Discrimination between deterministic trend and stochastic trend processes

Jorge Caiado and Nuno Crato

2005

Discrimination between deterministic trend and stochastic trend processes

Jorge Caiado[1] and Nuno Crato[2]

[1] Escola Superior de Ciências Empresariais,
Instituto Politécnico de Setúbal and CEMAPRE,
Rua do Quelhas 6, 1200-781 Lisboa, Portugal,
(e-mail: jcaiado@esce.ips.pt)

[2] Instituto Superior de Economia e Gestão,
Universidade Técnica de Lisboa and CEMAPRE,
Rua do Quelhas 6, 1200-781 Lisboa, Portugal,
(e-mail: ncrato@iseg.utl.pt)

Abstract. Most of economic and financial time series have a nonstationary behavior. There are different types of nonstationary processes, such as those with stochastic trend and those with deterministic trend. In practice, it can be quite difficult to distinguish between the two processes. In this paper, we compare random walk and determinist trend processes using sample autocorrelation, sample partial autocorrelation and periodogram based metrics.
Keywords: autocorrelation, classification, determinist trend, Kullback-Leibler, periodogram, stochastic trend, time series.

1 Introduction

There are different types of nonstationarity processes. One can consider a deterministic linear trend process $y_t = a+bt+\varepsilon_t$ (with ε_t a white noise term), that can be transformed into a stationary process by subtracting the trend $a + bt$, and a stochastic linear trend process such as the so-called random walk model $(1 - B)y_t = \varepsilon_t$ or $y_t = y_{t-1} + \varepsilon_t$. An interesting, but some times difficult problem is to determine whether a linear process contains a trend, and whether a linear process exhibits a deterministic or a stochastic trend. In particular, it is useful to distinguish between a random walk plus drift $y_t = \mu + y_{t-1} + \varepsilon_t$ and a deterministic trend in the form $y_t = a + \mu t + \varepsilon_t$.

The problem of classifying and clustering time series has been studied by Piccolo (1990), Tong and Dabas (1990), Shaw and King (1992), Kakizawa, Shumway and Taniguchi (1998), Maharaj (2000, 2002), Caiado, Crato and Peña (2005), Xiong and Yeung (2004), among others. In this paper, we use sample autocorrelation, sample partial autocorrelation and periodogram ordinate based metrics to compare deterministic trend and stochastic trend processes.

2 Classification Methods

A fundamental problem in clustering and classification analysis is the choice of a relevant metric. We know that the Euclidean distance is not a good metric for classifying time series since it is invariant to permutation of the coordinates and so it does not take into account the information about the autocorrelations.

Let $X = (x_{1,t}, \ldots, x_{k,t})'$ be a vector time series and $\widehat{\rho}_i = (\widehat{\rho}_{i,1}, \ldots, \widehat{\rho}_{i,m})$ be a vector of the sample autocorrelations of the time series i for some m such that $\widehat{\rho}_k \cong 0$ for $k > m$. A distance between two time series x and y can be defined by $d(x,y) = \sqrt{(\widehat{\rho}_x - \widehat{\rho}_y)'\Omega(\widehat{\rho}_x - \widehat{\rho}_y)}$, where Ω is some matrix of weights (see Galeano and Peña, 2000) . Caiado, Crato e Peña (2004) proposed three possible ways of computing a distance by using the sample autocorrelation function (ACF). The first uses the Euclidean distance between the sample autocorrelations coefficient vectors with uniform weights (ACFU metric),

$$d_{ACFU}(x,y) = \sqrt{\sum_{j=1}^{L} (\widehat{\rho}_{j,x} - \widehat{\rho}_{j,y})^2}, \quad (1)$$

where L is the number of autocorrelations. The second uses the Euclidean distance with geometric weights decaying with the lag (ACFG metric),

$$d_{ACFG}(x,y) = \sqrt{\sum_{j=1}^{L} f_j (\widehat{\rho}_{j,x} - \widehat{\rho}_{j,y})^2}, \quad (2)$$

where $f_j = pq^j$ for $i = 1, 2, ..., L$, $p = 1 - q$ and $0 < p < 1$. The third uses the Mahalanobis distance between the autocorrelations (ACFM metric),

$$d_{ACFM}(x,y) = \sqrt{(\widehat{\rho}_x - \widehat{\rho}_y)'\Omega^{-1}(\widehat{\rho}_x - \widehat{\rho}_y)}, \quad (3)$$

where Ω is the sample covariance matrix of the autocorrelation coefficients given by Bartlett's formula (see Brockwell and Davis, 1991, p. 221-222). A metric based on the sample partial autocorrelation function (PACF) is defined by

$$d_{PACF}(x,y) = \sqrt{(\widehat{\phi}_x - \widehat{\phi}_y)'\Omega(\widehat{\phi}_x - \widehat{\phi}_y)}, \quad (4)$$

where $\widehat{\phi}_{ii}$ are the sample partial autocorrelations and Ω is also some matrix of weights.

A measure based on the Kullback-Leibler (KL) information for time series classification can be defined by

$$d_{KL}(x,y) = tr(R_x R_y^{-1}) - \log \frac{|R_x|}{|R_y|} - n, \quad (5)$$

where R_x and R_y are the sample autocorrelation matrices of time series x and y. Since $d_{KL}(x, y) \neq d_{KL}(y, x)$, one can define a symmetric distance or quase-distance (KLJ metric), known as the *J divergence*, as,

$$d_{KLJ}(x, y) = \frac{1}{2} d_{KL}(x, y) + \frac{1}{2} d_{KL}(y, x), \tag{6}$$

which satisfies all the usual properties of a metric except the triangle inequality.

Caiado, Crato and Peña (2004) introduced also a periodogram-based metric. Let x and y be observed time series with periodograms, $P_x(w_j) = n^{-1} |\sum_{t=1}^{n} x_t e^{-itw_j}|^2$ and $P_y(w_j) = n^{-1} |\sum_{t=1}^{n} y_t e^{-itw_j}|^2$ at frequencies $w_j = 2\pi_j/n,\ j = 1, ..., m$ (with $m = [(n-1)/2]$) in the range 0 to π, and let $NP(w_j) = P(w_j)/\widehat{\gamma}_0$ be the normalized periodogram (with $\widehat{\gamma}_0$ the sample variance). Since the variance of periodogram ordinates is proportional to the spectrum value at the corresponding frequencies, Caiado, Crato and Peña (2004) proposed a metric based on the logarithm of the normalized periodograms (LNPER metric),

$$d_{LNPER}(x, y) = \sqrt{\sum_{j=1}^{m} \left[\log NP_x(w_j) - \log NP_y(w_j)\right]^2}. \tag{7}$$

3 Monte Carlo Simulations

For the Monte Carlo simulations we chose the determinist trend and random walk plus drift models studied by Enders (1995, p. 252),

$$y_t = 1 + 0.02t + \varepsilon_t$$

and

$$y_t = 0.02 + y_{t-1} + \varepsilon_t/3,$$

with ε_t a zero mean and unit variance white noise. These processes were discussed by Enders since it is quite difficult to distinguish between them, as we can see in Figure 1. We performed 250 replicated simulations of five deterministic trend models and five random walk models with those specifications, with sample sizes of 50, 100, 200, 500 and 1000 observations. We used the previously discussed metrics to compute the distance matrices among the 10 time series and to aggregate them into two clusters (determinist trend and stochastic trend) using an hierarchical clustering algorithm (complete linkage method).

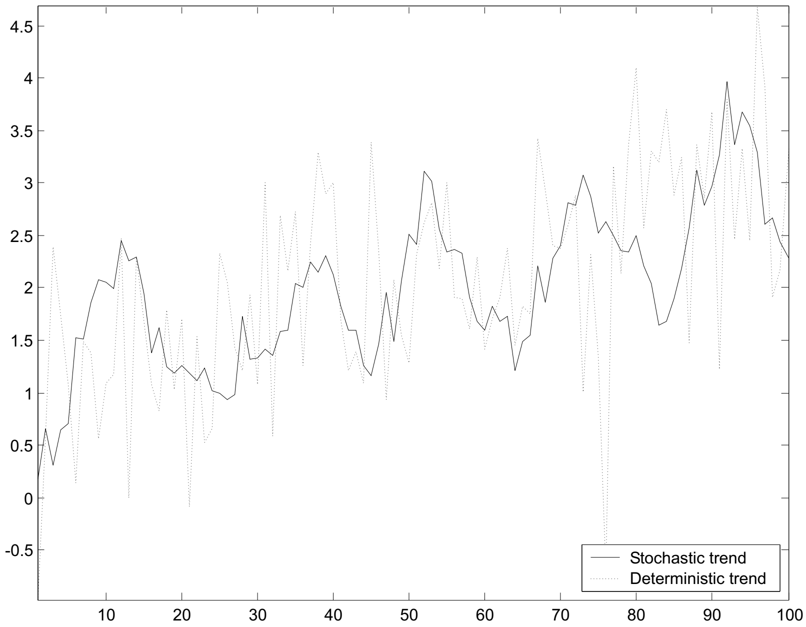

Fig. 1. Simulated stochastic trend and deterministic trend processes.

Table 1 presents the percentage of sucesses obtained in the comparison between the two processes, where n is the sample size, L is the autocorrelation lenght, the sample autocorrelation and sample partial autocorrelation metrics (ACFG and PACFG metrics) uses a geometric decay of $p = 0.05$, in the LNPER metric F for low frequencies corresponds to periodogram ordinates from 1 to $\sqrt{n}$ and F for high frequencies corresponds to periodogram ordinates from $\sqrt{n+1}$ to $n/2$.

The ACF based metrics can discriminate quite well between the deterministic trend models and random walk models. This is particularly evident for the first few autocorrelations, since the ACF of the random walk process is close to unity and the ACF of the deterministic trend tends to approach to zero. Because the PACF of the random walk has a very large first lag and cuts off after lag 1, while the PACF of the deterministic trend exhibits a pattern of a white noise process, the discrimination between the two models based on the first partial autocorrelations is striking. The KLJ metric perform quite well for all data sample sizes and the LNPER metric seems to perform better for periodogram ordinates dominated by high frequencies, which concerns the short-term information of the processes.

n	L	ACFU	ACFG	ACFM	PACFG	KLJ	F	LNPER
50	5	97.28	97.60	99.31	99.27	97.87	low	85.04
	10	92.12	94.88	99.56	99.46	98.53	high	95.24
	25	92.12	91.52	98.01	64.00	97.33	all	94.48
100	5	99.28	98.92	100.0	100.0	98.47		
	10	95.68	97.28	99.73	100.0	98.93	low	92.48
	25	88.16	89.84	96.44	100.0	99.47	high	99.04
	50	85.08	91.80	94.67	70.73	98.53	all	98.72
200	5	99.56	99.36	100.0	100.0	99.72		
	10	95.40	97.36	96.55	100.0	99.49	low	96.08
	20	87.80	91.20	92.22	100.0	99.60	high	99.28
	50	72.76	81.80	87.79	100.0	99.47	all	99.20
	100	70.56	82.56	na	94.37	99.20		
500	5	97.68	97.64	100.0	100.0	98.13		
	10	89.52	92.12	99.28	100.0	99.15	low	94.32
	20	78.00	81.28	96.32	100.0	98.81	high	98.56
	50	68.24	70.32	82.58	100.0	98.13	all	98.16
	125	68.72	70.04	80.97	100.0	99.20		
	250	67.92	70.12	na	na	na		
1000	5	94.48	94.60	100.0	100.0	98.31		
	10	83.04	83.56	95.26	100.0	98.81	low	90.40
	20	72.52	73.92	94.21	100.0	99.32	high	96.72
	50	67.36	68.65	72.97	100.0	97.12	all	93.92
	100	67.52	67.86	70.18	100.0	96.27		
	500	65.12	67.36	na	na	na		

Table 1. Percentage of sucess in the comparison between random walk plus drift and deterministic trend processes.

4 Discussion

In this paper we use different dependence metrics for comparison of a particular type of nonstationary time series models. Simulation results show that the metrics based on the sample autocorrelations, the sample partial autocorrelations, the Kullback-Leibler information measure and the normalized periodogram can distinguish quite well between deterministic trend and stochastic trend processes. In particularly, we point out the performance of the sample partial autocorrelation metric in this type of comparison. For the autocorrelation-based metrics we note that short lags L provide better results. This can be explained by the structure of these models, since the main differences arise for the first ACF and PACF values. Contrarily to what could be expected, the performance of ACF methods decreases with sample size. This does not happen with the PACF method. Kullback-Leibler method shows a remarkable good performance and stability across sample sizes and ACF orders considered. The periodogram-based metric compares well to Kullback-Leibler and is computationally simpler.

Acknowledgment: We thank Daniel Peña of Universidad Carlos III de Madrid for contributions in the development of the methods used in this paper. This research was supported by a grant from the Fundação para a Ciência e Tecnologia (POCTI/FCT).

References

1.Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Springer, New York.

2.Caiado, J., Crato, N. and Peña, D. (2005). "A periodogram-based metric for time series classification", *Computational Statistics* & *Data Analysis* (forthcoming).

3.Enders, W. (1995). *Applied Econometric Time Series*, Wiley, New York.

4.Galeano, P. and Peña, D. (2000). "Multivariate analysis in vector time series", *Resenhas*, 4, 383-404.

5.Kakizawa, Y., Shumway, R. H. and Taniguchi, M. (1998). "Discrimination and clustering for multivariate time series", *Journal of the American Statistical Association*, 93, 328-340.

6.Maharaj, E. A. (2000). "Clusters of time series", *Journal of Classification*, 17, 297-314.

7.Maharaj, E. A. (2002). "Comparison of non-stationary time series in the frequency domain", *Computational Statistics* & *Data Analysis*, 40, 131-141.

8.Piccolo, D. (1990). "A distance measure for classifying ARIMA models", *Journal of Time Series Analysis*, 11, 152-164.

9.Shaw, C. T. e King, G. P. (1992). "Using cluster analysis to classify time series", *Physica D*, 58, 288-298.

10.Tong, H e Dabas, P. (1990). "Cluster of time series models: an example", *Journal of Applied Statistics*, 17, 187-198.

11.Xiong, Y. e Yeung, D. (2004). "Time series clustering with ARMA mixtures", *Pattern Recognition* (in press).