A DISTANCE MEASURE FOR CLASSIFYING ARIMA MODELS

By Domenico Piccolo[1]

Universitá di Napoli

First version received June 1986

Abstract. In a number of practical problems where clustering or choosing from a set of dynamic structures is needed, the introduction of a distance between the data is an early step in the application of multivariate statistical methods. In this paper a parametric approach is proposed in order to introduce a well-defined metric on the class of autoregressive integrated moving-average (ARIMA) invertible models as the Euclidean distance between their autoregressive expansions. Two case studies for clustering economic time series and for assessing the consistency of seasonal adjustment procedures are discussed. Finally, some related proposals are surveyed and some suggestions for further research are made.

Keywords. Distance; ARIMA models; clustering time series; seasonal adjustment.

1. INTRODUCTION

The introduction of a metric for linear models is an important step in the application of multivariate statistical methods to the clustering of dynamic structures. In this paper we are concerned with a real situation where autogressive integrated moving-average (ARIMA) models are fitted to a large number of time series for the purpose of forecasting and seasonal adjustment. From an empirical point of view there are many similarities between fitted models, so that classifying them into a small number of groups may be useful for detecting a few representative models. Moreover, models which appear dissimilar because of the presence of different orders and/or parameters may in fact have similar properties. Thus it is worth investigating a measure of dissimilarities between models which is appropriate to their applications.

In a different context, the problem of choosing a statistical procedure for the seasonal adjustment of several economic time series provides further motivation for our proposal. It is well known that some widely used procedures such as X-11 and X-11-ARIMA give better results for series generated by models which are not very dissimilar from the optimal model implied by the procedure itself. This problem can also be reduced to a comparison of time series models.

It is important to stress that the comparison of models cannot be uniquely defined since any distance is an arbitrary measure of diversity which satisfies well-known axioms. Thus we shall choose a method which is effective with respect to our purpose and is statistically consistent, simple to calculate and generally applicable.

0143-9782/90/02 153–12 $02.50/0

© 1990 Domenico Piccolo

In the next section we define a distance measure for the class of ARIMA admissible models. Then in sections 3 and 4 we present some examples of classifying economic time series and discuss the implications of the approach for the comparison of seasonal adjustment procedures. Finally we briefly review related proposals and make some suggestions for further research.

2. DEFINITION AND PROPERTIES OF A DISTANCE MEASURE

Let a_t be a Gaussian white noise and Z_t a zero-mean stochastic process such that $Z_t \sim \text{ARIMA}(p, d, q) \times (P, D, Q)_s$ and hence, following standard notation (Box and Jenkins 1976), $\varphi(B)Z_t = \theta(B)a_t$. If $\theta(B) = 0 \rightarrow |B| > 1$, then $Z_t \in \mathfrak{L}$ where $\mathfrak{L}$ is the class of ARIMA invertible models. An equivalent statement is that if $Z_t \in \mathfrak{L}$ then $W_t = \nabla^d \nabla_s{}^D Z_t$ is an admissible (i.e. stationary and invertible) Gaussian autogressive moving-average (ARMA) process.

Now, it is well known that $Z_t \in \mathfrak{L}$ then

$$Z_t = \bar{Z}_{t-1} + a_t$$

$$\bar{Z}_{t-1} \perp a_t \qquad \bar{Z}_{t-1} = \sum_{j=1}^{\infty} \pi_j Z_{t-j} = \{1 - \pi(B)\} Z_t$$

where the AR(∞) operator is defined by $\pi(B) = \varphi(B)\theta^{-1}(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots$.

In this respect, we recall that *given initial values and known orders*, any process $Z_t \in \mathfrak{L}$ is fully characterized by the sequence $\boldsymbol{\pi}' = (\pi_1, \pi_2, \ldots)$ which specifies completely the distribution of the processes W_t and Z_t when a_t is a Gaussian white noise process. In a sense, the $\boldsymbol{\pi}$ sequence conveys all useful information about the stochastic structure of the process since any other information needed to specify Z_t (except for initial values) is just a_t, which is unpredictable at time $t - 1$. Then a measure of structural diversity between $X_t \in \mathfrak{L}$ and $Y_t \in \mathfrak{L}$ can be obtained by comparing the respective $\boldsymbol{\pi}$ sequences.

The previous discussion leads to the assignment of a metric on $\mathfrak{L}$ by the distance

$$d(X,Y) = \left\{ \sum_{j=1}^{\infty} (\pi_{j,x} - \pi_{j,y})^2 \right\}^{1/2} . \tag{1}$$

Since $\Sigma\pi_j$, $\Sigma|\pi_j|$ and hence $\Sigma\pi_j^2$, are well-defined quantities, it is straightforward to show that $d(X,Y)$ always exists for any process in $\mathfrak{L}$ and satisfies the classical properties of a distance, i.e. non-negativity, symmetry and triangularity. It is useful to list some properties of the distance (1).

(i) The proposed distance is completely general and can be computed even if some orders of the models to be compared are zero. In fact, the metric depends upon the coefficients of the AR(∞) operator which always converges. Its interpretation could be that the $\boldsymbol{\pi}$ sequence uniquely determines the forecast function for future values given present and past values.

(ii) The metric (1) does not take account of the residual variance since this is purely a scale parameter and is not relevant to the kind of comparisons which are required.

(iii) The zero element (origin) of $\mathfrak{L}$ is any white noise process which is characterized by the null $(0,0,\ldots)$ sequence. Then, the distance between X_t and the origin O is the norm of X_t, i.e.

$$d(X,\mathrm{O}) = (\Sigma\pi_{j,x}^2)^{1/2} = \|X\|.$$

(iv) There is an isometry between the class of non-seasonal ARIMA models and the corresponding class of seasonal ARIMA models. Thus, if X_t and Y_t are associated with operators $\pi_x(B)$ and $\pi_y(B)$ respectively, and X_t' and Y_t' are associated with operators $\pi_x(B^s)$ and $\pi_y(B^s)$, then $d(X,Y) = d(X',Y')$.

(v) If the diameter of a class $\mathcal{B} \subset \mathfrak{L}$ is defined as

$$\operatorname{diam}(\mathcal{B}) = \sup\left\{d(X,\ Y);\ X_t \in \mathcal{B},\ Y_t \in \mathcal{B}\right\},$$

it can be shown that, for a fixed p, the class of AR(p) processes is bounded. Conversely, the MA class is unbounded since the norm of MA processes tends to infinity as the parameters approach the borderline of the non-invertibility region. For instance, the MA(1) process $Z_t = a_t - \theta a_{t-1}$ possesses a sequence $\pi_j = -\theta^j$ $(j = 1, 2, \ldots)$ whose norm is $\{\theta^2/(1-\theta^2)\}^{1/2}$ which tends to $+\infty$ as $|\theta| \rightarrow 1$. In contrast, a subclass of IMA(1, 1) processes is bounded as we shall show later.

(vi) If we restrict our attention to the admissible ARMA processes, we can define a dual metric by

$$\delta(X,Y) = \left\{\sum_{j=1}^{\infty}(\psi_{j,x} - \psi_{j,y})^2\right\}^{1/2}$$

where the ψ sequence defines the MA(∞) operator as $\psi(B) = \theta(B)\psi^{-1}(B) = \pi^{-1}(B)$. However, this metric is inadequate for integrated processes, and hence it can not be applied to the analysis of non-stationary economic time series.

We discuss a simple example of the proposed distance between two ARIMA(1, 0, 1) processes $X_t \in \mathfrak{L}$, $Y_t \in \mathfrak{L}$. In standard ARIMA notation, it is straightforward to obtain $\pi_{j,x} = (\phi_x - \theta_x)\theta_x^{j-1}$, $\pi_{j,y} = (\phi_y - \theta_y)\theta_y^{j-1}$ and deduce

$$d^2(X,Y) = \frac{(\phi_x - \theta_x)^2}{1-\theta_x^2} + \frac{(\phi_y - \theta_y)^2}{1-\theta_y^2} - 2\frac{(\phi_y - \theta_y)^2}{1-\theta_x\theta_y}.$$

This result can be specialized to particular cases (for instance, the comparison of AR(1) or ARIMA (0, 1, 1) models). More interestingly, when $\phi_x = \phi_y = 1$ we are comparing two IMA(1, 1) models and the formula reduces to

$$d^2(X,Y) = \frac{2\,(\theta_x - \theta_y)^2}{(1+\theta_x)(1+\theta_y)(1-\theta_x\theta_y)}.$$

This shows that, even if $\theta_y = 1$ say, $d^2(X, Y) = (1 - \theta_x)(1 + \theta_x)^{-1}$ is finite for $\theta_x > 0$, and thus the important subclass of IMA(1, 1) processes with positive θ, corresponding to the exponentially weighted MA predictor, is bounded with respect to our metric even though the class of MA(1) processes is unbounded.

3. CLUSTERING TIME SERIES MODELS

The definition of a distance between time series models immediately allows applications to clustering algorithms. As an example we discuss a real case study aimed at forecasting monthly industrial production indices in Italy. The purpose of the analysis is to investigate a possible similarity in the behaviour of industrial production in different sectors. If this were the case, it would mean that some areas possess a similar dynamic structure during economic cycles and/or that they react similarly to economic or policy interventions. The features that we are comparing are a mixture of seasonality and inertia components. In fact, the production series present a seasonal component which is superimposed on the long term behaviour.

The building of models for the sectorial time series leads to the class of structures

$$\phi_i(B)\nabla\nabla_{12} \log X_{i,t} = \theta_i(B)a_{i,t} \qquad (i = 1, 2, \ldots, 14)$$

whose estimates (which we omit for brevity) were used to obtain the π sequences and the distance matrix (Table I) by using metric (1). Then, by a complete linkage method, we obtain the dendrogram shown in Figure 1. The data are clearly split into the clusters $\mathscr{C}_1 = (\mathrm{N, O, P, E})$ and $\mathscr{C}_2 = (\mathrm{I, F, D, L, C, B, H, G})$, except for series A and M which have unusual production

TABLE I

Distance Matrix for the ARIMA Models of Industrial Production Series

	A	B	C	D	E	F	G	H	I	L	M	N	O	P
A	0.00													
B	2.13	0.00												
C	1.99	0.54	0.00											
D	1.93	0.45	0.45	0.00										
E	1.96	0.52	0.69	0.61	0.00									
F	1.76	0.55	0.43	0.25	0.61	0.00								
G	1.75	0.68	0.59	0.47	0.72	0.35	0.00							
H	1.72	0.66	0.65	0.58	0.67	0.44	0.54	0.00						
I	1.78	0.54	0.45	0.22	0.61	0.11	0.39	0.47	0.00					
L	1.86	0.51	0.45	0.21	0.62	0.16	0.40	0.50	0.23	0.00				
M	1.46	1.03	1.02	1.03	0.79	0.86	0.89	0.76	0.91	0.94	0.00			
N	1.77	0.62	0.64	0.63	0.44	0.49	0.58	0.52	0.56	0.51	0.56	0.00		
O	1.82	0.66	0.69	0.70	0.43	0.58	0.66	0.59	0.64	0.58	0.57	0.09	0.00	
P	1.73	0.69	0.73	0.72	0.40	0.59	0.66	0.57	0.64	0.63	0.45	0.18	0.18	0.00

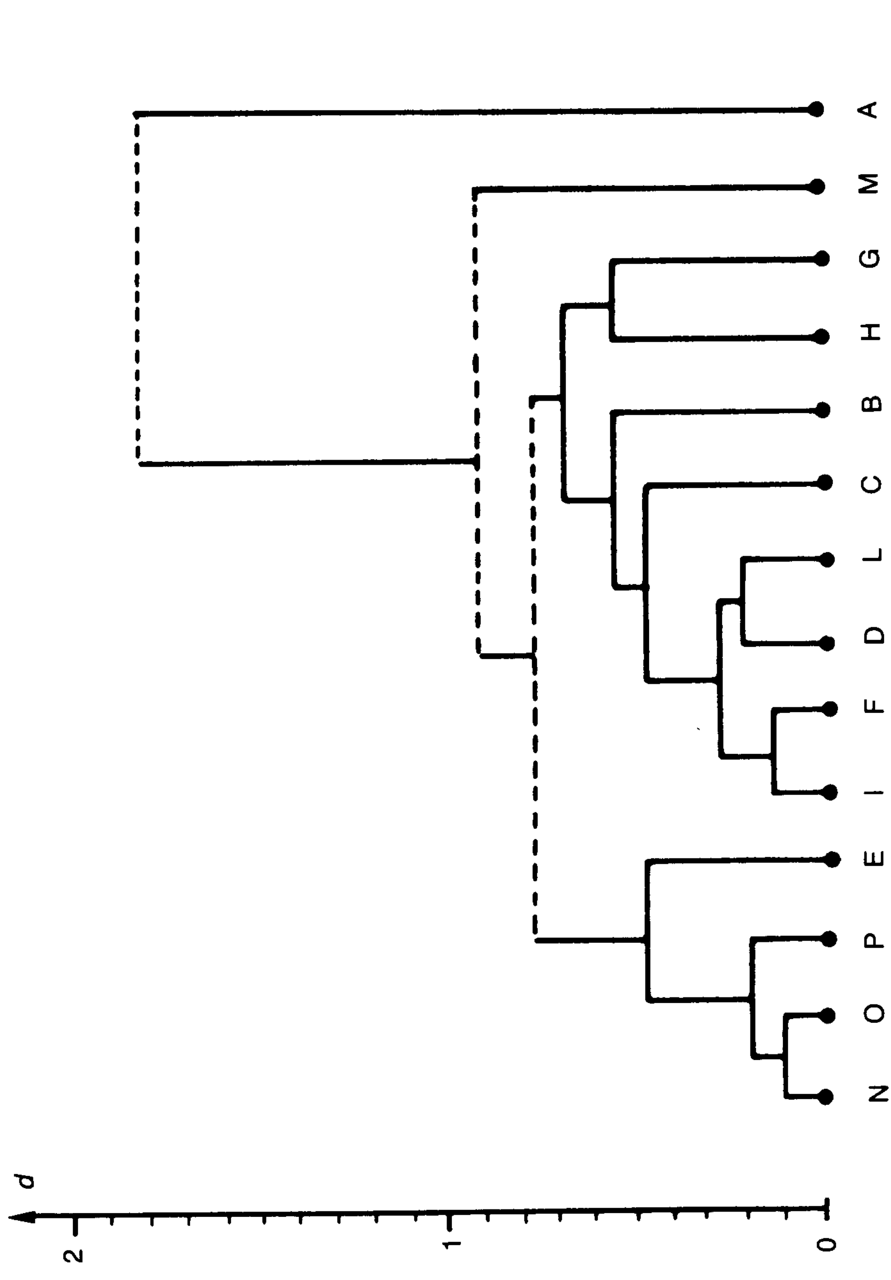


FIGURE 1. Dendrogram for 14 ARIMA models of industrial production series in Italy.

cycles. From a close comparison of the dendrogram and the structure of the ARIMA models, we can infer that the dissimilarity between the classes mainly originates in the nature of the seasonal component (more regular for the series in $\mathcal{C}_2$) and the inertia component (with a longer cycle and a simpler structure in $\mathcal{C}_1$). This behaviour may be induced by the effects of exogenous economic shocks in the composition of stock and technological factors which can strongly affect the production output of each sector.

An alternative derivation of these results and a well-resolved graphic representation can be obtained by applying the classical solution of multidimensional scaling (MDS) to the distance matrix (Table I) which solves, in the metric case, in a principal coordinate analysis (Mardia *et al.*, 1979). In MDS we are searching for a configuration of points (i.e. the time series models) in a convenient space where the interpoint distance matrix reproduces the original matrix as closely as possible. This solution can depict the projections of the points in fewer dimensions (m = 1,2,3), providing immediate information about their relationships (closeness, similarity, anomaly, etc.). When $m = 2$, the MDS solution for the models of industrial production in Italy confirms the existence of clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ as well as the anomalous structures A and M previously identified. The graphical representation of the data (Figure 2) also shows excellent agreement between the original and the reduced representation (confirmed by the Mardia index $\alpha_2 = 99.1\%$).

Finally, this approach can be used to select characteristic series representing the different structures. For instance, in the clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ we can choose the series O and F since they satisfy the criterion

$$\sum_{j\in\mathcal{C}_k} d^2(X_{(0)}, X_{(j)}) = \text{minimum} \qquad (k = 1,2).$$

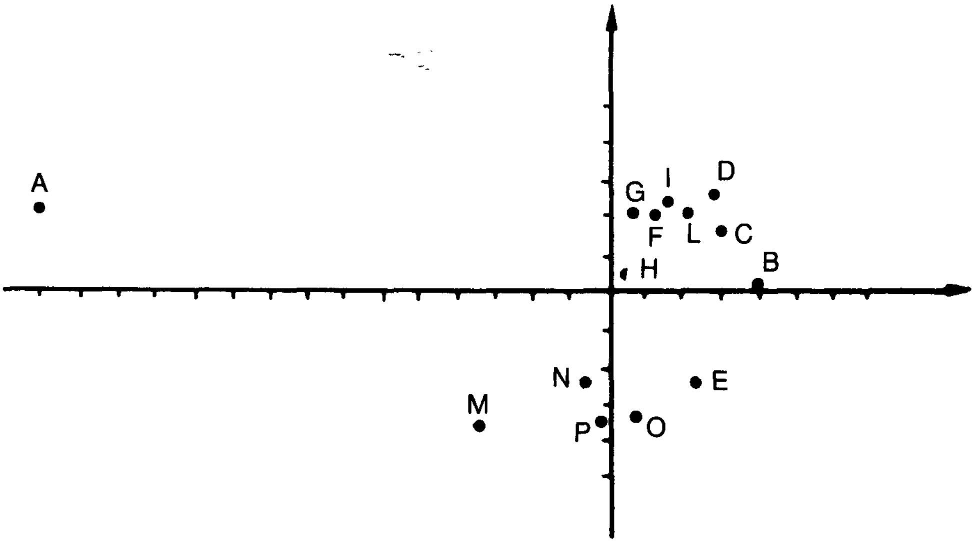


FIGURE 2. Multidimensional scaling for the ARIMA models.

In this way, we conclude that (O, F, A, M) are the most significant patterns in our set and could be used for subsequent analyses.

4. COMPARISON OF SEASONAL ADJUSTMENT PROCEDURES

As shown by Cleveland and Tiao (1976), the linear filter of the X-11 procedure is optimal when the data are generated by a particular ARIMA model. This fact makes the comparison of ARIMA models a relevant issue in the debate on seasonal adjustment, as confirmed by Bell and Hillmer (1984).

In order to simplify the problem, the main question 'Is the procedure $\mathcal{P}_0$ adequate for analysing the x_t series?' can be transformed into the more definite question 'Is the process X_t generating the series x_t *sufficiently close* to the process X_t^0 implied by the procedure $\mathcal{P}_0$?'. Thus, as shown in Figure 3, the problem can be solved by assessing a measure of closeness between two ARIMA models: the first is implied by the procedure, and the second is deduced from the data. We are therefore suggesting that a procedure $\mathcal{P}_0$ is adequate for a series x_t when the distance $d(X, X^0)$ is 'small enough'. Of course, when comparing two procedures (or two variants of the same procedure), we prefer $\mathcal{P}_1$ to $\mathcal{P}_2$ for the series x_t when $d(X, X^1) < d(X, X^2)$.

In this regard it is worth noting that the distance (1) relies on important aspects of seasonal decomposition since it is a function of the π weights. As Hillmer *et al.* (1983) and Maravall (1984) have pointed out, the decomposition of ARIMA models is quite sensitive to the MA parameters: even a small variation in the θ values can produce quite different decomposition results. In fact, it is well known that the filter producing $\hat{S}_t$ from $Z_t = S_t + N_t$ is proportional to

$$w(B) = \frac{\pi(B)\pi(B^{-1})}{\pi_s(B)\pi_s(B^{-1})}$$

where the π operators are the AR formulations for the Z_t and S_t processes respectively. Thus comparing π weights by means of the distance (1) is also a way of comparing different results in the decomposition analyses. This fact strongly supports the previous proposal of comparing seasonal adjustment procedures by means of a distance between ARIMA models, as will be comfirmed in the next example.

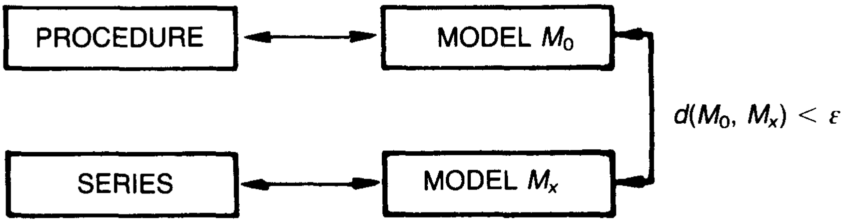


FIGURE 3. Distance as a check of adequacy for seasonal adjustment procedures.

Cleveland and Tiao (1976) seasonally adjusted the 'airline data' x_t (Box and Jenkins, 1986) and the 'telephone data' y_t (Thompson and Tiao, 1971) using the standard options of the X-11 procedure. They concluded that the standard procedure gives poor results for the telephone data. The proposed models for the logged series and for the X-11 procedure with a 13-term Henderson filter were as follows:

$$\begin{aligned}
&\textit{airline} && \nabla\nabla_{12}X_t = (1 - 0.4\,B)(1 - 0.6\,B)\,a_t \\
&\textit{telephone} && (1 - 0.490B)(1 - 1.005B)\,Y_t \\
& && \qquad = (1 - 0.230B^9 - 0.334B^{12} - 0.170B^{13})b_t \\
&X\!-\!1\,1 && \nabla\nabla_{12}Z_t = \theta(B)c_t
\end{aligned}$$

where $\theta(B)$ is a polynomial of degree 25 (Figure 4). Cleveland and Tiao adjusted the telephone data without any non-standard options although, as reported by Hillmer (1982), this series should be corrected daily prior to being seasonally adjusted. In Figure 4 (Corduas, 1984) we show the π coefficients for the models, their distance matrix and a geometric characterization, which is particularly simple, in this example. Thus we confirm Cleveland and Tiao's findings using a descriptive measure and a geometric device.

Finally, the previous argument suggests a simple strategy for seasonal adjustment analysis of a large number of time series:

(i) fit ARIMA models to the time series;

(ii) measure the distance between the fitted models and that implied by the seasonal adjustment procedure with standard options;

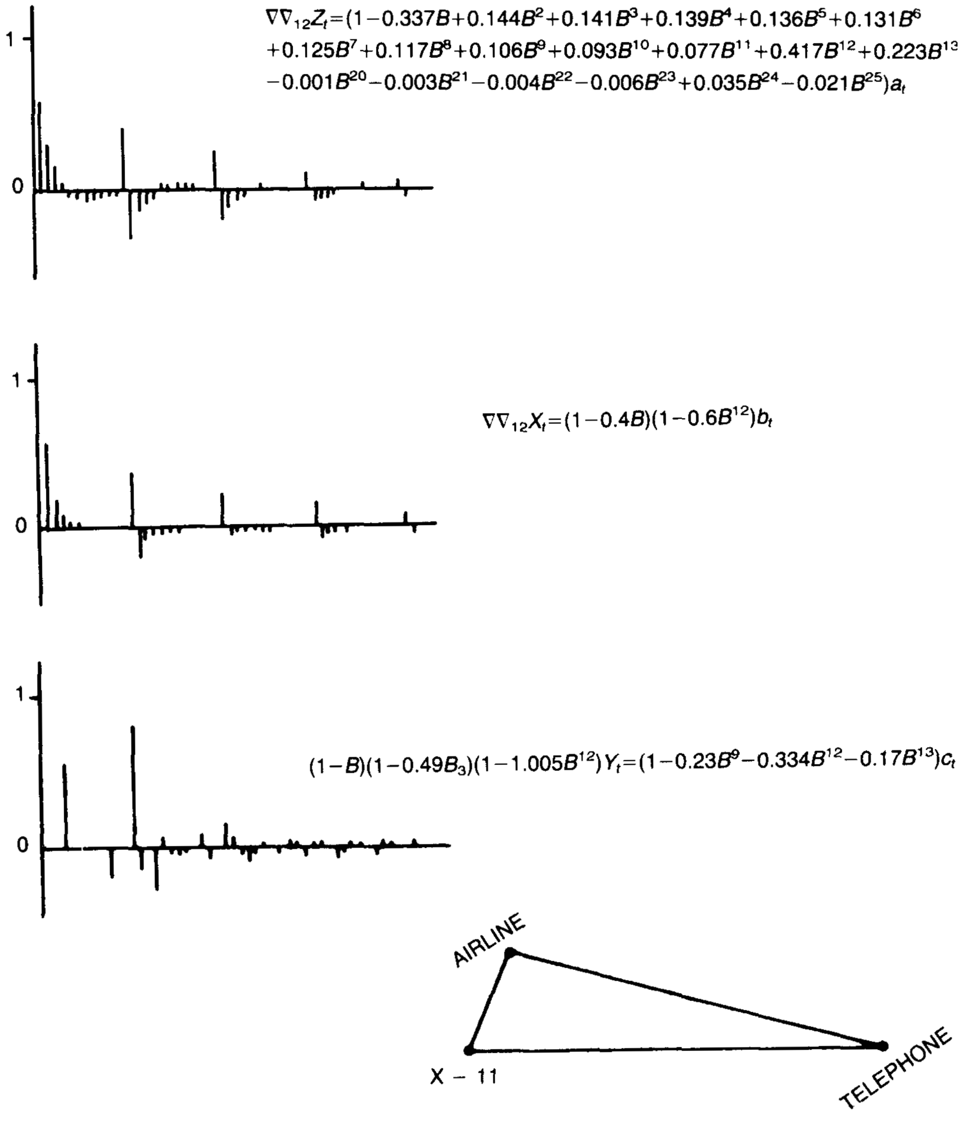
(iii) reject the procedure (or check for different options or try a model-based approach) when the distance in (ii) is greater than a target value established by experimental trials, simulations, further statistical investigations and so on.

5. DISCUSSION

The problem of classifying and discriminating time series is not new, and therefore it is useful to discuss related proposals briefly in order to assess the merits and shortcomings of the distance previously introduced.

Some proposals are derived from sample autocorrelations (as in Bohte *et al.* (1980) who weighted the normalized absolute deviations between estimated autocorrelations) and cross-correlations (as in Zani (1983) who proposed a lag distance for clustering common time lags in the estimated cross-correlations). In this respect, it is worth noting that Euclidean distances on autocovariances and inverse autocovariances are immediately related to the distances between spectra and inverse spectra respectively.

Following a more formal approach, it has been found that the Kullback–Leibler divergence for k-variate Gaussian processes with equal means reduces to



$$\nabla\nabla_{12}Z_t=(1-0.337B+0.144B^2+0.141B^3+0.139B^4+0.136B^5+0.131B^6$$
$$+0.125B^7+0.117B^8+0.106B^9+0.093B^{10}+0.077B^{11}+0.417B^{12}+0.223B^{13}$$
$$-0.001B^{20}-0.003B^{21}-0.004B^{22}-0.006B^{23}+0.035B^{24}-0.021B^{25})a_t$$

$$\nabla\nabla_{12}X_t=(1-0.4B)(1-0.6B^{12})b_t$$

$$(1-B)(1-0.49B_3)(1-1.005B^{12})Y_t=(1-0.23B^9-0.334B^{12}-0.17B^{13})c_t$$

$d(.,.)$	X − 11	Airline	Telephone
X − 11	0.0		
Airline	0.279	0.0	
Telephone	1.056	1.017	0.0

FIGURE 4. Comparisons of airline, telephone and X-11 structures; π weights, distance matrix and geometric representation. (From Corduas, 1984)

$$J(X,Y) = \mathrm{tr}(\Sigma_x{}^{-1}\,\Sigma_y + \Sigma_y{}^{-1}\,\Sigma_x) - 2k$$

and the limiting discriminating rate of J when $k \to +\infty$ is equivalent to

$$J'(X,Y) = \int_0^{\pi} \{f_y(\omega)f_x^{-1}(\omega) + f_x(\omega)f_y^{-1}(\omega) - 2\}\, d\omega$$

(see Sh'ımway and Unger (1974) and Dargahi-Noubary and Laycock (1981) for applications). The J' rate is invariant for any common linear filter to X_t and Y_t and thus can also be used for comparing ARIMA models with the same order of differencing. Moreover, it can be shown that $J'(X, Y) = J'(X^{\circ}, Y^{\circ})$ where X_t°, Y_t° are the inverse processes of X_t, Y_t.

Since the J' measure is a function of both the π and ψ sequences, it is obviously related to our metrics, but we should note the following.

(i) The divergence is not a distance (since the triangular inequality does not hold) and consequently we need some rescaling for applying metric multivariate methods.

(ii) We cannot compare a set of ARIMA models by J' measures if at least one has a different order of differencing.

(iii) The discriminating rate seems to be less sensitive to the model distinction, which we believe to be important in real applications. For instance, J' for two AR(1) processes equals J' for two MA(1) processes with the same parameters, whereas if the parameters are not small we prefer to have a greater distance between the MA(1) processes.

We are now in a position to list some relative merits of the distance (1). First, the d metric can be computed for the comparison of any ARIMA models. Second, it depends on the AR structure which is directly relevant in forecasting and decomposition analyses. Third, the proposed distance can discriminate effectively between AR and MA models. In contrast, a limitation of the proposal seems to be the need for *ad hoc* ARIMA modelling of several time series. This fact could be overcome by automatic modelling of AR structures by means of a definite criterion (Akaike information criterion, CAT, Schwarz criterion etc.). Since the fitting of sequential AR models is quite simple and efficient (via the Durbin–Levinson algorithm for example), an automatic clustering of time series can be obtained by grouping fitted AR models directly from a distance matrix computed using our method.

6. CONCLUSION

The general philosophy of this paper can be summarized as follows. Given a collection of logically connected time series, the best way to compare them is via ARIMA modelling which is the most effective and parsimonious parameterization for general analyses. In this class, the π sequence conveys all the useful information about the stochastic dynamic structure of the process. Thus, a comparison using π weights seems to be a simple, well-defined and powerful method for introducing a distance into the class of invertible

ARIMA models. Since distance is a ubiquitous concept in statistics, its introduction extends the operational capabilities of ARIMA modelling, allowing for clustering, discriminating, choosing from a set, selecting anomalous structures etc. In this area, there is room for more detailed analyses mainly from an inferential viewpoint, However, the relevant areas of application and the preliminary results reported here confirm the usefulness of the approach.

ACKNOWLEDGEMENTS

We gratefully acknowledge an anonymous referee whose numerous suggestions greatly improved the presentation of the paper, and the constant cooperation of Dr Marcella Corduas, who developed the ideas and the implications of a distance measure for time series. The concept of a metric between ARIMA models has recently been proposed by Professor Daniel Peña, University of Madrid, as a measure for detecting influential observations in time series. We proposed the distance (1) for the first time at a seminar at the Department of Statistics, University of Wisconsin, Madison, in September 1983. A preliminary version of this paper appeared as Technical Report 21/1984, Desec Project, Centro Specializzazione e Ricerche, Portici, Naples, with the partial support of Consiglio Nazionale delle Ricerche.

NOTES

[1] Present address: Centro di Specializzazione e Ricerche, Via Universita' 96, 80055 Portici (Napoli), Italy.

REFERENCES

Bell, W. R. and Hillmer, S. C. (1984) Issues involved with the seasonal adjustment of economic time series (with discussion). *J. Bus. Econ. Statist.* 2 (4), 291–349.

Bohte, Z., Cepar, D. and Kosmelij, K. (1980) Clustering of time series. *Compstat 80*, pp. 587–93.

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control* (revised edn). San Francisco, CA: Holden-Day.

Cleveland, W. P. and Tiao, G. C. (1976) Decomposition of seasonal time series: a model for the Census X-11 program. *J. Am. Statist. Assoc.* 71, 581–87.

Corduas, M. (1984) Distanza tra Modelli: problemi metodologici e indici statistici. *Statistica* 44(3), 513–24.

Dragahi-Noubary, G. R. and Laycock, P. J. (1981) Spectral ratio discrimination and information theory. *J. Time Ser. Anal.* 2(2), 71–86.

Hillmer, S. C. (1982) Forecasting time series with trading day variation. *J. Forecast.* 1, 385–95.

——, Bell, W. R. and Tiao, G. C. (1983) Modeling considerations in the seasonal adjustment of economic time series, In *Applied Time Series Analysis of Economic Data* (ed. A. Zellner). Washington, DC, pp. 74–100.

Maravall, A. (1984) Model-based treatment of a manic-depressive series. In *Computer Science* *and Statistics: Symposium on the Interface*. Amsterdam: North-Holland.

Mardia, K. A., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.

Shumway, R. H. and Unger, A. N. (1974) Linear discriminant functions for stationary time series. *J. Am. Statist. Assoc.* 69, 948–56, and papers cited therein.

Thompson, H. E. and Tiao, G. C. (1971) Analysis of telephone data: a case study of forecasting seasonal time series. *Bell J. Econ. Manag. Sci.* 2, 515–41.

Zani, S. (1983) Osservazioni sulle serie storiche multiple e l'analisi dei gruppi. In *Analisi Moderna delle Serie Storiche* (ed. D. Piccolo) Milan: Angeli, pp. 263–74.