Working Paper 01-24
Statistics and Econometrics Series 15
March 2000

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

MULTIVARIATE ANALYSIS IN VECTOR TIME SERIES

Pedro Galeano and Daniel Peña *

Abstract

This paper reviews the applications of classical multivariate techniques for discrimination, clustering and dimension reduction for time series data. It is shown that the discrimination problem can be seen as a model selection problem. Some of the results obtained in the time domain are reviewed. Clustering time series requires the definition of an adequate metric between univariate time series and several possible metrics are analyzed. Dimension reduction has been a very active line of research in the time series literature and the dynamic principal components or canonical analysis of Box and Tiao (1977) and the factor model as developed by Peña and Box (1987) and Peña and Poncela (1998) are analyzed. The relation between the nonstationary factor model and the cointegration literature is also reviewed.

Keywords: Canonical Analysis; Cluster Analysis; Classification; Dynamic Factor Model; Discriminant Analysis; Principal Components.

*Galeano, Department of Statistics and Econometrics, Universidad Carlos III de Madrid, C/ Madrid, 126 Getafe (Madrid), e-mail: pgaleano@est-econ.uc3m.es; Peña, Department of Statistics and Econometrics, Universidad Carlos III de Madrid, e-mail: dpena@est-econ.uc3m.es

Multivariate Analysis in Vector Time Series

Pedro Galeano and Daniel Peña
Department of Statistics & Econometrics
Universidad Carlos III de Madrid

March 27, 2001

Abstract

This paper reviews the applications of classical multivariate techniques for discrimination, clustering and dimension reduction for time series data. It is shown that the discrimination problem can be seen as a model selection problem. Some of the results obtained in the time domain are reviewed. Clustering time series requires the definition of an adequate metric between univariate time series and several possible metrics are analyzed. Dimension reduction has been a very active line of research in the time series literature and the dynamic principal components or canonical analysis of Box and Tiao (1977) and the factor model as developed by Peña and Box (1987) and Peña and Poncela (1998) are analyzed. The relation between the nonstationary factor model and the cointegration literature is also reviewed.

Key words: Canonical Analysis, Cluster Analysis, Classification, Dynamic Factor Model, Discriminant Analysis, Principal Components.

1 Introduction

Standard multivariate analysis includes, among others, procedures for discrimination among several populations, classification (pattern recognition) of multivariate data into groups, either hierarchical or not, and dimension reduction. These problems are also important in multivariate time series. The discrimination problem appears as follows. Suppose that we know that a set of time series can be generated by one of several possible models, M_i, $i = 1, ..., k$ and we assume that these models are known. Now, we observe a new time series and the problem is to decide which of the models, M_i, has generated this time series. This problem is an important area of research in different disciplines. For instance, in seismology it is important to be able to discriminate between data from earthquakes and nuclear explosions (Dargahi-Noubary, 1992, Dargahi-Noubary and Laycock, 1981, Kakizawa, et al., 1998, Shumway and Unger, 1974). In medicine the information from the electroencephalographic time series (EEG) can be used for discriminating between different stages of sleep (Alagón, 1989, Gersch et al., 1979). In engineering it is important to discriminate between a pattern generated by a signal plus noise and a pattern generated by a noise alone, for example, to detect a radar signal for determining the position of a moving target. In Economics we are interested in classifying the economic situation as expansion or depression by considering the values of some time series economic

indicators. Finally, in Business a company can be classified as successful or in potential trouble by looking at some time series indicators of its economic activity.

The problem of making clusters of set of time series appears also in many scientific fields but most of the published examples of cluster analysis in time series have been made with environmental data. We have time series from different locations and we want to make groups with locations with the same behavior. See for instance Bohte et al. (1980), Cowpertwait and Cox (1992), Gantert (1994), Walden (1994) and Macchiato et al. (1995). There are several problems not completely solved in the application of cluster analysis in time series. The standard approach for splitting a sample of multivariate data into clusters is to assume that the multivariate observations have been generated by a mixture of multivariate normal distributions with different means and covariance matrices and unknown mixture probabilities. If the number of populations were known, the parameters can be estimated by the EM algorithm or by MC^2 Bayesian methods. As the number of population is unknown, a model selection procedure, such as the BIC or AIC criteria is applied to select the number of populations involved. The generalization of these approach to time series is to assume that data has been generated by some set of possible multivariate time series models or data generating processes, $M_1, ..., M_k$, with unknown probabilities, and then the cluster problem is closely related to the discrimination problem. However, this approach has not yet been fully explored.

The problem of dimensionality reduction is very important for dynamic data since for vector ARMA models, as well as for simultaneous equations econometric models, the number of parameters to estimate grows rapidly with the number of observed variables. An interesting extension of the idea of principal components for time series is the canonical analysis of Box and Tiao (1977). Instead of finding linear combination of maximum (or minimum) variability these authors studied the problem of finding linear combinations of maximum (or minimum) predictability. They showed that the canonical variables are useful for understanding and simplifying the dynamic structure present in the vector of time series. Factor analysis of time series was studied by Geweke and Singleton (1981), Brillinger (1981), Engle and Watson (1981), Molenaar (1985), Peña and Box (1987), Molenaar et al. (1992), Peña and Poncela (1998) among others. An alternative approach to dimension reduction is the reduced rank approach by Velu et al. (1986) and Ahn and Reinsel (1988). In the nonstationary case estimating the nonstationary factors is equivalent to testing for cointegration in the econometrics field (whose vast literature we do not pretend to review here), since the number of cointegration relations among the components of a vector of time series is the dimension of the vector minus the number of nonstationary common factors (see Escribano and Peña, 1994). An alternative useful approach for model simplification is the scalar components approach by Tiao and Tsay (1989). Finally the state space approach to time series includes procedures for dimension reduction (Hannan and Deistler, 1988, Aoki, 1990).

This paper describes some of the development of these procedures in the time domain. The reader interested in the development in the frequency domain is advised to read chapter 5 of Shumway and Stoffer (2000), which contains a good review of this field. The article is organized as follows. In the next section the problem of discrimination in time series is presented. The standard discriminant analysis is seen as a model fitting exercise and it is shown that in practice, when the parameters are unknown, discriminant analysis for time series is closely related to the model selection problem that has been the subject of an important area of research in time series. In Section 3 we present the clustering problem and discuss some of the measures of distance among time series that have been proposed in the literature. Some suggestions for

further research in this field are also included. We have decided to consider only in Section 4 the extensions of standard multivariate methods, as the literature on model simplification and dimension reduction is very large. Thus, in the section we present the extension of the principal component idea of Box and Tiao (1977) and the Dynamic factor model. The relationship between both approaches is discussed and we also relate the nonstationary factor model and the cointegration literature. Section 5 presents some concluding remarks.

2 Discrimination in time series

2.1 Linear Discrimination

Discriminant analysis has been mainly studied for Gaussian processes. The classical approach is as follows. Suppose a series with T observations, denoted by $x = (x_1, ..., x_T)'$, which follows a Gaussian process with vector of marginal means $\mu_j = (\mu_{j1}, ..., \mu_{jT})'$ for $j = 1, 2$. Assume that the process $x - \mu_j$ is a zero mean stationary process with covariance matrix $\Sigma_j = \{\sigma_j(s-t) : s, t = 1, ..., T\}$. Thus, under the hypothesis H_j, $x \sim N_T(\mu_j, \Sigma_j)$, for $j = 1, 2$. Then, the probability density function for this process is,

$$p(x/H_j) = (2\pi)^{-\frac{T}{2}} |\Sigma_j|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu_j)' \Sigma_j^{-1}(x-\mu_j)\right), \tag{1}$$

The classical approach suppose that both covariance matrices are equal, $\Sigma_1 = \Sigma_2 = \Sigma$, but the means are unequal. Thus, we assume that the difference between the two marginal means is due to some deterministic function. For instance, if $\mu_{ji} = b_{oj} + b_{1j}i$, the series have a different deterministic trend and if $b_{1j} = 0$ the series have a different marginal mean. The Neyman-Pearson lemma for the hypothesis $H_1 : x \in M_1$ versus $H_2 : x \in M_2$, leads to the following rule for accepting H_1 :

$$\frac{p(x/H_1)}{p(x/H_2)} > K, \tag{2}$$

for some value K that takes into account the probabilities of misclassifying the time series. Assuming that the costs of missclassification are the same and that the a priori probabilities of each model are also the same, we will classify the observation in the model that have the maximum likelihood. This is equivalent to accept the hypothesis H_1 if

$$(x-\mu_1)' \Sigma^{-1}(x-\mu_1) < (x-\mu_2)' \Sigma^{-1}(x-\mu_2)$$

that is, if we denote by $D_i = (x-\mu_i)' \Sigma^{-1}(x-\mu_i)$ to the Mahalanobis distance between the data and the vector of marginal means, x is classified in the first population if $D_2 > D_1$. An alternative interpretation of this rule can be obtained by writing this equation as:

$$(\mu_1-\mu_2)' \Sigma^{-1} x > (\mu_1-\mu_2)' \Sigma^{-1} \frac{1}{2}(\mu_1+\mu_2), \tag{3}$$

that implies that the scalar measure $v = \alpha' x$ is built, where

$$\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$$

calling $m_1 = \alpha'\mu_1$ and $m_2 = \alpha'\mu_2$ the series is classified in M_1 if $v > \frac{(m_1+m_2)}{2}$. If we denote for D_{12}, the Mahalanobis distance between the means of both populations,

$$D_{12} = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) = m_1 - m_2 \tag{4}$$

then, the linear discrimination function, v, is normally distributed with mean m_1 under H_1, and m_2 under H_2. The variance is D_{12} in both cases. Thus we classify on M_1 if the scalar variable v is closer to m_1 than to m_2.

Note that this rule, obtained by the likelihood ratio test, it is equivalent to fitting the time series by both models and then choosing the model that leads to a smaller residual variance. This result is clear from (1) because note that $e_j = x - \mu_j$, $j = 1, 2$ are the residuals from the deterministic fit $\widehat{x} = \mu_j$ and $a_j = \Sigma^{-1/2} e_j$ corresponds to the residuals taking into account the stationary structure. Note that the errors a_j have an identity covariance matrix. Thus

$$a_j' a_j = e_j' \Sigma^{-1} e_j = (x - \mu_i)' \Sigma^{-1} (x - \mu_i) \tag{5}$$

and minimum Mahalanobis distance is equivalent to minimum residual sum of squares. Another way to look at this property is by noting that if e_j follows a zero mean linear process the likelihood $f(e_j)$ can be written, by using the prediction error decomposition, as

$$f(e_j) = f(e_{j1}) f(e_{j2}/e_{j1}) .... f(e_{jT}/e_{j1} .... e_{jT-1})$$

and the likelihood will only depend on the one-step ahead forecasting errors that are equal to the residuals a. Note that for linear time series we can write the zero mean process $\pi(B) e_t = \epsilon_t$, where $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - ....$ as $\Pi e = \epsilon$, where :

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -\pi_p & \ddots & \ddots & \ddots & \vdots \\ \ddots & \ddots & \ddots & \ddots & 0 \\ -\pi_{T-1} & \ddots & -\pi_p & \cdots & 1 \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_T \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix}.$$

Suppose that the covariance matrix of ϵ is $\sigma^2 I$. Then, calling Σ to the covariance matrix of e, we have,

$$\Pi \Sigma \Pi' = \sigma^2 I$$

and, therefore,

$$\Sigma^{-1} = \frac{1}{\sigma^2} \Pi' \Pi$$

and

$$e' \Sigma^{-1} e = e' \frac{1}{\sigma^2} (\Pi' \Pi) e = \frac{1}{\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} a_i^2$$

in agreement with (5). Thus, discriminant analysis can be viewed as assigning the observed time series x to the model (population) that when fitted to the time series produces the smallest one step ahead squared forecast error.

2.2 Unequal Covariance Matrices

A more relevant case in time series discrimination is when the covariance matrices are unequal. Suppose that we want to discriminate between two time series models. Both imply Gaussian populations but with different covariance matrices and, for simplicity, we will assume that the marginal means are in both cases equal to 0. Then, the rule (2) says that we accept H_1 if

$$Q(x) = x' \left(\Sigma_2^{-1} - \Sigma_1^{-1}\right) x > K$$

which is a quadratic form. This discriminant rule has a simple interpretation in term of prediction errors, because, as before, $x'\Sigma_i^{-1}x$ is the residual sum of squares of the fitted model. Thus, the likelihood ratio test leads to fitting the observed time series with both models and choosing the one with the smallest one step ahead forecast error.

An alternative interesting interpretation of the discriminant rule is that it assigns the series to the model producing the smallest interpolation error. The best linear interpolator of a time series is given by (see for instance Peña and Maravall, 1991)

$$\widehat{x}_s = E\left[x_s | x_t, t \neq s\right] = -\sum_{i=1}^{\infty} \rho_i^D \left(x_{s-i} + x_{s+i}\right)$$

where ρ_i^D are the coefficients of the dual autocorrelation function of the model given by:

$$\rho^D(B) = \sigma^2 \frac{\pi(B)\,\pi(F)}{v_D}$$

with $\pi(B)$, the autoregressive form of the model, $v_D = \sigma^2 \sum_{i=0}^{\infty} \pi_i^2$ and $F = B^{-1}$, the forward operator. Then, Galeano and Peña (2001) showed that :

$$x'\Sigma_x^{-1}x = (x - \widehat{x})' D_{\widehat{x}}^{-1} (x - \widehat{x})$$

where $D_{\widehat{x}}$ denotes a diagonal matrix with the variance of the interpolation errors. That is, the series x_t is assigned to the model that produces the smallest interpolation error or, in other words, the model that better adjusts the data.

As the distribution of $Q(x)$ is difficult to find, Shumway (1982) suggests that under H_j, $j = 1, 2$, and for large values of T, $Q(x)$ can be approximated by a normal distribution with mean $tr\left(\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)\Sigma_j\right)$ and variance $2 \cdot tr\left(\left(\Sigma_2^{-1} - \Sigma_2^{-1}\right)\Sigma_j\right)^2$, where tr denotes trace. This method has the principal inconvenient that the eigenvalues must be obtained numerically, being the matrices $\left(\Sigma_2^{-1} - \Sigma_2^{-1}\right)\Sigma_j$ very large, which makes a numerical solution very difficult to obtain.

When the covariance matrix are different the optimum discriminant rule is not linear. An alternative approach in these situations is to obtain a good linear discriminant rule according to some criteria. This is the idea of admissible linear procedures introduced by Anderson and Bahadur (1962). For Gaussian populations, under H_i, a linear discriminant rule, $\alpha' x$, has a univariate normal distribution with mean $\alpha'\mu_i$ and variance $\alpha'\Sigma_i\alpha$. Therefore the probability of misclassifying an observation are given by

$$\begin{aligned} \Pr\left(\alpha' x < K | x \in M_1\right) &= \Phi\left(\frac{K - \alpha'\mu_1}{\sqrt{\alpha'\Sigma_1\alpha}}\right) \\ \Pr\left(\alpha' x > K | x \in M_2\right) &= \Phi\left(\frac{\alpha'\mu_2 - K}{\sqrt{\alpha'\Sigma_2\alpha}}\right) \end{aligned}$$

where $\Phi(x)$ is the cdf of the $N(0,1)$ distribution. The objective is to make these values as small as possible, and this is equivalent to make the values, $y_1 = \frac{K-\alpha'\mu_1}{\sqrt{\alpha'\Sigma_1\alpha}}$ and $y_2 = \frac{\alpha'\mu_2-K}{\sqrt{\alpha'\Sigma_2\alpha}}$, small. The set of desirable procedures are those that: (1) minimize the probability of one error when the other is specified, or (2) minimize the maximum probability of error, or (3) minimize the probability of error when a priori probabilities of the two populations are specified. The solutions to these problems are the set of admissible linear procedures. The set of solutions that minimizes y_1 for each given y_2 is characterized by,

$$\alpha = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}(\mu_2 - \mu_1) \tag{6}$$

where the values t_1 and t_2 verify that $K = \alpha'\mu_1 + t_1\alpha'\Sigma_1\alpha = \alpha'\mu_2 - t_2\alpha'\Sigma_2\alpha$.

Information measures usually leads to admissible linear procedures. For instance, Kullback (1959) considered the Kullback-Leibler discrimination information for discriminating in favor of H_1 over H_2. It is given by,

$$\begin{aligned} I(1:2,\alpha'x) &= \frac{1}{T}E_1\left(\log\frac{p_1(\alpha'x)}{p_2(\alpha'x)}\right) = \\ &= \frac{1}{2T}\left[Tr\left(\Sigma_1\Sigma_2^{-1}\right) - \log\frac{|\Sigma_1|}{|\Sigma_2|} - T + (\mu_1-\mu_2)'\Sigma_2^{-1}(\mu_1-\mu_2)\right] \end{aligned}$$

Another useful measure is the divergence that for discriminating is defined by,

$$J(1:2,\alpha'x) = I(1:2,\alpha'x) + I(2:1,\alpha'x)$$

The values of α that maximize $I(1:2,\alpha'x)$, $I(2:1,\alpha'x)$ or $J(1:2,\alpha'x)$ are of the form $\Sigma_1\alpha - \lambda\Sigma_2\alpha = \gamma\delta$, where $\delta = (\mu_1 - \mu_2)$, for some values of the scalars λ and γ. As a consequence of this, the procedures based in the Kullback-Leibler information and in the divergence are admissible linear procedures. Chaudhuri, et al. (1991) obtain linear discriminant procedures through the maximization of the Bhattacharyya distance for Gaussian processes with unequal covariance matrices. If (Ω,β,υ) is a measure space and $\wp$ is the set of all the probability measures on β which are absolutely continuous with respect to υ, then the Bhattacharyya distance between two probability measures with density functions p_1 and p_2 belonging to $\wp$, is defined by,

$$-\ln\rho(p_1,p_2) = -\ln\int_\Omega\sqrt{p_1p_2}d\upsilon$$

Under $H_1 : x \in N(\mu_1,\Sigma_1)$ and $H_2 : x \in N(\mu_2,\Sigma_2)$, the linear discriminant function obtained maximizing $-\ln\rho(p_1,p_2)$ is,

$$\alpha'x = (\mu_1-\mu_2)'(\Sigma_1-\Sigma_2)^{-1}x$$

Chaudhuri (1992) considered the problem of classifying a complex normal time series through the maximization of the previous distance.

When the parameters of the models are unknown they must be estimated from the data. Although in principle we can plug in the estimates and use the same criteria that in the known

parameter case this is not a good solution when the number of parameters in both models are very different. For instance, suppose that one of the possible model is an AR(1) and the other is an AR(5) with four complex roots, that is, we are checking if an observed time series presents pseudo-cycles. Then if we use the plug in procedure of obtaining the estimates and introducing them in the discriminant function we will always get that the model with a larger number of parameters provides a better fit. Thus we have to take into account the difference between in sample fit and out of sample forecast.

Several criteria has been proposed for selecting time series models since the seminal work of Akaike (1969, 1974). Among them are the Bayesian Information criteria BIC of Schwarz (1978) and Akaike (1979), the penalty methods of Hannan and Quinn (1979), the predictive least squares criterion of Rissanen (1986), extended by Lai and Lee (1997), and the modified AIC of Hurvich and Tsai (1989) and Cavanaugh and Shumway (1997). Surveys on the performance of these criteria for ARMA order selection can be found in Bhansali (1993) and Postcher and Srinivasan (1994).

These criteria have the general form

$$C = -2(\log \text{max imized likelihood}) + f(\text{number of parameters}) \tag{7}$$

where the function f depend on the criteria. For instance, for ARMA models the AIC of Akaike is

$$AIC = n \log \widehat{\sigma}^2 + 2(p+q)$$

where $(p+q)$ is the number of parameters in the model. The BIC criteria due to Schwarz (1978) is

$$BIC = -2(\log \text{max imized likelihood}) + (\log n)(\text{number of parameters}). \tag{8}$$

This last criterion has been showed to have a very good performance in many model selection problems.

Some Bayesian approaches have been proposed that do not adopt a formal Bayes rules via a loss function. For example, Broemeling and Son (1987) consider how to assign an observed time series to one of several possible autoregressive sources with a common known order and unknown parameters and error variance. Using a vague prior density for the parameters and the variance, the observed time series data is assigned to one class by using the marginal posterior mass function of a classification vector,

$$\lambda = (\lambda_1, ..., \lambda_k)'$$

with k mass points $(1, 0, ..., 0)'$,..., $(0, ..., 0, 1)'$. The realization is assigned to the process i if the posterior mass function of λ has its largest value at the $i-th$ mass point. Marco, et al. (1988) consider the case of different autoregressive classes. The data is assigned to the class k, if it has the greater predictive probability.

Finally, there exist other approaches for discrimination in which discriminant functions are not used. For instance Kedem and Slud (1982) proposed transform a stationary time series into binary arrays that retain only the signs of the $j-th$ difference series. This binary series are used for discriminating among different models. Li (1996) proposed a generalization of this method through the use of parametric filtering, i.e., using a family of filters indexes by a parameter. The series is filtered and the information provided by the autocorrelation function is used for discriminate the series into different models.

3 Clustering time series

Suppose that we have a large set of time series following different models. In a nonparametric approach each series is considered as a point in $\Re^T$, where T is the length of the series. A straightforward generalization of the standard cluster methods is to obtain groups of series by looking at the distance between these points in the space. In order to identify groups we can work directly with a distance metric in $\Re^T$, or we can try to work in a smaller space by projecting the points according to some optimality criterion. This criterion should be related to the possibility of identifying clusters in the projected cloud of points.

A parametric approach would proceed by first fitting time series models to the data and then representing the series by the vector of estimated parameters. If the dimension of the vector of parameters is p these parameter vectors will be points in $\Re^p$, and again we can try to find points that are close in this space. In both cases we can define a measure of distance and then use a standard k-means type algorithm. Thus, an important first step is obtaining an appropriate metric for measuring the similarity between points.

In the parametric ARIMA approach each time series is represented by the vector of parameters models. For instance, if the series are fitted by

$$\phi_i(B)(1-B)^d x_{it} = \theta_i(B)\epsilon_{it}, \qquad i = 1, ..., k$$

where $\phi_i(B) = 1 - \phi_{i1}B - ... - \phi_{ip}B^p$, and $\theta_i(B) = 1 - \theta_{i1}B - ... - \theta_{iq}B^q$, we can represent the series by the autoregressive and moving-average parameters including all of them in a vector

$$\beta_i = \left(\phi_{i1}, ..., \phi_{ip}, \theta_{i1}, ..., \theta_{iq}\right)^T$$

and then defining a measure of distance by

$$D(x_i, x_j) = (\beta_i - \beta_j)'\Sigma_\beta^{-1}(\beta_i - \beta_j)$$

where Σ_β is an appropriate matrix to define the metric that, in particular, it can be the identity. Bruce and Martin (1989) consider a similar measure of distance between ARIMA models. However, as indicated by Peña (1989) this measure has three main principal problems. The first is that it cannot compare ARIMA models with different degrees of differencing. The second is that it does not take into account the possibility of cancellation between AR and MA. For instance the models $(1 - .9B)x_t = (1 - .89B)\epsilon_t$ is almost exactly the same as the model $x_t = \epsilon_t$ whereas with this metric both will seem very different. The third is that it does not allow for the duality between the AR and MA forms. A more convenient measure is defined through the comparison between the coefficients of the polynomial $\pi(B)$, obtained from $\theta(B)\pi(B) = \phi(B)(1-B)^d$.

Piccolo (1990) introduced a metric for ARIMA models that can be used for classifying and clustering time series. Let x_t is a zero mean stochastic processes following an ARIMA(p, d, q) model in the usual notation, $\phi(B)x_t = \theta(B)\epsilon_t$, where ϵ_t is Gaussian white noise. When x_t is invertible, it is possible to define the autoregressive operator $\pi(B) = \theta^{-1}(B)\phi(B) = 1 - \pi_1 B - \pi_2 B^2 - ....$ The coefficients of $\pi(B)$ conveys all usual information about the stochastic structure given initial values and the order of the process. If $\pounds$ denotes the set of invertible processes, we can define a measure of structural diversity between processes in $\pounds$ comparing

their respective π sequences. The metric on $\pounds$ is defined by the distance,

$$d(x,y) = \left\{ \sum_{j=1}^{\infty} \left(\pi_{j,x} - \pi_{j,y}\right)^2 \right\}^{\frac{1}{2}}$$

which always exists for every x, $y \in \pounds$, and being the zero element the sequence $(0, ..., 0)$. We notice that a dual metric can be defined by:

$$d(x,y) = \left\{ \sum_{j=1}^{\infty} \left(\psi_{j,x} - \psi_{j,y}\right)^2 \right\}^{\frac{1}{2}}$$

where the ψ sequence defines the $MA(\infty)$ operator as $\psi(B) = \phi(B)^{-1}\theta(B) = \pi^{-1}(B)$. However this metric can not be computed for integrated processes.

This definition of distance allows to perform applications to clustering algorithms, through the study of similarities between time series. Piccolo (1990) applies the following method to study a possible similarity in the behavior of industrial production series in different sectors. The algorithm starts defining a model for each series considered and in base of this model the distances between all the time series are computed. Then a dendrogram based on the similarities is built and that gives us the different clusters formed by the models. An alternative procedure, also used in the paper, is the classical solution of multidimensional scaling to the distance matrix previously obtained, that is, obtaining a configuration of points in a convenient space where the interpoint distance reproduces the similarity matrix. The results found in the two ways are very similar.

Nonparametric clustering techniques for time series have been less studied because of the difficulties of defining a general measure of distance between stationary time series sequences. In order to illustrate a possible procedure suppose that we have n zero mean and unit variance stationary time series sequences, $X_1, ..., X_n$. We assume first that the data has been centered and scaled. A possible distance metric among the points X_i is the euclidean metric. However, this metric is invariant to transformations which modify the order of the observation over time in the two series that are compared and, therefore, it does not take into account the correlation structure of the stationary data. That is, given the original set of time series $X_i = (x_{i1}, ..., x_{iT})$ for $i = 1, ..., k$, if we know built a new set of time series sequences $X_i^* = (x_{i1}^*, ..., x_{iT}^*)$ by using the same permutation of the time observation for all the series the euclidean distance between the elements in the second set are identical to those in the first, whereas the correlation structure of the second set can be arbitrarily distorted. Thus, the euclidean distance does not take into account the autocorrelation structure.

The distance measure to be used depends on the kind of similarities we are interested in. We may be interested in (a) Finding series with a similar correlation structure or (b) Finding series with a similar noise structure. In case (a) a straightforward measure of distance is to compute the autocorrelation coefficients $r_i = (r_i(1), ..., r_i(h))$ for some h such that $r_i(j) \simeq 0$ for $j > h$ and then use

$$D(X_i, X_j) = (r_i - r_j)' W_r (r_i - r_j)$$

for some weighting function W_r that can be used to give weights to the coefficients that decrease with the lag. This measure is related to the parametric approach as the parameters of the autoregressive approximation are computed from the autocorrelation coefficients.

For (b), a model is fitted to each series and the residuals $\widehat{\epsilon}$ are obtained. Then a measure of distance between them is built by

$$D(X_{i,}X_j) = (\widehat{\epsilon}_i - \widehat{\epsilon}_j)'W(\widehat{\epsilon}_i - \widehat{\epsilon}_j),$$

where, for instance, the matrix W can be used to give more weight to the recent values than to the oldest values in the time series. Both procedures can be combined to define a measure of distance that takes into account both sources of variability by

$$D(X_{i,}X_j) = \lambda_1(r_i - r_j)'W_r(r_i - r_j) + \lambda_2(\widehat{\epsilon}_i - \widehat{\epsilon}_j)'W(\widehat{\epsilon}_i - \widehat{\epsilon}_j)$$

where λ_i, $i = 1, 2$ are normalizing constants. This idea does not seem to have been yet explored in the literature.

4 Dimension reduction

4.1 Canonical Analysis

The canonical analysis of time series was introduced by Box and Tiao (1977) and can be considered as a principal component analysis of time series. This work was very important because (1) it leads to a clear solution of the dimension reduction problem in terms of prediction, (2) it introduces, for the first time, the idea that linear combination of nonstationary time series can be stationary, that is, the idea of cointegration.

Suppose a $m \times 1$ vector x_t that follows a stationary VAR(p) model

$$\phi(B)\, x_t = \varepsilon_t,$$

we can always write the orthogonal decomposition

$$x_t = \widehat{x}_{t-1}(1) + \varepsilon_t$$

where $\widehat{x}_{t-1}(1)$ is the one step ahead prediction. Corresponding to this decomposition we can also split the covariance matrix, $E\left[x_t x_t^{'}\right] = \Gamma_x(0)$, as

$$\Gamma_x(0) = F_x(0) + \Sigma$$

where $E(\varepsilon_t \varepsilon_t') = \Sigma$ and $E\left[\widehat{x}_{t-1}(1)\widehat{x}_{t-1}(1)\right] = F_x(0)$. We are interested in finding a linear combination of x_t

$$z_{1t} = m' x_t$$

such that it has maximum predictability. The variance of this linear combination is $m'\Gamma_x(0)m$ and this variance is decomposed into an explained variability, $m'F_x(0)m$, and a residual variability, $m'\Sigma m$. We want to maximize

$$\lambda = \frac{m'F_x(0)m}{m'\Gamma_x(0)m}$$

and the value that maximize this equation is

$$\Gamma_x(0)^{-1} F_x(0) m = \lambda m.$$

Thus, m must be the largest eigenvector of the matrix obtained as product of the matrix of explained variability and the inverse of the matrix of total variability,

$$Q = \Gamma_x(0)^{-1} F_x(0) = \Gamma_x(0)^{-1} \left(\Gamma_x(0) - \Sigma\right).$$

The procedure can be extended to find other linear combinations by choosing as m the ordered eigenvectors of the matrix Q. Thus, in practice the canonical decomposition consists in finding the eigenvalues and eigenvectors of the matrix

$$Q m_i = \lambda_i m_i,$$

the eigenvectors provide the required linear combination and the eigenvalues the predictability of these linear combinations. Building the matrix $M = [m_1 ... m_p]$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ and the transformation

$$z_t = M' x_t$$

a new vector of time series is obtained with components ordered from most to least predictable. The components are contemporaneously uncorrelated because it is easy to show that the matrices $M'\Gamma_0 M$ and $M'\Sigma M$ are both diagonal.

4.2 The Dynamic Factor model

The factor model has a straightforward extension to the dynamic case. Considered the possible nonstationary vector process x_t. The dynamic factor model assumes that this time series vector, which we assume has dimension m, has been generated by the equation

$$x_t = P f_t + n_t, \tag{9}$$

where P is a $m \times r$ loading matrix that we assumed is normalized in such a way that $P'P = I$. Thus, all the common dynamic structure comes through the common factors, f_t, and the n_t includes the independent idiosyncratic components. We suppose that the vector of common factors follows a VARIMA(p, q) model

$$\Phi(B) f_t = \Theta(B) a_t \tag{10}$$

where $\Phi(B) = I - \phi_1 B - ... - \phi_p B^p$ and $\Theta(B) = I - \theta_1 B - ... - \theta_q B^q$ are polynomial matrices $r \times r$ and the roots of $|\Phi(B)|$ are on or outside the unit circle and those of $|\Theta(B)|$ are outside the unit circle. The sequence a_t is serially uncorrelated with zero mean and covariance matrix Σ_a. The components of the vector of common factors can be either stationary or nonstationary.

The specific dynamic structure associated with each of the observed series is included in the vector n_t of idiosyncratic components. Some components of this vector can be white noise, while other ones can have stationary dynamic structure. In general, we assume that n_t follows the vector ARMA model

$$\Phi_n(B) n_t = \Theta_n(B) e_t, \tag{11}$$

where $\Phi_n(B)$ and $\Theta_n(B)$ include $m \times m$ diagonal matrices. The sequence of vectors e_t are normally distributed, have zero mean and diagonal covariance matrix Σ_e. Therefore, each component follows an univariate ARMA(p_i, q_i), $i = 1, 2, \cdots, m$, being p=max(p_i) and q=max(q_i), $i = 1, 2, \cdots, m$. We assume that the noises from the common factors and specific components are also uncorrelated for all lags, $\mathrm{E}(a_t e'_{t-h}) = 0$, $\forall h$.

The model as stated is not identified, because for any $r \times r$ non singular matrix H the observed series can be expressed in terms of a new set of factors,

$$x_t = PH^{-1}Hf_t + n_t = P^* f_t^* + n_t$$

where $f_t^* = Hf_t$, and

$$\Phi^*(B) f_t^* = \Theta^*(B) a_t^*$$

where $a_t^* = Ha_t$. With this transformation the old system matrices are related to the new system matrices by

$$\begin{aligned} \Phi^*(B) &= H\Phi(B)H^{-1} \\ \Theta^*(B) &= H\Theta(B)H^{-1} \\ \Sigma_a^* &= H\Sigma_a H' \end{aligned}$$

To solve this identification problem, we can always choose either $\Sigma_a = I$ or $P'P = I$. Note that as

$$P^{*'}P^* = (H^{-1})'P'PH^{-1}$$

if $P'P = I$ then $P^{*'}P^* = (H^{-1})'H^{-1}$ that will only be the identity matrix if H is orthogonal. Therefore the model is not yet identified under rotations, and we need to introduce a restriction to estimate the model. The standard restriction used to solve this problem in static factor analysis is that $P'\Sigma_n^{-1}P$ should be diagonal. Harvey (1989) imposes that $p_{ij} = 0$, for $j > i$, where $P = [p_{ij}]$. This condition is not restrictive, since the factor model can be rotated for a better interpretation when needed (see Harvey, 1989, for a brief discussion about it).

Peña and Poncela (1998) showed that the model presented is fairly general and includes also the case where lagged factors are present in equation (9). For instance, for ease of exposition assume a stationary model with no specific components, but with lagged factors on the observation equation, such as

$$x_t = Pv(B)F_t + n_t$$

where $v(B) = I + v_1 B + \cdots + v_l B^l$, $l < \infty$ and F_t follows a VARMA model

$$F_t = \Psi(B)a_t, \ \Psi_0 = I$$

This model can be rewritten in the standard form presented in (9) with

$$f_t = F_t + v_1 F_{t-1} + \cdots + v_l F_{t-l}$$

following the VARMA model $f_t = \widetilde{\Psi}(B)a_t$ where $\widetilde{\Psi}(B) = \sum_{i=1}^{\infty} \widetilde{\Psi}_i B^i$ and $\widetilde{\Psi}_i$ satisfies $\widetilde{\Psi}_i = \Psi_i + v_1\Psi_{i-1} + \cdots + v_l\Psi_{i-l}$ with $\Psi_j = 0_{r\times r}$ if $j < 0$ and $\widetilde{\Psi}_0 = I$. Since matrices v_i are

of constants coefficients, $\|v_i\| < \infty$ and equation $f_t = \widetilde{\Psi}(B)a_t$ also represents a VARMA stationary process. Therefore the standard formulation presented in (9) can include important complex relationships between the series and the factors.

In the particular case that all the factors are stationary and the component n_t is white noise, ϵ, the dynamic factor model reduces to the model studied by Peña and Box (1987). In this case, assuming $E(x_t) = 0$ and calling $\Gamma_x(k) = E\left[x_t x_{t-k}^{'}\right]$ and $\Gamma_f(k) = E\left[f_t f_{t-k}'\right]$ we have that

$$\begin{aligned}\Gamma_x(0) &= P\Gamma_f(0)P' + \Sigma_\epsilon \\ \Gamma_x(k) &= P\Gamma_f(k)P', k \geq 1\end{aligned}$$

which implies that the columns of P are eigenvectors of the matrix $\Gamma_x(k)$ for all $k \geq 1$. To show this note that

$$E\left[x_t x_{t-k}^{'}\right] = E\left[(Pf_t + \epsilon_t)(Pf_{t-k} + \epsilon_{t-k})'\right]$$

and,

$$E\left[Pf_t f_{t-k}'P' + \epsilon_t\epsilon_{t-k}' + Pf_t\epsilon_{t-k}' + \epsilon_t f_{t-k}'P'\right] = PE\left[f_t f_{t-k}'\right]P' = P\Gamma_f(k)P'$$

Thus, the eigenvalues of $\Gamma_x(k)$ are the covariance of the factors, $k \geq 1$. Peña and Box (1987) proposed the following procedure to recover the factors:

(1) Compute eigenvalues and eigenvectors of $\Gamma_x(k)$ for $k \geq 1$.

(2) Obtain the number of common factors by the rank of the matrices $\Gamma_x(k)$. Assume that the common rank is r, the number of common factors.

(3) Use the non zero eigenvectors of $\Gamma_x(k)$, for $k \geq 1$, in order to estimate the loading matrix P.

(4) Build the transformation $M = [P \;\; V]$, where $P'V = 0$, V belong to the null space of P and apply it to the x_t in order to recover the factors. As $P'P = I$ we have

$$P'x_t = f_t + P'\epsilon_t$$

and

$$V'x_t = V'\epsilon_t$$

then, the transformation

$$z_t = M'x_t$$

gives r linear combinations of the time series components that will recover the factors and $m-r$ combinations that will be white noise.

This model has been studied in the nonstationary case by Peña and Poncela (1998). They showed that the identification of the nonstationary $I(d)$ factors can be made through the common eigenstructure of some generalized covariance matrices, properly normalized. The number of common nonstationary factors is the number of nonzero eigenvalues, Thus, a similar identification procedure can be applied in the stationary and in the nonstationary case. Once we have a preliminary estimation of the dimension of the system we can estimate the factor loading matrix and the parameters of the VARIMA factor representation by writing the model in the state space from and use the EM algorithm.

4.3 Relationship between Canonical Analysis and the Dynamic factor model

Let us show the relationship between the dynamic principal components or canonical analysis and the standard principal component approach that considers the eigenstructure of $\Gamma_x(0)$. In the canonical analysis we obtain eigenvectors from $Q = \Gamma_x(0)^{-1}(\Gamma_x(0) - \Sigma) = I - \Gamma_x(0)^{-1}\Sigma$. Note that:

(1) Q, $\Gamma_x(0)^{-1}\Sigma$, and $\Sigma^{-1}\Gamma_x(0)$ have the same eigenvectors;

(2) The largest eigenvalue of $I - \Gamma_x(0)^{-1}\Sigma$ is the smallest of $\Gamma_x(0)^{-1}\Sigma$;

(3) The smallest eigenvalue of $\Gamma_x^{-1}(0)\Sigma$ is the largest of $\Sigma^{-1}\Gamma_x(0)$.

Then the canonical analysis can be interpreted as obtaining eigenvectors from $\Sigma^{-1}\Gamma_x(0)$, whereas the standard principal component approach uses directly the matrix $\Gamma_x(0)$.

To understand better this difference, suppose that the factorial model hold and

$$\Gamma_x(0) = P\Gamma_f(0)P' + \Sigma$$

then

$$\Sigma^{-1}\Gamma_x(0) = \Sigma^{-1}P\Gamma_f(0)P' + I$$

and if V is such that $P'V = 0$ then

$$\Sigma^{-1}\Gamma_x(0)V = V$$

and therefore a transformation based on the eigenvectors of $\Sigma^{-1}\Gamma_x(0)$ will also separate the factors from white noise.

Let us consider now the relationship between the canonical analysis and the identification procedure in the factor model as developed by Peña and Box (1987) and Peña and Poncela (1998). Consider the stationary case to simplify. Then the factor are initially estimated by computing eigenvalues and eigenvectors of $\Gamma_x(k)$ for $k \geq 1$. Thus this identification depends only on P. whereas in the canonical analysis the components obtained depend on both P and Σ.

4.4 Cointegration and the factor model

Suppose that x_t follows a nonstationary model $\phi(B)\nabla x_t = \varepsilon_t$. Then we say that x_t is $I(1)$. There will be cointegration among the components if we can find linear combinations that are stationary. That is, we will say that the components of x_t are cointegrated if there exits a $m \times p$ matrix β such that

$$\beta' x_t = stationary$$

that is, $x_t \sim I(1)$ but $\beta' x_t \sim I(0)$. The matrix β is called the cointegration matrix and there will be p linear combinations that lead to stationary processes. In order to see the implications of this property, suppose the simplest $I(1)$ model $x_t = x_{t-1} + \varepsilon_t$. We can write this model as

$$\nabla x_t = \Pi x_{t-1} + \epsilon_t \tag{12}$$

and note that if x_t follows the multivariate random walk the value of Π in this equation is cero and this implies no cointegration. However, if the process is a stationary VAR(1) process we can write the model as

$$x_t = (\Pi - I)x_{t-1} + \epsilon_t$$

in this equation Π is a full rank matrix because $\Pi = \phi + I$ where ϕ is the AR matrix that must have eigenvalues smaller than one for the process to be stationary. Thus saying that Π is a full rank matrix implies that x_t follows and VAR(1), all the components are stationary, or they are $I(0)$. A third intermediate possibility is that Π is neither a zero matrix nor a full rank matrix but it has rank p. Let us show that this implies cointegration, that is, in this case some linear combinations of the vector of time series are stationary whereas some others will be nonstationary. To show this property note that if Π has rank p it can be written as

$$\Pi = \alpha\beta'$$

where α and β are $m \times p$ matrices of rank $p < m$. Now if we multiply (12) by β' , we have

$$\nabla\beta' x_t = (\beta'\alpha)\beta' x_{t-1} + \beta'\epsilon_t$$

and calling $z_t = \beta' x_t$ we have that

$$\nabla z_t = \Pi^* z_{t-1} + u_t$$

and Π^* is a full rank matrix and z_t stationary. Thus, the p linear combinations $\beta' x_t$ will be stationary whereas if the matrix $m \times (m-p)$, $\alpha_\perp$ belongs to the null space of α, that is, it verifies $\alpha'_\perp \alpha = 0$, we have that the $m - p$ combinations $\alpha'_\perp x_t$ are nonstationary.

There is a close connection between cointegration and the factor model. Escribano and Peña (1994) showed that the following two proposition are equivalent:

(1) The individual components of x_t are $I(1)$ but there are p cointegration relationships, $\beta' x_t$, that are $I(0)$.

(2) x_t can be written as generated by $m - p$ common factors that are $I(1)$.

Thus, cointegration implies common factor and common nonstationary factors implies cointegration. From the practical point of view if the dimension m is large it is simpler to look for a few factors than for many cointegration relations.

5 Conclusions

As any stationary time series is a sample from some multivariate distribution one could expect that multivariate classical methods will be widely applied in time series. However, in practice

the time series analysis is made without any reference to multivariate analysis by using the special structure implied by the ordering of the observations on time. Some univariate time series identification methods have been based on canonical correlation analysis (see Tiao and Tsay, 1985) but in general the use of multivariate methods in univariate time series is small. However, with vector time series multivariate techniques are of key importance. Discrimination is related to the problem of model selection, clustering methods appear in a natural way when working with large set of time series and methods for dimension reduction are a clear need for practical model building. In fact, it was shown by Peña and Box (1987) that building a VARMA model ignoring the possible common factors is a sure method to look for trouble: the MA and AR parameter matrices are not identified when common factors are present and so we could end up building a very complicated multivariate VARMA model when in fact the data generating process is very simple. Also Tiao and Tsay (1989) have shown the usefulness of linear combinations of the vector of observed time series for model simplification.

We have seen that the discrimination problem is closely related to the model selection problem, and the criteria to choose models can be applied to select the data generating process in discriminant analysis. In time series cluster methods, more research is needed in order to have meaningful procedures that search for useful configurations taking into account the autocorrelation structure and new algorithms need to be developed to implement them. Although research in model simplification and dimension reduction has been very large, still more research is needed in order to compare the advantages and drawbacks of the different procedures available. We expect that this review can stimulate further developments in this area in the future.

Acknowledgment

This work has been partially supported by DGES, grant PB96-0111, and Cátedra BBVA de Métodos para la Mejora de la Calidad.

References

[1] Ahn, S. K. and Reinsel, G. C. (1988) " Nested reduced-rank autoregressive models for multiple time series" *J. Am. Statist. Ass.,* 83, 849-856.

[2] Akaike, H. (1969) " Fitting autorregresive models for prediction" *Annals Inst. Stat. Math.,* 21, 343-347.

[3] Akaike, H. (1974) " A new look at the statistical model identification" *I.E.E.E. Trans. Aut. Contr. AC,* 19, 203-217.

[4] Akaike, H. (1979) " A Bayesian extension of the minimum AIC procedure of autorregressive model fitting" *Biometrika*, 66, 2 237-242.

[5] Alagón, J. (1989) " Spectral discrimination for two groups of time series" *J. Time Ser. Anal.,* 10, 3, 203-214.

[6] Anderson, T. W. and Bahadur, R. R. (1962) " Classification into two multivariate normal distributions with different covariance matrices" *Ann. Math. Stat.* 33, 420-431.

[7] Aoki, M (1990) *State Space Modeling of Time Series*. Springer: Berlin.

[8] Bhansali, R. J. (1993) " Order selection for linear time series models: a review". In T. Subba Rao (eds). *Developments in Time Series Analysis*. Chapman and Hall, London, 50-66.

[9] Bohte, Z., Cepar, D. and Kosmelj, K. (1980). " Clustering of time series" *COMPSTAT 1980, Proceedings in Computational Statistics.* 587-593.

[10] Box, G. and Tiao, G. (1977) " A canonical analysis of multiple time series" *Biometrika*, 64, 355-65.

[11] Brillinger, D. R. (1981) *Time series Data Analysis and Theory*, expanded edition. San Francisco: Holden-Day.

[12] Broemeling, L. D. and Son, M. S. (1987) " The classification problem with autoregressive processes" *Communications in statistics (theory and methods),* 16, 927-936.

[13] Bruce, A. G. and Martin R. D. (1989) " Leave-k-out Diagnostics for time series (with discussion)" *J. R. Statist. Soc. B,* 51, 3, 363-424.

[14] Cavanaugh, J. E. and Shumway, R. H. (1997) " A Bootstrap variant of AIC for state space model selection" *Statistica Sinica*, 7, 473-496.

[15] Cowpertwait, P. S. P., and Cox, T. F. (1992). " Clustering population means under heterogeneity of variance with an application to a rainfall time series problem" *The Statistician,* 41, 113-121.

[16] Chaudhuri, G., Borwankar, J. D. and Rao, P. R. K. (1991) " Bhattacharyya Distance based linear discriminant function for stationary time series" *Communications in statistics (theory and methods),* 20, 2195-2205.

[17] Chaudhuri, G. (1992) " Linear discriminant function for complex normal time series" *Statist. Probab. Lett.,* 15, 277-279.

[18] CIS Extended Data Base Publication.

[19] Dargahi-Noubary, G. R. (1992) " Discrimination between gaussian time series based on their spectral differences" *Communications in statistics (theory and methods),* 21, 2439-2458.

[20] Dargahi-Noubary, G. R. and Laycock, P. J. (1981) " Spectral ratio discriminants and information theory" *J. Time Ser. Anal.,* 2, 2, 71-86.

[21] Engle, R. F. and Watson, M. W. (1981) " A one-factor multivariate time series model of metropolitan wage rates" *J. Am. Statist. Ass.,* 76, 774-781.

[22] Escribano, A. and Peña, D. (1994) " Cointegration and common factors" *J. Time Ser. Anal.,* 15, 577-586.

[23] Galeano, P. and Peña, D. (2001) " A note on Maximum Likelihood Estimation and Optimal Interpolation in Time Series" *Mimeo, Universidad Carlos III de Madrid.*

[24] Gantert, C. (1994) " Classification of trends via the linear state space model" *Biometrical Journal. Journal of Mathematical Methods in Biosciences,* 36, 825-839.

[25] Gersch, W., Martinelli, F., Yonemoto, J., Low, M. D. and McEwan, J. A. (1979) " Automatic classification of electroencephalograms: Kullback-Leibler nearest neighbor rules" *Science,* 205, 193-195.

[26] Geweke, J. F. and Singleton K. J. (1981) " Maximum Likelihood Confirmatory Analysis of Economic Time Series" *International Economic Review*, 22, 37-54.

[27] Hannan, E. J. and Quinn, B. J. (1979) " The determination of the order of an autoregression" *J. R. Statist. Soc. B,* 41, 190-195.

[28] Hannan E. J. and Deistler, M. (1988) *The Statistical Theory of Linear Systems.* New York: John Wiley.

[29] Harvey, A. (1989) *Forecasting Structural Time Series Models and the Kalman Filter* (2nd edn). Cambridge: Cambridge University Press.

[30] Hurvich, C. M. and Tsai, C. L. (1989) " Regression and Time series model selection in small samples". *Biometrika*, 76, 297-307.

[31] Kakizawa, Y., Shumway, R. H. and Taniguchi, M.(1998) " Discrimination and clustering for multivariate time series" *J. Am. Statist. Ass.,* 93, 328-340.

[32] Kedem, B. and Slud, E. (1982) " Time series discrimination by higher order crossings" *Annals of Statistics,* 10, 786-794.

[33] Kullback, S. (1959) *Information theory and statistics*. Smith, Gloucester, MA.

[34] Lai, T. L. and Lee, C. P.(1997) " Information and Prediction criteria for model selection in sthocastic regression and ARMA models" *Statistica Sinica*, 7, 285-309.

[35] Li, T. (1996) " Discrimination of time series by parametric filtering" *J. Am. Statist. Ass.,* 91, 284-293.

[36] Macchiato, M. F., La Rotonda, L., Lapenna, V., Ragosta, M. (1995) " Time modelling and spatial clustering of daily ambient temperature: An application in Southern Italy" *EnvironMetrics,* 6: 31-53.

[37] Marco, V. R., Young, D. M. and Turner, D. W. (1988) " Predictive discrimination for autoregressive processes" *Pattern Recognition Letters,* 7, 145-149.

[38] Molenaar, P. C. M.(1985) " A dynamic factor model for the analysis of multivariate time series" *Psycometrika* 50, 181- 202.

[39] Molenaar, P. C. M, De Gooijer, J. and Schmitz, B. (1992) " A dynamic factor analysis of nonstationary multivariate time series" *Psycometrika* 57, 333- 349.

[40] Peña, D. (1989) Discussion of " Leave-k-out Diagnostics for time series". *J. R. Statist. Soc. B,* 51, 3, 414-415.

[41] Peña, D. and Box, G. (1987) " Identifying a simplifying structure in time series" *J. Am. Statist. Ass.,* 82, 836-843.

[42] Peña, D. and Maravall, A. (1991) " Interpolation, outliers and inverse autocorrelations" *Communications in statistics (theory and methods),* 20, 3175-3186.

[43] Peña, D. and Poncela, P. (1998). " Nonstationary Dynamic factor Analysis" Working paper, Universidad Carlos III de Madrid.

[44] Piccolo, D. (1990) " A distance measure for classifying ARIMA models" *J. Time Ser. Anal.,* 11, 2, 153-164.

[45] Postcher, B. M. and Srinivasan, S. (1994). " A comparison of order determination procedures for ARMA models". *Statistica Sinica*, 4, 29-50.

[46] Rissanen, J. (1986) " Sthocastic complexity and Modelling" *Annals of Statistics*, 14, 3, 1080- 1100.

[47] Schwarz, G. (1978) " Estimating the dimension of a model" *Annals of Statistics*, 6, 2, 461-464.,

[48] Shumway, R. H. (1982) " Discriminant analysis for time series" *Handbook of Statistics,* 2, 1-46.

[49] Shumway, R. H. and Unger, A. N. (1974) " Linear discriminant functions for stationary time series" *J. Am. Statist. Ass.,* 69, 948-956.

[50] Shumway, R. H. and Stoffer, D. S. (2000) *Time Series Analysis and Its Applications.* New York: Springer.

[51] Tiao, G. C. and Tsay, R. S. (1989) " Model specification in multivariate time series" *J. R. Statist. Soc.* B, 51, 157-213.

[52] Tiao, G. C. and Tsay, R. S. (1985) " Use of canonical analysis in time series model identification" *Biometrika* 72, 299-315.

[53] Velu, R. P., Reinsel, G. C. and Wichern, D. W. (1986) " Reduced rank models for multiple time series" *Biometrika*, 73, 105-118.

[54] Walden, A. T. (1994) " Spatial clustering: Using simple summaries of seismic data to find the edge of an oil-field" *Applied Statistics,* 43: 385-398.