A Test for Independence between Two Infinite Order Autoregressive Processes

Eunhee Kim [1] and Sangyeol Lee[2]

Abstract

This paper considers the independence test for two stationary infinite order autoregressive processes. For a test, we follow the empirical process method devised by Hoeffding (1948) and Blum, Kiefer and Rosenblatt (1961), and construct the Cramér-von Mises type test statistics based on the least squares residuals. It is shown that the proposed test statistics behave asymptotically the same as those based on true errors.

Key words: Independence test, infinite order autoregressive processes, the Cramér-von Mises test, residual empirical process, weak convergence.

1. Introduction

In this paper, we consider the problem of testing the independence of two stationary time series. For the past two decades, the issue has drawn much attention from many researchers. For instance, Haugh (1976) proposed an independence test for the errors in ARMA models based on the sum of squares of residual cross correlations. Later, adopting his idea, Pierce (1977), Geweke (1981) and Hong (1996) studied the independence test for two stationary time series. In fact, the method using the cross correlations has been much popular in the time series context since it is a crucial task to figure out the dependence structure of given time series in a correct manner, and any model selection procedures require a step for diagnostics to set up a true model. However, the cross correlation method merely guarantees the uncorrelatedness of observations and does not ensure the independence. Moreover, the cross correlation just checks the linear relationship but cannot find a nonlinear dependence. Therefore, instead of it, here we employ the empirical process method devised by Hoeffding (1948) and Blum, Kiefer and Rosenblatt (1961). Their test statistics essentially fall into the category of Cramér-von Mises (CV) statistics, and have been applied under a variety of circumstances. Recently, their metohd has been adopted by Skaug and Tjøstheim (1993), Delgado (1996), Hong (1998) and Delgado and Mora (2000) aimed at developing a serial independence test.

In this paper, we focus on the independence test for two stationary infinite order autoregressive processes. We adopted autoregressive processes since they include the most popular

[1] Ph.D. Student, Department of Statistics, Seoul National University.
[2] Associate Professor, Department of Statistics, Seoul National University.

ARMA processes in time series analysis, and a method based on residuals usually discards correlation effects. For a test, we construct the CV statistic based on residuals. It will be seen that the limiting distribution of the residual based CV statistic is the same as the CV statistic based on the true errors.

In fact, the CV statistic is designed for testing the independence for a specific lag k. However, in real situations one should check the independence for several lags, say, $|k| \leq K$, where K is a positive integer larger than 1, since the test for only one lag is not sufficient to ensure the independence between the two times series. To this end, we consider three types of test statistics. First, we consider the summation of the CV statistics based on residuals, which Skaug and Tjøstheim (1993) suggested as a test statistic for testing serial independence. Second, we consider the weighted summation of CV statistics which was proposed in Hong (1998) for testing serial independence, since the summation type test statistic may suffer from severe size distortions as K increases. Finally, we propose as a test statistic the maximum of the CV test statistics. We consider this because it is less affected by CV statistics with large values and will have more stability compared to other tests. Our simulation study shows that the method based on CV statistic with residuals turns out to be suitable for the independence test of two stationary time series.

2. Main results

Suppose that $\{X_t\}$ and $\{Y_t\}$ satisfy the following difference equations:

$$X_t - \mu - \sum_{j=1}^{\infty} \phi_j (X_{t-j} - \mu) = \epsilon_t, \quad t = 1, \cdots, n$$

and

$$Y_t - \nu - \sum_{j=1}^{\infty} \theta_j (Y_{t-j} - \nu) = \eta_t, \quad t = 1, \cdots, n,$$

where (ϵ_t, η_t) are random vectors with common distribution F, ϵ_t and η_t are iid r.v.'s with marginal distributions F_1 and F_2, respectively, and $E\epsilon_t^4 + E\eta_t^4 < \infty$. Furthermore, both $A_1(z) := 1 - \sum_{j=1}^{\infty} \phi_j z^j$ and $A_2(z) := 1 - \sum_{j=1}^{\infty} \theta_j z^j$ are assumed to be analytic on an open neighborhood of the closed unit disk D in the complex plane and have no zeroes on D. It can be easily seen that the last condition implies

$$| \phi_j | + | \theta_j | \leq C\rho^j, \quad C > 0, \ 0 < \rho < 1,$$

(cf. Lee and Wei, 1999). It is well-known that the AR(∞) process covers a broad class of stationary processes including invertible ARMA processes (cf. Brockwell and Davis, 1990).

Suppose that one wishes to test the hypotheses

$$H_0 : \ \{X_t\} \text{ and } \{Y_t\} \text{ are independent. vs. } H_1 : \text{ not } H_0.$$

The above is equivalent to testing

$$H_0^{'}: \ \{\epsilon_t\} \text{ and } \{\eta_t\} \text{ are independent. vs. } H_1^{'}: \text{ not } H_0^{'}$$

since the independence of the processes themselves implies that of the error processes and the converse is also true. For testing $H_0^{'}$ vs. $H_1^{'}$, we consider employing the CV test statistic (cf. Hoeffding, 1948),

$$B_{nk} = \begin{cases} (n-|k|)^{-1}\sum_{i=|k|+1}^{n} S_{nk}^2(\epsilon_{i-k}, \eta_i), & k \geq 0, \\ (n-|k|)^{-1}\sum_{i=|k|+1}^{n} S_{nk}^2(\epsilon_i, \eta_{i-|k|}), & k < 0, \end{cases} \tag{2.1}$$

with

$$S_{nk}(x,y) = \begin{cases} (n-|k|)^{-1}\sum_{t=|k|+1}^{n} I(\epsilon_{t-k} \leq \ x)I(\eta_t \leq \ y) \\ \quad -(n-|k|)^{-2}\sum_{t=|k|+1}^{n} I(\epsilon_{t-k} \leq x)\sum_{t=|k|+1}^{n} I(\eta_t \leq y), & k \geq 0 \\ (n-|k|)^{-1}\sum_{t=|k|+1}^{n} I(\epsilon_t \leq \ x)I(\eta_{t-|k|} \leq \ y) \\ \quad -(n-|k|)^{-2}\sum_{t=|k|+1}^{n} I(\epsilon_t \leq x)\sum_{t=|k|+1}^{n} I(\eta_{t-|k|} \leq y), & k < 0. \end{cases} \tag{2.2}$$

Note that since verifying the same lag independence itself is not enough to ensure the independence between two time series, we consider the test statistic based on the empirical distribution of (ϵ_{t-k}, η_t). Here, utilizing the similar method of Skaug and Tjøstheim (1993) and the result of Carlstein (1988), one can show that under H_0 for each k,

$$(n-k)B_{nk} \xrightarrow{d} \mathcal{W}_k := \sum_{i,j=1}^{\infty} \lambda_{ij} W_{ijk}^2, \tag{2.3}$$

where $W_{ijk}, i,j = 1,2,\ldots$ are iid $N(0,1)$ r.v.'s and λ_{ij} are the numbers in Theorem 2 of Skaug and Tjøstheim (1993). It is well known that $\lambda_{ij} = (ij\pi^2)^{-2}$ for continuous type r.v.'s.

In order to test H_0 vs. H_1, however, true errors should be replaced by residuals since they are unknown in actual practice. For this task, we fit finite order autoregressive models to the observations $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Let $p = p_n$ and $q = q_n$ be certain sequences of positive integers that diverge to ∞ and satisfy $p^3/n \to 0$ and $q^3/n \to 0$ as $n \to \infty$. Writing

$$X_t - \mu = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu) + r_{1t} + \epsilon_t$$

and

$$Y_t - \nu = \sum_{j=1}^{q} \theta_j (Y_{t-j} - \nu) + r_{2t} + \eta_t,$$

where

$$r_{1t} = \sum_{j=p+1}^{\infty} \phi_j (X_{t-j} - \mu) \ \text{ and } \ r_{2t} = \sum_{j=q+1}^{\infty} \theta_j (Y_{t-j} - \nu),$$

we estimate $\boldsymbol{\phi}_n = (\phi_1, \cdots, \phi_p)'$ and $\boldsymbol{\theta}_n = (\theta_1, \cdots, \theta_q)'$ by the least squares estimates $\tilde{\boldsymbol{\phi}}_n = (\tilde{\phi}_1, \cdots, \tilde{\phi}_p)'$ and $\tilde{\boldsymbol{\theta}}_n = (\tilde{\theta}_1, \cdots, \tilde{\theta}_q)'$, i.e.,

$$\tilde{\boldsymbol{\phi}}_n = \{ \sum_{t=p+1}^{n} (\mathbf{X}_{t-1} - \hat{\mu}_n \mathbf{1}_p)(\mathbf{X}_{t-1} - \hat{\mu}_n \mathbf{1}_p)' \}^{-1} \sum_{t=p+1}^{n} (\mathbf{X}_{t-1} - \hat{\mu}_n \mathbf{1}_p)(X_t - \hat{\mu}_n) \tag{2.4}$$

and

$$\tilde{\boldsymbol{\theta}}_n = \{ \sum_{t=q+1}^{n} (\mathbf{Y}_{t-1} - \hat{\nu}_n \mathbf{1}_q)(\mathbf{Y}_{t-1} - \hat{\nu}_n \mathbf{1}_q)' \}^{-1} \sum_{t=q+1}^{n} (\mathbf{Y}_{t-1} - \hat{\nu}_n \mathbf{1}_q)(Y_t - \hat{\nu}_n), \tag{2.5}$$

where $\mathbf{X}_t = (X_t, \cdots, X_{t-p+1})'$, $\mathbf{Y}_t = (Y_t, \cdots, Y_{t-q+1})'$, $\mathbf{1}_r$, $r \geq 1$, denotes the vector in R^r whose components are all equal to one, and $\hat{\mu}_n$ and $\hat{\nu}_n$ are suitable estimates of μ and ν. Then calculating the residuals

$$\tilde{\epsilon}_t = X_t - \hat{\mu}_n - \tilde{\boldsymbol{\phi}}_n'(\mathbf{X}_{t-1} - \hat{\mu}_n \mathbf{1}_p)$$

and

$$\tilde{\eta}_t = Y_t - \hat{\nu}_n - \tilde{\boldsymbol{\theta}}_n'(\mathbf{Y}_{t-1} - \hat{\nu}_n \mathbf{1}_q),$$

we define

$$\tilde{B}_{nk} = \begin{cases} (n-|k|)^{-1} \sum_{i=|k|+1}^{n} \tilde{S}_{nk}^2(\tilde{\epsilon}_{i-k}, \tilde{\eta}_i), & k \geq 0, \\ (n-|k|)^{-1} \sum_{i=|k|+1}^{n} \tilde{S}_{nk}^2(\tilde{\epsilon}_i, \tilde{\eta}_{i-|k|}), & k < 0, \end{cases}$$

where $\tilde{S}_{nk}(x,y)$ is defined the same as $S_{nk}(x,y)$ in (2.2) with ϵ_t and η_t replaced by $\tilde{\epsilon}_t$ and $\tilde{\eta}_t$. Throughout this paper, we assume that

(C1) $n^{-1}(p^5 + q^5)(\log n)^2 \to 0$ and $n^2(p\rho^p + q\rho^q) \to 0$ for all $\rho \in (0,1)$ as $n \to \infty$;

(C2) $\sup_z \left| \frac{\partial F_i(z)}{\partial z} \right| < \infty$, and $\sup_z \left| \frac{\partial^2 F_i(z)}{\partial z^2} \right| < \infty$, $i = 1, 2$;

(C3) $n^{1/2}(\hat{\mu}_n - \mu) = O_p(1)$ and $n^{1/2}(\hat{\nu}_n - \nu) = O_p(1)$.

The first condition in (C1) implies that the rate of p and q is not so fast; otherwise, we are in a situation that there are too many parameters to be estimated, while the second condition requires those to be large enough for a good approximation. A typical example of p and q is $p = q = [C(\log n)^2]$ for some $C > 0$. Then it can be shown that under H_0, for each nonnegative integer k,

$$\tilde{B}_{nk} - B_{nk} = o_P(n^{-1}), \tag{2.6}$$

proof of which result is omitted.

However, with $\tilde{B}_{nk}$, it is just tested that (ϵ_{t-k}, η_t) are dependent for given k. In order to testing the independence for k's, one should consider the test statistic based on more than one $\tilde{B}_{nk}$'s. If the true errors were known, typically one could consider the summation type test statistic $G_{nK} := n\sum_{k=-K}^{K} B_{nk}$ as Skaug and Tjøstheim (1993) examined this for testing the serial independence of random observation. In fact, similarly to Serfling (1983), pages 194-199, it can be shown that under H_0,

$$G_{nK} \xrightarrow{d} \sum_{i,j=1}^{\infty} \lambda_{ij} C_{ij}(2K+1). \tag{2.7}$$

where $C_{ij}(K), i, j = 1, 2, \dots$ are independent chi-square r.v.'s with K degrees of freedom. Combining (2.7) and (2.6), we obtain the following.

Theorem 2.1 *Assume that (C1)-(C3) hold, and let* $\tilde{G}_{nK} = n\sum_{k=-K}^{K} \tilde{B}_{nk}$, *where* K *is a nonnegative integer. Then under* H_0,

$$\tilde{G}_{nK} \xrightarrow{d} \mathcal{G}_K := \sum_{i,j=1}^{\infty} (ij\pi^2)^{-2} C_{ij}(2K+1) \quad \text{as } n \to \infty,$$

where $C_{ij}(K)$ *are independent chi-square r.v.'s with* K *degrees of freedom.*

Remark. Under the assumption of Theorem 2.1, we have that $\tilde{V}_{nK} := \sum_{k=-K}^{K}(n-k)\tilde{B}_{nk}$ has the same limiting distribution as $\tilde{G}_{nK}$.

Although the idea of using the summation type statistic sounds quite natural, Hong (1998) pointed out that it suffers from severe size distortions as K increases and suggested a weight sum of CV test statistic for testing the serial independence. Similarly, in our set-up, we can also employ weighted sum of $\tilde{B}_{nk}$'s:

$$\tilde{H}_{nK_n} := \{2\tilde{V}_0 \sum_{k=2-n}^{n-2} g^4(k/K_n)\}^{-1/2} \sum_{k=1-n}^{n-1} g^2(k/K_n)\{(n-k)\tilde{B}_{nk} - \tilde{M}_{0n}\},$$

where g is a kernel function, $\{K_n\}$ is a sequence of positive real numbers,

$$\tilde{M}_{0n} = \frac{1}{n}\sum_{t=1}^{n} \tilde{F}_{1n}(\tilde{\epsilon}_t)\{1 - \tilde{F}_{1n}(\tilde{\epsilon}_t)\}\frac{1}{n}\sum_{t=1}^{n} \tilde{F}_{2n}(\tilde{\eta}_t)\{1 - \tilde{F}_{2n}(\tilde{\eta}_s)\}$$

and

$$\tilde{V}_{0n} = \frac{1}{n^2}\sum_{s,t=1}^{n} \{\tilde{F}_{1n}(\tilde{\epsilon}_s \wedge \tilde{\epsilon}_t) - \tilde{F}_{1n}(\tilde{\epsilon}_s)\tilde{F}_{1n}(\tilde{\epsilon}_t)\}^2 \frac{1}{n^2}\sum_{s,t=1}^{n} \{\tilde{F}_{2n}(\tilde{\eta}_s \wedge \tilde{\eta}_t) - \tilde{F}_{2n}(\tilde{\eta}_s)\tilde{F}_{2n}(\tilde{\eta}_t)\}^2$$

with $\tilde{F}_{1n}(u) = \frac{1}{n}\sum_{t=1}^{n} I(\tilde{\epsilon}_t \le u)$ and $\tilde{F}_{2n}(u) = \frac{1}{n}\sum_{t=1}^{n} I(\tilde{\eta}_t \le u)$. Then we have the following result, of which proof is omitted.

Theorem 2.2 *Assume that (C1)-(C3) hold and the function* $g: R \to [-1, 1]$ *is symmetric, continuous at 0 and all except a finite number of points, with* $g(0) = 1$, $\int_{-\infty}^{\infty} g(z)^2 dz < \infty$

and $|g(z)| \leq C|z|^{-b}$ *as* $z \to \infty$ *for some* $b > \frac{1}{2}$ *and* $0 < C < \infty$, *and* $K_n = cn^{\nu}$ *for some* $0 < \nu < 1$ *and* $0 < c < \infty$. *Then under* H_0, *we have*

$$\tilde{H}_{nK_n} \xrightarrow{d} N(0,1). \tag{2.8}$$

Notice that the truncated ($g(z) = I(|z| < 1)$), Bartlett ($g(z) = (1 - |z|)I(|z| < 1)$) and Daniell ($g(z) = sin(\pi z)/(\pi z)$) kernels satisfy the above conditions (cf. Hong (1998) and Priestley (1981)). As it will be seen in our simulation study, $\tilde{H}_{nK_n}$ cures the drawback of $\tilde{G}_{nK}$.

Now, we propose another test statistic, which is obtained as the maximum of $\tilde{B}_{nk}$'s, namely,

$$\tilde{M}_{nK} = n \max_{|k| \leq K} \tilde{B}_{nk}.$$

It will be seen in our simulation study that the maximum type statistic is the most stable among the test statistics considered here. The proof of the following theorem is omitted.

Theorem 2.3 *Assume that (C1)-(C3) hold. Then under* H_0,

$$\tilde{M}_{nK} \xrightarrow{d} \mathcal{M}_K := \max_{|k| \leq K} \mathcal{W}_k \quad \text{as } n \to \infty,$$

where $\mathcal{W}_k$ *is as defined in (2.3) with* $\lambda_{ij} = (ij\pi^2)^{-2}$.

REFERENCES

Billingsley, P. (1968). *Convergence of Probability Measure.* Wiley, New York.

Blum, J. R., Kiefer, J. and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.* **32**, 485-498.

Brockwell, P. J. and Davis, R. A. (1990). *Time series: Theory and Models.* 2nd ed. Springer-Verlag, New York.

Carlstein, E. (1988). Degenerate U-statistic based on non-independent observations. *Calcutta Statist. Assoc. Bull.* **37**, 55-65.

Delgado, M. A. (1996). Testing serial independence using the sample distribution function. *J. Time Ser. Anal.* **17**, 271-286.

Delgado, M. A. and Mora, J. (2000). A nonparametric test for serial independence of regression errors. *Biometrika* **87**, 228-234.

Geweke, J. (1981). A comparison of tests of the independence of two covariance-stationary time series. *J. Amer. Statist. Assoc.* **76**, 363-373.

Haugh, L. D. (1976). Checking the independence of two covariance-stationary time series: A univariate residual cross-correlation approach. *J. Amer. Statist. Assoc.* **71**, 378-385.

Hoeffding, W. (1948). A nonparametric test of independence. *Ann. Math. Statist.* **26**, 189-211.

Hong, Y. (1996). Testing for independence between two covariance stationary time series. *Biometrika* **83**, 615-625.

Hong, Y. (1998). Testing for pairwise serial independence via the empirical distribution function. *J. Roy. Statist. Soc.* **60**, 429-453.

Lee, S. and Wei, C. Z. (1999). On residual empirical processes of stochastic regression models with applications to time sereis. *Ann. Statist.* **27**, 237-261.

Pierce, D. A. (1977). Relationships-and the lack thereof-between economic time series, with special reference to money and interest rates. *J. Amer. Statist. Assoc.* **72**, 11-22.

Priestley, M. B. (1991). *Spectral Analysis and Time Series*, 1, 2. Academic Press, London.

Shorack, G. R. and Wellner, J.A. (1986) *Empirical Processes with Applications to Statistics.* Wiley, New York.

Skaug, H. J. and Tjøstheim, D. (1993). A nonparametic test of serial; independence based on the empirical distribution function, *Biometrika* **80**, 591-602.