Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

# Interpreting the asymptotic increment of Jeffrey's divergence between some random processes

Eric Grivel [a], Mahdi Saleh [a,b,*], Samir-Mohamad Omar [b]

[a] Bordeaux University, Bordeaux INP ENSEIRB-MATMECA, IMS, UMR CNRS 5218, Talence, France
[b] Department of Physics and Electronics, Faculty of Sciences I, Lebanese University Hadath, Beirut, Lebanon

ARTICLE INFO

*Article history:*
Available online xxxx

*Keywords:*
Jeffrey's divergence
Kullback–Leibler divergence
Model comparison
ARMA process
Asymptotic analysis
Physical meaning

ABSTRACT

In signal and image processing, Jeffrey's divergence (JD) is used in many applications for classification, change detection, etc. The previous studies done on the JD between ergodic wide-sense stationary (WSS) autoregressive (AR) and/or moving average (MA) processes state that the asymptotic JD increment, which is the difference between two JDs based on $k$ and $(k-1)$-dimensional random vectors when $k$ becomes high, tends to a constant value, except JDs which involve a 1st-order MA process whose power spectral density (PSD) is null for one frequency. In this paper, our contribution is threefold. We first propose an interpretation of the asymptotic JD increment for ergodic WSS ARMA processes: it consists in calculating the power of the first process filtered by the inverse filter associated with the second process and conversely. This explains the atypical cases identified in previous works and generalizes them to any ergodic WSS ARMA process of any order whose PSD is null for one or more frequencies. Then, we suggest comparing other random processes such as noisy sums of complex exponentials (NSCE) by using the JD. In this case, the asymptotic JD increment and the convergence speed towards the asymptotic JD are useful to compare the processes. Finally, NSCE and $p$th-order AR processes are compared. The parameters of the processes, especially the powers of the processes, have a strong influence on the asymptotic JD increment.



## 1. Introduction

In the field of signal and image processing or even in the field of control, models or processes are often compared. This is, for instance, the case when dealing with identification issues where the estimated model parameters are compared with the true ones in order to analyze the estimation accuracy [1] [2]. Model comparison also occurs when designing a Bayesian estimation approach based on Kalman filtering, $H_\infty$ filtering or particle filtering [3]. In this case, *a priori* modeling the system under study is necessary and leads to the state space representation (SSR) of the system. However, several problems may arise. On the one hand, the performance of the estimation algorithm depends on how good the SSR fits the system. As it is not necessarily easy to set it properly, estimation approaches combining different models, or equivalently different SSRs, can be considered. This leads to multiple-model approaches such as the interactive multiple models [4] [5]. In this case, selecting dissimilar models is suggested by Bar-Shalom in [6]. Therefore, a way to compare models *a priori* has to be designed. On the other hand, the practitioners may prefer models whose parameter estimation may be easier and which lead to SSRs that can be written in a simple way. Thus, they can propose to use an autoregressive (AR) model instead of a moving average (MA) model. Therefore, comparing AR and MA models can be useful especially when the power spectral density (PSD) is not null at some frequencies or does not exhibit resonances. Process comparison can also be of interest, especially in the field of image processing when textures are compared [7]. In biomedical applications or flood forecast, change detection can be useful. In this latter case, the problem is to detect whether the statistical properties of a process change over time. In the above situations, statistical properties have to be analyzed.

To address this issue, one could consider the spectral distance measures which include the log-spectral distance (LSD), the Itakura–Saito divergence (ISD), the Itakura divergence (ID), the model distance measure proposed by Itakura and their symmetric versions as well as parametric spectral distances such as the cepstral distance [8]. In [9], a comparative study was recently done between them. The $l$-norm (with $l = 1, 2, \infty$) between the true

* Corresponding author at: Bordeaux University, Bordeaux INP ENSEIRB-MATMECA, IMS, UMR CNRS 5218, Talence, France.
*E-mail addresses:* eric.grivel@ims-bordeaux.fr (E. Grivel), mahdi.saleh@u-bordeaux.fr (M. Saleh), samir.omar@liu.edu.lb (S.-M. Omar).

https://doi.org/10.1016/j.dsp.2018.01.006
1051-2004/

model parameter vector and the estimated one could be also used, as well as the COSH distance [10]. As an alternative, general distance measures [8] can be considered. After the pioneering works of Pearson [11] in 1900, Hellinger [12] in 1909, Bhattacharyya in 1943 [13] and Shannon [14] in 1948 where the measure of entropy and the mutual information were introduced, several researchers focused their attentions on quantifying how close two distributions are from one another: Kullback–Leibler (KL) divergence [15], also known as the relative entropy, generalized the notion of mutual information. In 1952, Chernoff [16] introduced another measure of divergence called Chernoff distance of order $\lambda$. In 1961, Rényi [17] suggested an extension of the entropy of order $\alpha$ for discrete probabilities. Then, the Rényi divergence of order $\alpha$, also called $\alpha$-divergence, was introduced. Then, in 1960$s$, another degree of generalization was proposed through the so-called $f$-divergences where the probability density function (pdf) ratio is weighted by a function $f$. They are also known as Csiszar $f$-divergences, Csiszar–Morimoto divergences or Ali–Silvey distances. Depending on the choice of the function $f$, one can retrieve specific cases. Finally, $f$-dissimilarities have been introduced when more than two pdfs are considered. The reader may refer to [18] [19] for more details and for information about recent works.

Among the above measures, the KL divergence remains one of the most popular. Several authors analyzed it in various fields of applications, for classification, identification or change detection [7], [20], [9], [21] and [22]. Meanwhile, the estimations of the KL between two pdfs that are not necessarily Gaussian, by using sets of data, were studied. See [23] [24].

However, when dealing with Gaussian processes, the expression of the KL depends on the logarithm of the ratio between the covariance matrix determinants. Secondly, the KL divergence is not a distance: it is not symmetric and does not satisfy the triangular inequality. For the above reasons, a great deal of interest has been paid to the symmetric KL divergence, known as Jeffrey's divergence (JD) [25]. When dealing with the JD, the symmetry conditions are satisfied. As the logarithms compensate each other, they no longer appear in the expression of the JD for the Gaussian case. Given all these considerations, we will focus our attention on the JD in the following.

When dealing with $k$-dimensional Gaussian random vectors of size $k$, the JD amounts to computing the sum of two traces of matrices, that can be expressed as the $k \times k$ covariance matrix of the first process pre-multiplied by the inverse of the covariance matrix of the second process. When $k$ increases, the resulting computational cost of the JD increases because the standard computational burden of a generic $k \times k$ matrix inversion is usually $O(k^3)$ [26]. To address this problem with processes that are ergodic WSS autoregressive (AR) and/or moving average (MA), eigenvalue decomposition could be considered *a priori*. Analytical expressions of the eigenvalues and the eigenvectors exist for 1st-order MA processes [27]. Concerning 1st-order AR processes, estimates of eigenvalues have been proposed for a large correlation matrix [28]. However, to the best of our knowledge, these estimates do not exist for higher-order AR processes. For this reason, for a $p$th-order AR process, a LDL factorization could be rather used and requires the parameters of the AR process where the order varies between 1 and $p$. Its computational cost is of $O(\frac{2k^3}{3})$ [26]. Moreover, alternative approaches have been proposed. Taking advantage of the Markovian properties of the AR process, the JD between the pdfs of the $k$ successive samples of two $p$th-order time-varying AR (TVAR) processes or AR processes can be recursively computed [29]. In this case, the expression of the JD for $k$-dimensional vectors only depends on matrices of size $p$, which significantly reduces the computational cost. Then, this method has been used to classify more than two AR processes in different subsets [30]. The analytical expression of the JD between ergodic WSS 1st-order MA processes, that can be real or complex, noise-free or disturbed by additive white Gaussian noises, has also been studied in [31]. For this purpose, the authors use the analytical expression of each element of the tridiagonal-correlation-matrix inverse [32]. Unlike $p$th-order AR processes, no recursive expression of the JD can be obtained for 1st-order MA processes. Finally, comparing ergodic WSS 1st-order AR and ergodic WSS 1st-order MA processes by using the JD has been proposed in [33]. It is based on the expression of the inverses of the AR correlation-matrices [34].

Concerning the above cases, we can summarize the results we obtained as follows:

1) Links with Rao distance [35] have been proposed when it was possible. It was confirmed that the square of the Rao distance was approximately twice the value of the JD, except when a 1st-order MA process is considered whose zero is close to the unit-circle in the z-plane.

2) The JD tends to have a stationary regime. The difference between two JDs computed for $k$ and $(k-1)$-dimensional random vectors tends to a constant when $k$ increases, except for a 1st-order MA process when the zero is on the unit-circle in the z-plane. This difference is called asymptotic JD increment when $k$ becomes high. In previous papers [29] [31] [33], analytical expressions of the asymptotic JD increment are provided for AR and/or MA processes. They depend on the parameters of the processes.

3) The asymptotic JD increment can be used to compare the random processes instead of the JD between $k$ successive samples of the processes. The resulting computation cost is smaller since the analytical expression of the asymptotic JD increment is a known function of the parameters of the processes. Large-size matrix inversions for different values of $k$ and trace computations are no longer required.

4) As the asymptotic JD increment does not depend on $k$, the selection of the number of variates $k$ is no longer a problem for the practitioner.

Although various particular cases have been studied by taking advantage of the expressions of the inverses of the correlation matrices [36][34], the JD between ergodic WSS ARMA processes of orders $p$ and $q$, denoted as ARMA($p$,$q$), has not been addressed yet as there is no explicit expression for the inverses of the correlation matrices. Therefore, it could be of interest to find an alternative approach. In this paper, we suggest giving an interpretation of the asymptotic JD increment in order to better understand the influence of each process parameter on the JD and to generalize the results we obtained to ARMA($p$,$q$) processes. For this reason, our first purpose is to provide a new way to derive the asymptotic JD increment between WSS ARMA processes. It amounts to calculating the power of the first process filtered by the inverse filter associated with the second process and conversely. The second contribution of this paper is to study the JD that involves sums of complex exponentials that are disturbed by additive white noises (NSCE). Furthermore, at the end of the paper, a comparison between an AR process and an NSCE process is addressed by using the JD.

This paper is organized as follows: in section 2, we briefly recall the definitions and properties of the processes under study. In section 3, the expression of the JD is introduced and our contributions are presented. In section 4, we apply our interpretation of the asymptotic JD increment to various random processes. Illustrations are proposed in some cases, especially those based on NSCE processes. Finally, in section 5 we conclude our work. It is followed by an appendix that reveals some necessary derivations.

In the following, $I_k$ is the identity matrix of size $k$ and $Tr$ is the trace of a matrix. The upper-scripts $^T$ and $^H$ denote the transpose and the Hermitian of a matrix.

$x_{k_1:k_2} = \left(x_{k_1}, ..., x_{k_2}\right)$ is the collection of samples from time $k_1$ to $k_2$. $l$ is the label of the process under study. $l = 1, 2$.

Please cite this article in press as: E. Grivel et al., Interpreting the asymptotic increment of Jeffrey's divergence between some random processes, Digit. Signal Process. (2018), https://doi.org/10.1016/j.dsp.2018.01.006