

Spectral decomposition of the AR metric

Maria Iannario and Domenico Piccolo

Abstract This work investigates a spectral decomposition of the AR metric proposed as a measure of structural dissimilarity among ARIMA processes. Specifically, the metric will be related to the variance of a stationary process so that its behaviour in the frequency domain will help to detect how unobserved components generated by the parameters of both phenomena concur in specifying the obtained distance. The main consequences of this result are discussed and some specific formal decomposition are examined in order to improve the interpretative content of the AR metric.

Key words: AR metric, Spectral decomposition, ARIMA processes

1 Introduction

In recent years, we register an increasing interest towards dissimilarity measures among dynamic phenomena for clustering and discrimination purposes. This interest is often involved with *data mining* procedures operating on massive data archives originated from various fields of study: Economics and Finance, Signal processing, Environmental Sciences, Medicine, Demography, Hydrology, Meteorology, Physics, Astronomy, Geology, and so on.

In this respect, extensive reviews of the current research on clustering and discrimination of time series can be found in [12, 11, 18, 9]. Moreover, innovative

Maria Iannario

Department of Statistical Sciences, University of Naples Federico II, Via L. Rodinò 22, I-80138 Naples, Italy, e-mail: maria.iannario@unina.it

Domenico Piccolo

Department of Statistical Sciences, University of Naples Federico II, Via L. Rodinò 22, I-80138 Naples, Italy e-mail: domenico.piccolo@unina.it

proposals aimed at extending the use of AR metric to more complex econometric models have been discussed by [13, 14, 19].

In this paper, moving from a formal result about the AR metric between linear processes, a spectral decomposition is considered. Specifically, the metric will be related to the variance of a convenient stationary process obtained by the parameters of both processes to be compared; so its behaviour in the frequency domain will help to detect how unobserved components of this new process concur in specifying the obtained distance. From a logical point of view, this approach is similar to results of [1] who obtained a frequential decomposition of the linear determinism index introduced by [2].

As a first contribution on this topic, this work is mainly devoted to obtain useful results based on a full knowledge of the probability structure of the processes to be compared; further researches should translate such achievements in an inferential framework when models are estimated on real time series data sets. However, the availability of consistent and efficient maximum likelihood estimators should make such further extensions immediate.

The paper is organized as follows: in the next section, we summarize the foundations for the metric. In section 3 we establish the connection between the AR metric and the variance of a suitable stationary process, and derive a spectral decomposition of such measure. Section 4 is concerned with the study and comments of the decomposition of some processes often encountered in the applications. Some concluding remarks end the paper.

2 The foundations for the AR metric

In order to set a convenient notation, for a given stochastic process X_t , we consider ARIMA models for a transformed process $Z_t = g(X_t) - f_t$, which is obtained after a possibly non-linear transformation of X_t and then by removing any deterministic component f_t : trading days, calendar effects, outliers and mathematical functions of time, including constants. Such transformations are completely general and should improve both linearity and Gaussianity of Z_t .

Hereafter, Z_t is a zero-mean invertible ARIMA(p, d, q) process¹, that is:

$$\phi(B) \nabla^d Z_t = \theta(B) a_t, \quad a_t \sim WN(0, \sigma_a^2),$$

where a_t is a sequence of zero-mean uncorrelated and homoscedastic random variables (defined White Noise) and the backshift operator B applies to any function h_t of t in such a way that $B^k h_t = h_{t-k}$, $\forall k = 0, \pm 1, \dots$ [3].

The AR and MA polynomials $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$, respectively, have *no common factors* and, moreover, all the roots of $\phi(B)\theta(B) = 0$ lie outside the unit circle. Finally, we denote by $\mathcal{L}$ the

¹ In order to simplify notation, in this paper we are not formally considering multiplicative seasonal ARIMA processes; however, all the subsequent results straightforwardly follow.

class of all linear *causal and invertible* stochastic processes $Z_t \sim \text{ARIMA}(p, d, q)$. Notice that $\mathcal{L}$ includes also (long-memory) fractional difference processes $Z_t \sim \text{ARFIMA}(p, d + d^*, q)$, as long as $d^* \in (-0.5, 0.5)$.

Then, for convenient initial values and a constant $\sigma_a^2 < +\infty$, any process $Z_t \in \mathcal{L}$ admits a well defined AR operator $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \dots$ characterized by the square-converging sequence $\{\pi_1, \pi_2, \dots\}$, which specifies the *forecast function* $F_t = \mathbb{E}(Z_t | Z_{t-1}, Z_{t-2}, \dots)$. Indeed, for all $Z_t \in \mathcal{L}$, the following orthogonal representation holds:

$$F_t = \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \dots; \quad Z_t = F_t + a_t, \quad F_t \perp a_t.$$

In this context, a fundamental theorem [4, 130-133] assesses that: “for *any* stationary process with a continuous spectrum $f(\omega)$ there exists a finite order $\text{AR}(p)$ whose spectrum $f_{\text{AR}}(\omega)$ is as close as possible in absolute value to $f(\omega)$ uniformly on $[-\pi, \pi]$ ”.

Taking these results into account, Piccolo [15, 16] introduced a measure of *structural dissimilarity* among two processes $X_t \in \mathcal{L}$ and $Y_t \in \mathcal{L}$ by defining as *AR metric* the Euclidean distance among the coefficients $\pi_x = (\pi_{1,x}, \pi_{2,x}, \dots)'$ and $\pi_y = (\pi_{1,y}, \pi_{2,y}, \dots)'$, obtained by the AR expansions of X_t and Y_t , respectively.

With obvious notation, we get:

$$d(X, Y) = [(\pi_x - \pi_y)'(\pi_x - \pi_y)]^{\frac{1}{2}} = \sqrt{\sum_{j=1}^{\infty} (\pi_{j,x} - \pi_{j,y})^2}. \quad (1)$$

Then, Piccolo [16] and Corduas [6, 7] derived the asymptotic distributions of the corresponding estimator for AR and ARIMA processes, respectively. Other results on these topics are reported by [18] and [9], who provide also efficient algorithms for computation.

The main interpretation of the AR metric (valid also for ARFIMA processes: [8]) is the following: *the distance between two ARIMA processes is zero if, given the same set of initial values, the corresponding models produce the same forecasts*. Notice that, as a measure of structural dissimilarity, the AR metric cannot take WN variances into account since they are just scale factors and do not affect the dynamic of the processes to be compared. This feature is shared by cepstral distances [10].

A useful example of the AR metric is obtained by comparing $X_t \sim \text{ARMA}(1, 1)$ and $Y_t \sim \text{ARMA}(1, 1)$, both belonging to $\mathcal{L}$. Thus, with obvious notation, we get:

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

If in this formula we let some parameters equal to 0 and/or 1, we find distances for several ARIMA subclasses including ARIMA(0,1,1) processes implied by the *exponential smoothing* procedures.

Finally, we observe that the introduction of $d(X, Y)$ transforms $\mathcal{L}$ in a *metric space* and this supports several implications which are also important for statistical

consequences (cluster analysis, discrimination, multidimensional scaling, choosing representative or anomalous structures, and so on): for instance, the diameter of a class of processes seems relevant for the multivariate application of the metric.

3 Frequential analysis of the AR metric

The starting point of our derivation is a theorem obtained by Corduas [5] who proposed to compute the AR metric by means of the variance of a related ARMA process, by using efficient algorithms. We slightly modify her result for saving the symmetry of the metric and for a more immediate fitting to our objectives.

We denote by

$$\delta_j = |\pi_{j,x} - \pi_{j,y}|, \quad j = 1, 2, \dots$$

the absolute difference of the AR expansion of processes $X_t \in \mathcal{L}$ and $Y_t \in \mathcal{L}$. Then, we introduce the stationary process:

$$W_t = \eta_t + \alpha_1 \eta_{t-1} + \alpha_2 \eta_{t-2} + \dots = \alpha(B) \eta_t, \quad (2)$$

where $\eta_t \sim WN(0, \delta_1^2)$ and the coefficients α_j are obtained as²:

$$\alpha_j = \frac{\delta_{j+1}}{\delta_1}, \quad j = 1, 2, \dots$$

Then, it is immediate to show that

$$\text{Var}(W_t) = \sum_{j=1}^{\infty} |\pi_{j,x} - \pi_{j,y}|^2 = d^2(X, Y). \quad (3)$$

The process W_t is obtained by filtering the WN process η_t by means of a linear operator $\alpha(B)$ where the differences of the AR expansions of both processes X_t and Y_t are taken into account. It is evident that the time dynamics of W_t reflects characteristics of both processes to be compared. As a matter of fact, two processes X_t and Y_t with similar behaviour produce small α_j coefficients while processes with quite dissimilar structures give large coefficients α_j . Then, the statistical analysis of W_t might help in investigating the structural dissimilarity among X_t and Y_t as synthesized by the AR metric.

The process W_t may be non-invertible but it is *always stationary*. Thus, it admits a symmetric spectral function defined by:

$$g(\omega) = \frac{\delta_1^2}{\pi} |\alpha(e^{-i\omega})|^2, \quad 0 < \omega < \pi.$$

² In fact, to be correct we should define as δ_1 the first non-zero absolute difference between the AR expansions of X_t and Y_t processes. This point is important when comparing pure AR seasonal processes, for instance; however, we are omitting it for simplicity of notation.

Then, the spectral decomposition of the AR metric immediately follows:

$$d^2(X, Y) = \int_0^\pi g(\omega) d\omega. \quad (4)$$

This result allows to ascertain whether some components at low, high, periodic, seasonal frequencies contribute to the determination of the value of $d^2(X, Y)$.

In addition, since $g(\omega)/d^2(X, Y)$ shares all the properties of a well defined density function over $(0, \pi)$, we may also consider the spectral distribution function:

$$G(\omega) = \int_0^\omega \frac{g(z)}{d^2(X, Y)} dz.$$

In this way, for any fixed angular frequency $\omega_0 \in (0, \pi)$, the function $G(\omega_0)$ measures the relative proportion of distance explained by angular frequencies less or equal to ω_0 .

4 Discussion of some specific decompositions

We study some specific comparisons by examining the spectral decomposition implied when the AR metric is computed between the most common processes.

- *AR processes.* If $X_t \sim AR(p_x)$ and $Y_t \sim AR(p_y)$, then W_t is a pure MA process of order $q_w = \max(p_x, p_y)$ characterized by the coefficients:

$$\alpha_j = \frac{|\phi_{j+1,x} - \phi_{j+1,y}|}{|\phi_{1,x} - \phi_{1,y}|}, \quad j = 1, 2, \dots$$

As a consequences, we get a constant spectrum $g(\omega)$ when $\max(p_x, p_y) = 1$; this results apply also to the comparison of pure seasonal AR(1) processes because of the isometry of the AR metric.

A different situation happens in environmental studies, where it may be of interest to compare phenomena well synthesized by AR(2) processes with different stochastic periods $P_x \geq 2$ and $P_y \geq 2$, and amplitudes $\rho_x > 1$ and $\rho_y > 1$, respectively. In these cases, the AR coefficients for X_t are:

$$\phi_{1,x} = 2\rho_x^{-1} \cos\left(\frac{2\pi}{P_x}\right); \quad \phi_{2,x} = -\rho_x^{-2};$$

and similarly for Y_t . Then, the W_t process is completely specified by δ_1 , δ_2 , and $\alpha_1 = \delta_2/\delta_1$.

Thus, the spectral decomposition (4) of $d^2(X, Y)$ derives from the spectrum:

$$g(\omega) = \frac{\delta_1^2}{\pi} [1 + \alpha_1^2 + 2\alpha_1 \cos(\omega)], \quad 0 < \omega < \pi,$$

As an instance, Figure 1 shows the spectra of two AR(2) processes with periods 12 and 6, respectively (left panel), and the corresponding spectral decomposition of the AR distance between them (right panel). It is evident that the main contributions to the distance among the two processes are given by low frequencies (that is high periods) of the composite process W_t .

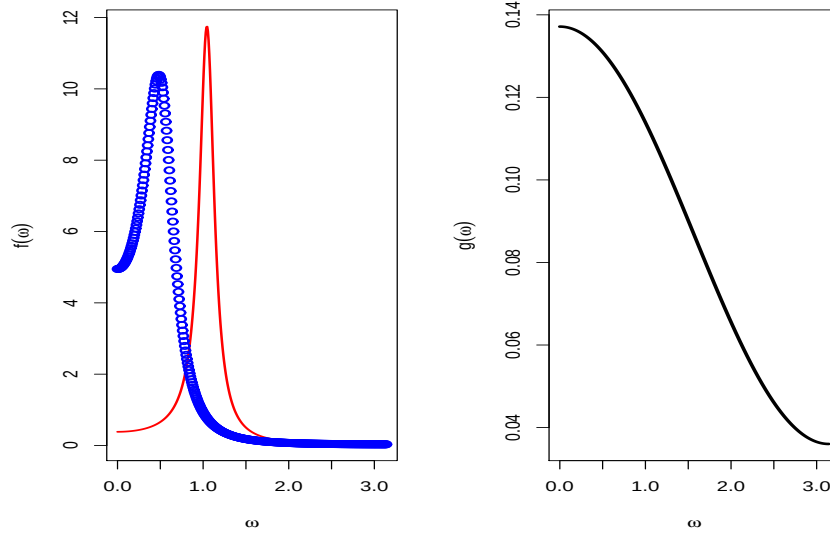


Fig. 1 Spectra of two AR(2) processes and frequential decomposition of the AR metric

Further considerations may be derived from these results: for instance, if two processes are characterized by stochastic periods such that $P_x = 2P_y/(P_y - 2)$, then $g(\omega)$ is uniformly constant, since this implies $\phi_{2,x} = \phi_{2,y}$. As a consequence, two AR(2) processes with stochastic periods are more and more distant inasmuch as these periods are different from 4 units.

• *ARMA(1,1) processes.* With standard notation, the AR expansions of two $X_t \sim ARMA(1,1)$ and $Y_t \sim ARMA(1,1)$ processes, both belonging to $\mathcal{L}$, are given by:

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

Then, $\delta_1 = |(\phi_x - \theta_x) - (\phi_y - \theta_y)|$ and the sequence of coefficients:

$$\alpha_j = \delta_1^{-1} \left| (\phi_x - \theta_x) \theta_x^{j-1} - (\phi_y - \theta_y) \theta_y^{j-1} \right|; \quad j = 1, 2, \dots,$$

converges to 0 when $\delta_1 \neq 0$. Then, we can approximate the $\alpha(B)$ operator by omitting coefficients for $j > M$ for an integer M such that $\max(\theta_x^M, \theta_y^M) < \varepsilon$, for a small $\varepsilon > 0$. In this way, we may well consider as a convenient approximation for $g(\omega)$ the spectral decomposition $g_M(\omega)$ based on a truncated $\alpha_M(B)$ operator.

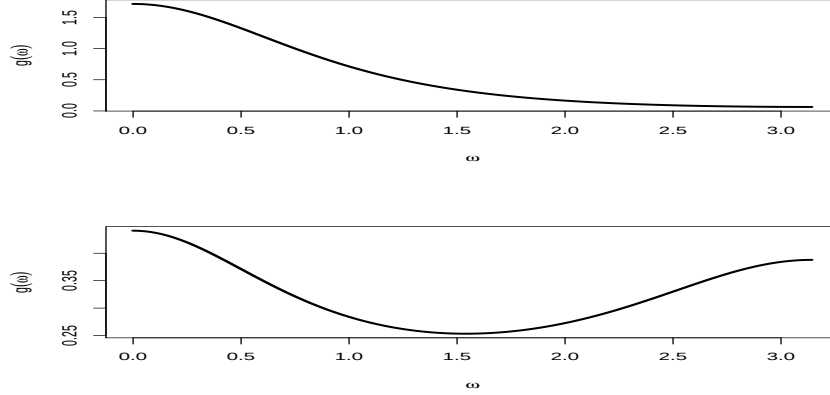


Fig. 2 Spectral decompositions of the AR metric for two ARMA(1,1) processes.

Figure 2 shows the spectral decompositions obtained when we compare two ARMA(1,1) processes which are quite similar, with $\phi_x = 0.9$; $\theta_x = 0.3$; $\phi_y = 0.7$; $\theta_y = 0.2$ (upper panel), and two very dissimilar ones, with $\phi_x = -0.9$; $\theta_x = 0.3$; $\phi_y = 0.6$; $\theta_y = -0.4$ (lower panel), respectively. The function $g(\omega)$ is mainly concentrated on the low frequencies in the first instance whereas it is shared almost evenly between high and low frequencies in the second case.

5 Concluding remarks

This paper provides a frequential analysis of the AR metric in order to evaluate the relevance and persistence of possible unobserved components in determining the observed distance. Among the possible fields of applications we quote the comparison of monthly indexes of industrial production (for allocating similarities among series to short or long movement in the economic cycle), the study of EEG time series data (for discriminating among safe and risky patterns), and so on.

Apparently, these issues are all achievable by some spectral metric (as those proposed by many Authors: see [18], for some comments). However, spectral metrics

are not always well defined for all processes belonging to $\mathcal{L}$; this is the case when processes with difference operators of different orders are to be compared, for instance. Thus, it seems worthwhile to explore also the spectral features of a metric always well defined on $\mathcal{L}$.

In this respect, some empirical work on real data set seems encouraging; however, further research is necessary for using such results in an effective way from an operational point of view.

Acknowledgements The research has been supported by Department of Statistical Sciences, University of Naples Federico II. Authors thanks CFEPSR, Portici, for the availability of scientific structures.

References

1. Barbieri M.M., Decomposizione frequenziale dell'indice di determinismo lineare, *Atti della XXXIV Riunione Scientifica della SIS*, vol. 2, 59–66, Nuova Immagine Editrice, Siena, (1988).
2. Battaglia F., Inverse autocovariances and a measure of linear determinism for a stationary process, *Journal of Time Series Analysis*, 4, 79–87, (1983).
3. Box G.E.P., Jenkins G.M., *Time series analysis: forecasting and analysis*, revised edition, 1976, Holden-Day, S.Francisco, (1970).
4. Brockwell, P.J., Davies, R.A., *Time series: theory and methods*, 2nd edition, Springer-Verlag, New York, (1991).
5. Corduas M., *Misure di distanza tra serie storiche e modelli parametrici*, Quaderni dell'Istituto Economico Finanziario, n.3, Università di Napoli Federico II, (1992).
6. Corduas M., Uno studio sulla distribuzione asintotica della metrica Autoregressiva, *STATISTICA*, LVI, 321–332, (1996).
7. Corduas M., La metrica autoregressiva tra modelli ARIMA: una procedura in linguaggio GAUSS, *Quaderni di Statistica*, 2, 1–37, (2000).
8. Corduas M., Classifying daily streamflow time series by AR metric, 4th *International Workshop on Space-Temporal Modelling*, EDES, Sassari, 171–176, (2008).
9. Corduas M., Piccolo D., Time series clustering and classification by the Autoregressive metric, *Computational Statistics & Data Analysis*, 52, 1860–1872, (2007).
10. Gray, A., Markel, J., Distance measures for speech processing, *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24, 380–391, (1976).
11. Liao, T., Clustering time series data: A survey, *Pattern Recognition*, 38, 1857–1874, (2005).
12. Maharaj, E.A., Clusters of time series, *Journal of Classification*, 17, 297–314, (2000).
13. Otranto, E., Classifying the markets volatility with ARMA distance measures, *Quaderni di Statistica*, 6, 1–19, (2004).
14. Otranto E., Clustering heteroskedastic time series by model-based procedures, *Computational Statistics & Data Analysis*, 52, 4685–4698, (2007).
15. Piccolo D. Una topologia per la classe dei processi ARIMA, *STATISTICA*, XLIV, 47–59, (1984).
16. Piccolo D., On a measure of dissimilarity between ARIMA models, *Proceedings of the ASA Meetings, Business and Economic Section*, Washington, DC, 231–236, (1989).
17. Piccolo D., A distance measure for classifying ARIMA models, *Journal of Time Series Analysis*, 11, 153–164, (1990).
18. Piccolo D., Statistical issues on the AR metric in time series analysis, *Proceedings of the SIS 2007 Conference on "Risk and Prediction"*, CLEUP, Padova, 221–232, (2007).
19. Triacca U., Feedback, causality and distance between ARMA models, *Mathematics and Computers in Simulation*, 64, 679–685, (2004).