}$. Assumption A2 is required only under $\mathbb{H}_0$, which states that the MDS $\{\epsilon_t\}$ can be approximated by a q-dependent MDS process $\{\epsilon_t\}$ arbitrarily well when q is sufficiently large. Because $\{\epsilon_t\}$ is a MDS, Assumption A2 essentially imposes restrictions on the serial dependence in higher order moments of $\{\epsilon_t\}$. It covers GARCH and stochastic volatility processes as special cases; see *e.g.* Hong and Lee (2005). Assumption A3 requires a $\sqrt{T}$-consistent estimator $\hat{\theta}$, which may not be asymptotically most efficient. It can be a conditional least squares estimator or a conditional quasi-maximum likelihood estimator.

Assumption A4 is a start-up value condition. It ensures that the impact of initial values assumed in the observed information set is asymptotically negligible. Assumption A5 is a regularity condition on the kernel $k(.)$. It includes all commonly used kernels in practice. For kernels with bounded support, such as the Bartlett and Parzen kernels, we have $b=\infty$: For kernels with unbounded support, b is some finite positive real number. Assumption A6 is a condition on the weighting function $W(.)$ for the transform parameter v. It is satisfied by the CDF of any symmetric continuous distribution with a finite fourth moment. Assumption A7 provides some covariance and fourth order cumulant conditions on $\{\frac{\partial\epsilon_{t-1}}{\partial\theta},\epsilon_t\}$, which restricts the degree of serial dependence in $\{\frac{\partial\epsilon_{t-1}}{\partial\theta},\epsilon_t\}$. Finally, Assumption A8 impose a condition on the serial dependence in $\{\epsilon_t\}$. The asymptotic properties of the test statistic is stated in the following theorem. The proof is similar to the univariate case of Hong and Lee (2005), and for the sake of space is given in the online Appendix.

Proposition 4.1: Let $h=cT^{\lambda}$ for $0<\lambda<(3+\frac{1}{4b-2})^{-1}$ and $0<c<\infty$. Then:

(a) Under Assumptions A1-A7 and $\mathbb{H}_0$, $\hat{M} \xrightarrow{d} N(0,1)$.

(b) Under Assumptions A1-A8 and $\mathbb{H}_1$, $\lim_{T\to\infty} P[\hat{M} > C(T)] = 1$ for any sequence $C(T) = o(T/h^{1/2})$.

Under the null, $\hat{M}$ has a simple standard normal distribution. Under the alternative hypothesis, $\mathrm{E}(\epsilon_t|\epsilon_{t-j}) \neq 0$ a.s., at some lag $j > 0$. Then we have $\int \|\sigma_j^{(1,0)}(0,v)\|^2 d\mathcal{W}(v) > 0$ for any weighting function $\mathcal{W}(\cdot)$ that is positive, monotonically increasing and continuous, with unbounded support on $\mathbb{R}$. Therefore, $\hat{M}$ has asymptotic unit power at any given significance level.

An important feature of $\hat{M}$ is that the use of the estimated residuals $\{\hat{\epsilon}_t\}$ in place of the true errors $\{\epsilon_t\}$ has no impact on the limit distribution of $\hat{M}$. The reason is that the convergence rate of the parametric parameter estimator $\hat{\theta}$ to θ_0 is faster than that of the nonparametric kernel estimator $\hat{f}^{(0,1,0)}(w,0,v)$ to $f^{(0,1,0)}(w,0,v)$. Consequently, the limit distribution of $\hat{M}$ is solely determined by $\hat{f}^{(0,1,0)}(w,0,v)$, and replacing θ_0 by $\hat{\theta}$ has no impact asymptotically.

4 Monte Carlo evidence and empirical application

4.1 Simulation study

In order to assess the finite sample performance of our proposed test, we conduct a Monte Carlo study. To investigate the empirical size and power of $\hat{M}$, we consider AR (or VAR) processes with *iid* centralized log-normal errors as follows:

1. (DGP1): Univariate, causal AR(1) process

$$y_t = 0.5y_{t-1} + \xi_t, \quad \xi_t \sim lognorm(0,1)$$

Table 1: Empirical size of the test: univariate case (DGP1)

	$T = 100$			$T = 250$			$T = 400$		
$\bar{h}$	10%	5%	1%	10%	5%	1%	10%	5%	1%
A: Bartlett									
5	6.2	4.4	2.3	7.3	5.3	1.8	7.7	4.9	1.7
10	7.5	5.2	2.5	7.8	5.8	2.2	7.9	4.7	1.9
15	8.6	5.8	2.9	8.4	5.3	2.5	8.2	4.8	1.9
B: Parzen									
5	5.5	3.8	2.0	6.0	4.2	1.2	6.5	4.4	1.4
10	5.7	4.4	2.1	6.4	5.2	1.6	7.2	4.6	1.7
15	6.3	4.3	2.3	7.9	5.1	1.6	8.0	4.7	1.6

Notes: (1) $\bar{h}$ is the preliminary lag order used in a plug-in method to select a data-driven lag order $\hat{h}_0$; (2) The number of replication is 1000.

2. (DGP2): Univariate, noncausal AR(1) process

$$y_t = 0.5y_{t+1} + \xi_t, \quad \xi_t \sim lognorm(0,1)$$

3. (DGP3): Bivariate, causal VAR(1) process

$$\begin{bmatrix} x_{t,1} \\ x_{t,2} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.1 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} x_{t-1,1} \\ x_{t-1,2} \end{bmatrix} + \begin{bmatrix} \xi_{t,1} \\ \xi_{t,2} \end{bmatrix}$$

4. (DGP4): Bivariate, noncausal VAR(1) process

$$\begin{bmatrix} x_{t,1} \\ x_{t,2} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.1 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} x_{t+1,1} \\ x_{t+1,2} \end{bmatrix} + \begin{bmatrix} \xi_{t,1} \\ \xi_{t,2} \end{bmatrix}$$

Some comments are in order. First, $\hat{M}$ involves $d-$ and $2d-$ dimensional numerical integration, which can be computationally cumbersome when d is large. In practice, one may approximate the integrals by choosing a finite number of grid points symmetric about zero or generate a finite number of points drawn from the

Table 2: Empirical power of the test: univariate case (DGP2)

	$T = 100$			$T = 250$			$T = 400$		
$\bar{h}$	10%	5%	1%	10%	5%	1%	10%	5%	1%
A: Bartlett									
5	68.8	61.2	47.2	93.9	89.9	82.1	98.8	98.1	95.6
10	63.2	56.9	42.9	91.0	86.6	78.0	98.5	97.2	94.1
15	58.7	52.2	37.5	88.5	83.3	72.7	97.8	96.1	91.0
B: Parzen									
5	68.8	62.3	47.1	94.0	90.4	83.2	98.9	98.1	95.9
10	67.2	59.1	46.2	93.3	89.1	80.8	98.4	97.6	95.0
15	64.9	57.0	44.1	91.9	88.1	79.1	97.8	96.9	94.5

Notes: (1) $\bar{h}$ is the preliminary lag order used in a plug-in method to select a data-driven lag order $\hat{h}_0$; (2) The number of replication is 1000.

Table 3: Empirical size of the test: bivariate case (DGP3)

	$T = 100$			$T = 250$			$T = 400$		
$\bar{h}$	10%	5%	1%	10%	5%	1%	10%	5%	1%
A: Bartlett									
5	2.4	0.8	0.0	2.8	1.0	0.2	5.8	3.6	1.2
10	2.2	0.8	0.0	2.4	1.0	0.4	6.2	3.4	1.0
15	2.0	1.0	0.2	2.8	0.8	0.4	5.6	3.0	1.0
B: Parzen									
5	2.8	1.2	0.4	3.2	1.6	0.2	6.4	4.0	1.0
10	2.6	1.4	0.4	2.8	1.8	0.2	6.0	3.6	1.0
15	2.6	1.2	0.2	2.8	1.2	0.4	5.4	3.2	0.8

Notes: (1) $\bar{h}$ is the preliminary lag order used in a plug-in method to select a data-driven lag order $\hat{h}_0$; (2) The number of replication is 500.

Table 4: Empirical power of the test: bivariate case (DGP4)

	$T = 100$			$T = 250$			$T = 400$		
$\bar{h}$	10%	5%	1%	10%	5%	1%	10%	5%	1%
A: Bartlett									
5	48.4	28.6	10.4	94.0	87.0	65.4	99.2	98.4	96.0
10	44.0	24.8	8.6	91.0	82.2	56.8	99.0	98.4	95.8
15	39.4	21.2	7.6	87.6	76.8	51.8	99.0	97.8	92.2
B: Parzen									
5	50.2	27.8	12.8	96.8	88.2	64.8	100.0	99.0	95.8
10	48.4	26.2	11.8	92.2	86.8	58.6	99.8	98.2	95.4
15	47.6	24.4	11.0	88.0	84.4	55.8	99.2	98.0	94.6

Notes: (1) $\bar{h}$ is the preliminary lag order used in a plug-in method to select a data-driven lag order $\hat{h}_0$; (2) The number of replication is 500.

uniform distribution on $[-1, 1]^d$. Alternatively, for some weighting functions there is a closed form expression for the test statistics. In this paper, we use a closed form solution obtained by choosing $d\mathcal{W}(\cdot)$ as the $d-$dimensional Gaussian CDF.

Second, a practical issue in implementing the test is the choice of the bandwidth parameter $\hat{h}$. Following Hong and Lee (2005), one can choose a data-driven bandwidth $\hat{h} = \hat{c}_0 T^{\frac{1}{2q+1}}$ via the plug-in method, which lets data themselves determine an appropriate lag.[5] The data-driven bandwidth $\hat{c}_0$, involves the choice of a preliminary bandwidth $\bar{h}$, which can be fixed or grow with the sample size T. Applying the data-driven method to choose the bandwidth, while considering a wide range of the bandwidth, $\bar{h} \in \{4, \cdots, 16\}$, the simulation results show that the test is not sensitive to the choice of preliminary bandwidth. For the sake of space, we only report the results for $\bar{h} = 5, 10$ and 15, using the Bartlett and Parzen kernels. Simulations suggest that the choice of $k(\cdot)$ has little impact on both the level and the power of the test.

Table 1 reports the empirical rejections probabilities of $\hat{M}$ under DGP1 at the 10%, 5% and 1% levels for the sample size $T = 100$; 250 and 400. Overall, the

[5] q is called the characteristic exponent of $k(.)$. For Bartlett kernel, $q = 1$; for Daniell, Parzen, QS, and Tukey kernels, $q = 2$.

size of the test under the null of causality is appropriate and is robust to the choice of kernel and preliminary bandwidth $\bar{h}$. Table 3 reports the empirical power of $\hat{M}$ against the noncausal univariate AR process. Overall,$\hat{M}$ is powerful against DGP3. The power is robust to the choice of kernel and bandwidth parameter $\bar{h}$.

Appendix A

I first prove Lemma 1, which is an extension of Theorem 5.4.1 Rosenblatt (2000), by dropping the identically distribution assumption. In Lemma 2, I use Lemma 1 to prove the univariate case of Proposition 2.1, and then show that under Assumption 1 the multivariate case can be reduced to the univariate case. **Lemma 1:** Consider a univariate causal and non-invertible VARMA(p,q) model, that is, $r_\Phi = r_p$ and $r_\Theta < r_q$. Let $\phi^t(\tau)$ denote the characteristic function of ξ_t and $\phi^t_{\tau_0}(\cdot) = \frac{\partial \phi^t(\cdot)}{\partial \tau_0}$. Then linearity of the best predictor in mean square implies that

$$\sum_{k=-\infty}^{\infty} \big(\gamma_k - \sum_{l=1}^{\infty} \beta_l \gamma_{k-l}\big) h^{t-k}\big(\sum_{l=1}^{\infty} \tau_l \gamma_{k-l}\big) = 0 \tag{A.7}$$

where $h^t(\vartheta) = \frac{\phi^t_{\tau_0}(\vartheta)}{\phi^t(\vartheta)}$ and β_l's are the coefficients of the best linear predictor of x_t in mean square

$$x_t^* = \sum_{l=1}^{\infty} \beta_l x_{t-l}$$

Proof of Lemma 1: Writing (1) in the MA form we have:

$$x_t = \sum_{k=-\infty}^{\infty} \gamma_k \xi_{t-k}, \quad \gamma_k = 0 \quad \forall k < 0 \tag{A.8}$$

The joint characteristic function of $\{x_{t-j}, j \geq 0\}$ is given by

$$\begin{aligned} \eta^t(\tau_0, \tau_1, \cdots, \tau_p, \cdots) &= E\Big\{exp\big(i \sum_{l=0}^{\infty} \tau_l x_{t-l}\big)\Big\} \\ &= \prod_{k=-\infty}^{\infty} \phi^{t-k}\big(\sum_{l=0}^{\infty} \tau_l \gamma_{t-l}\big) \end{aligned} \tag{A.9}$$

while the joint characteristic function of $\{x_{t-j}, j \geq 1\}$ is

$$\tilde{\eta}^t(\tau_1, \cdots, \tau_p, \cdots) = \prod_{k=-\infty}^{\infty} \phi^{t-k}\big(\sum_{l=1}^{\infty} \tau_l \gamma_{t-l}\big) \tag{A.10}$$

Differentiating $\eta^t(\tau_0, \tau_1, \cdots, \tau_p, \cdots)$ w.r.t. τ_0 we have

$$\begin{aligned}
\frac{\partial}{\partial \tau_0} \eta^t(\tau_0, \tau_1, \cdots, \tau_p, \cdots)|_{\tau_0=0} &= \eta^t_{\tau_0}(0, \tau_1, \cdots, \tau_p, \cdots) \\
&= \int i x_t \; exp(i \sum_{l=1}^{\infty} \tau_l x_{t-l}) \; dF^t(x_t, x_{t-1}, \cdots, x_{t-p}, \cdots) \\
&= i \int E[x_t | x_{t-s}, s > 0] \; exp(i \sum_{l=1}^{\infty} \tau_l x_{t-l}) \; dF^t(x_{t-1}, \cdots, x_{t-p}, \cdots)
\end{aligned} \tag{A.11}$$

where $F^t(x_t, x_{t-1}, \cdots, x_{t-p}, \cdots)$ is the joint cumulative distribution function of $x_{t-j}, j \geq 0$. Also by differentiating the logarithm of (A.9) w.r.t. τ_0 we get:

$$\frac{\eta^t_{\tau_0}(0, \tau_1, \cdots, \tau_p, \cdots)}{\eta^t(0, \tau_1, \cdots, \tau_p, \cdots)} = \sum_{k=-\infty}^{\infty} \gamma_k h^{t-k} (\sum_{l=1}^{\infty} \tau_l \gamma_{k-l}). \tag{A.12}$$

Similarly, differentiating the logarithm of $\tilde{\eta}^t(\tau_1, \cdots, \tau_p, \cdots)$ w.r.t. τ_j, $j = 1, 2, \cdots$, we have

$$\frac{\partial}{\partial \tau_j} log \; \tilde{\eta}^t(\tau_1, \cdots, \tau_p, \cdots) = \sum_{k=-\infty}^{\infty} \gamma_{k-j} h^{t-k} (\sum_{l=1}^{\infty} \tau_l \gamma_{k-l}), \;\; j = 1, 2, \cdots \tag{A.13}$$

If the best predictor in mean square is linear we must have

$$\eta^t_{\tau_0}(0, \tau_1, \cdots) = \sum_{k=1}^{\infty} \beta_k \tilde{\eta}^t_{\tau_k}(\tau_1, \tau_2, \cdots) \tag{A.14}$$

which implies

$$\sum_{k=-\infty}^{\infty} \big(\gamma_k - \sum_{l=1}^{\infty} \beta_l \gamma_{k-l}\big) h^{t-k} (\sum_{l=1}^{\infty} \tau_l \gamma_{k-l}) = 0. \tag{A.15}$$

□

Lemma 2: Let Assumption 1 hold. The univariate non-Gaussian AR model (1) is

causal if and only if the Wold innovations $\{\epsilon_t\}$ are MDS.

Proof of Lemma 2: A standard result for AR processes is that any AR(p) process $\{x_t\}$ which is non-causal with respect to the noise sequence $\{\xi_t\}$ can also be modeled as a causal AR(p) with respect to a new noise sequence $\{\epsilon_t\}$ defined by[6]

$$\epsilon_t = \frac{\prod_{r<i\leq q}(1-b_iL)}{\prod_{r<i\leq q}(1-b_i^{-1}L)}\xi_t, \qquad |b_i|<1. \tag{A.16}$$

which can be written as:

$$\sum_{i=0}^{q-r}\alpha_i\epsilon_{t-i} = e_t \tag{A.17}$$

where $e_t = \sum_{i=0}^{q-r}\beta_i\xi_{t-i}$. Then (A.17) Lemma 1 and Corollary 5.4.2 of Rosenblatt (2000) implies that the best one-step predictor of ϵ_t is non-linear, i.e., $E[\epsilon_t|\epsilon_{t-s}, s\geq 1]$ is non-linear. If ϵ_t were a MD, i.e. $E[\epsilon_t|\epsilon_{t-s}, s\geq 1]=0$, Lemma 1 implies that:

$$\sum_{k=-\infty}^{\infty}\gamma_k h^{t-k}(\sum_{l=1}^{\infty}\tau_l\gamma_{k-l}) = 0 \tag{A.18}$$

Since $\mu_{a+1}\neq 0$, we have

$$\sum_{k=-\infty}^{\infty}\gamma_k\gamma_{k-l_1}\cdots\gamma_{k-l_a} = 0, \;\; l_1,\cdots,l_a = 1,2,\cdots. \tag{A.19}$$

For the ath order partial derivative of the expression (A.18) w.r.t $\tau_{l_1},\cdots,\tau_{l_a}$ at $\tau_{l_1}=\cdots=\tau_{l_a}=0$, $i^{a+1}\mu_{a+1}a!$ is multiplied by the expression (A.19) on the left. Since

$$(1-bz)(1-b^{-1}z)^{-1} = b^2+(b^2-1)\sum_{j=1}^{\infty}b^jz^{-j}$$

we have $\gamma_k=0$ for $k>0$. Therefore (A.19) is equal to

$$\sum_{k=0}^{\infty}\gamma_{-k}\gamma_{-k-l_1}\cdots\gamma_{-k-l_a} = 0, \;\; l_1,\cdots,l_a = 1,2,\cdots. \tag{A.20}$$

[6]See Brockwell and Davis (1991), page 103.

Also

$$\gamma_{-k} = \sum_{j=r+1}^{p} \alpha_j b_j^k, \;\; k > 0$$

for some coefficients $\alpha_j \neq 0, \; j = r+1, \cdots, p$. Therefore, equations (A.20) can be written as

$$\sum_{j_1, \cdots, j_a = r+1}^{p} \alpha_{j_1} \cdots \alpha_{j_a} b_{j_1}^{l_1} \cdots b_{j_a}^{l_a} \sum_{k=0}^{\infty} \gamma_{-k} (b_{j_1} \cdots b_{j_a})^k = 0$$

$l_1, \cdots, l_a = 1, \cdots, p$. Consider the set of equations obtained by letting $l_1, \cdots, l_a = 1, \cdots, s$. The matrix of this set of equations is

$$M = (M_{j,l}) = \{\alpha_{j_1} \cdots \alpha_{j_a} b_{j_1}^{l_1} \cdots b_{j_a}^{l_a}\}$$

where $j = (j_1, \cdots, j_a), \; l = (l_1, \cdots, l_a), \; j_1, \cdots, j_a = r+1, \cdots, p, \; l_1, \cdots, l_a = 1, \cdots, s$. The determinant of this matrix is $(\prod_{u=r+1}^{p} \alpha_u)^{2a}$ multiplied by the $2a$-th power of the Vandermonde determinant

$$|b_j^l; j = r+1, \cdots, q, \; l = 1, \cdots, s|$$

Since the determinant is nonzero, we must have

$$\gamma(b_{j_1}, \cdots, b_{j_a}) = \sum_{k=0}^{\infty} \gamma_k (b_{j_1}, \cdots, b_{j_a})^k$$

This implies $(b_{j_1} \cdots b_{j_a})$, for $j_1, \cdots, j_a = r+1, \cdots, p$ are also zeros of $\gamma(z)$, a clear contradiction. Therefore the assumption that $E[\epsilon_t | \epsilon_{t-s}, s > 0] = 0$ cannot hold. ■

Proof of Proposition 2.1: The proof is similar to the Corollary 2.1 in Hamidi Sahneh (June, 2015).□

References

Box, G. E. and Pierce, D. A., 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526.

Breidt, F. J., Davis, R. A., Lh, K.-S., and Rosenblatt, M., 1991. Maximum likelihood estimation for noncausal autoregressive processes. *Journal of Multivariate Analysis*, 36(2):175–198.

Breidt, F. and Davis, R., 1992. Time-reversibility, identifiability and independence of innovations for stationary time series. *Journal of Time Series Analysis*, 13(5): 377–390.

Brockwell, P. and Davis, R., 1991. *Time Series: Theory and Methods: Theory and Methods*. Springer series in statistics. Springer.

Chan, K.-S., Ho, L.-H., and Tong, H., 2006. A note on time-reversibility of multivariate linear processes. *Biometrika*, 93(1):221–227.

Domínguez, M. A. and Lobato, I. N., 2003. Testing the martingale difference hypothesis. *Econometric Reviews*, 22(4):351–377.

Gonçalves, S. and Kilian, L., 2004. Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, 123(1):89–120.

Hamedani, G., 2013. Sub-independence: An expository perspective. *Communications in Statistics-Theory and Methods*, 42(20):3615–3638.

Hamidi Sahneh, M. Are the shocks obtained from svar fundamental? Technical report, University Library of Munich, Germany, June, 2015.

Harvey, C. R. and Siddique, A., 1999. Autoregressive conditional skewness. *Journal of financial and quantitative analysis*, 34(04):465–487.

Harvey, C. R. and Siddique, A., 2000. Conditional skewness in asset pricing tests. *The Journal of Finance*, 55(3):1263–1295.

Hong, Y., 1999. Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. *Journal of the American Statistical Association*, 94(448):1201–1220.

Hong, Y. and Lee, Y.-J., 2005. Generalized spectral tests for conditional mean models in time series with conditional heteroscedasticity of unknown form. *The Review of Economic Studies*, 72(2):499–541.

Jondeau, E. and Rockinger, M., 2003. Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. *Journal of Economic Dynamics and Control*, 27(10):1699–1737.

Lanne, M. and Saikkonen, P., 2011. Noncausal autoregressions for economic time series. *Journal of Time Series Econometrics*, 3(3).

Lii, K. and Rosenblatt, M., 1982. Deconvolution and estimation of transfer function phase and coefficients for nongaussian linear processes. *The Annals of Statistics*, pages 1195–1208.

Lippi, M. and Reichlin, L., 1994. Var analysis, nonfundamental representations, blaschke matrices. *Journal of Econometrics*, 63(1):307–325.

Ljung, G. M. and Box, G. E., 1978. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Rosenblatt, M., 2000. *Gaussian and non-Gaussian linear time series and random fields*. Springer.