Entropy **2012**, *14*, 1606-1626; doi:10.3390/e14091606

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Kullback–Leibler Divergence Measure for Multivariate Skew-Normal Distributions

Javier E. Contreras-Reyes [1,2,*] and Reinaldo B. Arellano-Valle [3]

[1] División de Investigación Pesquera, Instituto de Fomento Pesquero, Almte, Manuel Blanco Encalada 839, Valparaíso, 2361827, Chile

[2] Departamento de Estadística, Universidad de Valparaíso, Gran Bretaña 1111, Playa Ancha, Valparaíso, 2360102, Chile

[3] Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, 7820436, Chile; E-Mail: reivalle@mat.puc.cl

* Author to whom correspondence should be addressed; E-Mail: jecontrr@mat.puc.cl; Tel.: +56-032-215-1455.

Received: 16 July 2012; in revised form: 25 August 2012 / Accepted: 27 August 2012 / Published: 4 September 2012

Abstract: The aim of this work is to provide the tools to compute the well-known Kullback–Leibler divergence measure for the flexible family of multivariate skew-normal distributions. In particular, we use the Jeffreys divergence measure to compare the multivariate normal distribution with the skew-multivariate normal distribution, showing that this is equivalent to comparing univariate versions of these distributions. Finally, we applied our results on a seismological catalogue data set related to the 2010 Maule earthquake. Specifically, we compare the distributions of the local magnitudes of the regions formed by the aftershocks.

Keywords: skew-normal; cross-entropy; Kullback–Leibler divergence; Jeffreys divergence; earthquakes; nonparametric clustering

---

1. Introduction

The first notion of entropy of a probability distribution was addressed by [1], thus becoming a measure widely used to quantify the level of aleatority present on instrumental variables, which has been commonly used in different engineering areas. Posteriorly, several notions of entropies have been proposed in order to generalize the Shannon entropy, such as Rényi entropy and Tsallis entropy. At this time, [2] introduced the so called Kullback–Leibler divergence (KL-divergence) measures, which is a pseudo-distance (or discriminant function) between two probability distributions and the most common divergence measures used in practical works.

In a recent application about polarimetric synthetic aperture radar (PolSAR) images, Frery *et al.* [3] make use of the complex Wishart distribution (see e.g., [4]) for modeling radar backscatter from forest and pasture areas. There, they conclude that the KL-divergence measure is the best one with respect to Bhattacharyya, Chi-square, Hellinger or Rényi's distances. The studies in [3] and [4] conclude that it is necessary to have appropriate statistics to compare multivariate distributions such as the Wishart one. In addition, [5] gives an extended theoretical analysis of the most important aspects on information theory for the multivariate normal and Student-t distributions, among other distributions commonly used in the literature. Divergence measures have also been used to examine data influences and model perturbations; see, e.g., [6] for a unified treatment and [7] for a review and some extensions of previous works on Bayesian influence measures based on the KL-divergence. In addition, this measure has been considered in selection model analysis by [8], where the authors recommend this criterion because, in contrast to other criteria such as AIC (Akaike's Information Criterion), it does not assume the existence of the true model. However, [8] considers the AIC as a good approximation of the KL-divergence for selection model analysis. On the another hand, asymptotic approximations of the KL-divergence for the multivariate linear model are given in [9], whereas asymptotic approximations of the Jeffreys divergence or simply J-divergence for the multivariate linear model are given in [10]. Another example is the Rényi's divergence and its special case, where the KL-divergence has been recently successfully applied to region-of-interest tracking in video sequences [11], independent subspace analysis [12], image registration [13], and guessing moments [14].

On the other hand, extensive literature has been developed on non-symmetric families of multivariate distributions as the multivariate skew-normal distribution [15–18]. More recently, a study due to [19] computes the mutual information index of the multivariate skew-normal distribution in terms of an infinite series. Next, this work was extended for the full class of multivariate skew-elliptical distributions by [20], where a real application for optimizing an atmospheric monitoring network is presented using the entropy and mutual information indexes of the multivariate skew-normal, among other related family distributions. Several statistical applications to real problems using multivariate skew-normal distributions and others related families can be found in [21].

In this article, we explore further properties of the multivariate skew-normal distribution. This distribution provides a parametric class of multivariate probability distributions that extends the multivariate normal distribution by an extra vector of parameters that regulates the skewness, allowing for a continuous variation from normality to skew-normality [21]. In an applied context, this multivariate family appears to be very important, since in the multivariate case there are not many distributions

available for dealing with non-normal data, primarily when the skewness of the marginals is quite moderate. Considering the multivariate skew-normal distribution as a generalization of the multivariate normal law is a natural choice in all practical situations in which there is some skewness present. For this reason, the main motivation of this article is to analyze some information measures in multivariate observations under the presence of skewness.

Specifically, we propose a theory based on divergence measures for the flexible family of multivariate skew-normal distributions, thus extending the respective theory based on the multivariate normal distribution. For this, we start with the computation of the entropy, cross-entropy, KL-divergence and J-divergence for the multivariate skew-normal distribution. As a byproduct, we use the J-divergence to compare the multivariate skew-normal distribution with the multivariate normal one. Posteriorly, we apply our findings on a seismic catalogue analyzed by [22]. They estimate regions using nonparametric clustering (NPC) methods based on kernel distribution fittings [23], where the spatial location of aftershock events produced by the well-known 2010 earthquake in Maule, Chile is considered. Hence, we compare the skew-distributed local magnitudes among these clusters using KL-divergence and J-divergence; then, we test for significant differences between MLE's parameter vectors [17].

The organization of this paper is as follows. Section 2 presents general concepts of information theory as entropy, cross-entropy and divergence. Section 3 presents the computation of these concepts for multivariate skew-normal distributions, including the special case of the J-divergence between multivariate skew-normal *versus* multivariate normal distribution. Section 4 reports numerical results of a real application of seismic events mentioned before and finally, this paper ends with a discussion in Section 5. Some proofs are presented in Appendix A.

2. Entropy and Divergence Measures

Let $\mathbf{Z} \in \mathbb{R}^k$ be a random vector with probability density function (pdf) $f_{\mathbf{Z}}(\mathbf{z})$. The *Shannon entropy*—also named *differential entropy*—which was proposed earlier by [1] is

$$H(\mathbf{Z}) = -E[\log\ f_{\mathbf{Z}}(\mathbf{Z})] = -\int_{\mathbb{R}^k} f_{\mathbf{Z}}(\mathbf{z}) \log\ f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} \tag{1}$$

Here $E[g(\mathbf{Z})]$ denotes the *expected information* in $\mathbf{Z}$ of the random function $g(\mathbf{Z})$. Hence, the Shannon's entropy is the expected value of $g(\mathbf{Z}) = -\log\ f_{\mathbf{Z}}(\mathbf{Z})$, which satisfies $g(\mathbf{1}) = 0$ and $g(\mathbf{0}) = \infty$.

Suppose now that $\mathbf{X}$, $\mathbf{Y} \in \mathbb{R}^k$ are two random vectors with pdf's $f_{\mathbf{X}}(\mathbf{x})$ and $f_{\mathbf{Y}}(\mathbf{y})$, respectively, which are assumed to have the same support. Under these conditions, the *cross-entropy*—also called *relative entropy*—associated to Shannon entropy (1) is related to compare the information measure of $\mathbf{Y}$ with respect to $\mathbf{X}$, and is defined as follows

$$CH(\mathbf{X}, \mathbf{Y}) = -E[\log\ f_{\mathbf{Y}}(\mathbf{X})] = -\int_{\mathbb{R}^k} f_{\mathbf{X}}(\mathbf{x}) \log\ f_{\mathbf{Y}}(\mathbf{x}) d\mathbf{x} \tag{2}$$

where the expectation is defined with respect to the pdf $f_{\mathbf{X}}(\mathbf{x})$ of the random vector $\mathbf{X}$. It is clear from (2) that $CH(\mathbf{X}, \mathbf{X}) = H(\mathbf{X})$. However, $CH(\mathbf{X}, \mathbf{Y}) \neq CH(\mathbf{Y}, \mathbf{X})$ at least that $\mathbf{X} \overset{d}{=} \mathbf{Y}$, *i.e.*, $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

Related to the entropy and cross-entropy concepts we can also find divergence measures between the distributions of $\mathbf{X}$ and $\mathbf{Y}$. The most well-known of these measures is the so called Kullback–Leibler (KL) divergence proposed by [2] as

$$D_{\mathrm{KL}}(\mathbf{X},\mathbf{Y}) = E\left[\log\left\{\frac{f_{\mathbf{X}}(\mathbf{X})}{f_{\mathbf{Y}}(\mathbf{X})}\right\}\right] = \int_{\mathbb{R}^k} f_{\mathbf{X}}(\mathbf{x}) \log\left\{\frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{x})}\right\} d\mathbf{x} \tag{3}$$

which measures the divergence of $f_{\mathbf{Y}}$ from $f_{\mathbf{X}}$ and where the expectation is defined with respect to the pdf $f_{\mathbf{X}}(\mathbf{x})$ of the random vector $\mathbf{X}$. We note that (3) comes from (2) as $D_{\mathrm{KL}}(\mathbf{X},\mathbf{Y}) = CH(\mathbf{X},\mathbf{Y}) - H(\mathbf{X})$. Thus, we have $D_{\mathrm{KL}}(\mathbf{X},\mathbf{X}) = 0$, but again $D_{\mathrm{KL}}(\mathbf{X},\mathbf{Y}) \neq D_{\mathrm{KL}}(\mathbf{X},\mathbf{Y})$ at least that $\mathbf{X} \overset{d}{=} \mathbf{Y}$, *i.e.*, the KL-divergence is not symmetric. Also, it is easy to see that it does not satisfy the triangular inequality, which is another condition of a proper distance measure (see [24]). Hence it must be interpreted as a pseudo-distance measure only.

A familiar symmetric variant of the KL-divergence is the J-divergence (e.g., [7]), which is defined by

$$J(\mathbf{X},\mathbf{Y}) = E\left[\left\{\frac{f_{\mathbf{X}}(\mathbf{X})}{f_{\mathbf{Y}}(\mathbf{X})} - 1\right\}\log\left\{\frac{f_{\mathbf{X}}(\mathbf{X})}{f_{\mathbf{Y}}(\mathbf{X})}\right\}\right],$$

where the expectation is defined with respect to the pdf $f_{\mathbf{X}}(\mathbf{x})$ of the random vector $\mathbf{X}$. It could be expressed in terms of the KL-divergence d_{KL} [25] as $J(\mathbf{X},\mathbf{Y}) = 2d_{\mathrm{KL}}(\mathbf{X},\mathbf{Y})$ and the KL-divergence as

$$J(\mathbf{X},\mathbf{Y}) = D_{\mathrm{KL}}(\mathbf{X},\mathbf{Y}) + D_{\mathrm{KL}}(\mathbf{Y},\mathbf{X}) \tag{4}$$

As is pointed out in [24], this measure does not satisfy the triangular inequality of distance and hence it is a pseudo-distance measure. The J-divergence has several practical uses in statistics, e.g., for detecting influential data in regression analysis and model comparisons (see [7]).

3. The Multivariate Skew-Normal Distribution

The multivariate skew-normal distribution has been introduced by [18]. This model and its variants have focalized the attention of an increasing number of research. For simplicity of exposition, we consider here a slight variant of the original definition (see [16]). We say that a random vector $\mathbf{Z} \in \mathbb{R}^k$ has a skew-normal distribution with location vector $\boldsymbol{\xi} \in \mathbb{R}^k$, dispersion matrix $\boldsymbol{\Omega} \in \mathbb{R}^{k\times k}$, which is considered to be symmetric and positive definite, and shape/skewness parameter $\boldsymbol{\eta} \in \mathbb{R}^k$, denoted by $\mathbf{Z} \sim SN_k(\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\eta})$, if its pdf is

$$f_{\mathbf{Z}}(\mathbf{z}) = 2\phi_k(\mathbf{z};\boldsymbol{\xi},\boldsymbol{\Omega})\Phi[\boldsymbol{\eta}^\top\boldsymbol{\Omega}^{-1/2}(\mathbf{z}-\boldsymbol{\xi})], \quad \mathbf{z} \in \mathbb{R}^k \tag{5}$$

where $\phi_k(\mathbf{z};\boldsymbol{\xi},\boldsymbol{\Omega}) = |\boldsymbol{\Omega}|^{-1/2}\phi_k(\mathbf{z}_0)$, with $\mathbf{z}_0 = \boldsymbol{\Omega}^{-1/2}(\mathbf{z}-\boldsymbol{\xi})$, is the pdf of the k-variate $N_k(\boldsymbol{\xi},\boldsymbol{\Omega})$ distribution, $\phi_k(\mathbf{z}_0)$ is the $N_k(\mathbf{0},\boldsymbol{I}_k)$ pdf, and Φ is the univariate $N_1(0,1)$ cumulative distribution function. Here $\boldsymbol{\Omega}^{-1/2}$ represents the inverse of the square root $\boldsymbol{\Omega}^{1/2}$ of $\boldsymbol{\Omega}$.

To simplify the computation of the KL-divergence, the following properties of the multivariate skew-normal distribution are useful.

Lemma 1 *Let* $\mathbf{Z} \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta})$*, and consider the vector* $\boldsymbol{\delta} = \boldsymbol{\Omega}\boldsymbol{\eta}/\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta}}$*. Then:*

(i) $\mathbf{Z} \stackrel{d}{=} \boldsymbol{\xi} + \boldsymbol{\delta}|U_0| + \mathbf{U}$*, where* $U_0 \sim N(0,1)$ *and* $\mathbf{U} \sim N_k(\mathbf{0}, \boldsymbol{\Omega} - \boldsymbol{\delta}\boldsymbol{\delta}^\top)$ *and they are independent;*

(ii) $E(\mathbf{Z}) = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\,\boldsymbol{\delta}, \quad \mathrm{var}(\mathbf{Z}) = \boldsymbol{\Omega} - \frac{2}{\pi}\boldsymbol{\delta}\boldsymbol{\delta}^\top$*;*

(iii) For every vector $\mathbf{a} \in \mathbb{R}^k$ *and symmetric matrix* $\mathbf{B} \in \mathbb{R}^{k\times k}$*,*

$$E\{(\mathbf{Z}-\mathbf{a})^\top\mathbf{B}(\mathbf{Z}-\mathbf{a})\} = \mathrm{tr}(\mathbf{B}\boldsymbol{\Omega}) + (\boldsymbol{\xi}-\mathbf{a})^\top\mathbf{B}(\boldsymbol{\xi}-\mathbf{a}) + 2\sqrt{\frac{2}{\pi}}(\boldsymbol{\xi}-\mathbf{a})^\top\mathbf{B}\boldsymbol{\delta};$$

(iv) For every vectors $\tilde{\boldsymbol{\eta}}$, $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^k$*,*

$$\tilde{\boldsymbol{\eta}}^\top(\mathbf{Z}-\tilde{\boldsymbol{\xi}}) \sim SN_1\left(\tilde{\boldsymbol{\eta}}^\top(\boldsymbol{\xi}-\tilde{\boldsymbol{\xi}}), \tilde{\boldsymbol{\eta}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{\eta}}, \frac{\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\delta}}{\sqrt{\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{\eta}} - (\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\delta})^2}}\right).$$

For a proof of Property (i) in Lemma 1, see, e.g., [15,16]. The results in (ii) are straightforward from property (i). Property (iii) comes from (ii) and the well-known fact that $E\{(\mathbf{Z}-\mathbf{a})^\top\mathbf{B}(\mathbf{Z}-\mathbf{a})\} = \mathrm{tr}\{\mathbf{B}E(\mathbf{Z}\mathbf{Z}^\top)\} - 2\mathbf{a}^\top\mathbf{B}E(\mathbf{Z}) + \mathbf{a}^\top\mathbf{B}\mathbf{a}$, see also [26]. For a sketch of the proof of part (iv), see Appendix A.

The calculus of the *cross-entropy* $CH(\mathbf{X},\mathbf{Y})$ when $\mathbf{X} \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\eta}_1)$ and $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\eta}_2)$, requires of the expectation of the functions $(\mathbf{X}-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}(\mathbf{X}-\boldsymbol{\xi}_2)$ and $\log[\Phi\{\boldsymbol{\eta}_2^\top(\mathbf{X}-\boldsymbol{\xi}_2)\}]$. Therefore, the properties (iii) and (iv) in Lemma 1 allow the simplification of the computations of these expected values as is shown in the proof of the lemma given next, and where the following skew-normal random variables will be considered:

$$W_{ij} \sim SN_1\left(\boldsymbol{\eta}_i^\top(\boldsymbol{\xi}_j-\boldsymbol{\xi}_i), \boldsymbol{\eta}_i^\top\boldsymbol{\Omega}_j\boldsymbol{\eta}_i, \frac{\boldsymbol{\eta}_i^\top\boldsymbol{\delta}_j}{\sqrt{\boldsymbol{\eta}_i^\top\boldsymbol{\Omega}_j\boldsymbol{\eta}_i - (\boldsymbol{\eta}_i^\top\boldsymbol{\delta}_j)^2}}\right) \tag{6}$$

where $\boldsymbol{\delta}_j = \boldsymbol{\Omega}_j\boldsymbol{\eta}_j/\sqrt{1+\boldsymbol{\eta}_j^\top\boldsymbol{\Omega}_j\boldsymbol{\eta}_j}$ for $j = 1,2$. Note for $i = j$ that $W_{ii} \sim SN_1\left(0, \boldsymbol{\eta}_i^\top\boldsymbol{\Omega}_i\boldsymbol{\eta}_i, (\boldsymbol{\eta}_i^\top\boldsymbol{\Omega}_i\boldsymbol{\eta}_i)^{1/2}\right)$, with $i = 1,2$. Also, we note that (6) can be expressed as $W_{ij} \stackrel{d}{=} \boldsymbol{\eta}_i^\top(\boldsymbol{\xi}_j-\boldsymbol{\xi}_i) + (\boldsymbol{\eta}_i^\top\boldsymbol{\Omega}_j\boldsymbol{\eta}_i)^{1/2}U_{ij}$, where $U_{ij} \sim SN_1(0,1,\tau_{ij})$, with $\tau_{ij} = \boldsymbol{\eta}_i^\top\boldsymbol{\delta}_j/\sqrt{\boldsymbol{\eta}_i^\top\boldsymbol{\Omega}_j\boldsymbol{\eta}_i - (\boldsymbol{\eta}_i^\top\boldsymbol{\delta}_j)^2}$.

Lemma 2 *The cross-entropy between* $\mathbf{X} \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\eta}_1)$ *and* $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\eta}_2)$ *is given by*

$$CH(\mathbf{X},\mathbf{Y}) = CH(\mathbf{X}_0,\mathbf{Y}_0) + \sqrt{\frac{2}{\pi}}(\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}\boldsymbol{\delta}_1 - E[\log\{2\Phi(W_{21})\}],$$

where

$$CH(\mathbf{X}_0,\mathbf{Y}_0) = \frac{1}{2}\left\{k\log(2\pi) + \log|\boldsymbol{\Omega}_2| + \mathrm{tr}(\boldsymbol{\Omega}_2{}^{-1}\boldsymbol{\Omega}_1) + (\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}(\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)\right\}$$

is the cross-entropy between $\mathbf{Y}_0 \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \mathbf{0})$ *and* $\mathbf{X}_0 \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \mathbf{0})$*, and by (6)*

$$W_{21} \sim SN_1\left(\boldsymbol{\eta}_2^\top(\boldsymbol{\xi}_1-\boldsymbol{\xi}_2), \boldsymbol{\eta}_2^\top\boldsymbol{\Omega}_1\boldsymbol{\eta}_2, \frac{\boldsymbol{\eta}_2^\top\boldsymbol{\delta}_1}{\sqrt{\boldsymbol{\eta}_2^\top\boldsymbol{\Omega}_1\boldsymbol{\eta}_2 - (\boldsymbol{\eta}_2^\top\boldsymbol{\delta}_1)^2}}\right).$$

Proposition 1 *The KL-divergence between* $\mathbf{X} \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\eta}_1)$ *and* $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\eta}_2)$ *is given by*

$$D_{\mathrm{KL}}(\mathbf{X}, \mathbf{Y}) = D_{\mathrm{KL}}(\mathbf{X}_0, \mathbf{Y}_0) + \sqrt{\frac{2}{\pi}}(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)^\top \boldsymbol{\Omega}_2^{-1}\boldsymbol{\delta}_1 + E[\log\{2\Phi(W_{11})\}] - E[\log\{2\Phi(W_{21})\}],$$

where

$$D_{\mathrm{KL}}(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2}\left\{\log\left(\frac{|\boldsymbol{\Omega}_2|}{|\boldsymbol{\Omega}_1|}\right) + \mathrm{tr}(\boldsymbol{\Omega}_2^{-1}\boldsymbol{\Omega}_1) + (\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)^\top \boldsymbol{\Omega}_2^{-1}(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) - k\right\}$$

is the KL-divergence between $\mathbf{X}_0 \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \mathbf{0})$ *and* $\mathbf{Y}_0 \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \mathbf{0})$*, and the* W_{ij} *defined as in (6).*

The proofs of Lemma 2 and Proposition 1 are included in Appendix A. In the following proposition, we give the J-divergence between two multivariate skew-normal distributions. Its proof is immediate from (4) and Proposition 1.

Proposition 2 *Let* $\mathbf{X} \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\eta}_1)$ *and* $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\eta}_2)$*. Then,*

$$\begin{aligned} J(\mathbf{X}, \mathbf{Y}) &= J(\mathbf{X}_0, \mathbf{Y}_0) + \sqrt{\frac{2}{\pi}}(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)^\top(\boldsymbol{\Omega}_2^{-1}\boldsymbol{\delta}_1 - \boldsymbol{\Omega}_1^{-1}\boldsymbol{\delta}_2) \\ &\quad + E[\log\{2\Phi(W_{11})\}] - E[\log\{2\Phi(W_{12})\}] \\ &\quad + E[\log\{2\Phi(W_{22})\}] - E[\log\{2\Phi(W_{21})\}], \end{aligned}$$

where

$$J(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2}\{\mathrm{tr}(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1}) + \mathrm{tr}(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2) + 2(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)^\top(\boldsymbol{\Omega}_1^{-1} + \boldsymbol{\Omega}_2^{-1})(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) - 2k\}$$

is the J-divergence between the normal random vectors $\mathbf{X}_0 \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \mathbf{0})$ *and* $\mathbf{Y}_0 \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \mathbf{0})$*, and by (6) we have that*

$$\begin{aligned} W_{11} &\sim SN_1\left(0, \boldsymbol{\eta}_1^\top\boldsymbol{\Omega}_1\boldsymbol{\eta}_1, (\boldsymbol{\eta}_1^\top\boldsymbol{\Omega}_1\boldsymbol{\eta}_1)^{1/2}\right), \\ W_{21} &\sim SN_1\left(\boldsymbol{\eta}_2^\top(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2), \boldsymbol{\eta}_2^\top\boldsymbol{\Omega}_1\boldsymbol{\eta}_2, \frac{\boldsymbol{\eta}_2^\top\boldsymbol{\delta}_1}{\sqrt{\boldsymbol{\eta}_2^\top\boldsymbol{\Omega}_1\boldsymbol{\eta}_2 - (\boldsymbol{\eta}_2^\top\boldsymbol{\delta}_1)^2}}\right), \\ W_{12} &\sim SN_1\left(\boldsymbol{\eta}_1^\top(\boldsymbol{\xi}_2 - \boldsymbol{\xi}_1), \boldsymbol{\eta}_1^\top\boldsymbol{\Omega}_2\boldsymbol{\eta}_1, \frac{\boldsymbol{\eta}_1^\top\boldsymbol{\delta}_2}{\sqrt{\boldsymbol{\eta}_1^\top\boldsymbol{\Omega}_2\boldsymbol{\eta}_1 - (\boldsymbol{\eta}_1^\top\boldsymbol{\delta}_2)^2}}\right), \\ W_{22} &\sim SN_1\left(0, \boldsymbol{\eta}_2^\top\boldsymbol{\Omega}_2\boldsymbol{\eta}_2, (\boldsymbol{\eta}_2^\top\boldsymbol{\Omega}_2\boldsymbol{\eta}_2)^{1/2}\right). \end{aligned}$$

In what follows we present the KL-divergence and J-divergence for some particular cases only. We start by considering the case where $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ and $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$. Hence, the KL and J divergences compares the location vectors of two multivariate skew-normal distributions, which is essentially equivalent to comparing their mean vectors. For this case we also have that $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2$, $W_{ii} \overset{d}{=} (\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2}W$, $W_{12} \overset{d}{=} -\boldsymbol{\eta}^\top(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) + (\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2}W$ and $W_{21} \overset{d}{=} \boldsymbol{\eta}^\top(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) + (\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2}W$, where $W \sim SN_1(0, 1, (\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2})$. With this notation, the results in Propositions 1 and 2 are simplified in this case as follows.

Corollary 1 *Let* $\mathbf{X} \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}, \boldsymbol{\eta})$ *and* $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}, \boldsymbol{\eta})$. *Then,*

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{X}, \mathbf{Y}) &= D_{\mathrm{KL}}(\mathbf{X}_0, \mathbf{Y}_0) + \sqrt{\frac{2}{\pi}}(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)^\top \boldsymbol{\Omega}^{-1}\boldsymbol{\delta} + E[\log\{2\Phi(\tau W)\}] - E[\log\{2\Phi(\gamma + \tau W)\}], \\ J(\mathbf{X}, \mathbf{Y}) &= J(\mathbf{X}_0, \mathbf{Y}_0) + 2E[\log\{2\Phi(\tau W)\}] - E[\log\{2\Phi(\tau W - \gamma)\}] - E[\log\{2\Phi(\tau W + \gamma)\}], \end{aligned}$$

where $D_{\mathrm{KL}}(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2}\left\{(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) - k + 1\right\}$ *and* $J(\mathbf{X}_0, \mathbf{Y}_0) = 2(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)$ *are, respectively, the KL and J divergences between* $\mathbf{X}_0 \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}, \mathbf{0})$ *and* $\mathbf{Y}_0 \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}, \mathbf{0})$, $\tau = (\boldsymbol{\eta}^\top \boldsymbol{\Omega}\boldsymbol{\eta})^{1/2}$, $\gamma = \boldsymbol{\eta}^\top(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)$ *and* $W \sim SN_1(0, 1, \tau)$.

When $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2$ and $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$, the KL and J divergence measures compare the dispersion matrices of two multivariate skew-normal distributions. For this case we have also that $W_{11} \stackrel{d}{=} W_{21}$ and $W_{12} \stackrel{d}{=} W_{22}$. Consequently, the KL and J divergences does not depend on the skewness/shape vector $\boldsymbol{\eta}$, *i.e.*, it reduces to the respective KL-divergence and J-divergence between two multivariate normal distribution with the same mean vector but different covariance matrices, as is established next.

Corollary 2 *Let* $\mathbf{X} \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}_1, \boldsymbol{\eta})$ *and* $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}_2, \boldsymbol{\eta})$. *Then,*

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{X}, \mathbf{Y}) &= D_{\mathrm{KL}}(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2}\left\{\log\left(\frac{|\boldsymbol{\Omega}_2|}{|\boldsymbol{\Omega}_1|}\right) + \mathrm{tr}(\boldsymbol{\Omega}_2^{-1}\boldsymbol{\Omega}_1) - k\right\}, \\ J(\mathbf{X}, \mathbf{Y}) &= J(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2}\{\mathrm{tr}(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1}) + \mathrm{tr}(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2) - 2k\}, \end{aligned}$$

where $\mathbf{X}_0 \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}_1, \mathbf{0})$ *and* $\mathbf{Y}_0 \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}_2, \mathbf{0})$.

Finally, if $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2$ and $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$, then the KL-divergence and J-divergence compares the skewness vectors of two multivariate skew-normal distributions.

Corollary 3 *Let* $\mathbf{X} \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}_1)$ *and* $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}_2)$. *Then,*

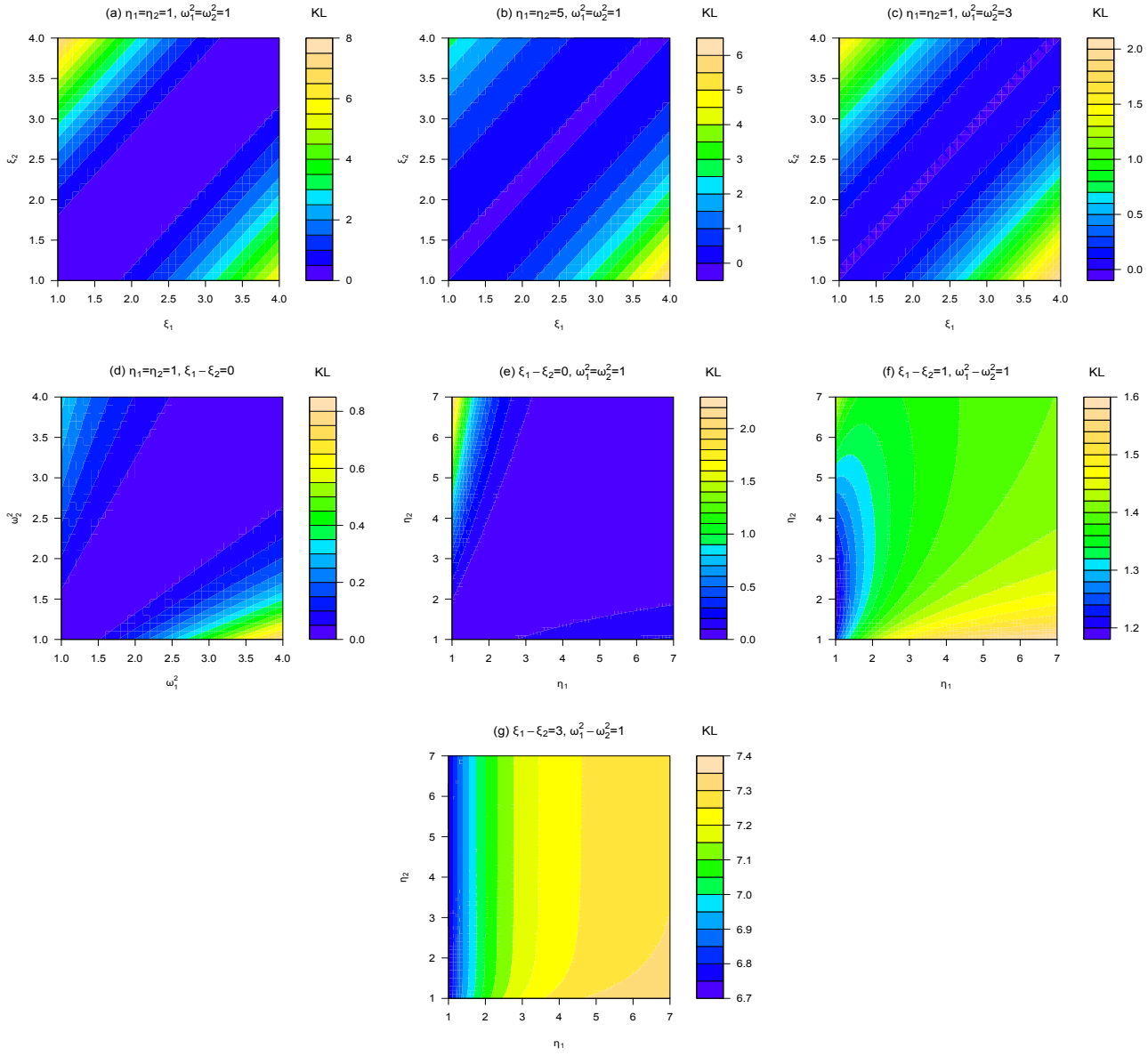
$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{X}, \mathbf{Y}) &= E[\log\{2\Phi(W_{11})\}] - E[\log\{2\Phi(W_{21})\}], \\ J(\mathbf{X}, \mathbf{Y}) &= E[\log\{2\Phi(W_{11})\}] - E[\log\{2\Phi(W_{12})\}] + E[\log\{2\Phi(W_{22})\}] - E[\log\{2\Phi(W_{21})\}], \end{aligned}$$

where

$$\begin{aligned} W_{11} &\sim SN_1\left(0, \boldsymbol{\eta}_1^\top \boldsymbol{\Omega}\boldsymbol{\eta}_1, (\boldsymbol{\eta}_1^\top \boldsymbol{\Omega}\boldsymbol{\eta}_1)^{1/2}\right), \\ W_{21} &\sim SN_1\left(0, \boldsymbol{\eta}_2^\top \boldsymbol{\Omega}\boldsymbol{\eta}_2, \frac{\boldsymbol{\eta}_2^\top \boldsymbol{\delta}_1}{\sqrt{\boldsymbol{\eta}_2^\top \boldsymbol{\Omega}\boldsymbol{\eta}_2 - (\boldsymbol{\eta}_2^\top \boldsymbol{\delta}_1)^2}}\right), \\ W_{12} &\sim SN_1\left(0, \boldsymbol{\eta}_1^\top \boldsymbol{\Omega}\boldsymbol{\eta}_1, \frac{\boldsymbol{\eta}_1^\top \boldsymbol{\delta}_2}{\sqrt{\boldsymbol{\eta}_1^\top \boldsymbol{\Omega}\boldsymbol{\eta}_1 - (\boldsymbol{\eta}_1^\top \boldsymbol{\delta}_2)^2}}\right), \\ W_{22} &\sim SN_1\left(0, \boldsymbol{\eta}_2^\top \boldsymbol{\Omega}\boldsymbol{\eta}_2, (\boldsymbol{\eta}_2^\top \boldsymbol{\Omega}\boldsymbol{\eta}_2)^{1/2}\right). \end{aligned}$$

Figure 1 illustrates the numerical behavior of the KL-divergence between two univariate skew-normal distributions under different scenarios for the model parameters. More specifically, we can observe from there the behavior of the KL-divergence and J-divergence given in Proposition 1 and Corollaries 1–3 for the univariate special cases described below.

Figure 1. Plots of KL-divergence between $\mathbf{X} \sim SN_1(\xi_1, \omega_1^2, \eta_1)$ and $\mathbf{Y} \sim SN_1(\xi_2, \omega_2^2, \eta_2)$. The panels (**a**), (**b**) and (**c**) show that this J-divergence increases mainly with the distance between the location parameters ξ_1 and ξ_2. In the panel (**c**), we can observe that larger values of ω^2 produce the smallest values of KL-divergence, independently of the values of η. In the panel (**d**) is illustrated the case (2) for $\xi = \eta = 1$. In panel (**e**) is illustrated the case (3) for $\xi = 1$ and $\omega^2 = 1$. The panels (**f**) and (**g**) correspond to the KL-divergence of Proposition 1 for $k = 1$.



(1) $X \sim SN_1(\xi_1, \omega^2, \eta)$ *versus* $Y \sim SN_1(\xi_2, \omega^2, \eta)$:

$$D_{\mathrm{KL}}(X,Y) = \frac{1}{2}\left(\frac{\xi_1 - \xi_2}{\omega}\right)^2 + \sqrt{\frac{2}{\pi}}\frac{\gamma}{\sqrt{1+\tau^2}} + E[\log\{2\Phi(\tau W)\}] - E[\log\{2\Phi(\tau W + \gamma)\}],$$

$$J(X,Y) = 2\left(\frac{\xi_1 - \xi_2}{\omega}\right)^2 + 2E[\log\{2\Phi(\tau W)\}] - E[\log\{2\Phi(\tau W - \gamma)\}] - E[\log\{2\Phi(\tau W + \gamma)\}],$$

where $\tau = \omega|\eta|$, $\gamma = \eta(\xi_1 - \xi_2)$ and $W \sim SN_1(0, 1, \tau)$.

(2) $X \sim SN_1(\xi, \omega_1^2, \eta)$ *versus* $Y \sim SN_1(\xi, \omega_2^2, \eta)$:

$$\begin{aligned} D_{\mathrm{KL}}(X, Y) &= \frac{1}{2}\left\{\log\left(\frac{\omega_2^2}{\omega_1^2}\right) + \frac{\omega_1^2}{\omega_2^2} - 1\right\}, \\ J(X, Y) &= \frac{1}{2}\left(\frac{\omega_1^2}{\omega_2^2} - \frac{\omega_2^2}{\omega_1^2} - 2\right), \end{aligned}$$

where $\delta_1 = \eta\omega_1^2/\sqrt{1 + \eta^2\omega_1^2}$.

(3) $X \sim SN_1(\xi, \omega^2, \eta_1)$ *versus* $Y \sim SN_1(\xi, \omega^2, \eta_2)$:

$$D_{\mathrm{KL}}(X, Y) = E[\log\{2\Phi(W_{11})\}] - E[\log\{2\Phi(W_{21})\}],$$

$$J(X, Y) = E[\log\{2\Phi(W_{11})\}] - E[\log\{2\Phi(W_{12})\}] + E[\log\{2\Phi(W_{22})\}] - E[\log\{2\Phi(W_{21})\}],$$

where $W_{ij} \sim SN_1(0, \tau_i^2, s_{ij}\tau_j)$, with $\tau_i = |\eta_i|\omega$ and $s_{ij} = \mathrm{sign}(\eta_i\eta_j)$, $i, j = 1, 2$.

3.1. J-Divergence between the Multivariate Skew-Normal and Normal Distributions

By letting $\boldsymbol{\eta}_1 = \boldsymbol{\eta}$ and $\boldsymbol{\eta}_2 = \mathbf{0}$ in Proposition 2, we have the J-divergence between a multivariate skew-normal and normal distributions, $J(\mathbf{X}, \mathbf{Y}_0)$ say, where $\mathbf{X} \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ and $\mathbf{Y}_0 \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \mathbf{0})$. For this important special case, we find in Corollary 3 that $J(\mathbf{X}_0, \mathbf{Y}_0) = 0$, the random variable W_{21} and W_{22} are degenerate at zero, and $W_{11} \stackrel{d}{=} (\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2}W$, with $W \sim SN_1(0, 1, (\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2})$, and $W_{12} \stackrel{d}{=} (\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2}W_0$, with $W_0 \sim N_1(0, 1)$. This proves the following results.

Corollary 4 *Let* $\mathbf{X} \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ *and* $\mathbf{Y}_0 \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \mathbf{0})$. *Then,*

$$J(\mathbf{X}, \mathbf{Y}_0) = E[\log\{2\Phi(\tau W)\}] - E[\log\{2\Phi(\tau W_0)\}],$$

where $\tau = (\boldsymbol{\eta}^\top\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2}$, $W \sim SN_1(0, 1, \tau)$ *and* $W_0 \sim N_1(0, 1)$.

It follows from Corollary 4 that the J-divergence between the multivariate skew-normal and normal distributions is simply the J-divergence between the univariate $SN_1(0, \tau^2, \tau)$ and $N_1(0, \tau^2)$ distributions. Also, considering that $E[\log\{2\Phi(\tau W)\}] = E[\{2\Phi(\tau W_0)\}\log\{2\Phi(\tau W_0)\}]$, an alternative way to compute the $J(\mathbf{X}, \mathbf{Y}_0)$-divergence is

$$J(\mathbf{X}, \mathbf{Y}_0) = E[\{2\Phi(\tau W_0) - 1\}\log\{2\Phi(\tau W_0)\}] \tag{7}$$

It is interesting to notice that for $\tau = 1$ in (7) we have $J(\mathbf{X}, \mathbf{Y}_0) = E\{(2U_0 - 1)\log(2U_0)\}$, where U_0 is a random variable uniformly distributed on $(0, 1)$, or $J(\mathbf{X}, \mathbf{Y}_0) = E\{U\log(1 + U)\}$, with U being uniformly distributed on $(-1, 1)$. The following remark is suitable when used to compute the expected value $E[\log\{2\Phi(Z)\}]$ for $Z \sim SN_1(\xi, \omega^2, \alpha)$.

Remark 1: If $Z \sim SN_1(\xi, \omega^2, \alpha)$, $\omega > 0$, $\alpha \in \mathbb{R}$, then

$$\begin{aligned} E[\log\{2\Phi(Z)\}] &= E[2\Phi(\alpha Z_0)\log\{2\Phi(\omega Z_0 + \xi)\}] \\ &= E[\Phi(-\alpha S_0)\log\{2\Phi(-\omega S_0 + \xi)\}] + E[\Phi(\alpha S_0)\log\{2\Phi(\omega S_0 + \xi)\}], \end{aligned}$$

where $Z_0 \sim N_1(0,1)$, $S_0 = |Z_0| \sim HN_1(0,1)$, and HN_1 is the univariate half-normal distribution with density $2\phi(s)$, $s > 0$. Since the function $\log[2\Phi(x)]$ is negative on $(-\infty, 0)$ and non-negative on $[0, \infty)$, the last expression is more convenient for the numerical integration.

3.2. Maximum Entropy

We also explore the necessary inequalities to determine the bounds for the entropy of a variable distributed multivariate skew-normal. By [5], for any density $f_{\mathbf{X}}(\mathbf{x})$ of a random vector $\mathbf{X} \in \mathbb{R}^k$—not necessary Gaussian—with zero mean and variance-covariance matrix $\mathbf{\Sigma} = E[\mathbf{X}\mathbf{X}^\top]$, the entropy of $\mathbf{X}$ is upper bounded as

$$H(\mathbf{X}) \leq \frac{1}{2}\log\left\{(2\pi e)^k|\mathbf{\Sigma}|\right\} \tag{8}$$

and

$$H(\mathbf{X}_0) = \frac{1}{2}\log\left\{(2\pi e)^k|\mathbf{\Omega}|\right\} \tag{9}$$

is the entropy of $\mathbf{X}_0 \sim N_k(\mathbf{0}, \mathbf{\Omega})$, *i.e.*, the entropy is maximized under normality. Let $\mathbf{X} \sim SN_k(\boldsymbol{\xi}, \mathbf{\Omega}, \boldsymbol{\eta})$, our interest is now to give an alternative approximation of the entropy of the skew-normal random vector $\mathbf{X}$. By [20] or by Lemma 2, we have that the skew-normal entropy is

$$H(\mathbf{X}) = \frac{1}{2}\log\left\{(2\pi e)^k|\mathbf{\Omega}|\right\} - E\left[\log\left\{2\Phi(\tau W)\right\}\right],$$

where $W \sim SN_1(0, 1, \tau)$ with $\tau = (\boldsymbol{\eta}^\top\mathbf{\Omega}\boldsymbol{\eta})^{1/2}$. By (8), (9) and Property (ii) of Lemma 1, we have that

$$\begin{aligned} H(\mathbf{X}) &\leq \frac{1}{2}\log(2\pi e)^k + \frac{1}{2}\log\left|\mathbf{\Omega} - \frac{2}{\pi}\boldsymbol{\delta}\boldsymbol{\delta}^\top\right| \\ &= H(\mathbf{X}_0) + \frac{1}{2}\log\left(1 - \frac{2}{\pi}\boldsymbol{\delta}^\top\mathbf{\Omega}^{-1}\boldsymbol{\delta}\right) \\ &= H(\mathbf{X}_0) + \frac{1}{2}\log\left(1 - \frac{2}{\pi}\frac{\tau^2}{1+\tau^2}\right), \end{aligned}$$

since $\boldsymbol{\delta} = \mathbf{\Omega}\boldsymbol{\eta}/\sqrt{1 + \boldsymbol{\eta}^\top\mathbf{\Omega}\boldsymbol{\eta}}$ and so $\boldsymbol{\delta}^\top\mathbf{\Omega}^{-1}\boldsymbol{\delta} = \tau^2/(1+\tau^2)$. Therefore, we obtain a lower bound for the following expected value

$$E\left[\log\left\{2\Phi(\tau W)\right\}\right] \geq -\frac{1}{2}\log\left(1 - \frac{2}{\pi}\frac{\tau^2}{1+\tau^2}\right).$$

Note from this last inequality that $E\left[\log\left\{2\Phi(\tau W)\right\}\right]$ is always positive, because $2\tau^2/\pi(1+\tau^2) < 1$.

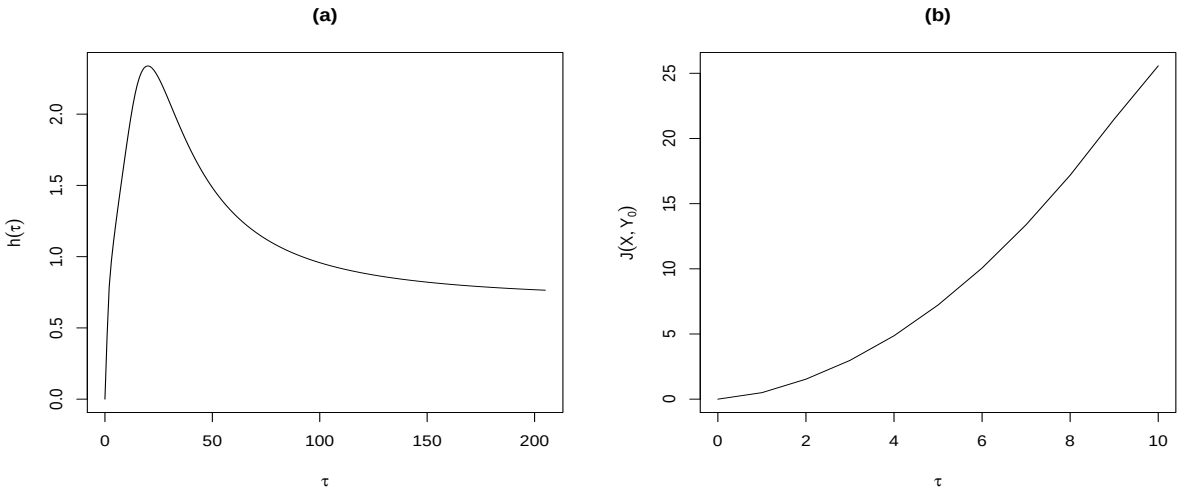
On the other hand, [27] uses the *Negentropy* H_N to quantify the non-normality of a random variable $\mathbf{X}$, which defined as

$$H_N(\mathbf{X}) = H(\mathbf{X}_0) - H(\mathbf{X}),$$

where $\mathbf{X}_0$ is a normal variable with the same variance as that of $\mathbf{X}$. The *Negentropy* is always nonnegative, and will become even larger as the random variable and is farther from the normality. Then, the *Negentropy* of $\mathbf{X} \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ coincides with $E[\log\{2\Phi(\tau W)\}]$. As is well-known, the entropy is a measure attributed to uncertainty of information, or a randomness degree of a single variable. Therefore, the *Negentropy* measures the departure from the normality of the distribution of the random variable $\mathbf{X}$. To determine a symmetric difference of a Gaussian random variable with respect to its skewed version, *i.e.*, that preserves the location and dispersion parameters but incorporates a shape/skewness parameter, the J-divergence presented in Section 3.2 is a useful tool to analyze this fact.

Figure 2 shows in panel (a) several values of $h(\tau) = E[\log\{2\Phi(\tau W)\}]$ for $\tau = 0, 1, ..., 200$. It is interesting to notice that the maximum value of this expected value is approximately equal to 2.339. In the panel (b) this figure shows the values of J-divergence between $X \sim SN_1(0, \tau^2, \tau)$ and $Y_0 \sim N(0, \tau^2)$ computed in (7) for $\tau = 0, 1, ..., 10$.

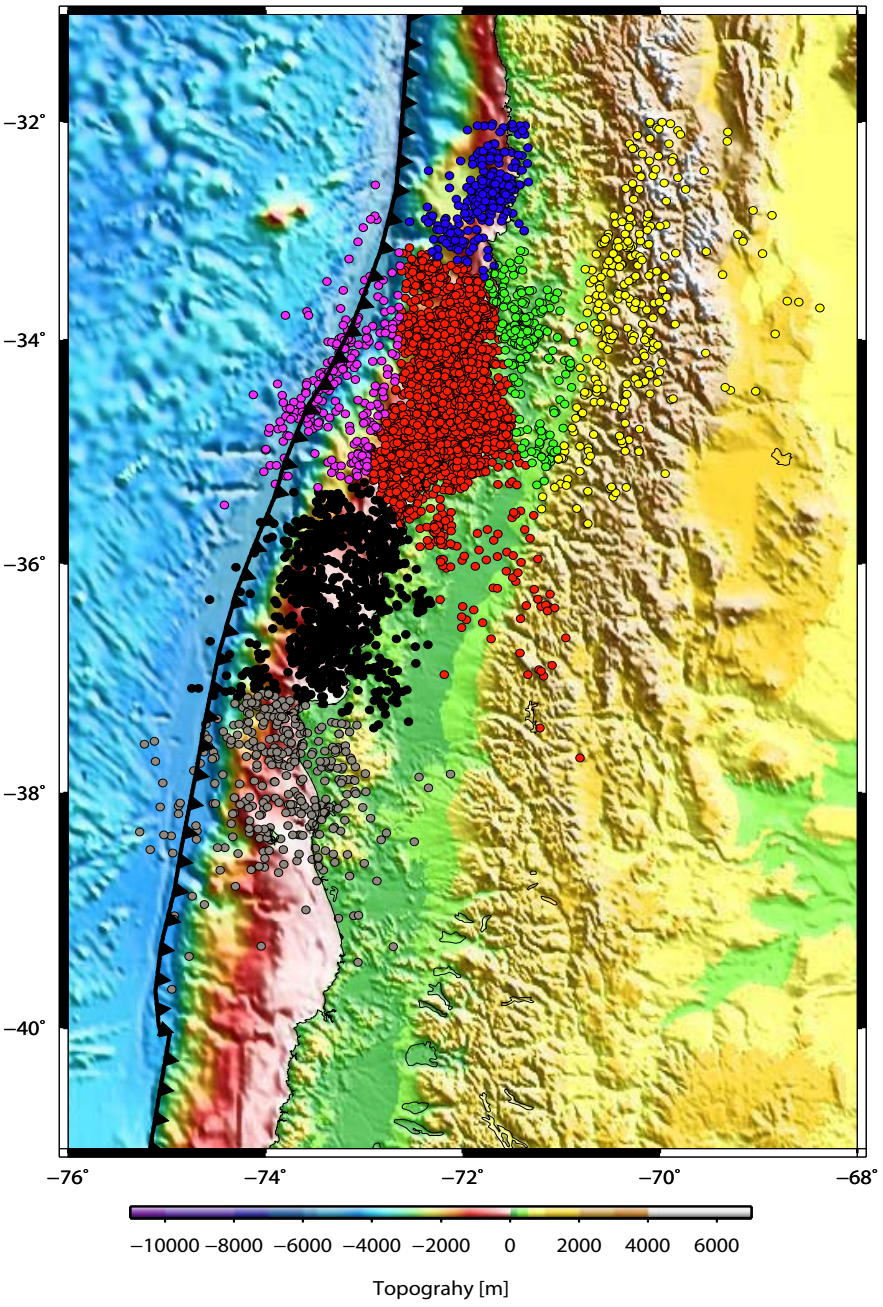
Figure 2. (**a**) Behavior of $h(\tau) = E[\log\{2\Phi(\tau W)\}]$ for $\tau = 0, 1, ..., 200$. (**b**) Behavior of $J(X, Y_0)$ for $\tau = 0, 1, ..., 10$.



4. Statistical Application

For a statistical application of this paper, we consider the seismic catalogue of the Servicio Sismológico Nacional of Chile (SSN, [28]) analyzed by [22], containing 6,714 aftershocks on a map [32–40°S]×[69–75.5°E] for a period between 27 February 2010 to 13 July 2011. Our main goal is to compare the aftershock distributions of local and moment magnitudes (M_l and M_w, respectively) using the KL-divergence and J-divergence between clusters detected by nonparametric clustering (NPC) method developed by [23] (See Figure 3). This method allows the detection of subsets of points forming clusters associated with high density areas which hinge on an estimation of the underlying probability density function via a nonparametric kernel method for each of these clusters. Consequently, this methodology has the advantage of not requiring some subjective choices on input, such as the number of existing clusters. The aftershock clusters analyzed by [22] consider the high density areas with respect to its map positions, *i.e.*, they consider the bi-dimensional distribution of latitude-longitude joint variable to be estimated by the kernel method. For more details about the NPC method, see also [23].

Figure 3. Left: Map of the Chile region analyzed for post-seismicity correlation with clustering events: black (1), red (2), green (3), blue (4), violet (5), yellow (6) and gray (7).



Depending on the case, we consider it pertinent to analyze the measures of J-divergences between a cluster sample fitted by a skew-normal distribution *versus* the same sample fitted by a normal distribution, where the fits are previously diagnosed by QQ-plots (see e.g., [20]). The MLE's of the model parameters are obtained by using the *sn* package developed by [29] and described later in Section 4.1; the entropies, cross-entropies, KL-divergence and J-divergences presented in the previous sections are computed using *skewtools* package developed by [30]. Both packages are implemented in R software [31]. In Section 4.2 we present the Kupperman test [32] based on asymptotic approximation of the KL-divergence statistic to chi-square distribution with degrees of freedom depending on the dimension of the parametric space.

4.1. Likelihood Function

In order to examine the usefulness of the KL-divergence and J-divergence between multivariate skew-normal distributions developed in this paper, we consider the MLE's of the location, dispersion and shape/skewness parameters proposed by [17] (and considered for a meteorological application in [20]) for a sample of independent observations $\mathbf{Z}_i \sim SN_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta})$, $i = 1, ..., N$. We estimate the parameters by numerically maximizing the log-likelihood function:

$$\log\ L(\boldsymbol{\theta}) \propto -\frac{N}{2}\log\ |\boldsymbol{\Omega}| - \frac{N}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-1}\mathbf{V}) + \sum_{i=1}^{N}\log\ [\Phi\{\boldsymbol{\eta}^{\top}\boldsymbol{\Omega}^{-1/2}(\mathbf{z}_i - \boldsymbol{\xi})\}],$$

where

$$\mathbf{V} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \boldsymbol{\xi})(\mathbf{z}_i - \boldsymbol{\xi})^{\top}$$

and $\boldsymbol{\theta} = \{\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}\}$. Then, we can obtain $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\eta}}\} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ using the Newton–Rhapson method. This method works well for distributions with a small number of k-variables. Other similar methods such as the EM algorithm tend to have more robust relativity but run slower than MLE.

4.2. Asymptotic Test

Following [32,33], in this section we consider the asymptotic properties of the likelihood estimator of the J-divergence between the distributions of two random vectors $\mathbf{X}$ and $\mathbf{Y}$. For this, it is assumed that $\mathbf{X}$ and $\mathbf{Y}$ have pdf indexed by unknown parameters vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, which belong to the same parameters space. Let $\widehat{\boldsymbol{\theta}}_1 = (\widehat{\theta}_{11}, \ldots, \widehat{\theta}_{1p})^{\top}$ and $\widehat{\boldsymbol{\theta}}_2 = (\widehat{\theta}_{21}, \ldots, \widehat{\theta}_{2p})^{\top}$ be the MLE's of the parameter vectors $\boldsymbol{\theta}_1 = (\theta_{11}, \ldots, \theta_{1p})^{\top}$ and $\boldsymbol{\theta}_2 = (\theta_{21}, \ldots, \theta_{2p})^{\top}$, respectively, based on independent samples of size N_1 and N_2 from the distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Denote by $J(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ the J-divergence between the distributions of $\mathbf{X}$ and $\mathbf{Y}$, and consider the statistic defined by

$$S_{KL}(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2) = \frac{N_1 N_2}{N_1 + N_2} J(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2).$$

Under the regularity conditions discussed by [33], it follows that if $\frac{N_1}{N_1+N_2} \underset{N_1,N_2\to\infty}{\longrightarrow} \lambda$, with $0 < \lambda < 1$, then under the homogeneity null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$,

$$S_{\mathrm{KL}}(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2) \underset{N_1,N_2\to\infty}{\xrightarrow{d}} \chi^2_p \tag{10}$$

where "$\xrightarrow{d}$" means convergence in distribution.

Based on (10) an asymptotic statistical hypothesis tests for the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ can be derived. Consequently, it can be implemented in terms of the J-divergence (or the KL-divergence, as in [3]) between the multivariate skew-normal distributions $\mathbf{X} \sim SN_k(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\eta}_1)$ and $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\eta}_2)$, for which $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the corresponding $p = 2k + k(k+1)/2$ different unknown parameters in $\{\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\eta}_1\}$ and $\{\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\eta}_2\}$, respectively. Hence, (10) allows testing through the P-value if the homogeneity null hypothesis $H_0 : \boldsymbol{\xi}_1 = \boldsymbol{\xi}_2, \boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2, \boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$ is rejected or not. Thus, if for large values of N_1, N_2 we observe $S_{\mathrm{KL}}(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2) = s$, then the homogeneity null hypothesis can be rejected at level α if $P(\chi^2_p > s) \leq \alpha$.

4.3. Main Results

The skew-normal KL-divergence values of Proposition 1 for each pair of clusters are reported in Table 1. By (4), we can obtain the symmetrical skew-normal J-divergences to compare the parametric differences between these clusters. The MLE's of the unknown parameters for the distribution of each cluster are shown in Table 2 with its respective descriptive statistics and estimated skew-normal model parameters. In the Appendix B, we attach the Figures A1 and A2 which indicate the performance of the fitted models, where we also include the QQ-plots for the normal and skew-normal cases. These QQ-plots represent the dispersion of the Mahalanobis distances related to the theoretical parameters, with respect to the empirical percentiles of the chi-square distribution. It follows from there that as the dispersion line is fitted by the theoretical line in a greater degree, the skew-normal fit will have better performance. The diagnostic QQ-plots are also possible to obtain by using the *sn* package developed by [29].

Table 1. KL-divergences for pairs of clusters.

Cluster	**black (1)**	**red (2)**	**green (3)**	**blue (4)**	**violet (5)**	**yellow (6)**	**gray (7)**
black (1)	0	0.178	0.149	0.008	0.262	0.041	0.835
red (2)	0.219	0	0.743	0.267	0.018	0.455	0.273
green (3)	0.181	0.601	0	0.102	0.909	0.038	1.688
blue (4)	0.015	0.234	0.095	0	0.374	0.015	0.981
violet (5)	0.212	0.018	0.721	0.269	0	0.437	0.216
yellow (6)	0.053	0.350	0.031	0.020	0.530	0	1.194
gray (7)	0.978	0.224	1.887	1.032	0.274	1.398	0

We can see from Table 1 that the grey (7) cluster has the larger discrepancy with respect to the other clusters, except with respect to red (2) and violet (5) clusters, which is mainly due to the location and shape fitted parameters (see Table 2). A counterpart case is found for the green (3) cluster, which presents greater differences with respect to these two aforementioned clusters. On the other hand, the diagnostic QQ-plots show good performance of the skew-normal fit with respect to the normal case, although we should observe here that the red (2) cluster is being affected by an outlier observation corresponding to the greater magnitude $M_w = 8.8$. However, this fit considers that the probability of a similar occurrence in the future of a great event like this is practically zero.

Given that the seismic observations have been classified by the NPC method considering their positions on the map, the KL-divergence and J-divergence based on magnitudes proposed in this paper are not valid tools to corroborate the clustering method. Nevertheless, these measures corroborate some similarities in the distributions of those clusters localized away from the epicenter as, e.g., red (2)–violet (5) and green (3)–yellow (6), as well as some discrepancies in the distributions of some clusters as, e.g., red (2)–green (3), red (2)–blue (4) and gray (7)–black (1). All of these similarities and discrepancies were evaluated through the Kupperman test (10). Table 3 reports the statistic test values and the corresponding P-values obtained by comparing the asymptotic reference chi-square distribution with $p = 3$ degrees of freedom ($k = 1$). We can see that this test corroborates the similarities in the distribution of the clusters red (2)–violet (5) and green (3)–yellow (6), but this test also suggests

similarities for the black (1)–blue (4) and blue (4)–yellow (6) clusters. These results are consistent with the values of the fitted parameters, as we can see in Table 2. In this last table we have also presented the values of the parameter $\tau = (\eta^2\Omega)^{1/2}$ for each cluster and the divergence $J(X, Y_0)$ between skew-normal and normal distributions defined in Equation (7). Specifically, since the shape/skewness parameters of red (2) and gray (7) clusters are the smallest, it is then evident that the lower values for the divergence $J(X, Y_0)$ correspond to these clusters, a result that is consistent with the panel (b) Figure 2.

Table 2. Mean and standard deviation (sd) from the normal fit, minimum (min), maximum (max) and number of observations (N) for each cluster and for the full data (see "Total" below); skew-normal MLE's and their respective standard deviations (in brackets) for each and the full cluster; τ and $J(X, Y_0)$ values for each and the full cluster.

Cluster	**Descriptive Statistics**					**Skew-normal fit**				
	mean	**sd**	**min**	**max**	N	ξ	Ω	η	τ	$J(X, Y_0)$
black (1)	3.427	0.655	2.0	6.6	4182	3.430 (0.010)	0.651 (0.008)	0.756 (0.020)	0.610	0.211
red (2)	3.924	0.769	2.1	8.8	962	3.927 (0.025)	0.766 (0.019)	0.445 (0.068)	0.389	0.092
green (3)	3.085	0.615	2.0	5.2	265	3.081 (0.038)	0.618 (0.030)	0.711 (0.105)	0.559	0.181
blue (4)	3.339	0.729	2.0	6.1	280	3.337 (0.043)	0.730 (0.035)	0.697 (0.101)	0.595	0.202
violet (5)	3.852	0.682	2.6	6.8	265	3.858 (0.041)	0.673 (0.033)	0.820 (0.067)	0.673	0.252
yellow (6)	3.215	0.666	2.1	5.2	215	3.201 (0.047)	0.683 (0.040)	0.805 (0.128)	0.665	0.247
gray (7)	4.447	0.695	2.7	6.9	332	4.447 (0.038)	0.694 (0.029)	0.453 (0.124)	0.378	0.087
Total	3.539	0.743	2.0	8.8	6584	3.539 (0.009)	0.743 (0.007)	0.731 (0.018)	0.629	0.224

Table 3. J-divergences for each pair of clusters. The statistic values and P-values of the asymptotic test are given in brackets. Those marked in bold correspond to the P-values higher than a probability 0.04 related to a 4% significance level.

Cluster	**black (1)**	**red (2)**	**green (3)**	**blue (4)**	**violet (5)**	**yellow (6)**	**gray (7)**
black (1)	0	0.397	0.330	0.023	0.475	0.093	1.814
	(0; 1)	(311; 0)	(82; 0)	(**6.1**; **0.106**)	(118; 0)	(19; 0)	(558; 0)
red (2)	0.397	0	1.344	0.501	0.037	0.805	0.497
	(311; 0)	(0; 1)	(279; 0)	(109; 0)	(**7.6**; **0.055**)	(142; 0)	(123; 0)
green (3)	0.330	1.344	0	0.197	1.630	0.069	3.575
	(82; 0)	(279; 0)	(0; 1)	(27; 0)	(216; 0)	(**8.1**; **0.043**)	(527; 0)
blue (4)	0.023	0.501	0.197	0	0.642	0.035	2.014
	(**6.1**; **0.106**)	(109; 0)	(27; 0)	(0; 1)	(88; 0)	(**4.2**; **0.239**)	(306; 0)
violet (5)	0.475	0.037	1.630	0.642	0	0.967	0.490
	(118; 0)	(**7.6**; **0.055**)	(216; 0)	(88; 0)	(0; 1)	(115; 0)	(72; 0)
yellow (6)	0.093	0.805	0.069	0.035	0.967	0	2.593
	(19; 0)	(142; 0)	(**8.1**; **0.043**)	(**4.2**; **0.239**)	(115; 0)	(0; 1)	(338; 0)
gray (7)	1.814	0.497	3.575	2.014	0.490	2.593	0
	(558; 0)	(123; 0)	(527; 0)	(306; 0)	(72; 0)	(338; 0)	(0; 1)

5. Conclusions

We have presented a methodology to compute the Kullback–Leibler divergence for multivariate data presenting skewness, specifically, for data following a multivariate skew-normal distribution. The calculation of this measure is semi-analytical, since it is the sum of two analytical terms, one corresponding to the multivariate normal Kullback–Leibler divergence and the other depending on the location, dispersion and shape parameters, and a third term which must be computed numerically and which was reduced from a multidimensional integral to an integral in only one dimension. Numerical experiments have shown that the performance of this measure is consistent with its theoretical properties. Additionally, we have derived expressions for the J-divergence between different multivariate skew-normal distributions, and in particular for the J-divergence between the skew-normal and normal distributions. The presented entropy and KL-divergence concepts for this class of distributions are necessary to compute other information tools as mutual information.

This work has also presented a statistical application related to aftershocks produced by the Maule earthquake which occurred on 27 February 2010. The results shown that the proposed measures are useful tools for comparing the distributions of magnitudes of events related to the regions near the epicenter. We also consider an asymptotic homogeneity test for the cluster distributions under the skew-normality assumption and, consequently, confirm the found results in a consistent form.

Acknowledgments

Arellano-Valle's research work was partially supported by FONDECYT, Chile, with grant No. 1120121. The authors wish to thank the editor and two anonymous referees for some helpful comments.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
2. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
3. Frery, A.C.; Nascimento, A.; Cintra, R. Information theory and image understanding: An application to polarimetric SAR imagery. *Chil. J. Stat.* **2011**, *2*, 81–100.
4. Gupta, M.; Srivastava, S. Parametric Bayesian estimation of differential entropy and relative entropy. *Entropy* **2010**, *12*, 818–843.
5. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley & Son, Inc.: New York, NY, USA, 1991; p. 774.
6. Weiss, R., Cook, D. A graphical case statistic for assessing posterior influence. *Biometrika* **1992**, *79*, 51–55.
7. Arellano-Valle, R.B.; Galea-Rojas, M.; Iglesias, P. Bayesian sensitivity analysis in elliptical linear regression models. *J. Stat. Plan. Infer.* **2000**, *86*, 175–199.
8. Burnham, K.P.; Anderson, D.R. *Model Selection and Inference: A Practical Information Theoretic Approach*; Springer-Verlag: New York, NY, USA, 1998; p. 353.
9. Seghouane, A.K. New AIC corrected variants for multivariate linear regression model selection. *IEEE Trans. Aerosp. Electron. Syst.* **2011**, *47*, 1154–1165.

10. Seghouane, A.K. Multivariate regression model selection from small samples using Kullback's symmetric divergence. *Signal Process.* **2006**, *86*, 2074–2084.
11. Boltz, S.; Debreuve, E.; Barlaud, M. High-dimensional statistical measure for region-of-interest tracking. *Trans. Img. Proc.* **2009**, *18*, 1266–1283.
12. Póczos, B.; Szabó, Z.; Schneider, J. Nonparametric divergence estimators for independent subspace analysis. In Proceedings of the European Signal Processing Conference, Barcelona, Spain, 2 September 2011; pp. 1849–1853.
13. Durrani, T.S.; Zeng, X. Copula based divergence measures and their use in image registration. In Proceedings of the European Signal Processing Conference, Glasgow, Scotland, 24–28 August 2009; pp. 1309–1313.
14. van Erven, T.; Harremoës, P. Rényi divergence and majorization. In Proceedings of the 2010 IEEE International Symposium on Information Theory Proceedings, Amsterdam, The Netherlands, 13–18 June 2010; pp. 1335–1339.
15. Arellano-Valle, R.B.; Azzalini, A. On the unification of families of skew-normal distributions. *Scand. J. Stat.* **2006**, *33*, 561–574.
16. Arellano-Valle, R.B.; Genton, M.G. On fundamental skew distributions. *J. Multivariate Anal.* **2005**, *96*, 93–116.
17. Azzalini, A.; Capitanio, A. Statistical applications of the multivariate skew normal distributions. *J. Roy. Stat. Soc. Ser. B* **1999**, *61*, 579–602.
18. Azzalini, A.; Dalla-Valle, A. The multivariate skew-normal distribution. *Biometrika* **1996**, *83*, 715–726.
19. Javier, W.R.; Gupta, A.K. Mutual information for certain multivariate distributions. *Far East J. Theor. Stat.* **2009**, *29*, 39–51.
20. Arellano-Valle, R.B.; Contreras-Reyes, J.E.; Genton, M.G. Shannon entropy and mutual information for multivariate skew-elliptical distributions. *Scand. J. Stat.* **2012**, doi:10.1111/j.1467-9469.2011.00774.x.
21. Genton, M.G. *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2004; Edited Volume, p. 416.
22. Contreras-Reyes, J.E.; Azzalini, A. On the spatial correlation between areas of high coseismic slip and aftershock clusters of the Maule earthquake Mw=8.8. *arXiv* **2012**, arXiv:1208.1517v1.
23. Azzalini, A.; Torelli, N. Clustering via nonparametric density estimation. *Stat. Comput.* **2007**, *17*, 71–80.
24. Ullah, A. Entropy, divergence and distance measures with econometric applications. *J. Stat. Plan. Infer.* **1996**, *49*, 137–162.
25. Kullback, S. The Kullback–Leibler distance. *Am. Stat.* **1987**, *41*, 340–341.
26. Genton, M.G.; He, L.; Liu, X. Moments of skew-normal random vectors and their quadratic forms. *Stat. Probabil. Lett.* **2001**, *51*, 319–325.
27. Gao, J.-H.; Zhang, B. Estimation of seismic wavelets based on the multivariate scale mixture of Gaussians model. *Entropy* **2010**, *12*, 14–33.
28. Servicio Sismológico. Departamento de Geofísica, Universidad de Chile, Chile. Available online: http://ssn.dgf.uchile.cl/ (accessed on 23 March 2012).

29. Azzalini, A. R package sn: The skew-normal and skew-t distributions (version 0.4-6). Universitá di Padova, Padova, Italy, 2008. Available online: http://cran.r-project.org/web/packages/sn (accessed on 6 May 2012).
30. Contreras-Reyes, J.E. R package skewtools: Tools for analyze skew-elliptical distributions and related models (version 0.1.1). Instituto de Fomento Pesquero, Valparaíso, Chile, 2012. Available online: http://cran.rproject.org/web/packages/skewtools (accessed on 6 May 2012).
31. R Development Core Team. *R: A Language and environment for statistical Computing*; R Foundation for Statistical Computing, Vienna, Austria, 2012; ISBN 3-900051-07-0. Available online: http://www.R-project.org (accessed on 5 May 2012).
32. Kupperman, M. *Further applications of information theory to multivariate analysis and statistical Inference*. Ph.D. dissertation, The George Washington, University, Washington, DC, USA, 1957, p. 270.
33. Salicrú, M.; Menéndez, M.L.; Pardo, L.; Morales, D. On the applications of divergence type measures in testing statistical hypothesis. *J. Multivariate Anal.* **1994**, *51*, 372–391.

Appendices

Appendix A. Some Proofs

Proof of part (iv) of Lemma 1: From Lemma 1(i) we have

$$\tilde{\boldsymbol{\eta}}^\top(\mathbf{Z}-\tilde{\boldsymbol{\xi}}) \stackrel{d}{=} \tilde{\boldsymbol{\eta}}^\top(\boldsymbol{\xi}-\tilde{\boldsymbol{\xi}}) + \tilde{\boldsymbol{\eta}}^\top\boldsymbol{\delta}|U_0| + \tilde{\boldsymbol{\eta}}^\top\mathbf{U}.$$

Since $\tilde{\boldsymbol{\eta}}^\top\mathbf{U} \sim N_1(0, \tilde{\boldsymbol{\eta}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{\eta}} - (\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\delta})^2)$, which is independent of U_0, we can write $\sqrt{\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{\eta}}}\,\tilde{\boldsymbol{\eta}}^\top\mathbf{U} \stackrel{d}{=} \sqrt{1-\delta_0^2}\,U_1$, where $\delta_0 = \tilde{\boldsymbol{\eta}}^\top\boldsymbol{\delta}/\sqrt{\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{\eta}}}$ and $U_1 \sim N_1(0,1)$ and is independent of U_0. Hence, we obtain

$$\tilde{\boldsymbol{\eta}}^\top(\mathbf{Z}-\tilde{\boldsymbol{\xi}}) \stackrel{d}{=} \tilde{\boldsymbol{\eta}}^\top(\boldsymbol{\xi}-\tilde{\boldsymbol{\xi}}) + \sqrt{\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{\eta}}}\,Z_0,$$

where $Z_0 = \delta_0|U_0| + \sqrt{1-\delta_0^2}\,U_1$. Since $Z_0 \sim SN_1(0,1,\eta_0)$, where

$$\eta_0 = \frac{\delta_0}{\sqrt{1-\delta_0^2}} = \frac{\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\delta}}{\sqrt{\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{\eta}} - (\tilde{\boldsymbol{\eta}}^\top\boldsymbol{\delta})^2}},$$

the proof follows.

Proof of Lemma 2: By (5) we have for the logarithm of the pdf of $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\eta}_2)$ that

$$\begin{aligned}\log f_{\mathbf{Y}}(\mathbf{x}) &= \log\phi_k(\mathbf{x};\boldsymbol{\xi}_2,\boldsymbol{\Omega}_2) + \log[2\Phi\{\boldsymbol{\eta}_2^\top(\mathbf{x}-\boldsymbol{\xi}_2)\}] \\ &= -\frac{1}{2}\left\{k\log(2\pi) + \log|\boldsymbol{\Omega}_2| + (\mathbf{x}-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}(\mathbf{x}-\boldsymbol{\xi}_2)\right\} + \log[2\Phi\{\boldsymbol{\eta}_2^\top(\mathbf{x}-\boldsymbol{\xi}_2)\}].\end{aligned}$$

Thus, since by (2) $CH(\mathbf{X},\mathbf{Y}) = -E[\log f_{\mathbf{Y}}(\mathbf{X})]$, we have by applying the Lemma 1(iii) with $\mathbf{Z}$ replaced by $\mathbf{X}$, $\mathbf{a} = \boldsymbol{\xi}_2$ and $\mathbf{B} = \boldsymbol{\Omega}_2^{-1}$ that

$$\begin{aligned} CH(\mathbf{X},\mathbf{Y}) &= \frac{1}{2}\left\{k\log(2\pi) + \log|\boldsymbol{\Omega}_2| + E\{(\mathbf{X}-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}(\mathbf{X}-\boldsymbol{\xi}_2)\}\right\} - E[\log\{2\Phi(\boldsymbol{\eta}_2^\top(\mathbf{X}-\boldsymbol{\xi}_2))\}] \\ &= \frac{1}{2}\left\{k\log(2\pi) + \log|\boldsymbol{\Omega}_2| + \mathrm{tr}({\boldsymbol{\Omega}_2}^{-1}\boldsymbol{\Omega}_1) + (\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}(\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)\right. \\ &\quad \left. +2\sqrt{\frac{2}{\pi}}(\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}\boldsymbol{\delta}_1\right\} - E[\log\{2\Phi(\boldsymbol{\eta}_2^\top(\mathbf{X}-\boldsymbol{\xi}_2))\}] \\ &= \frac{1}{2}\left\{k\log(2\pi) + \log|\boldsymbol{\Omega}_2| + \mathrm{tr}({\boldsymbol{\Omega}_2}^{-1}\boldsymbol{\Omega}_1) + (\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}(\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)\right\} \\ &\quad +\sqrt{\frac{2}{\pi}}(\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}\boldsymbol{\delta}_1 - E[\log\{2\Phi(\boldsymbol{\eta}_2^\top(\mathbf{X}-\boldsymbol{\xi}_2))\}]. \end{aligned}$$

From Lemma 1(iii) we find that the random variable $\boldsymbol{\eta}_2^\top(\mathbf{X}-\boldsymbol{\xi}_2)$ has the same distribution of W_{21} in (6). Thus, the proof follows by noting that

$$\begin{aligned} CH(\mathbf{X}_0,\mathbf{Y}_0) &= -E[\log\phi_k(\mathbf{X}_0;\boldsymbol{\xi}_2,\boldsymbol{\Omega}_2)] \\ &= \frac{1}{2}\left\{k\log(2\pi) + \log|\boldsymbol{\Omega}_2| + \mathrm{tr}({\boldsymbol{\Omega}_2}^{-1}\boldsymbol{\Omega}_1) + (\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)^\top\boldsymbol{\Omega}_2^{-1}(\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)\right\}. \end{aligned}$$

Proof of Proposition 1: Note first by Lemma 2 that $H(\mathbf{X}) = CH(\mathbf{X},\mathbf{X})$ is given by

$$\begin{aligned} H(\mathbf{X}) &= \frac{1}{2}\{k + k\log(2\pi) + \log|\boldsymbol{\Omega}_1|\} - E[\log\{2\Phi(\boldsymbol{\eta}_1^\top(\mathbf{X}-\boldsymbol{\xi}_1))\}] \\ &= H(\mathbf{X}_0) - E[\log\{2\Phi(\boldsymbol{\eta}_1^\top(\mathbf{X}-\boldsymbol{\xi}_1))\}], \end{aligned}$$

where $H(\mathbf{X}_0) = CH(\mathbf{X}_0,\mathbf{X}_0) = \frac{1}{2}\{k + k\log(2\pi) + \log|\boldsymbol{\Omega}_1|\}$ and by the property (iii) of the Lemma 1 we have $\boldsymbol{\eta}_1^\top(\mathbf{X}-\boldsymbol{\xi}_1) \overset{d}{=} W_{11}$. Thus, the proof follows from the fact that $D_{\mathrm{KL}}(\mathbf{X},\mathbf{Y}) = CH(\mathbf{X},\mathbf{Y}) - H(\mathbf{X})$.

Appendix B. Complementary Figures

Figure A1. Plots of Skew-normal fits (in red) and QQ-plots of Normal and Skew-Normal distributions for clusters black (1), red (2), green (3) and blue (4).

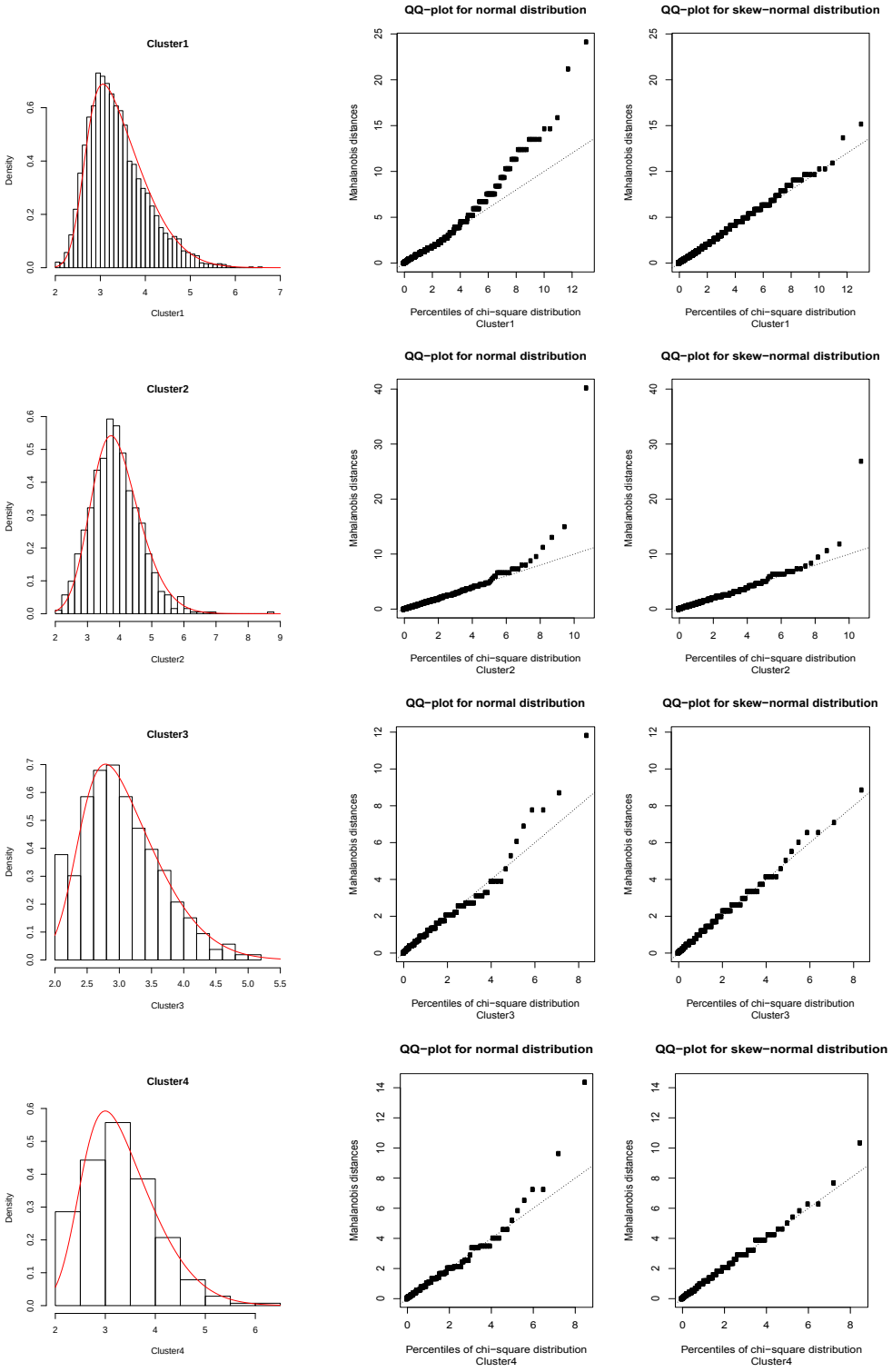


Figure A2. Plots of Skew-normal fits (in red) and QQ-plots of Normal and Skew-Normal distributions for clusters violet (5), yellow (6), gray (7) and all observations.