Scandinavian Journal of Statistics Theory and Applications

doi: 10.1111/j.1467-9469.2011.00774.x

Shannon Entropy and Mutual Information for Multivariate Skew-Elliptical Distributions‡

REINALDO B. ARELLANO-VALLE

Departamento de Estadística, Pontificia Universidad Católica de Chile

JAVIER E. CONTRERAS-REYES

Departamento de Ingeniería Matemática, Universidad de Chile

MARC G. GENTON

Department of Statistics, Texas A&M University

ABSTRACT. The entropy and mutual information index are important concepts developed by Shannon in the context of information theory. They have been widely studied in the case of the multivariate normal distribution. We first extend these tools to the full symmetric class of multivariate elliptical distributions and then to the more flexible families of multivariate skew-elliptical distributions. We study in detail the cases of the multivariate skew-normal and skew-*t* distributions. We implement our findings to the application of the optimal design of an ozone monitoring station network in Santiago de Chile.

Key words: elliptical distribution, entropy, information theory, optimal network design, Shannon, skew-normal, skew-*t*

1. Introduction

The mathematical theory of communication introduced by Shannon (1948) describes logarithmic measures of information and has stimulated a tremendous amount of study in engineering fields on the subject of information theory. It is a branch of applied probability and statistics that is relevant to statistical inference and therefore should be of basic interest to statisticians (Kullback, 1978). Information theory seeks the quantification of information. One goal of information theory is the development of coding schemes that provide good performance in comparison with the optimal performance given by the theory. It works under the assumption of a strongly stationary random process to define an information quantity contained in a multivariate probability density function, for example, the multivariate normal distribution (Kullback, 1978; Silva & Quiroz, 2003; Misra *et al.*, 2005; Cover & Thomas, 2006). This quantity allows to measure the cumulative information of a multivariate data set, or more specifically, to quantify the mutual information between two random variables or vectors. On the other hand, the entropy is a notion of information provided by a random process about itself and it is sufficient to study the reproduction of a marginal process through a noiseless environment. For a systematic and comprehensive account of these and related concepts, see, for example, Cover & Thomas (2006).

Defined according to Cover & Thomas (2006) from discrete to continuous variables, we consider the following concepts of entropy and mutual information index. Let $\mathbf{X} \in \mathbb{R}^n$

‡This article was published online on [27 February 2012]. Errors were subsequently identified. This notice is included in the online and print versions to indicate that both have been corrected [04 April 2012].

and $\mathbf{Y} \in \mathbb{R}^m$ be two random vectors with joint and marginal probability density functions $p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})$, $p_{\mathbf{X}}(\mathbf{x})$ and $p_{\mathbf{Y}}(\mathbf{y})$, respectively. The *mutual information index* between $\mathbf{X}$ and $\mathbf{Y}$ is defined by

$$I_{\mathbf{XY}} = E\left[\log\left\{\frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y})}{p_{\mathbf{X}}(\mathbf{X})p_{\mathbf{Y}}(\mathbf{Y})}\right\}\right] = \int_{\mathbb{R}^m}\int_{\mathbb{R}^n} \log\left\{\frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})p_{\mathbf{Y}}(\mathbf{y})}\right\} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}. \tag{1}$$

Moreover the (differential) *entropy* of a random vector $\mathbf{Z} \in \mathbb{R}^k$ with probability density function $p_{\mathbf{Z}}(\mathbf{z})$ is defined by

$$H_{\mathbf{Z}} = -E\left[\log\{p_{\mathbf{Z}}(\mathbf{Z})\}\right] = -\int_{\mathbb{R}^k} \log\{p_{\mathbf{Z}}(\mathbf{z})\} p_{\mathbf{Z}}(\mathbf{z})\, \mathrm{d}\mathbf{z}. \tag{2}$$

The entropy concept is attributed to uncertainty of information or mathematical contrariety of information. From (1) and (2), it is straightforward to see that the mutual information index $I_{\mathbf{XY}}$ between $\mathbf{X}$ and $\mathbf{Y}$ can be computed as

$$I_{\mathbf{XY}} = H_{\mathbf{X}} + H_{\mathbf{Y}} - H_{\mathbf{XY}}, \tag{3}$$

where $H_{\mathbf{XY}}$, $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ are joint and marginal entropies of $(\mathbf{X}, \mathbf{Y})$, $\mathbf{X}$ and $\mathbf{Y}$, respectively. By definition, $I_{\mathbf{XY}} = 0$ when the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent; otherwise, this index is positive (Cover & Thomas, 2006), and it increases with the degree of dependence between the components of $\mathbf{X}$ and $\mathbf{Y}$. In other words, the mutual information index provides a generalized measure of association between $\mathbf{X}$ and $\mathbf{Y}$, which is particularly convenient in those models where the correlation is not defined.

Several research studies have used information theory to design a network in situations where specific objectives are hard to define or may be unforeseen, especially in monitoring networks design (Silva & Quiroz, 2003; Ainslie *et al.*, 2009), transmission of bits information (Shannon, 1948) and other applications. However, studies such as Silva & Quiroz (2003) have assumed a multivariate normality condition on the variables under transformations which restricts the real information given by the data set.

Although recently, Javier & Gupta (2008, 2009) have studied the mutual information for non-normal distributions belonging to the family of continuous multivariate location-scale distributions, their results are concerned only with three specific distributions and are derived in terms of infinite series. Some other entropy expressions for multivariate distributions can be found in Ahmed & Gokhale (1989). We propose in this paper a general and unified theory of the mutual information for flexible and tractable families of continuous multivariate distributions, in which the multivariate normal and further well-known symmetric distributions, such as the Student's t, are particular members. Specifically, we consider the multivariate elliptical and skew-elliptical families of distributions; see the books by Fang *et al.* (1990) and Genton (2004), respectively. We give special attention to the particular cases of the multivariate skew-normal and skew-t distributions that allow to model skewness.

The organization of this paper is as follows. Section 2 presents the entropy and mutual information index for multivariate elliptical distributions, with the multivariate normal and Student's t distributions as special cases. Section 3 presents the entropy and mutual information index for multivariate skew-elliptical distributions, with the multivariate skew-normal and skew-t distributions as special cases. Section 4 reports numerical results on the evaluation of the skew-normal and skew-t entropies as a function of the skewness. Section 5 presents the application of our results to the optimal design of an ozone monitoring station network in Santiago de Chile. This paper ends with a discussion in section 6.

2. Entropy and mutual information for multivariate elliptical distributions

2.1. Location-scale models

In this paper, our interest lies in the computation of the mutual information index for location-scale models. In this sense, lemma 1 shows that for any distribution in this class, the entropy (and hence the mutual information index) does not depend on where it is localized. In other words, for these distributions, the location parameter is irrelevant to compute the entropy and the mutual information index.

Lemma 1. *Let* $p_{\mathbf{Z}}(\mathbf{z}) = |\mathbf{\Omega}|^{-1/2} p_{\mathbf{Z}_0}\{\mathbf{\Omega}^{-1/2}(\mathbf{z}-\boldsymbol{\xi})\}$ *be a location-scale probability density function, where* $\boldsymbol{\xi} \in \mathbb{R}^k$ *is the location vector and* $\mathbf{\Omega} \in \mathbb{R}^{k\times k}$ *is the dispersion matrix. Let* $\mathbf{Z}_0 = \mathbf{\Omega}^{-1/2}(\mathbf{Z}-\boldsymbol{\xi})$ *be a standardized version of* $\mathbf{Z}$*, with standardized probability density function* $p_{\mathbf{Z}_0}(\mathbf{z}_0)$ *that does not depend on* $(\boldsymbol{\xi}, \mathbf{\Omega})$*. Then,*

$$H_{\mathbf{Z}} = \frac{1}{2}\log|\mathbf{\Omega}| + H_{\mathbf{Z}_0}, \tag{4}$$

where $H_{\mathbf{Z}_0} = -E[\log\{p_{\mathbf{Z}_0}(\mathbf{Z}_0)\}]$ *is the entropy of the standardized random vector* $\mathbf{Z}_0$*.*

Proof. The result is immediate from $E[\log\{p_{\mathbf{Z}}(\mathbf{Z})\}] = -(1/2)\log|\mathbf{\Omega}| + E[\log\{p_{\mathbf{Z}_0}(\mathbf{Z}_0)\}]$.

2.2. Multivariate elliptical distributions

The multivariate elliptical family of distributions defines one of the most important classes of symmetric location-scale models. It contains the normal model and preserves most of its main properties. For a systematic review of this family, see, for example, Fang *et al.* (1990). In this section, we give the ingredients to compute the elliptical mutual information index.

Let $\mathbf{Z} \sim \mathrm{EC}_k(\boldsymbol{\xi}, \mathbf{\Omega}, h^{(k)})$ be an elliptical random vector in $\mathbb{R}^k$, with location vector $\boldsymbol{\xi} \in \mathbb{R}^k$, dispersion matrix $\mathbf{\Omega} \in \mathbb{R}^{k\times k}$ and density generator function $h^{(k)}$, whose probability density function is

$$p_{\mathbf{Z}}(\mathbf{z}) \equiv f_k(\mathbf{z}; \boldsymbol{\xi}, \mathbf{\Omega}, h^{(k)}) = |\mathbf{\Omega}|^{-1/2} h^{(k)}\{(\mathbf{z}-\boldsymbol{\xi})^T \mathbf{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\xi})\}, \quad \mathbf{z} \in \mathbb{R}^k.$$

Here, the density generator function $h^{(k)}$ is a non-negative real-valued function such that

$$g(s) = \frac{\pi^{k/2}}{\Gamma(k/2)} s^{k/2-1} h^{(k)}(s), \quad s > 0,$$

is a valid probability density function. Note that $p_{\mathbf{Z}}(\mathbf{z}) = |\mathbf{\Omega}|^{-1/2} h^{(k)}(\mathbf{z}_0^T \mathbf{z}_0)$, where $\mathbf{z}_0 = \mathbf{\Omega}^{-1/2}(\mathbf{z}-\boldsymbol{\xi})$. Hence, for this class, the *standardized* random vector $\mathbf{Z}_0 = \mathbf{\Omega}^{-1/2}(\mathbf{Z}-\boldsymbol{\xi})$ has a spherical probability density function $p_{\mathbf{Z}_0}(\mathbf{z}_0) = h^{(k)}(\mathbf{z}_0^T \mathbf{z}_0)$, $\mathbf{z}_0 \in \mathbb{R}^k$, for which

$$H_{\mathbf{Z}_0}^{\mathrm{EC}_k} = -E[\log\{h^{(k)}(\mathbf{Z}_0^T \mathbf{Z}_0)\}].$$

This expectation depends on the distribution of the squared radial random variable $S = \mathbf{Z}_0^T \mathbf{Z}_0 = (\mathbf{Z}-\boldsymbol{\xi})^T \mathbf{\Omega}^{-1}(\mathbf{Z}-\boldsymbol{\xi})$, which has the probability density function $g(s)$ given above. As in Arellano-Valle *et al.* (2006b), we call the distribution of S a squared-radial distribution and we denote it by $\mathcal{R}^2(h^{(k)})$. Hence, for the entropy of $\mathbf{Z}_0$ we have

$$H_{\mathbf{Z}_0}^{\mathrm{EC}_k} = -E[\log\{h^{(k)}(S)\}] = -\int_0^\infty [\log\{h^{(k)}(s)\}] g(s)\, \mathrm{d}s.$$

Thus, the entropy of $\mathbf{Z} \sim \mathrm{EC}_k(\boldsymbol{\xi}, \mathbf{\Omega}, h^{(k)})$ can be obtained from (4). Moreover, the mutual information between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with elliptical joint distribution can be

computed using (3) and considering that if

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim \mathrm{EC}_{n+m}\left(\begin{pmatrix}\boldsymbol{\xi}_{\mathbf{X}}\\ \boldsymbol{\xi}_{\mathbf{Y}}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Omega}_{\mathbf{XX}} & \boldsymbol{\Omega}_{\mathbf{XY}}\\ \boldsymbol{\Omega}_{\mathbf{YX}} & \boldsymbol{\Omega}_{\mathbf{YY}}\end{pmatrix}, h^{(n+m)}\right),$$

then the respective marginal random vectors are distributed as $\mathbf{X} \sim \mathrm{EC}_n(\boldsymbol{\xi}_{\mathbf{X}}, \boldsymbol{\Omega}_{\mathbf{XX}}, h^{(n)})$ and $\mathbf{Y} \sim \mathrm{EC}_m(\boldsymbol{\xi}_{\mathbf{Y}}, \boldsymbol{\Omega}_{\mathbf{YY}}, h^{(m)})$. In fact, it is clear from the above results that the respective marginal and joint entropies are

$$H_{\mathbf{X}}^{\mathrm{EC}_n} = \frac{1}{2}\log|\boldsymbol{\Omega}_{\mathbf{XX}}| - E[\log\{h^{(n)}(S_{\mathbf{X}})\}],$$
$$H_{\mathbf{Y}}^{\mathrm{EC}_m} = \frac{1}{2}\log|\boldsymbol{\Omega}_{\mathbf{YY}}| - E[\log\{h^{(m)}(S_{\mathbf{Y}})\}],$$
$$H_{\mathbf{XY}}^{\mathrm{EC}_{n+m}} = \frac{1}{2}\log|\boldsymbol{\Omega}| - E[\log\{h^{(n+m)}(S_{\mathbf{XY}})\}],$$

where $S_{\mathbf{X}} \sim \mathcal{R}^2(h^{(n)})$, $S_{\mathbf{Y}} \sim \mathcal{R}^2(h^{(m)})$ and $S_{\mathbf{XY}} \sim \mathcal{R}^2(h^{(n+m)})$. The determinant of the joint dispersion matrix $\boldsymbol{\Omega}$ can be computed as

$$|\boldsymbol{\Omega}| = \begin{vmatrix}\boldsymbol{\Omega}_{\mathbf{XX}} & \boldsymbol{\Omega}_{\mathbf{XY}}\\ \boldsymbol{\Omega}_{\mathbf{YX}} & \boldsymbol{\Omega}_{\mathbf{YY}}\end{vmatrix} = |\boldsymbol{\Omega}_{\mathbf{YY}}||\boldsymbol{\Omega}_{\mathbf{XX}}||\mathbf{I}_n - \mathbf{B}_{\mathbf{X}\cdot\mathbf{Y}}\mathbf{B}_{\mathbf{Y}\cdot\mathbf{X}}|,$$

where $\mathbf{B}_{\mathbf{X}\cdot\mathbf{Y}} = \boldsymbol{\Omega}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}_{\mathbf{XY}}$ and $\mathbf{B}_{\mathbf{Y}\cdot\mathbf{X}} = \boldsymbol{\Omega}_{\mathbf{YY}}^{-1}\boldsymbol{\Omega}_{\mathbf{YX}}$ are the matrices of regression coefficients associated with the regression functions of $\mathbf{X} \mid \mathbf{Y} = \mathbf{y}$ and $\mathbf{Y} \mid \mathbf{X} = \mathbf{x}$, respectively. Note here that $0 \le |\mathbf{I}_n - \mathbf{B}_{\mathbf{X}\cdot\mathbf{Y}}\mathbf{B}_{\mathbf{Y}\cdot\mathbf{X}}| \le 1$. Hence, we obtain that the elliptical mutual information index between $\mathbf{X}$ and $\mathbf{Y}$ is

$$I_{\mathbf{XY}}^{\mathrm{EC_{n+m}}}(\boldsymbol{\Omega}, h) = E[\log\{h^{(n+m)}(S_{\mathbf{XY}})\}] - E[\log\{h^{(n)}(S_{\mathbf{X}})\}] - E[\log\{h^{(m)}(S_{\mathbf{Y}})\}] - \frac{1}{2}\log|\mathbf{I}_n - \mathbf{B}_{\mathbf{X}\cdot\mathbf{Y}}\mathbf{B}_{\mathbf{Y}\cdot\mathbf{X}}|. \quad (5)$$

The last term in (5) represents the information due the dispersion matrix $\boldsymbol{\Omega}$, which is the same for the whole elliptical class. A similar fact occurs with the correlation matrix induced by $\boldsymbol{\Omega}$, which means that within the elliptical family, the correlation does not depend on the specific elliptical density generator h. As a consequence from (5), the elliptical mutual information depends on both $\boldsymbol{\Omega}$ and h, allowing differences for the association between $\mathbf{X}$ and $\mathbf{Y}$ through the different elliptical joint distributions.

The multivariate normal distribution, namely $\mathbf{Z} \sim N_k(\boldsymbol{\xi}, \boldsymbol{\Omega})$, is a particular member of the elliptical family. In this case, $E(\mathbf{Z}) = \boldsymbol{\xi}$ and $\mathrm{var}(\mathbf{Z}) = \boldsymbol{\Omega}$. Moreover, for the normal density generator function, we have $h_N^{(k)}(s) = (2\pi)^{-k/2}\,\mathrm{e}^{-s/2}$, $s > 0$, and for the distribution of the normal squared radial random variable, we have $S = (\mathbf{Z} - \boldsymbol{\xi})^T\boldsymbol{\Omega}^{-1}(\mathbf{Z} - \boldsymbol{\xi}) \sim \chi_k^2$, the chi-squared distribution with k degrees of freedom. Another important member is the multivariate Student's t distribution $\mathbf{Z} \sim T_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, v)$, where $v > 0$ is the degrees of freedom, for which $E(\mathbf{Z}) = \boldsymbol{\xi}$ for $v > 1$ and $\mathrm{var}(\mathbf{Z}) = v/(v-2)\boldsymbol{\Omega}$ for $v > 2$. Also, for the Student's t distribution, we have $h_T^{(k)}(s) = \Gamma\{(v+k)/2\}/\{\Gamma(v/2)(v\pi)^{k/2}\}\left(1 + \frac{s}{v}\right)^{-(v+k)/2}$ and $S/k \sim F_{k,v}$, the Fisher distribution with k and v degrees of freedom. Further properties of these distributions can be found in the book of Fang *et al.* (1990) and in Arellano-Valle *et al.* (2006b). For the particular case of the Student's t distributions, see Arellano-Valle & Bolfarine (1995). The normal and Student's t distributions are, however, particular cases of the so-called scale mixtures of normal distributions, a subclass of elliptical distributions, for which the density generator function can be represented as

$$h^{(k)}(u) = \int_0^\infty v^{k/2}h_N^{(k)}(\sqrt{v}u)\,\mathrm{d}F(v),$$

where $h_N^{(k)}$ is the aforementioned normal density generator function and F is a cumulative distribution function on $(0,\infty)$ that does not depend on k. This is equivalent to representing stochastically the spherical random vector $\mathbf{Z}_0=\boldsymbol{\Omega}^{-1/2}(\mathbf{Z}-\boldsymbol{\xi})$ as $\mathbf{Z}_0\overset{d}{=}V^{-1/2}\mathbf{Z}_{0N}$, where $V\sim F$, $\mathbf{Z}_{0N}\sim N_k(\mathbf{0},\mathbf{I}_k)$ and they are independent. As a consequence of this fact, we have $S\overset{d}{=}V^{-1}S_N$, where $S_N\sim\chi^2_k$ and is independent of V. We study the normal and Student's t special cases in the next sections.

2.3. The multivariate normal distribution

We give an alternative proof of the multivariate normal Shannon entropy (Kullback, 1978; Misra *et al.*, 2005; Cover & Thomas, 2006) by considering lemma 1. Let $\mathbf{Z}\sim N_k(\boldsymbol{\xi},\boldsymbol{\Omega})$ denote a k-dimensional normal random vector, with mean vector $E(\mathbf{Z})=\boldsymbol{\xi}\in\mathbb{R}^k$ and covariance matrix $\mathrm{var}(\mathbf{Z})=\boldsymbol{\Omega}\in\mathbb{R}^{k\times k}$. We have $\mathbf{Z}=\boldsymbol{\xi}+\boldsymbol{\Omega}^{1/2}\mathbf{Z}_0$, where $\mathbf{Z}_0\sim N_k(\mathbf{0},\mathbf{I}_k)$. The probability density function of $\mathbf{Z}_0$ is $p_{\mathbf{Z}_0}(\mathbf{z}_0)=\phi_k(\mathbf{z}_0)=(2\pi)^{-k/2}\exp\{-(1/2)\mathbf{z}_0^T\mathbf{z}_0\}$. Thus, since in this case $S=\mathbf{Z}_0^T\mathbf{Z}_0\sim\chi^2_k$, and so $E(S)=k$, we have

$$H_{\mathbf{Z}_0}^{N_k}=\frac{k}{2}\log(2\pi)+\frac{1}{2}E(S)=\frac{k}{2}\{1+\log(2\pi)\}.$$

Therefore, by lemma 1:

$$H_{\mathbf{Z}}^{N_k}=\frac{1}{2}\log|\boldsymbol{\Omega}|+\frac{k}{2}\{1+\log(2\pi)\}. \tag{6}$$

Now let

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\sim N_{n+m}\left(\begin{pmatrix}\boldsymbol{\xi}_{\mathbf{X}}\\ \boldsymbol{\xi}_{\mathbf{Y}}\end{pmatrix},\begin{pmatrix}\boldsymbol{\Omega}_{\mathbf{XX}} & \boldsymbol{\Omega}_{\mathbf{XY}}\\ \boldsymbol{\Omega}_{\mathbf{YX}} & \boldsymbol{\Omega}_{\mathbf{YY}}\end{pmatrix}\right).$$

It is well-known that the marginal distributions are $\mathbf{X}\sim N_n(\boldsymbol{\xi}_{\mathbf{X}},\boldsymbol{\Omega}_{\mathbf{XX}})$ and $\mathbf{Y}\sim N_m(\boldsymbol{\xi}_{\mathbf{Y}},\boldsymbol{\Omega}_{\mathbf{YY}})$. Hence, $S_{\mathbf{X}}\sim\chi^2_n$, $S_{\mathbf{Y}}\sim\chi^2_m$ and $S_{\mathbf{XY}}\sim\chi^2_{n+m}$, and therefore

$$\begin{aligned}H_{\mathbf{X}}^{N_n}&=\frac{1}{2}\log|\boldsymbol{\Omega}_{\mathbf{XX}}|+\frac{n}{2}\{1+\log(2\pi)\},\\ H_{\mathbf{Y}}^{N_m}&=\frac{1}{2}\log|\boldsymbol{\Omega}_{\mathbf{YY}}|+\frac{m}{2}\{1+\log(2\pi)\},\\ H_{\mathbf{XY}}^{N_{n+m}}&=\frac{1}{2}\log|\boldsymbol{\Omega}|+\frac{n+m}{2}\{1+\log(2\pi)\}.\end{aligned}$$

Thus, we obtain from (3), or directly from (5), that the normal mutual information index between $\mathbf{X}$ and $\mathbf{Y}$ is

$$I_{\mathbf{XY}}^{N_{n+m}}(\boldsymbol{\Omega})=\frac{1}{2}\log\left(\frac{|\boldsymbol{\Omega}_{\mathbf{XX}}||\boldsymbol{\Omega}_{\mathbf{YY}}|}{|\boldsymbol{\Omega}|}\right)=-\frac{1}{2}\log|\mathbf{I}_n-\mathbf{B}_{\mathbf{X}\cdot\mathbf{Y}}\mathbf{B}_{\mathbf{Y}\cdot\mathbf{X}}|.$$

Hence, the normal mutual information and Shannon entropy depend only on the covariance matrix $\boldsymbol{\Omega}$. That is, similar to the correlation coefficients, Shannon's mutual information index measures multivariate linear dependence between $\mathbf{X}$ and $\mathbf{Y}$.

2.4. The multivariate Student's t distribution

Let $\mathbf{Z}\sim T_k(\boldsymbol{\xi},\boldsymbol{\Omega},\nu)$ denote a k-dimensional Student's t random vector with location vector $\boldsymbol{\xi}\in\mathbb{R}^k$, dispersion matrix $\boldsymbol{\Omega}\in\mathbb{R}^{k\times k}$ and ν degrees of freedom, that is, with probability density function

$$p_{\mathbf{Z}}(\mathbf{z}) \equiv t_k(\mathbf{z}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)|\boldsymbol{\Omega}|^{1/2}(\nu\pi)^{k/2}}\left(1+\frac{1}{\nu}\mathbf{z}_0^T\mathbf{z}_0\right)^{-(\nu+k)/2}, \quad \mathbf{z}\in\mathbb{R}^k,$$

where as before $\mathbf{z}_0 = \boldsymbol{\Omega}^{-1/2}(\mathbf{z}-\boldsymbol{\xi})$. For this case, we have $\mathbf{Z}_0 = \boldsymbol{\Omega}^{-1/2}(\mathbf{Z}-\boldsymbol{\xi}) \sim T_k(\mathbf{0}, \mathbf{I}_k, \nu)$ and $\mathbf{Z}_0^T\mathbf{Z}_0 \sim kF_{k,\nu}$. Thus, considering that

$$\log\{h^{(k)}(s)\} = \log\left\{\Gamma\left(\frac{\nu+k}{2}\right)\right\} - \log\left\{\Gamma\left(\frac{\nu}{2}\right)\right\} - \frac{k}{2}\log(\nu\pi) - \frac{\nu+k}{2}\log\left(1+\frac{s}{\nu}\right),$$

we have $H_{\mathbf{Z}_0}^{T_k} = E[\log\{h^{(k)}(S)\}]$ where $S \sim kF_{k,\nu}$. Using now the well-known fact that $S \stackrel{d}{=} k(S_1/k)/(S_2/\nu)$, where $S_1 \sim \chi_k^2$, $S_2 \sim \chi_\nu^2$, and they are independent, and consequently $S_1+S_2 \sim \chi_{k+\nu}^2$, it is straightforward to see that

$$E\left\{\log\left(1+\frac{S}{\nu}\right)\right\} = E\{\log(S_1+S_2)\} - E\{\log(S_2)\} = \psi\left(\frac{\nu+k}{2}\right) - \psi\left(\frac{\nu}{2}\right),$$

where $\psi(x) = \mathrm{d}/\mathrm{d}x\log\{\Gamma(x)\}$ is the digamma function. We find for the entropy of $\mathbf{Z}_0 \sim T_k(\mathbf{0}, \mathbf{I}_k, \nu)$ that

$$H_{\mathbf{Z}_0}^{T_k} = -\log\left\{\frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{k/2}}\right\} + \frac{\nu+k}{2}\left\{\psi\left(\frac{\nu+k}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right\}. \tag{7}$$

Now let

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim T_{n+m}\left(\begin{pmatrix}\boldsymbol{\xi}_{\mathbf{X}}\\ \boldsymbol{\xi}_{\mathbf{Y}}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Omega}_{\mathbf{XX}} & \boldsymbol{\Omega}_{\mathbf{XY}}\\ \boldsymbol{\Omega}_{\mathbf{YX}} & \boldsymbol{\Omega}_{\mathbf{YY}}\end{pmatrix}, \nu\right).$$

From Arellano-Valle & Bolfarine (1995), the marginal distributions are $\mathbf{X} \sim T_n(\boldsymbol{\xi}_{\mathbf{X}}, \boldsymbol{\Omega}_{\mathbf{XX}}, \nu)$ and $\mathbf{Y} \sim T_m(\boldsymbol{\xi}_{\mathbf{Y}}, \boldsymbol{\Omega}_{\mathbf{YY}}, \nu)$. Hence, from (3) and (7), we deduce that the mutual information index for the Student's t case is

$$\begin{aligned} I_{\mathbf{XY}}^{T_{n+m}}(\boldsymbol{\Omega}, \nu) = I_{\mathbf{XY}}^{N_{n+m}}(\boldsymbol{\Omega}) + \log\left[\frac{\Gamma(\nu/2)\,\Gamma\{(\nu+n+m)/2\}}{\Gamma\{(\nu+n)/2\}\,\Gamma\{(\nu+m)/2\}}\right] - \frac{\nu+m}{2}\psi\left(\frac{\nu+m}{2}\right)\\ - \frac{\nu+n}{2}\psi\left(\frac{\nu+n}{2}\right) + \frac{\nu+n+m}{2}\psi\left(\frac{\nu+n+m}{2}\right) + \frac{\nu}{2}\psi\left(\frac{\nu}{2}\right). \end{aligned}$$

It is interesting to notice that the information due to $\boldsymbol{\Omega}$ arises only from $I_{\mathbf{XY}}^{N_k}(\boldsymbol{\Omega})$, and the information due to ν comes from the remaining terms. It is also clear that as ν increases, the Student's t mutual information converges to the normal mutual information.

3. Entropy and mutual information for multivariate skew-elliptical distributions

3.1. Multivariate skew-elliptical distributions

A flexible class of location-scale models is defined by the so-called skew-elliptical family of distributions; see Branco & Dey (2001), Azzalini & Capitanio (1999, 2003), Arellano-Valle & Azzalini (2006), Arellano-Valle & Genton (2005, 2010a, 2010b), and the book edited by Genton (2004). It allows for modelling skewness in the distribution of the data. In this section, we extend the previous results to this more general class.

We say that a random vector $\mathbf{Z} \in \mathbb{R}^k$ has a skew-elliptical distribution, with location vector $\boldsymbol{\xi} \in \mathbb{R}^k$, dispersion matrix $\boldsymbol{\Omega} \in \mathbb{R}^{k\times k}$, shape/skewness parameter $\boldsymbol{\eta} \in \mathbb{R}^k$ and density generator function $h^{(k+1)}$, denoted by $\mathbf{Z} \sim \mathrm{SE}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}, h^{(k+1)})$, if its probability density function is

$$p_{\mathbf{Z}}(\mathbf{z}) = 2f_k(\mathbf{z}; \boldsymbol{\xi}, \boldsymbol{\Omega}, h^{(k)})F(\boldsymbol{\eta}^T(\mathbf{z}-\boldsymbol{\xi}); h_s^{(1)}), \quad \mathbf{z}\in\mathbb{R}^k, \tag{8}$$

where $f_k(\mathbf{z};\boldsymbol{\xi},\boldsymbol{\Omega},h^{(k)})=|\boldsymbol{\Omega}|^{-1/2}h^{(k)}(s)$ with $s=\mathbf{z}_0^T\mathbf{z}_0$ and $\mathbf{z}_0=\boldsymbol{\Omega}^{-1/2}(\mathbf{z}-\boldsymbol{\xi})$, that is, the probability density function of an $\mathrm{EC}_k(\boldsymbol{\xi},\boldsymbol{\Omega},h^{(k)})$ distribution, and $F(x;h_s^{(1)})=\int_{-\infty}^{x}h_s^{(1)}(w)\,\mathrm{d}w$ is the univariate cumulative distribution function induced by the conditional density generator function $h_s^{(1)}(u)=h^{(k+1)}(s+u)/h^{(k)}(s)$.

Let $\bar{\boldsymbol{\eta}}=\boldsymbol{\Omega}^{1/2}\boldsymbol{\eta}$. In terms of $\mathbf{z}_0=\boldsymbol{\Omega}^{-1/2}(\mathbf{z}-\boldsymbol{\xi})$, the skew-elliptical probability density function (8) can be rewritten as $p_{\mathbf{Z}}(\mathbf{z})=|\boldsymbol{\Omega}|^{-1/2}p_{\mathbf{Z}_0}(\mathbf{z}_0)$, where

$$p_{\mathbf{Z}_0}(\mathbf{z}_0)=2h^{(k)}(\mathbf{z}_0^T\mathbf{z}_0)F(\bar{\boldsymbol{\eta}}^T\mathbf{z}_0;h_s^{(1)})$$

is the probability density function of $\mathbf{Z}_0=\boldsymbol{\Omega}^{-1/2}(\mathbf{Z}-\boldsymbol{\xi})\sim\mathrm{SE}_k(\mathbf{0},\mathbf{I}_k,\bar{\boldsymbol{\eta}},h^{(k+1)})$; see, for example, Arellano-Valle & Genton (2010a). Then, lemma 1 yields the following result.

Proposition 1. *The entropy of a skew-elliptical random vector* $\mathbf{Z}\sim\mathrm{SE}_k(\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\eta},h^{(k+1)})$ *is*

$$H_{\mathbf{Z}}^{\mathrm{SE}_k}=H_{\mathbf{Z}_{\mathrm{EC}}}^{\mathrm{EC}_k}-E[\log\{2F(\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0;h_S^{(1)})\}],$$

where $H_{\mathbf{Z}_{\mathrm{EC}}}^{\mathrm{EC}_k}$ *is the entropy of* $\mathbf{Z}_{\mathrm{EC}}\sim\mathrm{EC}_k(\boldsymbol{\xi},\boldsymbol{\Omega},h^{(k)})$, $\mathbf{Z}_0\sim\mathrm{SE}_k(\mathbf{0},\mathbf{I}_k,\bar{\boldsymbol{\eta}},h^{(k+1)})$ *and* $S=\mathbf{Z}_0^T\mathbf{Z}_0$.

It follows from proposition 1 that to compute the entropy $H_{\mathbf{Z}_0}^{\mathrm{SE}_k}$, we need only the joint distribution of $U=\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0$ and $S=\mathbf{Z}_0^T\mathbf{Z}_0$, where $\mathbf{Z}_0\sim\mathrm{SE}_k(\mathbf{0},\mathbf{I}_k,\bar{\boldsymbol{\eta}},h^{(k+1)})$. For this, the next result is necessary, the proof of which is given in the Appendix.

Proposition 2. *Let* $U=\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0$ *and* $S=\mathbf{Z}_0^T\mathbf{Z}_0$, *where* $\mathbf{Z}_0\sim\mathrm{SE}_k(\mathbf{0},\mathbf{I}_k,\bar{\boldsymbol{\eta}},h^{(k+1)})$. *Then,* $(U,S)\overset{d}{=}(\|\bar{\boldsymbol{\eta}}\|W,S)$, *where* $\|\bar{\boldsymbol{\eta}}\|=\bar{\boldsymbol{\eta}}^T\bar{\boldsymbol{\eta}}=(\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta})^{1/2}$, *and for* $k\geq 2$, *the joint probability density function of* (W,S) *can be computed as* $p_{W,S}(u,s)=p_{W|S=s}(u)p_S(s)$, *where*

$$p_{W|S=s}(u)=\frac{2}{\sqrt{s}}\left(1-\frac{u^2}{s}\right)^{\frac{k-1}{2}-1}F(\|\bar{\boldsymbol{\eta}}\|u,h_s^{(1)}),\quad |u|<\sqrt{s},$$

and

$$p_S(s)\equiv g(s)=\frac{\pi^{k/2}}{\Gamma(\frac{k}{2})}s^{\frac{k}{2}-1}h^{(k)}(s),\quad s>0.$$

3.2. The multivariate skew-normal distribution

The multivariate skew-normal distribution has been introduced by Azzalini & Dalla Valle (1996). 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. 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 $\boldsymbol{\Omega}\in\mathbb{R}^{k\times k}$ and shape/skewness parameter $\boldsymbol{\eta}\in\mathbb{R}^k$, denoted by $\mathbf{Z}\sim\mathrm{SN}_k(\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\eta})$, if its probability density function is

$$p_{\mathbf{Z}}(\mathbf{z})=2\phi_k(\mathbf{z};\boldsymbol{\xi},\boldsymbol{\Omega})\Phi\{\boldsymbol{\eta}^T(\mathbf{z}-\boldsymbol{\xi})\},\quad \mathbf{z}\in\mathbb{R}^k,\tag{9}$$

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

We can rewrite (9) as

$$p_{\mathbf{Z}}(\mathbf{z})=|\boldsymbol{\Omega}|^{-1/2}p_{\mathbf{Z}_0}(\mathbf{z}_0),\ \text{with}\ p_{\mathbf{Z}_0}(\mathbf{z}_0)=2\phi_k(\mathbf{z}_0)\Phi(\bar{\boldsymbol{\eta}}^T\mathbf{z}_0),\tag{10}$$

where $\bar{\boldsymbol{\eta}}=\boldsymbol{\Omega}^{1/2}\boldsymbol{\eta}$. Hence, we can apply lemma 1 to obtain the entropy for the skew-normal model. For this, we need the following preliminary result to simplify the computation of this

entropy. Its proof can be found in Arellano-Valle & Genton (2010a, c). Let $\mathbf{Z}_0 \sim \mathrm{SN}_k(\boldsymbol{\alpha}) \equiv \mathrm{SN}_k(\mathbf{0}, \mathbf{I}_k, \boldsymbol{\alpha})$ denote the standardized k-variate skew-normal distribution with probability density function $p_{\mathbf{Z}_0}(\mathbf{z}_0) = 2\phi_k(\mathbf{z}_0)\Phi(\boldsymbol{\alpha}^T\mathbf{z}_0)$.

Lemma 2. *Let* $\mathbf{Z} \sim \mathrm{SN}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ *and* $\mathbf{Z}_0 = \boldsymbol{\Omega}^{-1/2}(\mathbf{Z} - \boldsymbol{\xi})$. *Let also* $\mathbf{Z}_{0N} \sim N_k(\mathbf{0}, \mathbf{I}_k)$ *and* $W \sim \mathrm{SN}_1(\|\bar{\boldsymbol{\eta}}\|)$. *Then,* $\mathbf{Z}_0 \sim \mathrm{SN}_k(\mathbf{0}, \mathbf{I}_k, \bar{\boldsymbol{\eta}}) \equiv \mathrm{SN}_k(\bar{\boldsymbol{\eta}})$, $\boldsymbol{\eta}^T(\mathbf{Z} - \boldsymbol{\xi}) = \bar{\boldsymbol{\eta}}^T\mathbf{Z}_0 \stackrel{d}{=} \|\bar{\boldsymbol{\eta}}\| W$ *and* $g(\mathbf{Z}_0) \stackrel{d}{=} g(\mathbf{Z}_{0N})$ *for any even function* g.

From lemmas 1 and 2, we have the following result.

Proposition 3. *The entropy of a skew-normal random vector* $\mathbf{Z} \sim \mathrm{SN}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ *is*

$$H_{\mathbf{Z}}^{\mathrm{SN}_k} = H_{\mathbf{Z}_N}^{N_k} - E\left[\log\{2\Phi(\|\bar{\boldsymbol{\eta}}\| W)\}\right],$$

where $H_{\mathbf{Z}_N}^{N_k}$ *is the entropy of* $\mathbf{Z}_N \sim N_k(\boldsymbol{\xi}, \boldsymbol{\Omega})$ *given in* (6), *and* $W \sim \mathrm{SN}_1(\|\bar{\boldsymbol{\eta}}\|)$.

Proof. Let $\mathbf{Z}_0 = \boldsymbol{\Omega}^{-1/2}(\mathbf{Z} - \boldsymbol{\xi})$. By lemmas 1 and 2 and by (10), we have

$$\begin{aligned} E[\log\{p_{\mathbf{Z}}(\mathbf{Z})\}] &= -(1/2)\log|\boldsymbol{\Omega}| + E[\log\{2\phi_k(\mathbf{Z}_0)\Phi(\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0)\}] \\ &= -(1/2)\log|\boldsymbol{\Omega}| + E[\log\{\phi_k(\mathbf{Z}_0)\}] + E[\log\{2\Phi(\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0)\}] \\ &= \underbrace{-(1/2)\log|\boldsymbol{\Omega}| + E[\log\{\phi_k(\mathbf{Z}_{0N})\}]}_{-H_{\mathbf{Z}_N}^{N_k}} + E[\log\{2\Phi(\|\bar{\boldsymbol{\eta}}\| W)\}], \end{aligned}$$

because $E[\log\{\phi_k(\mathbf{Z}_0)\}] = E[\log\{\phi_k(\mathbf{Z}_{0N})\}]$, since the function ϕ_k is even, and because $E[\log\{2\Phi(\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0)\}] = E[\log\{2\Phi(\|\bar{\boldsymbol{\eta}}\| W)\}]$, since $\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0 \stackrel{d}{=} \|\bar{\boldsymbol{\eta}}\| W$.

To derive the mutual information index of the multivariate skew-normal distribution, we need the following result about its marginal distributions. Let

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathrm{SN}_{n+m}\left(\begin{pmatrix} \boldsymbol{\xi}_{\mathbf{X}} \\ \boldsymbol{\xi}_{\mathbf{Y}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{XX}} & \boldsymbol{\Omega}_{\mathbf{XY}} \\ \boldsymbol{\Omega}_{\mathbf{YX}} & \boldsymbol{\Omega}_{\mathbf{YY}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\eta}_{\mathbf{X}} \\ \boldsymbol{\eta}_{\mathbf{Y}} \end{pmatrix}\right).$$

Then $\mathbf{X} \sim \mathrm{SN}_n(\boldsymbol{\xi}_{\mathbf{X}}, \boldsymbol{\Omega}_{\mathbf{XX}}, \boldsymbol{\eta}_{\mathbf{X(Y)}})$ and $\mathbf{Y} \sim \mathrm{SN}_m(\boldsymbol{\xi}_{\mathbf{Y}}, \boldsymbol{\Omega}_{\mathbf{YY}}, \boldsymbol{\eta}_{\mathbf{Y(X)}})$ where

$$\boldsymbol{\eta}_{\mathbf{X(Y)}} = \frac{\boldsymbol{\eta}_{\mathbf{X}} + \boldsymbol{\Omega}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}_{\mathbf{XY}}\boldsymbol{\eta}_{\mathbf{Y}}}{\sqrt{1 + \boldsymbol{\eta}_{\mathbf{Y}}^T\boldsymbol{\Omega}_{\mathbf{YY\cdot X}}\boldsymbol{\eta}_{\mathbf{Y}}}} \quad \text{and} \quad \boldsymbol{\eta}_{\mathbf{Y(X)}} = \frac{\boldsymbol{\eta}_{\mathbf{Y}} + \boldsymbol{\Omega}_{\mathbf{YY}}^{-1}\boldsymbol{\Omega}_{\mathbf{YX}}\boldsymbol{\eta}_{\mathbf{X}}}{\sqrt{1 + \boldsymbol{\eta}_{\mathbf{X}}^T\boldsymbol{\Omega}_{\mathbf{XX\cdot Y}}\boldsymbol{\eta}_{\mathbf{X}}}}.$$

Consequently, by proposition 1, we obtain the following results for the marginal and joint skew-normal entropies.

Proposition 4. *Let*

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathrm{SN}_{n+m}\left(\begin{pmatrix} \boldsymbol{\xi}_{\mathbf{X}} \\ \boldsymbol{\xi}_{\mathbf{Y}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{XX}} & \boldsymbol{\Omega}_{\mathbf{XY}} \\ \boldsymbol{\Omega}_{\mathbf{YX}} & \boldsymbol{\Omega}_{\mathbf{YY}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\eta}_{\mathbf{X}} \\ \boldsymbol{\eta}_{\mathbf{Y}} \end{pmatrix}\right).$$

Then,

$$\begin{aligned} H_{\mathbf{X}}^{\mathrm{SN}_n} &= \frac{1}{2}\log|\boldsymbol{\Omega}_{\mathbf{XX}}| + \frac{n}{2}\{1 + \log(2\pi)\} - E\left[\log\{2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{X(Y)}}\| W_{\mathbf{X}})\}\right], \\ H_{\mathbf{Y}}^{\mathrm{SN}_m} &= \frac{1}{2}\log|\boldsymbol{\Omega}_{\mathbf{YY}}| + \frac{m}{2}\{1 + \log(2\pi)\} - E\left[\log\{2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{Y(X)}}\| W_{\mathbf{Y}})\}\right], \\ H_{\mathbf{XY}}^{\mathrm{SN}_{n+m}} &= \frac{1}{2}\log|\boldsymbol{\Omega}| + \frac{n+m}{2}\{1 + \log(2\pi)\} - E\left[\log\{2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{XY}}\| W_{\mathbf{XY}})\}\right], \end{aligned}$$

with $W_{\mathbf{X}} \sim \mathrm{SN}_1(\|\bar{\boldsymbol{\eta}}_{\mathbf{X}(\mathbf{Y})}\|)$, $W_{\mathbf{Y}} \sim \mathrm{SN}_1(\|\bar{\boldsymbol{\eta}}_{\mathbf{Y}(\mathbf{X})}\|)$ *and* $W_{\mathbf{XY}} \sim \mathrm{SN}_1(\|\bar{\boldsymbol{\eta}}_{\mathbf{XY}}\|)$, *where*

$$\|\bar{\boldsymbol{\eta}}_{\mathbf{X}(\mathbf{Y})}\| = (\boldsymbol{\eta}_{\mathbf{X}(\mathbf{Y})}^T \boldsymbol{\Omega}_{\mathbf{XX}} \boldsymbol{\eta}_{\mathbf{X}(\mathbf{Y})})^{1/2}, \quad \|\bar{\boldsymbol{\eta}}_{\mathbf{Y}(\mathbf{X})}\| = (\boldsymbol{\eta}_{\mathbf{Y}(\mathbf{X})}^T \boldsymbol{\Omega}_{\mathbf{YY}} \boldsymbol{\eta}_{\mathbf{Y}(\mathbf{X})})^{1/2}$$

and

$$\|\bar{\boldsymbol{\eta}}_{\mathbf{XY}}\| = (\boldsymbol{\eta}_{\mathbf{X}(\mathbf{Y})}^T \boldsymbol{\Omega}_{\mathbf{XX}} \boldsymbol{\eta}_{\mathbf{X}(\mathbf{Y})} + \boldsymbol{\eta}_{\mathbf{Y}(\mathbf{X})}^T \boldsymbol{\Omega}_{\mathbf{YY}} \boldsymbol{\eta}_{\mathbf{Y}(\mathbf{X})} + 2\boldsymbol{\eta}_{\mathbf{X}(\mathbf{Y})}^T \boldsymbol{\Omega}_{\mathbf{XY}} \boldsymbol{\eta}_{\mathbf{Y}(\mathbf{X})})^{1/2}.$$

Thus, we obtain from (3) that the skew-normal mutual information index between $\mathbf{X}$ and $\mathbf{Y}$ is

$$I_{\mathbf{XY}}^{\mathrm{SN}_{n+m}} = I_{\mathbf{XY}}^{N_{n+m}}(\boldsymbol{\Omega}) + E\left\{\log\left(\frac{\mathrm{V}_{\mathbf{XY}}}{\mathrm{V}_{\mathbf{X}(\mathbf{Y})}\mathrm{V}_{\mathbf{Y}(\mathbf{X})}}\right)\right\},$$

where $V_{\mathbf{XY}} = 2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{XY}}\| W_{\mathbf{XY}})$, $V_{\mathbf{X}(\mathbf{Y})} = 2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{X}(\mathbf{Y})}\| W_{\mathbf{X}(\mathbf{Y})})$ and $V_{\mathbf{Y}(\mathbf{X})} = 2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{Y}(\mathbf{X})}\| W_{\mathbf{Y}(\mathbf{X})})$. An alternative way to compute the term $E[\log\{2\Phi(\alpha W_{\mathrm{SN}})\}]$, where $W_{\mathrm{SN}} \sim \mathrm{SN}_1(\alpha)$, is given by the following result.

Proposition 5. *Let* $W_{\mathrm{SN}} \sim \mathrm{SN}_1(\alpha)$ *and* $W_N \sim N_1(0,1)$. *Then,*

$$E[\log\{2\Phi(\alpha W_{\mathrm{SN}})\}] = E[2\Phi(\alpha W_N)\log\{2\Phi(\alpha W_N)\}].$$

Proof. We have directly that

$$E[\log\{2\Phi(\alpha W)\}] = \int_{-\infty}^{\infty} \log\{2\Phi(\alpha w)\} 2\Phi(\alpha w)\phi(w)\, \mathrm{d}w = E[2\Phi(\alpha W_N)\log\{2\Phi(\alpha W_N)\}].$$

From proposition 5, we then have

$$I_{\mathbf{XY}}^{\mathrm{SN}_{n+m}} = I_{\mathbf{XY}}^{N_{n+m}}(\boldsymbol{\Omega}) + E\left\{\log\left(\frac{U_{\mathbf{XY}}}{U_{\mathbf{X}(\mathbf{Y})}U_{\mathbf{Y}(\mathbf{X})}}\right)\right\},$$

where

$$U_{\mathbf{XY}} = 2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{XY}}\| W_N)\log\{2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{XY}}\| W_N)\},$$
$$U_{\mathbf{X}(\mathbf{Y})} = 2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{X}(\mathbf{Y})}\| W_N)\log\{2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{X}(\mathbf{Y})}\| W_N)\},$$
$$U_{\mathbf{Y}(\mathbf{X})} = 2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{Y}(\mathbf{X})}\| W_N)\log\{2\Phi(\|\bar{\boldsymbol{\eta}}_{\mathbf{Y}(\mathbf{X})}\| W_N)\}.$$

In Fig. 1, we can see that $f(\alpha, w) = \log\{2\Phi(\alpha w)\}2\Phi(\alpha w)\phi(w)$ is 0 when $w \to \pm\infty$ for any α. In particular for $\alpha > 0$, the probability density function $\phi(w)$ tends to 0 when $w \to \pm\infty$. Hence, this allows the convergence of the integral in proposition 3.

3.3. The multivariate skew-t distribution

We say that a random vector $\mathbf{Z} \in \mathbb{R}^k$ has a skew-*t* distribution with location vector $\boldsymbol{\xi} \in \mathbb{R}^k$, dispersion matrix $\boldsymbol{\Omega} \in \mathbb{R}^{k\times k}$, shape/skewness parameter $\boldsymbol{\eta} \in \mathbb{R}^k$ and $\nu > 0$ degrees of freedom, denoted by $\mathbf{Z} \sim \mathrm{ST}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu)$, if its probability density function is

$$p_{\mathbf{Z}}(\mathbf{z}) = 2|\boldsymbol{\Omega}|^{-1/2} t_k(\mathbf{z}_0; \nu) T\left(\sqrt{\frac{\nu + k}{\nu + \|\mathbf{z}_0\|^2}}\, \bar{\boldsymbol{\eta}}^T \mathbf{z}_0; \nu + k\right),$$

where as before $\mathbf{z}_0 = \boldsymbol{\Omega}^{-1/2}(\mathbf{z} - \boldsymbol{\xi})$, $t_k(\mathbf{x}; \nu)$ is the $t_k(\mathbf{0}, \mathbf{I}_k, \nu)$ probability density function and $T(x; \nu + k)$ is the $T_1(0, 1, \nu + k)$ cumulative distribution function; see Branco & Dey (2001), Azzalini & Capitanio (2003), Gupta (2003) and Ma & Genton (2004).

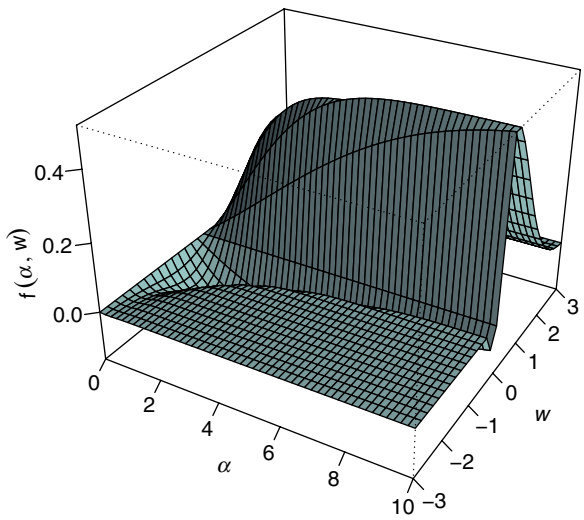

Fig. 1. Behaviour of $f(\alpha, w) = \log\{2\Phi(\alpha w)\}2\Phi(\alpha w)\phi(w)$.

For this model, we have $\mathbf{Z}_0 \sim \mathrm{ST}_k(\mathbf{0}, \mathbf{I}_k, \bar{\boldsymbol{\eta}}, \nu)$, with probability density function $p_{\mathbf{Z}_0}(\mathbf{z}_0) = 2t_k(\mathbf{z}_0; \nu)T\left(\sqrt{\frac{\nu+k}{\nu+\|\mathbf{z}_0\|^2}}\bar{\boldsymbol{\eta}}^T\mathbf{z}_0; \nu+k\right)$, so that in (4) we obtain

$$H^{\mathrm{ST}_k}_{\mathbf{Z}_0} = H^{T_k}_{\mathbf{Z}_0} - E\left[\log\left\{2T\left(\sqrt{\frac{\nu+k}{\nu+\|\mathbf{Z}_0\|^2}}\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0; \nu+k\right)\right\}\right], \quad \begin{array}{c}\text{[Correction made here after}\\ \text{initial online publication]}\end{array} \tag{11}$$

where $H^{T_k}_{\mathbf{Z}_0}$ is given by formula (7). To compute the last factor in the skew-t entropy (11) by integration in only one dimension, we need the following result whose proof is given by Arellano-Valle (2010).

Lemma 3. *Let* $\mathbf{Z}_0 \sim \mathrm{ST}_k(\mathbf{0}, \mathbf{I}_k, \bar{\boldsymbol{\eta}}, \nu)$. *Then*,

$$\sqrt{\frac{\nu+k}{\nu+\|\mathbf{Z}_0\|^2}}\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0 \overset{d}{=} \frac{\sqrt{\nu+k}\|\bar{\boldsymbol{\eta}}\|W_{\mathrm{ST}}}{\sqrt{\nu+k-1+W^2_{\mathrm{ST}}}},$$

where $W_{\mathrm{ST}} \sim \mathrm{ST}_1(0, 1, \|\bar{\boldsymbol{\eta}}\|, \nu+k-1)$.

Since for large values of ν the multivariate Student's t, and hence the skew-t, distributions converge to the normal and skew-normal ones, respectively, it is straightforward to see from (11) that the $H^{\mathrm{ST}_k}_{\mathbf{Z}_0}$ entropy converges to $H^{\mathrm{SN}_k}_{\mathbf{Z}_0}$ for any values of $\bar{\boldsymbol{\eta}}$ as $\nu \to \infty$. The behaviour of this convergence and the respective entropies are simulated/reported for several values of $\alpha = \|\bar{\boldsymbol{\eta}}\|$ and ν in the next section.

The following results are related with the marginal distributions from a multivariate skew-t distribution. For the proof of the latter, see Lee *et al.* (2010) or Arellano-Valle & Genton (2010a). Let

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim \mathrm{ST}_{n+m}\left(\begin{pmatrix}\boldsymbol{\xi}_{\mathbf{X}}\\ \boldsymbol{\xi}_{\mathbf{Y}}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Omega}_{\mathbf{XX}} & \boldsymbol{\Omega}_{\mathbf{XY}}\\ \boldsymbol{\Omega}_{\mathbf{YX}} & \boldsymbol{\Omega}_{\mathbf{YY}}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\eta}_{\mathbf{X}}\\ \boldsymbol{\eta}_{\mathbf{Y}}\end{pmatrix}, \nu\right).$$

Then $\mathbf{X} \sim \mathrm{ST}_n(\boldsymbol{\xi}_{\mathbf{X}}, \boldsymbol{\Omega}_{\mathbf{XX}}, \boldsymbol{\eta}_{\mathbf{X(Y)}}, \nu)$ and $\mathbf{Y} \sim \mathrm{ST}_m(\boldsymbol{\xi}_{\mathbf{Y}}, \boldsymbol{\Omega}_{\mathbf{YY}}, \boldsymbol{\eta}_{\mathbf{Y(X)}}, \nu)$. By lemma 3, we have

$$E\left[\log\left\{2T\left(\sqrt{\frac{\nu+k}{\nu+\|\mathbf{Z}_0\|^2}}\bar{\boldsymbol{\eta}}^T\mathbf{Z}_0;\nu+k\right)\right\}\right] = E\left[\log\left\{2T\left(\frac{\sqrt{\nu+k}\|\bar{\boldsymbol{\eta}}\|W_{\mathrm{ST}}}{\sqrt{\nu+k-1+W_{\mathrm{ST}}^2}};\nu+k\right)\right\}\right],$$

where $W_{\mathrm{ST}} \sim \mathrm{ST}_1(0,1,\|\bar{\boldsymbol{\eta}}\|,\nu+k-1)$. So, considering the above results, we can deduce the mutual information index for the skew-t case as follows.

Proposition 6. *Let*

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim \mathrm{ST}_{n+m}\left(\begin{pmatrix}\boldsymbol{\xi}_{\mathbf{X}}\\ \boldsymbol{\xi}_{\mathbf{Y}}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Omega}_{\mathbf{XX}} & \boldsymbol{\Omega}_{\mathbf{XY}}\\ \boldsymbol{\Omega}_{\mathbf{YX}} & \boldsymbol{\Omega}_{\mathbf{YY}}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\eta}_{\mathbf{X}}\\ \boldsymbol{\eta}_{\mathbf{Y}}\end{pmatrix}, \nu\right).$$

Then,

$$I_{\mathbf{XY}}^{\mathrm{ST}_{n+m}} = I_{\mathbf{XY}}^{T_{n+m}} + E\left\{\log\left(\frac{C_{\mathbf{XY}}}{C_{\mathbf{X(Y)}}C_{\mathbf{Y(X)}}}\right)\right\},$$

where

$$C_{\mathbf{XY}} = 2T\left(\frac{\sqrt{\nu+n+m}\|\bar{\boldsymbol{\eta}}_{\mathbf{XY}}\|W_{\mathrm{ST}}}{\sqrt{\nu+n+m-1+W_{\mathrm{ST}}^2}};\nu+n+m\right),$$

$$C_{\mathbf{X(Y)}} = 2T\left(\frac{\sqrt{\nu+n}\|\bar{\boldsymbol{\eta}}_{\mathbf{X(Y)}}\|W_{\mathrm{ST}}}{\sqrt{\nu+n-1+W_{\mathrm{ST}}^2}};\nu+n\right),$$

$$C_{\mathbf{Y(X)}} = 2T\left(\frac{\sqrt{\nu+m}\|\bar{\boldsymbol{\eta}}_{\mathbf{Y(X)}}\|W_{\mathrm{ST}}}{\sqrt{\nu+m-1+W_{\mathrm{ST}}^2}};\nu+m\right),$$

and $W_{\mathrm{ST}} \sim \mathrm{ST}_1(0,1,\|\bar{\boldsymbol{\eta}}\|,\nu+k-1)$.

4. Numerical results

A convenient and fast method to compute the entropies presented in this paper is based on the numerical integration Quadpack (a Subroutine Package for Automatic Integration) implemented in the integrate R (R Development Core Team, 2010) function.

The results are shown in Fig. 2 for the dimension $k=1$, dispersion matrix $\Omega=1$, skewness parameter $\alpha \in [0.1, 20]$, integration interval $[-10^3, 10^3]$ and degrees of freedom $\nu = 1, 2, \ldots, 185$ to illustrate the entropies:

$$H_X^{\mathrm{SN}_1}(\alpha) = \frac{1}{2}\{1+\log(2\pi)\} - E\left[\log\{2\Phi(\alpha W_{\mathrm{SN}})\}\right], \tag{12}$$

$$H_X^{\mathrm{ST}_1}(\alpha,\nu) = -\log\left\{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\right\} + \frac{\nu+1}{2}\left\{\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right\} - E\left[\log\left\{2T\left(\sqrt{\frac{\nu+1}{\nu+W_{\mathrm{ST}}^2}}\alpha W_{\mathrm{ST}};\nu+1\right)\right\}\right], \tag{13}$$

where, as were defined before, $W_{\mathrm{SN}} \sim \mathrm{SN}_1(\alpha)$ and $W_{\mathrm{ST}} \sim \mathrm{ST}_1(0,1,\alpha,\nu)$.

We can see in Fig. 2 that the numerical implementation suggests the convergence of the integrals involved in (12) and (13). The Quadpack method is more precise and efficient in terms of computational time than other methods such as Monte Carlo.

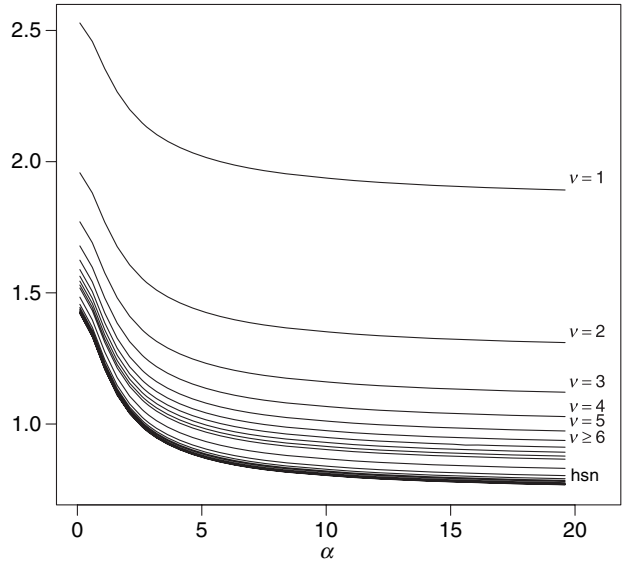

Fig. 2. Plots of the skew-*t* entropy $H_X^{\mathrm{ST}_1}(\alpha, \nu)$ for $\nu = 1, 2, \ldots, 185$ degrees of freedom and the skew-normal entropy $H_X^{\mathrm{SN}_1}(\alpha)$ (denoted by hsn) as a function of α.

The convergence of the skew-*t* entropy to the skew-normal entropy is obtained quickly for values of $\nu \geq 12$. In other words, greater values of the marginal skew-*t* entropy are produced by small values of ν. As expected, for the normal and Student's *t* marginal cases ($\alpha = 0$), we have that the entropy is maximized and decreasing for greater values of α. When $\alpha \to \infty$, the entropy tends to a constant and for $\alpha > 20$, it is almost that of a half-normal distribution already.

5. Application to a network design

The design of optimal networks is a crucial problem in engineering and environmental pollutant analysis. Among several existing methods, the computation of the Shannon information index (Silva & Quiroz, 2003) and Bayesian entropy (Ainslie *et al.*, 2009) are useful to design a meteorological monitoring network. A practical illustration of our methodology is provided in this section on a subset (see reference about the MACAM network in Seremi de Salud, 2006) of time series of ozone concentrations at seven monitoring stations denoted by $\mathbf{X}Y = \{\mathrm{F}, \mathrm{L}, \mathrm{M}, \mathrm{N}, \mathrm{O}, \mathrm{P}, \mathrm{Q}\}$ with $n = 7 \times 24 \times 31 = 5208$ hourly observations in March 2006. In this case, the pollutant data contain abnormalities in the observations, specifically skewness in the empirical distribution. Therefore, standard distributions are very limited to represent such data. In this study, we proceed to analyse the optimization of this monitoring network as follows:

1. We define the moving average smoothing (MA_s) with seasonal parameter s for station j at time t:

$$T_{t,j}^{s} = \frac{1}{s} \sum_{i=t-s}^{t} y_{ij},$$

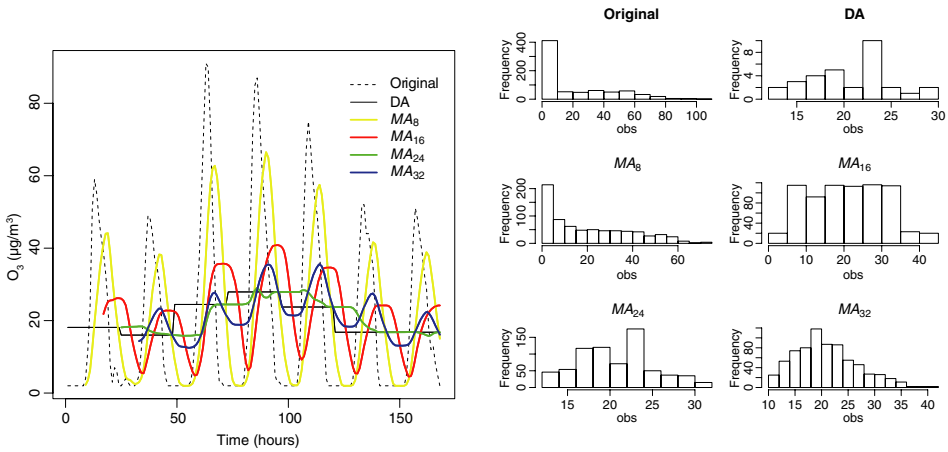

Fig. 3. Left: Graphic of original data ($s=1$) with the transformations of moving average (MA_s) for $s=\{8, 16, 24, 32\}$ hours and daily average (DA) for 1 March 2006 to 7 March 2006 of station L. Right: Several histograms for the transformed ozone data mentioned before.

where y_{ij} is the observation for station j at the ith time. For small values of s, the smoothing detects the influence of the minimum and maximum values; however, for larger values of s, the transformation $T^s_{t,j}$ decreases the variance of the time series (see Fig. 3, left panel).

2. In this application, we consider a multivariate data set $\mathbf{X}Y$ of seven stations, a subset $\mathbf{X}$ of six monitoring stations and we choose one non-monitoring station Y to be removed from $\mathbf{X}Y$ for each value of s. We compute the mutual information index $I_{\mathbf{X}Y}$ related to multivariate normal, skew-normal and skew-t distributions.

3. To find or not evidence to reject the null hypothesis about the marginal variable Y having skew-normal or skew-t distributions, it is possible to compute the p-values according to the goodness-of-fit test proposed by Kolmogorov–Smirnov, for all s and the variables defined in step 2. Alternatively, it is possible to create a PP-plot and compare the performance of these fitted distributions.

4. We calculate the maximum likelihood estimators (MLEs) of the location, dispersion and shape/skewness parameters using the sn library of R (Azzalini, 2008) for variables defined in the previous steps, for each value of s. From Azzalini & Capitanio (1999), for a sample of independent observations $\mathbf{Z}_i \sim \mathrm{SN}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta})$, $i=1,\ldots,n$, we estimate the parameters by numerically maximizing the log-likelihood function:

$$\log L(\boldsymbol{\Theta}_{\mathrm{SN}}) \propto -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-1}\tilde{\mathbf{V}}) + \sum_{i=1}^{n}\log[\Phi\{\boldsymbol{\eta}^T(\mathbf{z}_i-\boldsymbol{\xi})\}],$$

where $\boldsymbol{\Theta}_{\mathrm{SN}}=\{\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}\}$ and $\tilde{\mathbf{V}}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i-\boldsymbol{\xi})(\mathbf{z}_i-\boldsymbol{\xi})^T$. Now, if $\mathbf{Z}_i \sim \mathrm{ST}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu)$, $i=1,\ldots,n$, we use the reparameterization and log-likelihood of Azzalini & Capitanio (2003). Let $\boldsymbol{\Omega}=(\mathbf{A}^T\mathbf{D}\mathbf{A})^{-1}$, where $\mathbf{A}$ is an upper triangular $k\times k$ matrix with diagonal terms equal to 1, $\mathbf{D}=\mathrm{diag}(\mathrm{e}^{-2\rho})$ and $\boldsymbol{\rho}\in\mathbb{R}^k$. For the parameter set $\boldsymbol{\Theta}_{\mathrm{ST}}=\{\boldsymbol{\xi}, \mathbf{A}, \boldsymbol{\rho}, \boldsymbol{\eta}, \log(\nu)\}$, we obtain

$$\log L(\boldsymbol{\Theta}_{\mathrm{ST}}) \propto \frac{n}{2}\log|\mathbf{D}| + \sum_{i=1}^{n}\log\{t_k(\mathbf{z}_i-\boldsymbol{\xi};\nu)\} + \sum_{i=1}^{n}\log\left\{T\left(\boldsymbol{\eta}^T(\mathbf{z}_i-\boldsymbol{\xi})\sqrt{\frac{\nu+k}{\nu+s_i}};\nu+k\right)\right\},$$

where $s_i=(\mathbf{z}_i-\boldsymbol{\xi})^T\boldsymbol{\Omega}^{-1}(\mathbf{z}_i-\boldsymbol{\xi})$. So, from $\hat{\boldsymbol{\Theta}}_{\mathrm{ST}}=\arg\max_{\boldsymbol{\Theta}_{\mathrm{ST}}}\{\log L(\boldsymbol{\Theta}_{\mathrm{ST}})\}$, we can obtain the MLEs $\{\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\eta}}, \hat{\nu}\}$.

Table 1. *P-values for Mardia's goodness-of-fit test of multivariate normality applied to the joint* $\mathbf{X}Y$ *and to the* $\mathbf{X}$ *multivariate variables, and p-values for Shapiro's test for the* Y *marginal variables. All tests are for daily average transformation of the original data. The p-values higher than the probability* (0.05) *related to a* 5 *per cent confidence level (marked in* **bold***) lead to multivariate normality used in the last column to compute the mutual information index* ($I_{\mathbf{X}Y}^{N_{6+1}}$) *for this distribution (the first- and second-largest values are marked in* **bold***)*

Monitored stations		$H_0: \beta_{1,k} = 0$		$H_0: \beta_{2,k} = k(k+2)$		Shapiro's test	Normal
Yes ($\mathbf{X}$)	No (Y)	$\mathbf{X}Y$	$\mathbf{X}$	$\mathbf{X}Y$	$\mathbf{X}$	Y	$I_{\mathbf{X}Y}^{N_{6+1}}$
L, M, N, O, P, Q	F		**0.429**		**0.140**	**0.115**	0.970
F, M, N, O, P, Q	L		**0.710**		**0.136**	**0.481**	0.963
F, L, N, O, P, Q	M		**0.299**		**0.218**	**0.991**	0.514
F, L, M, O, P, Q	N	**0.765**	**0.785**	**0.056**	**0.059**	0.025	**1.107**
F, L, M, N, P, Q	O		**0.935**		0.028	**0.096**	0.769
F, L, M, N, O, Q	P		**0.927**		**0.078**	**0.706**	0.312
F, L, M, N, O, P	Q		**0.468**		**0.138**	**0.275**	**1.280**

Table 2. *P-values for the Kolmogorov–Smirnov goodness-of-fit test of multivariate skew-normality applied to marginal variables. The p-values marked in* **bold** *are higher than the probability* (0.05) *related to a* 5 *per cent confidence level*

Skew-normal							
s	F	L	M	N	O	P	Q
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.004	0.000
16	0.005	0.017	0.026	0.013	0.026	0.001	0.009
24	**0.060**	0.000	**0.096**	0.000	**0.102**	0.002	0.000
32	**0.913**	**0.382**	**0.945**	**0.297**	**0.534**	**0.361**	**0.167**
40	**0.170**	**0.711**	**0.944**	**0.483**	**0.746**	**0.255**	**0.770**

5. For the variables selected in step 2, let $p_{\mathbf{X},Y}$ represent the multivariate normal, skew-normal or skew-t joint probability density function between $\mathbf{X}$ and Y. The multivariate Student's t case is not included in this application because the estimation of the skewness parameter in the skew-t case is clearly larger than zero. Let $p_{\mathbf{X}}$ and p_Y be the corresponding marginal densities. Then, the Shannon mutual information index has been derived in sections 2 and 3. If $\mathbf{X}$ and Y are independent, then $I_{\mathbf{X}Y} = 0$. We can interpret this as when the monitoring stations do not provide information on the chosen non-monitoring station and vice versa (Silva & Quiroz, 2003). Then, from the MLEs obtained in step 4, we can obtain the terms $\|\bar{\boldsymbol{\eta}}_{\mathbf{X}(Y)}\|$, $\|\bar{\boldsymbol{\eta}}_{Y(\mathbf{X})}\|$ and $\|\bar{\boldsymbol{\eta}}_{\mathbf{X}Y}\|$ mentioned in proposition 4. So, we can compute the skew-normal and skew-t mutual information index for all s values from the MLEs in step 4 according to propositions 2 and 3.

6. We compare our approach with the normal case used by Silva & Quiroz (2003). That study analysed daily averaged data at eight stations during July 1998 (there exists an extra B station until 2003). However, other authors such as Ainslie *et al.* (2009) used a moving average according to government policies of their countries. In this work, we analyse an updated data in the Summer of 2006, because the ozone produces its minimum and maximum variabilities in that season. So, we proceed as follows: (a) we calculate the daily average (DA) of observations corresponding to fixed average of 24 hours; (b) we use the Box–Cox transformation to obtain near multivariate normality in the data:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(\lambda), & \text{if } \lambda = 0; \end{cases}$$

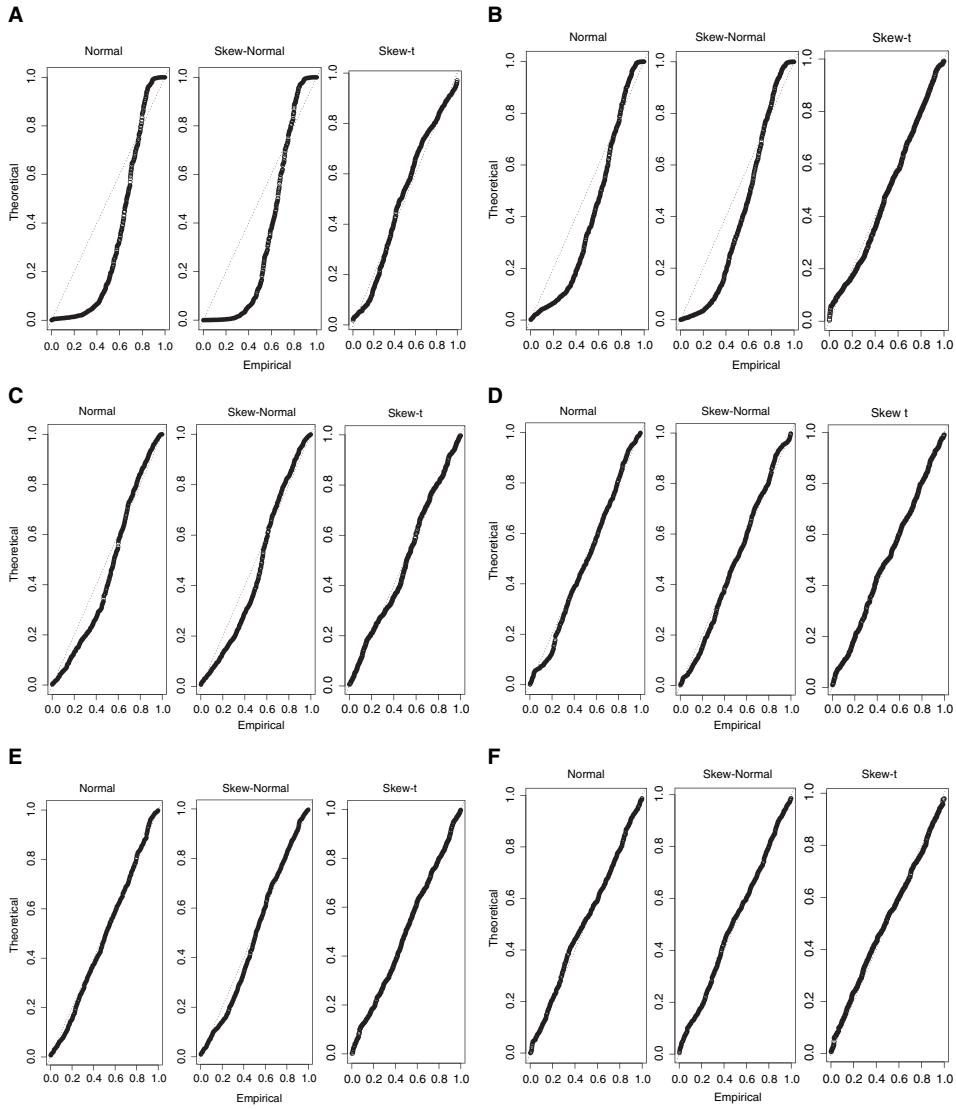

Fig. 4. Multivariate normal, skew-normal and skew-t PP-plots of (A) original data ($s=1$); and transformed data with (B) $s=8$, (C) $s=16$, (D) $s=24$, (E) $s=32$ and (F) $s=40$.

(c) we test multivariate normality according to Mardia (1985)'s test based on the measures of multivariate skewness ($\beta_{1,k}$) and multivariate kurtosis ($\beta_{2,k}$) with $k=7$ ($\mathbf{X}Y$: complete monitoring network with seven stations), $k=6$ ($\mathbf{X}$: monitoring network with six stations and one station removed) and $k=1$ (Y: removed station); and (d) we compute the multivariate normal Shannon index for this case according to section 2.3.

Figure 3 illustrates the behaviour of the transformations DA and MA, and the original data. Given that the period of the time series is 24 hours, values of s less than 24 preserve the variance of the original data but values higher than 24 decrease the variability. The amplitude of the data increases for the case of moving average $s=16$ and 32. About the distribution of the data, small values of s present heavy tails in the data, specifically for a moving average of $s=1$ and 8. For the cases of $s=24$ and DA, the distribution tends to be normal, and finally,

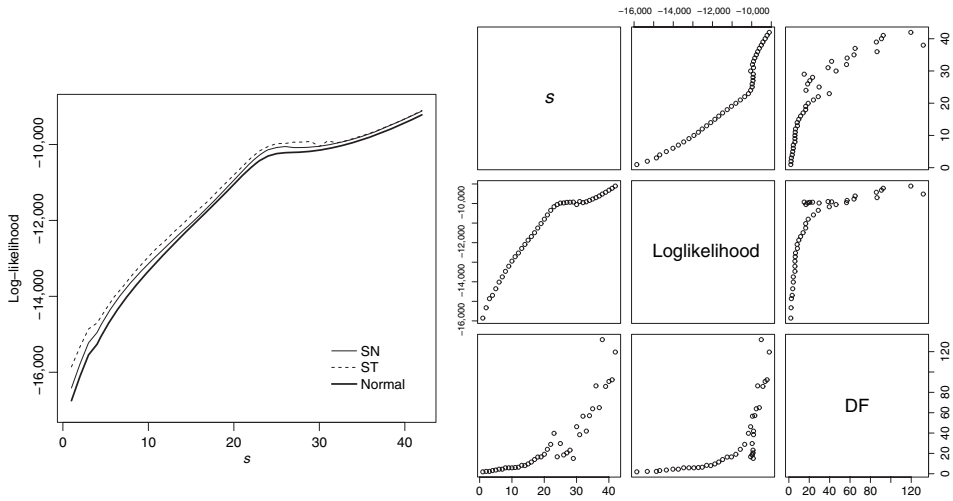

Fig. 5. Left: Graphic of log-likelihoods of multivariate skew-normal, skew-*t* and normal fits. Right: Scatter plots between s, log-likelihood of multivariate skew-*t* fit and degrees of freedom (DF) parameters.

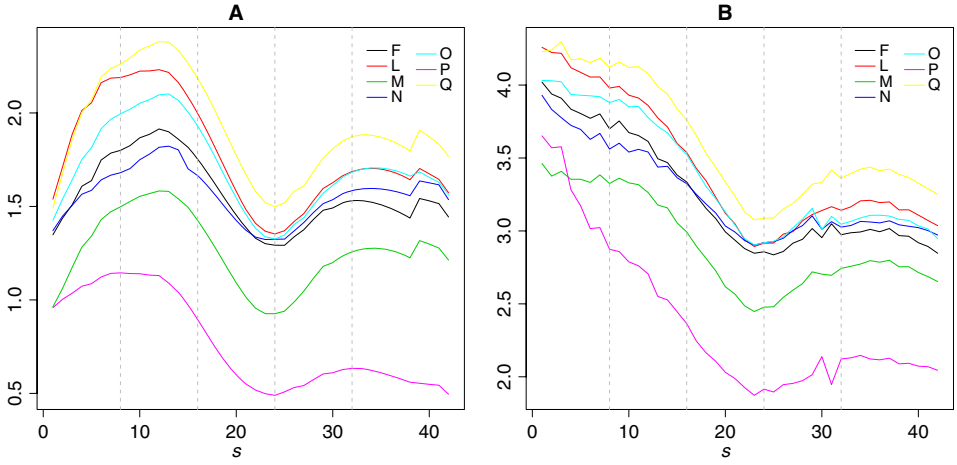

Fig. 6. Graphics of (A) $I^{\mathrm{SN}_{6+1}}_{\mathbf{X}Y}$ and (B) $I^{\mathrm{ST}_{6+1}}_{\mathbf{X}Y}$ when the $Y=$F, L, M, N, O, P or Q station is removed from the network $\mathbf{X}$. The vertical dotted grey lines correspond to $s=\{8,16,24,32\}$.

the case $s=32$ presents skewness and light tail in the distribution. These considerations are submitted to an analysis of distribution fit.

The results of the multivariate and univariate tests for joint and marginal variables for several transformations s described in steps 1 and 2 are shown in Tables 1 and 2 and in Fig. 4. They illustrate the flexibility of the class of the skew-*t* over skew-normal and normal distributions. We can see in Table 2 that for the transformations MA_s with $s=\{32,40\}$, the Kolmogorov–Smirnov test's p-values are higher than 0.05 in all marginal variable cases for the skew-normal distribution. On the other hand, the PP-plot (see Fig. 4) shows that the skew-*t* presents a better performance in the fit to the empirical distributions. However, in the cases of $s=\{1,8,16,24\}$, the null hypothesis is rejected in some marginal variables for the

Table 3. *Summary of results for skew-normal and skew-t distributions and different network configurations. The highest values for each transformation are marked in* **bold**

	Skew-normal							Skew-*t*						
s	F	L	M	N	O	P	Q	F	L	M	N	O	P	Q
1	1.348	**1.539**	0.963	1.370	1.425	0.959	1.492	4.021	**4.259**	3.463	3.930	4.032	3.652	4.230
8	1.800	2.189	1.498	1.680	1.995	1.145	**2.261**	3.700	3.980	3.325	3.561	3.880	2.874	**4.122**
16	1.752	1.997	1.425	1.665	1.931	0.895	**2.187**	3.338	3.543	2.993	3.327	3.529	2.376	**3.766**
24	1.293	1.353	0.926	1.324	1.326	0.490	**1.498**	2.857	2.916	2.477	2.920	2.919	1.915	**3.088**
32	1.530	1.687	1.257	1.585	1.683	0.634	**1.873**	2.974	3.142	2.743	3.026	3.046	2.122	**3.362**
40	1.530	1.674	1.299	1.626	1.655	0.548	**1.872**	2.932	3.115	2.716	3.021	3.042	2.072	**3.341**

skew-normal case. In addition, the multivariate and univariate data are normally distributed for the DA transformation according to Mardia's (joint cases) and Shapiro's (marginal cases) tests (see Table 1).

We can see in Fig. 5 (left panel) through the log-likelihood values that the $\mathrm{AIC} = -2 \times \{\log L(\boldsymbol{\Theta}) - n_p\}$ of the skew-*t* model are smaller than the skew-normal model values (the number of parameters n_p is irrelevant for these quantities). However, between $s = 32$ and 42, both log-likelihoods tend to be equal. Indeed, in Fig. 5 (right panel) when the period s increases, the v parameter increases too but for the values $s = \{20, \ldots, 42\}$, v increases quickly to 120. This behaviour may be explained by the good fit of the skew-normal for these s values according to the performance of the skew-*t* distribution fit (see Fig. 4D–4F).

The mutual information index is maximized when the station L is removed from the network $\mathbf{X}Y$ for the original data and for both skew-normal and skew-*t* distributions (see Fig. 6 and Table 3). However, for the cases of $s = \{8, 16, 24, 32, 40\}$, when the station Q is removed from the network, the mutual information index is maximized in both distribution cases; then, this induces a constant decision. It is then interesting to note that the mutual information index is maximum for the first value of s. As expected, both mutual information indexes have a seasonal effect of 24 hours in relation with the diurnal change of the ozone pollutant. However, if we look for the second-largest mutual information index for $s = 1$, we have that it is the Q station that needs to be removed. According to the procedure of Silva & Quiroz (2003) in the case of the normal distribution, the largest mutual information index is when the station Q is removed and, in the second place, when the station N is removed.

Different meteorological factors are not considered in this study but may be important in the decision to design an optimal network. However, our statistical tool uses the contained information of a selected appropriate data set and preserves some features of the data distribution such as skewness and heavy tails, necessary to make a better decision.

6. Conclusions

We have proposed an alternative way to compute the Shannon entropy and mutual information index for data with skewness and heavy tails. The calculation of this index produces a similar expression as for the normal and Student's t cases except for a new term represented by a one dimensional integral that can easily and quickly be computed by standard numerical methods. Moreover, a numerical study showed the convergence of this integral and in fact of the skew-normal and skew-*t* mutual information indexes. Finally, an analysis of an optimal network design of a classical pollutant was presented. The principal objective is to choose a network design in an optimal way through established methods of maximizing Shannon's index. We conclude from this analysis that the consideration of skewness and heavy tails in the model to fit the untransformed data produces different conclusions/

decisions than those obtained by applying the normal model to the transformed data. Moreover, data transformation to achieve multivariate normality is known to be challenging. The correct fit of the original data ensures the optimal maximization of the mutual information index and determines a better optimization network design. In this paper, we have given the tools to compute this new information index. Other methods could be derived from this index, for example, the *effectiveness index*, or to remove more than one station at a time (Silva & Quiroz, 2003).

The skew-elliptical entropy and mutual information index can be explored further by considering the whole class of selection elliptical distributions introduced by Arellano-Valle *et al.* (2006a). In fact, since a selection random vector $\mathbf{Z} \in \mathbb{R}^k$ is defined by $\mathbf{Z} \stackrel{d}{=} (\mathbf{V} \mid \mathbf{U} \in C)$, where $\mathbf{U} \in \mathbb{R}^l$ and $\mathbf{V} \in \mathbb{R}^k$ are correlated vectors and $C \subset \mathbb{R}^l$ is a proper selection set, we have that the probability density function $p_{\mathbf{Z}}$ of $\mathbf{Z}$ having a selection distribution (SLCT) is (provided that $\mathbf{V}$ has a density $p_{\mathbf{V}}$):

$$p_{\mathbf{Z}}(\mathbf{z}) = p_{\mathbf{V}}(\mathbf{z}) \frac{P(\mathbf{U} \in C \mid \mathbf{V} = \mathbf{z})}{P(\mathbf{U} \in C)}.$$

Therefore, we have from (2) that the entropy of $\mathbf{Z}$ is

$$H_{\mathbf{Z}}^{\mathrm{SLCT_k}} = H_{\mathbf{V}} - E[\log\{P(\mathbf{U} \in C \mid \mathbf{V})\} - \log\{P(\mathbf{U} \in C)\}].$$

The last term in the above selection entropy is justly the contribution of the selection mechanism. For selection skew-elliptical distributions, for example, we have $\mathbf{V} \sim \mathrm{EC}_k(\mathbf{x}_{\mathbf{V}}, \mathbf{\Omega}_{\mathbf{VV}}, h^{(k)})$ and $\mathbf{U} \mid \mathbf{V} \sim \mathrm{EC}_l(\boldsymbol{\xi}_{\mathbf{U}\cdot\mathbf{V}}, \mathbf{\Omega}_{\mathbf{UU}\cdot\mathbf{V}}, h_{S_{\mathbf{V}}}^{(l)})$, where $\boldsymbol{\xi}_{\mathbf{U}\cdot\mathbf{V}} = \boldsymbol{\xi}_{\mathbf{U}} + \mathbf{\Omega}_{\mathbf{UV}}\mathbf{\Omega}_{\mathbf{VV}}^{-1}(\mathbf{V} - \boldsymbol{\xi}_{\mathbf{V}})$, $\mathbf{\Omega}_{\mathbf{UU}\cdot\mathbf{V}} = \mathbf{\Omega}_{\mathbf{UU}} - \mathbf{\Omega}_{\mathbf{UV}}\mathbf{\Omega}_{\mathbf{VV}}^{-1}\mathbf{\Omega}_{\mathbf{VU}}$ and $S_{\mathbf{V}} = (\mathbf{V} - \boldsymbol{\xi}_{\mathbf{V}})^T \mathbf{\Omega}_{\mathbf{VV}}^{-1}(\mathbf{V} - \boldsymbol{\xi}_{\mathbf{V}})$. Hence, $H_{\mathbf{V}} = H_{\mathbf{V}}^{\mathrm{EC}_k} = (1/2)\log|\mathbf{\Omega}_{\mathbf{VV}}| - E\{\log h^{(k)}(S_{\mathbf{V}})\}$, while the computation of the contribution of the selection mechanism requires the specification of the selection set C. In our case, we have $l = 1$, $\boldsymbol{\xi}_{\mathbf{U}} = 0$, $\mathbf{\Omega}_{\mathbf{UU}} = 1$, $\mathbf{\Omega}_{\mathbf{VU}} = \boldsymbol{\delta}$ and $C = (0, \infty)$, so that the probability density function of the selection random vector $\mathbf{Z}$ reduces to (8), with $\boldsymbol{\xi} = \mathbf{x}_{\mathbf{V}}$, $\mathbf{\Omega} = \mathbf{\Omega}_{\mathbf{VV}}$ and $\boldsymbol{\eta} = \mathbf{\Omega}^{-1}\boldsymbol{\delta}/\sqrt{1 - \boldsymbol{\delta}^T\mathbf{\Omega}^{-1}\boldsymbol{\delta}}$, and therefore $H_{\mathbf{Z}}^{\mathrm{SLCT}_k}$ becomes $H_{\mathbf{Z}}^{SE_k}$ as in proposition 1.

Acknowledgements

Arellano-Valle's research was partially supported by grant FONDECYT 1085241-Chile. Contreras-Reyes's research was partially supported by a grant from the Inter-American Institute for Global Change Research (IAI) CRN II 2017, which is supported by the US National Science Foundation (Grant GEO-0452325). Genton's research was partially supported by NSF grant DMS-1007504. This publication is based in part on work supported by Award No. KUS-C1-016-04 made by King Abdullah University of Science and Technology (KAUST). The authors thank the editor, an associate editor, a referee and Zdenek Hlávka for their helpful comments and suggestions.

References

Ahmed, N. A. & Gokhale, D. V. (1989). Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. Inform. Theory* **35**, 688–692.

Ainslie, B., Reuten, C., Steyn, D. G., Le, N. D. & Zidek, J. V. (2009). Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants. *J. Environ. Manage.* **90**, 2715–2729.

Arellano-Valle, R. B. (2010). On the information matrix of the multivariate skew-*t* model. *Metron* **68**, 371–386.

Arellano-Valle, R. B. & Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scand. J. Statist.* **33**, 561–574.

Arellano-Valle, R. B. & Bolfarine, H. (1995). On some characterizations of the t-distribution. *Statist. Probab. Lett.* **25**, 179–185.

Arellano-Valle, R. B. & Genton, M. G. (2005). On fundamental skew distributions. *J. Multivariate Anal.* **96**, 93–116.

Arellano-Valle, R. B. & Genton, M. G. (2010a). Multivariate extended skew-*t* distributions and related families. *Metron* **68**, 201–234.

Arellano-Valle, R. B. & Genton, M. G. (2010b). Multivariate unified skew-elliptical distributions. Special issue 'Tribute to Pilar Loreto Iglesias Zuazola'. *Chil. J. Stat.* **1**, 17–33.

Arellano-Valle, R. B. & Genton, M. G. (2010c). An invariance property of quadratic forms in random vectors with a selection distribution, with application to sample variogram and covariogram estimators. *Ann. Inst. Statist. Math.* **62**, 363–381.

Arellano-Valle, R. B., Branco, M. D. & Genton, M. G. (2006a). A unified view on selection distributions. *Canad. J. Statist.* **33**, 561–574.

Arellano-Valle, R. B., del Pino, G. & Iglesias, P. (2006b). Bayesian inference in spherical linear models: robustness and conjugate analysis. *J. Multivariate Anal.* **97**, 179–197.

Azzalini, A. (2008). *R package sn: the skew-normal and skew-t distributions (version 0.4-6)*. Universitá di Padova, Italia.

Azzalini, A. & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distributions. *J. Roy. Statist. Soc. Ser. B* **61**, 579–602.

Azzalini, A. & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. Roy. Statist. Soc. Ser. B* **65**, 367–389.

Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.

Branco, M. D. & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *J. Multivariate Anal.* **79**, 99–113.

Cover, T. M. & Thomas, J. A. (2006). *Elements of information theory*. Second edn. Wiley, New Jersey.

Fang, K.-T., Kotz, S. & Ng, K.-W. (1990). *Symmetric multivariate and related distributions*. Monographs on Statistics and Applied Probability 36. Chapman and Hall, Ltd., London.

Genton, M. G. (2004). *Skew-elliptical distributions and their applications: a journey beyond normality*. Edited Volume, Chapman & Hall/CRC, Boca Raton, FL, 416 pp.

Gupta, A. K. (2003). Multivariate skew-*t* distribution. *Statistics* **37**, 359–363.

Javier, W. R. & Gupta, A. K. (2008). Mutual information for the mixture of two multivariate normal distributions. *Far East J. Theor. Stat.* **26**, 47–58.

Javier, W. R. & Gupta, A. K. (2009). Mutual information for certain multivariate distributions. *Far East J. Theor. Stat.* **29**, 39–51.

Kullback, S. (1978). *Information theory and statistics*. Dover Edition, Gloucester.

Lee, S., Genton, M. G. & Arellano-Valle, R. B. (2010). Perturbation of numerical confidential data via skew-*t* distributions. *Manag. Sci.* **56**, 318–333.

Ma, Y. & Genton, M. G. (2004). A flexible class of skew-symmetric distributions. *Scand. J. Statist.* **31**, 459–468.

Mardia, K. V. (1985). Mardia's test of multinormality. In *Encyclopedia of statistical sciences* (eds N. Johnson & S. Kotz), 217–221. Wiley, New York.

Misra, N., Singh, H. & Demchuk, E. (2005). Estimation of the entropy of a multivariate normal distribution. *J. Multivariate Anal.* **92**, 324–342.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available on http://www.R-project.org.

Seremi de Salud (2006). Indices de Calidad del Aire, Santiago de Chile. Available on http://www.seremisaludrm.cl/sitio/pag/aire/indexjs3aireindgasesdemo-prueba.asp.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423, 623–656.

Silva, C. & Quiroz, A. (2003). Optimization of the atmospheric pollution monitoring network at Santiago de Chile. *Atmos. Environ.* **37**, 2337–2345.

Received February 2011, in final form October 2011

Marc G. Genton, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.
E-mail: genton@stat.tamu.edu

Appendix: Proof of proposition 2

Let $\mathbf{Z}_c = \mathbf{\Gamma}\mathbf{Z}_0$ and $W = \boldsymbol{\gamma}^T\mathbf{Z}_c$, where $\mathbf{\Gamma} \in \mathbb{R}^{k\times k}$ is an orthogonal matrix such that $\mathbf{\Gamma}\bar{\boldsymbol{\eta}} = \|\bar{\boldsymbol{\eta}}\|\boldsymbol{\gamma}$ and the vector $\boldsymbol{\gamma} \in \mathbb{R}^k$ is such that $\boldsymbol{\gamma}^T\bar{\boldsymbol{\eta}} = \|\bar{\boldsymbol{\eta}}\|$ and $\|\boldsymbol{\gamma}\| = 1$. Note that $W = \boldsymbol{\gamma}^T\mathbf{Z}_c = \boldsymbol{\gamma}^T\mathbf{Z}_0$ and $S_c = \|\mathbf{Z}_c\|^2 = \|\mathbf{Z}_0\|^2 = S$ since $\mathbf{\Gamma}\mathbf{\Gamma}^T = \mathbf{I}_k$. Thus, considering also that $\mathbf{Z}_0 = \mathbf{\Gamma}^T\mathbf{Z}_c$ and the absolute value of the determinant of $\mathbf{\Gamma}^{\mathrm{T}}$ equals 1, the Jacobian method yields $f_{\mathbf{Z}_c}(\mathbf{z}) = 2h^{(k)}(s) \times F(\|\bar{\boldsymbol{\eta}}\|w, h_s^{(1)})$, where $s = \|\mathbf{z}^T\mathbf{z}\|^2$ and $w = \boldsymbol{\gamma}^T\mathbf{z}$, that is, $\mathbf{Z}_c \sim \mathrm{SE}_k(\mathbf{0}, \mathbf{I}_k, \|\bar{\boldsymbol{\eta}}\|\boldsymbol{\gamma}, h^{(k+1)})$, and by proposition 4.1 in Arellano-Valle & Genton (2010a), we have $W = \boldsymbol{\gamma}^T\mathbf{Z}_c \sim \mathrm{SE}_1(0, 1, \|\bar{\boldsymbol{\eta}}\|, h^{(k+1)})$. On the other hand, since (see, e.g., Arellano-Valle & Azzalini, 2006; Arellano-Valle *et al.*, 2006a)

$$\mathbf{Z}_c \overset{d}{=} \bar{\boldsymbol{\delta}}|X_0| + (\mathbf{I}_k - \bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}^T)^{1/2}\mathbf{X},$$

where

$$\bar{\boldsymbol{\delta}} = \frac{\bar{\boldsymbol{\eta}}}{\sqrt{1+\|\bar{\boldsymbol{\eta}}\|^2}} \text{ and } \begin{pmatrix} \mathbf{X} \\ X_0 \end{pmatrix} \sim \mathrm{EC}_{k+1}\left(\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix}, h^{(k+1)}\right),$$

we have $W \overset{d}{=} \|\bar{\boldsymbol{\delta}}\||X_0| + \sqrt{1-\|\bar{\boldsymbol{\delta}}\|^2}X_1$, where $X_1 = \boldsymbol{\gamma}^T\mathbf{X} \sim \mathrm{EC}_1(0, 1, h^{(1)})$ is distributed as the first component of $\mathbf{X}$. Thus, since we can assume without loss of generality that $\boldsymbol{\gamma}$ is the first column of $\mathbf{\Gamma}$, we find

$$\mathbf{Z}_c \overset{d}{=} \boldsymbol{\gamma}W + (\mathbf{I}_k - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{1/2}\mathbf{X} = (W, X_2, \ldots, X_k)^T,$$

where $X_2, \ldots, X_k$ are the last $k-1$ components of $\mathbf{X}$. Consider now the transformation

$$U_1 = W, \quad U_j = X_j, \quad 2 \le j \le k-1, \text{ and } R = \sqrt{W^2 + \sum_{j=2}^{k} X_j^2}.$$

This transformation has two inverses given by

$$w = u_1, \quad x_j = u_j, \quad 2 \le j \le k-1, \text{ and } x_k = \pm\sqrt{r^2 - \sum_{j=1}^{k-1} u_j^2}.$$

The corresponding Jacobians are

$$J_1 = \frac{r}{\sqrt{r^2 - \sum_{j=1}^{k-1} u_j^2}},$$
$$J_2 = -J_1.$$

Thus, we obtain for $k \ge 2$ that

$$f_{U_1, U_2, \ldots, U_{k-1}, R}(u_1, u_2, \ldots, u_{k-1}, r) = \frac{4rh^{(k)}(r^2)F^{(1)}(\|\bar{\boldsymbol{\eta}}\|u_1; h_{r^2}^{(1)})}{\sqrt{(r^2 - u_1^2)(r^2 - \sum_{j=1}^{k-1} u_j^2)}},$$

where $\sum_{j=2}^{k-1} u_j^2 < r^2 - u_1^2$, $|u_1| < r$ and $r > 0$. Considering now the change of variables

$$w_j = \frac{u_j}{\sqrt{r^2 - u_1^2}} = \frac{u_j}{\sum_{j=2}^{k} u_j^2}, \quad 2 \le j \le k-1, \quad k \ge 2,$$

we have

$$
\begin{aligned}
f_{U_1,R}(u_1,r) &= 4r^{k-2}\left(1-\frac{u_1^2}{r^2}\right)^{\left(\frac{k-1}{2}-1\right)} h^{(k)}(r^2)F(\|\bar{\boldsymbol{\eta}}\|u_1;h_{r^2}^{(1)}) \\
&\quad \times \int_{\{w_2,\ldots,w_{k-1}:0<\sum_{j=2}^{k-1}w_j^2<1\}} \left(1-\sum_{j=2}^{k-1}w_j^2\right)^{-1/2} \mathrm{d}w_2\cdots\mathrm{d}w_{k-1} \\
&= \frac{4\pi^{k/2-1}}{\Gamma(\frac{k}{2})} r^{k-2}\left(1-\frac{u_1^2}{r^2}\right)^{\left(\frac{k-1}{2}-1\right)} h^{(k)}(r^2)F(\|\bar{\boldsymbol{\eta}}\|u_1;h_{r^2}^{(1)}).
\end{aligned}
$$

Thus, the change of variables $(W,S)=(U_1,R^2)$ implies the result.