Statistical issues on the AR metric in time series analysis[1]

La metrica Autoregressiva nell'analisi statistica delle serie storiche

Domenico Piccolo [2]

Dipartimento di Scienze Statistiche
Università degli Studi di Napoli Federico II, E-mail: domenico.piccolo@unina.it

Abstract: The AR metric was firstly introduced in 1983 as a tool for choosing a representative element from a large collection of time series and for clustering temporal data. The proposal has been extended to many contexts and has raised currently an increasing interest in time series data mining. The main results concerning the AR metric, its asymptotic distribution and some operational and methodological issues are presented. A comparison of the merits of this distance criterion and some caveats about its usage in practical applications conclude the paper.

Keywords: AR metric, Distance between $ARFIMA$ processes, Asymptotic distribution.

1. Introduction

In the last decade, many factors have stimulated an increasing interest towards dissimilarity measures for time series data; in particular:

- thousands of financial time series are regularly collected and their analysis requires fast and effective methods for clustering, discriminating and for selecting representative elements of homogenous groups of series;
- time series data mining is a research area where an effective metric is needed;
- computers speed and numerical algorithms efficiency allow monitoring and control of a large amount of dynamic phenomena in real time.

In this line, it is useful to deepen the foundations of the Autoregressive (AR) metric as a simple and effective measure of dissimilarity among time series, in order to spread current researches and to suggest new developments. In this regard, since we will focus mainly on the AR metric, we refer to Maharaj (2000), Liao (2005), Corduas (2003, 2007), Corduas and Piccolo (2007) for extensive reviews of the current research on clustering and discrimination of time series.
This paper is organized as follows: after a brief outline of the genesis of the AR metric, the rationale of the proposal, its main properties and a typical example are discussed in sections 3-5. Then, the main result about the distribution of the maximum likelihood estimator of the metric is presented. Some applications are illustrated in sections 7-8. In section 9, a spectral decomposition of the metric is introduced and in section 10 some alternatives to this metric are briefly examined. Some concluding remarks end the paper.

[1]This research has been supported by Dipartimento di Scienze Statistiche and CFEPSR, Portici.

[2]Address of correspondence: Dipartimento di Scienze Statistiche, Via Leopoldo Rodinò 22, 80138, Napoli.

2. Genesis of the AR metric

The official birth of the AR metric –as a distance measure among any $ARIMA$ model– was on 14 September, 1983 when I was invited by prof. George Box to give a Seminar on *"A Distance Measure among ARIMA Models"*, during my visit at the Department of Statistics, University of Wisconsin, Madison (USA).
Indeed, at that time, I had already experienced this problem in a previous research involving the construction of an $ARIMA$ model for the monthly wholesale prices index in Italy (Piccolo, 1972). In this circumstance, I compared two models for P_t:

$$\begin{array}{lllll} \mathcal{M}_1: & (1-1.30\,B+0.40\,B^2) & \nabla\,\log(P_t) & = (1-0.12\,B)\,a_t \\ \mathcal{M}_2: & (1-0.45\,B) & \nabla^2\,\log(P_t) & = (1-0.45\,B+0.05\,B^2)\,b_t \end{array}$$

by plotting the Autoregressive coefficients (π_j) *obtained from the* $ARIMA$ *operators*. In fact, I was comparing an $ARMA(2,1)$ model for the *inflation rate* $\nabla\,\log(P_t)$ with an $ARMA(1,2)$ model for the *acceleration rate* $\nabla^2\,\log(P_t)$ of the wholesale prices.
However, this genuine idea came out as an explicit methodological issue only during the *DESEC* project (1982-85), a national research aimed at leading the Italian public Institutions towards a shared and unique seasonal adjustment procedure.
In that context, I faced the following problem: "How to choose few *representative series* from a large data set in order to reduce time and costs of statistical analyses?" Formally, the problem is:

"Given a class of time series $X_{j,t} \in \mathcal{C}$, select a series X_t^* such that:

$$distance\,(X_{j,t},\,X_t^*) = min! \quad \forall X_{j,t} \in \mathcal{C}\,."$$

Since the series were generated from different areas and had different time lengths, the investigation was aimed at introducing a completely general metric for time series. In fact, this question motivated the study of a metric defined on the class of all real time series object of a statistical analysis.
The journal "*Statistica*" published the first paper on the new proposed metric (Piccolo, 1984a) and the distribution of the maximum likelihood (ML) estimator of the metric for pure AR processes was presented at the ASA Conference in Washington (Piccolo, 1989). Then, after a long revision process, a contribution on this topic (submitted in 1986) appeared in "*Journal of Time Series Analysis*" (Piccolo, 1990) and became the standard international reference for the AR metric.
It is worth considering that other criteria are based on AR coefficients. In speech recognition analyses, the AR coefficients (=LPC, Linear Predictor Coding) were used in order to synthesize the voice signals for a specific word (Gray and Markel, 1976); however, the method was finalized *only* to fitting and testing purposes. In this respect, De Souza (1977) and Thomson and De Souza (1985) introduced the Mahalanobis distance between AR models and derived its distributional properties. Many recent medical applications in ECGs and EEGs classifications still refer to this kind of approach (Kosěc, 2000; Ge et al., 2002).
For many years, the AR metric has been used in several fields and many statistical papers have been published (mainly in Italy, Spain and USA). In this respect, a significant progress on these topics has been achieved by Corduas (1996, 2000b) who assessed distributional properties of the AR metric.

3. The AR metric: logical and statistical foundations

Dictionaries defines *distance* as: "the property created by the space between two objects or points; the size of the gap between two places; the interval between two times; ..." Instead, *metric* is: "a system of related measures that facilitates the quantification of some particular characteristic; a function of a topological space that gives, for any two points in the space, a value equal to the distance between them, ..."
Then, a preliminary remark applies: distance is a *concept* that may be transformed into an operational tool by means of some conventional *measure*. Thus, it is correct to argue in favor or against a specific metric, since a metric is strictly determined by the purpose of a research: different metrics are acceptable if different objectives are pursued.
When *objects* to be compared are *time series*, an effective metric has to satisfy the following requirements:

- it is simple to compute and provides meaningful interpretation of data;;
- it is dependent on the stochastic structure of the generating process;
- it is implemented for both stationary and non-stationary time series;
- it is not dependent on the length and on the unit of measurement of the time series.

In addition, it is relevant to point out that our approach stems from the fundamental paradigma which relates a *time series* to the generating *stochastic process* via a *statistical model*. As a matter of fact, a metric on the space of admissible statistical models (which is able to account for almost any real time series) has two important features: it is meaningful for applications and, also, it is robust with respect to the presence of anomalous behaviour in the data.
From a statistical point of view, the AR metric is justified by a fundamental theorem: "for *any* stationary process with a continuous spectrum $f(\omega)$ there exists a finite order $AR(p)$ whose spectrum $f_{AR}(\omega)$ is as close as possible in absolute value to $f(\omega)$ uniformly on $[-\pi, \pi]$": Brockwell and Davis (1991, 130-133).
The theorem is extended to moving average (MA) processes and, for numerical efficiency, to mixed $ARMA$ structures. Thus, the AR operator provides the simplest and effective approximation for any stationary process or for any process that may be transformed to stationary form. In fact, the theorem applies to both linear and not linear processes.
Specifically, given the process X_t, we consider the *ARIMA* model for $Z_t = g(X_t) - f_t$, where Z_t is obtained after g-transforming X_t (in order to reduce asymmetries, improve Gaussianity and take into account of non-linearities) and after removing any deterministic components f_t (such as trading days, calendar effects, outliers and mathematical functions of time, including constants).
Hereafter, we will refer to Box and Jenkins (1970) standard notation and we will assume that Z_t is a zero mean invertible $ARIMA$ process defined as:

$$\varphi(B)\nabla^d\nabla_s^D Z_t = \vartheta(B)a_t,$$

where a_t is a White Noise (WN) process with constant variance $\sigma_a^2 < +\infty$. If a_t is a Gaussian process, given the initial values, the operators $\varphi(B), \vartheta(B)$ and the WN variance σ_a^2 characterize the probability distribution of the process Z_t.
The polynomials $\varphi(B) = \phi(B)\Phi(B^s)$ and $\vartheta(B) = \theta(B)\Theta(B^s)$, for any $s \geq 0$, have no common factors, and all the roots of $\varphi(B)\vartheta(B) = 0$ lie outside the unit circle. Thus, we denote by $\mathcal{L}$ the class of *invertible* linear stochastic processes $Z_t \sim ARIMA$ such that the MA operators have *all the roots outside the unit circle*.

The *invertibility* assumption ensures the absolute (and squared) convergence of the π_j coefficients so that Z_t can be represented in terms of its past values according to:

$$\pi(B)Z_t = a_t \quad \Longleftrightarrow \quad Z_t = \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \cdots + a_t \,,$$

where: $\pi(B) = (1-B)^d(1-B^s)^D\varphi(B)\vartheta^{-1}(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$.
For any $Z_t \in \mathcal{L}$, the *forecast function* $F_t = \mathbb{E}\{Z_t \mid Z_{t-1},\, Z_{t-2}, \dots\}$ is given by:

$$F_t = \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \dots \,,$$

whereas a corresponding orthogonal representation is: $Z_t = F_t + a_t,\ F_t \perp a_t,\ \forall t$.
Let $X_t \in \mathcal{L}$ and $Y_t \in \mathcal{L}$ be invertible processes whose forecast functions may be expressed via the corresponding AR coefficients:

$$\boldsymbol{\pi}_x = (\pi_{1,x},\, \pi_{2,x}, \dots, \pi_{j,x} \dots,)'; \quad \boldsymbol{\pi}_y = (\pi_{1,y},\, \pi_{2,y}, \dots, \pi_{j,y} \dots,)' \,.$$

Then, given the absolute convergence of the π-sequences in $\mathcal{L}$, Piccolo (1984a, 1990) introduced a metric between two $ARIMA$ processes, X_t and Y_t, with given orders, as the Euclidean distance between the π-weights of their corresponding $AR(\infty)$ formulation:

$$d(X_t, Y_t) \;=\; [(\boldsymbol{\pi}_x - \boldsymbol{\pi}_y)'(\boldsymbol{\pi}_x - \boldsymbol{\pi}_y)]^{\frac{1}{2}} \;=\; \sqrt{\sum_{j=1}^{\infty} (\pi_{j,x} - \pi_{j,y})^2} \,.$$

The most immediate and convincing interpretation of the AR metric stems from the following result: given the same set of initial values, *the distance between two $ARIMA$ processes is zero if and only if the corresponding models produce the same forecasts*.
The distance $d(X_t, Y_t)$ is a well defined measure of *structural dissimilarity* among any processes belonging to $\mathcal{L}$ and its value is determined by all the components of the processes to be compared. Notice that, if both $X_t \in \mathcal{L}$ and $Y_t \in \mathcal{L}$, then $d(X_t, Y_t)$ is always well defined irrespective of the fact that one or both the processes are stationary or non-stationary.
A recurrent objection against the AR metric is that it does not take the WN variance into account. Indeed, this quantity is a mere scale factor depending on the measurement unit: it is well known that, for stationary linear processes, the noise-to-series variances ratio is a function of the process parameters. In this respect, in order to detect influential observations, Peña (1990) considered the squared Mahalanobis AR distance to assess how the parameters of a model change when each observation is in turn removed from the time series and replaced by the estimated missing value. Such measure is not a metric; it explicitly depends on the WN variance, and it results in the squared Euclidean distance between the one step ahead forecast values of the series. Consequently, it is strongly affected by the scale unit.
Similar considerations apply to the proposal of Tong and Dabas (1990), which introduced *similarity and dissimilarity measures* for clustering the residuals obtained from various statistical models fitted to the same time series, and to Maharaj (1996, 1999, 2000) works aimed at classifying and clustering time series data.
These criteria are effective for the purposes that the Authors considered (outliers detection, clustering of several homogeneous time series, comparing residuals from different models fitted to the same series, etc.) but they cannot be used to compare in general two different time series. For instance, these proposals are not *even* useful to compare a model fitted to a time series with a model fitted to the logarithm of the same series.

4. Properties of the AR metric

The introduction of $d(X_t, Y_t)$ over $\mathcal{L}$ transforms $\mathcal{L}$ in a *metric space*, and any sub-class of $\mathcal{L}$ (e.g. the AR, MA, $ARMA$, IMA classes) is a well defined metric space with respect to the same metric. Then, any WN process is the *origin* for the metric space $\mathcal{L}$, and for any $Z_t \in \mathcal{L}$, the *norm* is defined by: $\sum_j \pi_j^2 < +\infty$. Moreover, the *angle* α between two processes $X_t \in \mathcal{L}$, $Y_t \in \mathcal{L}$ is defined by: $\cos(\alpha) = \sum_j \pi_{j,x}\pi_{j,y} \left(\sum_j \pi_{j,x}^2 \sum_j \pi_{j,y}^2 \right)^{-1/2}$. Notice that the metric space $\mathcal{L}$ is *isometric* with respect to seasonal processes, in the sense that, for any $s > 0$:

$$d\left(\pi_x(B)X_t, \pi_y(B)Y_t\right) = d\left(\pi_x(B^s)X_t, \pi_y(B^s)Y_t\right).$$

For multivariate applications of the metric it is important to define the distance of a single series from a given class, and the diameter of a class of time series models.
Given a series $X_t \in \mathcal{L}$ and a class of series $\mathcal{B} \subset \mathcal{L}$, the distance of X_t from $\mathcal{B}$ is defined by:

$$dist(X_t, \mathcal{B}) = \inf \{d(X_t, Y_t),\ Y_t \in \mathcal{B}\}\ .$$

For any class $\mathcal{B} \subset \mathcal{L}$, the *diameter* is defined by:

$$diam(\mathcal{B}) = \sup \{d(X_t, Y_t),\ X_t \in \mathcal{B},\ Y_t \in \mathcal{B}\}\ .$$

Finally, it may be shown that the sub-class of AR processes has a *finite* diameter. Notice that the size of the diameter has an immediate impact on the reliability of the selection of a representative time series from a given set.

5. A prototypical example for the AR metric

Let $X_t \sim ARMA(1,1)$ and $Y_t \sim ARMA(1,1)$ be two processes both belonging to $\mathcal{L}$. Then, from a formal expansion of the corresponding π_j coefficients:

$$\pi_{j,x} = (\phi_x - \theta_x)\,\theta_x^{j-1}; \quad \pi_{j,y} = (\phi_y - \theta_y)\,\theta_y^{j-1}; \qquad j = 1, 2, \ldots$$

we obtain:

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

This result is completely general for computing the distance between processes belonging to the sub-classes $AR(1), MA(1), ARIMA(0,1,1), ARIMA(0,1,1)$: it suffices, in the previous formula, to let some parameters equal to 0 and/or 1. For instance, by letting $\phi_x = \phi_y = 0$ and $\theta_x = 1 - \lambda_x$; $\theta_y = 1 - \lambda_y$, we obtain the distance between the $ARIMA(0,1,1)$ processes implied by the so-called *exponential smoothing* procedure.

6. Statistical inference for the ML estimator of the AR metric

If one needs to compare several $ARIMA$ processes, the estimation of the models parameters may be obtained by *ad hoc* modelling (for small/moderate size of the data set) or

by *automatic* modeling via AIC or BIC criteria (for large data set). In any case, the distribution of the distance estimator is needed in order to assess significant dissimilarities.
A preliminary result, concerning the comparison of AR models based on ML estimators, was obtained by Piccolo (1989). Instead, Sarno (2001) discusses asymptotic distribution of the metric derived from least squares estimators when MA processes are involved.
The distribution of the metric for any $ARIMA$ processes in $\mathcal{L}$ has been fully derived by Corduas (1996, 2000b), together with some efficient approximating distributions. Briefly, assuming that the ML method is used for estimating the $k = p + q + P + Q$ parameters of the $ARMA$ models to be compared, it can be shown that, under the null hypothesis $H_0 : \boldsymbol{\pi}_x = \boldsymbol{\pi}_y$:

$$\hat{d}^2(X_t, Y_t) \sim \sum_{j=1}^{k} \lambda_j \, \chi^2_{g_j} \, ,$$

where $\chi^2_{g_j}$ are independent Chi-square random variables, with g_j degrees of freedom given by the multiplicity of each eigenvalue (usually, $g_j \equiv 1$) and λ_j are the eigenvalues of a convenient matrix $\boldsymbol{C}_0$ of order $(k \times k)$.
In this regard, we notice that we may write the non-stationary $\boldsymbol{\pi}$ coefficients as the linear transformation: $\boldsymbol{\pi} = \boldsymbol{A}\,\boldsymbol{\pi}^A + \boldsymbol{v}$, of the stationary coefficients $\boldsymbol{\pi}^A$, for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{v}$. Then, for $ARIMA$ models, the matrix $\boldsymbol{C}_0$ is defined by:

$$\boldsymbol{C}_0 = \left(n_x^{-1} + n_y^{-1}\right) \, \boldsymbol{ABVB'A'} \, ,$$

where n_y and n_x are the lengths of the realizations of X_t and Y_t, respectively, and the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{V}$ can be derived from the models operators via effective algorithms.
For computing critical values, standard results may be applied to this problem: the approximation of a linear combination of Chi-square random variables by mean of a linear transformation of a Chi-square, with convenient degrees of freedom (as proposed by Corduas, 1996, 2000b), or the power transformation of the estimator $\hat{d}$ in order to improve its convergence to Normality (as proposed by D'Elia, 2000).

7. Extensions and generalizations of the AR metric

Firstly, the AR metric may be generalized in order to compare (long-memory) fractional difference processes $Z_t \sim ARFIMA(p, d, q)$, when $\mid d \mid < 0.5$. In this case, for the fractional difference operator $\nabla^d = (1 - B)^d$ we get:

$$\pi_i(d) = (-1)^{i+1}\binom{d}{i}, \quad i = 1, 2, \ldots; \qquad \sum_{i=1}^{\infty} \pi_i^2(d) = \frac{\Gamma(1 + 2d)}{\Gamma^2(1 + d)} - 1 < +\infty \, .$$

Notice that $Z_t \in \mathcal{L}$ as long as $d > -0.5$. Then, the Euclidean distance between two π-sequences is well defined even if one or both are generated by $ARFIMA$ operators. In this way, the AR metric is generalized to the class od invertible $ARFIMA$ processes and may be applied also if one or both processes express long-memory behaviour.
A second relevant extension of the AR metric has been proposed by Otranto (2004) for the classification of the volatility of financial time series generated by GARCH models. In particular, the AR metric can be used to measure the distance between squared noise processes. Then, cluster algorithms are applied in order to classify the volatility of several

stock prices and to study their interdependence. The joint application of the AR metric to both $ARIMA$ and $GARCH$ model components is a further proposal for clustering and discrimination of real time series.
A third extension of the AR metric concerns its usage as an estimation method, as proposed by Corduas (2000a). The objective is to find the fractional value of d such that the process $\nabla^d Z_t = a_{zt}$ is as close as possible to the $X_t \sim ARMA(1,1)$ process defined by: $(1-\phi B)X_t = (1-\theta B)a_{xt}$, where *the closeness is measured by the AR metric*. Then, given the estimates $\hat{\boldsymbol{\beta}} = (\hat{\phi}, \hat{\theta})'$, the problem is to find d such that:

$$G(d) = \sum_{i=1}^{\infty} \left[\pi_i(d) - \pi_i(\hat{\boldsymbol{\beta}})\right]^2 \simeq \sum_{i=1}^{L} \left[(-1)^i \binom{d}{i} - (\hat{\phi} - \hat{\theta})\,(\hat{\theta})^{i-1}\right]^2 = \min!$$

for some fixed $L = 100, 150$, say. Some related results were obtained by Corduas and Piccolo (2001, 2003, 2006); D'Elia and Piccolo (2002a, 2002b).
Further methodological issues concerning the AR metric include the computation of power functions in time series analysis (Gonzalo and Lee, 1996); the consequences on the metric when the series are correlated (Corduas, 1992a); the relationship between feedback in stochastic systems and Granger causality (Triacca, 2004a); a test of parallelism between two $ARIMA$ processes (Triacca, 2004b).

8. Main fields of applications

The AR metric has been applied in several scientific fields such as Economic and Finance, Demography, Medicine, Linguistic, Signal processing, Environmental Sciences, Hydrology and Meteorology, Sismology, Astronomy, and so on.
In this respect, we limit ourselves to mention some applications: the convergence of inflation rates in the EU countries (Sarno and Zazzaro, 2002); the information redundancy in environmental monitoring networks (Sarno, 2005); data mining problems (Agrawal et al., 1993, 1994); validation of seasonal adjustment procedures (Corduas and Piccolo, 1999); plotting time series as objects in a multidimensional scaling space (Piccolo, 1984b); the representativeness of an aggregated index (Caceres et al., 1993); the comparison of stochastic components of a time series (Corduas and Piccolo, 1995); similarity among original series and canonical components (Quilis, 2004); clustering time series (Piccolo, 1984; Cano et al., 1992; Corduas and Piccolo, 1996; Maharaj, 1996, 2000; Grimaldi, 2004; Liao, 2005); classification of multivariate time series (Maharaj, 1999; Galeano and Peña, 2000); detecting extreme (anomalous) behaviors in a homogeneous time series data set (Corduas and Piccolo, 1999); discriminating time series (Corduas, 2004); selecting between direct and indirect model-based seasonal adjustment (Otranto and Triacca, 2002).

9. A proposal for a spectral decomposition of the AR metric

The computation of the AR metric has been greatly simplified by a theorem obtained by Corduas (1992b) who showed that: *the squared AR metric is always the variance of a well defined stationary process*.
Given $X_t \in \mathcal{L}$ and $Y_t \in \mathcal{L}$, and a standardized White Noise process $\epsilon_t \sim WN(0,1)$, we

may define a dummy *stationary* process W_t as:

$$W_t = [\pi_X(B) - \pi_Y(B)]\, \epsilon_t, .$$

Then, the squared AR metric $d^2(X_t, Y_t)$ is exactly the variance of W_t. Thus, for the computation of the AR metric, efficient numerical algorithms may be applied via a state-space representation of the $ARMA$ processes (as proposed by Anderson and Moore, 1979) or via the autocovariance generating function of the W_t process (as proposed by Tunnicliffe Wilson, 1979).
In this regard, we are currently exploiting the Corduas' theorem in order to perform a spectral decomposition of the AR metric. As a matter of fact, if $d^2(X_t, Y_t)$ is the variance of a stationary process W_t, then there is a a well defined spectrum $g_W(\omega)$ whose components exhibit the contributions of each angular frequency $\omega \in [-\pi, \pi]$ to the dissimilarity among the processes X_t and Y_t, according to the spectral decomposition:

$$d^2(X_t, Y_t) = Var\,(W_t) = \int_{-\pi}^{\pi} g_W(\omega)\, d\omega\,.$$

This decomposition should improve the interpretative content of the AR metric since it would enhance the stochastic components (or a convenient range of them) which are relevant for explaining the observed distance among the processes.

10. Some alternatives to the AR metric

We defer to the literature for an extensive review of the many alternatives to the AR metric and we limit ourselves to quote the pioneering work of Zani (1983) and the results of some Italian researchers whose contributions are relevant in this area. Specifically, Baragona (2001) and Baragona et al. (2001) introduced genetic algorithms to measure diversity in time series data. Moreover, Ingrassia et al. (2003) applied functional analysis and Cerioli et al. (2004) performed clustering by means of symbolic analysis.
Instead, a technique that in our opinion is strongly related to the AR metric has been proposed by several Authors and stems from the definition of *cepstrum* by Bogert at al. (1962). The cepstrum coefficients c_j are obtained by the parametric expansion of the logarithm of the spectrum $f_Z(\omega)$ of a stationary process Z_t, so that:

$$c_j \;=\; \frac{1}{2\,\pi} \int_{-\pi}^{\pi} \log[f(\omega)]\, e^{-\iota\omega j}\, d\omega, \quad j = 1, 2, \ldots$$

and $c_0 = \log(\sigma^2/(2\,\pi))$ by Kolmogorov identity. Gray and Markel (1976) proposed the Euclidean distance $\delta(X_t, Y_t)$ among the cepstral coefficients of two stationary processes X_t and Y_t as an effective metric among time series data. It has been applied with some interesting results by Kang et al. (1995) and Kalpakis et al. (2001). Recently, a weighted version of $\delta(X_t, Y_t)$ has been reformulated by Martin (2000) who completely ignores the existence of any other metric.
Indeed, cepstral coefficients have several properties, including a relationship with the partial autocorrelation function (Li, 2004). Moreover, all the proposed metrics correctly exclude the c_0 coefficients arguing that it is a scale factor. It is worth to observe that, with obvious notations, the following identity holds:

$$\int_{-\pi}^{\pi} |\, \log[f_X(\omega)] - \log[f_Y(\omega)]\, |^2 = \sum_{j=-\infty}^{\infty} (c_{xj} - c_{yj})^2 = 2\,\delta^2(X_t, Y_t) + \log\left(\frac{\sigma_{ax}^2}{\sigma_{ax}^2}\right).$$

As a consequence, if the processes have different degree of non-stationarity the cepstral metric is not defined. A further problem is generated by the fact that, after few lags, the cepstral coefficients are nearly zero; thus, the cepstral metric is heavily determined only by few AR parameters.

11. Concluding remarks

Time and/or spectral properties are the stylized features of any time series generated by $ARIMA$ models: these characteristics are fully conveyed by the forecast function. As a consequence, the AR formulation is a fundamental issue for any time series analysis, both for stationary and non-stationary processes. Notice that non-linearity and deterministic trends are excluded from our approach; thus, when these dynamics are relevant, we suggest to move towards non-parametric metrics, as for instance those discussed by Zhang and Taniguchi (1995).
In this regard, it is important to realize that any metric both enhances and hides several aspects of the compared objects. Then, the AR *metric* is a useful tool for many statistical objectives if the whole *structural diversity* is the main point of the analysis. On the other hand, a *spectral metric* could be more effective if the comparison involves *local features* (e.g. long memory, periodic patterns, seasonality, etc.).
As a matter of fact, in the AR metric, the contribution of each π_j coefficient to the stochastic components of the process is spread over all the angular frequencies; thus, if our concept of closeness is related to some specific component this metric should not be applied. For instance, Caiado et al. (2006) proved by simulation that a periodogram-based measure may result more effective than the AR metric for detecting a non-stationary behaviour. Similarly, Piccolo and Corduas (2006) preferred a spectral metric when the angular frequencies around 0 are the central issue for assessing the similarity among stationary and fractional difference processes.

Finally, we notice that the AR metric is well defined for stationary and non stationary, for short and long memory, for seasonal and non seasonal processes. Thus, according to our opinion and experiences, the AR metric is a powerful and wide applicable tool to study and understand the relationships among time series, but it also helps to produce new ideas, genuine proposals and innovative developments.

References

Agrawal R., Faloutsos C. and Swami A. (1994) Efficient similarity search in sequence databases, in: *4th Proceedings of F.O.D.O. '93*, Springer Verlag, New York, number 730 in Lecture notes in Computer Science, 69–84.

Anderson B. and Moore J. (1979) *Optimal filtering*, Prentice Hall, Englewood Cliffs.

Baragona R. (2001) A simulation study on clustering time series with metaheuristic methods, *Quaderni di Statistica*, 3, 1–26.

Bogert B., Healy M. and Tukey J. (1962) The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-ceptstrum and saphe cracking, in: *Proceeding of the Symposium on Time Series Analysis*, Rosenblatt M., ed., J. Wiley & Sons, New York, 209–263.

Box G. and Jenkins G. (1970) *Time series analysis: forecasting and control (revised edition, 1976)*, Holden-Day, San Francisco.

Brockwell P. and Davies R. (1991) *Time series: theory and methods (2nd edition)*, Springer-Verlag, New York.

Caceres J., Cano V. and Martin F. (1993) Analysis de la representatividad del I.P.I. agregado, Documento de trabajon n.45, Universidad de la Laguna, Tenerife.

Caiado J., Crato N. and Peña D. (2006) A periodogram-based metric for time series classification, *Computational Statistics & Data Analysis*, 50, 2668–2684.

Cano V., Martin F. and Caceres J. (1992) Medida de distancia entre modelos ARIMA. una aplicacion a los indices de precios percibidos por los agricultores, *Investigacion Agraria*, 7, 33–45.

Cerioli A., Ingrassia S. and Corbellini A. (2004) Classificazione simbolica di dati funzionali: un'applicazione al monitoraggio ambientale, in: *Data mining e analisi simbolica*, Lauro C.N. and Davino C., eds., F. Angeli, Milano.

Corduas M. (1992a) *Misure di distanza tra serie storiche e modelli parametrici*, Quaderni dell'Istituto Economico Finanziario, n.3, Università di Napoli Federico II.

Corduas M. (1992b) Una nota sulla distanza tra modelli ARIMA per serie storiche correlate, *Statistica*, LII, 512–520.

Corduas M. (1996) Uno studio sulla distribuzione asintotica della metrica Autoregressiva, *Statistica*, LVI, 321–332.

Corduas M. (2000a) La metrica autoregressiva tra modelli ARIMA: una procedura in linguaggio GAUSS, *Quaderni di Statistica*, 2, 1–37.

Corduas M. (2000b) Preliminary estimation of ARFIMA models, in: *Proceedings in Computational Statistics*, Betlehem J. and van der Heijden P., eds., Physica Verlag, Heidelberg, 247–252.

Corduas M. (2003) Il confronto tra serie storiche nell'analisi statistica di dati dinamici, in: *Atti della Riunione SIS*, Rocco Curto editore, Napoli, 213–224.

Corduas M. (2004) Time series discrimination using AR metric, in: *Proceedings of XLII Riunione Scientifica SIS*, CLEUP, Padova, 143–146.

Corduas M. (2007) Comparing time series: shape-based or structural similarities?, in: *Proceedings of CLADAG-2007 Meeting*, University of Macerata.

Corduas M. and Piccolo D. (1995) Mutamenti strutturali della natalità e differenziazioni regionali, in: *Atti del Convegno SIS: "Continuità e discontinuità nei fenomeni demografici"*, Università degli Studi della Calabria, Editore Rubettino, 315–322.

Corduas M. and Piccolo D. (1996) Time series clustering of the Italian consumer price indices: a model approach, in: *Quaderni di Ricerca ISTAT*, Istituto Nazionale di Statistica, Roma.

Corduas M. and Piccolo D. (1999) On the use of AR metric for seasonal adjustment, in: *Proceedings of the International Conference CLADAG-99*, University of Rome "La Sapienza", 1–4.

Corduas M. and Piccolo D. (2001) Fractional differencing models estimations: some new approaches, in: *Metodi Statistici e Matematici per l'Analisi delle Serie Idrologiche*, Piccolo D. and Ubertini L., eds., CNR-GNDCI n.2136, Roma, 73–79.

Corduas M. and Piccolo D. (2003) Determinazione del lag ottimale nelle stime di minima distanza del parametro alle differenze frazionarie, in: *Metodi Statistici e Matematici per l'Analisi delle Serie Idrologiche*, Piccolo D. and Ubertini L., eds., CNR-GNDCI n.2818, Roma, 73–80.

Corduas M. and Piccolo D. (2007) Time series clustering and classification by the Au-

toregressive metric, submitted for publication, Dipartimento di Scienze Statistiche.
De Souza P. (1977) Statistical tests and distance measures for LPC coefficients, *IEEE Transactions on Acoustics, Speech, and Signal processing*, ASSP-25, 6, 554–559.
D'Elia A. (2000) Uno studio sull'asimmetria dello stimatore della metrica Autoregressiva, *Quaderni di Statistica*, 2, 59–84.
D'Elia A. and Piccolo D. (2002a) A comparison among several methods for estimating the fractional differencing parameter, in: *Proceedings of the Compstat 2002 Conference*, Kinke S. and Ahrend P. Richter L., eds., Humboldt-Universitat zu Berlin.
D'Elia A. and Piccolo D. (2002b) Stimatori di minima distanza del parametro alle differenze frazionarie, *Quaderni di Statistica*, 4, 115–138.
Galeano P. and Peña D. (2000) Multivariate analysis in vector time series, *Resenhas*, 4, 383–404.
Ge D., Srinivasan N. and S.M. K. (2002) Cardiac arrhythmia classification using autoregressive modeling, *Biomedical Engineering OnLine*, http:\\www.biomedical-engineering-online.com.
Gonzalo J. and Lee T. (1996) Relative power of t type tests for stationary and unit root processes, *Journal of Time Series Analysis*, 17, 37–47.
Gray A. and Markel J. (1976) Distance measures for speech processing, *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24, 380–391.
Grimaldi S. (2004) Linear parametric models applied on daily hydrological series, *Journal of Hydrological Engineering*, 9, 383–391.
Ingrassia S., Cerioli A. and Corbellini A. (2003) Some issues on clustering of functional data, in: *Between Data Science and Applied Data Analysis*, Schader M., Gaul W. and Vichi M., eds., Springer, Berlin, 49–56.
Kalpakis K., Gada D. and Puttagunta V. (2001) Distance measures for effective clustering of ARIMA time series, in: *Proceedings of the IEEE International Conference on Data Mining*, ICDM'01, San Jose, California, 273–280.
Kang W., Cheng C., Lai J. and Tsao H. (1995) The application of cepstral coefficients and maximum likelihood method in EGM pattern recognition, *IEEE Transactions on Biomedical Engineering*, 42, 777–785.
Kosĕc D. (2000) Parametric estimation of continuous non stationary spectrum and its dynamics in surface EMG studies, *International Journal of Medical Informatics*, 58/59, 59–69.
Kovačić Z. (1996) Classification of time series with application to the leading indicator selection, in: *Proceedings of the fifth Conference of IFCS*, number 2, 204–207.
Li L. (2004) Some notes on mutual information between past and future, *Journal of Time Series Analysis*, 27, 309–322.
Liao T. (2005) Clustering time series data - a survey, *Pattern Recognition*, 38, 1857–1874.
Maharaj E. (1996) A significance test for classifying ARMA models, *Journal of Statistical Computation and Simulation*, 54, 305–331.
Maharaj E. (1999) Comparison and classification of stationary multivariate time series, *Pattern Recognition*, 32, 1129–1138.
Maharaj E. (2000) Clusters of time series, *Journal of Classification*, 17, 297–314.
Martin R. (2000) A metric for ARMA processes, *IEEE Transactions on Signal Processing*, 48, 1164–1170.
Otranto E. (2004) Classifying the markets volatility with ARMA distance measures, *Quaderni di Statistica*, 6, 1–19.
Otranto E. and Triacca U. (2002) Measures to evaluate the discrepancy between direct and

indirect model-based seasonal adjustment, *Journal of Official Statistics*, 18, 511–530.
Peña D. (1990) Influential observation in time series, *Journal of Business and Economic Statistics*, 8, 235–242.
Piccolo D. (1972) Analisi statistica dei prezzi all'ingrosso in Italia:1956-71, *Rassegna Economica*, XXXVI, 1555–1599.
Piccolo D. (1984a) Una topologia per la classe dei processi ARIMA, *Statistica*, XLIV, 47–59.
Piccolo D. (1984b) Una rappresentazione multidimensionale per modelli statistici dinamici, in: *Atti della XXXII Riunione Scientifica della SIS*, volume II, 149–160.
Piccolo D. (1989) On a measure of dissimilarity between ARIMA models, in: *Proceedings of the A.S.A. Meetings, Business and Economic Statistics Section*, ASA, Washington D.C., 231–236.
Piccolo D. (1990) A distance measure for classifying ARIMA models, *Journal of Time Series Analysis*, 11, 153–164.
Piccolo D. and Corduas M. (2006) Spectral approximation to the fractional differencing operator, in: *Metodi Statistici e Matematici per l'Analisi delle Serie Idrologiche*, Piccolo D. and Ubertini L., eds., CNR-GNDCI n.2908, Roma, 11–23.
Quilis E. (1990) Una aplicación de los modelos BVAR estacionales, *Economía, Instituto Nacional de Estadistíca, Madrid*, 4, 207–214.
Sarno E. (2001) Further results on the asymptotic distribution of the Euclidean distance between MA models, *Quaderni di Statistica*, 3, 165–175.
Sarno E. (2005) Testing information redundancy in environmental monitoring networks, *Environmetrics*, 16, 71–79.
Sarno E. and Zazzaro A. (2002) An index of dissimilarity among time series: an application to the inflation rates of the EU countries, in: *Proceedings of COMPSTAT 2002*, Klinke S., Ahrend P. and Richter L., eds., Springer, Berlin.
Thomson P. and De Souza P. (1985) Speech recognition using LPC distance measures, in: *Handbook of Statistics*, Hannan E., Krishnaiah P. and Rao M., eds., North Holland, Amsterdam, volume 5, 389–412.
Tong H. and Dabas P. (1990) Cluster of time series, *Journal of Applied Statistics*, 17, 187–198.
Tran-Luu T. and DeClaris N. (1997) Visual heuristics for data clustering, *IEEE Transactions on Systems, Man and Cybernetics*, 1, 19–24.
Triacca U. (2004aa) Feedback, causality and distance between ARMA models, *Mathematics and Computers in Simulation*, 64, 679–685.
Triacca U. (2004ab) A note on distance and parallelism between two ARIMA processes, *Quaderni di Statistica*, 6, 21–29.
Tunnicliffe Wilson G. (1979) Some efficient computational procedure for high order ARMA models, *Journal of Statistical Computation and Simulation*, 8, 301–309.
Zani S. (1983) Osservazioni sulle serie storiche multiple e l'analisi dei gruppi, in: *Analisi moderna delle serie storiche, Convegno nazionale 1981*, Piccolo D., ed., F. Angeli, Milano, 263–274.
Zhang G. and Taniguchi M. (1995) Nonparametric approach for discriminant analysis in time series, *Nonparametric Statistics*, 5, 91–101.