Jeffrey’s Divergence Between Autoregressive Moving-Average Processes

Léo Legrand†, Éric Grivel† and Audrey Giremus‡
†Bordeaux university - INP Bordeaux ENSEIRB-MATMECA - IMS - UMR CNRS 5218, Talence, FRANCE
‡Thales Systèmes Aéroportés, Mérignac, FRANCE

Abstract—Various works have been carried out about the Jeffrey’s divergence (JD) which is the symmetric version of the Kullback-Leibler (KL) divergence. An expression of the JD for Gaussian processes can be deduced from the definition of the KL divergence and the expression of the Gaussian-multivariate distributions of k-dimensional random vectors. It depends on the $k \times k$ Toeplitz covariance matrices of the stationary processes. However, the resulting computational cost may be high as these matrices must be inverted and it is all the higher as $k$ increases. To circumvent this problem, a recursive expression can be obtained for real 1st-order autoregressive (AR) processes. When they are disturbed by additive uncorrelated white noises, we showed that when $k$ becomes large, the derivative of the JD with respect to $k$ tends to be constant. This constant is sufficient to compare the noisy AR processes. In this paper, we propose to extend our work to AR moving-average (MA) processes with one AR term and one MA term. Some examples illustrate the theoretical analysis.

Index Terms—Jeffrey’s divergence, Kullback-Leibler divergence, autoregressive moving-average processes.

I. INTRODUCTION

Comparing stochastic processes such as autoregressive moving-average (ARMA) processes can be useful in many applications, especially when classification is required and when change detection is of interest. One intuitive approach would be to compare the process-parameter vectors by computing the 2-norm or the ∞-norm of the vector difference. Cepstral distance can be also useful. It makes it possible to distinguish EEG signals recorded a few minutes after an asphyxic cardiac arrest injury [1]. Another way is to compare the power spectra of the data. Thus, the spectral distortion between the natural and the synthetic speech can be based on the COSH distance which is defined from the power spectra of the data to be compared [2]. The log-spectral distance and the Itakura-Saito distance are also widely computed, especially in speech processing. As an alternative, divergences that aim at measuring the similarity between the processes are. Among the questions that can arise, the practitioner may wonder which value of $k$ is relevant. For these reasons, we propose to give an expression of the JD depending on the ARMA parameters. We study how the JD evolves when $k$ increases and we show that after a transient period, the JD is incremented by the same value between two consecutive values of $k$. This phenomenon is always true except when the MA parameter of one or the two processes is equal to 1 or -1. This analysis can help the practitioner see the influences of the ARMA parameters on the JD.

This paper is organized as follows: in sections II and III, we briefly recall the main properties of the ARMA processes and the JD between 1st-order AR processes as well as the JD between two noisy AR processes. In section IV, the JD between ARMA processes is addressed. In section V, theoretical results are illustrated by some examples. In the following, $I_k$ is the identity matrix of size $k$. $(Q)_{a,b}^k$ is the element of the matrix $Q$ which is at the $a$th row and the $b$th column. $Tr$ is the trace of a matrix. The superscript $T$ denotes the transpose.

II. ABOUT REAL AUTOREGRESSIVE MOVING-AVERAGE PROCESSES

A. Definition and properties of an ARMA process

Let us define the ARMA process, $ARMA(1)_{a,b,\sigma^2}$, as follows:

$$x_k = -ax_{k-1} + u_k + bu_{k-1}, \hspace{1cm} (1)$$

where $a$ and $b$ are the process parameters whereas $u$ is the zero-mean driving process with variance $\sigma^2$. It should be noted that $a$, $b$ and $\sigma^2$ are used for the sake of simplicity instead of $a_1$, $b_1$ and $\sigma^2_1$. When $|b| = 1$, the power spectral density of the ARMA process is equal to 0 either at the normalized angular frequency $\theta = 0$ or $\theta = \pm \pi$. Moreover, $0 < |a| < 1$ to guarantee stability. If $a = 0$, the resulting process is a 1st-order MA process. If $b$ is equal to 0, it reduces to a 1st-order AR process, denoted as $AR(1)_{a,\sigma^2}$, and the correlation function is equal to $r_{AR(a,\sigma^2)}(\tau) = \frac{(1-a^2)\sigma^2}{1-2a\cos(\theta) + a^2}$ with $\tau$ the time lag. When both $a$ and $b$ are equal to 0, $x_k$ reduces to a white noise. The correlation function of the ARMA process, denoted as $r_{xx}(\tau)$, satisfies:

$$\begin{align*}
r_{xx}(0) &= r_{AR(a,\sigma^2)}(0) + \frac{1}{\sigma^2_1} (b - 2a)\sigma^2_1, \\
r_{xx}(1) &= r_{xx}(-1) = -ar_{AR(a,\sigma^2)}(0) + \frac{1}{\sigma^2_1} (1 + a^2 - ab)\sigma^2_1, \\
r_{xx}(\tau) &= r_{xx}(-\tau) = -ar_{xx}(\tau - 1), \text{ otherwise.}
\end{align*}$$

(2) In the following, the Toepellization correlation matrix of the vector storing $k$ consecutive values of the ARMA process $ARMA(1)_{a,b,\sigma^2}$ is denoted as $Q_{ARMA(a,b,\sigma^2),k}$ whereas the one related to the process $AR(1)_{a,\sigma^2}$ is denoted as $Q_{AR(1)_{a,\sigma^2},k}$. If the process $AR(1)_{a,\sigma^2}$ is disturbed by an additive white zero-mean Gaussian noise $u_k$ with variance $\sigma_n^2$ and uncorrelated with the driving process $u_a$, the Toepellization correlation matrix of the noisy AR process is $Q_{NAR(a,\sigma^2,\sigma_n^2),k}$. Since there is no straightforward expression of the inverse of $Q_{ARMA(a,b,\sigma^2),k}$ it
can be shown that it can take the form of the correlation matrix of a "noisy" AR as follows:

\[
Q_{ARMA(a,b),k} = \begin{bmatrix}
a - (1 + a^2)b + ab^2 & \frac{b}{a} \sigma^2 I_k \end{bmatrix}
\begin{bmatrix}
Q_{AR(a),k} + \frac{b}{a} \sigma^2 I_k
\end{bmatrix}
\frac{1}{a}
\]

(3)

When \( a = b \) or \( a = \frac{1}{b} \), the gain \( a - (1 + a^2)b + ab^2 \) is equal to 0. In this case, \( Q_{ARMA(a,b),k} \) reduces to the diagonal matrix \( \frac{1}{a} \sigma^2 I_k \). Besides, when \( \frac{1}{a} \sigma^2 > 0 \), this perturbation corresponds to a zero-mean white noise sequence with variance \( \frac{1}{a} \sigma^2 \). More generally, this disturbance can be seen in the frequency domain as an offset equal to \( \frac{1}{a} \sigma^2 \) which can be positive or negative because \( b \) and \( a \) can take any value except zero (otherwise, this is no longer an ARMA process).

In the following, we suggest rewriting (3) as follows:

\[
Q_{ARMA(a,b),k} = Q_{AR(a),k} + \Psi \chi = Q_{NAR(a),k},
\]

(4)

where \( \Psi = \left[ \begin{array}{c}
\frac{a - (1 + a^2)b + ab^2}{a} \\
\frac{b}{a} \sigma^2
\end{array} \right] \) and \( \chi = \frac{1}{a} \sigma^2 \chi \). It is true that they are not necessarily positive and hence cannot be considered properly as variances of white sequences. It should be noted that the AR parameter \( a \) is unchanged after this transformation.

B. Parameter estimation methods

In the noise-free or noisy case, both the AR parameters and the MA parameters as well as the driving-process variances can be predefined or estimated from a set of data. In this latter case and for the noise-free case, the AR parameters can be estimated in many ways: Yule-Walker (YW) equations, correlation method, adaptive filtering, etc. Concerning the MA parameter estimations, one can use the Durbin algorithm, the approach combining the inverse Fourier transform of the inverse of the MA power spectral density (PSD) and the YW equations [6], the “vocariance” methods [7] [8], the one based on higher-order statistics [9], the covariance fitting approaches [10] and the spectral factorization based on the estimation of the outer factor in the PSD such as [11].

In the noisy case, adaptive filters such as the \( \gamma \)-LMS [12] and the \( \rho \)-LMS [13] and extended Kalman filter can estimate the AR parameters from noisy observations. As an alternative, higher-order YW equations, iterative bias compensation schemes such as those proposed by [14], Davila’s method [15] or errors-in-variables (EIV) approaches [16] [17] can be considered. In [18] and [19], the estimations of the ARMA parameters are addressed from noisy observations. More particularly, in [18], Fattah et. al suggest combining approaches that were initially proposed by Davila [15] for the AR-parameter estimations and Stoica [10] for the MA-parameter estimations.

III. DEFINITION OF JEFFREY’S DIVERSION BETWEEN STOCHASTIC PROCESSES

To analyze the dissimilarities between the two random processes \( x \) and \( y \), the JD between the joint distributions of \( k \) successive values of both processes can be computed. It is deduced by symmetrizing the KL expression as follows:

\[
JD_{k}^{(x,y)} = \frac{1}{2} \left( KL_{k}^{(x,y)} + KL_{k}^{(y,x)} \right),
\]

(5)

where \( KL_{k}^{(x,y)} \) denotes the KL divergence between the multivariate densities \( p \) and \( q \). When dealing with Gaussian processes, the KL divergence between two real multivariate Gaussian densities \( p \) and \( q \) with means \( \mu_{x,k} \) and \( \mu_{y,k} \) and covariance matrices \( Q_{x,k} \) and \( Q_{y,k} \) [20] can be obtained by combining the definition of the KL and the expressions of the Gaussian distributions as follows:

\[
KL_{k}^{(x,y)} = \frac{1}{2} \left[ \text{Tr}(Q_{y,k}^{-1}Q_{x,k}) - k - \ln \det Q_{x,k} \right] + (\mu_{x,k} - \mu_{y,k})^{T}Q_{y,k}^{-1}(\mu_{y,k} - \mu_{x,k}).
\]

(6)

Given (5) and (6), the JD becomes for zero-mean processes:

\[
JD_{k}^{(x,y)} = -k + \frac{1}{2} \left[ \text{Tr}(Q_{y,k}^{-1}Q_{x,k}) + \text{Tr}(Q_{x,k}^{-1}Q_{y,k}) \right].
\]

(8)

In this case, the JD is only defined from the \( k \times k \) covariance matrices of the processes. Nevertheless, these matrices must be inverted. When dealing with AR or MA processes, the resulting correlation matrix have specific structures. Various works have been done about the expressions of their inverses. See [21] and [22]. As a consequence, analytical expressions of the JD can be obtained. They depend on the process parameters, but it still implies multiplication of matrices of size \( k \times k \). As an alternative, in [3] and [4], the JD between joint densities of \( k \) samples of AR processes, namely \( \text{AR}_1(a_{1},\sigma_{1}^{2}) \) and \( \text{AR}_2(a_{2},\sigma_{2}^{2}) \), can be deduced recursively as follows:

\[
JD_{AR,k+1} = JD_{AR,k} + A + B,
\]

(9)

with:

\[
A = -1 + \frac{1}{2} \left( R_{ar} + \frac{1}{R_{ar}} \right),
\]

(10)

and:

\[
B = \frac{(a_{2} - a_{1})^{2}}{2} \left( 1 - \frac{1}{(a_{1})^{2}} \right) R_{ar} + \frac{1}{2} - \frac{(a_{2} - a_{1})^{2}}{(a_{1})^{2}} R_{ar}.
\]

(11)

where the driving-process variance ratio is equal to \( R_{ar} = \frac{\sigma_{2}^{2}}{\sigma_{1}^{2}} \), and \( A \) is the JD between the distributions of one sample of each driving process. The difference between two consecutive JD is hence a constant equal to \( A + B \) and can be defined as follows:

\[
\Delta JD_{AR} = JD_{AR,k+1} - JD_{AR,k} = A + B
\]

(12)

This increment \( \Delta JD_{AR} \) is sufficient to compare the AR processes. In a recent work, we suggested studying if the JD follows the same type of property when dealing with noisy AR processes, namely \( \text{NAR}_1(a_{1},\sigma_{1}^{2};\sigma_{n1}^{2}) \) and \( \text{NAR}_2(a_{2},\sigma_{2}^{2};\sigma_{n2}^{2}) \). For this reason, we proposed to deduce this JD denoted as \( JD_{NAR,k}^{(x,y)} \) from the one computed when the processes are noise-free. More particularly, using the inversion matrix lemma [23] we showed that:

\[
JD_{NAR,k+1}^{(1,2)} = JD_{NAR,k+1}^{(1,2)} + C_{k} \frac{1}{2}
\]

(13)

where the offset \( C_{k} \) due to the white additive noises is the sum of six terms:

\[
D_{k,1} = \sigma_{2}^{2} \frac{(a_{2})^{2}}{2} \text{Tr} \left( Q_{AR(a_{1}),\sigma_{1}^{2}}^{-1} \right),
\]

(14)

\[
E_{k,1} = -\sigma_{n1}^{2} \frac{(a_{2})^{2}}{2} \text{Tr} \left( Q_{AR(a_{1}),\sigma_{1}^{2}}^{-1} \right)
\]

(15)

\[
\left( I_{k} + \sigma_{n1}^{2} Q_{AR(a_{1}),\sigma_{1}^{2}}^{-1} \right),
\]

(16)

and three other terms, denoted as \( D_{k,2}, E_{k,2} \) and \( F_{k,2} \) similarly defined as \( D_{k,1}, E_{k,1} \) and \( F_{k,1} \) but where the subscripts are switched. We showed that there is a transient period before the derivatives of these terms with respect to the number of variates \( k \) become constant.

- For \( k \geq 2 \), \( \Delta D_{1} = D_{k+1} - D_{k} \) satisfies:

\[
\Delta D_{1} = \frac{\sigma_{2}^{2}}{\sigma_{1}^{2}} \left( 1 + (a_{1})^{2} \right)
\]

(17)
When $k$ becomes large, $\Delta E_1 = E_{k+1,1} - E_{k,1}$ is equal to:

$$\Delta E_1 = -\frac{\sigma^2_{(2)}}{\sigma^2_{(1)}} \frac{Z}{a_1(1) Z^2 - 1} \left[ (1 + 4a_1)^2 + 4a_1 \right]$$

and:

$$\Delta E_1 = -\frac{\sigma^2_{(2)}}{\sigma^2_{(1)}} \frac{Z}{a_1(1) Z^2 - 1} \left[ (1 + 4a_1)^2 + 4a_1 \right]$$

for $0 < a_1 < 1$, (18)

where:

$$Z = \frac{1}{2} \left[ -\left( \frac{1}{a_1(1)} (1 + \rho(1)) + a_1 \right) \right]$$

and $\rho(1) = \frac{\sigma^2_{(1)}}{\sigma^2_{(n)}} > 0$. (21)

We can also deduce $\Delta F_1$ in the asymptotical case. For this purpose, let us first introduce $S_k^{(1)} = (I_k + \sigma^2_{(1)} Q_{AR(n(1), \sigma^2_{(1)}), k})$. One can show that:

$$S_k^{(1)} = \frac{-\sigma^2_{(1)}}{\sigma^2_{(1)} \alpha(1)} \frac{Z^{2g} - 1 + a_1(1) Z (Z^{2g-2} - 1)}{Z^{2k+2} - 1 + a_2(1) Z (Z^{2k+2} - 1)} \times \frac{\sigma^2_{(2)}}{\sigma^2_{(2)} \alpha(1)} \frac{Z^{2g} - 1 + a_1(1) Z (Z^{2g-2} - 1)}{Z^{2k+2} - 1 + a_2(1) Z (Z^{2k-2} - 1)}$$

for $g \leq h$ and $S_k^{(1)} = S_k^{(1)} h^h$, when $g > h$.

According to (22), the range of values taken by $Z$ varies and depends on the value of the AR parameter $a(1)$. This leads to:

$$S_k^{(1)} = \lim_{k \to +\infty} S_k^{(1)} h^h$$

where $Z > 1$, $a(1) > 0$.

It should be noted that $(Z^2 - 1)$ is necessarily different from 0 as well as $(1 + a(1) Z)$ or $(1 + a(1) Z^{-1})$.

Now, let us consider $T_k^{(1)} = Q_{AR(n(1), \sigma^2_{(1)}), k} S_k^{(1)} Q_{AR(n(1), \sigma^2_{(1)}), k}^{-1}$.

The notation should be $Z^{(1)}$, but we have decided to omit the superscript (1) in the section in order to simplify the presentations of the equations.

where for $g = 2, ..., k - 1$ and $h = 2, ..., k - 1$, it can be shown that:

$$\lim_{k \to +\infty} T_k^{(1)} = T_{lim}^{(1)}$$

(24)

where $T_{lim}^{(1)}$ is the ratio between the variance of the driving process and the variance of the additive white noise:

$$\rho(1) = \frac{\sigma^2_{(1)}}{\sigma^2_{(n)}} > 0.$$
must be adjusted. Indeed, in the JD between noisy AR processes, \( \rho(1) \) defined in (21) was always positive. As \( Z \) could be greater than 1 or between -1 and 0, two expressions of \( S^{(1)}_{\text{ARMA}} \) were proposed in (23). For the ARMA case, \( \rho(1) \) is no longer necessary positive. Therefore, we have to analyze how the properties of \( Z \) evolve in the next subsections.

A. About the properties of \( \rho(1) \) in the ARMA case

When dealing with ARMA processes, according to (4), \( \rho(1) \) becomes equal to:

\[
\rho(1) = \frac{\Psi(1)}{T(1)} = \frac{a(1) - (1 + a^2(1))b(1) + a(1)b(1)}{b(1)}.
\]

(29)

This quantity is hence positive or negative depending on the value of the ARMA parameters. It should be also noted that the same value of \( \rho(1) \) can be obtained by the couples \( (a(1), b(1)) \) and \( (a(1), \frac{1}{b(1)}) \).

B. About the properties of \( Z \) in the ARMA case

Let us now analyze the values of \( Z \) that can take. First of all, due to the new properties of \( \rho(1) \) in the ARMA case, one has to check whether the variable \( Z \) is real or complex. Due to the definition (20) of \( Z \) and the new value of \( \rho(1) \) in (29), it depends on the sign of \( F(\rho(1)) = \left( \frac{1}{\sigma(1)} (1 + \rho(1) + a(1)) \right)^2 - 4 \). If \( \rho(1)_{\min} \) and \( \rho(1)_{\max} \) respectively denotes the minimum and the maximum roots of \( F(\rho(1)) \) among \( (1 + a(1))^2 \) and \( (1 - a(1))^2 \), \( F(\rho(1)) \) would be negative and \( Z \) would be complex for any value of \( a(1) \) in the interval \(-1,1\). Provided \( \rho(1)_{\min} < \rho(1) < \rho(1)_{\max} \). This is however not possible for any \( a(1) \) and \( b(1) \). As a consequence, \( Z \) is always real. Then, it can be easily shown that two cases can happen. Either \( Z > 1 \) or \(-1 < Z < 0 \). See Figure 1.

D. About the expressions of \( \Delta E_1 \) and \( \Delta F_1 \) in the ARMA case

The expression of \( \Delta E_1 \) initially introduced in (18) must be replaced by:

\[
\Delta E_1 = -\frac{\Psi(2)}{\Psi(1)} \frac{Z}{a(1)} \left[ (1 + 4a(1)^2 + a(1)^4) \right]
\]

(31)

for \( \{ a(1) > 0 \} \) or \( \{ b(1) < 0 \} \),

\[
\Delta F_1 = -\frac{\Psi(2)}{\Psi(1)} \frac{Z}{a(1)} \left[ (1 + 4a(1)^2 + a(1)^4) \right]
\]

(32)

for \( \{ a(1) < 0 \} \) or \( \{ b(1) > 0 \} \).

Finally, the expression of \( \Delta F_1 \) initially introduced in (25) must be replaced by:

\[
\Delta F_1 = -\frac{\Psi(2)}{\Psi(1)} \frac{Z}{a(1)} \left[ (1 + 4a(1)^2 + a(1)^4) \right]
\]

(33)

for \( \{ a(1) < 0 \} \) or \( \{ b(1) > 0 \} \).

E. Comments

Except when the MA parameter of one or the two ARMA processes is equal to 1 or -1, the derivative of the JD between ARMA processes with respect to \( k \) becomes a constant when \( k \) is getting higher. As for the AR and the MA case, the resulting asymptotic increment is sufficient to compare the processes. In the next section, let us illustrate these theoretical results by some examples.

V. ILLUSTRATIONS AND COMMENTS

Let us first consider two ARMA processes where \( a(1) = 0.7 \), \( b(1) = 2 \), \( \sigma^2_{a(1)} = 2 \) and \( a(2) = -0.9 \), \( b(2) = 1.1 \) and \( \sigma^2_{a(2)} = 3 \). The results are presented on Figure 2. It is confirmed that the JD between both ARMA processes tends to be constant after a transient period. In addition, on Figure 2 (b), \( JD_{\text{ARMA}+1} - JD_{\text{ARMA}+1} \) is compared with \( JD_{\text{ARMA}} \). This particular case illustrates how the difference between the JDs for two successive instants tends to the constant derived in the theoretical part.

Depending on the ARMA parameters, the convergence speed towards the asymptotic increment is not the same. This phenomenon can be seen by comparing Figure 2 (b) and Figure 3, where \( a(1) = 0.9 \), \( a(2) = 0.7 \), \( b(1) = 2.5 \), \( b(2) = -0.5 \) - \( \sigma^2_{a(1)} = 1 \), \( \sigma^2_{a(2)} = 2 \). Note that the value of the increment is not the same also. In addition, we can see the increment as a function of the AR parameter \( a(2) \), on Figure 4 where \( a(1) = 0.7 \), \( b(1) = 2 \), \( \sigma^2_{a(1)} = 2 \), \( b(2) = 1.1 \). \( \sigma^2_{a(2)} = 3 \), \( k \) is set at 50 whereas \( a(2) \) varies in \([-0.9,0.9]\). On this example, \( k \) is sufficiently large and \( JD_{\text{ARMA}} \) fits \( JD_{\text{ARMA}+1} - JD_{\text{ARMA}} \).

VI. CONCLUSIONS AND PERSPECTIVES

In this paper, we have studied the JD between two real ARMA processes with one AR parameter and one MA parameter. As the correlation matrix of the ARMA process has a structure similar to the one of the AR process, we suggested analyzing how to benefit of the study we did for the JD between AR processes disturbed by additive white noises. This study is also useful to select
We are currently investigating the case of noisy ARMA processes. Asymptotic increment is given and depends on the ARMA parameters. An expression of this asymptotic increment is given and depends on the ARMA parameters. We are currently investigating the case of noisy ARMA processes.

Thus, due to the transient regime, it is better to avoid computing the JD for small values of \( k \). Secondly, the derivative of the JD with respect to the number of variates \( k \) tends to be a constant. The resulting asymptotic increment is relevant of the behavior of the JD and is sufficient to compare ARMA processes. An expression of this asymptotic increment is given and depends on the ARMA parameters. We are currently investigating the case of noisy ARMA processes.

Fig. 4. Evolutions of \( JD_{k+1, k} - JD_{ARMA,k} \) and \( \Delta JD_{ARMA} \) when \( a(2) \) varies.

\[ a(1) = 0.7, a(2) \text{ varies in } [-0.9, 0.9], \]

\[ b(1) = 2, b(2) = 1.1 - \sigma_u^2(1) = 2, \sigma_u^2(2) = 3, k = 50. \]