

Generalized skew-normal negentropy and its applications



Javier Contreras Reyes

Supervisor: Prof. Milan Stehlík

Advisor: Prof. Reinaldo B. Arellano Valle

Instituto de Estadística
Universidad de Valparaíso

This dissertation is submitted for the degree of
Doctor en Estadística

Executive Summary

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. Information theory seeks the quantification of information and the development of coding schemes that provide good performance in comparison with the optimal performance given by the theory. The *Shannon entropy* is one of the most used measures to quantify the information of a random variable and is attributed to uncertainty of information or mathematical contrariety of information. In order to quantify the mutual information between two random variables, the most used quantity is the mutual information index that provides a generalized measure of association between both random variables, say X and Y , which is particularly convenient in those models where the correlation is not defined. Additionally, the Rényi entropy (Rényi, 1970) corresponds to a generalization of Shannon entropy that provides some additional mathematical advantages, such as include an α -order for compare two Rényi entropies. Kullback and Leibler (1951) proposed the Kullback–Leibler (KL) divergence to compare the information of Y with respect to X , or the disparity of the Y 's probability density function (pdf) with respect to X 's pdf. KL divergence is strictly positive and non-symmetrical, i.e., the order of X and Y affects the information. Also KL divergence is considered a good indicator of the correlation degree between two finite sets of data, specially if they are affected by noise produced by the interaction of the system. Other divergences that provides additional mathematical properties are: Jeffreys divergence (symmetrical with respect to X and Y) and Jensen–Shannon distance (symmetrical and accomplish the triangular inequality). Focusing in normal distribution, the *Negentropy* becomes the KL divergence and measures the departure from the normality of the X 's pdf. The *Negentropy* is always nonnegative, and will become even larger as the random variable and is farther from the normality.

Frequently, several authors have been computed information measures for a large list of univariate and multivariate distributions, beyond the multivariate normal distribution (Zografos and Nadarajah, 2005). In the last three decades, non-normal distributions have received substantial interest within scientific literature, especially in regards to skew-elliptical distributions, a class of flexible distributions characterized for account with skewness and heavy-tails as an extra parameters respect to elliptical distributions as normal and t (Branco and Dey, 2001), behind the multivariate skew-normal (SN, Azzalini and Dalla-Valle, 1996) and multivariate skew- t (ST, Azzalini and Capitanio, 2003) distributions. The family of SN distributions has been popularized by Azzalini (1985) and ever since it has been discussed extensively in the literature. Such discussions include a wide variety of skewed models

in addition to having gaussian distribution as a special case and flexibility in capturing skewness in the data. In this sense, González-Farías et al. (2004) present the closed SN (CSN) distribution as an extension of the SN case, but closed under operations such as sums, marginalization, and linear conditioning. Another generalization of the SN distribution is the extended SN distribution (ESN, Capitanio et al., 2003) that adds a fourth real parameter to accommodate both skewness and heavy tails. In some cases where observed variables can be simultaneously skewed and restricted to a fixed interval, the truncated SN distribution is a good choice for those applications (Flecher et al., 2010).

In this work, we propose a general and unified theory of the information theory measures mentioned early for flexible and tractable families of continuous multivariate distributions, in which the multivariate normal and further well-known symmetric distributions, such as the t , are particular members. Specifically, we consider the multivariate elliptical, skew-elliptical, closed SN and generalized SN families of distributions. We give special attention to the particular cases of the multivariate SN and ST distributions that allow to model skewness. In particular, we provide the following results.

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 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 SN and ST mutual information indexes. As an application, an analysis of an optimal network design of a classical pollutant was presented. 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. The correct fit of the original data ensures the optimal maximization of the mutual information index and determines a better optimization network design.

In addition, we have presented a methodology to compute the KL divergence for multivariate data presenting skewness, specifically, for data following a multivariate SN distribution. The calculation of this measure is semi-analytical, since it is the sum of two analytical terms, one corresponding to the multivariate normal KL 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 SN distributions, and in particular for the J divergence between the SN and normal distributions. The proposed SN KL divergence is

applied to aftershocks produced by the Maule earthquake which occurred on February 27 of 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 founded results in a consistent form. However, this asymptotic homogeneity test do not satisfy regularity condition when shape parameter is close to zero, i.e., where the Fisher information matrix is singular. This problem is solved by considering a modified version of SN distribution, as is describer later. Besides, a measure to compare two multivariate ST densities is presented based on the approximated KL divergence. This has some advantages such as: the detection of skewness presence and heavy-tails within the data; it is of rapid and easy computational implementation, because it is an explicit form of the divergence where there is an expected value representing an integral in one dimension.

Some solutions to compute the Rényi entropy with discrete α -order and for a wide range of asymmetric distributions are presented. Specifically, we find a closed expression for SN, ESN, and TSN distributions. Additional inequalities for SN and ESN entropies were reported. Lower and upper bounds of the Shannon and Rényi entropies for finite mixtures of SN (FMSN) distributions were derived. Using such a pair of bounds some kind of confidence interval for the approximate entropy value can be calculated, where the average between these values can be used as an approximation of the entropy. We presented practical (bounds) and theoretical (bounds and asymptotic expression) results for Rényi entropy. In the case of practical results, the first upper bound deals only with the density parameters and the second one with the density and mixing weights parameters. In the case of theoretical results, the bounds and approximations are based on L^p space metric and multinomial coefficients. The results presented are valid for the SN case, taking the shape parameters set equal zero, for integer values of α (Contreras-Reyes, 2015). In addition, the proportioned results are also valid for other continuous densities where the Rényi entropies of the component exist. We hope the Rényi entropy developments in finite mixtures of densities can stimulate more research in the future, for more flexible densities such as ST distribution (Azzalini and Capitanio, 2003). As an application of SN Rényi entropy, we applied 2-dimensional length-weight data for the determination of swordfish age. We considered a length-weight dataset instead of the usual length (considered by Roa-Ureta, 2010) to determine the number of clusters, and posteriorly we compared it with the real observations obtained by the procedure of Cerna (2009). The best results were obtained using the Rényi entropy, as an average between upper and lower bounds, over Shannon entropy and information criteria.

Finally, we consider asymptotic expansions of moments and cumulants for the negentropy of two particular cases: the SN and Modified SN (MSN) distributions. Given that SN

distribution do not accomplished regularity condition of Fisher information matrix at when skewness parameter is zero, normality is tested based on the MSN distribution (Arrué et al., 2016). This allow to implement an asymptotic normality test for testing significance of skewness parameter. We have presented the methodology to compute the Shannon entropy, the negentropy, and the KL and J divergences for a broad family of asymmetric distributions with normal kernel called Generalized SN distributions. Our method considers asymptotic expansions regarding moments and cumulants for two particular cases: the SN and MSN distributions. We then measured the degrees of disparity of these distributions from the normal distribution by using exact expressions for the negentropy in terms of moments and cumulants. Numerical results showed that the Shannon entropy and negentropy of the MSN distribution is better approximated than SN one, at least for a wider range of the shape parameter. For small of the asymmetry parameter, where the approximations are appropriate, we find that expansions series converge from the fourth moment/cummulant to greater, as in Gram–Charlier and Edgeworth expansion methods. For large values of the skewness parameter, where the expansions are inappropriate, the functions related to negentropy are not well approximated by Taylor expansions around zero, produced by a divergence in the moment and cummulant terms. When this happens, the normal cumulative density function tends to 1, since according to the stochastic representation for large values of skewness parameter, the distribution converges to the standardized half-normal distribution.

We also compared asymptotic test with Likelihood Ratio test and asymptotic normality test given by Arrué et al. (2016). Given the regularity conditions accomplished by the MSN distribution, normality was tested based on the modified SN distribution. This test considered the asymptotic behavior of the KL divergence, which is determined by the negentropy for normality disparity. However, the normality test considered in the application used skewness parameters inside the appropriate range. Proposed asymptotic test is applied to condition factor time series of anchovy. The results show that the proposed methodology serves to detect non-normal events in these time series, which produces an empirical distribution with high well presence of skewness. The proposed test for normality is therefore useful to detect anomalies in condition factor time series, linked to food deficit (positive shape parameter) or food abundance (negative shape parameter) influenced by environmental conditions.

I dedicate this dissertation to my family and friends who have supported me throughout the process.

Special thanks to my parents for their endless patience, support and encouragement.

“It is not true that people stop pursuing dreams because they grow old, they grow old because they stop pursuing dreams”

Gabriel García Márquez

Acknowledgements

I would like to express my sincere gratitude to my advisor, Dr. Milan Stehlík. This research would not have been possible without his talented insightful guidance and continuing motivation. I am thankful to my second advisor, Dr. Reinaldo Arellano, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject.

I am also grateful to my friends and partners who have supported me along the way. Specially to Alessandra, Daniel, Ivonne, Monica, Pablo and Sebastian.

Abstract

Spanish Version

El problema de medir la disparidad de una función de densidad de probabilidad con respecto a una densidad normal ha sido estudiado en varios estudios recientes. La técnica más usada para lidiar con este problema ha sido mediante expresiones exactas usando medidas de información de ciertas distribuciones. En esta tesis consideramos una clase de distribuciones asimétricas con un kernel normal, llamada distribuciones normal-sesgadas generalizadas (GSN). Medimos el grado de disparidad de esas distribuciones con respecto a la distribución normal usando expresiones exactas de la negentropía GSN en términos de cumulantes. Específicamente, nos focalizamos en las distribuciones normal-sesgada y normal-sesgada modificada. Luego, establecemos las divergencias de Kullback–Leibler entre cada distribución GSN y la normal en términos de sus negentropías para desarrollar un test de hipótesis de normalidad. Finalmente, aplicamos este resultado a series de tiempo de factor de condición de anchovetas de la costa del norte de Chile.

English Version

The problem of measuring the disparity of a particular probability density function from a normal one has been addressed in several recent studies. The most used technique to deal with the problem has been exact expressions using information measures over particular distributions. In this thesis, we consider a class of asymmetric distributions with a normal kernel, called Generalized Skew-Normal (GSN) distributions. We measure the degrees of disparity of these distributions from the normal distribution by using exact expressions for the GSN negentropy in terms of cumulants. Specifically, we focus on skew-normal and modified skew-normal distributions. Then, we establish the Kullback–Leibler divergences between each GSN distribution and the normal one in terms of their negentropies to develop hypothesis testing for normality. Finally, we apply this result to condition factor time series of anchovies off northern Chile.

Table of contents

List of figures	xvii
List of tables	xxi
1 Background of Information theory	1
1.1 Shannon entropy, mutual information and negentropy	1
1.2 Rényi entropy and complexity measure	3
1.3 Cross-entropy and related divergences	4
2 Asymmetric and heavy-tailed class of distributions	7
2.1 Multivariate elliptical distributions	8
2.1.1 Multivariate normal distribution	9
2.1.2 Multivariate t distribution	9
2.1.3 Scale mixture of normal	10
2.2 Multivariate skew-elliptical distributions	10
2.2.1 Multivariate skew-normal distribution	11
2.2.2 Finite mixtures of multivariate skew-normal distributions	12
2.2.3 Multivariate skew- t distribution	13
2.2.4 Scale mixture of skew-normal	14
2.3 Multivariate closed skew-normal distributions	14
2.3.1 Multivariate skew-normal distribution	15
2.3.2 Multivariate extended skew-normal distribution	15
2.4 Generalized skew-normal distribution	16
2.4.1 Skew-normal distribution	17
2.4.2 Truncated skew-normal distribution	18
2.4.3 Modified skew-normal distribution	19
3 Information measures for symmetric and asymmetric distributions	21
3.1 Information measures for multivariate elliptical distributions	21

3.1.1	Multivariate normal distribution	22
3.1.2	Multivariate t distribution	23
3.2	Information measures for multivariate skew-elliptical distributions	24
3.2.1	Multivariate skew-normal distribution	25
3.2.2	Finite mixture of multivariate skew-normal distributions	34
3.2.3	Multivariate skew- t distribution	40
3.3	Information measures for multivariate closed skew-normal distributions	43
3.3.1	Multivariate skew-normal distribution	44
3.3.2	Multivariate extended skew-normal distribution	44
3.4	Information measures for generalized skew-normal distributions	45
3.4.1	Skew-normal distribution	48
3.4.2	Modified skew-normal distribution	50
3.5	J Divergence between generalized skew-normal distributions	51
3.6	Asymptotic tests for generalized skew-normal distributions	52
4	Simulations	55
4.1	Skew-normal and skew- t distributions	55
4.1.1	Shannon and Rényi entropies	55
4.1.2	Maximum entropy, Jeffrey's and Kullback-Leibler divergences	56
4.1.3	Jensen-Shannon distance	57
4.2	Finite mixture of multivariate skew-normal distributions	58
4.3	Skew-normal and modified skew-normal distributions	60
5	Applications	77
5.1	Santiago's monitoring network design	77
5.1.1	Data treatment	77
5.1.2	Main results	79
5.2	Chile's seismological catalogue	81
5.2.1	Asymptotic test	81
5.2.2	Main results	82
5.3	Swordfish age-length-weight data	83
5.3.1	Data and software	84
5.3.2	Length-weight relationship	85
5.3.3	Clustering and model selection	85
5.4	Fish condition factor time series	87
5.4.1	Condition factor time series modelling	88
5.4.2	Skew-normal process	89

5.4.3	Skew-normal Jeffrey's divergence and Jensen–Shannon distance . .	90
5.4.4	Normality test	91
5.5	Other applications	92
6	Concluding remarks	117
6.1	Skew-elliptical distributions	117
6.2	Finite mixture of skew-normal distributions	119
6.3	Generalized skew-normal distributions	120
	References	123
	Appendix A	133
	Appendix B	143

List of figures

2.1	Relationships among the skew-elliptical distributions: the skew- t (ST) contains the t (T), skew-normal (SN) and normal (N) distributions as special cases.	11
4.1	Plots of the ST entropy for $\nu = 1, 2, \dots, 185$ degrees of freedom as a function of shape parameter α . The red and blue lines correspond to normal and skew-normal entropies, respectively.	64
4.2	(a) Behavior of $h(\tau) = E[\log\{2\Phi(\tau W)\}]$ for $\tau = 0, 1, \dots, 200$. (b) Behavior of $J(X, Y_0)$ for $\tau = 0, 1, \dots, 10$	67
4.3	3D-plots of exact univariate SN Rényi entropy (Proposition 5), $R_\alpha(Z)$, $Z \sim SN_1(0, 1, \eta)$, with $\alpha = 2, \dots, 15$ and $\eta = 0, \dots, 10$. Panels (a) and (b) correspond to different angles of 3D-plots.	68
4.4	Plots of KL-divergence between $\mathbf{X} \sim SN_1(\xi_1, \omega_1^2, \eta_1)$ and $\mathbf{Y} \sim SN_1(\xi_2, \omega_2^2, \eta_2)$	69
4.5	Plot of exact and asymptotic KL divergence between $X \sim ST_1(0, 1.5, 2, 5)$ and $Y \sim ST_1(0, 2.5, 3, \nu_2)$, with $\nu_2 = 3, \dots, 150$; and its respective absolute errors.	70
4.6	3D-plots of asymptotic KL divergence between (a) $X \sim ST_1(0, 1.5, 2, \nu_1)$ and $Y \sim ST_1(0, 2.5, 3, \nu_2)$, and (b) $X \sim T_1(0, 1.5, \nu_1)$ and $Y \sim T_1(0, 2.5, \nu_2)$	71
4.7	JS distance for shape parameters $\eta_1 \times \eta_2 = [-7, 7] \times [-7, 7]$	72
4.8	Comparisons of Lin's (blue lines) and Crook's (red lines) upper bounds with JS distance for skew-gaussian distributions and shape parameters $\eta_1 = \{0.1, 0.2, \dots, 7\}$ and (a) $\eta_2 = 0.1$, (b) $\eta_2 = 0.5$, (c) $\eta_2 = 0.8$, (d) $\eta_2 = 1.2$, (e) $\eta_2 = 3$, and (f) $\eta_2 = 7$	73
4.9	Examples (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, and (f) 6.	74
4.10	Shannon entropy and negentropy for (A)-(B) SN and (C)-(D) MSN cases. The blue and red lines correspond to numerical integration and cumulant expansion series methods, respectively.	75

4.11	KL divergence, $K_0 = K(Z_\tau, Z_0)$, between SN and normal (solid line) and MSN and normal (dotted line).	76
5.1	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 01/03/2006 to 07/03/2006 of station L. Right: Several histograms for the transformed ozone data mentioned before.	105
5.2	Multivariate normal, SN, and ST 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) $s = 40$	106
5.3	Left: Graphic of log-likelihoods of multivariate SN, ST and normal fits. Right: Scatter plots between s , log-likelihood of multivariate ST fit, and degrees of freedom (DF) parameters.	107
5.4	Graphics of a) SN and b) ST mutual information, $I(\mathbf{X}, Y)$, when the $Y = F, L, M, N, O, P$ or Q station is removed from the network \mathbf{X} . The vertical dotted gray lines correspond to $s = \{8, 16, 24, 32\}$	107
5.5	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).	108
5.6	Plots of SN fits (in red) and QQ-plots of Normal and SN distributions for clusters black (1), red (2), green (3) and blue (4).	109
5.7	Plots of SN fits (in red) and QQ-plots of Normal and SN distributions for clusters violet (5), yellow (6), gray (7) and all observations.	110
5.8	Scheme of steps (a)–(d) for swordfish data.	111
5.9	Length-weight log-transformed relationship and regression fit (red solid line) for (a) male and (b) female swordfish.	111
5.10	Logarithmic of the average between upper and lower bounds for Shannon and Rényi entropies, for (a, c) male and (b, d) female swordfish, respectively.	112
5.11	Selected FMSN fits for (a) male and (b) female swordfish. Each color corresponds to each mixture component from a total of $m = 7$	113
5.12	Length-weight log-transformed relationship and linear regression fit (red solid line) for (a) male and (b) female anchovies.	113
5.13	(a) Stochastic representations of Z for $ \eta = \{0, 3, 10, 30\}$, and its respective estimated (b) Hurst exponents and (c) fractal dimensions.	114
5.14	Condition factor time series and histograms with SN fits (solid line) of anchovies by length classes $L = 12, \dots, 18$, All for (a) male and (b) female anchovies.	115

5.15	SN J divergence for (a) males and (b) females, and JS distance for (c) males and (d) females. This measures consider the shape parameter estimates of Table 5.10 for each group j and length L	116
------	--	-----

List of tables

4.1	Numerical and exact (Proposition 5) method for univariate SN Rényi entropy, $R_\alpha(Z)$, $Z \sim SN_1(0, 1, \eta)$, with $\alpha = 2, \dots, 6$ and $\eta = 0.2, \dots, 1, 1.5, 2, \dots, 6$.	63
4.2	Simulated Shannon and Rényi entropy bounds. Rényi entropy bounds are computed for $\alpha = 2, \dots, 5$. For each model and number of clusters m the AIC and BIC criteria, misclassification (MC), normal skill (NSS), Heidke skill (HSS), and Hanssen–Kuipers (HK) scores appear.	65
4.3	Observed power (both in %) of the proposed normality test using MLE of MSN model from 2000 simulations for nominal level 5%, and various values of shape parameter η and sample size n .	66
5.1	P-values for Mardia’s goodness-of-fit test of multivariate normality applied to the joint \mathbf{XY} 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% confidence level lead to multivariate normality used in the last column to compute the mutual information index, $I(\mathbf{X}, Y)$, for this distribution (the first and second largest values are marked in bold).	94
5.2	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% confidence level.	95
5.3	Summary of results for SN and ST distributions and different network configurations. The highest values for each transformation is marked in bold.	96
5.4	KL divergences for pairs of clusters.	97
5.5	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); SN 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.	98

5.6	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 per cent significance level.	99
5.7	Summary of estimates $\alpha' = \log \alpha$ and β with their respective standard errors in parenthesis, for each length-weight log-transformed relationships of Eq. (5.2) and sex.	100
5.8	Summary of FMSN and FMN clustering for swordfish data. For each model and number of clusters m the misclassification (MC), normal skill (NSS), Heidke skill (HSS), and Hanssen–Kuipers (HK) scores appear.	101
5.9	Linear regression fit parameters, $\alpha' = \log \alpha$ and β of Eq. (5.2), for length-weight log-transformed relationships with its respective standard errors in parenthesis, for each length class L and group j	102
5.10	Shape parameter estimates ($\hat{\eta}$) of SN and MSN models for each sex and length class L , together with its respective standard deviations (s.d). Sample size (n), empirical skewness ($\sqrt{\hat{b}_1}$) and kurtosis (\hat{b}_2), and log-likelihood function $\ell(\hat{\eta})$ for each model fit are also reported.	103
5.11	MSN Shannon entropy (H) and negentropy (N) for each sex and length class L using expansion series of cumulants. For each time series, the KL divergence $K_0 = K(Z_{\hat{\tau}}, Z_0)$, statistic $2nK_0$ of Eq. (3.37), Likelihood Ratio Test (LRT) statistic, and its respective p-values are reported. All values reported consider estimates $\hat{\eta}$ (for $\hat{\tau} = \hat{\eta} $) and sample size n from Table 5.10.	104

Chapter 1

Background of Information theory

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 in order to define an information quantity contained in a multivariate probability density function (pdf), for example such as the multivariate normal distribution (Cover and Thomas, 2006, Kullback, 1978, Silva and Quiroz, 2003) and the exponential family (Stehlík, 2003). 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 e.g. Cover and Thomas, 2006).

In this chapter, we review a few important basic aspects of information theory.

1.1 Shannon entropy, mutual information and negentropy

Let $\mathbf{Z} \in \mathbb{R}^k$ be a random vector with pdf $f(\mathbf{z})$. The *Shannon entropy* -also named *differential entropy*-which was proposed earlier by Shannon (1948) is

$$H(\mathbf{Z}) = -E[\log f(\mathbf{Z})] = - \int_{\mathbb{R}^k} f(\mathbf{z}) \log f(\mathbf{z}) d\mathbf{z}, \quad (1.1)$$

where $E[g(\mathbf{Z})]$ denote the *expected information* in \mathbf{Z} for a function $g(\mathbf{z})$. In this case, Shannon's entropy is the expected value of the function $g(\mathbf{z}) = -\log f(\mathbf{z})$, which satisfies $g(\mathbf{1}) = 0$ and $g(\mathbf{0}) = \infty$. We extend this notation in all expected values expressed on this work. The Shannon entropy is attributed to uncertainty of information or mathematical contrariety of information, see Cover and Thomas (2006) for additional properties.

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

$$H(\mathbf{Z}) = \frac{1}{2} \log |\Omega| + H(\mathbf{Z}_0), \quad (1.2)$$

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

Defined according to Gray (1990), we consider the following concept of mutual information index. Let $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^m$ be two random vectors with joint and marginal pdfs $f(\mathbf{x}, \mathbf{y})$, $f(\mathbf{x})$ and $f(\mathbf{y})$, respectively. The *mutual information index* between \mathbf{X} and \mathbf{Y} is defined by

$$I(\mathbf{X}, \mathbf{Y}) = E \left[\log \left\{ \frac{f(\mathbf{X}, \mathbf{Y})}{f(\mathbf{X})f(\mathbf{Y})} \right\} \right] = \int_{\mathbb{R}^m} \int_{\mathbb{R}^n} \log \left\{ \frac{f(\mathbf{x}, \mathbf{y})}{f(\mathbf{x})f(\mathbf{y})} \right\} f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \quad (1.3)$$

From (1.3) and (1.1) it is straightforward to see that the mutual information index $I(\mathbf{X}, \mathbf{Y})$ between \mathbf{X} and \mathbf{Y} can be computed as

$$I(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{XY}), \quad (1.4)$$

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, $I(\mathbf{X}, \mathbf{Y}) = 0$ when the random vectors \mathbf{X} and \mathbf{Y} are independent, otherwise this index is positive (Cover and 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.

The Shannon entropy of a localization-scale random variable $\mathbf{X} = \mu + \Sigma^{1/2}\mathbf{Z}$, does not depend on μ and is such that $H(\mathbf{X}) = (1/2) \log |\Sigma| + H(\mathbf{Z})$ (see e.g. Arellano-Valle et al., 2013, Dembo et al., 1991). The SE could serve to define a measure of disparity from normality, the so-called negentropy (Hyvärinen et al., 2001), which is zero for a normal variable and positive for any distribution. It is defined by

$$N(\mathbf{Z}) = H(\mathbf{Z}'_0) - H(\mathbf{Z}), \quad (1.5)$$

where \mathbf{Z}'_0 being a normal random variable with the same mean and variance as that of \mathbf{Z} . Eq. (1.5) express the negentropy in terms of the standardized version of \mathbf{Z} , say \mathbf{Z}^* , as $N(\mathbf{Z}) = H(\mathbf{Z}_0) - H(\mathbf{Z}^*) = N(\mathbf{Z}^*)$, which \mathbf{Z}^* has zero mean and unit variance. Thus, negentropy measures essentially the amount of information that departs from the normal entropy. Also, clearly the negentropy becomes the KL divergence (see Eq. (1.13) below) between \mathbf{Z}^* and \mathbf{Z}_0 .

1.2 Rényi entropy and complexity measure

The α th-order Rényi entropy (Rényi, 1970) of $\mathbf{Z} \in \mathbb{R}^k$ is defined by

$$R_\alpha(\mathbf{Z}) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^k} [f(\mathbf{z})]^\alpha d\mathbf{z}. \quad (1.6)$$

Golshani and Pasha (2010) provide the following important properties of the Rényi entropy:

1. $R_\alpha(\mathbf{Z})$ can be negative,
2. $R_\alpha(\mathbf{Z})$ is invariant under a location transformation,
3. $R_\alpha(\mathbf{Z})$ is not invariant under a scale transformation, and
4. for any $\alpha_1 < \alpha_2$, we have $R_{\alpha_1}(\mathbf{Z}) \geq R_{\alpha_2}(\mathbf{Z})$, which are equal if and only if \mathbf{Z} is uniformly distributed.

From (1.6), the Shannon entropy is obtained by the limit

$$H(\mathbf{Z}) = \lim_{\alpha \rightarrow 1} R_\alpha(\mathbf{Z}) = - \int_{\mathbb{R}^k} f(\mathbf{z}) \log f(\mathbf{z}) d\mathbf{z} \quad (1.7)$$

by applying l'Hôpital's rule to $R_\alpha(\mathbf{Z})$ with respect to α (Rényi, 1970). In addition, the Rényi entropy corresponds to a generalization of Shannon entropy that provides some additional mathematical advantages, such as include an α -order for compare two Rényi entropies (property 4). Song (2001) recall on Rényi entropy's connection to the log-likelihood for a continuous random variable.

Example 1. (Cover and Thomas, 2006, Dembo et al., 1991). Let $\mathbf{Z} \in \mathbb{R}^k$ be a normal random vector with mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance matrix Σ of dimension $k \times k$ with determinant $|\Sigma| > 0$. Then, the Rényi and Shannon entropies of \mathbf{Z} are given by

$$R_\alpha(\mathbf{Z}) = \frac{1}{2} \log \{ (2\pi)^k |\Sigma| \} + \frac{k \log \alpha}{2(\alpha - 1)}, \quad 1 < \alpha < \infty, \quad (1.8)$$

$$H(\mathbf{Z}) = \frac{1}{2} \log \{ (2\pi e)^k |\Sigma| \}, \quad (1.9)$$

respectively.

Another important concept is the statistical complexity that measures the randomness and structural correlations of a known system (Carpi et al., 2011). López-Ruiz et al. (1995) proposed a measure of statistical complexity (LMC) in order to determine the *disequilibrium* of the system attributed to entropy measure (Anteneodo and Plastino, 1996, Sánchez-Moreno et al., 2014). LMC measure is defined as

$$LMC(\mathbf{Z}) = e^{H(\mathbf{Z}) - R_2(\mathbf{Z})}, \quad (1.10)$$

where $R_2(\mathbf{Z})$ is the quadratic Rényi entropy of \mathbf{Z} ($\alpha = 2$). Yamano (2004) provide an extensive entropy instead of an additive Shannon entropy in (1.10), characterized as a difference between the α th-order Rényi entropy and quadratic Rényi entropy as

$$C_\alpha(\mathbf{Z}) = e^{R_\alpha(\mathbf{Z}) - R_2(\mathbf{Z})}. \quad (1.11)$$

Note that $C_\alpha(\mathbf{Z})$ reflects the shape of the distribution of \mathbf{Z} and takes unity for all distributions when $\alpha = 2$. In addition, C_α satisfies a great variety of interesting mathematical and physical properties. Let us just recall here the following properties:

1. $C_\alpha(\mathbf{Z}) > 1$, $\forall \alpha \leq 2$, and, $0 < C_\alpha(\mathbf{Z}) \leq 1$, $\forall \alpha > 2$;
2. $C_\alpha(\mathbf{Z})$ is invariant under a location and scale transformation in the distribution of \mathbf{Z} ;
3. $C_\alpha(\mathbf{Z})$ is invariant under replications of the original distribution of \mathbf{Z} .

1.3 Cross-entropy and related divergences

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* (CE)-also called *relative entropy*- associated to Shannon entropy (1.1) is related to compare the information measure of \mathbf{Y} with respect to \mathbf{X} , and is defined as follows

$$CH(\mathbf{X}, \mathbf{Y}) = -E[\log f_{\mathbf{Y}}(\mathbf{X})] = - \int_{\mathbb{R}^k} f_{\mathbf{X}}(\mathbf{x}) \log f_{\mathbf{Y}}(\mathbf{x}) d\mathbf{x}. \quad (1.12)$$

It is clear from (1.12) that $CH(\mathbf{X}, \mathbf{X}) = H(\mathbf{X})$. However, $CH(\mathbf{X}, \mathbf{Y}) \neq CH(\mathbf{Y}, \mathbf{X})$ at least that $\mathbf{X} \stackrel{d}{=} \mathbf{Y}$, i.e., \mathbf{X} and \mathbf{Y} have the same distribution.

Related to the entropy and CE 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 Kullback and Leibler (1951) as

$$K(\mathbf{X}, \mathbf{Y}) = E \left[\log \left\{ \frac{f_{\mathbf{X}}(\mathbf{X})}{f_{\mathbf{Y}}(\mathbf{X})} \right\} \right] = \int_{\mathbb{R}^k} f_{\mathbf{X}}(\mathbf{x}) \log \left\{ \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{x})} \right\} d\mathbf{x}. \quad (1.13)$$

which measures the divergence of $f_{\mathbf{Y}}$ from $f_{\mathbf{X}}$. KL is a pseudo distance (or discriminant function) between two distributions and which is the most common divergence measures used in practical works. Also KL divergence is considered a good indicator of the correlation degree between two finite sets of data, specially if they are affected by noise produced by the interaction of the system, or the measurement error (Tumminello et al., 2007).

We note that (1.13) comes from (1.12) as

$$K(\mathbf{X}, \mathbf{Y}) = CH(\mathbf{X}, \mathbf{Y}) - H(\mathbf{X}). \quad (1.14)$$

Thus, we have $K(\mathbf{X}, \mathbf{X}) = 0$, but again $K(\mathbf{X}, \mathbf{Y}) \neq K(\mathbf{Y}, \mathbf{X})$ at least that $\mathbf{X} \stackrel{d}{=} \mathbf{Y}$, 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 Ullah, 1996). Hence it must be interpreted as a pseudo-distance measure only.

A familiar symmetric variant of the KL-divergence is the Jeffreys divergence or simply J-divergence (e.g Jeffreys, 1946), which is defined by

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

It could be expressed in terms of the KL-divergence as

$$J(\mathbf{X}, \mathbf{Y}) = K(\mathbf{X}, \mathbf{Y}) + K(\mathbf{Y}, \mathbf{X}). \quad (1.15)$$

As is pointed in Ullah (1996), 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 statistic as, e.g., for detecting influential data in regression analysis and model comparisons (see Arellano-Valle et al., 2000).

The Jensen–Shannon (JS) distance (Lin, 1991) corresponds to the capability of a random variable with two entries and giving as output the pdfs of \mathbf{X} and \mathbf{Y} . This also provides important properties in the study of decision-making problems (Lin, 1991) and complex

networks (Carpi et al., 2011). JS is defined in terms of Shannon entropy as

$$JS(\mathbf{X}, \mathbf{Y}) = H\left(\frac{\mathbf{X} + \mathbf{Y}}{2}\right) - \frac{1}{2}\{H(\mathbf{X}) + H(\mathbf{Y})\}, \quad (1.16)$$

and has the properties of a metric given by:

1. non-negativity, $JS(\mathbf{X}, \mathbf{Y}) \geq 0$;
2. disappears if and only if the two densities are equal almost everywhere, $JS(\mathbf{X}, \mathbf{Y}) = 0 \Leftrightarrow \mathbf{X} = \mathbf{Y}$;
3. symmetry, $JS(\mathbf{X}, \mathbf{Y}) = JS(\mathbf{Y}, \mathbf{X})$;
4. triangular inequality, $JS(\mathbf{X}, \mathbf{Y}) \leq JS(\mathbf{X}, \mathbf{Z}) + JS(\mathbf{Z}, \mathbf{Y})$, $\forall \mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^k$ (Briët and Harremoës, 2009).

JS distance has some advantages over the KL (symmetry and triangular inequality) and J (triangular inequality) divergences.

Chapter 2

Asymmetric and heavy-tailed class of distributions

In the last three decades, non-normal distributions have received substantial interest within scientific literature, especially in regards to skew-elliptical distributions. This class of flexible distribution is characterized for account with skewness and heavy-tails as an extra parameters respect to elliptical distributions as normal and t (Arellano-Valle and Bolfarine, 1995, Arellano-Valle et al., 1994, Fang et al., 1990, Genton, 2004), behind the multivariate skew-normal (SN, Azzalini and Capitanio, 1999, Azzalini and Dalla-Valle, 1996) and multivariate skew- t (ST, Arellano-Valle and Genton, 2005, Azzalini and Capitanio, 2003, Branco and Dey, 2001, Gupta, 2003) distributions. They have been successfully applied to numerous datasets from a wide range of fields including biological sciences, geophysics, astronomy, engineering and economics. Some recent applications of ST models include those by Genton (2004), Lee et al. (2010), Ghizzoni et al. (2010) and Eling (2012).

The family of SN distributions has been popularized by Azzalini (1985) and ever since it has been discussed extensively in the literature. Such discussions include a wide variety of skewed models in addition to having normal distribution as a special case and flexibility in capturing skewness in the data (Azzalini and Capitanio, 2013). In this sense, González-Farías et al. (2004) present the closed skew-normal distribution as an extension of the SN case, but closed under operations such as sums, marginalization, and linear conditioning (Rezaie et al., 2014). Another generalization of the skew-normal distribution is the extended SN distribution (ESN, Capitanio et al., 2003) that adds a fourth real parameter to accommodate both skewness and heavy tails. All these distributions are extended and unified by Arellano-Valle and Genton (2005), Arellano-Valle and Azzalini (2006), Arellano-Valle et al. (2006) and Arellano-Valle (2010). In some cases where observed variables can be simultaneously skewed and restricted to a fixed interval, the truncated SN distribution (TSN, Jamalizadeh et al., 2009) is a good

choice for those applications, especially for environmental and biological variables in which the observations are positives (Flecher et al., 2010).

In this section, we review five class of asymmetric and heavy-tailed distributions. Specifically the multivariate elliptical distributions that contains the normal and t cases; the multivariate skew-elliptical (SE) distributions that contains the SN and ST cases, and with the finite mixture of SN (FMSN) distributions as another case; the closed skew-normal (CSN) distribution that contain the SN, ESN, and TSN cases; and the generalized skew-normal (GSN) distribution that contains the skew-normal (SN) and modified skew-normal (MSN) cases.

2.1 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, e.g., Arellano-Valle et al. (1994), Fang et al. (1990). In this section, we give the ingredients to compute the elliptical mutual information index.

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

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

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

$$g(s) = \frac{\pi^{k/2}}{\Gamma(k/2)} s^{k/2-1} h^{(k)}(s), \quad s > 0, \quad (2.1)$$

is a valid pdf. Note that $f(\mathbf{z}) = |\Omega|^{-1/2} h^{(k)}(\mathbf{z}_0^T \mathbf{z}_0)$, where $\mathbf{z}_0 = \Omega^{-1/2}(\mathbf{z} - \xi)$.

As in Arellano-Valle et al. (2006), we call the distribution of S a squared-radial distribution and we denote it by $\mathcal{R}^2(h^{(k)})$. Considering that if

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim EC_{n+m} \left(\begin{pmatrix} \xi_{\mathbf{X}} \\ \xi_{\mathbf{Y}} \end{pmatrix}, \begin{pmatrix} \Omega_{\mathbf{XX}} & \Omega_{\mathbf{XY}} \\ \Omega_{\mathbf{YX}} & \Omega_{\mathbf{YY}} \end{pmatrix}, h^{(n+m)} \right),$$

then the respective marginal random vectors are distributed as $\mathbf{X} \sim EC_n(\xi_{\mathbf{X}}, \Omega_{\mathbf{XX}}, h^{(n)})$ and $\mathbf{Y} \sim EC_m(\xi_{\mathbf{Y}}, \Omega_{\mathbf{YY}}, h^{(m)})$. The determinant of the joint dispersion matrix Ω can be computed

as

$$|\Omega| = \begin{vmatrix} \Omega_{XX} & \Omega_{XY} \\ \Omega_{YX} & \Omega_{YY} \end{vmatrix} = |\Omega_{YY}| |\Omega_{XX}| |I_n - \mathbf{B}_{X \cdot Y} \mathbf{B}_{Y \cdot X}|, \quad (2.2)$$

where $\mathbf{B}_{X \cdot Y} = \Omega_{XX}^{-1} \Omega_{XY}$ and $\mathbf{B}_{Y \cdot X} = \Omega_{YY}^{-1} \Omega_{YX}$ are the matrices of regression coefficients associated with the regression functions of $\mathbf{X}|\mathbf{Y} = \mathbf{y}$ and $\mathbf{Y}|\mathbf{X} = \mathbf{x}$, respectively. Note here that $0 \leq |I_n - \mathbf{B}_{X \cdot Y} \mathbf{B}_{Y \cdot X}| \leq 1$.

The normal and 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

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

where

$$h_N^{(k)}(s) = (2\pi)^{-k/2} e^{-s^2/2}, \quad s > 0,$$

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

2.1.1 Multivariate normal distribution

The multivariate normal distribution, namely $\mathbf{Z} \sim N_k(\xi, \Omega)$, is a particular member of the elliptical family. In this case, $E(\mathbf{Z}) = \xi$ and $\text{Var}(\mathbf{Z}) = \Omega$. Moreover, for the normal density generator function and for the distribution of the normal squared radial random variable we have $S = (\mathbf{Z} - \xi)^T \Omega^{-1} (\mathbf{Z} - \xi) \sim \chi_k^2$, the chi-squared distribution with k degrees of freedom.

2.1.2 Multivariate t distribution

Another important member is the multivariate t distribution $\mathbf{Z} \sim T_k(\xi, \Omega, \nu)$, where $\nu > 0$ is the degrees of freedom, for which $E(\mathbf{Z}) = \xi$ for $\nu > 1$ and $\text{Var}(\mathbf{Z}) = \frac{\nu}{\nu-2} \Omega$ for $\nu > 2$. For the Student- t distribution, we have

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

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 Fang et al. (1990), Arellano-Valle et al. (1994) and in Arellano-Valle et al. (2006). For the particular case of the t distributions, see Arellano-Valle and Bolfarine (1995).

2.1.3 Scale mixture of normal

For this distribution (Andrews and Mallows, 1974), we have

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

for some cdf F on $(0, \infty)$. By considering the following stochastic representation:

$$\mathbf{Z} \stackrel{d}{=} \xi + V^{-1/2} \mathbf{Z}_N,$$

where $V \sim F$ is independent of $\mathbf{Z}_N \sim N_k(\mathbf{0}, \Omega)$; we have the scale mixture of normal (SMN) random variable $S_{SMN} \stackrel{d}{=} V^{-1} S_N$, where $V \sim F$ is independent of $S_N \sim \chi_k^2$. This class of distributions will not be addressed in Chapter 3.

2.2 Multivariate skew-elliptical distributions

A flexible class of location-scale models is defined by the so-called SE family of distributions; see Arellano-Valle and Azzalini (2006), Arellano-Valle and Genton (2005, 2010), Arellano-Valle et al. (2006), Azzalini and Capitanio (1999, 2003), Branco and Dey (2001) and Genton (2004). It allows for modeling 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 SE distribution, with location vector $\xi \in \mathbb{R}^k$, dispersion matrix $\Omega \in \mathbb{R}^{k \times k}$, shape/skewness parameter $\eta \in \mathbb{R}^k$ and density generator function $h^{(k+1)}$, denoted by $\mathbf{Z} \sim SE_k(\xi, \Omega, \eta, h^{(k+1)})$, if its pdf is

$$f(\mathbf{z}) = 2f_k(\mathbf{z}; \xi, \Omega, h^{(k)})F(\eta^T(\mathbf{z} - \xi); h_s^{(1)}), \quad \mathbf{z} \in \mathbb{R}^k, \quad (2.3)$$

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

Let $\bar{\eta} = \Omega^{1/2}\eta$. In terms of $\mathbf{z}_0 = \Omega^{-1/2}(\mathbf{z} - \xi)$, the SE pdf (2.3) can be rewritten as $f(\mathbf{z}) = |\Omega|^{-1/2}f(\mathbf{z}_0)$, where

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

is the pdf of $\mathbf{Z}_0 = \Omega^{-1/2}(\mathbf{Z} - \xi) \sim SE_k(\mathbf{0}, I_k, \bar{\eta}, h^{(k+1)})$ (see, e.g., Arellano-Valle and Genton, 2010).

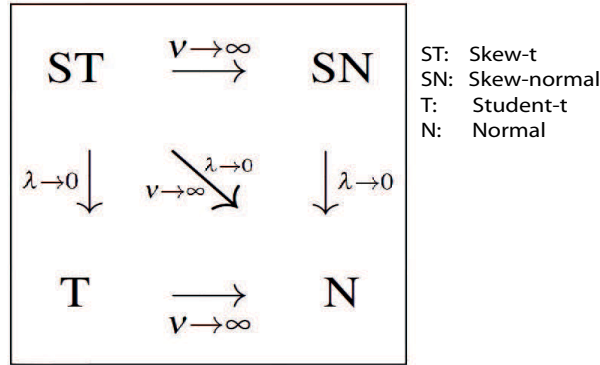


Fig. 2.1 Relationships among the skew-elliptical distributions: the skew- t (ST) contains the t (T), skew-normal (SN) and normal (N) distributions as special cases.

For the univariate case ($\eta = \lambda$), see Figure 2.1 for relationships among skew-elliptical distributions considered in this study. We study the SN and ST special cases in the next sections.

2.2.1 Multivariate skew-normal distribution

The SN distribution has been introduced by Azzalini and 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 as considered in Arellano-Valle and Genton (2005). We say that a random vector $\mathbf{Z} \in \mathbb{R}^k$ has a SN distribution with location vector $\xi \in \mathbb{R}^k$, dispersion matrix $\Omega \in \mathbb{R}^{k \times k}$ and shape/skewness parameter $\eta \in \mathbb{R}^k$, denoted by $\mathbf{Z} \sim SN_k(\xi, \Omega, \eta)$, if its pdf is

$$f(\mathbf{z}) = 2\phi_k(\mathbf{z}; \xi, \Omega)\Phi\{\eta^T(\mathbf{z} - \xi)\}, \quad \mathbf{z} \in \mathbb{R}^k, \quad (2.4)$$

where $\phi_k(\mathbf{z}; \xi, \Omega) = |\Omega|^{-1/2}\phi_k(\mathbf{z}_0)$ is the $N_k(\xi, \Omega)$ pdf, $\mathbf{z}_0 = \Omega^{-1/2}(\mathbf{z} - \xi)$, $\phi_k(\mathbf{z}_0)$ is the $N_k(\mathbf{0}, I_k)$ pdf, and Φ is the univariate $N_1(0, 1)$ cdf.

We can rewrite (2.4) as

$$f(\mathbf{z}) = |\Omega|^{-1/2} f(\mathbf{z}_0), \quad \text{with} \quad f(\mathbf{z}_0) = 2\phi_k(\mathbf{z}_0)\Phi(\bar{\eta}^T \mathbf{z}_0), \quad (2.5)$$

where $\bar{\eta} = \Omega^{1/2}\eta$.

The stochastic representation of \mathbf{Z} is given by

$$\mathbf{Z} \stackrel{d}{=} \xi + \delta|U_0| + \mathbf{U}, \quad (2.6)$$

here $U_0 \sim N(0, 1)$ and $\mathbf{U} \sim N_d(\mathbf{0}, \Omega - \delta\delta^\top)$, $\delta = \Omega\eta/\sqrt{1 + \eta^\top\Omega\eta}$, $|\delta| < 1$, which are independent. $|U_0|$ represents the absolute value of U_0 , i.e., it is half-normal distributed. From (2.6), Azzalini and Capitanio (1999) derived the mean vector and covariance matrix of \mathbf{Z} :

$$E[\mathbf{Z}] = \xi + \sqrt{\frac{2}{\pi}}\delta, \quad (2.7)$$

$$\text{Var}[\mathbf{Z}] = \Omega - \frac{2}{\pi}\delta\delta^\top. \quad (2.8)$$

2.2.2 Finite mixtures of multivariate skew-normal distributions

Let us consider the definition of Frühwirth-Schnatter and Pyne (2010) for finite mixtures of SN (FMSN) distributions. The pdf of a m -component mixture model with parameter vector set $\tilde{\boldsymbol{\theta}} = (\tilde{\xi}, \tilde{\Omega}, \tilde{\eta})$, where $\tilde{\xi} = (\xi_1, \dots, \xi_m)$ is a set of m location vector parameters, $\tilde{\Omega} = (\Omega_1, \dots, \Omega_m)$ is a set of m dispersion matrices, $\tilde{\eta} = (\eta_1, \dots, \eta_m)$ is a set of shape vector, is

$$f(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi}) = \sum_{i=1}^m \pi_i f(\mathbf{y}; \boldsymbol{\theta}_i), \quad (2.9)$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_m)$ is a vector of mixing weights π_i , with $\pi_i \geq 0$, $\sum_{i=1}^m \pi_i = 1$, and $f(\mathbf{y}; \boldsymbol{\theta}_i)$ are defined as in (2.4) with $\boldsymbol{\theta}_i = (\xi_i, \Omega_i, \eta_i)$, $i = 1, \dots, m$. Additional details about the log-likelihood function of an FMSN model are described in Lin (2009). Let $\mathbf{S} = (S_1, \dots, S_n)$ be a set of n latent allocations for the distribution of a FMSN random vector $\mathbf{Y} = (Y_1, \dots, Y_n)$, $f(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi}) = \prod_{j=1}^n f(s_j; \tilde{\boldsymbol{\theta}})$, where $\Pr(S_j = i | \boldsymbol{\pi}) = \pi_i$. Then, conditionally on $S_j = i$, an stochastic representation of the j -th component in (2.6) is

$$\mathbf{Y}_j | (S_j = i) \stackrel{d}{=} \xi_i + \delta_i |U_{0j}| + \sqrt{\Omega_i - \delta_i \delta_i^\top} \mathbf{U}_j, \quad j = 1, \dots, n, \quad (2.10)$$

where $U_{0j} \sim N(0, 1)$ and $\mathbf{U}_j \sim N_k(0, I_k)$ and they are mutually independent and $\delta_i = \Omega_i \eta_i / \sqrt{1 + \eta_i^\top \Omega_i \eta_i}$, $i = 1, \dots, m$. Considering the stochastic representation (2.10), (2.7)

and (2.8); we obtain for the mean vector and covariance matrix of \mathbf{Y} that

$$E[\mathbf{Y}] = \sum_{i=1}^m \pi_i \left(\xi_i + \sqrt{\frac{2}{\pi}} \delta_i \right), \quad (2.11)$$

$$\text{Var}[\mathbf{Y}] = \sum_{i=1}^m \pi_i \left[\Omega_i - \frac{2}{\pi} \delta_i \delta_i^\top + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top \right], \quad (2.12)$$

with $\boldsymbol{\mu}_i = \xi_i + \sqrt{\frac{2}{\pi}} \delta_i - E[\mathbf{Y}]$, $i = 1, \dots, m$.

2.2.3 Multivariate skew- t distribution

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

$$f(\mathbf{z}) = 2|\Omega|^{-1/2} t_k(\mathbf{z}_0; \nu) T \left(\sqrt{\frac{\nu + k}{\nu + \|\mathbf{z}_0\|^2}} \bar{\eta}^T \mathbf{z}_0; \nu + k \right), \quad (2.13)$$

where as before $\mathbf{z}_0 = \Omega^{-1/2}(\mathbf{z} - \xi)$, $t_k(\mathbf{x}; \nu)$ is the $t_k(\mathbf{0}, I_k, \nu)$ pdf, and $T(\mathbf{x}; \nu + k)$ is the $T_1(0, 1, \nu + k)$ cdf; see Branco and Dey (2001), Azzalini and Capitanio (2003), Gupta (2003), Ma and Genton (2004) and Arellano-Valle and Genton (2005). For this model, we have $\mathbf{Z}_0 \sim ST_k(\mathbf{0}, I_k, \bar{\eta}, \nu)$, with pdf

$$f(\mathbf{z}_0) = 2t_k(\mathbf{z}_0; \nu) T \left(\sqrt{\frac{\nu + k}{\nu + \|\mathbf{z}_0\|^2}} \bar{\eta}^T \mathbf{z}_0; \nu + k \right).$$

The MST distribution includes the MT distribution as a special case and is related to the MSN distribution by the equation

$$\mathbf{Z} \stackrel{d}{=} \xi + V^{-1/2} \mathbf{Z}_0, \quad (2.14)$$

where $\mathbf{Z}_0 \sim SN_k(\mathbf{0}, \Omega, \eta)$ and $V \sim \chi^2/\nu$, and they are independent, where χ^2 denote the univariate chi-square distribution (Azzalini and Capitanio, 2003). Also is well-know that $E(V^{-m/2}) = K_m(\mathbf{Z})$, $m \geq 1$, where $V \sim \chi^2/\nu$ and

$$K_m(\mathbf{Z}) = \left(\frac{\nu}{2} \right)^{m/2} \frac{\Gamma\left(\frac{\nu-m}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \quad \nu > m,$$

is the m -cumulant function of \mathbf{Z} .

Remark 1. An asymptotic expression for gamma function given by Jose and Naik (2008) allow the approximation of the function

$$B_k(v) = \frac{\Gamma\left(\frac{v+k}{2}\right)}{\Gamma\left(\frac{v}{2}\right)(v\pi)^{k/2}}, \quad v > 0.$$

Let $a < \infty$, then $\Gamma(x+a) \approx \sqrt{2\pi}x^{x+a-1/2}e^{-x}$, as $|x| \rightarrow \infty$. Therefore, $B_k(v)$ can be approximated by

$$B_k(v) \approx (2\pi)^{-k/2}, \quad (2.15)$$

as $v \rightarrow \infty$.

2.2.4 Scale mixture of skew-normal

As in Section 2.1.3, we can define the scale mixture of SN distributions (SMSN; see e.g. Contreras-Reyes and Arellano-Valle, 2013) by replacing $\mathbf{Z}_N \sim N_k(\mathbf{0}, \Omega)$ with $\mathbf{Z}_{SN} \sim SN_k(\mathbf{0}, \Omega, \eta)$ in the stochastic representation. This class contains several distributions: SMN, SN, ST, skew contaminated normal and skew-slash distributions. This class of distributions will not be addressed in Chapter 3.

2.3 Multivariate closed skew-normal distributions

Concerning the definition of González-Farías et al. (2004) (see also Arellano-Valle and Azzalini, 2006), let $\mathbf{Y} \in \mathbb{R}^k$ be a random vector with CSN distribution denoted as $CSN_{k,s}(\mu, \Omega, \mathbf{D}, \varphi, \mathbf{A})$ and with pdf

$$f(\mathbf{y}) = \phi_k(\mathbf{y}; \mu, \Omega) \frac{\Phi_s(\mathbf{D}^\top(\mathbf{y} - \mu); \varphi, \mathbf{A})}{\Phi_s(\mathbf{0}; \varphi, \mathbf{A} + \mathbf{D}^\top \Omega \mathbf{D})}, \quad (2.16)$$

where $\mu \in \mathbb{R}^k$, $\varphi \in \mathbb{R}^s$, $\Omega \in \mathbb{R}^{k \times k}$ and $\mathbf{A} \in \mathbb{R}^{s \times s}$ are both covariance matrices, $\mathbf{D} \in \mathbb{R}^{k \times s}$, $\phi_k(\mathbf{y}; \mu, \Omega)$ and $\Phi_k(\mathbf{y}; \mu, \Omega)$ are the pdf and cdf, respectively, of the k -dimensional normal distribution with mean vector μ and variance matrix Ω . The CSN distribution is closed under translations, scalar multiplications, and full, row rank linear transformations (Genton, 2004, González-Farías et al., 2004). The CSN is redefined by Arellano-Valle et al. (2006) as unified skew-normal (SUN) distribution. SUN distribution presents a more flexible conditioning mechanism than CSN, by relaxing the conditions $\Omega > 0$ and $\mathbf{A} > 0$. In addition, in CSN case the skewness parameter \mathbf{D} is not invariant to changes of scale. Once the above aspects of the parametrization are adjusted, the SUN and CSN classes are equivalent, on setting $\bar{\Omega} = \Omega$, $\Gamma = \mathbf{A} + \mathbf{D}^\top \Omega \mathbf{D}$ and $\Delta = \Omega \mathbf{D}$ in Eqs. (8) and (10) of Arellano-Valle et al. (2006).

Let $\mathbf{T} \in \mathbb{R}^{n \times k}$ be a matrix with rank n such that $k \leq n$, then

$$\mathbf{TY} = \text{CSN}_{n,s}(\mathbf{T}\boldsymbol{\mu}, \tilde{\boldsymbol{\Omega}}, \tilde{\mathbf{D}}, \boldsymbol{\varphi}, \tilde{\mathbf{T}}) \quad (2.17)$$

where $\tilde{\boldsymbol{\Omega}} = \mathbf{T}^\top \boldsymbol{\Omega} \mathbf{T}$, $\tilde{\mathbf{D}} = \mathbf{D}^\top \boldsymbol{\Omega} \mathbf{T} \tilde{\boldsymbol{\Omega}}^{-1}$, and $\tilde{\mathbf{T}} = \mathbf{A} + \mathbf{D}^\top \boldsymbol{\Omega} \mathbf{D} - \tilde{\mathbf{D}}^\top \tilde{\boldsymbol{\Omega}} \tilde{\mathbf{D}}$ (see Proposition 2.3.1 of Genton, 2004).

A particular case of (2.17) is the standardised random vector $\mathbf{z}_0 = \boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})$. In this case, Eq. (2.16) is rewritten as

$$f(\mathbf{z}_0) = \phi_k(\mathbf{z}_0) \frac{\Phi_s(\mathbf{D}^\top \boldsymbol{\Omega}^{1/2} \mathbf{z}_0; \boldsymbol{\varphi}, \mathbf{A})}{\Phi_s(\mathbf{0}; \boldsymbol{\varphi}, \mathbf{A} + \mathbf{D}^\top \boldsymbol{\Omega} \mathbf{D})}. \quad (2.18)$$

Given that the CSN distribution is closed under translations and by property (2.17), the standardised random vector \mathbf{Z}_0 follows $\text{CSN}_{k,s}(\mathbf{0}, I_k, \mathbf{D}^\top \boldsymbol{\Omega}^{1/2}, \boldsymbol{\varphi}, \mathbf{A})$, where I_k denotes the k -dimensional identity matrix. For more details, see Flecher et al. (2009) and Genton (2004). For the moment generating function of the CSN distribution, see González-Farías et al. (2004).

Lemma 2. (Flecher et al., 2009). Let \mathbf{Y} be a $\text{CSN}_{k,s}(\boldsymbol{\mu}, \boldsymbol{\Omega}, \mathbf{D}, \mathbf{0}, \mathbf{A})$, r a positive integer and $h(\mathbf{y}) = h(y_1, \dots, y_k)$ be any real valued function such that $E[h(\mathbf{Y})]$ is finite, then

$$E[h(\mathbf{Y})\{\Phi_k(\mathbf{Y}; \mathbf{0}, I_k)\}^r] = E[h(\tilde{\mathbf{Y}})] \frac{\Phi_{rk+s}(\mathbf{0}; \tilde{\boldsymbol{\varphi}}, \tilde{\mathbf{A}} + \tilde{\mathbf{D}}^\top \boldsymbol{\Omega} \tilde{\mathbf{D}})}{\Phi_s(\mathbf{0}; \mathbf{0}, \mathbf{A} + \mathbf{D}^\top \boldsymbol{\Omega} \mathbf{D})}, \quad (2.19)$$

where $\tilde{\mathbf{Y}} \sim \text{CSN}_{k, rk+s}(\boldsymbol{\mu}, \boldsymbol{\Omega}, \tilde{\mathbf{D}}, \tilde{\boldsymbol{\varphi}}, \tilde{\mathbf{A}})$ with $\tilde{\mathbf{D}} = (\mathbf{E}^\top, \mathbf{D}^\top)$, \mathbf{E} a $k \times rk$ matrix defined by $\mathbf{E} = (I_k, \dots, I_k)$, $\tilde{\boldsymbol{\varphi}} = (-\boldsymbol{\mu}, \dots, -\boldsymbol{\mu}, \mathbf{0}_s)$ a $(rk+s)$ vector and $\tilde{\mathbf{A}} = \begin{pmatrix} I_{rk} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{pmatrix}$.

2.3.1 Multivariate skew-normal distribution

A special case of CSN is the normal density when $\mathbf{D} = \mathbf{0}$ as is defined in Section 2.2.1. When $s = 1$ and $\mathbf{D} = \boldsymbol{\eta}$, the multivariate SN density function is obtained (Azzalini and Capitanio, 1999, 2013, Azzalini and Dalla-Valle, 1996). For more details about MSN distribution, see Section 2.2.1.

2.3.2 Multivariate extended skew-normal distribution

The ESN distribution is another special case of the CSN or SUN distribution which is obtained for $s = 1$. Consider a slight variant of the ESN distribution proposed by Capitanio et al. (2003). Let $\mathbf{Z} \sim \text{ESN}_k(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau})$, $\mathbf{Z} \in \mathbb{R}^k$, with mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$, variance matrix

$\Omega \in \mathbb{R}^{k \times k}$, shape/skewness parameter $\eta \in \mathbb{R}^k$, extended parameter $\tau \in \mathbb{R}$, and with pdf given by:

$$f(\mathbf{z}) = \frac{1}{\Phi_1(\tau)} \phi_k(\mathbf{z}; \mu, \Omega) \Phi_1[\eta^\top (\mathbf{z} - \mu) + \tilde{\tau}], \quad (2.20)$$

where $\mathbf{z} \in \mathbb{R}^k$ and $\tilde{\tau} = \tau \sqrt{1 + \eta^\top \Omega \eta}$. The mean vector and the variance matrix of \mathbf{Z} are

$$E[\mathbf{Z}] = \mu + \delta \zeta_1(\tau), \quad (2.21)$$

$$\text{Var}[\mathbf{Z}] = \Omega - \zeta_1(\tau) [\tau + \zeta_1(\tau)] \delta \delta^\top, \quad (2.22)$$

respectively; where $\zeta_1(\mathbf{z}) = \phi(\mathbf{z})/\Phi_1(\mathbf{z})$ is the *zeta* function (Azzalini and Capitanio, 1999, Capitanio et al., 2003). The SN distribution is a particular case of ESN distribution by setting $\tau = 0$ in (2.20). From Arellano-Valle and Azzalini (2006), a stochastic representation of the ESN distribution is

$$\mathbf{Z} \stackrel{d}{=} \mathbf{W} + \delta U, \quad (2.23)$$

where $\delta = \Omega \eta / \sqrt{1 + \eta^\top \Omega \eta}$, $U \sim \text{LTN}_{(-\tau, \infty)}(0, 1)$, which is independent of $\mathbf{W} \sim N_k(\xi, \Sigma)$, $\Sigma = \Omega - \delta \delta^\top$, where $\text{LTN}_{(-\tau, \infty)}(0, 1)$ represents the unit normal distribution truncated below the point $-\tau$ and “ $\stackrel{d}{=}$ ” denotes equality in terms of distribution. From the stochastic representation (2.23) it follows that $\mathbf{Z} \stackrel{d}{=} \mathbf{W} + \delta W_\tau$, where $W_\tau \stackrel{d}{=} (W_0 \mid W_0 + \tau > 0)$ and

$$\begin{pmatrix} W_0 \\ \mathbf{W} \end{pmatrix} \sim N_{1+k} \left(\begin{pmatrix} 0 \\ \xi \end{pmatrix}, \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \Omega \end{pmatrix} \right).$$

Note that W_0 and \mathbf{W} are independent. This fact must be interpreted as the non-normality effect of W_0 on \mathbf{Z} , where W_0 is the so-called *unobserved confounder* variable.

2.4 Generalized skew-normal distribution

An attractive class of skew-symmetric (SS) distributions defined in terms of pdf appears in Azzalini (1985), Azzalini and Capitanio (1999) and Gupta et al. (2002)

$$f(z; \eta) = 2f(z)G\{w(z; \eta)\}, \quad z \in \mathbb{R}, \quad (2.24)$$

where $\eta \in \mathbb{R}$ represents a skewness/shape parameter, f and G are the respective pdf and cdf of symmetrical continuous distributions, and $w(z; \eta)$ an odd function of z , with $w(0; \eta) = 0$ for any fixed value of η . Also, we assume that $w(z; \eta_0) = 0$ for all z and some value η_0 of η (typically $\eta_0 = 0$), so that $f(z; \eta_0) = f(z)$, thus recovering symmetry.

The notation $Z \sim SS(\eta; f, G, w)$ expresses that random variable Z has distribution with pdf given by (2.24). If $f(z) = \phi(z)$ represents the pdf of the standardized normal distribution, denoted by $N(0, 1)$, then (2.24) becomes a family of SS distributions generated by the normal kernel, also known as GSN family. In this case, $Z \sim GSN(\eta; G, w)$ emerges. An important property of the GSN random variable Z is that all its moments are finite. In particular, it possess the same even moments of $Z_0 \sim N(0, 1)$. For instance, $E(Z^2) = 1$ and so $Var(Z) = 1 - \mu_z^2$, where $\mu_z = E(Z)$. The most popular GSN distribution is SN (Azzalini and Capitanio, 2013), for which $w(z; \eta) = \eta z$ and $G(z) = \Phi(z)$ is the cdf of the standardized normal distribution. So $Z \sim SN(\eta)$ expresses that Z follows an SN distribution. The location-scale extension of the SS pdf in (2.24) follows by applying the Jacobian method to the linear random variable $X = \mu + \sigma Z$, where $\mu \in \mathbb{R}$ and $\sigma > 0$. In this case, we state that X follows a SS distribution with location parameter μ , scale parameter σ and shape/skewness parameter η and get $X \sim SS(\mu, \sigma^2, \eta; f, G, w)$. Furthermore, we write $X \sim GSN(\mu, \sigma^2, \eta; G, w)$ if $f = \phi$, $X \sim GSN(\mu, \sigma^2, \eta; G)$ if $f = \phi$ and $w(z; \eta) = \eta z$, and $X \sim SN(\mu, \sigma^2, \eta)$ if $f = \phi$, $w(z; \eta) = \eta z$ and $G = \Phi$.

Two other members of the GSN family that have been studied recently are the Skew-Normal-Cauchy (SNC) distribution (Arrué et al., 2010, Nadarajah and Kotz, 2003), which follows from (2.24) by taking $f(z) = \phi(z)$, $w(z; \eta) = \eta z$ and $G(z) = 1/2 + (1/\pi)\arctan(z)$, and the MSN distribution (Arrué et al., 2016), for which $f(z) = \phi(z)$, $w(z; \eta) = \eta z/\sqrt{1+z^2}$ and $G(z) = \Phi(z)$. Nadarajah and Kotz (2003) recall that SNC distribution appears to attain a higher degree of sharpness than normal distribution, i.e., disparity exists from the common normal distribution produced by the skewness parameter η . A random variable Z with SNC or MSN distribution is denoted, respectively, by $Z \sim SNC(\eta)$ or $Z \sim MSN(\eta)$, and by $X \sim SNC(\mu, \sigma^2, \eta; G)$ or $X \sim MSN(\mu, \sigma^2, \eta; G)$ for their respective location-scale extensions.

2.4.1 Skew-normal distribution

If $Z \sim SN(\eta)$ or $Z \sim SN(0, 1, \eta)$ represents a SN random variable, then its pdf is

$$f(z; \eta) = 2\phi(z)\Phi(\eta z), \quad z \in \mathbb{R}, \eta \in \mathbb{R}. \quad (2.25)$$

Clearly, if $\eta = \eta_0 = 0$, then (2.25) reduces to the $N(0, 1)$ -pdf. The SN random variable Z has the following properties (Azzalini, 1985):

1. As $\eta \rightarrow \infty$, $f(z; \eta)$ tends to the half-normal density.
2. If $Z \sim SN(\eta)$, then $-Z \sim SN(-\eta)$.
3. The density (2.25) is strongly unimodal, i.e. $\log f(z; \eta)$ is a concave function of z .

The SN random variable Z can be conveniently represented as a linear combination of half-normal and normal variables through the following stochastic representation (Henze, 1986):

$$Z \stackrel{d}{=} \delta|U_0| + \sqrt{1 - \delta^2}U, \quad (2.26)$$

where $\delta = \eta/\sqrt{1 + \eta^2}$, U_0 and U are independent and identically distributed with unit normal distribution. In particular, since the half-normal random variable $|U_0|$ has mean $b = \sqrt{2/\pi}$ and variance 1, it follows from (2.26) that the mean and variance of Z_τ , $\tau = |\eta|$, are given by

$$E(Z_\tau) = b\delta_\tau \quad \text{and} \quad \text{Var}(Z_\tau) = 1 - (b\delta_\tau)^2, \quad (2.27)$$

where $\delta_\tau = \tau/\sqrt{1 + \tau^2}$.

The moments $\mu_i = E(Z_\tau^i)$ are given by

$$\begin{aligned} \mu_{2i} &= (2i-1)!! = \frac{(2i)!}{2^i i!} \quad (\text{even moments}), \\ \mu_{2i-1} &= \frac{(2i-1)! b \tau}{2^{i-1} (1 + \tau^2)^{(2i-1)/2}} \sum_{j=0}^{i-1} \frac{j! (2\tau)^{2j}}{(2j+1)!(i-j-1)!} \quad (\text{odd moments; Henze, 1986}). \end{aligned}$$

From Proposition 2 in Martínez et al. (2008), the odd moments can also be computed as

$$\mu_{2i-1} = \frac{b \sum_{m=1}^i a_i(m) \tau^{2m-1}}{(1 + \tau^2)^{(2i-1)/2}},$$

where the coefficient $a_i(m)$ is computed iteratively as follows:

$$\begin{aligned} a_1(1) &= 1, \\ a_i(1) &= (2i-1)a_{i-1}(1), \quad i \geq 2, \\ a_i(m) &= 2(i-1)\{a_{i-1}(m) + a_{i-1}(m-1)\}, \quad 1 < m < i, \quad i \geq 2, \\ a_i(i) &= 2(i-1)a_{i-1}(i-1), \quad i \geq 2. \end{aligned}$$

2.4.2 Truncated skew-normal distribution

The TSN pdf given by Jamalizadeh et al. (2009) and Flecher et al. (2010), considers the random variable $Z \sim SN(\mu, \sigma^2, \eta)$, $Z \in \mathbb{R}$, and the definition given in (2.25). Flecher et al. (2010) gives the expressions of the higher order and weighted moments of TSN distributions. We also consider the following definition based on (2.25) (when $k = 1$) for a TSN random

variable $W \in [a, b] \subset \mathbb{R}$, denoted by $W \sim TSN(\mu, \sigma^2, \eta)$, and with density

$$g(w) = \frac{f(w)}{[F(w)]_a^b}, \quad a < w \leq b, \quad (2.28)$$

where $f(z)$ is defined in (2.25) for $k = 1$ with $\Sigma = \sigma^2$ and $F(z)$ is the cdf of Z with

$$[F(w)]_a^b = F(b) - F(a) = \int_a^b f(u) du.$$

The following Remark allows the computation of $[F(w)]_a^b$ in terms of the normal cdf and a bivariate integral term.

Remark 2. (Azzalini, 1985, Owen, 1956). Let $Z \sim SN(\mu, \sigma^2, \eta)$, $F(z)$ can be computed as follows

$$F(z) = 2 \int_z^{-\infty} \int_{-\infty}^{\lambda s} \phi(s) \phi(t) dt ds = \Phi_1(z) - 2 \int_z^{\infty} \int_0^{\lambda s} \phi(s) \phi(t) dt ds. \quad (2.29)$$

Then, by replacing (2.29) in $[F(w)]_a^b$, we obtain

$$[F(w)]_a^b = \Phi_1(b) - \Phi_1(a) - 2 \int_a^b \int_0^{\lambda s} \phi(s) \phi(t) dt ds.$$

2.4.3 Modified skew-normal distribution

The pdf for a random variable Z with MSN distribution, denoted by $Z \sim MSN(\eta)$, is given by

$$f(z) = 2\phi(z)\Phi\{\eta u(z)\}, \quad z \in \mathbb{R},$$

where $u(z) = z/\sqrt{1+z^2}$. The following properties are straightforward from the results in Arellano-Valle et al. (2004):

1. If $Z \sim MSN(\eta)$, then $-Z \sim MSN(-\eta)$.
2. If $Z \sim MSN(\eta)$, then $|Z| \sim \sqrt{\chi_1^2} \equiv HN(0, 1)$.
3. If $Z|S = s \sim SN(s)$ and $S \sim N(\lambda, 1)$, then $Z \sim MSN(\eta)$.
4. $MSN(0) \equiv N(0, 1)$ and $MSN(\eta) \equiv GSN(\eta; G, w)$ if $G(z) = \Phi(z)$ and $w(z; \eta) = \eta z / \sqrt{1+z^2}$.

Similarly to the SN case, the MSN random variable $Z_\tau \sim MSN(\tau)$, $\tau = |\eta|$, has even moments equal to the corresponding even moments of the standardized normal random variable Z_0 (Arrué et al., 2016), i.e., $\mu_{2i} = (2i-1)!! = (2i)!/(2^i i!)$. Recalling that $b = \sqrt{2/\pi}$, the odd moments $\mu_{2i-1}(\tau)$, $i = 1, 2, \dots$, can be computed as

$$\mu_{2i-1}(\tau) = 2^{i-1}(i-1)!b\{2\xi_i(\tau) - 1\},$$

with

$$\xi_i(\tau) = \int_0^\infty \Phi\{\tau u(\sqrt{x})\} \frac{x^{i-1} e^{-x/2}}{2^i \Gamma(i)} dx, \quad i = 1, 2, \dots,$$

where $\Gamma(\cdot)$ denotes the usual gamma function. Note that $\xi_i(\tau) = E[\Phi\{\tau u(\sqrt{X_i})\}]$, where $X_i \sim \chi_{2i}^2$, thus $0 < 2\xi_i(\tau) - 1 < 1$ for all $i = 1, 2, \dots$ and $\tau > 0$. In particular, the first four moments are $\mu_1(\tau) = b\{2\xi_1(\tau) - 1\}$, $\mu_2 = 1$, $\mu_3(\tau) = 2b\{2\xi_2(\tau) - 1\}$, and $\mu_4 = 3$.

Chapter 3

Information measures for symmetric and asymmetric distributions

Frequently, several authors have been computed information measures for a large list of univariate and multivariate distributions (Zografos and Nadarajah, 2005). Cover and Thomas (2006) provides the properties of Shannon entropy and KL divergence measures for multivariate normal distribution, among others. Zografos and Nadarajah (2005) gives the entropy for multivariate t distribution and for several other distributions.

In this chapter, we propose 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 (Fang et al., 1990, 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.

3.1 Information measures for multivariate elliptical distributions

For this class, we note that if the *standardized* random vector $\mathbf{Z}_0 = \Omega^{-1/2}(\mathbf{Z} - \xi)$ has a spherical pdf $f(\mathbf{z}_0) = h^{(k)}(\mathbf{z}_0^\top \mathbf{z}_0)$, $\mathbf{Z}_0 \in \mathbb{R}^k$, $\mathbf{Z} \sim EC_k(\xi, \Omega, h^{(k)})$, then its entropy is given by

$$H(\mathbf{Z}_0) = -E[\log \{h^{(k)}(\mathbf{Z}_0^\top \mathbf{Z}_0)\}]. \quad (3.1)$$

This expectation depends on the distribution of the squared radial random variable $S = \mathbf{Z}_0^\top \mathbf{Z}_0 = (\mathbf{Z} - \xi)^\top \Omega^{-1}(\mathbf{Z} - \xi)$ which has the radial-square pdf $g(s)$ given in Eq. (2.1).

Considering an elliptical random vector (\mathbf{X}, \mathbf{Y}) with marginal distributions $\mathbf{X} \sim EC_n(\xi_{\mathbf{X}}, \Omega_{\mathbf{X}\mathbf{X}}, h^{(n)})$ and $\mathbf{Y} \sim EC_m(\xi_{\mathbf{Y}}, \Omega_{\mathbf{Y}\mathbf{Y}}, h^{(m)})$, it is clear from the above results that the respective marginal and joint entropies are

$$\begin{aligned} H(\mathbf{X}) &= \frac{1}{2} \log |\Omega_{\mathbf{X}\mathbf{X}}| - E[\log \{h^{(n)}(S_{\mathbf{X}})\}], \\ H(\mathbf{Y}) &= \frac{1}{2} \log |\Omega_{\mathbf{Y}\mathbf{Y}}| - E[\log \{h^{(m)}(S_{\mathbf{Y}})\}], \\ H(\mathbf{XY}) &= \frac{1}{2} \log |\Omega| - E[\log \{h^{(n+m)}(S_{\mathbf{XY}})\}], \end{aligned}$$

where $S_{\mathbf{X}} \sim \mathcal{R}^2(h^{(n)})$, $S_{\mathbf{Y}} \sim \mathcal{R}^2(h^{(m)})$ and $S_{\mathbf{XY}} \sim \mathcal{R}^2(h^{(n+m)})$.

Considering the determinant of the joint dispersion matrix Ω (2.2), the elliptical mutual information index between \mathbf{X} and \mathbf{Y} is

$$\begin{aligned} I(\mathbf{X}, \mathbf{Y}) &= E[\log \{h^{(n+m)}(S_{\mathbf{XY}})\}] - E[\log \{h^{(n)}(S_{\mathbf{X}})\}] - E[\log \{h^{(m)}(S_{\mathbf{Y}})\}] \\ &\quad - \frac{1}{2} \log |I_n - \mathbf{B}_{\mathbf{X}\cdot\mathbf{Y}} \mathbf{B}_{\mathbf{Y}\cdot\mathbf{X}}|. \end{aligned} \quad (3.2)$$

The last term in (3.2) represents the information due the dispersion matrix Ω , which is the same for the whole elliptical class. A similar fact occurs with the correlation matrix induced by Ω , which means that within the elliptical family, the correlation does not depend on the specific elliptical density generator h . As a consequence from (3.2), the elliptical mutual information depends on both Ω and h , allowing differences for the association between \mathbf{X} and \mathbf{Y} through the different elliptical joint distributions.

3.1.1 Multivariate normal distribution

Let $\mathbf{Z} \sim N_k(\xi, \Omega)$ denote a k -dimensional normal random vector, with mean vector $E(\mathbf{Z}) = \xi \in \mathbb{R}^k$ and covariance matrix $\text{var}(\mathbf{Z}) = \Omega \in \mathbb{R}^{k \times k}$. We have $\mathbf{Z} = \xi + \Omega^{1/2} \mathbf{Z}_0$, where $\mathbf{Z}_0 \sim N_k(\mathbf{0}, I_k)$. From (3.1), since in this case $S = \mathbf{Z}_0^\top \mathbf{Z}_0 \sim \chi_k^2$, and so $E(S) = k$, we have

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

Therefore, by Lemma 1:

$$H(\mathbf{Z}) = \frac{1}{2} \log |\Omega| + \frac{k}{2} \{1 + \log(2\pi)\}. \quad (3.3)$$

Now let

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

It is well-known that the marginal distributions are $\mathbf{X} \sim N_n(\xi_{\mathbf{X}}, \Omega_{\mathbf{XX}})$ and $\mathbf{Y} \sim N_m(\xi_{\mathbf{Y}}, \Omega_{\mathbf{YY}})$. Hence, $S_{\mathbf{X}} \sim \chi_n^2$, $S_{\mathbf{Y}} \sim \chi_m^2$ and $S_{\mathbf{XY}} \sim \chi_{n+m}^2$, and therefore

$$\begin{aligned} H(\mathbf{X}) &= \frac{1}{2} \log |\Omega_{\mathbf{XX}}| + \frac{n}{2} \{1 + \log(2\pi)\}, \\ H(\mathbf{Y}) &= \frac{1}{2} \log |\Omega_{\mathbf{YY}}| + \frac{m}{2} \{1 + \log(2\pi)\}, \\ H(\mathbf{XY}) &= \frac{1}{2} \log |\Omega| + \frac{n+m}{2} \{1 + \log(2\pi)\}. \end{aligned}$$

Thus, we obtain from (3.2), that the normal mutual information index between \mathbf{X} and \mathbf{Y} is

$$I(\mathbf{X}, \mathbf{Y}) = \frac{1}{2} \log \left(\frac{|\Omega_{\mathbf{XX}}| |\Omega_{\mathbf{YY}}|}{|\Omega|} \right) = -\frac{1}{2} \log |I_n - \mathbf{B}_{\mathbf{X} \cdot \mathbf{Y}} \mathbf{B}_{\mathbf{Y} \cdot \mathbf{X}}|.$$

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

3.1.2 Multivariate t distribution

From (3.1), we have $\mathbf{Z}_0 = \Omega^{-1/2}(\mathbf{Z} - \xi) \sim T_k(\mathbf{0}, I_k, \nu)$ and $\mathbf{Z}_0^\top \mathbf{Z}_0 \sim k F_{k, \nu}$. Thus, considering that

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

we have $H(\mathbf{Z}_0) = E[\log \{h^{(k)}(S)\}]$ where $S \sim k F_{k, \nu}$. 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 $\psi(x) = \frac{d}{dx} \log\{\Gamma(x)\}$ is the digamma function. We find for the entropy of $\mathbf{Z}_0 \sim T_k(\mathbf{0}, I_k, \nu)$ that

$$H(\mathbf{Z}_0) = -\log \left\{ \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) (\nu\pi)^{k/2}} \right\} + \frac{\nu+k}{2} \left\{ \psi\left(\frac{\nu+k}{2}\right) - \psi\left(\frac{\nu}{2}\right) \right\}. \quad (3.4)$$

Now let

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

From Arellano-Valle and Bolfarine (1995), the marginal distributions are $\mathbf{X} \sim T_n(\xi_{\mathbf{X}}, \Omega_{\mathbf{XX}}, \nu)$ and $\mathbf{Y} \sim T_m(\xi_{\mathbf{Y}}, \Omega_{\mathbf{YY}}, \nu)$. Hence, from (1.4) and (3.4), we deduce that the mutual information index for the Student's t case is

$$\begin{aligned} I(\mathbf{X}, \mathbf{Y}) &= I(\mathbf{X}_0, \mathbf{Y}_0) + \log \left[\frac{\Gamma(\nu/2) \Gamma\{(\nu+n+m)/2\}}{\Gamma\{(\nu+n)/2\} \Gamma\{(\nu+m)/2\}} \right] - \frac{\nu+m}{2} \psi\left(\frac{\nu+m}{2}\right) \\ &\quad - \frac{\nu+n}{2} \psi\left(\frac{\nu+n}{2}\right) + \frac{\nu+n+m}{2} \psi\left(\frac{\nu+n+m}{2}\right) + \frac{\nu}{2} \psi\left(\frac{\nu}{2}\right), \end{aligned} \quad (3.5)$$

where $\mathbf{X}_0 \sim N_n(\xi_{\mathbf{X}}, \Omega_{\mathbf{XX}})$ and $\mathbf{Y}_0 \sim N_m(\xi_{\mathbf{Y}}, \Omega_{\mathbf{YY}})$. It is interesting to notice that the information due to Ω arises from $I(\mathbf{X}_0, \mathbf{Y}_0)$ only, and the information due to ν comes from the remaining terms. It is clear also that, as ν increases, the Student's t mutual information converges to the normal mutual information.

3.2 Information measures for multivariate skew-elliptical distributions

Lemma 1 yields the following result.

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

$$H(\mathbf{Z}) = H(\mathbf{Z}') - E[\log\{2F(\bar{\eta}^\top \mathbf{Z}_0; h_S^{(1)})\}],$$

where $H(\mathbf{Z}')$ is the entropy of $\mathbf{Z}' \sim EC_k(\xi, \Omega, h^{(k)})$, $\mathbf{Z}_0 \sim SE_k(\mathbf{0}, I_k, \bar{\eta}, h^{(k+1)})$, $\bar{\eta} = \Omega^{1/2} \eta$, and $S = \mathbf{Z}_0^\top \mathbf{Z}_0$.

It follows from Proposition 1 that to compute the entropy $H(\mathbf{Z}_0)$, we need only the joint distribution of $U = \bar{\eta}^\top \mathbf{Z}_0$ and $S = \mathbf{Z}_0^\top \mathbf{Z}_0$, where $\mathbf{Z}_0 \sim SE_k(\mathbf{0}, I_k, \bar{\eta}, h^{(k+1)})$. For this, the next result is necessary, the proof of which is given in Arellano-Valle et al. (2013).

Proposition 2. Let $U = \bar{\eta}^\top \mathbf{Z}_0$ and $S = \mathbf{Z}_0^\top \mathbf{Z}_0$, where $\mathbf{Z}_0 \sim SE_k(\mathbf{0}, I_k, \bar{\eta}, h^{(k+1)})$. Then $(U, S) \stackrel{d}{=} (\|\bar{\eta}\|W, S)$, where $\|\bar{\eta}\| = \bar{\eta}^\top \bar{\eta} = (\eta^\top \Omega \eta)^{1/2}$, and for $k \geq 2$ the joint pdf of (W, S) can be computed as $f_{W,S}(u, s) = f_{W|S=s}(u)f_S(s)$, where

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

and

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

3.2.1 Multivariate skew-normal distribution

Shannon Entropy and Mutual Information

We can apply Lemma 1 to obtain the entropy for the skew-normal (SN) random model. For this, we need the following preliminary result in order to simplify the computation of this entropy. Its proof can be found in Arellano-Valle and Genton (2010).

Lemma 3. Let $\mathbf{Z} \sim SN_k(\xi, \Omega, \eta)$ and $\mathbf{Z}_0 = \Omega^{-1/2}(\mathbf{Z} - \xi)$. Let also $\mathbf{Z}_{0N} \sim N_k(\mathbf{0}, I_k)$ and $W \sim SN_1(\|\bar{\eta}\|)$. Then, $\mathbf{Z}_0 \sim SN_k(\mathbf{0}, I_k, \bar{\eta}) \equiv SN_k(\bar{\eta})$, $\eta^\top (\mathbf{Z} - \xi) = \bar{\eta}^\top \mathbf{Z}_0 \stackrel{d}{=} \|\bar{\eta}\|W$, and $g(\mathbf{Z}_0) \stackrel{d}{=} g(\mathbf{Z}_{0N})$ for any even function g .

From Lemmas 1 and 3, we have the following result.

Proposition 3. The Shannon entropy of a SN random vector $\mathbf{Z} \sim SN_k(\xi, \Omega, \eta)$ is

$$H(\mathbf{Z}) = H(\mathbf{Z}_N) - E[\log\{2\Phi(\|\bar{\eta}\|W)\}],$$

where $H(\mathbf{Z}_N)$ is the entropy of $\mathbf{Z}_N \sim N_k(\xi, \Omega)$ given in (3.3), and $W \sim SN_1(\|\bar{\eta}\|)$.

In order to derive the mutual information index of the SN distribution, we need the following result about its marginal distributions. Let

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim SN_{n+m} \left(\begin{pmatrix} \xi_{\mathbf{X}} \\ \xi_{\mathbf{Y}} \end{pmatrix}, \begin{pmatrix} \Omega_{\mathbf{XX}} & \Omega_{\mathbf{XY}} \\ \Omega_{\mathbf{YX}} & \Omega_{\mathbf{YY}} \end{pmatrix}, \begin{pmatrix} \eta_{\mathbf{X}} \\ \eta_{\mathbf{Y}} \end{pmatrix} \right).$$

Then $\mathbf{X} \sim SN_n(\xi_{\mathbf{X}}, \Omega_{\mathbf{XX}}, \eta_{\mathbf{X}(\mathbf{Y})})$ and $\mathbf{Y} \sim SN_m(\xi_{\mathbf{Y}}, \Omega_{\mathbf{YY}}, \eta_{\mathbf{Y}(\mathbf{X})})$ where

$$\eta_{\mathbf{X}(\mathbf{Y})} = \frac{\eta_{\mathbf{X}} + \Omega_{\mathbf{XX}}^{-1} \Omega_{\mathbf{XY}} \eta_{\mathbf{Y}}}{\sqrt{1 + \eta_{\mathbf{Y}}^\top \Omega_{\mathbf{YY}}^{-1} \Omega_{\mathbf{YX}} \eta_{\mathbf{X}}}} \quad \text{and} \quad \eta_{\mathbf{Y}(\mathbf{X})} = \frac{\eta_{\mathbf{Y}} + \Omega_{\mathbf{YY}}^{-1} \Omega_{\mathbf{YX}} \eta_{\mathbf{X}}}{\sqrt{1 + \eta_{\mathbf{X}}^\top \Omega_{\mathbf{XX}}^{-1} \Omega_{\mathbf{XY}} \eta_{\mathbf{Y}}}}.$$

Consequently, by Proposition 3, we obtain the following results for the marginal and joint SN entropies.

Proposition 4. *Let*

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim SN_{n+m} \left(\begin{pmatrix} \xi_{\mathbf{X}} \\ \xi_{\mathbf{Y}} \end{pmatrix}, \begin{pmatrix} \Omega_{\mathbf{XX}} & \Omega_{\mathbf{XY}} \\ \Omega_{\mathbf{YX}} & \Omega_{\mathbf{YY}} \end{pmatrix}, \begin{pmatrix} \eta_{\mathbf{X}} \\ \eta_{\mathbf{Y}} \end{pmatrix} \right).$$

Then:

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

with $W_{\mathbf{X}} \sim SN_1(\|\bar{\eta}_{\mathbf{X}(\mathbf{Y})}\|)$, $W_{\mathbf{Y}} \sim SN_1(\|\bar{\eta}_{\mathbf{Y}(\mathbf{X})}\|)$ and $W_{\mathbf{XY}} \sim SN_1(\|\bar{\eta}_{\mathbf{XY}}\|)$, where

$$\|\bar{\eta}_{\mathbf{X}(\mathbf{Y})}\| = (\eta_{\mathbf{X}(\mathbf{Y})}^{\top} \Omega_{\mathbf{XX}} \eta_{\mathbf{X}(\mathbf{Y})})^{1/2}, \quad \|\bar{\eta}_{\mathbf{Y}(\mathbf{X})}\| = (\eta_{\mathbf{Y}(\mathbf{X})}^{\top} \Omega_{\mathbf{YY}} \eta_{\mathbf{Y}(\mathbf{X})})^{1/2},$$

and

$$\|\bar{\eta}_{\mathbf{XY}}\| = (\eta_{\mathbf{X}(\mathbf{Y})}^{\top} \Omega_{\mathbf{XX}} \eta_{\mathbf{X}(\mathbf{Y})} + \eta_{\mathbf{Y}(\mathbf{X})}^{\top} \Omega_{\mathbf{YY}} \eta_{\mathbf{Y}(\mathbf{X})} + 2\eta_{\mathbf{X}(\mathbf{Y})}^{\top} \Omega_{\mathbf{XY}} \eta_{\mathbf{Y}(\mathbf{X})})^{1/2}.$$

Thus, we obtain from (1.4) that the SN mutual information index between \mathbf{X} and \mathbf{Y} is

$$I(\mathbf{X}, \mathbf{Y}) = I(\mathbf{X}_0, \mathbf{Y}_0) + E \left\{ \log \left(\frac{V_{\mathbf{XY}}}{V_{\mathbf{X}(\mathbf{Y})} V_{\mathbf{Y}(\mathbf{X})}} \right) \right\},$$

where $\mathbf{X}_0 \sim N_k(\xi_{\mathbf{X}}, \Omega_{\mathbf{XX}})$, $\mathbf{Y}_0 \sim N_k(\xi_{\mathbf{Y}}, \Omega_{\mathbf{YY}})$, $V_{\mathbf{XY}} = 2\Phi(\|\bar{\eta}_{\mathbf{XY}}\|W_{\mathbf{XY}})$, $V_{\mathbf{X}(\mathbf{Y})} = 2\Phi(\|\bar{\eta}_{\mathbf{X}(\mathbf{Y})}\|W_{\mathbf{X}(\mathbf{Y})})$ and $V_{\mathbf{Y}(\mathbf{X})} = 2\Phi(\|\bar{\eta}_{\mathbf{Y}(\mathbf{X})}\|W_{\mathbf{Y}(\mathbf{X})})$. The mutual information between two normals, $I(\mathbf{X}_0, \mathbf{Y}_0)$, is given in (3.2).

Maximum entropy

We also explore the necessary inequalities to determine the bounds for the entropy of a variable distributed SN. By Cover and Thomas (2006), for any density $f_{\mathbf{X}}(\mathbf{x})$ of a random vector $\mathbf{X} \in \mathbb{R}^k$ - not necessary normal- with zero mean and variance-covariance matrix $\Sigma = E[\mathbf{XX}^{\top}]$, the entropy of \mathbf{X} is upper bounded as

$$H(\mathbf{X}) \leq \frac{1}{2} \log \left\{ (2\pi e)^k |\Sigma| \right\}, \quad (3.6)$$

and

$$H(\mathbf{X}_0) = \frac{1}{2} \log \left\{ (2\pi e)^k |\Omega| \right\}, \quad (3.7)$$

is the entropy of $\mathbf{X}_0 \sim N_k(\mathbf{0}, \Omega)$, i.e., the entropy is maximized under normality. Let $\mathbf{X} \sim SN_k(\xi, \Omega, \eta)$, our interest is now to give an alternative approximation of the entropy of the SN random vector \mathbf{X} . By Proposition 3, we have that the SN entropy is

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

where $W \sim SN_1(\tau)$ with $\tau = (\eta^\top \Omega \eta)^{1/2}$. By (3.20), (3.7) and Property (ii) of Lemma 4 (see below), we have that

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

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

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

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

On the other hand, Gao and Zhang (2009) uses the *Negentropy* to quantify the non-normality of a random variable \mathbf{X} , which defined as

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

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 $\mathbf{X} \sim SN_k(\xi, \Omega, \eta)$ coincide 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 normal 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 the Section 1.3 is a useful tool to analyze this fact.

Rényi entropy and complexity measure

The next Proposition 5 allows to compute the Rényi entropy and complexity measure of a SN random variable.

Proposition 5. *Let $\mathbf{Z} \in \mathbb{R}^k$ be a random vector with pdf given by (2.4). Then:*

$$\int_{\mathbb{R}^k} [f(\mathbf{z})]^\alpha k\mathbf{z} = \psi_{\alpha,d}(\Omega) \frac{\Phi_{\alpha+1}(\mathbf{0}; \mathbf{0}, \tilde{\Omega})}{\Phi_1(0; 0, \omega)}, \quad \alpha \in \mathbb{N}, \alpha > 1, \quad (3.8)$$

where

$$\psi_{\alpha,k}(\Omega) = \frac{2^\alpha}{\alpha^{k/2}} [(2\pi)^k |\Omega|]^{(1-\alpha)/2},$$

$\tilde{\Omega} = I_{\alpha+1} + \|\tilde{\eta}\|^2 \tilde{\mathbf{D}}^\top \tilde{\mathbf{D}}$, $\tilde{\mathbf{D}} = (\mathbf{1}_\alpha, \|\tilde{\eta}\|)^\top$, $\mathbf{1}_\alpha$ is the α -dimensional vector of ones, $\omega = 1 + \|\tilde{\eta}\|^4$, $\|\tilde{\eta}\| = \tilde{\eta}^\top \tilde{\eta}$ and $\tilde{\eta} = \alpha^{-1/2} \Omega^{1/2} \eta$.

By (1.6) and (3.8), the Rényi entropy of a random variable $\mathbf{Z} \sim SN_k(\mu, \Omega, \eta)$ is retrieved. Taking $\eta = \mathbf{0}$ in (3.8), the Rényi entropy of the normal distribution given by (1.8) is obtained. Lemma 2 allows the computing of the expected value of the normal cdf. Considering the standardised CSN variable in (2.18), Proposition 5 is solved by (2.19), by setting $\mathbf{v} = \mathbf{0}$ and $\mathbf{A} = I_k$, with $k = s = 1$. However, the case $\mathbf{v} \neq \mathbf{0}$ and $\mathbf{A} \neq I_k$, $k > 1$, is still an open problem and, it is useful to find the Rényi entropy for CSN distributions.

Corollary 1. *Let $\mathbf{Z} \sim SN_1(\mu, \Omega, \eta)$, $\mathbf{Z}_N \sim N_1(\mu, \Omega)$, $\|\tilde{\eta}\| = \tilde{\eta}^\top \tilde{\eta}$ and $\tilde{\eta} = \Omega^{1/2} \eta$. Then,*

(i) $R_\alpha(\mathbf{Z}) = R_\alpha(\mathbf{Z}_N) - N_\alpha(\mathbf{Z})$, $\alpha \in \mathbb{N}$, $\alpha > 1$, where

$$N_\alpha[f] = \frac{1}{\alpha - 1} \log \left[2^\alpha \frac{\Phi_{\alpha+1}(\mathbf{0}; \mathbf{0}, \tilde{\Omega})}{\Phi_1(0; 0, \omega)} \right]$$

is the so-called Negentropy, $R_\alpha(\mathbf{Z})$ is given by (1.8), and $\tilde{\Omega}$ and ω are defined as in Proposition 5.

(ii) $\lim_{\alpha \rightarrow 1} N_\alpha(\mathbf{Z}) = E\{\log[2\Phi_1(\|\tilde{\eta}\|W)]\}$.

(iii) $H(\mathbf{Z}) = H(\mathbf{Z}_N) - E\{\log[2\Phi_1(\|\tilde{\eta}\|W)]\}$,

where $H(\mathbf{Z}_N)$ is given by (1.9) and $W \sim SN_1(0, 1, \|\tilde{\eta}\|)$.

(iv) $H(\mathbf{Z}_N) - \log(4e) \leq H(\mathbf{Z}) \leq H(\mathbf{Z}_N)$, $\forall \eta$.

Contreras-Reyes and Arellano-Valle (2012) define the negentropy as the departure from normality of the distribution of \mathbf{Z} . Therefore, the SN Rényi entropy corresponds to the difference between normal Rényi entropy and negentropy, that depends on the skewness parameter η . On the another hand, by setting $\mathbf{v} = \mathbf{0}$ and $\mathbf{A} = I_k$ in (3.24) with $k = s = 1$, we obtain the property (ii) of Corollary 1. By properties (iii) and (iv), $-0.967 \leq H(\mathbf{Z}_N) - \log(4e) \leq H(\mathbf{Z})$ because, the minimum value of normal Shannon entropy is obtained for $k = 1$ and, $0 \leq E\{\ln[2\Phi_1(\|\tilde{\eta}\|W)]\} \leq 2.386$, for all η . In addition, Contreras-Reyes and Arellano-Valle (2012) reported a maximum value of this expected value equal to 2.339, using numerical approximations. Considering (1.6), (1.11) and (3.8); the complexity measure for SN distribution is obtained.

Cross-Entropy, KL and J divergences

To simplify the computation of the KL divergence, the following properties of the SN distribution are useful.

Lemma 4. *Let $\mathbf{Z} \sim SN_k(\xi, \Omega, \eta)$, and consider the vector $\delta = \Omega\eta / \sqrt{1 + \eta^\top \Omega \eta}$. Then:*

- (i) $\mathbf{Z} \stackrel{d}{=} \xi + \delta|U_0| + \mathbf{U}$, where $U_0 \sim N(0, 1)$ and $\mathbf{U} \sim N_k(\mathbf{0}, \Omega - \delta\delta^\top)$ and they are independent;
- (ii) $E(\mathbf{Z}) = \xi + \sqrt{\frac{2}{\pi}}\delta$, $\text{var}(\mathbf{Z}) = \Omega - \frac{2}{\pi}\delta\delta^\top$;
- (iii) For every vector $\mathbf{a} \in \mathbb{R}^k$ and symmetric matrix $\mathbf{B} \in \mathbb{R}^{k \times k}$,

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

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

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

The calculus of the CE $CH(\mathbf{X}, \mathbf{Y})$ when $\mathbf{Y} \sim SN_k(\xi_2, \Omega_2, \eta_2)$ and $\mathbf{X} \sim SN_k(\xi_1, \Omega_1, \eta_1)$, require of the expectation of the functions $(\mathbf{X} - \xi_2)^\top \Omega_2^{-1}(\mathbf{X} - \xi_2)$ and $\log[\Phi\{\eta_2^\top (\mathbf{X} - \xi_2)\}]$. Therefore, the properties (iii) and (iv) in Lemma 4 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 SN random variables will be considered:

$$W_{ij} \sim SN_1 \left(\eta_i^\top (\xi_j - \xi_i), \eta_i^\top \Omega_j \eta_i, \frac{\eta_i^\top \delta_j}{\sqrt{\eta_i^\top \Omega_j \eta_i - (\eta_i^\top \delta_j)^2}} \right), \quad (3.9)$$

where $\delta_j = \Omega_j \eta_j / \sqrt{1 + \eta_j^\top \Omega_j \eta_j}$ for $j = 1, 2$. Note for $i = j$ that $W_{ii} \sim SN_1(0, \eta_i^\top \Omega_i \eta_i, (\eta_i^\top \Omega_i \eta_i)^{1/2})$, with $i = 1, 2$. Also, we note that (3.9) can be expressed as

$$W_{ij} \stackrel{d}{=} \eta_i^\top (\xi_j - \xi_i) + (\eta_i^\top \Omega_j \eta_i)^{1/2} U_{ij},$$

where $U_{ij} \sim SN_1(0, 1, \tau_{ij})$, with $\tau_{ij} = \eta_i^\top \delta_j / \sqrt{\eta_i^\top \Omega_j \eta_i - (\eta_i^\top \delta_j)^2}$.

Lemma 5. *The CE between $\mathbf{X} \sim SN_k(\xi_1, \Omega_1, \eta_1)$ and $\mathbf{Y} \sim SN_k(\xi_2, \Omega_2, \eta_2)$ is given by*

$$CH(\mathbf{X}, \mathbf{Y}) = CH(\mathbf{X}_0, \mathbf{Y}_0) + \sqrt{\frac{2}{\pi}} (\xi_1 - \xi_2)^\top \Omega_2^{-1} \delta_1 - E[\log \{2\Phi(W_{21})\}],$$

where

$$CH(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2} \left\{ k \log(2\pi) + \log |\Omega_2| + \text{tr}(\Omega_2^{-1} \Omega_1) + (\xi_1 - \xi_2)^\top \Omega_2^{-1} (\xi_1 - \xi_2) \right\}$$

is the CE between $\mathbf{Y}_0 \sim SN_k(\xi_2, \Omega_2, \mathbf{0})$ and $\mathbf{X}_0 \sim SN_k(\xi_1, \Omega_1, \mathbf{0})$, and by (3.9)

$$W_{21} \sim SN_1 \left(\eta_2^\top (\xi_1 - \xi_2), \eta_2^\top \Omega_1 \eta_2, \frac{\eta_2^\top \delta_1}{\sqrt{\eta_2^\top \Omega_1 \eta_2 - (\eta_2^\top \delta_1)^2}} \right).$$

Proposition 6. *The KL divergence between $\mathbf{X} \sim SN_k(\xi_1, \Omega_1, \eta_1)$ and $\mathbf{Y} \sim SN_k(\xi_2, \Omega_2, \eta_2)$ is given by*

$$K(\mathbf{X}, \mathbf{Y}) = K(\mathbf{X}_0, \mathbf{Y}_0) + \sqrt{\frac{2}{\pi}} (\xi_1 - \xi_2)^\top \Omega_2^{-1} \delta_1 + E[\log \{2\Phi(W_{11})\}] - E[\log \{2\Phi(W_{21})\}],$$

where

$$K(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2} \left\{ \log \left(\frac{|\Omega_2|}{|\Omega_1|} \right) + \text{tr}(\Omega_2^{-1} \Omega_1) + (\xi_1 - \xi_2)^\top \Omega_2^{-1} (\xi_1 - \xi_2) - k \right\}$$

is the KL-divergence between $\mathbf{X}_0 \sim SN_k(\xi_1, \Omega_1, \mathbf{0})$ and $\mathbf{Y}_0 \sim SN_k(\xi_2, \Omega_2, \mathbf{0})$, and the W_{ij} defined as in (3.9).

The proofs of Lemma 5 and Proposition 6 are included in Appendix A. In the following proposition, we give the J divergence between two SN distributions. Its proof is immediate from (1.15) and Proposition 6.

Proposition 7. *Let $\mathbf{X} \sim SN_k(\xi_1, \Omega_1, \eta_1)$ and $\mathbf{Y} \sim SN_k(\xi_2, \Omega_2, \eta_2)$. Then,*

$$\begin{aligned} J(\mathbf{X}, \mathbf{Y}) &= J(\mathbf{X}_0, \mathbf{Y}_0) + \sqrt{\frac{2}{\pi}}(\xi_1 - \xi_2)^\top (\Omega_2^{-1} \delta_1 - \Omega_1^{-1} \delta_2) \\ &\quad + E[\log \{2\Phi(W_{11})\}] - E[\log \{2\Phi(W_{12})\}] \\ &\quad + E[\log \{2\Phi(W_{22})\}] - E[\log \{2\Phi(W_{21})\}], \end{aligned}$$

where

$$J(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2} \{ \text{tr}(\Omega_1 \Omega_2^{-1}) + \text{tr}(\Omega_1^{-1} \Omega_2) + 2(\xi_1 - \xi_2)^\top (\Omega_1^{-1} + \Omega_2^{-1})(\xi_1 - \xi_2) - 2k \}$$

is the J divergence between the normal random vectors $\mathbf{X}_0 \sim SN_k(\xi_1, \Omega_1, \mathbf{0})$ and $\mathbf{Y}_0 \sim SN_k(\xi_2, \Omega_2, \mathbf{0})$, and by (3.9) we have that

$$\begin{aligned} W_{11} &\sim SN_1 \left(0, \eta_1^\top \Omega_1 \eta_1, (\eta_1^\top \Omega_1 \eta_1)^{1/2} \right), \\ W_{21} &\sim SN_1 \left(\eta_2^\top (\xi_1 - \xi_2), \eta_2^\top \Omega_1 \eta_2, \frac{\eta_2^\top \delta_1}{\sqrt{\eta_2^\top \Omega_1 \eta_2 - (\eta_2^\top \delta_1)^2}} \right), \\ W_{12} &\sim SN_1 \left(\eta_1^\top (\xi_2 - \xi_1), \eta_1^\top \Omega_2 \eta_1, \frac{\eta_1^\top \delta_2}{\sqrt{\eta_1^\top \Omega_2 \eta_1 - (\eta_1^\top \delta_2)^2}} \right), \\ W_{22} &\sim SN_1 \left(0, \eta_2^\top \Omega_2 \eta_2, (\eta_2^\top \Omega_2 \eta_2)^{1/2} \right). \end{aligned}$$

In that follows we present the KL-divergence and J-divergence of some particular cases. We start considering that case where $\Omega_1 = \Omega_2$ and $\eta_1 = \eta_2$. Hence, the KL and J divergences compares the location vectors of two SN distributions, which is essentially equivalent to comparing their mean vectors. For this case we also have that $\delta_1 = \delta_2$, $W_{ii} \stackrel{d}{=} (\eta^\top \Omega \eta)^{1/2} W$, $W_{12} \stackrel{d}{=} -\eta^\top (\xi_1 - \xi_2) + (\eta^\top \Omega \eta)^{1/2} W$ and $W_{21} \stackrel{d}{=} \eta^\top (\xi_1 - \xi_2) + (\eta^\top \Omega \eta)^{1/2} W$, where $W \sim SN_1(0, 1, (\eta^\top \Omega \eta)^{1/2})$. With this notation, the results in Propositions 6 and 7 are simplified in this case as follows.

Corollary 2. *Let $\mathbf{X} \sim SN_k(\xi_1, \Omega, \eta)$ and $\mathbf{Y} \sim SN_k(\xi_2, \Omega, \eta)$. Then,*

$$\begin{aligned} K(\mathbf{X}, \mathbf{Y}) &= K(\mathbf{X}_0, \mathbf{Y}_0) + \sqrt{\frac{2}{\pi}}(\xi_1 - \xi_2)^\top \Omega^{-1} \delta \\ &\quad + E[\log \{2\Phi(\tau W)\}] - E[\log \{2\Phi(\gamma + \tau W)\}], \end{aligned}$$

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

where

$$K(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2} \left\{ (\xi_1 - \xi_2)^\top \Omega^{-1} (\xi_1 - \xi_2) - k + 1 \right\} \text{ and} \\ J(\mathbf{X}_0, \mathbf{Y}_0) = 2(\xi_1 - \xi_2)^\top \Omega^{-1} (\xi_1 - \xi_2)$$

are, respectively, the KL and J divergences between $\mathbf{X}_0 \sim SN_k(\xi_1, \Omega, \mathbf{0})$ and $\mathbf{Y}_0 \sim SN_k(\xi_2, \Omega, \mathbf{0})$, $\tau = (\eta^\top \Omega \eta)^{1/2}$, $\gamma = \eta^\top (\xi_1 - \xi_2)$ and $W \sim SN_1(\tau)$.

When $\xi_1 = \xi_2$ and $\eta_1 = \eta_2$, the KL and J divergence measures compares the dispersion matrices of two SN distributions, which is equivalent to compare their covariance matrices. 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 η , 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 3. Let $\mathbf{X} \sim SN_k(\xi, \Omega_1, \eta)$ and $\mathbf{Y} \sim SN_k(\xi, \Omega_2, \eta)$. Then,

$$K(\mathbf{X}, \mathbf{Y}) = K(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2} \left\{ \log \left(\frac{|\Omega_2|}{|\Omega_1|} \right) + \text{tr}(\Omega_2^{-1} \Omega_1) - k \right\}, \\ J(\mathbf{X}, \mathbf{Y}) = J(\mathbf{X}_0, \mathbf{Y}_0) = \frac{1}{2} \{ \text{tr}(\Omega_1 \Omega_2^{-1}) + \text{tr}(\Omega_1^{-1} \Omega_2) - 2k \},$$

where $\mathbf{X}_0 \sim SN_k(\xi, \Omega_1, \mathbf{0})$ and $\mathbf{Y}_0 \sim SN_k(\xi, \Omega_2, \mathbf{0})$.

Finally, if $\xi_1 = \xi_2$ and $\Omega_1 = \Omega_2$, then the KL divergence and J divergence compares the skewness vectors of two SN distributions, which again is equivalent to comparing their mean vectors.

Corollary 4. Let $\mathbf{X} \sim SN_k(\xi, \Omega, \eta_1)$ and $\mathbf{Y} \sim SN_k(\xi, \Omega, \eta_2)$. Then,

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

where

$$W_{11} \sim SN_1 \left(0, \eta_1^\top \Omega \eta_1, (\eta_1^\top \Omega \eta_1)^{1/2} \right),$$

$$\begin{aligned}
W_{21} &\sim SN_1 \left(0, \eta_2^\top \Omega \eta_2, \frac{\eta_2^\top \delta_1}{\sqrt{\eta_2^\top \Omega \eta_2 - (\eta_2^\top \delta_1)^2}} \right), \\
W_{12} &\sim SN_1 \left(0, \eta_1^\top \Omega \eta_1, \frac{\eta_1^\top \delta_2}{\sqrt{\eta_1^\top \Omega \eta_1 - (\eta_1^\top \delta_2)^2}} \right), \\
W_{22} &\sim SN_1 \left(0, \eta_2^\top \Omega \eta_2, (\eta_2^\top \Omega \eta_2)^{1/2} \right).
\end{aligned}$$

J divergence between the multivariate skew-normal and normal distributions

By letting $\eta_1 = \eta$ and $\eta_2 = \mathbf{0}$ in Proposition 7, we have the J-divergence between a multivariate skew-normal and normal distributions, $J(\mathbf{X}, \mathbf{Y}_0)$ say, where $\mathbf{X} \sim SN_k(\xi, \Omega, \eta)$ and $\mathbf{Y}_0 \sim SN_k(\xi, \Omega, \mathbf{0})$. For this important special case, we find in Corollary 4 that $J(\mathbf{X}_0, \mathbf{Y}_0) = 0$, the random variable W_{21} and W_{22} are degenerate at zero, and $W_{11} = (\eta^\top \Omega \eta)^{1/2} W$, with $W \sim SN_1(0, 1, (\eta^\top \Omega \eta)^{1/2})$, and $W_{12} = (\eta^\top \Omega \eta)^{1/2} W_0$, with $W_0 \sim N_1(0, 1)$. This proves the following results.

Corollary 5. *Let $\mathbf{X} \sim SN_k(\xi, \Omega, \eta)$ and $\mathbf{Y}_0 \sim SN_k(\xi, \Omega, \mathbf{0})$. Then,*

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

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

It follows from Corollary 5 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

$$J(\mathbf{X}, \mathbf{Y}_0) = E[\{2\Phi(\tau W_0) - 1\} \log \{2\Phi(\tau W_0)\}]. \quad (3.10)$$

It is interesting to notice that for $\tau = 1$ in (3.10) 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 skew-distributed.

Remark 3. *If $Z \sim SN(\xi, \omega^2, \alpha)$, $\omega > 0$, $\alpha \in \mathbb{R}$, then*

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

$$+E[\Phi(\alpha S_0) \log\{2\Phi(\omega S_0 + \xi)\}],$$

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 $\log[2\Phi(\omega z + \xi)]$ is negative on $(-\infty, 0)$ and non-negative on $[0, \infty)$, the last expression is more convenient for the numerical integration.

Jensen-Shannon distance

The next Proposition 8 allows to compute the Jensen–Shannon distance between two SN random variables.

Proposition 8. *Let $X \sim SN_1(\eta_1)$ and $Y \sim SN_1(\eta_2)$ be two SN random variables. Then, the JS distance between X and Y is*

$$JS(X, Y) = \frac{1}{2}E \left[\log \left\{ \frac{2\Phi(\eta_1 x)}{\Phi(\eta_1 x) + \Phi(\eta_2 x)} \right\} \right] + \frac{1}{2}E \left[\log \left\{ \frac{2\Phi(\eta_1 y)}{\Phi(\eta_1 y) + \Phi(\eta_2 y)} \right\} \right].$$

3.2.2 Finite mixture of multivariate skew-normal distributions

Mixture models are in high demand for machine-learning analysis, due to their computational tractability and for offering a good approximation for continuous densities McLachlan and Peel (2000). In addition, mixture models are an important statistical tool for many applications in clustering (Celeux and Soromenho, 1996, Jenssen et al., 2003), discriminant analysis Amoud et al. (2007), image processing and satellite imaging (Caillol et al., 1997, Carreira-Perpinán, 2000). Celeux and Soromenho (1996) consider a Maximum Likelihood (ML)–based entropy criterion to estimate the number of clusters arising from a mixture model, and compare it with the classical AIC and BIC information criteria. Carreira-Perpinán (2000) deal with the problem of finding all modes of multi-dimensional data, assuming a mixture of normal densities. Specifically, he uses the gradient as a mode locator and for controlling the significance modes thus obtained, by measuring the sparseness of the densities mixture via the entropy. However, so far no analytical expressions, which consider bounds of Shannon entropy for the normal mixture entropy exist. Similarly, in the case of KL divergence an analytic evaluation of the differential entropy is impossible, too. Thus, approximate calculations become inevitable (Durrieu et al., 2012, Huber et al., 2008, Nielsen and Sun, 2016). Jenssen et al. (2003) use the Rényi entropy (Rényi, 1970) as a similarity measure between clusters. They consider the Parzen window density estimation for differential Rényi entropy clustering to identify the *worst* cluster and subsequently reduce the overall number of clusters by one.

Predominantly, the entropy applications mentioned above have been developed in the normal context, but several results of both Shannon and Rényi entropies for various multivariate distributions (see e.g. Contreras-Reyes, 2015, Zografos and Nadarajah, 2005) actually exist. Here, we consider the novel class of finite mixture of multivariate skew-normal mixture (FMSN) models (Frühwirth-Schnatter and Pyne, 2010). This class provides some advantages over the normal mixtures. For instance, the normal components allow an arbitrarily close modeling of any distribution by increasing the number of components and, in the context of supervised learning, groups of observations represented by asymmetrically distributed data can lead to wrong classification (Frühwirth-Schnatter and Pyne, 2010). The components of skew-normal mixture models however capture skewness due to their flexibility. Specifically, we present practical results of upper and lower bounds of Shannon and Rényi entropies for FMSN distributions in Section 3.2.2. Section 3.2.2 also presents theoretical results of upper and lower bounds of these concepts. For the sake of simplicity, we denote the Shannon entropy and negentropy of a FMSN variable $(\mathbf{Y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})$ as $H[\mathbf{Y}; \tilde{\boldsymbol{\theta}}]$ and $N[\mathbf{Y}; \tilde{\boldsymbol{\theta}}]$, respectively.

Shannon entropy bounds

As in the normal case, the Shannon entropy of mixture of SN distributions does not have a closed form. However, the following proposition presents some lower and upper bounds as an approximation of the entropy of finite mixture of SN densities.

Proposition 9. *Consider the FMSN density of $(\mathbf{Y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})$ defined in (2.9). Then, the following inequalities are accomplished:*

$$(i) \quad A_{lower} \leq H[\mathbf{Y}; \tilde{\boldsymbol{\theta}}] \leq A_{upper},$$

$$(ii) \quad B_{lower} \leq H[\mathbf{Y}; \tilde{\boldsymbol{\theta}}] \leq B_{upper},$$

where

$$\begin{aligned} A_{upper} &= \frac{1}{2} \ln\{(2\pi e)^k |\Sigma|\}, \\ A_{lower} &= A_{upper} - \sum_{i=1}^m \pi_i N[\mathbf{Y}; \boldsymbol{\theta}_i], \\ B_{upper} &= A_{lower} - \sum_{i=1}^m \pi_i \ln \pi_i, \\ B_{lower} &= - \sum_{i=1}^m \pi_i \ln \left(\sum_{s=1}^m \pi_s \int_{-\infty}^{\infty} f(t; \boldsymbol{\eta}_i) f(t; \boldsymbol{\eta}_s) dt \right), \end{aligned}$$

with $N[\mathbf{Y}; \boldsymbol{\theta}_i] = E [\ln\{2\Phi_1(\|\tilde{\boldsymbol{\eta}}_i\|W_i)\}] = \int_{-\infty}^{\infty} f(w_i; \|\tilde{\boldsymbol{\eta}}_i\|) \ln\{2\Phi_1(\|\tilde{\boldsymbol{\eta}}_i\|w_i)\} dw_i$, $W_i \sim SN_1(\|\tilde{\boldsymbol{\eta}}_i\|)$, $\|\tilde{\boldsymbol{\eta}}_i\| = \boldsymbol{\eta}_i^\top \boldsymbol{\Omega}_i \boldsymbol{\eta}_i$, and $\Sigma = \text{Var}[\mathbf{Y}]$ is defined by (2.12).

For the case $m = 1$, Contreras-Reyes and Arellano-Valle (2012) consider the upper bound of the property (i) of Proposition 9 to approximate the Shannon entropy of an SN distribution using the property (ii) of Proposition 9. In this Proposition 9(ii), the left side includes an integral related to a product of two SN densities. When $i = s$, these integrals correspond to an L^2 -norm and are represented by the quadratic Rényi entropy (Contreras-Reyes, 2015). For the case $i \neq s$, the integral does not have explicit form and requires numerical methods to be computed. Moreover, the right side corresponds to the sum of the entropy of a multinomial density with parameters π_1, \dots, π_m and a second term based on the weights and shape parameters of the SN density. Other refinements can be found in Huber et al. (2008). These are suitable for cases of several components (for example, $m \gg 5$), i.e., a SN density consisting of several and well separated clusters.

A lower bound can be found for $H[\mathbf{Y}; \tilde{\boldsymbol{\theta}}]$ using the L^2 -norm of an FMSN density and Jensen's inequality (Cover and Thomas, 2006):

$$H[\mathbf{Y}; \tilde{\boldsymbol{\theta}}] \geq -2 \ln \|p(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})\|_2 = -\ln \int_{\mathbb{R}^k} p(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})^2 d\mathbf{y}. \quad (3.11)$$

Considering (1.6), (3.11) and Proposition 2(i)–(ii), we obtain the additional inequalities $R_2[\mathbf{Y}; \tilde{\boldsymbol{\theta}}] \leq H[\mathbf{Y}; \tilde{\boldsymbol{\theta}}]$ and

$$B_{\text{lower}} < A_{\text{lower}} < B_{\text{upper}} < A_{\text{upper}}. \quad (3.12)$$

The next section provides upper and lower bounds for Rényi entropy of FMSN random vectors.

Rényi entropy bounds

For the sake of simplicity, we define the following function in terms of Rényi entropy and α as

$$P_\alpha[\mathbf{Y}; \tilde{\boldsymbol{\theta}}] = e^{(1-\alpha)R_\alpha[\mathbf{Y}; \tilde{\boldsymbol{\theta}}]}, \quad 0 < \alpha < \infty, \alpha \neq 1,$$

for the calculus of the bounds of $\int_{\mathbb{R}^k} p(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})^\alpha d\mathbf{y}$. By applying the function $\ln(\cdot)/(1-\alpha)$ to these integrals, we have the Rényi entropy of FMSN density.

As in the Shannon entropy case, the Rényi entropy can be upper bounded in terms of the dispersion matrix of the finite mixture random variable. Sánchez-Moreno et al. (2011)

derived a multidimensional upper bound using a variational approach,

$$R_\alpha[\mathbf{Y}; \tilde{\boldsymbol{\theta}}] \leq \frac{d}{2} \ln \left(\frac{|\Lambda|}{d} \right) + F_k(\alpha), \quad (3.13)$$

with $\Lambda = \text{Var}[\mathbf{Y}]$ defined in (2.12),

$$F_k(\alpha) = \frac{k}{2} \ln \left(\frac{\pi b}{\alpha - 1} \right) + \frac{1}{\alpha - 1} \ln \left(\frac{b}{2\alpha} \right) + \ln \Gamma \left(\frac{\alpha}{\alpha - 1} \right) - \ln \Gamma \left(\frac{b}{2(\alpha - 1)} \right),$$

$b = (2 + k)\alpha - k$, $\boldsymbol{\theta}_0 = (\mathbf{0}, I_k)$, and $\mathbf{W}_0 \sim N_k(\mathbf{0}, I_k)$. $H[\mathbf{W}_0; \boldsymbol{\theta}_0]$ is obtained using the Proposition 3.

The right side of the inequality (3.13) is equivalent to the maximum Shannon entropy of Proposition 9. The first term depends on the dispersion matrix and the shape parameters, and the second only depends on the α th order and dimension k .

The next Lemma presents a useful result to compute the lower bound for Rényi entropy of an FMSN random vector \mathbf{Y} in terms of each component.

Lemma 6. *Consider the FMSN density of $(\mathbf{Y}; \boldsymbol{\theta}, \pi)$ defined in (2.9). Then,*

$$\int_{\mathbb{R}^k} p(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \pi)^\alpha d\mathbf{y} \geq P_\alpha[\mathbf{Y}; \boldsymbol{\theta}_m] + \sum_{i=1}^{m-1} \left\{ \left(\sum_{k=1}^i \pi_k \right)^\alpha \left(P_\alpha[\mathbf{Y}; \boldsymbol{\theta}_i] - P_\alpha[\mathbf{Y}; \boldsymbol{\theta}_{i+1}] \right) \right\},$$

with $0 < \alpha < \infty$, $\alpha \neq 1$, and $m > 1$.

Some theoretical results

In this section, we develop some bounds and asymptotic approximation of Rényi entropy for FMSN densities. Considering the multinomial (MN) theorem (see e.g. Bennett, 1986) for m sums, the equality

$$\begin{aligned} \int_{\mathbb{R}^k} p(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \pi)^\alpha d\mathbf{y} &= \int_{\mathbb{R}^k} \left[\sum_{i=1}^m \pi_i f(\mathbf{y}; \boldsymbol{\theta}_i) \right]^\alpha d\mathbf{y} \\ &\stackrel{(\text{MN})}{=} \int_{\mathbb{R}^k} \sum_{k_i \in \mathcal{A}} \frac{\alpha!}{k_1! \cdots k_m!} \prod_{i=1}^m [\pi_i f(\mathbf{y}; \boldsymbol{\theta}_i)]^{k_i} d\mathbf{y} \end{aligned} \quad (3.14)$$

is accomplished under the condition

$$\sum_{k_i \in \mathcal{A}} \frac{\alpha!}{k_1! \cdots k_m!} = m^\alpha, \quad (3.15)$$

with $0 < \alpha < \infty$, $\alpha \neq 1$, and $\mathcal{A} = \{k_i \in \mathbb{N} : k_i > 0, \sum_{i=1}^m k_i = \alpha; i = 1, \dots, m\}$. By choosing $\alpha_i = k_i/\alpha$, $\sum_{i=1}^m \alpha_i = \alpha^{-1} \sum_{i=1}^m k_i = 1$. Then, by applying Stirling's approximation given by Jaynes (1982) in (3.14), the equality

$$\frac{1}{\alpha} \ln U_\alpha(\tilde{\alpha}|Q) = H(\tilde{\alpha}|Q) + \mathcal{O}(\alpha^{-1} \ln \alpha), \quad (3.16)$$

is derived with

$$U_\alpha(\tilde{\alpha}|Q) = \frac{\alpha!}{k_1! \cdots k_m!} \prod_{i=1}^m Q_i(\mathbf{y})^{k_i}, \quad H(\tilde{\alpha}|Q) = - \sum_{i=1}^m \alpha_i \ln \left(\frac{\alpha_i}{Q_i(\mathbf{y})} \right),$$

$\tilde{\alpha} = (\alpha_1, \dots, \alpha_m)$, and $Q_i(\mathbf{y}) = \pi_i f(\mathbf{y}; \boldsymbol{\theta}_i)$.

Lemma 7. *Considering the condition (3.15), the approximation*

$$\int_{\mathbb{R}^k} p(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})^\alpha d\mathbf{y} \approx \sum_{k_i \in \mathcal{A}} \left[\prod_{i=1}^m \left(\frac{\pi_i}{\alpha_i} \right)^{k_i} \right] P_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]$$

is accomplished as $\alpha \rightarrow \infty$.

The next Lemma presents two upper bounds for Rényi entropy of an FMSN random vector \mathbf{Y} in terms of each m -component using the multinomial theorem, given that the components and weights of the mixture models are non-negative.

Lemma 8. *Consider the FMSN density of $(\mathbf{Y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})$ defined in (2.9). Then, the inequalities*

$$(i) \int_{\mathbb{R}^k} p(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})^\alpha d\mathbf{y} \leq \sum_{k_i \in \mathcal{A}} \frac{\alpha!}{k_1! \cdots k_m!} \left(\prod_{i=1}^m \pi_i^{k_i} \right) \exp \left(\frac{1-\alpha}{m} \sum_{i=1}^m R_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i] \right)$$

$$(ii) \int_{\mathbb{R}^k} p(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi})^\alpha d\mathbf{y} \leq \sum_{k_i \in \mathcal{A}} \frac{\alpha!}{k_1! \cdots k_m!} \left(\prod_{i=1}^m \pi_i^{k_i} \right) \left(\frac{1}{m} \sum_{i=1}^m P_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]^{k_i} \right)^{m/\alpha}$$

are accomplished under the condition (3.15).

Given the condition $0 < \alpha < \infty$, $\alpha \neq 1$, the multinomial coefficients k_i can not equal zero, i.e., $\sum_{i=1}^m k_i = \alpha > 0$. Then, the Rényi entropies $R_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]$ exist and can be obtained using the property (ii) of Proposition 1 for $\alpha \in \mathbb{N}$, $\alpha > 1$. This means that the inequalities of Lemma 3 consider m individual Rényi entropies with integer order, whereas inequality (3.13) considers orders with real values. In addition, Lemma 3 allow to determine an upper bound for FMSN Rényi entropy in terms of $R_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]$ and $P_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]$, $i = 1, \dots, m$; respectively.

After the inequalities of Lemma 6 and 8 are derived, it is natural to look for an underlying identity. We consider Proposition 3 of Bennett (1986) based on Abel's identity and multinomial theorem. Let \mathcal{B} be a set of multinomial coefficients defined as $\mathcal{B} = \{k_s \in \mathbb{N} : 0 < k_s < \alpha, \sum_{i=1}^m k_i = \alpha, k_{s+1} = \dots = k_m = 0\}$. We have

$$\begin{aligned} p(\mathbf{y}_j; \boldsymbol{\theta}, \boldsymbol{\pi})^\alpha &= \left[\sum_{i=1}^m \pi_i f(\mathbf{y}_j; \boldsymbol{\theta}_i) \right]^\alpha \\ &= \left(\sum_{i=1}^m \pi_i \right)^\alpha f(\mathbf{y}_j; \boldsymbol{\theta}_m)^\alpha + \sum_{i=1}^m \left\{ \left(\sum_{k=1}^i \pi_k \right)^\alpha \left(f(\mathbf{y}_j; \boldsymbol{\theta}_i)^\alpha - f(\mathbf{y}_j; \boldsymbol{\theta}_{i+1})^\alpha \right) \right. \\ &\quad \left. \sum_{k_s \in \mathcal{B}} \frac{\alpha!}{k_1! \dots k_s!} \left(\prod_{s=1}^i \pi_s^{k_s} \right) \left[\left(\prod_{s=1}^{i-1} f(\mathbf{y}_j; \boldsymbol{\theta}_s)^{k_s} \right) - f(\mathbf{y}_j; \boldsymbol{\theta}_i)^{\alpha-k_i} \right] \right\}. \end{aligned}$$

By integrating the last equality in both sides and applying the condition $\sum_{i=1}^m \pi_i = 1$, we obtain

$$\begin{aligned} \int_{\mathbb{R}^k} p(\mathbf{y}_j; \boldsymbol{\theta}, \boldsymbol{\pi})^\alpha d\mathbf{y}_j &= P_\alpha[\mathbf{Y}_j; \boldsymbol{\theta}_m] + \sum_{i=1}^m \left\{ \left(\sum_{s=1}^i \pi_s \right)^\alpha \left(P_\alpha[\mathbf{Y}_j; \boldsymbol{\theta}_i] - P_\alpha[\mathbf{Y}_j; \boldsymbol{\theta}_{i+1}] \right) \right. \\ &\quad \left. \sum_{k_s \in \mathcal{B}} \frac{\alpha!}{k_1! \dots k_s!} \left(\prod_{s=1}^i \pi_s^{k_s} \right) \left[\int_{\mathbb{R}^k} \left(\prod_{s=1}^{i-1} f(\mathbf{y}_j; \boldsymbol{\theta}_s)^{k_s} \right) d\mathbf{y}_j \right. \right. \\ &\quad \left. \left. - P_{\alpha-k_i}[\mathbf{Y}_j; \boldsymbol{\theta}_i] \right] \right\}, \end{aligned} \quad (3.17)$$

with the multinomial coefficients of \mathcal{B} satisfying

$$\sum_{k_s \in \mathcal{B}} \frac{\alpha!}{k_1! \dots k_{i-1}!} = (i-1)^\alpha, \quad i = 1, \dots, m. \quad (3.18)$$

The equality (3.17) is difficult to compute, given that it does not have a more explicit expression for the integral of products of the mixture densities. However, an alternative inequality can be found using the Generalized Hölder (GH) inequality and the proof of Lemma 6, with $p_s = (\alpha - k_i)/k_s$, $s = 1, \dots, i-1$, $i = 1, \dots, m$. This ensures the assumptions of p_s given the condition (3.18). In contrast to condition (3.15), condition (3.18) corresponds to the sum of $i-1$ multinomial coefficients k_s , i.e., the sum is always less than m^α of (3.15) for any index $i = 1, \dots, m$.

3.2.3 Multivariate skew- t distribution

Shannon entropy and mutual information

Let $\mathbf{Z}_0 \sim T_k(\mathbf{0}, I_k, \nu)$ and $\mathbf{Z} \sim ST_k(\xi, \Omega, \eta, \nu)$, we obtain directly from Proposition 1 that

$$H(\mathbf{Z}) = H(\mathbf{Z}_0) - E \left[\log \left\{ 2T \left(\sqrt{\frac{\nu+k}{\nu + \|\mathbf{Z}_0\|^2}} \bar{\eta}^\top \mathbf{Z}_0; \nu+k \right) \right\} \right], \quad (3.19)$$

where $H(\mathbf{Z}_0)$ is given by formula (3.4). In order to compute the last factor in the skew- t (ST) entropy (3.19) by integration in only one dimension, we need the following result whose proof is given by Arellano-Valle (2010).

Lemma 9. *Let $\mathbf{Z}_0 \sim ST_k(\mathbf{0}, I_k, \bar{\eta}, \nu)$. Then:*

$$\sqrt{\frac{\nu+k}{\nu + \|\mathbf{Z}_0\|^2}} \bar{\eta}^\top \mathbf{Z}_0 \stackrel{d}{=} \frac{\sqrt{\nu+k} \|\bar{\eta}\| W_{ST}}{\sqrt{\nu+k-1 + W_{ST}^2}}$$

where $W_{ST} \sim ST_1(0, 1, \|\bar{\eta}\|, \nu+k-1)$.

Since for large values of ν the multivariate Student's t , and hence the ST, distributions converge to the normal and SN ones, respectively, it is straightforward to see from (3.19) that the ST entropy converges to SN entropy for any values of $\bar{\eta}$ as $\nu \rightarrow \infty$. The behavior of this convergence and the respective entropies are simulated/reported for several values of $\alpha = \|\bar{\eta}\|$ and ν in Section 4.1.2.

Considering the Jensen's inequality (Cover and Thomas, 2006, pp. 27), is possible to found a lower bound for ST entropy. Since the ST density function is log-convex (Domínguez-Molina and Rocha-Arteaga, 2007), we have

$$H(\mathbf{Z}) = E[\log f_{\mathbf{Z}}(\mathbf{z})] \geq -\log E[f_{\mathbf{Z}}(\mathbf{z})] = -2 \log \|f_{\mathbf{Z}}(\mathbf{z})\|_2,$$

where $\|\cdot\|_2$ denotes the L_2 -norm. However, if \mathbf{Z} is ST distributed, in order to found $\|f_{\mathbf{Z}}(\mathbf{z})\|_2$, additional characterizations of the Shannon entropy as the Rényi entropy are necessary.

For any density $f_{\mathbf{X}}(\mathbf{x})$ of a random vector $\mathbf{X} \in \mathbb{R}^k$ not necessary normal with zero mean and variance-covariance matrix $\Delta = E[\mathbf{X}\mathbf{X}^\top]$, the entropy of \mathbf{X} is upper bounded as

$$H(\mathbf{X}) \leq \frac{1}{2} \log \left\{ (2\pi e)^k |\Delta| \right\}. \quad (3.20)$$

This means that the normal distribution maximises the entropy over the ST distribution with the same variance (Cover and Thomas, 2006, pp. 663). For the ST case, by (3.20) and property (i) of Lemma 10 (see below) we have that

$$H(\mathbf{Z}) \leq \frac{1}{2} \log\{(2\pi e)^k\} + \frac{1}{2} \log \left| K_2(\mathbf{Z})\Omega + \xi \xi^\top + K_1(\mathbf{Z}) \left(\xi \delta^\top + \delta \xi^\top \right) \right|.$$

The following results are related with the marginal distributions from a multivariate ST distribution. For the proof of the latter, see Arellano-Valle and Genton (2010). Let

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim ST_{n+m} \left(\begin{pmatrix} \xi_{\mathbf{X}} \\ \xi_{\mathbf{Y}} \end{pmatrix}, \begin{pmatrix} \Omega_{\mathbf{XX}} & \Omega_{\mathbf{XY}} \\ \Omega_{\mathbf{YX}} & \Omega_{\mathbf{YY}} \end{pmatrix}, \begin{pmatrix} \eta_{\mathbf{X}} \\ \eta_{\mathbf{Y}} \end{pmatrix}, \nu \right).$$

Then $\mathbf{X} \sim ST_n(\xi_{\mathbf{X}}, \Omega_{\mathbf{XX}}, \eta_{\mathbf{X}(\mathbf{Y})}, \nu)$ and $\mathbf{Y} \sim ST_m(\xi_{\mathbf{Y}}, \Omega_{\mathbf{YY}}, \eta_{\mathbf{Y}(\mathbf{X})}, \nu)$. By Lemma 3 we have

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

where $W_{ST} \sim ST_1(0, 1, \|\bar{\eta}\|, \nu+k-1)$. So, considering the above results, we can deduce the mutual information index for the ST case as follows.

Proposition 10. *Let*

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim ST_{n+m} \left(\begin{pmatrix} \xi_{\mathbf{X}} \\ \xi_{\mathbf{Y}} \end{pmatrix}, \begin{pmatrix} \Omega_{\mathbf{XX}} & \Omega_{\mathbf{XY}} \\ \Omega_{\mathbf{YX}} & \Omega_{\mathbf{YY}} \end{pmatrix}, \begin{pmatrix} \eta_{\mathbf{X}} \\ \eta_{\mathbf{Y}} \end{pmatrix}, \nu \right).$$

Then:

$$I(\mathbf{X}, \mathbf{Y}) = I(\mathbf{X}_0, \mathbf{Y}_0) + E \left\{ \log \left(\frac{C_{\mathbf{XY}}}{C_{\mathbf{X}(\mathbf{Y})} C_{\mathbf{Y}(\mathbf{X})}} \right) \right\},$$

where $\mathbf{X}_0 \sim T_n(\xi_{\mathbf{X}}, \Omega_{\mathbf{XX}}, \nu)$, $\mathbf{Y}_0 \sim T_m(\xi_{\mathbf{Y}}, \Omega_{\mathbf{YY}}, \nu)$,

$$\begin{aligned} C_{\mathbf{XY}} &= 2T \left(\frac{\sqrt{\nu+n+m}\|\bar{\eta}_{\mathbf{XY}}\| W_{ST}}{\sqrt{\nu+n+m-1+W_{ST}^2}}; \nu+n+m \right), \\ C_{\mathbf{X}(\mathbf{Y})} &= 2T \left(\frac{\sqrt{\nu+n}\|\bar{\eta}_{\mathbf{X}(\mathbf{Y})}\| W_{ST}}{\sqrt{\nu+n-1+W_{ST}^2}}; \nu+n \right), \\ C_{\mathbf{Y}(\mathbf{X})} &= 2T \left(\frac{\sqrt{\nu+m}\|\bar{\eta}_{\mathbf{Y}(\mathbf{X})}\| W_{ST}}{\sqrt{\nu+m-1+W_{ST}^2}}; \nu+m \right), \end{aligned}$$

$W_{ST} \sim ST_1(0, 1, \|\tilde{\eta}\|, \nu + k - 1)$, and $I(\mathbf{X}_0, \mathbf{Y}_0)$ is given in (3.5).

Asymptotic cross-entropy and Kullback–Leibler divergence

We are interested in compute the CE for $\mathbf{X} \sim ST_k(\xi_1, \Omega_1, \eta_1, \nu_1)$ with respect to $\mathbf{Y} \sim ST_k(\xi_2, \Omega_2, \eta_2, \nu_2)$. The following two lemmas gives a useful properties for the calculus of the CE between two ST variables, for the posteriorly calculus of KL divergence measure.

Lemma 10. *Let $\mathbf{Z} \sim ST_k(\xi, \Omega, \eta, \nu)$ and $\delta = \Omega\eta / \sqrt{1 + \eta^\top \Omega \eta}$. Then:*

- (i) $E[\mathbf{Z}] = \xi + K_1(\mathbf{Z})\delta$, for $\nu > 1$,
 $Var[\mathbf{Z}] = K_2(\mathbf{Z})\Omega + \xi\xi^\top + K_1(\mathbf{Z})\left(\xi\delta^\top + \delta\xi^\top\right)$, for $\nu > 2$,
 with $K_1(\mathbf{Z}) = \sqrt{\nu/2}\Gamma(\frac{\nu-1}{2})/\Gamma(\frac{\nu}{2})$ and $K_2(\mathbf{Z}) = \frac{\nu}{\nu-2}$.
- (ii) Let $\mathbf{X} \sim ST_k(\xi_1, \Omega_1, \eta_1, \nu_1)$, $\mathbf{Y} \sim ST_k(\xi_2, \Omega_2, \eta_2, \nu_2)$, and $\mathbf{Q}_2 = (\mathbf{x} - \xi_2)^\top \Omega_2^{-1}(\mathbf{x} - \xi_2)$. Then

$$\begin{aligned} E[\mathbf{Q}_2] &= K_2(\mathbf{X})\text{tr}(\Omega_2^{-1}\Omega_1) + (\xi_1 - \xi_2)^\top \Omega_2^{-1}(\xi_1 - \xi_2) \\ &\quad + 2K_1(\mathbf{X})(\xi_1 - \xi_2)^\top \Omega_2^{-1}\delta_1, \end{aligned}$$

where $\text{tr}(A)$ denotes the trace of the matrix A and $\delta_1 = \Omega_1\eta_1 / \sqrt{1 + \eta_1^\top \Omega_1 \eta_1}$.

- (iii) For every vectors $\tilde{\eta}, \tilde{\xi} \in \mathbb{R}^k$,

$$\tilde{\eta}^\top (\mathbf{Z} - \tilde{\xi}) \sim ST_1\left(\tilde{\eta}^\top (\xi - \tilde{\xi}), \tilde{\eta}^\top \Omega \tilde{\eta}, \frac{\tilde{\eta}^\top \delta}{\sqrt{\tilde{\eta}^\top \Omega \tilde{\eta} - (\tilde{\eta}^\top \delta)^2}}, \nu\right).$$

The expected value of the quadratic form of Lemma 10(ii), is a very important property of quadratic forms of ST distributions where it allows the computing of the CE between \mathbf{X} and \mathbf{Y} . The following Lemma 11 provides the CE between two ST variables to obtain the KL divergence.

Lemma 11. *Let $\mathbf{X} \sim ST_k(\xi_1, \Omega_1, \eta_1, \nu_1)$ and $\mathbf{Y} \sim ST_k(\xi_2, \Omega_2, \eta_2, \nu_2)$. Then:*

$$CH(\mathbf{X}, \mathbf{Y}) \approx \frac{1}{2} \log\{(2\pi)^k |\Omega_2|\} + \frac{1}{2} \left(\frac{\nu_2 + k}{\nu_2} \right) E[\mathbf{Q}_2] - E \left[\log \left\{ 2T \left(\tilde{\eta}_2^\top (\mathbf{x} - \xi_2); \nu_2 + k \right) \right\} \right], \quad (3.21)$$

as $v_2 \rightarrow \infty$, where $\tilde{\eta}_2^\top = \eta_2^\top \sqrt{\frac{v_2+k}{v_2+Q_2}}$, and $E[\mathbf{Q}_2]$ is given by Lemma 10(ii).

The expected value of (3.21) is not directly computable because $\mathbf{x} \in \mathbb{R}^k$, i.e., the computing of a multivariate integral is required. However, by replacing the terms $\tilde{\xi} = \xi_2$, $v = v_2$, and $\tilde{\eta} = \tilde{\eta}_2$ in property (iii) of Lemma 10, it is possible to integrate the last term of (3.21) in one dimension. Therefore, the asymptotic KL divergence for ST distributions is obtained by replacing the formulas (3.4), (3.19) and (3.21) in the identity (1.14). This measure consider an asymptotic behavior for $CH(\mathbf{X}, \mathbf{Y})$ as $v_2 \rightarrow \infty$. Then, the density $f_{\mathbf{Y}}$ tend to be a SN density and, $K(\mathbf{X}, \mathbf{Y})$ represents the divergence of a SN density $f_{\mathbf{Y}}$ from a ST density $f_{\mathbf{X}}$, for large values of v_2 and; a ST density $f_{\mathbf{Y}}$ from a ST density $f_{\mathbf{X}}$, for small values of v_2 . The entropy $H(\mathbf{X})$ given by (3.19) has an explicit form and only depends on Ω_1 and v_1 .

Given that the t distribution is a special case of ST distribution when $\eta = 0$ in (2.13), is possible to obtain an asymptotic CE between two t distributions as follows. Taking $\eta_1 = \eta_2 = 0$ in Lemma 10, we have $\mathbf{X} \sim T_k(\xi_1, \Omega_1, v_1)$ and $\mathbf{Y} \sim T_k(\xi_2, \Omega_2, v_2)$. Then:

$$CH(\mathbf{X}, \mathbf{Y}) \approx \frac{1}{2} \log\{(2\pi)^k |\Omega_2|\} + \frac{1}{2} \left(\frac{v_2+k}{v_2} \right) E[\mathbf{Q}_2], \quad (3.22)$$

as $v_2 \rightarrow \infty$, $v_1 > 2$, where $E[\mathbf{Q}_2] = K_2(\mathbf{X}) \text{tr}(\Omega_2^{-1} \Omega_1) + (\xi_1 - \xi_2)^\top \Omega_2^{-1} (\xi_1 - \xi_2)$. If $\xi_1 = \xi_2$; (1.14), (3.4) and (3.22) yields

$$\begin{aligned} K(\mathbf{X}, \mathbf{Y}) &\approx \frac{1}{2} \log \left(\frac{|\Omega_2|}{|\Omega_1|} \right) + \frac{1}{2} \left(\frac{v_2+k}{v_2} \right) \left(\frac{v_1}{v_1-2} \right) \text{tr}(\Omega_2^{-1} \Omega_1) \\ &\quad - \left(\frac{v_1+k}{2} \right) \left[\psi \left(\frac{v_1+k}{2} \right) - \psi \left(\frac{v_1}{2} \right) \right]. \end{aligned} \quad (3.23)$$

Exact expressions of KL and J divergences between a ST and t distribution has been developed in Corollary 2.1 and Proposition 2.3 of Godoi et al. (2017), respectively.

3.3 Information measures for multivariate closed skew-normal distributions

By (1.6) and (2.18), the Shannon entropy for CSN distributions is rewritten as

$$\begin{aligned} H(\mathbf{Z}) &= -E\{\log[f_{k,s}(\mathbf{Y})]\} \\ &= \frac{1}{2} \log |\Omega| - \log [\Phi_s(\mathbf{0}; v, \mathbf{A} + \mathbf{D}^\top \Omega \mathbf{D})] - E\{\ln[\phi_k(\mathbf{Z}_0) \Phi_s(\tilde{\mathbf{D}}^\top \mathbf{Z}_0; v, \mathbf{A})]\} \\ &= H(\mathbf{Z}_0) - \log [\Phi_s(\mathbf{0}; v, \mathbf{A} + \mathbf{D}^\top \Omega \mathbf{D})] - E\{\log [\Phi_s(\tilde{\mathbf{D}}^\top \mathbf{Z}_0; v, \mathbf{A})]\}, \end{aligned} \quad (3.24)$$

where \mathbf{Z}_0 is the standardised normal random variable and $H(\mathbf{Z}_0) = (1/2) \ln(2\pi e)$.

3.3.1 Multivariate skew-normal distribution

See Section 3.2.1.

3.3.2 Multivariate extended skew-normal distribution

The next Proposition 11 allows to compute the Rényi entropy and complexity measure of a ESN random variable.

Proposition 11. *Let \mathbf{Z} be a $ESN_k(\mu, \Omega, \eta, \tau)$, $\mathbf{z} \in \mathbb{R}^k$. Then:*

$$\int_{\mathbb{R}^k} [f(\mathbf{z})]^\alpha d\mathbf{z} = \psi_{\alpha,d}(\Omega) E \left\{ \left[\frac{\Phi_1(W)}{2\Phi_1(\tau)} \right]^\alpha \right\}, \quad \alpha \in \mathbb{N}, \alpha > 1, \quad (3.25)$$

where $\psi_{\alpha,d}(\Omega)$ is defined as in Proposition 5 and $W = \tilde{\eta}^\top \mathbf{Z}_0 + \tilde{\tau} \sim ESN_1(\tilde{\tau}, \|\tilde{\eta}\|^2, \|\tilde{\eta}\|, \tau)$, $\|\tilde{\eta}\| = \tilde{\eta}^\top \tilde{\eta}$, and $\tilde{\eta} = \Omega^{1/2} \eta$.

Corollary 6. *Let $\mathbf{Z} \sim ESN_d(\mu, \Omega, \eta, \tau)$, $\mathbf{Z}_N \sim N_d(\mu, \Omega)$ and W are defined as in Proposition 11. Then,*

$$(i) \quad R_\alpha(\mathbf{Z}) = R_\alpha(\mathbf{Z}_N) - N_\alpha(\mathbf{Z}), \quad \alpha \in \mathbb{N}, \alpha > 1,$$

where

$$N_\alpha(\mathbf{Z}) = \frac{1}{\alpha-1} \log \left(E \left\{ \left[\frac{\Phi_1(W)}{\Phi_1(\tau)} \right]^\alpha \right\} \right),$$

and $R_\alpha(\mathbf{Z}_0)$ is given by (1.8).

$$(ii) \quad R_\alpha(\mathbf{Z}) \leq R_\alpha(\mathbf{Z}_0) + \frac{\alpha}{1-\alpha} \log \left[\frac{\Phi_1(\tilde{\tau} + \tilde{\delta} \zeta_1(\tau))}{\Phi_1(\tau)} \right],$$

where $\tilde{\delta} = \|\tilde{\eta}\|^3 / \sqrt{1 + \|\tilde{\eta}\|^4}$.

$$(iii) \quad H(\mathbf{Z}) = H(\mathbf{Z}_0) - E \left\{ \ln \left[\frac{\Phi_1(W)}{\Phi_1(\tau)} \right] \right\}.$$

$$(iv) \quad H(\mathbf{Z}) \geq H(\mathbf{Z}_0) + \log[\Phi_1(\tau)] - \Phi_1 \left(\frac{\tilde{\tau}}{\sqrt{1 + \eta \eta^\top}} \right) \text{ and}$$

$$H(\mathbf{Z}) \leq \frac{1}{2} \log \left[(2\pi e)^k \left| \Omega - \zeta_1(\tau) [\tau + \zeta_1(\tau)] \delta \delta^\top \right| \right],$$

$$(v) \lim_{\alpha \rightarrow 1} N_{\alpha}(\mathbf{Z}) = E \left\{ \log \left[\frac{\Phi_1(W)}{\Phi_1(\tau)} \right] \right\}.$$

Pourahmadi (2007) illustrated the behaviour of $\zeta_1(\tau)$, $\tau \in \mathbb{R}$. This function is strictly decreasing for any $\tau \in \mathbb{R}$, tends to 0 when $\tau \rightarrow +\infty$, and diverge when $\tau \rightarrow -\infty$. For $\tau = 0$, the property (iv) of Corollary 6 becomes property (iii) of Corollary 1. By properties (iii) of Corollary 6 and (ii) of Corollary 1, the negentropy of an extended skew-normal (ESN) random vector is always larger than the negentropy of a SN random vector. Therefore, we obtain the following relationship among the Shannon entropies of normal ($H(\mathbf{Z}_0)$), SN ($H(\mathbf{Z}_N)$), and ESN ($H(\mathbf{Z})$) distributions: $H(\mathbf{Z}_0) \geq H(\mathbf{Z}_N) \geq H(\mathbf{Z})$. Considering (1.6), (1.11) and (3.25); the complexity measure for ESN distribution is obtained.

3.4 Information measures for generalized skew-normal distributions

Recent studies deal with the problem of measuring the disparity of a particular pdf from the normal one (Stehlík et al., 2014). Typical technique to deal with the problem has been exact expressions using information measures over particular distributions. For example, Vidal et al. (2006) measure the sensitivity of the skewness parameter using the L_1 distance between symmetric and asymmetric distributions. Stehlík (2003, 2012) proven results on decomposition of KL divergences in the gamma and normal family for divergence between maximum likelihood estimator (MLE) of canonical parameter and canonical parameter of regular exponential family. Gómez-Villegas et al. (2013) assessed the effect of kurtosis deviations from normality on conditional distributions, such as the multivariate exponential power family. Dette et al. (2017) characterizes the “disparity” between the skew-symmetric models and their symmetric counterparts in terms of the total variation distance, which is later used to construct priors. The paper provides additional insights, to those provided in Vidal et al. (2006), on the interpretation of this distance, and also discusses the usage of the KL divergence among several other distances.

Some recent applications of measuring the disparity of a particular pdf from the normal one using negentropy include those by Gao and Zhang (2009) and Wang et al. (2014), where negentropy method has been successfully applied to seismic wavelet estimation. Pires and Ribeiro (2017) considered the negentropy to measure the distance of non-normal information from normal one in independent components, with application to Northern Hemispheric winter monthly variability of a high-dimensional quasi-geostrophic atmospheric model. Also,

Pires and Hannachi (2017) used a tensorial invariant approximation of the multivariate negentropy in terms of a linear combination of squared coskewness and cokurtosis. Then, the method is applied to global sea surface temperature anomalies, after data anomalies being tested through a non-Gaussian distribution.

In this section we develop procedure, based on KL divergences, to test the significance of the skewness parameter in the Generalized SN (GSN) distributions, a flexible class of distributions that includes the SN and normal ones as particular cases (Arellano-Valle et al., 2017). We consider the Shannon entropy for the GSN subclass, i.e., the Shannon entropy of $Z \sim GSN(\eta; G, w)$. Thus, assuming a normal kernel in (2.24), we get the GSN-SE given by

$$H(Z) = H(Z_0) - E(\log[2G\{w(Z; \eta)\}]), \quad (3.26)$$

where $H(Z_0) = (1/2) \log(2\pi e)$ is the Shannon entropy of Z_0 . It is assumed that a specific skewness value η_0 exists so that $w(z; \eta_0) = 0$ and so $G\{w(z; \eta_0)\} = 1/2$, thus recovering symmetry at $\eta = \eta_0$. Therefore, at $\eta = \eta_0$, Z and Z_0 have the same distribution and thus the same Shannon entropy.

Let $\mu_z = E(Z)$ and $\sigma_z^2 = \text{Var}(Z)$ be the mean and variance of $Z \sim GSN(\eta; G, w)$, respectively, which must constitute functions of the skewness parameter η . Since $\sigma_z^2 = 1 - \mu_z^2$ and $H(Z'_0) = \log \sigma_z + H(Z_0)$, we get from (1.5) that the negentropy of Z becomes

$$N(Z) = \frac{1}{2} \log(1 - \mu_z^2) + E(\log[2G\{w(Z; \eta)\}]). \quad (3.27)$$

Since at $\eta = \eta_0$, we have by symmetry $\mu_z = 0$ and $w(Z; \eta_0) = 0$, so the negentropy in this case is null, as expected. Clearly, Shannon entropy and negentropy depend on the choice of the functions $G(\cdot)$ and $w(\cdot; \eta)$. In this Section, we consider both families of GSN distributions for which $\eta_0 = 0$, with $w(z; 0) = w(0; \eta) = 0$ and $w(-z; \eta) = w(z; -\eta)$, thus following that $-Z \sim GSN(-\eta; G, w)$ and recovering the normality at $\eta = 0$. Examples of this type of functions are $w(z; \eta) = \eta u(z)$ and $w(z; \eta) = u(\eta z)$ for some odd function $u(z)$, with $u(0) = 0$. In this case, recalling that $\eta Z \stackrel{d}{=} \tau Z_\tau$, where $Z_\tau \sim GSN(\tau; G, w)$, $\tau = |\eta|$, and “ $\stackrel{d}{=}$ ” denotes equality in distribution, we obtain

$$E(\log[2G\{w(Z; \eta)\}]) = E(\log[2G\{w(Z_\tau; \tau)\}]), \quad (3.28)$$

thus $H(Z) = H(Z_\tau)$ and $N(Z) = N(Z_\tau)$. That is, the entropy and negentropy of $Z \sim SN(\eta)$ depend on the skewness parameter η only through its absolute value $\tau = |\eta|$.

We have interest in both KL and J divergences for a GSN distribution with respect to the normal distribution. That is, assuming in (1.13) and (1.15) that $Z_1 = Z \sim GSN(\eta; G, w)$ and $Z_2 = Z_0$. In this case, remembering that $\eta Z \stackrel{d}{=} \tau Z_\tau$, where $Z_\tau \sim GSN(\tau; G, w)$ and $\tau = |\eta|$,

we have $K(Z, Z_0) = K(Z_\tau, Z_0)$ and $K(Z_0, Z) = K(Z_0, Z_\tau)$, with:

$$K(Z_\tau, Z_0) = E(\log[2G\{w(Z_\tau; \tau)\}]) \quad \text{and} \quad K(Z_0, Z_\tau) = -E(\log[2G\{w(Z_0; \tau)\}]). \quad (3.29)$$

Therefore, $J(Z, Z_0) = J(Z_\tau, Z_0)$, with:

$$J(Z_\tau, Z_0) = E(\log[2G\{w(Z_\tau; \tau)\}]) - E(\log[2G\{w(Z_0; \tau)\}]). \quad (3.30)$$

We also develop asymptotic expansions of the J divergence for the SN and MSN distributions from the normal distribution. To do this, we consider the following preliminary result, which proof stems from (3.29) and (3.30) by using the Taylor expansion of $\zeta(z; \tau) = \log[2G\{w(z; \tau)\}]$ at $z = 0$, and also because of the facts that (a) all moments of $Z_\tau \sim \text{GSN}(\tau; G, w)$ are finite, and (b) Z_τ and Z_0 contain the same even moments.

Lemma 12. *Consider the composite function $\zeta(z; \tau) = \log[2G\{w(z; \tau)\}]$, $\tau > 0$, by assuming that both functions $G(z)$ and $w(z; \tau)$ are infinitely differentiable, hence also $\zeta(z; \tau)$ is infinitely differentiable at $z = 0$. If $Z_\tau \sim \text{GSN}(\tau; G, u)$, then:*

$$\begin{aligned} K(Z_\tau, Z_0) &= \sum_{i=1}^{\infty} \frac{\zeta^{(i)}(0; \tau)}{i!} E(Z_\tau^i), \\ J(Z_\tau, Z_0) &= \sum_{i=1}^{\infty} \frac{\zeta^{(2i-1)}(0; \tau)}{(2i-1)!} E(Z_\tau^{2i-1}), \end{aligned} \quad (3.31)$$

where $\zeta^{(i)}(z; \tau)$ is the i th derivative of $\zeta(z; \tau)$. Moreover, from (3.31), the expressions (3.26) and (3.27) for the Shannon entropy and negentropy of the GSN distributions have the forms:

$$\begin{aligned} H(Z_\tau) &= \frac{1}{2} \log(2\pi e) - \sum_{i=1}^{\infty} \frac{\zeta^{(i)}(0; \tau)}{i!} E(Z_\tau^i), \\ N(Z_\tau) &= \frac{1}{2} \log(1 - \mu_\tau^2) + \sum_{i=1}^{\infty} \frac{\zeta^{(i)}(0; \tau)}{i!} E(Z_\tau^i), \end{aligned}$$

respectively, where $\mu_\tau = E(Z_\tau)$.

Notice in Lemma 12 that the coefficient $\zeta^{(i)}(0; \tau)$ depends on the derivatives of $G(z)$ and $w(z; \tau)$ at $z = 0$, which change for different GSN distributions. Moreover, since the expansion of $\zeta(z; \tau)$ emerges around $z = 0$ by assuming a fixed τ , the approximations may not be reasonable for some values of τ .

3.4.1 Skew-normal distribution

The function $\zeta(z; \tau) = \zeta_0(\tau z) = \log\{2\Phi(\tau z)\}$ is also infinitely differentiable at $z = 0$, thus admitting a Taylor expansion about zero. Therefore, since $\zeta^{(i)}(z; \tau) = \tau^i \zeta_i(\tau z)$, where $\zeta_i(x)$ is the i th derivative of $\zeta_0(x) = \log\{2\Phi(x)\}$, the expansion (3.31) in Lemma 12 of $E\{\zeta(Z_\tau; \tau)\} = E\{\zeta_0(\tau Z_\tau)\}$, $Z_\tau \sim SN(\tau)$, becomes

$$E\{\zeta_0(\tau Z_\tau)\} = \sum_{i=1}^{\infty} \frac{\tau^i \kappa_i \mu_i}{i!}, \quad (3.32)$$

where the moments $\mu_i = E(Z_\tau^i)$ are given in Section 2.4.1 and the coefficients $\kappa_i = \zeta_i(0)$, $i = 1, 2, \dots$, are related to the cumulants of the half-normal random variable $V \sim \sqrt{\chi_1^2}$ given by $K(t) = (1/2)t^2 + \zeta_0(t)$ (see also Azzalini and Capitanio, 1999, 2013). Let $K_m(0) = \frac{d^m}{dt^m} K_t|_{t=0}$ be the m th cumulant of V and clearly $K_1(0) = \zeta_1(0) = \kappa_1$, $K_2(0) = 1 + \zeta_2(0) = 1 + \kappa_2$, and $K_m(0) = \zeta_m(0) = \kappa_m$, $m = 3, 4, \dots$. Also, from (Azzalini and Capitanio, 1999, 2013), it emerges that

$$\begin{aligned} \zeta_1(x) &= \frac{\phi(x)}{\Phi(x)}, \\ \zeta_2(x) &= -\zeta_1(x)\{x + \zeta_1(x)\}, \\ \zeta_3(x) &= -\zeta_2(x)\{x + \zeta_1(x)\} - \zeta_1(x)\{1 + \zeta_2(x)\}, \\ \zeta_4(x) &= -\zeta_3(x)\{x + \zeta_1(x)\} - 2\zeta_2(x)\{1 + \zeta_2(x)\} - \zeta_1(x)\zeta_3(x), \\ \zeta_5(x) &= -\zeta_4(x)\{x + \zeta_1(x)\} - 3\zeta_3(x)\{1 + \zeta_2(x)\} - 3\zeta_2(x)\zeta_3(x) - \zeta_1(x)\zeta_4(x), \\ &\vdots \end{aligned}$$

Recalling that $b = \sqrt{2/\pi}$, the first five coefficients $\kappa_i = \zeta_i(0)$, $i = 1, \dots, 5$, are

$$\begin{aligned} \kappa_1 &= b, \\ \kappa_2 &= -\kappa_1^2, \\ \kappa_3 &= -2\kappa_2\kappa_1 - \kappa_1, \\ \kappa_4 &= -2\kappa_3\kappa_1 - 2\kappa_2^2 - 2\kappa_2, \\ \kappa_5 &= -2\kappa_4\kappa_1 - 6\kappa_3\kappa_2 - 3\kappa_3, \\ &\vdots \end{aligned}$$

Thus, by letting $\kappa_0 = 1$, a recursive rule for these coefficients is obtained as follows:

$$\kappa_1 = b,$$

$$\begin{aligned}\kappa_{2i} &= -(2i-2)\kappa_{2i-2} - 2 \sum_{j=1}^i \binom{2i-2}{j-1} \kappa_j \kappa_{2i-j} + \binom{2i-2}{i-1} \kappa_i^2, \\ \kappa_{2i+1} &= -(2i-1)\kappa_{2i-1} - 2 \sum_{j=1}^i \binom{2i-1}{j-1} \kappa_j \kappa_{2i-j+1}, \quad i = 1, 2, \dots\end{aligned}$$

In summary, since the even moments of $Z_\tau \sim SN(\tau)$ are also the even moments of Z_0 , Eq. (3.32) can be rewritten as

$$E\{\zeta_0(\tau Z_\tau)\} = \underbrace{\sum_{i=1}^{\infty} \frac{\tau^{2i} \kappa_{2i}}{(2i)!} \mu_{2i}}_{E\{\zeta_0(\tau Z_0)\}} + \underbrace{\sum_{i=1}^{\infty} \frac{\tau^{2i-1} \kappa_{2i-1}}{(2i-1)!} \mu_{2i-1}}_{J(Z_\tau, Z_0)}.$$

Hence, considering also Eq. (2.27), we can compute for the SN case the results for the KL and J divergences, Shannon entropy and negentropy given in Lemma 12 using the following Proposition 12.

Proposition 12. *Let $Z_\tau \sim SN(\tau)$ and $Z_0 \sim N(0, 1)$. Then:*

$$\begin{aligned}K(Z_\tau, Z_0) &= \sum_{i=1}^{\infty} \frac{\tau^i \kappa_i \mu_i}{i!}, \\ J(Z_\tau, Z_0) &= b \sum_{i=1}^{\infty} \frac{\kappa_{2i-1} \delta_\tau^{2i-1}}{(2i-1)!} \sum_{m=1}^i a_i(m) \tau^{2m-1}, \\ H(Z_\tau) &= \frac{1}{2} \log(2\pi e) - K(Z_\tau, Z_0), \\ N(Z_\tau) &= \frac{1}{2} \log\{1 - (b\delta_\tau)^2\} + K(Z_\tau, Z_0),\end{aligned}\tag{3.33}$$

where the coefficients $a_i(m)$, $i = 1, 2, \dots$, are given in Section 2.4.1.

The next Proposition 13 allows to compute the Rényi entropy and complexity measure of a TSN random variable.

Proposition 13. *Let Z, W be a $SN(\mu, \sigma^2, \eta)$ and $TSN(\mu, \sigma^2, \eta)$, respectively, $\eta \neq 0$. Then:*

$$\int_a^b [g(w)]^\alpha dw = 2\psi_{\alpha,1}(\sigma^2) \Phi_{\alpha+1}(\mathbf{0}; \mathbf{0}, \tilde{\Omega}) \frac{[H(v)]_{a_0}^{b_0}}{([F(z)]_a^b)^\alpha},\tag{3.34}$$

where $\psi_{\alpha,1}(\sigma^2)$ is defined as in Proposition 5 with $k = 1$ and $\Omega = J$; $\tilde{\Omega} = I_{\alpha+1} + \tilde{\eta}^2 \tilde{\mathbf{D}}^\top \tilde{\mathbf{D}}$, $\tilde{\eta}^2 = \omega \eta^2 / \alpha$, $\tilde{\mathbf{D}} = (\mathbf{1}_\alpha, \tilde{\eta})^\top$ and $V \sim CSN_{1,2}(0, \tilde{\eta}^2, \tilde{\mathbf{B}}, \mathbf{0}, I_2)$ with cdf $H(v)$, $\tilde{\mathbf{B}} = (1, \tilde{\eta})^\top$, $a_0 = \eta(a - \mu)/\omega$ and $b_0 = \eta(b - \mu)/\omega$.

Remark 4. By Lemma 2.2.1 of Genton (2004), $H(v)$ is easily computable by a tri-variate normal cdf as

$$H(v) = \frac{\Phi_3 \left[\begin{pmatrix} v \\ \mathbf{0} \end{pmatrix}; \begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \tilde{\eta}^2 & -\tilde{\eta}^2 \tilde{\mathbf{B}} \\ -\tilde{\eta}^2 \tilde{\mathbf{B}}^\top & I_2 + \tilde{\eta}^2 \tilde{\mathbf{B}}^\top \tilde{\mathbf{B}} \end{pmatrix} \right]}{\Phi_2(\mathbf{0}; \mathbf{0}, I_2 + \tilde{\eta}^2 \tilde{\mathbf{B}}^\top \tilde{\mathbf{B}})},$$

where $\tilde{\eta}$ and $\tilde{\mathbf{B}}$ are defined as in Proposition 13.

Considering (1.6), (1.11) and (3.34); the complexity measure for TSN distribution is obtained.

3.4.2 Modified skew-normal distribution

In the MSN case, $G(w) = \Phi(w)$ and $w(z; \tau) = \tau u(z) = \tau z / \sqrt{1 + z^2}$, both of which are infinitely differentiable at $z = 0$. Thus, in Lemma 12 we have $\zeta(z; \tau) = \zeta_0\{\tau u(z)\}$, where $\zeta_0(x) = \log\{2\Phi(x)\}$ is also infinitely differentiable at $z = 0$. Thus, the series expansion of $E\{\zeta(Z_\tau; \tau)\} = E[\zeta_0\{\tau u(Z_\tau)\}]$, $Z_\tau \sim MSN(\tau)$, can be obtained from (3.31) for which we need the derivatives of the composite function $\zeta_0\{\tau u(z)\} = \log[2\Phi\{\tau u(z)\}]$. Another way to obtain these derivatives is to define random variable $Z_\tau^* = u(Z_\tau) = Z_\tau / \sqrt{1 + Z_\tau^2}$, and using (3.32) with Z_τ and $\mu_k = E(Z_\tau)$ replaced by Z_τ^* and $\mu_k^* = E\{(Z_\tau^*)^k\}$, respectively. Thus, we obtain the series expansion

$$E\{\zeta_0(\tau Z_\tau^*)\} = \sum_{i=1}^{\infty} \frac{\tau^i \kappa_i}{i!} \mu_i^*.$$

From Lemma 12, the KL and J divergences, Shannon entropy and negentropy for the MSN case can be computed using the following Proposition 14.

Proposition 14. Let $Z_\tau \sim MSN(\tau)$ and $Z_0 \sim N(0, 1)$. Then:

$$\begin{aligned} K(Z_\tau, Z_0) &= \sum_{i=1}^{\infty} \frac{\tau^i \kappa_i}{i!} \mu_i^*, \\ J(Z_\tau, Z_0) &= \sum_{i=1}^{\infty} \frac{\tau^{2i-1} \kappa_{2i-1}}{(2i-1)!} \mu_{2i-1}^*, \\ H(Z_\tau) &= \frac{1}{2} \log(2\pi e) - K(Z_\tau, Z_0), \\ N(Z_\tau) &= \frac{1}{2} \log(1 - [b\{2\xi_1(\tau) - 1\}]^2) + K(Z_\tau, Z_0). \end{aligned} \tag{3.35}$$

In order to compute the quantities given by Proposition 14, we need to calculate the new moments $\mu_i^* = E\{(Z_\tau^*)^i\}$, $i = 1, 2, \dots$. Since $Z_\tau^* = Z_\tau / \sqrt{1 + Z_\tau^2}$ is a random variable limited

to the interval $(-1, 1)$, all its moments are finite. In particular, Z_τ^* clearly has the same even moments as $Z_0^* = Z_0/\sqrt{1+Z_0^2}$. Moreover, from the Jacobian method, the pdf of Z_τ^* becomes

$$f^*(u) = \frac{2}{(1-u^2)^{3/2}} \phi\left(\frac{u}{\sqrt{1-u^2}}\right) \Phi(\tau u), \quad u \in (-1, 1).$$

Hence, the i th moment of Z_τ^* is

$$\mu_i^* = \int_{-1}^1 u^i f^*(u) du, \quad k = 1, 2, \dots,$$

which must be computed numerically.

3.5 J Divergence between generalized skew-normal distributions

In the previous sections, SN and MSN distributions were compared with the normal distribution by means of the J divergence measure. As byproduct, we were also computing the J divergence between the SN and MSN distributions, both with the same skewness parameter. This allows measuring the distance between these distributions with different $w(z; \eta)$'s. For this, we consider in Eq. (1.15) that $Z_1 \sim SN(\tau)$ and $Z_2 \sim MSN(\tau)$, and define the random variables $Z_i^* = Z_i/\sqrt{1+Z_i^2}$ for $i = 0, 1, 2$, where $Z_0 \sim N(0, 1)$. Let $\mu_{i,j} = E(Z_i^j)$ and $\mu_{i,j}^* = E\{(Z_i^*)^j\}$, $i = 0, 1, 2$. Recall that $\mu_{i,2j} = \mu_{0,2j}$ and $\mu_{i,2j}^* = \mu_{0,2j}^*$ for all $j = 1, 2, \dots$. Thus, using (1.15) and then the Taylor expansion of $\zeta_0(x) = \log\{\Phi(x)\}$ around $x = 0$, Proposition 15 is obtained:

Proposition 15. *Let $Z_0 \sim N(0, 1)$, $Z_1 \sim SN(\tau)$ and $Z_2 \sim MSN(\tau)$. Define the random variables $Z_i^* = Z_i/\sqrt{1+Z_i^2}$, $i = 1, 2$. Then:*

$$\begin{aligned} J(Z_1, Z_2) &= E\{\zeta_0(\tau Z_1)\} - E\{\zeta_0(\tau Z_1^*)\} + E\{\zeta_0(\tau Z_2^*)\} - E\{\zeta_0(\tau Z_2)\} \\ &= J(Z_1, Z_0) - J(Z_1^*, Z_0) + J(Z_2^*, Z_0) - J(Z_2, Z_0), \end{aligned}$$

where as before

$$\begin{aligned} J(Z_i, Z_0) &= \sum_{j=1}^{\infty} \tau^{2j-1} \frac{\mu_{i,2j-1} \kappa_{2j-1}}{(2j-1)!}, \\ J(Z_i^*, Z_0) &= \sum_{j=1}^{\infty} \tau^{2j-1} \frac{\mu_{i,2j-1}^* \kappa_{2j-1}}{(2j-1)!}, \quad i = 1, 2. \end{aligned}$$

Proposition 15 indicates that J divergence between SN and MSN distributions is decomposed to the divergences of the normal distribution with each of these distributions, which depends only on their odd moments and cumulants.

3.6 Asymptotic tests for generalized skew-normal distributions

Let $f(x; \boldsymbol{\theta})$, $x \in \mathcal{X}$, $\boldsymbol{\theta} \in \Theta$, be the pdf of a regular parametric class of distributions, i.e., for which the sample space \mathcal{X} does not depend on $\boldsymbol{\theta}$, the parametric space Θ is an open subset of \mathbb{R}^p , and the regularity conditions (i), (ii) (iii) stated in Salicrú et al. (1994) are satisfied. As in Salicrú et al. (1994), we denote the KL divergence between $f(x; \boldsymbol{\theta})$ and $f(x; \boldsymbol{\theta}')$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, by

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int_{\mathcal{X}} f(x; \boldsymbol{\theta}) \log \left\{ \frac{f(x; \boldsymbol{\theta})}{f(x; \boldsymbol{\theta}')} \right\} dx.$$

Consider the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1 \in \Theta_1 \subset \mathbb{R}^r$ and $\boldsymbol{\theta}_2 \in \Theta_2 = \Theta \cap \Theta_1^c \subset \mathbb{R}^{p-r}$. Let $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^0)$ and consider the null hypothesis $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^0$ for a known $\boldsymbol{\theta}_2^0 \in \Theta \cap \Theta_1^c$. Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ and $\hat{\boldsymbol{\theta}}' = (\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2^0)$ be the (unrestricted) MLE of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, respectively, both based on a random sample of size n from X with pdf $f(x; \boldsymbol{\theta})$. Under these conditions, we have from part b) of Theorem 2 presented in Salicrú et al. (1994) that

$$2nK(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}') = 2n \int_{\mathcal{X}} f(x; \hat{\boldsymbol{\theta}}) \log \left\{ \frac{f(x; \hat{\boldsymbol{\theta}})}{f(x; \hat{\boldsymbol{\theta}}')} \right\} dx \xrightarrow[n \rightarrow \infty]{d} \chi_{p-r}^2, \quad \forall \boldsymbol{\theta} \in \Theta, \quad (3.36)$$

where “ \xrightarrow{d} ” denotes convergence in distribution and χ_s^2 denotes the chi-squared distribution function with s degrees of freedom. From (3.36), the above null hypothesis can be tested by the statistic $2nK(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}')$, which is asymptotically chi-squared distributed with $p - r$ degrees of freedom. Specifically for large values of n , if we observe $K(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}') = K_0$, then H_0 is rejected at level α if $P(\chi_{p-r}^2 > 2nK_0) \leq \alpha$.

One-sample case: Test for normality

The result in (3.36) can be applied for example to construct a normality test from the KL divergence between a regular GSN distribution and the normal distribution. Specifically, consider a random sample X_1, \dots, X_n from $X \sim GSN(\mu, \sigma^2, \eta, G, w)$ and the null hypothesis $H_0 : \eta = \eta_0$ under which $G\{w(z; \eta_0)\} = G(0) = 1/2$, thus the GSN random variable X

becomes a $N(\mu, \sigma^2)$ random variable. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma^2}, \widehat{\eta})$ be the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2, \eta)$ and $\widehat{\boldsymbol{\theta}}' = (\widehat{\mu}, \widehat{\sigma^2}, \eta_0)$. Therefore, under $H_0 : \eta = \eta_0$ we have

$$2nK(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}') = 2nK(Z_{\widehat{\tau}}, Z_0) \xrightarrow[n \rightarrow \infty]{d} \chi_1^2, \quad (3.37)$$

where $K(Z_{\widehat{\tau}}, Z_0)$ is the MLE of $K(Z_{\tau}, Z_0)$, which is defined in Eq. (3.31) of Lemma 1 and depends only on $\widehat{\tau} = |\widehat{\eta}|$. As stated in the introduction, normality is typically obtained from the GSN class at $\eta_0 = 0$ or equivalently $\tau_0 = |\eta_0| = 0$.

Azzalini (1985), Arellano-Valle and Azzalini (2008) and Azzalini and Capitanio (2013) recall on the singularity of SN FIM at $\eta = 0$, preventing the asymptotic distribution of the above statistic tests. As suggested by Azzalini (1985), a solution to recover the non-singularity of the information matrix under the symmetry hypothesis comes from the use of the so-called centered parametrization defined in terms of the mean, variance and the skewness parameters of the SN distribution (see also Arellano-Valle and Azzalini, 2008, Chiogna, 1998). Otherwise, the FIM of the MSN model is non-singular at $\eta = 0$ (Arrué et al., 2016). Thus, this model satisfies all the standard regularity conditions of Salicrú et al. (1994), leading to consistence and asymptotic normality of the MLEs under the null hypothesis of normality. Therefore, the MSN model serves to test the null hypothesis of normality using (3.37). Hence, the symmetry null hypothesis $H_0 : \tau = 0$ is rejected at level α if $P(\chi_1^2 > 2nK_0) \leq \alpha$, with $K_0 = K(Z_{\widehat{\tau}}, Z_0)$.

Two-samples case

Consider two independent samples of sizes n_1 and n_2 from X_1 and X_2 , respectively; where $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta \subset \mathbb{R}^p$, and X_1 and X_2 have pdf's $f(x; \boldsymbol{\theta}_1)$ and $f(x; \boldsymbol{\theta}_2)$, respectively. Suppose partition $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2})$, $i = 1, 2$, and assume $\boldsymbol{\theta}_{21} = \boldsymbol{\theta}_{11} \in \Theta_1 \subset \mathbb{R}^r$, so that $\boldsymbol{\theta}_{i2} \in \Theta \cap \Theta_1^c \subset \mathbb{R}^{p-r}$, $i = 1, 2$. Let $\widehat{\boldsymbol{\theta}}_i = (\widehat{\boldsymbol{\theta}}_{i1}, \widehat{\boldsymbol{\theta}}_{i2})$ be the MLE of $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2})$, $i = 1, 2$, which correspond to the MLE of the full model parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ under null hypothesis $H_0 : \boldsymbol{\theta}_{21} = \boldsymbol{\theta}_{11}$. Thus, part b) of Corollary 1 in Salicrú et al. (1994) establishes that if the null hypothesis $H_0 : \boldsymbol{\theta}_{22} = \boldsymbol{\theta}_{12}$ holds and $\frac{n_1}{n_1+n_2} \xrightarrow[n_1, n_2 \rightarrow \infty]{} \lambda$, with $0 < \lambda < 1$, then

$$\frac{2n_1n_2}{n_1+n_2}K(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2) = \frac{2n_1n_2}{n_1+n_2} \int_{-\infty}^{\infty} f(x; \widehat{\boldsymbol{\theta}}_1) \log \left\{ \frac{f(x; \widehat{\boldsymbol{\theta}}_1)}{f(x; \widehat{\boldsymbol{\theta}}_2)} \right\} dx \xrightarrow[n_1, n_2 \rightarrow \infty]{d} \chi_{p-r}^2. \quad (3.38)$$

Thus, a test of level α for the above homogeneity null hypothesis consists of rejecting H_0 if

$$\frac{2n_1n_2}{n_1+n_2}K(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2) > \chi_{p-r, 1-\alpha}^2,$$

where $\chi_{p-r,\alpha}^2$ is the α th percentile of the χ_{p-r}^2 -distribution.

Contreras-Reyes and Arellano-Valle (2012) considered the result of Kupperman (1957) to develop an asymptotic test of complete homogeneity in terms of the J divergence between two SN distributions. The SN distribution satisfies all the aforementioned regularity conditions when skewness parameter $\eta \neq 0$. Thus, considering this condition we can also apply (3.36) and (3.38) to obtain, respectively, asymptotic tests with one or two samples of other hypotheses not covered by Kupperman's test.

Likelihood Ratio Test

The Likelihood Ratio Test (LRT) statistic (Chernoff, 1954) for a null hypotheses $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$, $\Theta_0 \subset \Theta$ (the parametric space), is given by

$$\text{LRT} = 2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_0)\},$$

where $\hat{\theta}$ is the unrestricted MLE of θ , $\hat{\theta}_0$ is the MLE of θ under H_0 , and the log-likelihood function $\ell(\theta)$ for MSN distributions is

$$\ell(\eta) = \frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n z_i^2 + \sum_{i=1}^n \log \zeta_0\{\eta u(z_i)\}, \quad (3.39)$$

for shape parameter η and a random sample of size n from Z (Arru  et al., 2016).

As before, normality is typically obtained from the GSN class at $\eta_0 = 0$. Because the MSN distribution satisfies the standard regularity conditions (Arru  et al., 2016), the LRT statistic is asymptotically χ_s^2 distributed under H_0 , with $s = \dim(\Theta) - \dim(\Theta_0) = 1$ degrees of freedom (Azzalini and Arellano-Valle, 2013). Hence, the p-value associated with the LRT is computed as $1 - \chi_s^2(\text{LRT})$, where $\chi_s^2(\text{LRT})$ denotes the χ_s^2 -distribution function evaluated at the observed value of the LRT statistic.

In order to test normality, we considered the particular null hypothesis $H_0 : \eta = 0$ versus $H_1 : \eta \neq 0$, with the rest of the parameters not specified. Therefore, by (3.39) the LRT statistic is given by

$$\text{LRT} = 2 \left[-\frac{1}{2} \sum_{i=1}^n \hat{z}_i^2 + \frac{1}{2} \sum_{i=1}^n \hat{z}_{i0}^2 + \sum_{i=1}^n \log \zeta_0\{\hat{\eta} u(\hat{z}_i)\} \right],$$

where $\hat{\eta}$ is the unrestricted MLE of η , \hat{z}_i and \hat{z}_{i0} are the unrestrited and restricted MLE of $z_i = (y_i - \mu)/\sigma$, respectively; and the p-value is computed as $1 - \chi_1^2(\text{LRT})$.

Chapter 4

Simulations

4.1 Skew-normal and skew- t distributions

4.1.1 Shannon and Rényi entropies

A convenient and fast method to compute the entropies presented in Proposition 3 and Eq.(3.19) is based on the numerical integration QUADPACK (a Subroutine Package for Automatic Integration) (Piessens et al., 1983) implemented in the `integrate R` (R Core Team, 2016) function. The results are shown in Figure 4.1 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 $\nu = 1, 2, \dots, 185$ to illustrate the entropies:

$$H(X) = \frac{1}{2} \{1 + \log(2\pi)\} - E[\log\{2\Phi(\alpha W_{SN})\}], \quad (4.1)$$

$$H(Y) = -\log\left\{\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}}\right\} + \frac{\nu+1}{2} \left\{\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right\} \\ - E\left[\log\left\{2T\left(\sqrt{\frac{\nu+1}{\nu+W_{ST}^2}}\alpha W_{ST}; \nu+1\right)\right\}\right], \quad (4.2)$$

where $X, W_{SN} \sim SN_1(\alpha)$ and $Y, W_{ST} \sim ST_1(0, 1, \alpha, \nu)$.

We can see in Figure 4.1 that the numerical implementation suggests the convergence of the integrals involved in (4.1) and (4.2). The QUADPACK method is more precise and efficient in terms of computational time than other methods such as Monte Carlo.

The convergence of the ST entropy to the skew-normal entropy is obtained quickly for values of $\nu \geq 7$. 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- t 's marginal cases ($\alpha = 0$), we have that the entropy is maximized and decreasing for greater values of α . When $\alpha \rightarrow \infty$,

the entropy tends to a constant and for $\alpha > 20$, it is almost that of a half-normal distribution already.

We also compare the numerical method based on QUADPACK routine with exact computation given by Proposition 5, for univariate SN Rényi entropy. The results are shown in Table 4.1 for the dimension $k = 1$, $Z \sim SN_1(0, 1, \eta)$, skewness parameter $\alpha \in [0.2, 6]$, and $\alpha = 2, \dots, 6$. We can see similar results for each Rényi entropy from the 5th to 7th decimal. In Figure 4.3, we can see that Rényi entropies decrease when α increase and η decrease, i.e., when the SN tend to or is maximized by normal distribution.

4.1.2 Maximum entropy, Jeffrey's and Kullback-Leibler divergences

Figure 4.2 shows in panel (a) several values of $h(\tau) = E[\log\{2\Phi(\tau W)\}]$ for $\tau = 0, 1, \dots, 200$. 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 (3.10) for $\tau = 0, 1, \dots, 10$.

Figure 4.4 illustrates the numerical behavior of the KL divergence between two univariate SN distributions under different scenarios for the model parameters. More specifically, we can observe from there the behavior of KL and J divergences given in Proposition 6 and Corollaries 2, 3 and 4 for the univariate special cases described below.

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

$$\begin{aligned} K(X, Y) &= \frac{1}{2} \left(\frac{\xi_1 - \xi_2}{\omega} \right)^2 + \sqrt{\frac{2}{\pi}} \frac{\gamma}{\sqrt{1 + \tau^2}} + E[\log\{2\Phi(\tau W)\}] \\ &\quad - E[\log\{2\Phi(\tau W + \gamma)\}], \\ J(X, Y) &= 2 \left(\frac{\xi_1 - \xi_2}{\omega} \right)^2 + 2E[\log\{2\Phi(\tau W)\}] - E[\log\{2\Phi(\tau W - \gamma)\}] \\ &\quad - E[\log\{2\Phi(\tau W + \gamma)\}], \end{aligned}$$

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

The panels (a), (b) and (c) of Figure 4.4 show that this J divergence increases mainly with the distance between the location parameters ξ_1 and ξ_2 . Looking for $K(X, Y)$ and panel (c), we can observe that larger values of ω^2 produce the smallest values of KL divergence, independently of the values of η .

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

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

where $\delta_1 = \eta \omega_1^2 / \sqrt{1 + \eta^2 \omega_1^2}$. In the plot (d) of Figure 4.4 is illustrated this case for $\xi = \eta = 1$.

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

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

where $W_{ij} \sim SN_1(0, \tau_i^2, s_{ij}\tau_j)$, with $\tau_i = |\eta_i|\omega$ and $s_{ij} = \text{sign}(\eta_i\eta_j)$, $i, j = 1, 2$. The behavior of this case is illustrated in panel (e) of Figure 4.4 for $\xi = 1$ and $\omega^2 = 1$. Finally, the panels (f) and (g) of Figure 4.4 correspond to the KL divergence of Proposition 6 for $k = 1$.

The found approximated KL divergence of ST distributions is compared with numerical methods for the univariate case. Figure 4.5 illustrates the numerical behavior (using QUADPACK routine) of the exact and asymptotic KL divergences between two univariate skew- t distributions for a fixed value $v_1 = 5$ and several values $v_2 = 3, \dots, 150$. Without loss of generality, $\xi_1 = \xi_2 = 0$ is chosen in this study. It can be observed that the asymptotic KL divergence tends to approximate the exact KL divergence when $v_2 \rightarrow \infty$. Hence, this allows the convergence of the CE formula in Lemma 11. Figure 4.6 shows the behavior of asymptotic KL divergences between two univariate skew- t distributions (panel (a)) and two univariate Student- t distributions (panel (b)) for several values of v_1 and v_2 . In panel (a), the KL divergence tends to increase when v_1 differs from v_2 . Whilst in panel (b), in absence of skewness, the KL divergence given by (3.23) tends to increase mainly for small values of v_1 and, small values of v_2 increase this measure.

4.1.3 Jensen-Shannon distance

For any $\eta_1, \eta_2 \neq 0$, the expected values of (8) are not directly computable. However, the integrals are evaluated numerically using the QUADPACK routine. We have shown that for values of η_1 between 0 and 1 and large values of η_2 , the largest values of KL divergence are produced. For larger values of η_1 , the KL divergence is close to zero. This shows

the asymmetry property of this divergence. On the other hand, Fig. 4.7 shows clearly the symmetry property of the JS distance. In addition, we see that the JS distance increases mainly with distance between the shape parameters η_1 and η_2 .

The inequalities

$$JS(X, Y) \leq \frac{1}{4}J(X, Y), \quad (4.3)$$

$$JS(X, Y) \leq \log \left\{ \frac{2}{1 + e^{-\frac{J(X, Y)}{2}}} \right\}, \quad (4.4)$$

provided by Lin (1991) and Crooks (2008), respectively, are very helpful to approximate the SN JS distance with upper bounds for any η_1 and η_2 . Fig. 4.8 compares the inequalities (4.3) and (4.4) with the skew-gaussian JS distance (8) for several values of η_1 and η_2 . The smallest value of η_1 and η_2 produces similar results for the two bounds, and large differences with respect to JS distance (panels (a)–(c)). However, panels (d)–(f) show that Lin's bound is larger than Crook's bound for large values of η_1 ($\eta_1 > 2$), and both are similar to JS distance. As is expected, for all cases the upper bounds and JS distance are equal when $\eta_1 = \eta_2 = 0$, where $JS(X, Y) = J(X, Y) = 0$.

4.2 Finite mixture of multivariate skew-normal distributions

To study the behavior of the Shannon entropy bounds of Proposition 9 and the Rényi entropy bounds of Eq. (3.13) and Lemma 6, some examples are simulated for the cases $k = 1, 2$ and 3:

Example 2. $k = 1, m = 2, \pi = (0.3, 0.7), \tilde{\xi} = (0.5, 5), \tilde{\Omega} = (3.5, 6), \text{ and } \tilde{\eta} = (0.5, 3.5)$.

Example 3. (Prates et al., 2013)

$k = 1, m = 3, \pi = (0.5, 0.2, 0.3), \tilde{\xi} = (2, 20, 35), \tilde{\Omega} = (9, 16, 9), \text{ and } \tilde{\eta} = (5, 3, 6)$.

Example 4. (Prates et al., 2013)

$k = 2, m = 2, \pi = (0.65, 0.35),$

$\tilde{\xi} = \left(\begin{pmatrix} 5 \\ 7 \end{pmatrix}, \begin{pmatrix} 2 \\ 5 \end{pmatrix} \right), \tilde{\Omega} = \left(\begin{pmatrix} 0.18 & 0.6 \\ 0.6 & 4 \end{pmatrix}, \begin{pmatrix} 0.15 & 1.15 \\ 1.15 & 4 \end{pmatrix} \right), \text{ and}$

$\tilde{\eta} = \left(\begin{pmatrix} 0.69 \\ 0.64 \end{pmatrix}, \begin{pmatrix} 4.3 \\ 2.7 \end{pmatrix} \right).$

Example 5. (*Prates et al., 2013*)

$$k = 2, m = 3, \pi = (0.25, 0.5, 0.25),$$

$$\tilde{\xi} = \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 2 \\ 8 \end{pmatrix} \right), \tilde{\Omega} = \left(\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 0.15 & 1.15 \\ 1.15 & 40 \end{pmatrix} \right), \text{ and}$$

$$\tilde{\eta} = \left(\begin{pmatrix} 4 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 4.3 \\ 2.7 \end{pmatrix} \right).$$

Example 6. (*Celeux and Soromenho, 1996*)

$$k = 3, m = 3, \pi = (0.22, 0.36, 0.42),$$

$$\tilde{\xi} = \left(\begin{pmatrix} 10 \\ 12 \\ 10 \end{pmatrix}, \begin{pmatrix} 8.5 \\ 10.5 \\ 8.5 \end{pmatrix}, \begin{pmatrix} 12 \\ 14 \\ 12 \end{pmatrix} \right), \tilde{\Omega} = \left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right),$$

$$\text{and } \tilde{\eta} = \left(\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 2 \\ 2 \end{pmatrix} \right).$$

Example 7. (*Lee and McLachlan, 2013*)

$$k = 3, m = 4, \pi = (0.125, 0.19, 0.135, 0.55),$$

$$\tilde{\xi} = \left(\begin{pmatrix} 420 \\ 360 \\ 425 \end{pmatrix}, \begin{pmatrix} 160 \\ 570 \\ 200 \end{pmatrix}, \begin{pmatrix} 320 \\ 540 \\ 260 \end{pmatrix}, \begin{pmatrix} 530 \\ 80 \\ 450 \end{pmatrix} \right),$$

$$\tilde{\Omega} = \left(\begin{pmatrix} 9160 & 5580 & 7000 \\ 5580 & 12105 & 7160 \\ 7000 & 7160 & 7250 \end{pmatrix}, \begin{pmatrix} 3870 & 1810 & 1770 \\ 1810 & 2900 & 1270 \\ 1770 & 1270 & 1320 \end{pmatrix}, \begin{pmatrix} 1695 & 1190 & 2280 \\ 1190 & 2780 & 2010 \\ 2280 & 2010 & 3720 \end{pmatrix}, \right.$$

$$\left. \begin{pmatrix} 1590 & 590 & 15 \\ 590 & 2425 & 415 \\ 15 & 415 & 1870 \end{pmatrix} \right), \text{ and } \tilde{\eta} = \left(\begin{pmatrix} 4.8 \\ 17 \\ 50 \end{pmatrix}, \begin{pmatrix} 4 \\ 80 \\ 60 \end{pmatrix}, \begin{pmatrix} 40 \\ 8 \\ 10 \end{pmatrix}, \begin{pmatrix} 60 \\ 90 \\ 6 \end{pmatrix} \right).$$

Figure 4.9 presents the examples mentioned in the settings above. Examples 2 and 3 are represented in histogram plots and examples 4 and 5 in contour plots, according to Prates et al. (2013). Examples 6 and 7 are represented in 3D plots, according to Lee and McLachlan (2013). For all simulations, a sample of $n = 500$ generations is considered, and then fixed using the function `smsn.mix` from `mixsmsn` package, developed by Prates et al. (2013) and implemented in an R environment (R Core Team, 2016). Prates et al. (2013) implemented

routines for ML estimation via the Expectation Maximization EM-type algorithm in FMSN models (among several others).

For each example, Table 4.2 summarizes the four Shannon, as well as the Rényi entropy bounds for $\alpha = 2, \dots, 5$ and $m = 2, \dots, 6$. Shannon and Rényi entropies are compared with AIC and BIC criteria (see e.g. Contreras-Reyes et al., 2014), misclassification (MC) rates and consistency scores: normal skill score (NSS), Heidke skill score (HSS), and Hanssen–Kuipers (HK) (see e.g. Contreras-Reyes, 2013). All these indicators show an optimal performance of model fit if they are near 1; except MC, which ideally should be close to 0 (i.e. $100(1 - \text{MC}) \approx 100\%$). For all examples, it is worth pointing out that these criteria are optimal for minimum AIC and BIC values (marked in gray). For examples 2–4, the misclassification rates are close to 0, and for examples 5–7 less than 0.46. This is because of the complexity of systems with high dimensions and parsimonious models fit (excess of parameters).

The information measures illustrated a similar effect. It can be seen that inequalities given in (3.12) are accomplished in the Shannon entropy case and the information increase for more parsimonious systems, where these 3D systems are characterized by a bigger set of components and dispersion matrices with large elements. With respect to Rényi entropies, the lower and upper bounds rather slowly increase with more components in examples 2–4, but rise faster with more components in examples 5–7. However, in examples 5–7 the lower and upper bounds are maximum for large α . Therefore, the Rényi information criterion is suitable for model fits with accurate classification of observations, i.e., wrong performance of Rényi entropy is related to inadequate selection of components in complex systems. Additionally, the Rényi entropy of FMSN is localized between the upper and lower bounds and an approximation should be given by the mean of these bounds.

4.3 Skew-normal and modified skew-normal distributions

In this section, we study the behavior of the series expansions of the Shannon entropy and negentropy for the SN and MSN distributions. In both cases, we compare the Shannon entropy and negentropies obtained from their series expansions with their corresponding “exact” versions computed from the QUADPACK numerical integration method (Figure 4.10). More precisely, the “exact” expected values $E\{\zeta_0(\tau Z_\tau)\}$ and $E\{\zeta_0(\tau Z_\tau^*)\}$ are computed using the QUADPACK method as in Arellano-Valle et al. (2013), Contreras-Reyes and Arellano-Valle (2012) or Contreras-Reyes (2016). From the series expansions, the Shannon entropy and negentropies were carried out for $k = 12$ as in Withers and Nadarajah (2014). However they tend to converge for $k = 4$ as in the Gram–Charlier and Edgeworth expansion methods (see

e.g. Hyvärinen et al. (2001) and Stehlík et al. (2014), respectively). All proposed methods are implemented with R software (R Core Team, 2016).

From Figure 4.10, we observe that the approximations by series expansions are better in the MSN case (panels C and D) than in the SN case (panels A and B). Furthermore, that series expansion approximations are quite exact for small to moderate values of the skewness parameter τ . More specifically, for $0 \leq \tau \leq 2$ in the SN case, and $0 \leq \tau \leq 4$ in the MSN case. Additionally, panels A and C show that the Shannon entropy decreases as τ increases, while panels B and D indicate that the negentropy increases with τ . Finally, as expected in both GSN models, the Shannon entropy is less than or equal to the Shannon entropy of the normal model, namely $H(Z_0) \approx 1.418$ (Contreras-Reyes, 2015, Contreras-Reyes and Arellano-Valle, 2012).

Panel A of Figure 4.11 shows, respectively, the behavior of the KL divergences of the SN and MSN distributions from the normal one obtained from the expansions in series given in Eqs. (3.33) and (3.35). As in Figure 4.10, the KL divergence between the SN and normal distributions increases smoothly for values of $\tau \in [0, 2]$, but rises sharply for $\tau > 2$. Meanwhile, the increase in KL divergence between the MSN and normal distributions seems more stable, at least for $\tau \in [0, 5]$. Crucially, for $\tau = |\eta| \geq 2$ the SN model is close to its maximum level of asymmetry, while the MSN model does it for $\tau = |\eta| \geq 5$ (see Figure 2 in Arrué et al., 2016).

Table 4.3 presents the observed power of the asymptotic test of normality obtained from Eq. (3.37) in Section 3.6, for different sample sizes and values of the skewness parameter. All these results were obtained from 2000 simulations for a nominal level of 5%. In each simulation, the MLE of $Z \sim MSN(\eta)$ was obtained by maximizing the log-likelihood function Eq. (3.39). Table 4.3 shows that the proposed test is considerably conservative since the observed rate of incorrect rejections of the normality hypothesis is always lower than the nominal level. The proposed test is also considerably more powerful in large samples ($n \geq 300$) and values of the skewness parameter far from zero ($|\eta| \geq 1.2$). As expected, the power of the test increases with sample size, particularly for small values of the skewness parameter (close to normality), given that statistic $2nK_0$ depends on n despite K_0 is small (Figure 4.11).

Now, we compare the proposed asymptotic test with two additional tests considered by Arrué et al. (2016) for null hypothesis $H_0 : \eta = \eta_0$ versus $H_1 : \eta \neq \eta_0$: the Likelihood Ratio Test (LRT) (Section 3.6), and the asymptotic normality-based test. Since the regularity condition on MSN's FIM at $\eta = 0$ is satisfied, the authors proposed a distributional normal

theory for testing H_0 , i.e., based on asymptotic normality of MLE given by

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N_3(0, i_{MSN}^{-1}(\theta_0)),$$

as $n \rightarrow \infty$, where $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2, \hat{\eta})^\top$ is the MLE of $\theta = (\mu, \sigma^2, \eta)^\top$, $\theta_0 = (\mu, \sigma^2, 0)^\top$, and $i_{MSN}^{-1}(\theta_0)$ is the inverse FIM component related to θ_0 . For asymptotic normality and LRT, they conclude that H_0 is rejected for large values of $\hat{\tau} = |\hat{\eta}|$ and, for large values of n , the coverage rate increase when $\hat{\eta}$ exists (H_0 is rejected) (see Table 3–5 of Arrué et al., 2016). Analogously, in Table 6 of Arrué et al. (2016) the coverage rate increase when $\hat{\eta}$ exists for large values of n .

Table 4.1 Numerical and exact (Proposition 5) method for univariate SN Rényi entropy, $R_\alpha(Z)$, $Z \sim SN_1(0, 1, \eta)$, with $\alpha = 2, \dots, 6$ and $\eta = 0.2, \dots, 1, 1.5, 2, \dots, 6$.

Method	η / α	2	3	4	5	6
Numerical	0.2	1.3738881	1.0273108	0.9117845	0.8540211	0.8193630
	0.4	1.3401742	0.9935483	0.8779974	0.8202192	0.7855512
	0.6	1.2932667	0.9464721	0.8308348	0.7730041	0.7383008
	0.8	1.2414279	0.8942994	0.7784870	0.7205485	0.6857722
	1	1.1904423	0.8428276	0.7267522	0.6686491	0.6337602
	1.5	1.0823640	0.7332379	0.6162880	0.5576157	0.5223270
	2	1.0047402	0.6542817	0.5364793	0.4772159	0.4414939
	3	0.9100381	0.5581948	0.4393628	0.3793110	0.3429680
	4	0.8574065	0.5052829	0.3861362	0.3257878	0.2891777
	5	0.8247224	0.4727606	0.3536358	0.2932526	0.2565831
Exact	6	0.8026627	0.4510083	0.3320375	0.2717340	0.2351037
	0.2	1.3738894	1.0273131	0.9117890	0.8540252	0.8193690
	0.4	1.3401764	0.9935511	0.8780049	0.8202300	0.7855672
	0.6	1.2932684	0.9464738	0.8308416	0.7730247	0.7383349
	0.8	1.2414287	0.8943016	0.7784961	0.7205797	0.6858232
	1	1.1904431	0.8428294	0.7267614	0.6686843	0.6338222
	1.5	1.0823631	0.7332380	0.6162918	0.5576375	0.5223843
	2	1.0047386	0.6542808	0.5364795	0.4772250	0.4415245
	3	0.9100384	0.5581964	0.4393615	0.3793087	0.3429700
	4	0.8574080	0.5052757	0.3861371	0.3257861	0.2891721
	5	0.8247224	0.4727601	0.3536374	0.2932580	0.2565860
	6	0.8026627	0.4510081	0.3320370	0.2717355	0.2351083

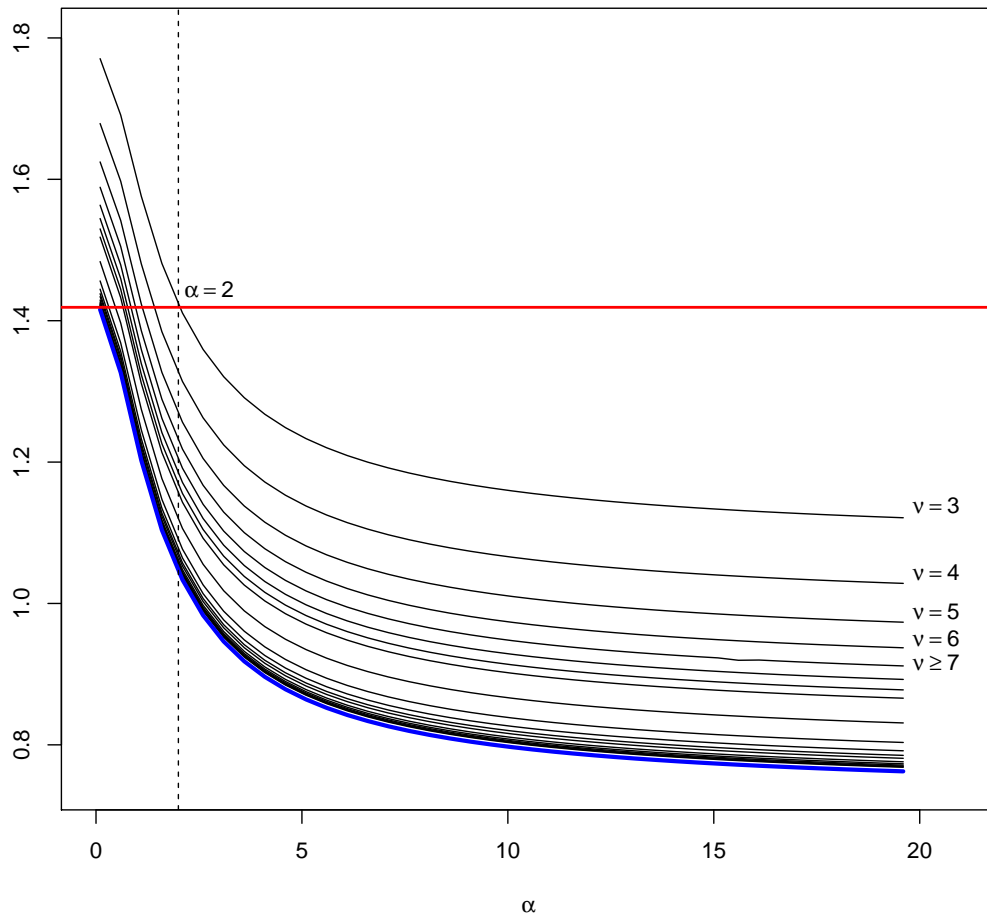


Fig. 4.1 Plots of the ST entropy for $v = 1, 2, \dots, 185$ degrees of freedom as a function of shape parameter α . The red and blue lines correspond to normal and skew-normal entropies, respectively.

Table 4.2 Simulated Shannon and Rényi entropy bounds. Rényi entropy bounds are computed for $\alpha = 2, \dots, 5$. For each model and number of clusters m the AIC and BIC criteria, misclassification (MC), normal skill (NSS), Heidke skill (HSS), and Hansen–Kuipers (HK) scores appear.

Example	m	MC	NSS	HSS	HK	H				R α lower					R α upper				
						A _{lower}	A _{upper}	B _{lower}	B _{upper}	2	3	4	5	2	3	4	5		
1	2	0.02	0.98	0.95	0.39	4831.88	4866.23	0.99	2.20	0.888	1.58	0.72	0.96	1.01	1.03	3.54	3.06	2.89	2.80
	3	0.03	0.97	0.93	0.42	4840.36	4894.34	1.89	2.79	1.58	2.92	0.58	0.84	0.91	0.94	3.52	3.04	2.87	2.78
	4	0.02	0.98	0.96	0.41	4846.62	4920.23	1.85	2.80	1.44	3.03	0.58	0.85	0.93	0.97	3.52	3.04	2.87	2.78
	5	0.61	0.39	0.01	0.01	4851.35	4944.59	2.15	2.91	1.93	3.59	0.50	0.63	0.65	0.65	3.50	3.01	2.85	2.76
	6	0.77	0.23	0.00	0.00	4858.87	4971.75	2.19	2.92	2.04	3.74	0.59	0.70	0.70	0.70	3.50	3.01	2.85	2.76
	2	0.01	0.99	0.98	0.40	6581.52	6615.87	2.57	3.92	0.66	3.26	1.18	1.29	1.32	1.33	4.93	4.45	4.28	4.19
2	3	0.00	1.00	1.00	0.62	6065.25	6119.23	2.96	4.25	0.75	3.99	1.06	1.32	1.38	1.40	4.97	4.49	4.32	4.23
	4	0.51	0.49	0.13	0.08	6071.55	6145.16	2.98	4.25	0.78	4.32	0.79	1.06	1.13	1.16	4.95	4.47	4.30	4.21
	5	0.60	0.40	0.00	0.00	6080.71	6173.95	3.28	4.26	1.55	4.81	0.83	1.06	1.11	1.12	4.95	4.47	4.30	4.21
	6	0.59	0.41	0.00	0.00	6090.82	6203.70	3.62	4.39	2.32	5.33	0.73	0.90	0.94	0.94	4.95	4.47	4.30	4.21
	2	0.00	1.00	1.00	0.46	5766.82	5840.43	3.44	3.94	2.84	4.09	1.68	1.61	1.55	1.50	5.27	4.67	4.46	4.35
	3	1.00	0.00	-0.96	-0.49	5778.71	5891.59	3.66	4.45	1.84	4.72	0.79	0.95	0.98	0.99	6.28	5.69	5.48	5.36
3	4	1.00	0.00	-0.98	-0.50	5785.43	5937.57	3.78	4.62	1.73	5.13	0.62	0.80	0.84	0.86	6.62	6.03	5.82	5.70
	5	1.00	0.00	-0.24	-0.19	5798.10	5989.51	3.89	4.66	1.92	5.31	0.72	0.84	0.85	0.84	6.71	6.11	5.90	5.79
	6	0.26	0.74	0.00	0.00	5758.53	5989.20	4.01	4.79	1.88	5.67	0.64	0.79	0.81	0.81	6.97	6.38	6.17	6.05
	2	0.76	0.24	-0.11	-0.07	8758.63	8832.25	3.35	4.61	0.82	3.92	1.64	1.94	2.02	2.05	6.60	6.00	5.79	5.68
	3	0.30	0.70	0.14	0.05	8282.84	8395.72	4.11	4.84	2.06	5.15	1.36	1.53	1.56	1.56	7.24	6.65	6.44	6.32
	4	0.36	0.64	0.46	0.31	8295.26	8447.40	4.51	5.24	2.06	5.71	1.73	1.83	1.84	1.83	7.86	7.27	7.06	6.94
4	5	0.36	0.64	0.47	0.32	8300.94	8492.34	4.41	5.24	1.78	5.79	1.27	1.39	1.40	1.40	7.86	7.27	7.06	6.95
	6	0.57	0.43	0.21	0.15	8246.46	8477.12	4.64	5.43	1.88	6.17	1.29	1.42	1.43	1.42	8.23	7.64	7.43	7.32
	2	0.64	0.36	-0.03	-0.02	9650.23	9772.92	5.57	6.52	1.61	6.25	2.43	2.54	2.52	2.48	10.26	9.56	9.30	9.17
	3	0.45	0.55	0.02	0.01	9510.53	9697.02	5.66	6.73	1.28	6.65	1.84	2.07	2.13	2.15	10.53	9.82	9.57	9.43
	4	0.53	0.47	0.19	0.13	9513.73	9764.02	5.83	6.78	1.43	7.19	1.68	1.90	1.96	1.98	11.06	10.36	10.10	9.97
	5	0.92	0.08	-0.23	-0.17	9539.02	9853.11	5.99	6.90	1.52	7.49	1.60	1.78	1.80	1.79	11.41	10.70	10.45	10.31
5	6	0.68	0.32	0.05	0.04	9550.44	9928.33	5.93	6.85	1.51	7.63	1.54	1.73	1.76	1.77	11.26	10.56	10.30	10.17
	2	0.62	0.38	-0.04	-0.02	33479.01	33601.70	15.56	16.93	0.64	16.25	7.53	7.70	7.75	7.77	41.50	40.79	40.54	40.40
	3	1.00	0.00	-0.47	-0.32	33019.83	33206.33	16.88	18.18	0.75	17.84	7.45	7.71	7.79	7.82	45.26	44.55	44.30	44.16
	4	0.45	0.55	0.29	0.18	32417.80	32668.10	17.31	18.62	0.73	18.65	7.45	7.77	7.87	7.92	45.82	45.11	44.86	44.72
	5	0.93	0.07	-0.15	-0.12	32346.29	32660.39	17.31	18.63	0.71	18.78	6.95	7.13	7.15	7.15	46.61	45.91	45.65	45.52
	6	0.56	0.44	0.20	0.14	32458.05	32835.95	17.54	18.85	0.73	19.20	7.05	7.27	7.32	7.33	47.26	46.56	46.30	46.17

Table 4.3 Observed power (both in %) of the proposed normality test using MLE of MSN model from 2000 simulations for nominal level 5%, and various values of shape parameter η and sample size n .

n/η	0	0.1	0.2	0.3	0.4	0.8	1.2	1.6	2
50	97.55	37.02	39.45	37.39	37.60	40.07	40.82	48.37	51.29
100	97.65	46.15	45.90	47.90	46.05	50.05	56.10	66.20	75.49
200	98.05	54.25	54.30	53.65	54.50	58.80	70.55	85.20	93.25
300	98.30	57.30	57.70	58.00	58.65	62.00	77.05	92.65	98.25
400	98.55	57.20	57.65	57.15	57.50	64.10	83.95	95.80	99.65
500	98.70	58.55	59.40	59.95	58.45	66.70	86.95	97.45	99.75

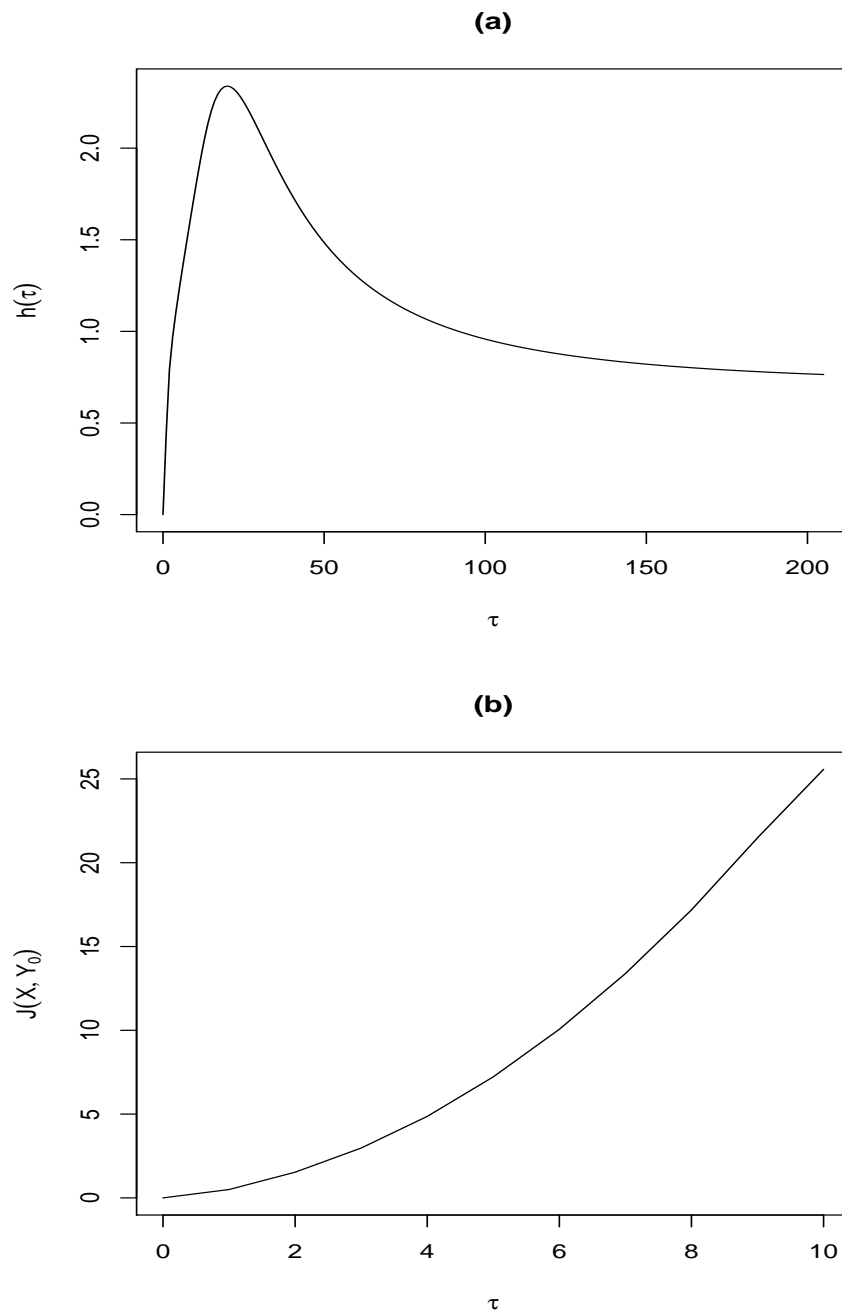


Fig. 4.2 (a) Behavior of $h(\tau) = E[\log\{2\Phi(\tau W)\}]$ for $\tau = 0, 1, \dots, 200$. (b) Behavior of $J(X, Y_0)$ for $\tau = 0, 1, \dots, 10$.

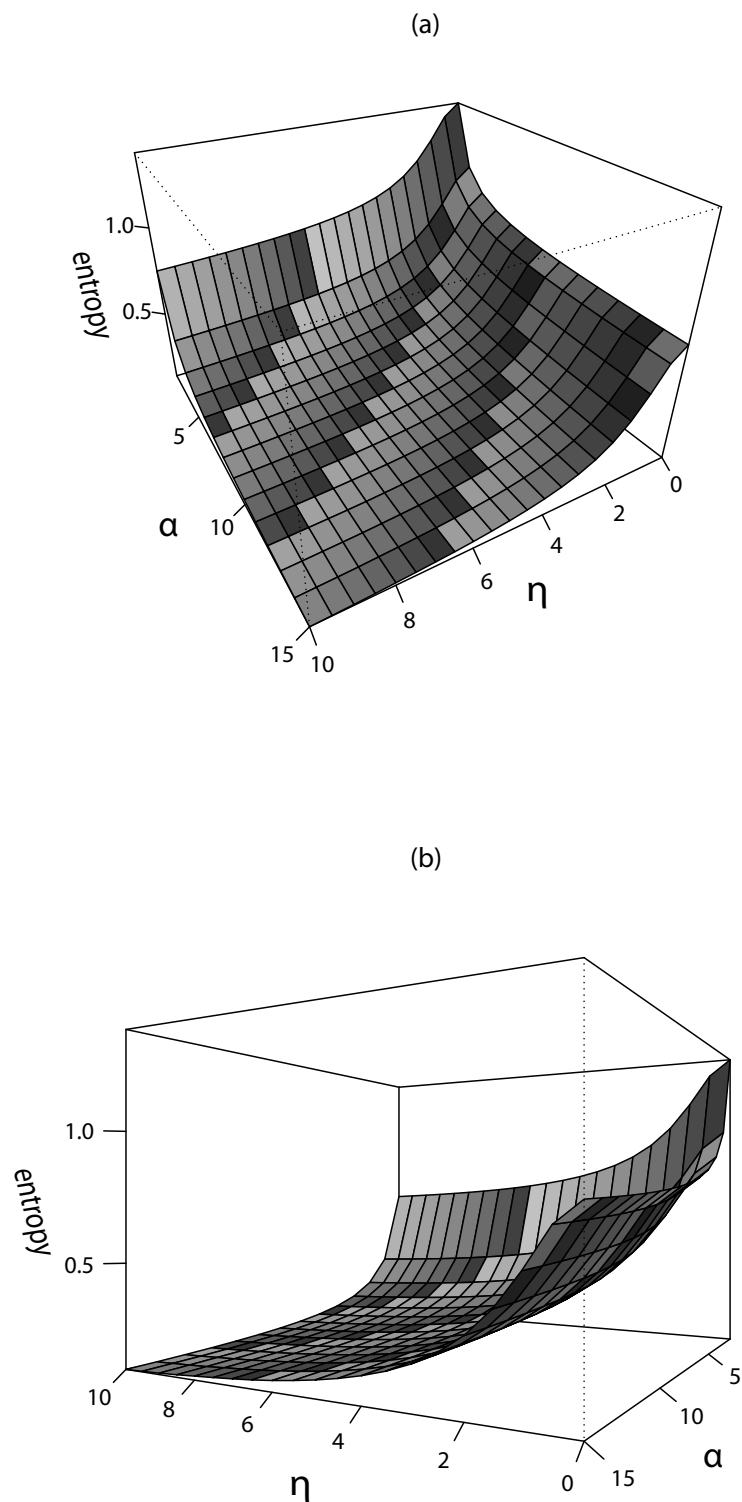


Fig. 4.3 3D-plots of exact univariate SN Rényi entropy (Proposition 5), $R_\alpha(Z)$, $Z \sim SN_1(0, 1, \eta)$, with $\alpha = 2, \dots, 15$ and $\eta = 0, \dots, 10$. Panels (a) and (b) correspond to different angles of 3D-plots.

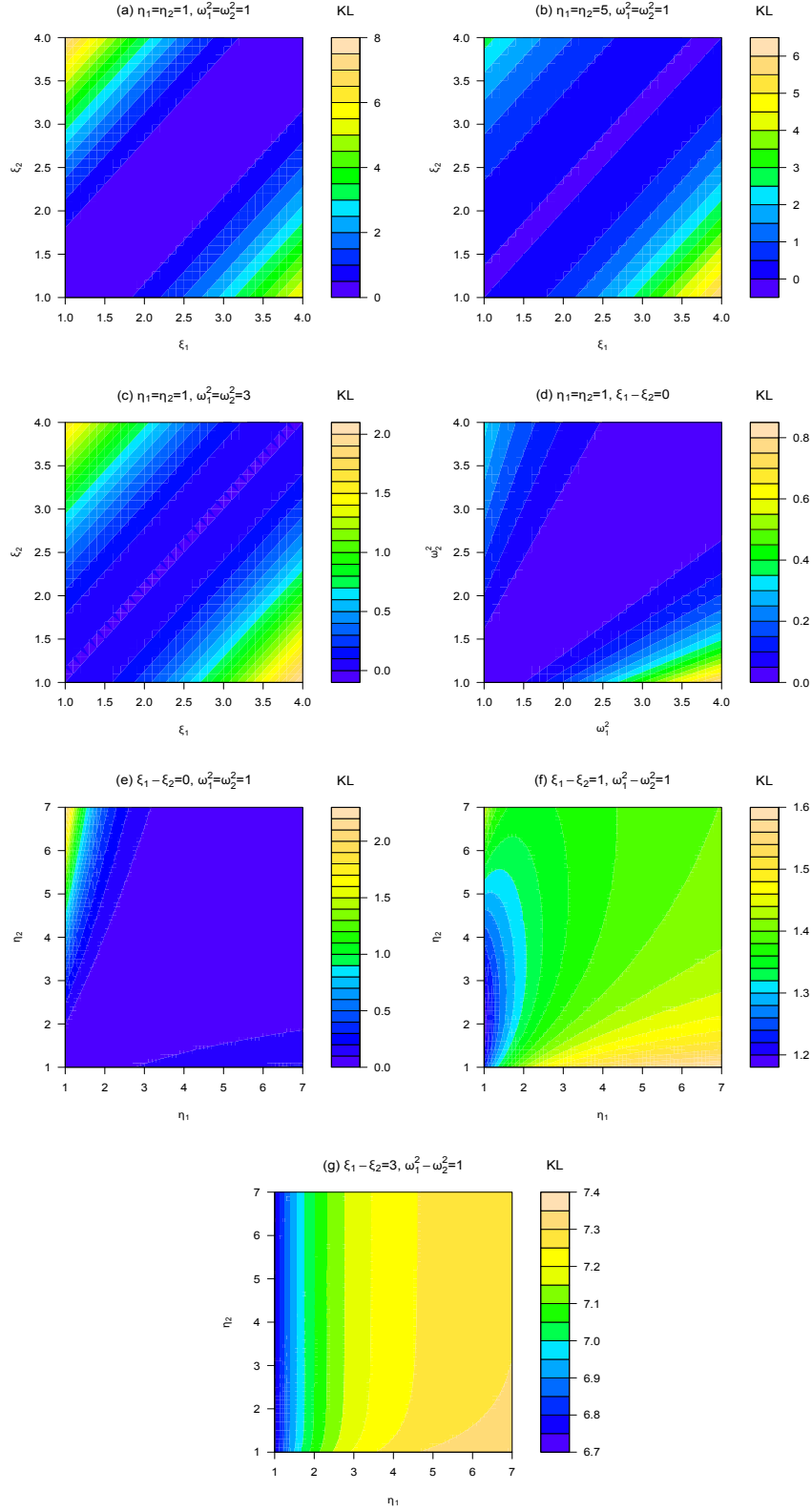


Fig. 4.4 Plots of KL-divergence between $\mathbf{X} \sim SN_1(\xi_1, \omega_1^2, \eta_1)$ and $\mathbf{Y} \sim SN_1(\xi_2, \omega_2^2, \eta_2)$.

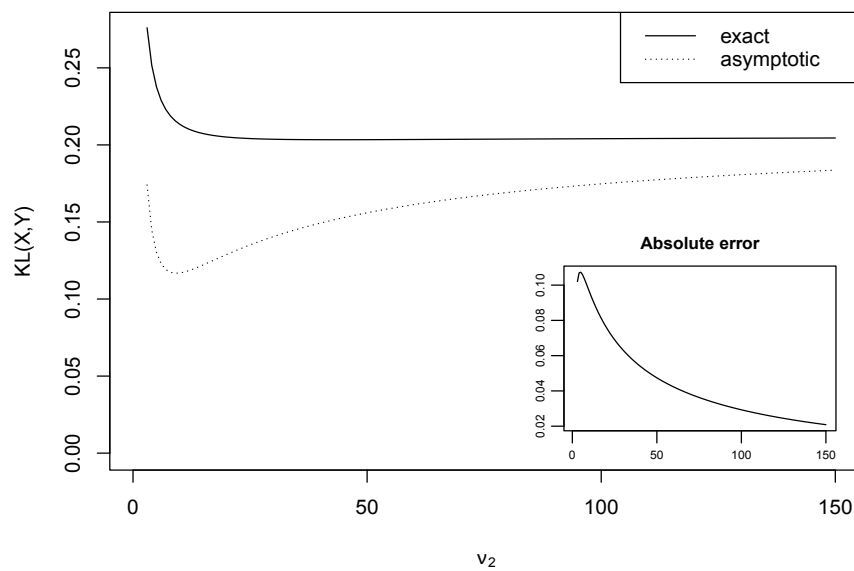


Fig. 4.5 Plot of exact and asymptotic KL divergence between $X \sim ST_1(0, 1.5, 2, 5)$ and $Y \sim ST_1(0, 2.5, 3, v_2)$, with $v_2 = 3, \dots, 150$; and its respective absolute errors.

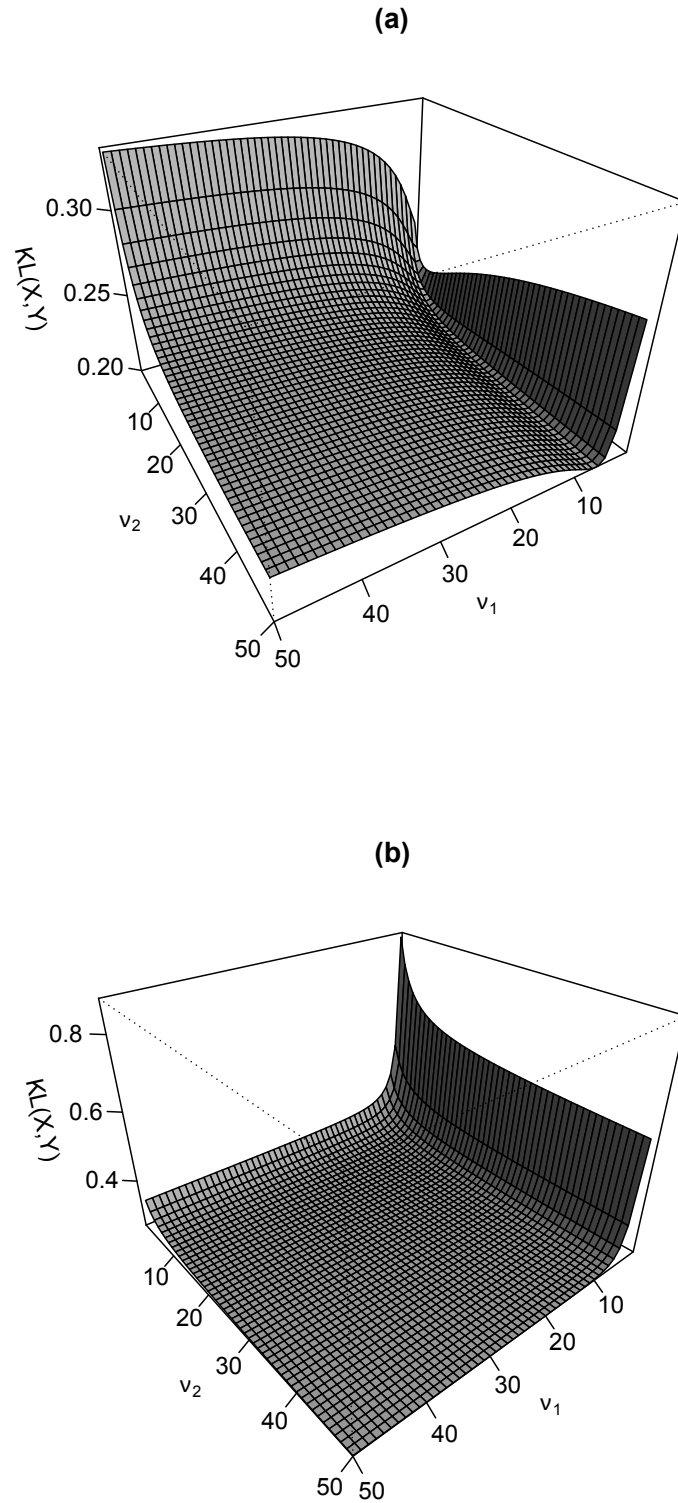


Fig. 4.6 3D-plots of asymptotic KL divergence between (a) $X \sim ST_1(0, 1.5, 2, v_1)$ and $Y \sim ST_1(0, 2.5, 3, v_2)$, and (b) $X \sim T_1(0, 1.5, v_1)$ and $Y \sim T_1(0, 2.5, v_2)$.

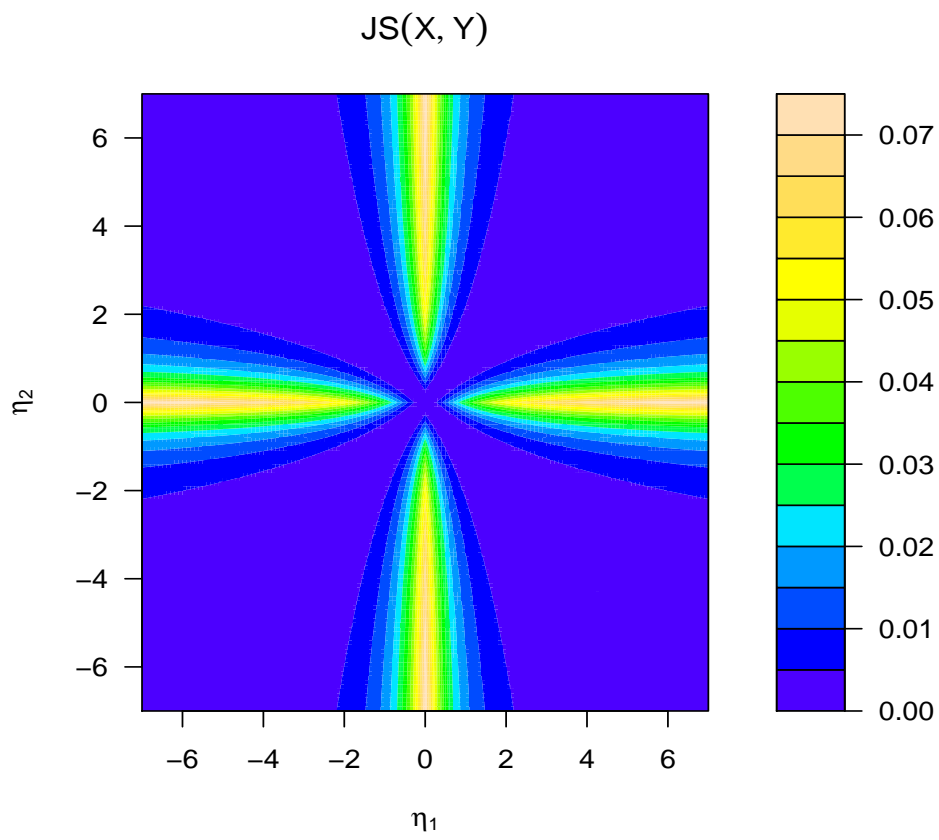


Fig. 4.7 JS distance for shape parameters $\eta_1 \times \eta_2 = [-7, 7] \times [-7, 7]$.

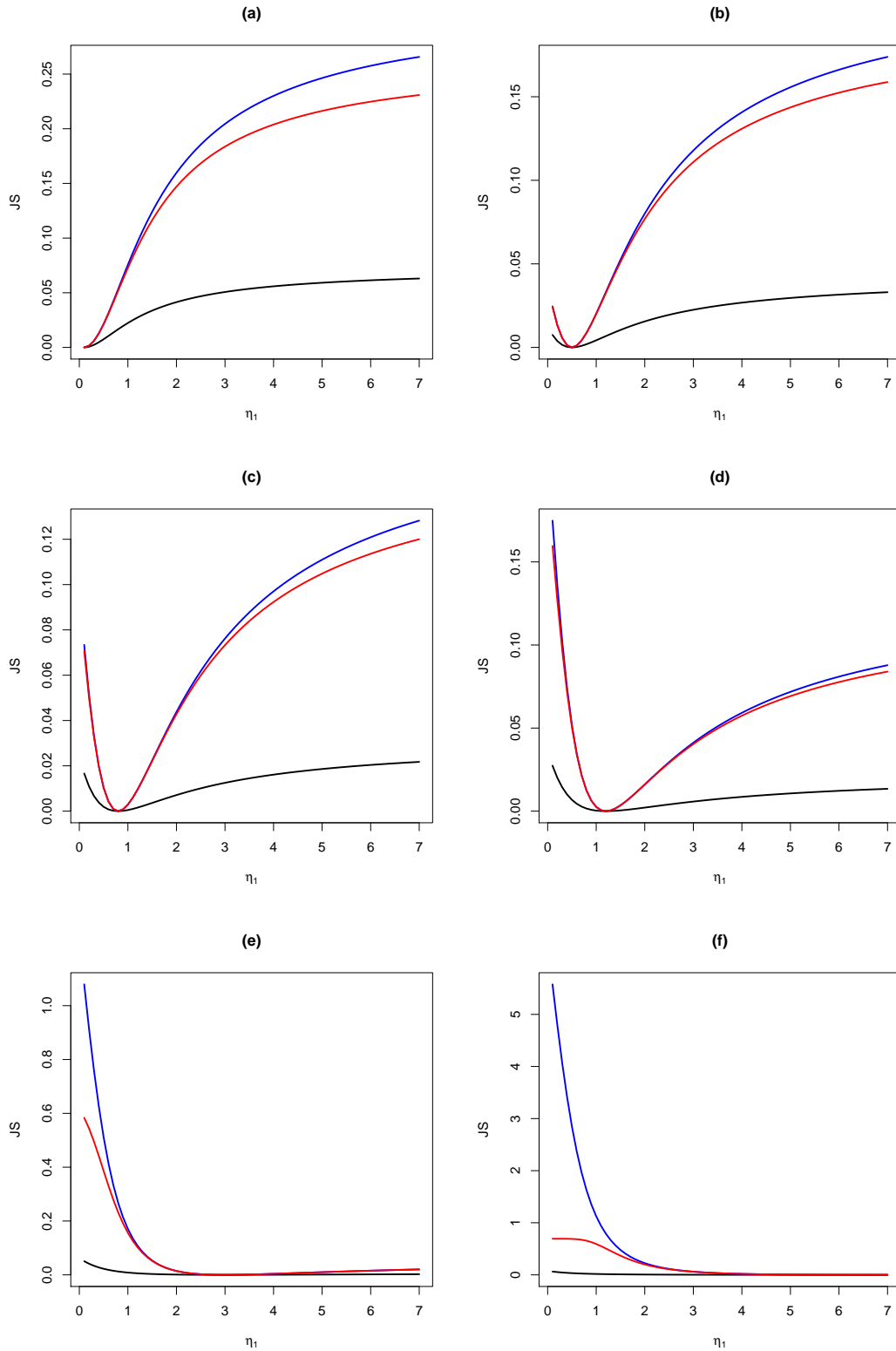


Fig. 4.8 Comparisons of Lin's (blue lines) and Crook's (red lines) upper bounds with JS distance for skew-gaussian distributions and shape parameters $\eta_1 = \{0.1, 0.2, \dots, 7\}$ and (a) $\eta_2 = 0.1$, (b) $\eta_2 = 0.5$, (c) $\eta_2 = 0.8$, (d) $\eta_2 = 1.2$, (e) $\eta_2 = 3$, and (f) $\eta_2 = 7$.

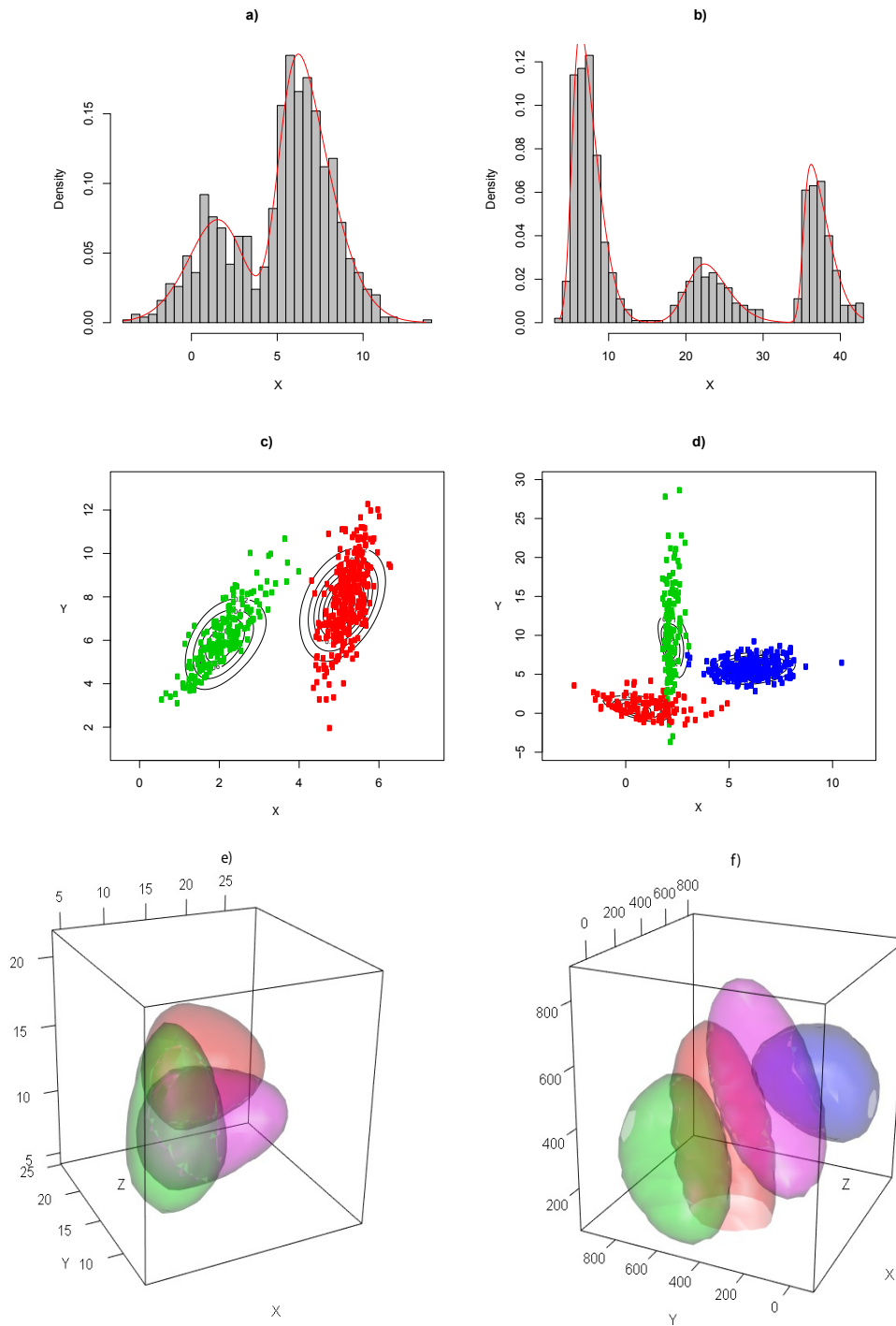


Fig. 4.9 Examples (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, and (f) 6.

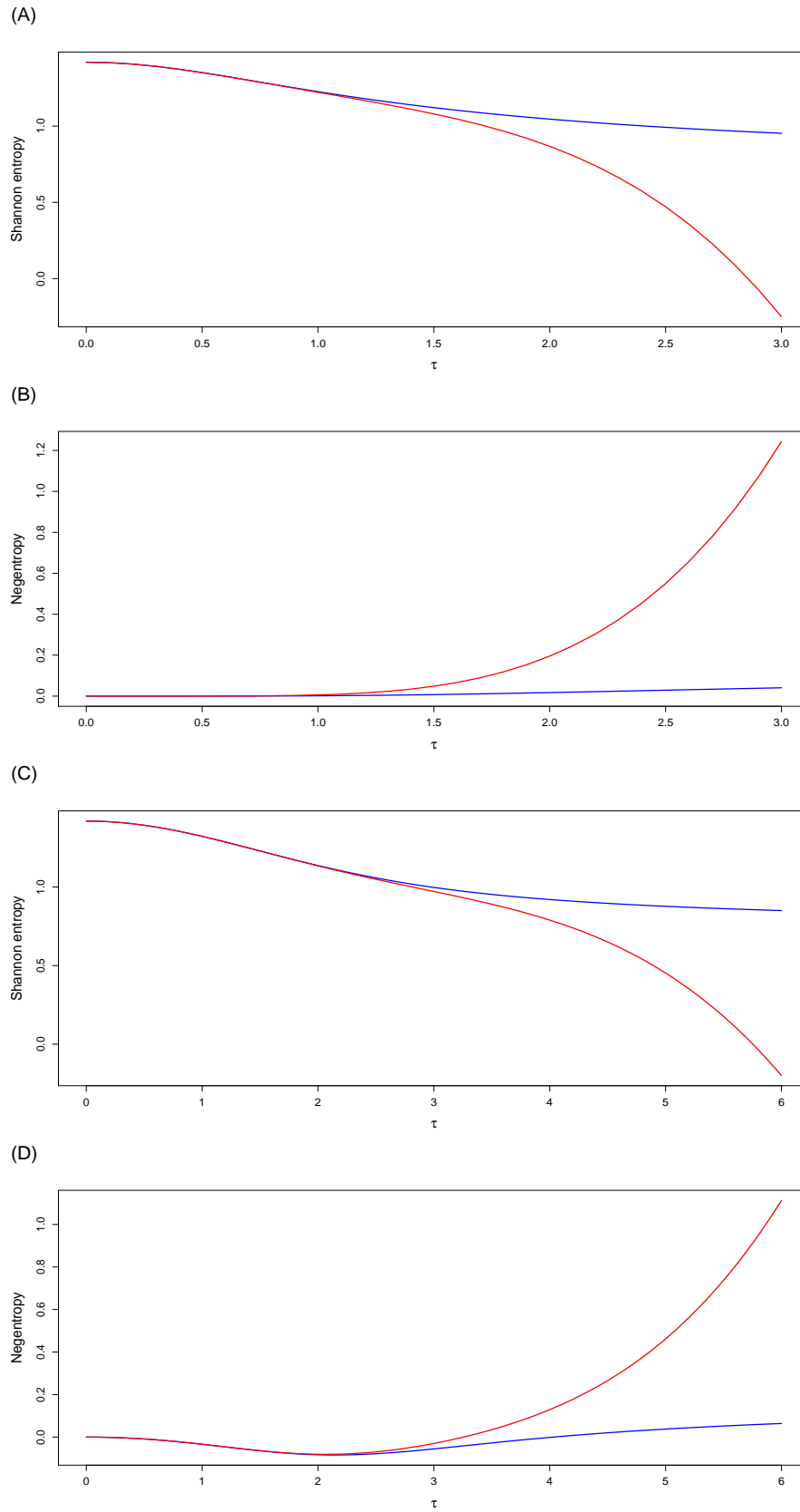


Fig. 4.10 Shannon entropy and negentropy for (A)-(B) SN and (C)-(D) MSN cases. The blue and red lines correspond to numerical integration and cumulant expansion series methods, respectively.

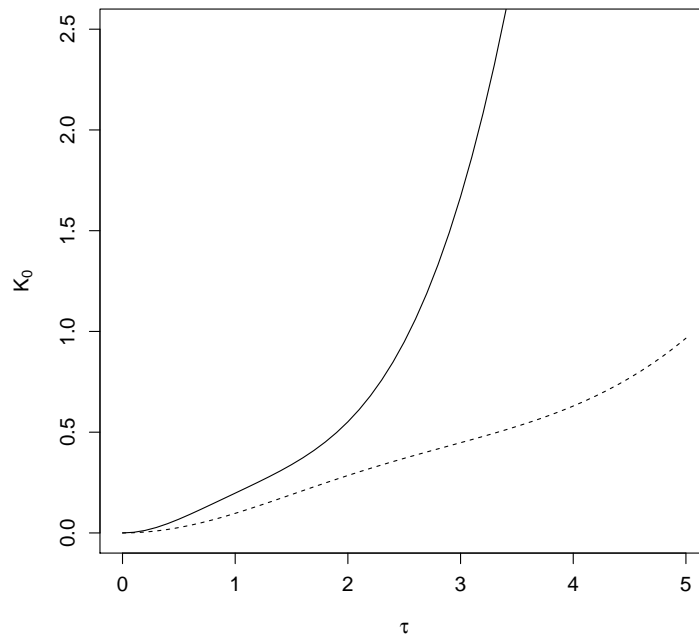


Fig. 4.11 KL divergence, $K_0 = K(Z_\tau, Z_0)$, between SN and normal (solid line) and MSN and normal (dotted line).

Chapter 5

Applications

5.1 Santiago's monitoring 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 and 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 (SEREMI, 2006) of time series of ozone concentrations at 7 monitoring stations denoted by $\mathbf{XY} = \{F, L, M, N, O, P, Q\}$ with $n = 7 \times 24 \times 31 = 5,208$ 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.

5.1.1 Data treatment

In this study, we proceed to analyze 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 :

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

where y_{ij} is the observation for station j at the i th 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 $T_{t,j}^s$ decreases the variance of the time series.

2. In this application, we consider a multivariate data set \mathbf{XY} of 7 stations, a subset \mathbf{X} of 6 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, SN, and ST variables ($W_{\mathbf{X}Y}, W_{\mathbf{X}}$ and W_Y).

3. To find or not evidence to reject the null hypothesis about the marginal variable W_Y having SN or ST 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 and Capitanio (1999), for a sample of independent observations $\mathbf{Z}_i \sim SN_k(\xi, \Omega, \eta)$, $i = 1, \dots, n$, we estimate the parameters by numerically maximizing the log-likelihood function:

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

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

$$\begin{aligned} \log L(\Theta_{ST}) = & n \log 2 + \log |\mathbf{D}|^{n/2} + \sum_{i=1}^n \log t_k(\mathbf{z}_i - \xi; \nu) \\ & + \sum_{i=1}^n \log T \left(\eta^T(\mathbf{z}_i - \xi) \sqrt{\frac{\nu + k}{\nu + s_i}}; \nu + k \right), \end{aligned}$$

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

5. For the variables selected in step 2, let $p_{\mathbf{X}, Y}$ represent the multivariate normal, SN or ST 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 ST case is clearly larger than zero. Let $p_{\mathbf{X}}$ and p_Y be the corresponding marginal densities. Then the mutual information index has been derived in Sections 3.2.1 and 3.2.3 for SN and ST distributions, respectively. If \mathbf{X} and Y are independent,

then $I_{XY} = 0$. We can interpret this as when the monitoring stations do not provide information on the chosen non-monitoring station and vice versa (Silva and Quiroz, 2003). Then, from the MLEs obtained in step 4, we can obtain the terms $\|\bar{\eta}_{X(Y)}\|$, $\|\bar{\eta}_{Y(X)}\|$ and $\|\bar{\eta}_{XY}\|$ mentioned in Proposition 4. So, we can compute the SN and ST 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 and Quiroz (2003). That study analyzed daily averaged data at 8 stations during July of 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 analyze 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:

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

- (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{XY} : complete monitoring network with 7 stations), $k = 6$ (\mathbf{X} : monitoring network with 6 stations and 1 station removed) and $k = 1$ (Y : removed station); and (d) We compute the multivariate normal Shannon index for this case according to the alternative proof showed in Section 3.1.1.

5.1.2 Main results

Figure 5.1 illustrates the behavior 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, 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 5.1 and 5.2, and

in Figure 5.2. They illustrate the flexibility of the class of the ST over SN and normal distributions. We can see in Table 5.2 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 SN distribution. On the other hand, the PP-plot (see Figure 5.2) shows that the ST 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 SN case. In addition, the multivariate and univariate data are normally distributed for the daily average transformation according to Mardia's (joint cases) and Shapiro's (marginal cases) tests (see Table 5.1).

We can see in Figure 5.3 (left panel) through the log-likelihood values that the Akaike's Information Criterion (AIC, Akaike, 1974), $AIC = -2\{\log L(\Theta) - n_p\}$, of the ST model are smaller than the SN 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 Figure 5.3 (right panel) when the period s increases, the v parameter increases too but for the values $s = \{20, \dots, 42\}$, v increases quickly to 120. This behavior may be explained by the good fit of the SN for these s values according to the performance of the ST distribution fit (see Figure 5.2d, 5.2e, 5.2f).

The mutual information index is maximized when the station L is removed from the network XY for the original data and for both SN and ST distributions (see Figure 5.4 and Table 5.3). 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 and 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.

5.2 Chile's seismological catalogue

For a statistical application of SN KL and J divergences, we consider the seismic catalogue of the SSN (2011) analyzed by Contreras-Reyes (2013), containing 6714 aftershocks on a map $[32-40^\circ\text{S}] \times [69-75.5^\circ\text{E}]$ for a period between February 27/2010 13 to July 13/2011. 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 Azzalini and Torelli (2007). 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 Contreras-Reyes (2013), 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 (Figure 5.5). For more details about the NPC method, see also Azzalini and Torelli (2007).

Depending on the case, we consider it pertinent to analyze the measures of J divergences between a cluster sample fitted by a SN distribution versus the same sample fitted by a normal distribution, where the fits are previously diagnosed by QQ-plots (see e.g Arellano-Valle et al., 2013). The MLE's of the model parameters are obtained by using the `sn` package developed by Azzalini (2008) and described in Section 5.1.1; the entropies, cross-entropies, KL and J divergences are computed using `skewtools` package developed by Contreras-Reyes (2012), both packages are implemented in R software (R Core Team, 2016). In Section 5.2.1 we present the Kupperman test (Kupperman, 1957) 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. In order to examine the usefulness of the KL and J divergences between SN distributions, we consider the MLE's of the location.

5.2.1 Asymptotic test

Following Kupperman (1957), Salicrú et al. (1994) and Frery et al. (2011), 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 $\hat{\boldsymbol{\theta}}_1 = (\hat{\theta}_{11}, \dots, \hat{\theta}_{1p})^\top$ and $\hat{\boldsymbol{\theta}}_2 = (\hat{\theta}_{21}, \dots, \hat{\theta}_{2p})^\top$ be the MLE's of the parameter vectors $\boldsymbol{\theta}_1 = (\theta_{11}, \dots, \theta_{1p})^\top$ and $\boldsymbol{\theta}_2 = (\theta_{21}, \dots, \theta_{2p})^\top$, 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

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

Under the regularity conditions discussed by Salicrú et al. (1994), it follows that if $\frac{N_1}{N_1 + N_2} \xrightarrow{N_1, N_2 \rightarrow \infty} \lambda$, with $0 < \lambda < 1$, then under the homogeneity null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$,

$$S_{KL}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) \xrightarrow[N_1, N_2 \rightarrow \infty]{d} \chi_p^2, \quad (5.1)$$

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

Based on (5.1) 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 Frery et al., 2011) between the multivariate SN distributions $\mathbf{X} \sim SN_k(\xi_1, \Omega_1, \eta_1)$ and $\mathbf{Y} \sim SN_k(\xi_2, \Omega_2, \eta_2)$, for which $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the corresponding $p = 2k + k(k+1)/2$ different unknown parameters in $\{\xi_1, \Omega_1, \eta_1\}$ and $\{\xi_2, \Omega_2, \eta_2\}$, respectively. Hence, (5.1) allows testing through the P -value if the homogeneity null hypothesis $H_0 : \xi_1 = \xi_2, \Omega_1 = \Omega_2, \eta_1 = \eta_2$ is rejected or not. Thus, if for large values of N_1, N_2 we observe $S_{KL}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = s$, then the homogeneity null hypothesis can be rejected at level α if $P(\chi_p^2 > s) \leq \alpha$.

5.2.2 Main results

The SN KL divergence values of Proposition 6 for each pair of clusters are reported in Table 5.4. By Proposition (7), we can obtain the symmetrical SN 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 5.5 with its respective descriptive statistics and estimated SN model parameters. Figures 5.6 and 5.7 indicate the performance of the fitted models, where the QQ-plots for the normal and SN cases are included. 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 SN fit will have better performance. The diagnostic QQ-plots are also possible to obtain by using the *sn* package developed by Azzalini (2008).

We can see from Table 5.4 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 due mainly

to the location and shape fitted parameters (see Table 5.5). A counterpart case is founded for the green(3) cluster, which presents the greater differences with respect to these two aforementioned clusters. On the other hand, the diagnostic QQ-plots show good performance of the SN 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 and J divergences based on magnitudes proposed in this Section 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 (5.1). Table 5.6 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 5.5. 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 SN and normal distributions defined in equation (3.10). 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) of Figure 4.2.

5.3 Swordfish age-length-weight data

Estimation of age from growth of swordfish (*Xiphias gladius Linnaeus*) is an important factor in assessing stock trends Quelle et al. (2014). The swordfish belongs to highly migratory pelagic species and has been observed in tropical to temperate waters (between 5 and 27°C), and in western and eastern Pacific and Atlantic (Cerna, 2009). A more detailed description of this species can be found in (Cerna, 2009).

Age and growth estimation of swordfish presents several difficulties (Quelle et al., 2014). For example, Cerna (2009) describes age estimates obtained by cross sections of the second anal fin ray (Sun et al., 2002), which appears an expensive procedure for age estimation.

Quelle et al. (2014) recall the inconclusive results obtained from the indirect validation test. (Roa-Ureta, 2010, , and references therein) maintain that since age is a latent variable, and thus extracting growth information objectively is difficult. He estimates growth parameters using a likelihood function approach underlying a normal mixture model to be applied on squat lobster length dataset, where age is unknown. The normal mixture model components are determined by Akaike's information criterion (AIC) which depends on the sample size and the number of parameters of the mixture.

This application is motivated by the determination of age-length relationship by sex group using information measures. This is presented in a framework format based on the following steps:

- (a) The matrix of data includes both length and weight ($k = 2$) for each observation. Because it is necessary to avoid colinearity, the length-weight regression is computed to show non-linear relationship among both columns.
- (b) Given that the number of components is unknown (age is unknown), the FMSN parameters are estimated considering the 2-dimensional matrix of the last step for several values m .
- (c) The number of components is determined by the bounds of information measures developed in Section 3.2.2 and then compared with AIC and BIC criteria.
- (d) The observed (measures obtained from the procedure of Cerna, 2009) and estimated (by selected mixture model) ages of all observations are compared using a misclassification analysis.

Section 5.3.1 describes the dataset used and Sections 5.3.2 and 5.3.3 describe the results for the steps mentioned (Figure 5.8).

5.3.1 Data and software

The dataset used for evaluating the performance of our findings corresponds to a sample of respectively 486 and 507 swordfish males and females length observations. The samples were collected in the southeastern Pacific off Chile during 2011 and were obtained using the routine sampling program of the fishery conducted by IFOP (2010). All these fish were measured to the nearest centimeter and the range of observed lengths. The catch included fish between 120–257 cm for males, and 110–299 cm for females. As is described in Section 4.2, the FMSN parameter estimates were computed using the `mixsmsn` package.

5.3.2 Length-weight relationship

Following Contreras-Reyes (2016, and references therein), we briefly describe the length-weight function. This function explains the increments in weight of species in terms of their length by the non-linear relationship

$$W(x) = \alpha x^\beta, \quad (5.2)$$

where $W(x)$ represents the observed weight at length x , α is the theoretical weight at length zero and β is the weight growth rate.

The model (5.2) is fitted to an empirical dataset, $(y_i, x_i) \in \mathbb{R}_+ \times \mathbb{R}_+$, $i = 1, \dots, n$. This can be described in terms of multiplicative structure the errors, $y_i = W(x_i)\varepsilon_i$, where ε_i are non-negative random errors and their transformations are given by $\varepsilon'_i = \log \varepsilon_i$. Here, we consider the residuals ε'_i iid and normal distributed, denoted by $N(0, \sigma^2)$, for a dispersion σ^2 parameter.

Figure 5.9 illustrates the linear regression fits of (5.2), for which we have a high value for the R^2 coefficient of determination for both sexes (Table 5.7). There exists a small number of observations of length classes larger than 210 and 250 cm for males and females, respectively, that tends to be isolated with respect to lighter weights. Given the good fitting of length-weight model, we can see that a non-linear relationship could be assumed between length and weight. Therefore, we consider a matrix with two columns constructed by these variables for the clustering modeling.

5.3.3 Clustering and model selection

As in Section 4.2, the length-weight data is evaluated with the FMSN model for several values m depending on the maximum age by sex. Some authors reported that maximum age in males and females reaches 9 and 11 years, respectively (Cerna, 2009, Quelle et al., 2014, Sun et al., 2002). One of the difficulties that anal-fins readers observed, was that they could find multiple annuli and disappearance of the first annulus in older fishes, thus careful interpretation is important (Quelle et al., 2014). Also, this species were aged as younger at given body lengths, i.e., it was difficult to find older fishes by selectivity (Cerna, 2009, Contreras-Reyes et al., 2016). We take into account these facts to discuss the optimal number of clusters for the classification of lengths into age classes.

To reduce the scale of the plots, in Fig. 5.10 appears the logarithmic of the average between upper and lower bounds for Shannon and Rényi entropies, for $m = 1, \dots, 9$ in males and $m = 1, \dots, 11$ in females. It is worth pointing out that the values related to Shannon

entropy (panels (a) and (c)) increase when the number of components increases in both sexes. Panels (b) and (d) show that values related to Rényi entropies increase until $m = 7$ and then decrease. This means that Rényi entropy bounds provide information of the models and help us to determine a criterion to choose a possible number of components on each sex group. There also exist some differences between α values, where the quadratic Rényi entropy ($\alpha = 2$) provides more information.

The results mentioned before are compared first with AIC and BIC criteria in Table 5.8. These criteria increase when the number of components increases, and minimum AIC and BIC values correspond to the simplest model $m = 2$. Table 5.8 also shows the misclassification (MC) rates and consistency scores considered in Section 4.2. All these indicators are applied over the assigned observations for each cluster and the observed age, for each FMSN and FMN (normal) model. The values corresponding to $m = 7$ clusters, marked in gray, provide the best results. The model has a classification rate of 71% and 65% for males and females, respectively; and the highest values of NSS, HSS and HK scores. The best FMN model corresponded to $m = 6$ and 8 for males and females, respectively; where its respective classifications rate was 57% and 55%.

The FMSN fits for length-weight of males and females are shown in Fig. 5.11. The lengths of the older species presents high variability compared to younger ones. The group of males has the parameters $\pi = (0.167, 0.117, 0.159, 0.257, 0.025, 0.084, 0.191)$,

$$\begin{aligned} \xi &= \left(\begin{pmatrix} 175.77 \\ 68.10 \end{pmatrix}, \begin{pmatrix} 192.98 \\ 93.99 \end{pmatrix}, \begin{pmatrix} 141.15 \\ 34.65 \end{pmatrix}, \begin{pmatrix} 155.55 \\ 41.27 \end{pmatrix}, \begin{pmatrix} 211.49 \\ 156.93 \end{pmatrix}, \right. \\ &\quad \left. \begin{pmatrix} 201.31 \\ 105.25 \end{pmatrix}, \begin{pmatrix} 162.76 \\ 53.97 \end{pmatrix} \right), \\ \tilde{\Omega} &= \left(\begin{pmatrix} 6.47 & 2.05 \\ 2.05 & 7.90 \end{pmatrix}, \begin{pmatrix} 8.74 & 4.18 \\ 4.18 & 10.74 \end{pmatrix}, \begin{pmatrix} 8.45 & 4.09 \\ 4.09 & 5.28 \end{pmatrix}, \begin{pmatrix} 7.66 & 1.81 \\ 1.81 & 6.73 \end{pmatrix}, \right. \\ &\quad \left. \begin{pmatrix} 21.63 & 12.29 \\ 12.29 & 14.91 \end{pmatrix}, \begin{pmatrix} 14.16 & 7.83 \\ 7.83 & 18.12 \end{pmatrix}, \begin{pmatrix} 7.20 & 2.34 \\ 2.34 & 6.94 \end{pmatrix} \right), \text{ and} \\ \tilde{\eta} &= \left(\begin{pmatrix} 0.87 \\ 1.06 \end{pmatrix}, \begin{pmatrix} 0.62 \\ -1.35 \end{pmatrix}, \begin{pmatrix} -1.28 \\ -0.99 \end{pmatrix}, \begin{pmatrix} -1.12 \\ 1.23 \end{pmatrix}, \begin{pmatrix} 0.91 \\ 1.19 \end{pmatrix}, \begin{pmatrix} -0.83 \\ 0.87 \end{pmatrix}, \right. \\ &\quad \left. \begin{pmatrix} 0.82 \\ 0.73 \end{pmatrix} \right); \end{aligned}$$

and for females, $\pi = (0.279, 0.058, 0.109, 0.070, 0.240, 0.015, 0.229)$,

$$\begin{aligned}
\tilde{\xi} &= \left(\begin{pmatrix} 193.98 \\ 82.98 \end{pmatrix}, \begin{pmatrix} 236.10 \\ 205.41 \end{pmatrix}, \begin{pmatrix} 207.07 \\ 120.38 \end{pmatrix}, \begin{pmatrix} 221.49 \\ 158.33 \end{pmatrix}, \begin{pmatrix} 153.72 \\ 43.52 \end{pmatrix}, \right. \\
&\quad \left. \begin{pmatrix} 264.01 \\ 283.60 \end{pmatrix}, \begin{pmatrix} 161.01 \\ 51.52 \end{pmatrix} \right), \\
\tilde{\Omega} &= \left(\begin{pmatrix} 11.23 & 4.31 \\ 4.31 & 13.57 \end{pmatrix}, \begin{pmatrix} 16.77 & 6.92 \\ 6.92 & 25.62 \end{pmatrix}, \begin{pmatrix} 9.68 & 4.03 \\ 4.03 & 15.32 \end{pmatrix}, \begin{pmatrix} 14.88 & 5.98 \\ 5.98 & 18.20 \end{pmatrix}, \right. \\
&\quad \left. \begin{pmatrix} 13.01 & 6.91 \\ 6.91 & 9.03 \end{pmatrix}, \begin{pmatrix} 16.87 & 13.23 \\ 13.23 & 58.63 \end{pmatrix}, \begin{pmatrix} 8.86 & 5.29 \\ 5.29 & 10.43 \end{pmatrix} \right), \text{ and} \\
\tilde{\eta} &= \left(\begin{pmatrix} -0.99 \\ 0.67 \end{pmatrix}, \begin{pmatrix} 1.13 \\ 1.25 \end{pmatrix}, \begin{pmatrix} 1.09 \\ 1.11 \end{pmatrix}, \begin{pmatrix} 1.18 \\ 1.33 \end{pmatrix}, \begin{pmatrix} -1.23 \\ -0.86 \end{pmatrix}, \begin{pmatrix} -0.20 \\ 1.49 \end{pmatrix}, \right. \\
&\quad \left. \begin{pmatrix} 0.95 \\ 0.96 \end{pmatrix} \right).
\end{aligned}$$

5.4 Fish condition factor time series

The database used in this study corresponded to the data collected from the routine biological sampling program of the anchovy (*Engraulis ringens*) landings in the North of Chile ($18^{\circ}20'LS$ – $24^{\circ}00'LS$), carried out by IFOP (2010). The samples are collected monthly from the landing ports and transferred to the laboratory for analysis. The sex of each individual in the sample is then identified and the weight (gr) and length (cm) of each individual is registered. The range of observed lengths in the catch included fish between 12 and 18 cm. The data covered the period from January 1990 to December 2010 and the numbers of individuals sampled by year and interval of length are summarized in Contreras-Reyes et al. (2016). The sample includes only 12 to 18 cm length range, because of the impact of length-specific implementation of a minimum landing size (fishing selectivity) (Contreras-Reyes et al., 2014). Given that samples are collected monthly, the CF process is defined in discrete time.

Individual variations from the general length–weight relationship have usually been considered more interesting than the relationship (5.2) itself (with length $x = L$), and have been frequently studied under the general name of *Condition Factor* (CF, Le Cren, 1951). This factor is calculated as a ratio between the observed weight and that expected from the observed length. Considering the Cubillos and Claramunt (2009) version, the CF (%) is computed:

$$CF(L) = \frac{W(L)}{\widehat{W}(L)} 100, \quad (5.3)$$

where $\widehat{W}(L)$ is a theoretical individual total weight obtained by estimating α and β from (5.2). This expression is commonly used for allometric growth when CF per length classes is analyzed. In addition, CF allows to determine the mean body condition or fatness present in a specific population. Hypothetically, values of CF near 100% indicate an ideal or equilibrated body condition, whereas CF values far from this limit are interpreted as abnormal body condition: food deficit in the ocean is $\ll 100\%$ and food abundance is $\gg 100\%$. Nevertheless, the CF of fish should also be affected by biological-environmental factors. For example, CF values lower and higher than 100% are produced by warm and cold events, in terms of sea superficial temperature (Cubillos and Claramunt, 2009). On the another hand, Kawabata et al. (2011) consider another version of CF for biometric data (not considered here), incorporating the cubic length and gonad weight.

Temporal variations in fatness related to growth have not yet been modeled in previous descriptive or correlation analysis studies (see e.g. Brosset et al., 2015a, Cubillos and Claramunt, 2009, Kawabata et al., 2011). However, it is well known that seasonal changes in gonad development, growth and stomach contents are deterministic of a species' condition. These changes emanate from the species' reproductive strategy (energy storage) (Cubillos and Claramunt, 2009). In addition, strong environmental events such as the ENSO cycle (El Niño-La Niña) and upwelling phenomena have been postulated as drivers of the biological processes (Yáñez et al., 2008). Yet, the current biological and ecological discussion ignores two crucial features: threshold due to food limitation, and delay due to development time (Tong, 1990). Given the discontinuity of the piecewise linear function to interpreting CF time series, a lack of ecological credibility also holds for these threshold autoregressive models (van der Meer et al., 2000). However, it fully characterizes the dynamical properties of the time series (Qian and Cuffney, 2012).

5.4.1 Condition factor time series modelling

Model (5.2) can be created as a model of non-linear regression with multiplicative error (Contreras-Reyes et al., 2014). Thus, the following linear regression can be obtained from a logarithmic transformation in both sides of (5.2):

$$\log \{W_{ij}(L_{ij})\} = \alpha^{(j)} + \beta^{(j)} \log L_{ij} + \varepsilon_{ij}, \quad (5.4)$$

where W_{ij} and L_{ij} are associated with the i th observation ($i = 1, \dots, n_j$), and the j th group (1: male, 2: female). We assume in (5.4) that ε_{ij} are independent and gaussian distributed random errors with zero mean and constant variance σ^2 , denoted as $N(0, \sigma^2)$. From (5.4), the estimated monthly body weight is determined by maximum likelihood and overall n_{ij}

observations:

$$\widehat{W}_j(L_j) = \exp(\widehat{\alpha}_j + \widehat{\beta}_j \log L_j). \quad (5.5)$$

Considering (5.3) and (5.5), the monthly CF time series are constructed for an specific group j and length $L = L_t^{(j)}$. Hereafter and for the sake of simplicity, the CF time series are denoted as $x_t \equiv \widehat{CF}_t^{(j)}(L_t^{(j)})$, $t = 1, \dots, n = 21 \times 12$. Linear regression of Eq. (5.4) was used to estimate the weight by group (see Table 5.9). Figure 5.12 illustrates the linear regression fits of (5.4) by group, for which we have in general a high value for the R^2 coefficient of determination. There exist a set of observations of length classes larger than 18 cm that tends to be isolated with respect to lighter weights. The estimates of length-weight regressions (Table 5.9) are used to obtain the CF time series.

5.4.2 Skew-normal process

Tong and Lim (1980, and references therein) consider that limit cycles play a central role in the modelling of cyclical data. More specifically, the authors considers a non-linear and recursive relationship

$$x_t = u(x_{t-1}), \quad (5.6)$$

where $u(\cdot)$ is a continuous function. From the general form (5.6), the threshold autoregressive model of order 1 is extracted:

$$x_t = \delta |x_{t-1}| + \varepsilon_t, \quad |\delta| < 1, \quad (5.7)$$

where $\{\varepsilon_t\}$ is white noise $N(0, 1 - \delta^2)$ and δ is the threshold parameter. According to (5.7), x_t is generated by one of two distinct autoregressive models depending on the level of the lagged variable x_{t-1} . The condition $|\delta| < 1$ ensures weak stationarity of the process (5.7). Therefore, the threshold parameter vary from the 0th to the 100th percentile of the empirical distribution of x_t .

Andêl et al. (1984) deduce the corresponding integral equation for (5.7):

$$g(v; \delta) = \int_{-\infty}^{\infty} \phi(v - \delta|t|; 0, 1 - \delta^2) g(t) dt,$$

and its solution given by (2.25). Therefore, the solution (2.25) corresponds to the standardized SN distribution (Azzalini, 1985). The connection of the process (5.7) with the density (2.25) is explained by its stochastic representation (2.26).

Figure 5.13 simulates the process (5.7) and presents the Hurst exponent (Peng et al., 1994) and fractal dimension (Gneiting et al., 2012) for several values of η . For larger values of η , δ increases and the variance $1 - \delta^2$ of the process decreases. When η is near zero, the Hurst exponent and fractal dimension decreases and increases, respectively, i.e., for small values of η , $|\delta|$ tends to 1 and the process $\{x_t\}$ becomes (weak) non-stationary.

5.4.3 Skew-normal Jeffrey's divergence and Jensen–Shannon distance

In this section, we emphasize the fact that environmental and biological events mentioned should produce extreme values in CF time series, which are represented by kurtosis and asymmetry in CF distribution. Therefore, the episodes of uncertainty could be determined if the proposed information measures identify weak or strong events in these time intervals. Since the time series associated with these phenomena are nonlinear, informational quantifiers could be employed to identify, classify, quantify and interpret occasional events (Carpi et al., 2011). It follows that CF time series should be controlled by the shape parameter of the SN distribution (Azzalini, 1985). For the fitted SN densities, uncertainty episodes are evaluated by J divergence as well as JS distance to compare length classes.

CF time series are related to each length class and group. Since the shape parameter η is not affected by a linear transformation of X (Azzalini and Capitanio, 2013), CF time series were previously standardized. Table 5.10 shows the shape parameter η estimates of SN fits related to these time series. In terms of the SN distribution, negative and positive $\hat{\eta}$ values correspond to asymmetry to the right and left, respectively (see SN fits of Figure 5.14 and property 2 of Section 2.4.1). This means that CF of the above mentioned classes are affected by extreme events. For easier comparisons of CF values with respect to 100%, Figure 5.14 illustrates this last fact for nonstandardized CF time series:

- a) For males, the smallest shape parameter estimate was $\hat{\eta} = -3.092$ for length $L = 17$ cm and with $|\hat{\delta}| = 0.951$, i.e., the process is close to be non-stationary. In this class, the smallest CF value (39.9%) was produced in 05/1992. The largest shape parameter estimate was $\hat{\eta} = 1.778$ for length $L = 13$ cm and with $\hat{\delta} = 0.872$, and the largest CF value (125%) was produced in 12/1996.
- b) For females, the smallest shape parameter estimate was $\hat{\eta} = -1.581$ for length $L = 18$ cm and with $|\hat{\delta}| = 0.845$. The smallest CF value (68.6%) was produced in 08/2008. Same as with males, the largest shape parameter estimate was $\hat{\eta} = 2.267$ for length $L = 13$ cm and with $\hat{\delta} = 0.915$, i.e., the process is close to be non-stationary. In this class, the largest CF value (129%) was produced in 11/1996.

In general, the SN fit works well for all classes and groups (Figure 5.14). Merging all observations for each group ($L = \text{ALL}$), the shape parameter estimates are near zero because these CF time series include normal and extreme events.

Comparing the J divergence with JS distance in Figure 5.15, the previous study of η estimates prevails for the magnitudes of these measures. Specifically for males, the length class $L = 17$ cm produces the higher values of J divergence with the length classes $L = 13, 16, 18$ cm (Figure 5.15a). In addition, JS distance highlights the discrepancy of the length class $L = 17$ with $L = 15$ cm (Figure 5.15b). For females, J divergence produces clear discrepancies for the length class $L = 18$ cm with respect to other classes (Figure 5.15c). JS distance also highlights the discrepancy of length class $L = 13$ cm with classes $L = 12, 14, 15, 17, \text{ALL}$ cm (Figure 5.15d), since the high value of η estimate (Table 5.10).

Our results indicate that the body condition of male anchovies with higher lengths ($L = 17$ cm) is susceptible to environmental variability and coincides with the ending of the moderate-strong El Niño event 91–92 reported by NOAA (2015). The reported CF value (39.9%) are far from the ideal body condition and is produced by the permanence of warm waters, which changed into cold waters after 1992. For both males and females, the smaller lengths ($L = 13$ cm) coincide with the beginning of the strong El Niño 97–98 event (NOAA, 2015). In this case, reported CF values (125% for males and 129% for females) are higher than the ideal body condition (fat condition), and appear before the warming.

This method could be applied to other species, too, depending on the fact that CF time series follows a threshold autoregressive (TAR) process. For the dataset used here, the SN model performed well under kurtosis and asymmetry presented in the CF time series. It can be noted that shape parameter estimate can be affected by the considered period of time series and the number of ENSO events. For the cases of the existence of seasonal patterns and with presence of more extreme events, the TAR process can be generalized with a drift term (Tong, 1990) and by considering a set of autoregressive covariables that involve external/environmental factors, such as sea surface temperature, chlorophyll, plankton, and others (Brosset et al., 2015b). Researchers should consider distributions that control extreme observations, such as ST distribution (Azzalini and Capitanio, 2003, Branco and Dey, 2001). However, the link of this distribution family with TAR process is not clear and requires more investigation.

5.4.4 Normality test

We applied hypothesis testing developed in Section 3.6 to monthly CF time series. CF were previously standardized, since the shape parameter η is not affected by a linear transformation of the CF (Azzalini and Capitanio, 2013). Table 5.10 shows the $\hat{\eta}$'s assuming an SN and

MSN distribution based on the MLE method of Azzalini (2008) and Arrué et al. (2016), respectively. For MSN, we considered the log-likelihood function of Eq. (3.39). In both models, negative and positive values of $\hat{\eta}$ correspond to asymmetry to the right and left, respectively Contreras-Reyes (see Figure 5 of 2016). This means that CF of the above mentioned classes are affected by extreme events. As expected, we find generally that for low values of the empirical skewness index, the shape parameter of both distributions is close to zero.

Since that SN model is not regular at $\eta = 0$, we used only the MSN model to perform the test of normality and LRT for each sample data. The results of this analysis appear in Table 5.11, and are not analogous for all the length classes in both groups. In fact, for the group of males, the null hypothesis $H_0 : \tau = 0$ is not rejected, only in length classes 15 (95% confidence level) and in class ALL (90% confidence level). In contrast, for the group of females the null hypothesis is not rejected for length classes 12, 15, 17 (95% confidence level) and in class ALL (90% confidence level). For both tests, we obtained similar decisions on each time series.

According to Contreras-Reyes (2016), the time series in which the shape parameter is close to zero or when the null hypothesis is not rejected, are influenced simultaneously by both normal and extreme events as in the length class ALL, where all fish population is included for the analysis. For length class 17 in males, for example, the CF is susceptible to some atypical events such as the moderate-strong El Niño event between 1991 and 1992 (high negative empirical skewness and high empirical kurtosis). For length class 13 in both sexes, the CF is susceptible to the strong El Niño event produced between 1997 and 1998.

5.5 Other applications

More recent applications have dealt with the calculus of Shannon entropy and mutual information for SE distributions. In this sense, Ormerod (2011) addresses the variational inference of SN distributions using the entropy measure as an approximation. Challis and Barber (2012) considers the KL divergence minimisation problem between a given target density and an approximated variational density, where the SN distribution is considered as a special case. The work of Eltoft et al. (2012) is highlighted that analyze polarimetric synthetic aperture radar (PolSAR) data assuming non-normal distributions, showing that non-gaussian entropy produces the clearest discrimination of buildings and man-made structures of a San Francisco area image. Kundu et al. (2013) considered the Lemma XX to develop the Shannon entropy of Generalized multivariate Birnbaum–Saunders distributions. Main et al. (2016) evaluated the local effect of asymmetry deviations from normality using the

KL divergence measure of SN distribution, and then compared the local sensitivity with Mardia's and Malkovich-Afifi's skewness indexes. They also agree on the use of the SN model to regulate the asymmetry of an empirical distribution because it reflects the deviation in a tractable way. Youssef et al. (2016) propose a fault detection and estimation approach using SN KL divergence to cope with the negative effects due dimension reduction while using Principal Component Analysis. De Queiroz et al. (2016) extended previous works presented in this chapter focuses on the study of the Shannon entropy and KL divergence of the multivariate log-canonical fundamental SN (LCFUSN) and canonical fundamental SN (CFUSN) families of distributions, with the log SN (LSN) as a special case. They used the Shannon entropy to compare models fitted to analyze monthly USA precipitation data, and the KL divergence to cluster regions in the Atlantic ocean according to their air humidity level. Godoi et al. (2017) presented a methodological application using exact expressions of KL and J divergences between a ST and Student- t distribution (developed in Corollary 2.1 and Proposition 2.3 using the Lemma XX). They used the concentration function to analyze departure from a symmetric baseline prior through multiplicative contamination prior distributions for the location parameter in a Gaussian model. Madani et al. (2017) measured the dissimilarity between original and estimated distributions through SN KL divergence for firefighter's tool, where the original distribution is related to target-items (i.e. fires and humans supposed attracting the firefighters' attention and thus their eye-fixation meeting points).

Table 5.1 P-values for Mardia's goodness-of-fit test of multivariate normality applied to the joint \mathbf{XY} 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% confidence level lead to multivariate normality used in the last column to compute the mutual information index, $I(\mathbf{X}, Y)$, 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{XY}	\mathbf{X}	\mathbf{XY}	\mathbf{X}	Y	$I(\mathbf{X}, Y)$
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 5.2 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% confidence level.

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

Table 5.3 Summary of results for SN and ST distributions and different network configurations. The highest values for each transformation is marked in bold.

s	SN							ST						
	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

Table 5.4 KL divergences for pairs of clusters.

	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

Table 5.5 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); SN 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					SN 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 5.6 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 per cent significance level.

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

Table 5.7 Summary of estimates $\alpha' = \log \alpha$ and β with their respective standard errors in parenthesis, for each length-weight log-transformed relationships of Eq. (5.2) and sex.

Sex	parameter	estimate (SE)	<i>t</i> value	<i>p</i> value	$R^2(\%)$
Male	α'	-11.619 (0.202)	-57.53	< 0.01	92.6
	β	3.064 (0.040)	77.58	< 0.01	
Female	α'	-12.413 (0.176)	-70.43	< 0.01	94.7
	β	3.218 (0.034)	94.95	< 0.01	

Table 5.8 Summary of FMSN and FMN clustering for swordfish data. For each model and number of clusters m the misclassification (MC), normal skill (NSS), Heidke skill (HSS), and Hanssen–Kuipers (HK) scores appear.

		Male						Female					
Model	m	MC	NSS	HSS	HK	AIC	BIC	MC	NSS	HSS	HK	AIC	BIC
FMSN	2	0.70	0.30	0.01	0.01	7742.24	7805.03	0.75	0.25	0.00	0.00	8834.89	8898.32
	3	0.77	0.23	-0.05	-0.04	7754.18	7850.46	0.87	0.13	-0.05	-0.04	8844.91	8942.17
	4	0.62	0.38	0.14	0.10	7741.22	7871.00	0.89	0.11	-0.09	-0.07	8838.63	8969.71
	5	0.42	0.58	0.42	0.30	7751.47	7914.73	0.90	0.10	-0.10	-0.08	8847.75	9012.67
	6	0.45	0.55	0.43	0.35	7760.76	7957.51	0.83	0.17	-0.04	-0.03	8864.74	9063.48
	7	0.29	0.71	0.61	0.46	7770.85	8001.10	0.35	0.65	0.56	0.46	8865.46	9098.03
	8	0.51	0.49	0.37	0.30	7783.31	8047.05	0.69	0.31	0.14	0.11	8879.20	9145.59
	9	0.65	0.35	0.22	0.18	7769.48	8066.70	0.49	0.51	0.42	0.35	8885.98	9186.20
	10	-	-	-	-	-	-	0.59	0.41	0.31	0.27	8897.56	9231.61
	11	-	-	-	-	-	-	0.65	0.35	0.26	0.22	8900.87	9268.75
	FMN	2	0.70	0.30	0.02	0.01	7818.79	7864.84	0.75	0.25	0.00	0.00	8914.32
3		0.78	0.22	-0.04	-0.03	7737.23	7808.40	0.88	0.12	-0.03	-0.03	8848.43	8920.31
4		0.52	0.48	0.28	0.20	7729.77	7826.05	0.87	0.13	-0.08	-0.06	8818.25	8915.51
5		0.43	0.57	0.43	0.32	7733.33	7854.73	0.82	0.18	-0.05	-0.04	8820.12	8942.74
6		0.43	0.57	0.46	0.36	7744.00	7890.52	0.70	0.30	0.11	0.09	8831.16	8979.16
7		0.53	0.47	0.35	0.29	7738.41	7910.04	0.66	0.34	0.17	0.14	8839.63	9013.00
8		0.52	0.48	0.36	0.29	7750.27	7947.02	0.45	0.55	0.47	0.39	8849.82	9048.56
9		0.85	0.15	-0.01	-0.01	7751.24	7973.11	0.50	0.50	0.41	0.35	8855.10	9079.21
10		-	-	-	-	-	-	0.78	0.22	0.11	0.10	8857.49	9106.97
11		-	-	-	-	-	-	0.62	0.38	0.29	0.25	8852.37	9127.22

Table 5.9 Linear regression fit parameters, $\alpha' = \log \alpha$ and β of Eq. (5.2), for length-weight log-transformed relationships with its respective standard errors in parenthesis, for each length class L and group j .

Group (j)	parameter	estimate (SE)	t value	p value	$R^2(\%)$
Males (1)	α'	-4.775 (0.005)	-942.1	< 0.01	83
	β	2.923 (0.002)	1569.6	< 0.01	
Females (2)	α'	-4.791 (0.005)	-955.9	< 0.01	84
	β	2.930 (0.002)	1602.9	< 0.01	

Table 5.10 Shape parameter estimates ($\hat{\eta}$) of SN and MSN models for each sex and length class L , together with its respective standard deviations (s.d). Sample size (n), empirical skewness ($\sqrt{\hat{b}_1}$) and kurtosis (\hat{b}_2), and log-likelihood function $\ell(\hat{\eta})$ for each model fit are also reported.

Sex	L	n	$\sqrt{\hat{b}_1}$	\hat{b}_2	$\hat{\eta}$	SN		MSN		
						s.d	$\ell(\hat{\eta})$	$\hat{\eta}$	s.d	$\ell(\hat{\eta})$
Male	12	213	-0.220	3.723	-1.065	0.147	-301.066	-1.717	0.662	-235.783
	13	238	0.658	5.120	1.778	0.134	-332.638	2.617	0.447	-359.291
	14	250	-0.147	2.885	-1.086	0.166	-367.935	0.459	0.542	-342.881
	15	251	-0.030	2.755	-0.442	0.150	-369.834	0.065	0.386	-450.489
	16	251	0.307	3.138	1.616	0.149	-367.542	2.565	0.415	-354.073
	17	221	-3.461	27.958	-3.092	0.080	-285.125	-3.597	0.524	-312.701
	18	180	0.264	3.001	1.368	0.201	-253.988	2.821	0.643	-269.193
	All	252	0.068	2.687	0.721	0.178	-371.192	-0.287	0.517	-131.335
Female	12	198	0.041	2.738	0.552	0.196	-280.434	-0.142	0.386	-209.551
	13	228	0.917	6.103	2.267	0.128	-315.305	3.101	0.524	-326.665
	14	250	0.190	2.907	1.242	0.157	-367.555	1.459	0.867	-331.373
	15	250	0.076	2.672	0.728	0.174	-368.344	-0.189	0.383	-425.212
	16	251	0.349	3.091	1.702	0.155	-367.160	2.631	0.475	-346.705
	17	246	-0.056	3.115	-0.689	0.149	-348.487	0.041	0.754	-373.068
	18	208	-0.539	4.349	-1.581	0.136	-291.484	-2.223	0.472	-313.160
	All	252	0.072	2.764	0.748	0.172	-371.174	-0.267	0.401	-110.108

Table 5.11 MSN Shannon entropy (H) and negentropy (N) for each sex and length class L using expansion series of cumulants. For each time series, the KL divergence $K_0 = K(Z_{\hat{\tau}}, Z_0)$, statistic $2nK_0$ of Eq. (3.37), Likelihood Ratio Test (LRT) statistic, and its respective p-values are reported. All values reported consider estimates $\hat{\eta}$ (for $\hat{\tau} = |\hat{\eta}|$) and sample size n from Table 5.10.

Sex	L	H	N	K_0	Asymptotic test		LRT	
					$2nK_0$	p-value	Statistic	p-value
Male	12	1.187	-0.074	0.232	98.950	< 0.001	98.304	< 0.001
	13	1.030	-0.063	0.389	185.087	< 0.001	180.469	< 0.001
	14	1.396	-0.008	0.023	11.699	0.001	11.693	0.001
	15	1.418	0.000	0.001	0.242	0.623	0.246	0.620
	16	1.038	-0.066	0.381	198.700	< 0.001	193.567	< 0.001
	17	0.873	0.050	0.546	241.472	< 0.001	211.924	< 0.001
	18	0.999	-0.047	0.420	151.341	< 0.001	146.427	< 0.001
	All	1.410	-0.003	0.009	4.682	0.031	4.673	0.030
Female	12	1.417	-0.001	0.002	0.873	0.350	0.903	0.342
	13	0.956	-0.019	0.463	211.320	< 0.001	198.516	< 0.001
	14	1.237	-0.062	0.183	94.892	< 0.001	93.300	< 0.001
	15	1.415	-0.001	0.004	2.027	0.155	2.071	0.150
	16	1.028	-0.062	0.391	204.117	< 0.001	199.015	< 0.001
	17	1.418	0.000	2e-04	0.091	0.763	0.089	0.765
	18	1.095	-0.080	0.324	134.967	< 0.001	133.629	< 0.001
	All	1.411	-0.003	0.008	4.058	0.044	4.112	0.042

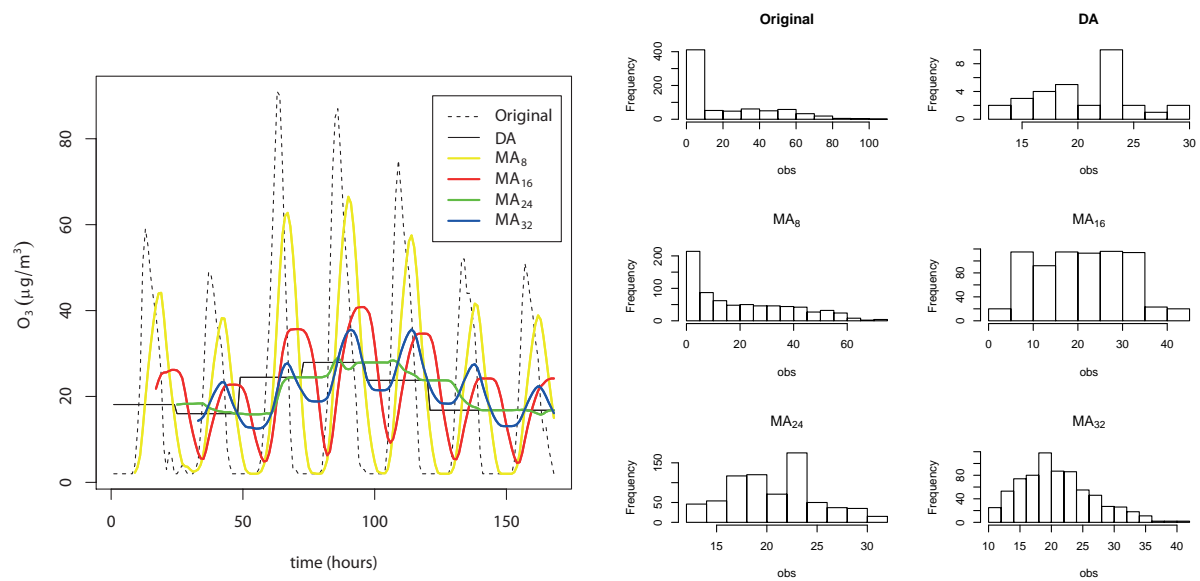


Fig. 5.1 Left: Graphic of Original Data ($s = 1$) with the transformations of moving average (MA_s) for $s = \{8, 16, 24, 32\}$ hours and daily average (DA) for 01/03/2006 to 07/03/2006 of station L. Right: Several histograms for the transformed ozone data mentioned before.

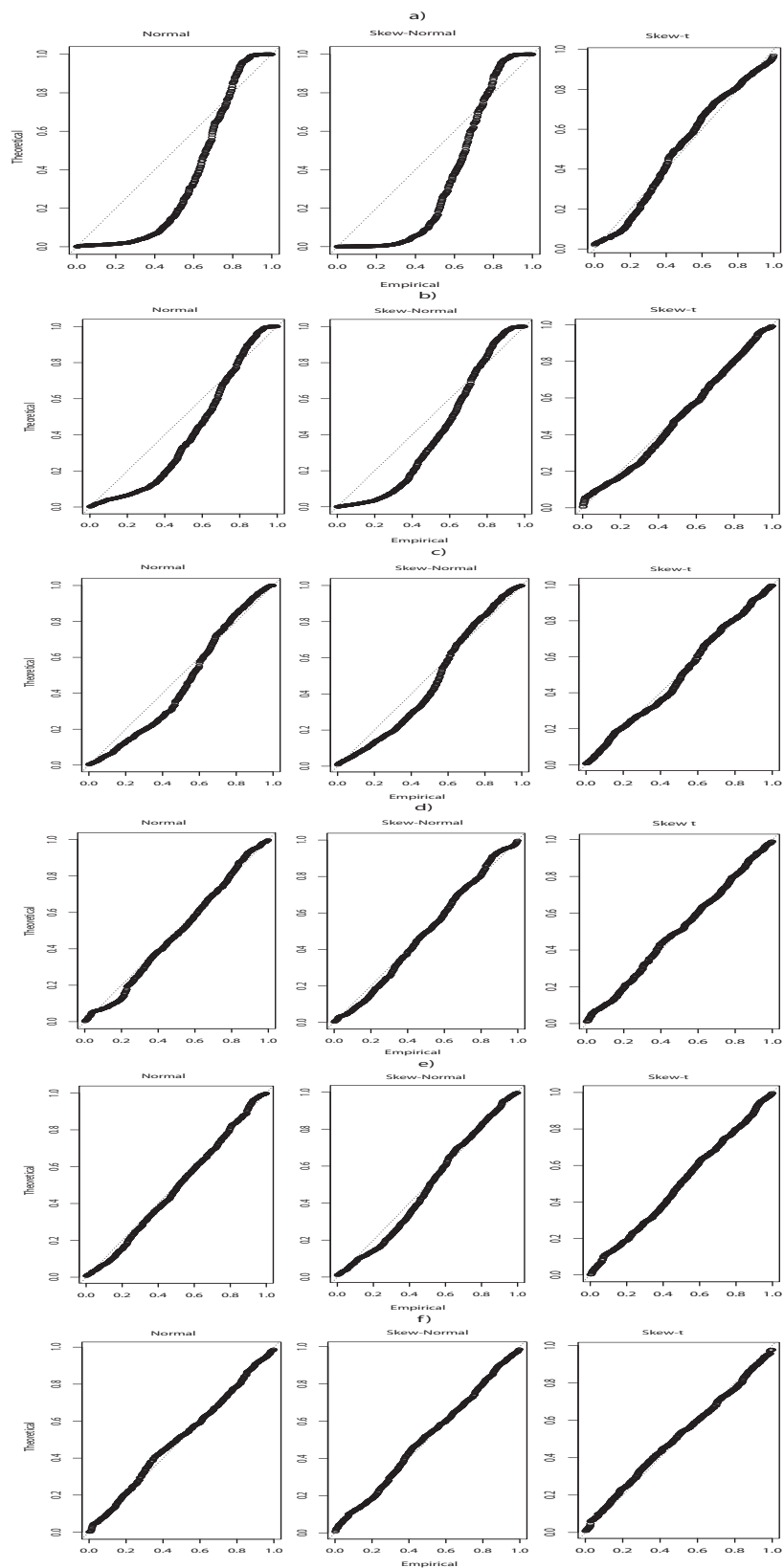


Fig. 5.2 Multivariate normal, SN, and ST PP-plots of: a) original data ($s = 1$); and transformed data with b) $s = 8$, c) $s = 16$, d) $s = 24$, e) $s = 32$ and f) $s = 40$.

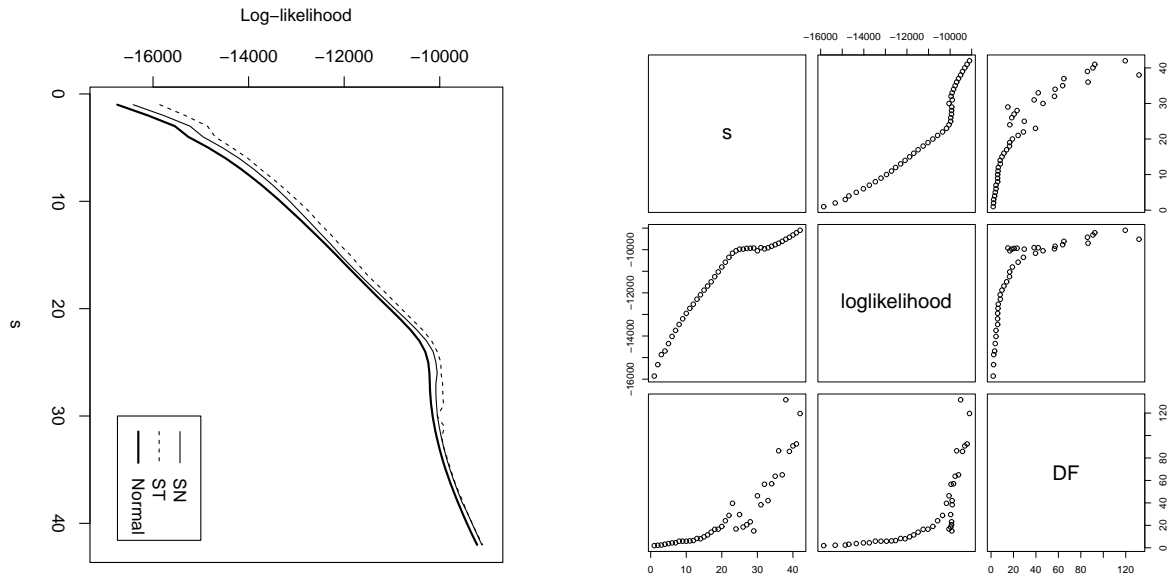


Fig. 5.3 Left: Graphic of log-likelihoods of multivariate SN, ST and normal fits. Right: Scatter plots between s , log-likelihood of multivariate ST fit, and degrees of freedom (DF) parameters.

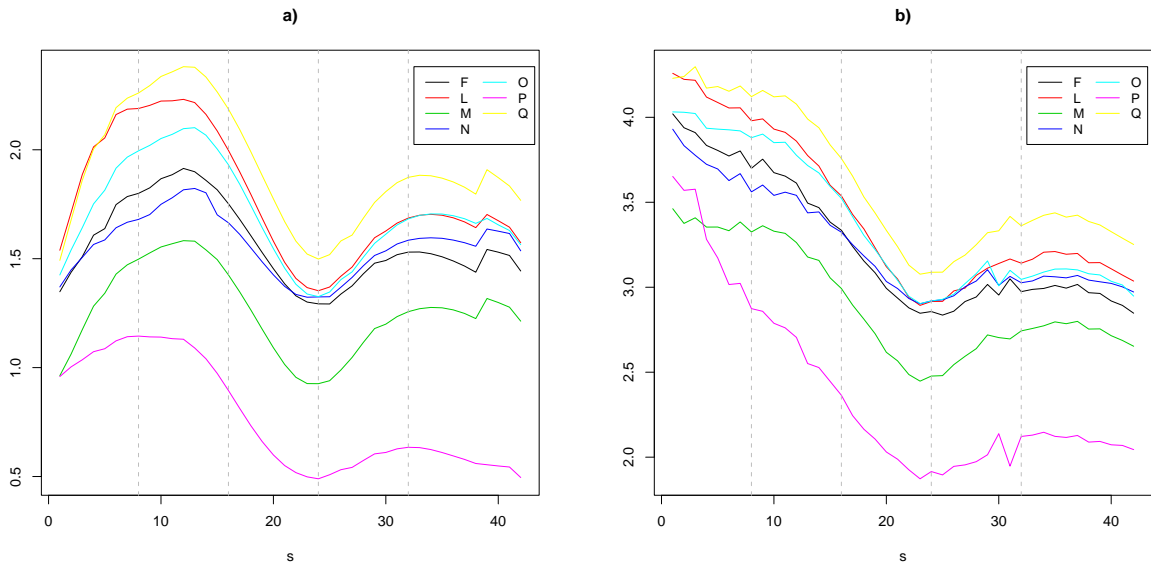


Fig. 5.4 Graphics of a) SN and b) ST mutual information, $I(\mathbf{X}, Y)$, when the $Y = F, L, M, N, O, P$ or Q station is removed from the network \mathbf{X} . The vertical dotted gray lines correspond to $s = \{8, 16, 24, 32\}$.

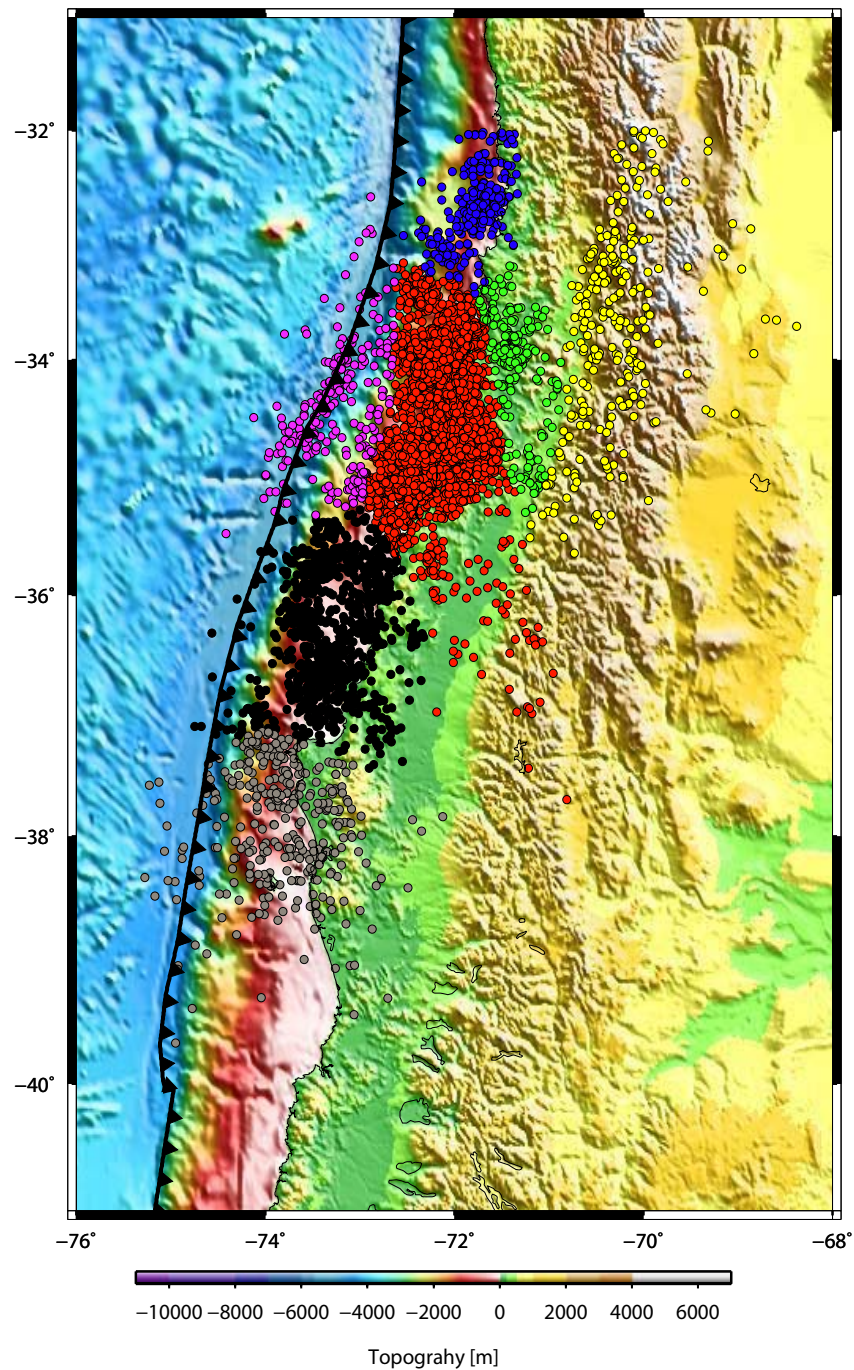


Fig. 5.5 Left: Map of the Chile region analyzed for post-seismicity correlation with clustering events: black(1), red(2), green(3), blue(4), violet(5), yellow(6) and gray(7).

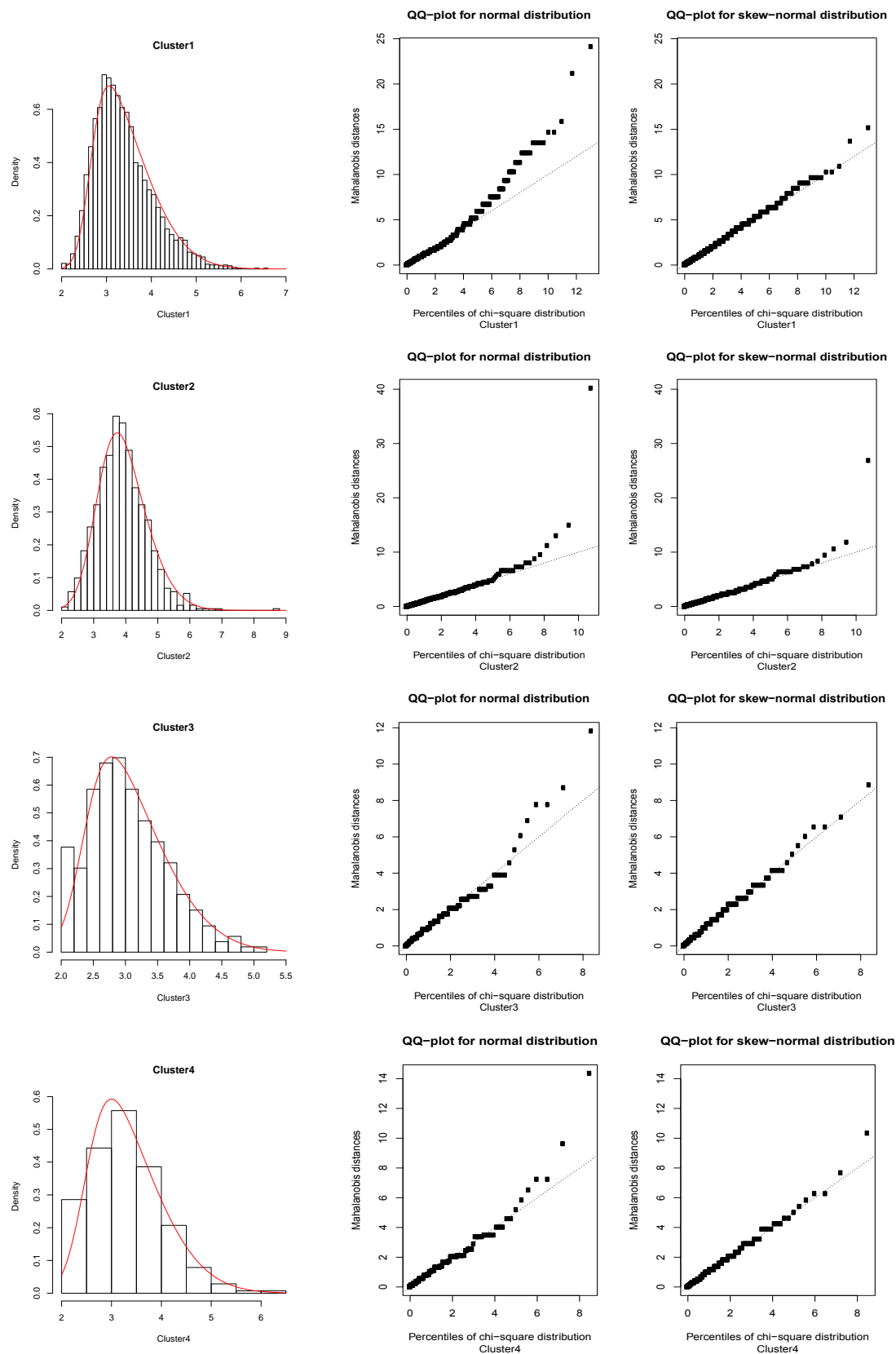


Fig. 5.6 Plots of SN fits (in red) and QQ-plots of Normal and SN distributions for clusters black (1), red (2), green (3) and blue (4).

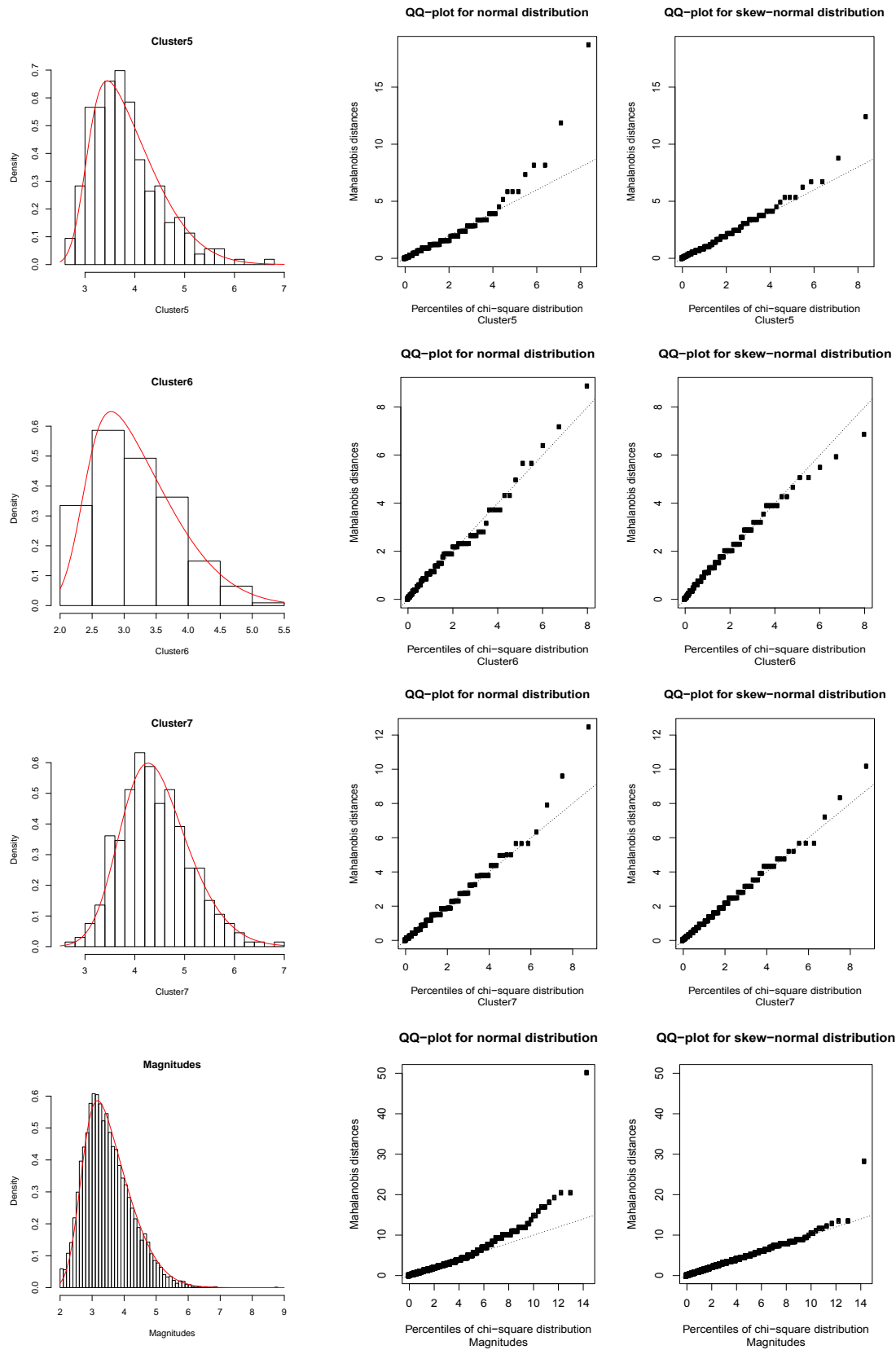


Fig. 5.7 Plots of SN fits (in red) and QQ-plots of Normal and SN distributions for clusters violet (5), yellow (6), gray (7) and all observations.

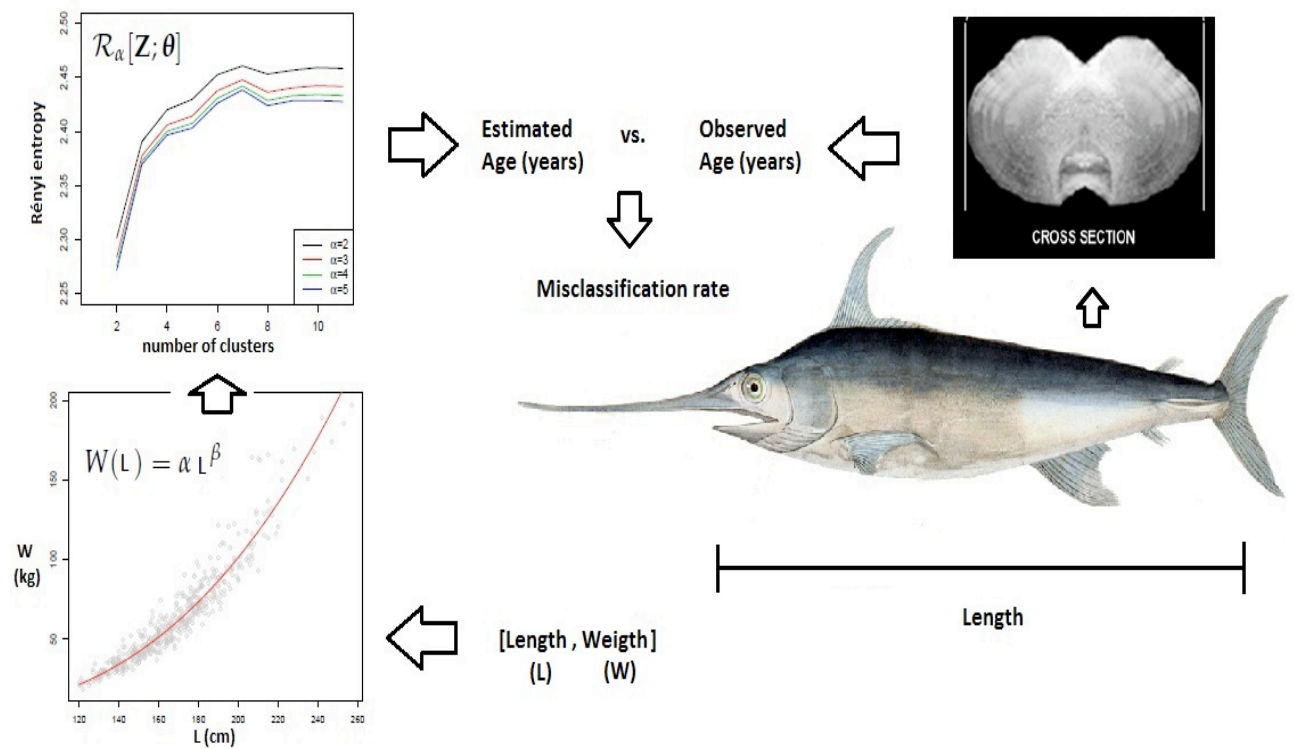


Fig. 5.8 Scheme of steps (a)–(d) for swordfish data.

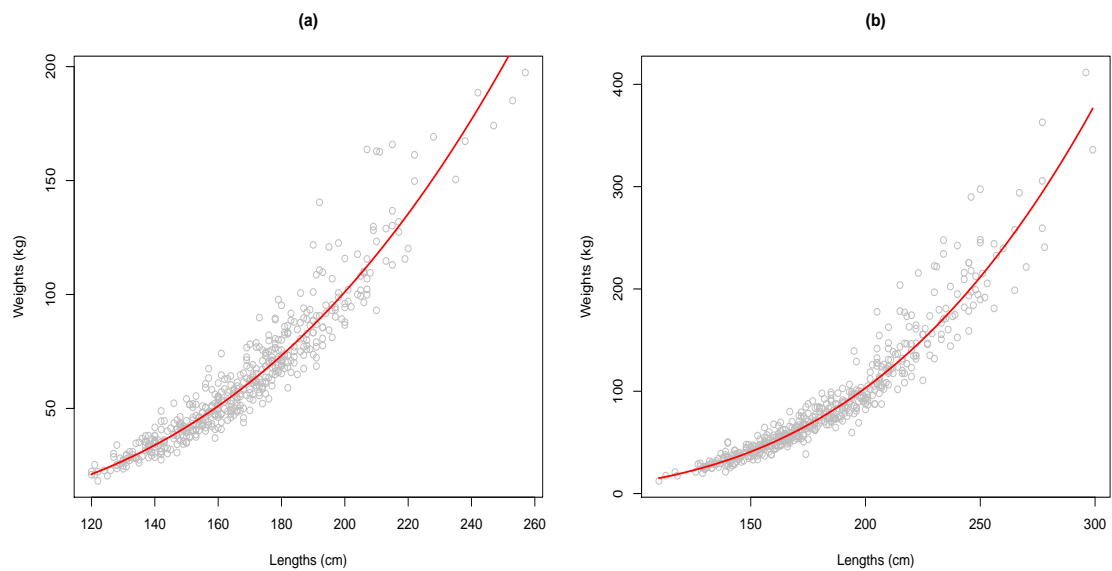


Fig. 5.9 Length-weight log-transformed relationship and regression fit (red solid line) for (a) male and (b) female swordfish.

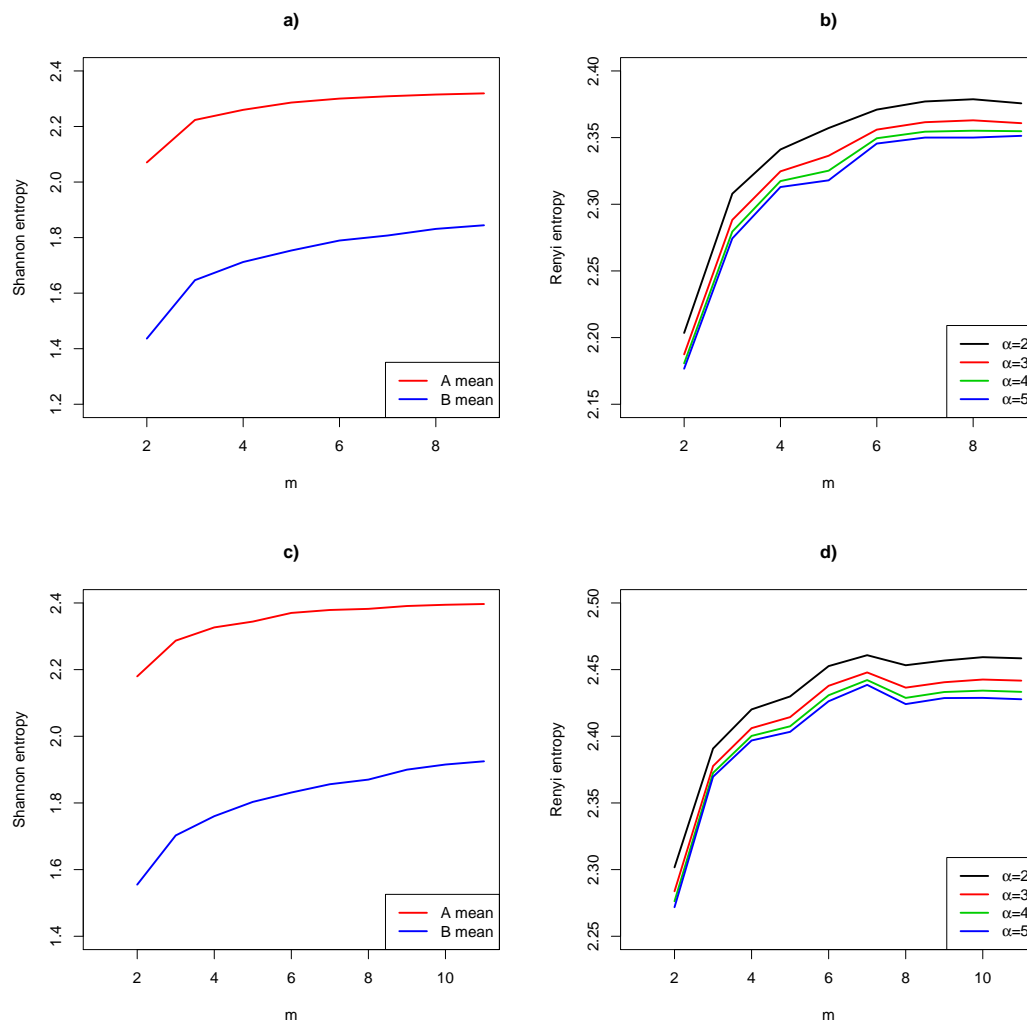


Fig. 5.10 Logarithmic of the average between upper and lower bounds for Shannon and Rényi entropies, for (a, c) male and (b, d) female swordfish, respectively.

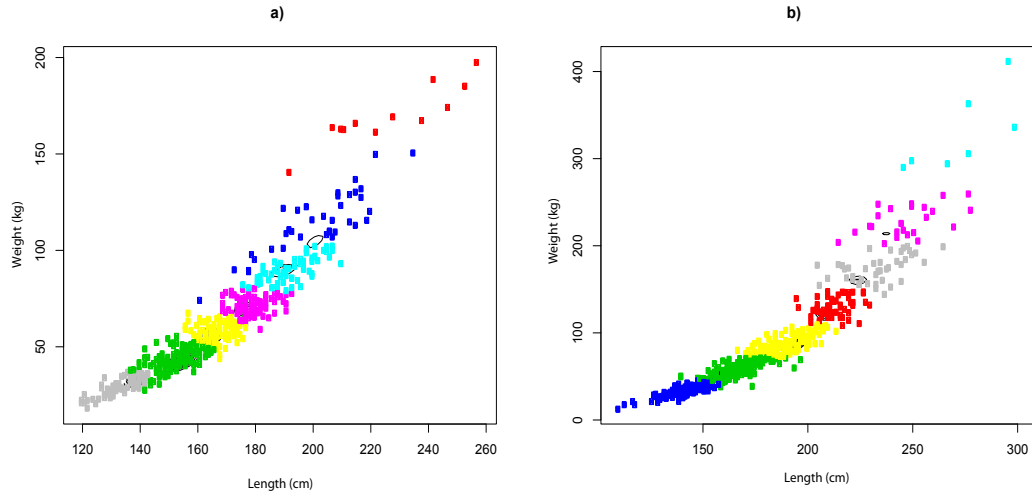


Fig. 5.11 Selected FMSN fits for (a) male and (b) female swordfish. Each color corresponds to each mixture component from a total of $m = 7$.

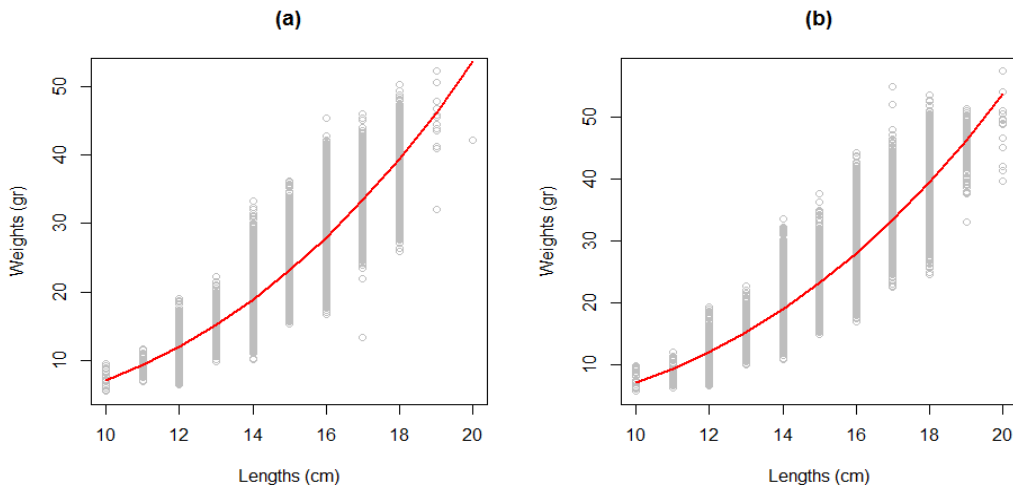


Fig. 5.12 Length-weight log-transformed relationship and linear regression fit (red solid line) for (a) male and (b) female anchovies.

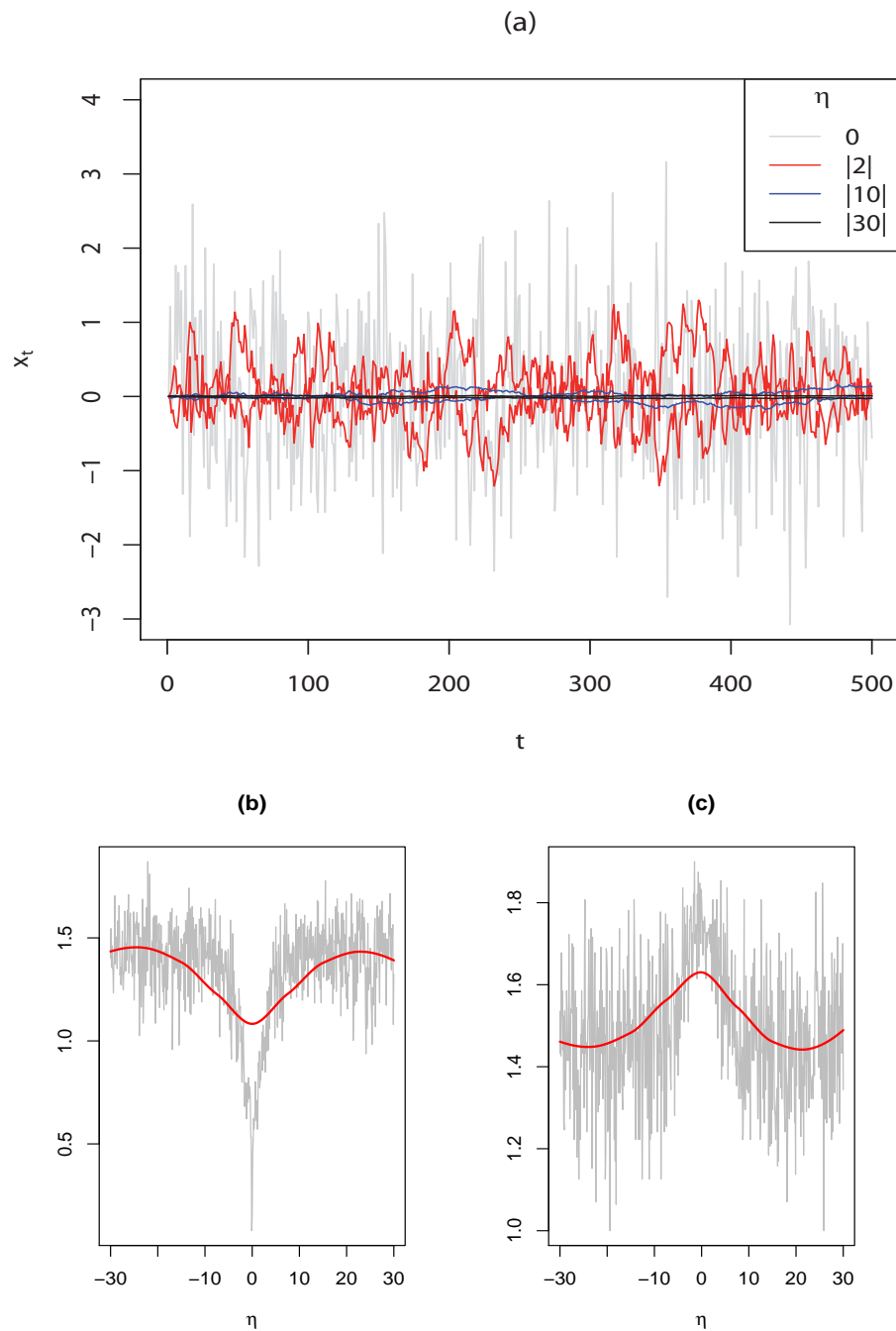


Fig. 5.13 (a) Stochastic representations of Z for $|\eta| = \{0, 3, 10, 30\}$, and its respective estimated (b) Hurst exponents and (c) fractal dimensions.

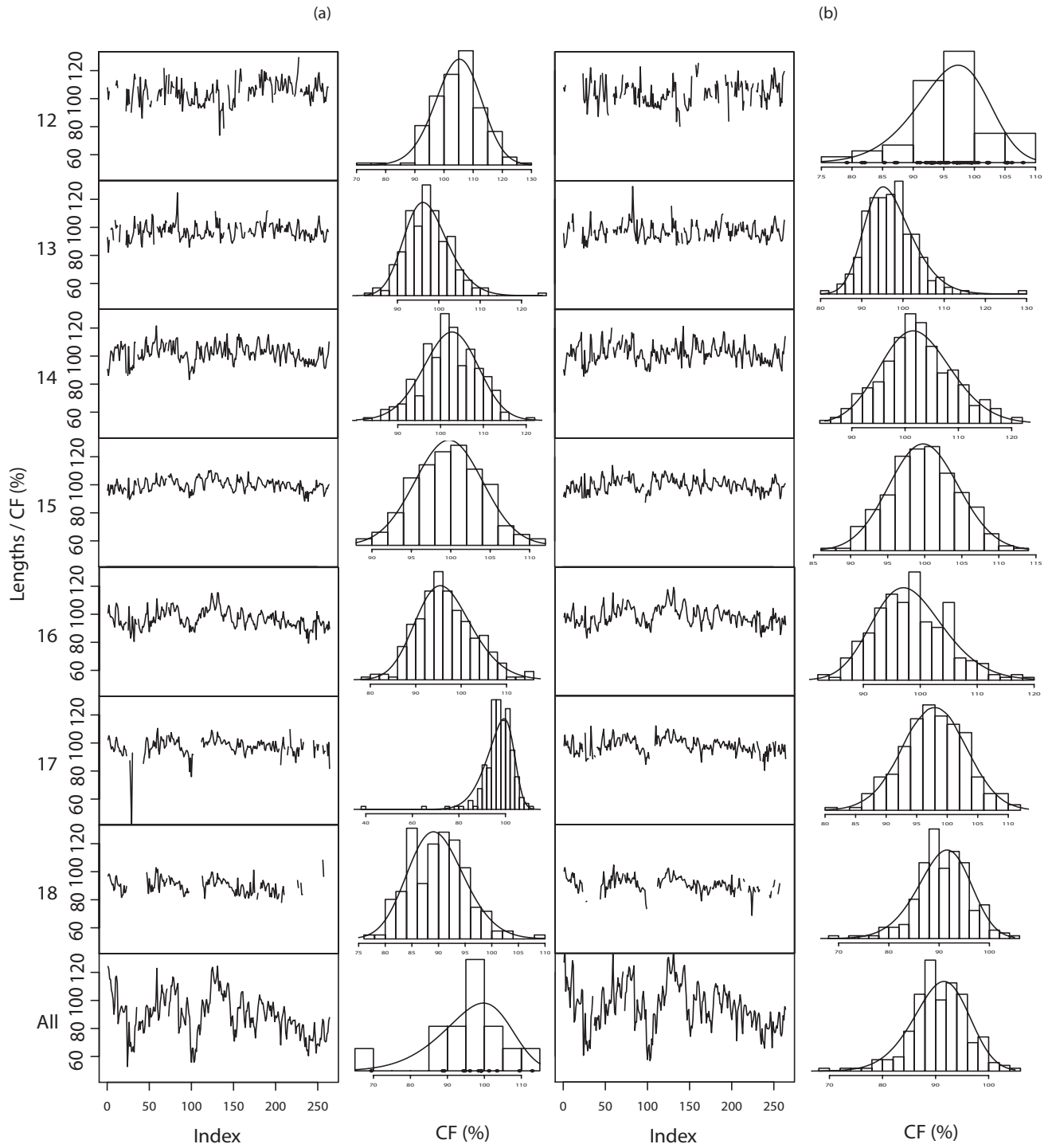


Fig. 5.14 Condition factor time series and histograms with SN fits (solid line) of anchovies by length classes $L = 12, \dots, 18, \text{All}$ for (a) male and (b) female anchovies.

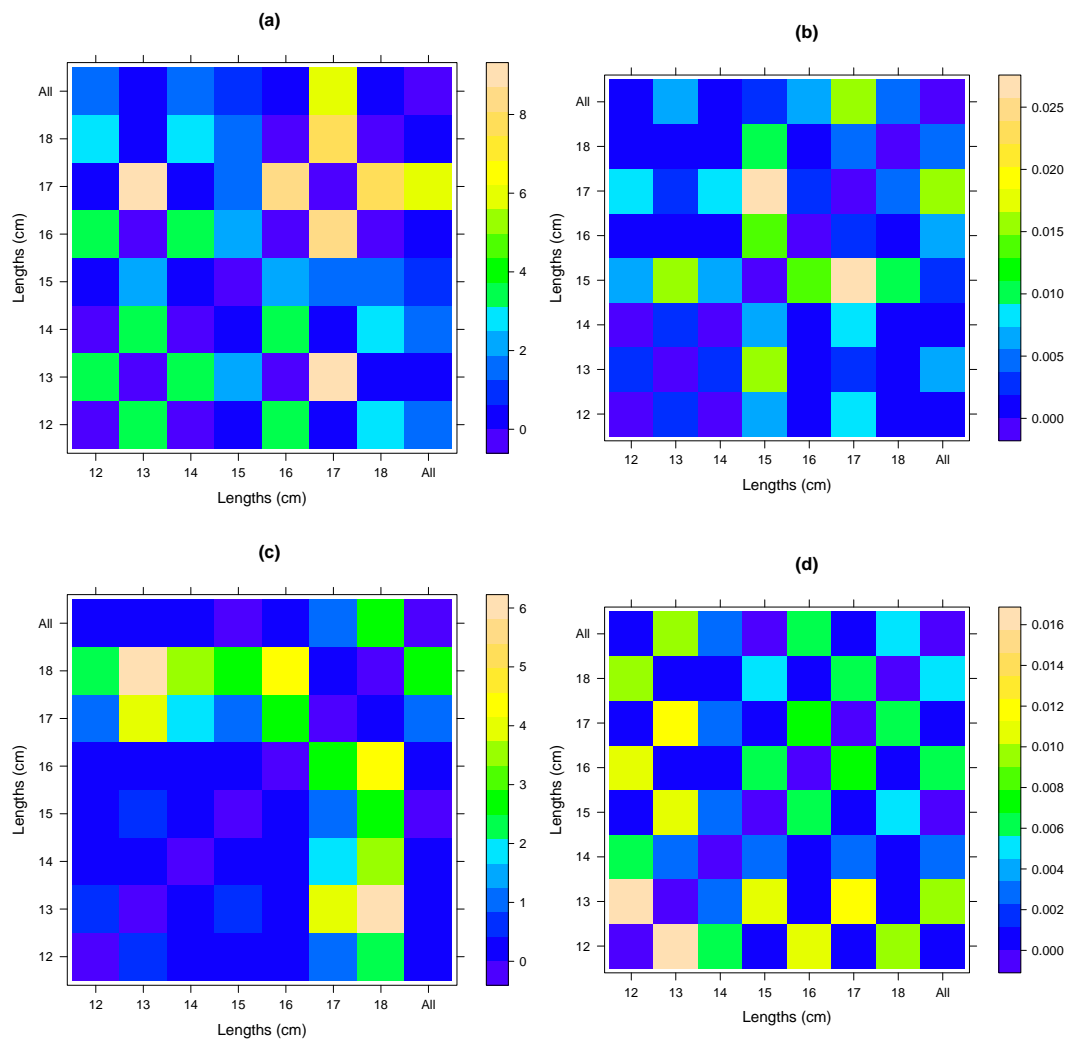


Fig. 5.15 SN J divergence for (a) males and (b) females, and JS distance for (c) males and (d) females. This measures consider the shape parameter estimates of Table 5.10 for each group j and length L .

Chapter 6

Concluding remarks

6.1 Skew-elliptical distributions

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 SN and ST 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 work, we have given the tools to compute this new information index. Others methods could be derived from this index, for example such as the *effectiveness index*, or to remove more than one station at a time (Silva and Quiroz, 2003).

The SE entropy and mutual information index can be explored further by considering the whole class of selection elliptical distributions introduced by Arellano-Valle et al. (2006). 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, the entropy of $\mathbf{Z} \sim SLCT_k$ is

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

The last term in the above selection entropy is justly the contribution of the selection mechanism. For selection SE distributions, e.g., we have $\mathbf{V} \sim EC_k(\mathbf{x}_{\mathbf{V}}, \Omega_{\mathbf{V}\mathbf{V}}, h^{(k)})$ and $\mathbf{U} \mid \mathbf{V} \sim EC_l(\xi_{\mathbf{U}|\mathbf{V}}, \Omega_{\mathbf{U}\mathbf{U}|\mathbf{V}}, h_{S_{\mathbf{V}}}^{(l)})$, where $\xi_{\mathbf{U}|\mathbf{V}} = \xi_{\mathbf{U}} + \Omega_{\mathbf{U}\mathbf{V}}\Omega^{-1}(\mathbf{V} - \xi_{\mathbf{V}})$, $\Omega_{\mathbf{U}\mathbf{U}|\mathbf{V}} = \Omega_{\mathbf{U}\mathbf{U}} - \Omega_{\mathbf{U}\mathbf{V}}\Omega_{\mathbf{V}\mathbf{V}}^{-1}\Omega_{\mathbf{V}\mathbf{U}}$ and $S_{\mathbf{V}} = (\mathbf{V} - \xi_{\mathbf{V}})^{\top}\Omega_{\mathbf{V}\mathbf{V}}^{-1}(\mathbf{V} - \xi_{\mathbf{V}})$. Hence,

$$H_{\mathbf{V}} = \frac{1}{2} \log |\Omega_{\mathbf{V}\mathbf{V}}| - E\{\log h^{(k)}(S_{\mathbf{V}})\},$$

while the computation of the contribution of the selection mechanism requires the specification of the selection set C . In our case, we have $l = 1$, $\xi_{\mathbf{U}} = 0$, $\Omega_{\mathbf{U}\mathbf{U}} = 1$, $\Omega_{\mathbf{V}\mathbf{U}} = \delta$ and $C = (0, \infty)$, so that the probability density function of the selection random vector \mathbf{Z} reduces to (2.3), with $\xi = \mathbf{x}_{\mathbf{V}}$, $\Omega = \Omega_{\mathbf{V}\mathbf{V}}$ and $\eta = \Omega^{-1}\delta/\sqrt{1 - \delta^{\top}\Omega^{-1}\delta}$, and therefore SLCT entropy becomes SE entropy as in Proposition 1.

In addition, we have presented a methodology to compute the KL 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 KL 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. The proposed methodology is applied to aftershocks produced by the Maule earthquake which occurred on February 27 of 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 founded results in a consistent form.

A measure to compare two multivariate ST densities is presented based on the approximated KL divergence. This has some advantages such as: the detection of skewness presence and heavy-tails within the data; it is of rapid and easy computational implementation, because it is an explicit form of the divergence where there is an expected value representing an integral in one dimension. On the another hand, many applications should be addressed as outlier detection, influential data analysis, hypothesis testing, variational inference, ensemble correlation matrices, discriminant analysis and clustering; and some applications on real data in the field of optimal network design, satellite image processing, region-of-interest tracking in video sequences and image registration.

6.2 Finite mixture of skew-normal distributions

Some solutions to compute the Rényi entropy with discrete α -order and for a wide range of asymmetric distributions are presented. Specifically, we find a closed expression for SN, ESN, and TSN distributions. Additional inequalities for SN and ESN entropies were reported. Lower and upper bounds of the Shannon and Rényi entropies for FMSN distributions were derived. Using such a pair of bounds some kind of confidence interval for the approximate entropy value can be calculated, where the average between these values can be used as an approximation of the entropy. We presented practical (bounds) and theoretical (bounds and asymptotic expression) results for Rényi entropy. In the case of practical results, the first upper bound deals only with the density parameters and the second one with the density and mixing weights parameters. In the case of theoretical results, the bounds and approximations are based on L^p space metric and multinomial coefficients.

Inserting the ML estimation (fixed) parameters represents the simplest evaluation of these bounds (Contreras-Reyes and Arellano-Valle, 2012). However, between the lower and upper Rényi entropy bounds exists a considerable distance. For this reason, further research must consider the exact expression and asymptotic approximation presented in this work. For these, an algorithm to identify the multinomial coefficients restricted to conditions (3.15) and (3.18), respectively, could be developed. In addition, the Bayesian approach allows for a direct estimation of the entropies, depending on the accuracy of prior parameters, where performance can be substantially improved compared to ML or nonparametric estimators (Gupta and Srivastava, 2010).

The results presented are valid for the SN case, taking the shape parameters set $\eta = (\mathbf{0}, \dots, \mathbf{0})$, for integer values of α (Contreras-Reyes, 2015). However, numerical algorithms can be applied for real values of α ($\alpha > 2$), but that requires more challenging computational work. In addition, Proposition 2 and Lemma 1, 3 and 9 are also valid for other continuous

densities where the Rényi entropies of the component exist. We hope the Rényi entropy developments in finite mixtures of densities can stimulate more research in the future, for more flexible densities such as ST distribution (Azzalini and Capitanio, 2003, Gupta, 2003).

We applied 2-dimensional length-weight data for the determination of swordfish age. We considered a length-weight dataset instead of the usual length (considered by Roa-Ureta, 2010) to determine the number of clusters, and posteriorly we compared it with the real observations obtained by the procedure of Cerna (2009). The best results were obtained using the Rényi entropy, as an average between upper and lower bounds, over Shannon entropy and information criteria. Additionally, the classification rates and consistency scores of FMSN models showed better results versus the FMN model.

Wrongly classified observations arise with older species because they produce higher variability in the length-weight relationship. Moreover, the age determination in these age classes is difficult to obtain for the reasons mentioned in Section 5.3. We encourage anal-fins readers to consider the proposed methodology to compare their results with this statistical methodology, especially for the revision of older species data.

6.3 Generalized skew-normal distributions

We have presented the methodology to compute the Shannon entropy, the negentropy, and the KL and J divergences for a broad family of asymmetric distributions with normal kernel called Generalized Skew-Normal distributions. Our method considers asymptotic expansions regarding moments and cumulants for two particular cases: the SN and MSN distributions. We then measured the degrees of disparity of these distributions from the normal distribution by using exact expressions for the negentropy in terms of moments and cumulants. Additionally, given the regularity conditions accomplished by the MSN distribution, normality was tested based on the modified skew-normal distribution. This test considered the asymptotic behavior of the KL divergence, which is determined by the negentropy for normality disparity.

Numerical results showed that the Shannon entropy and negentropy of the modified skew-normal distribution is better approximated than SN one, at least for a wider range of the shape parameter. For small to moderate values of the asymmetry parameter, where the approximations are appropriate, we find that expansions series converge from the fourth moment/cumulant to greater, as in Gram–Charlier and Edgeworth expansion methods (Hyvärinen et al., 2001). For large values of the skewness parameter, where the expansions are inappropriate, the functions related to negentropy are not well approximated by Taylor expansions around 0, produced by a divergence in the moment and cumulant terms, i.e.,

the Taylor expansions for the expected values of the functions $\zeta_0(\tau Z_\tau)$ and $\zeta_0\{\tau u(Z_\tau)\}$ (SN and MSN case, respectively) if $\tau = |\eta|$ is too large. When this happens, the normal cdf, $\Phi(\tau Z_\tau)$ and $\Phi\{\tau u(Z_\tau)\}$ (SN and MSN case, respectively), tends to 1, since according to the stochastic representation in (2.26), for large values of τ , the distribution of Z_τ converges to the standardized half-normal distribution (see property 1 of Section 2.4.1 or Beaver and Arnold, 2000).

However, the normality test considered in the application used skewness parameters inside the appropriate range. Furthermore, we plan to investigate the negentropy of the modified skew-normal-Cauchy distribution or similar models. In addition, although the approximations are appropriate over the range of variation of the asymmetry admitted by both models, more work should be done in order to improve the asymptotic approximations for a greater range of the skewness parameter values. Besides, this is not an easy task since generally it is difficult to approximate KL divergences involving asymmetric and heavy-tailed distributions (Stehlík et al., 2017).

The statistical application related to condition factor time series of anchovies off northern Chile is given. The results show that the proposed methodology serves to detect non-normal events in these time series, which produces an empirical distribution with high well presence of skewness (Contreras-Reyes, 2016). The proposed test for normality is therefore useful to detect anomalies in condition factor time series, linked to food deficit (positive shape parameter) or food abundance (negative shape parameter) influenced by environmental conditions.

References

- B. Ainslie, C. Reuten, D. G. Steyn, N. D. Le, and J. V. Zidek. Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants. *Journal of Environmental Management*, 90:2715–2729, 2009.
- H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.
- H. Amoud, H. Snoussi, D. Hewson, M. Doussot, and J. Duchêne. Intrinsic mode entropy for nonlinear discriminant analysis. *IEEE Signal Processing Letters*, 14:297–300, 2007.
- J. Andêl, I. Netuka, and K. Zvára. On threshold autoregressive processes. *Kybernetika*, 20: 89–106, 1984.
- D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, 36:99–102, 1974.
- C. Anteneodo and A. R. Plastino. Some features of the López-Ruiz-Mancini-Calbet (LMC) statistical measure of complexity. *Physics Letters A*, 223:348–354, 1996.
- R. B. Arellano-Valle. On the information matrix of the multivariate skew- t model. *Metron*, 68:371–386, 2010.
- R. B. Arellano-Valle and A. Azzalini. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33:561–574, 2006.
- R. B. Arellano-Valle and A. Azzalini. The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis*, 99:1362–1382. (Corrigendum: vol. 100 (2009), p. 816), 2008.
- R. B. Arellano-Valle and H. Bolfarine. On some characterizations of the t -distribution. *Statistics & Probability Letters*, 25:179–185, 1995.
- R. B. Arellano-Valle and M. G. Genton. On fundamental skew distributions. *Journal of Multivariate Analysis*, 96:93–116, 2005.
- R. B. Arellano-Valle and M. G. Genton. Multivariate extended skew- t distributions and related families. *Metron*, 68:201–234, 2010.
- R. B. Arellano-Valle, H. Bolfarine, and P. L. Iglesias. A predictivistic interpretation to the multivariate t distribution. *Test*, 3:221–236, 1994.

- R. B. Arellano-Valle, M. Galea-Rojas, and P. Iglesias. Bayesian sensitivity analysis in elliptical linear regression models. *Journal of Statistical Planning and Inference*, 86: 175–199, 2000.
- R. B. Arellano-Valle, H. W. Gómez, and F. A. Quintana. A new class of skew-normal distributions. *Communications in Statistics - Theory and Methods*, 33:1465–1480, 2004.
- R. B. Arellano-Valle, M. D. Branco, and M. G. Genton. A unified view on selection distributions. *Canadian Journal of Statistics*, 33:561–574, 2006.
- R. B. Arellano-Valle, J. E. Contreras-Reyes, and M. G. Genton. Shannon entropy and mutual information for multivariate skew-elliptical distributions. *Scandinavian Journal of Statistics*, 40:42–62, 2013.
- R. B. Arellano-Valle, J. E. Contreras-Reyes, and M. Stehlík. Generalized skew-normal negentropy and its application to fish condition factor time series. *Entropy*, 19:528, 2017.
- J. Arrué, H. W. Gómez, H. Varela, and H. Bolfarine. On the skew-normal-Cauchy distribution. *Communications in Statistics - Theory and Methods*, 40:15–27, 2010.
- J. Arrué, R. B. Arellano-Valle, and H. W. Gómez. Bias reduction of maximum likelihood estimates for a modified skew-normal distribution. *Journal of Statistical Computation and Simulation*, 86:2967–2984, 2016.
- A. Azzalini. A class of distributions which includes the Normal Ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.
- A. Azzalini. *R package sn: The skew-normal and skew-t distributions (version 0.4-6)*. Università di Padova, Padova, Italia, 2008.
- A. Azzalini and R. B. Arellano-Valle. Maximum penalized likelihood estimation for skew-normal and skew- t distributions. *Journal of Statistical Planning and Inference*, 143: 419–433, 2013.
- A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew normal distributions. *Journal of the Royal Statistical Society: Series B*, 61:579–602, 1999.
- A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t -distribution. *Journal of the Royal Statistical Society B*, 65:367–389, 2003.
- A. Azzalini and A. Capitanio. *The skew-normal and related families. Vol. 3*. John Wiley & Sons, Inc., Cambridge University Press, UK, 2013.
- A. Azzalini and A. Dalla-Valle. The multivariate skew-normal distribution. *Biometrika*, 83: 715–726, 1996.
- A. Azzalini and N. Torelli. Clustering via nonparametric density estimation. *Statistics and Computing*, 17:71–80, 2007.
- R. J. Beaver and B. C. Arnold. Hidden truncation models. *Sankhya A*, 62:23–35, 2000.

- G. Bennett. Lower bounds for matrices. *Linear Algebra and its Applications*, 82:81–98, 1986.
- M. D. Branco and D. K. Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79:99–113, 2001.
- J. Briët and P. Harremoës. Properties of classical and quantum Jensen–Shannon divergence. *Physical Review A*, 79:052311, 2009.
- P. Brosset, J. M. Fromentin, F. Ménard, F. Pernet, J. H. Bourdeix, J. L. Bigot, E. V. Beveren, M. A. P. Roda, S. Choy, and C. Saraux. Measurement and analysis of small pelagic fish condition: A suitable method for rapid evaluation in the field. *Journal of Experimental Marine Biology and Ecology*, 462:90–97, 2015a.
- P. Brosset, F. Ménard, J. M. Fromentin, S. Bonhommeau, C. Ulses, J. H. Bourdeix, J. L. Bigot, E. V. Beveren, D. Roos, and C. Saraux. Influence of environmental variability and age on the body condition of small pelagic fish in the Gulf of Lions. *Marine Ecology Progress Series*, 529:219–231, 2015b.
- H. Caillol, W. Pieczynski, and A. Hillion. Estimation of fuzzy gaussian mixture and unsupervised statistical image segmentation. *IEEE Transactions on Image Processing*, 6:425–440, 1997.
- A. Capitanio, A. Azzalini, and E. Stanghellini. Graphical models for skew-normal variates. *Scandinavian Journal of Statistics*, 30:129–144, 2003.
- L. C. Carpi, O. A. Rosso, P. M. Saco, and M. G. Ravetti. Analyzing complex networks evolution through Information Theory quantifiers. *Physics Letters A*, 375:801–804, 2011.
- M. A. Carreira-Perpinán. Mode-Finding for Mixtures of Gaussian Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1318–1323, 2000.
- G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13:195–212, 1996.
- J. F. Cerna. Age and growth of the swordfish (*xiphias gladius linnaeus*, 1758) in the southeastern Pacific off Chile. *Latin American Journal of Aquatic Research*, 37:59–69, 2009.
- E. Challis and D. Barber. Affine Independent Variational Inference. *Advances in Neural Information Processing Systems*, 25:2195–2203, 2012.
- H. Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25:573–578, 1954.
- M. Chiogna. Some results on the scalar skew-normal distribution. *Journal of the Italian Statistical Society*, 1:1–13, 1998.
- J. E. Contreras-Reyes. *R package skewtools: Tools for analyze skew-elliptical distributions and related models (version 0.1.1)*. Instituto de Fomento Pesquero, Valparaíso, Chile, 2012. URL <http://cran.rproject.org/web/packages/skewtools>.

- J. E. Contreras-Reyes. Nonparametric Assessment of Aftershock Clusters of the Maule Earthquake $m_w=8.8$. *Journal of Data Science*, 11:623–638, 2013.
- J. E. Contreras-Reyes. Asymptotic form of the Kullback–Leibler divergence for multivariate asymmetric heavy-tailed distributions. *Physica A*, 395:200–208, 2014.
- J. E. Contreras-Reyes. Rényi entropy and complexity measure for skew-gaussian distributions and related families. *Physica A*, 433:84–91, 2015.
- J. E. Contreras-Reyes. Analyzing fish condition factor index through skew-gaussian information theory quantifiers. *Fluctuation and Noise Letters*, 15:1650013, 2016.
- J. E. Contreras-Reyes and R. B. Arellano-Valle. Kullback–Leibler divergence measure for Multivariate Skew-Normal Distributions. *Entropy*, 14:1606–1626, 2012.
- J. E. Contreras-Reyes and R. B. Arellano-Valle. Growth estimates of cardinalfish (*epigonus crassicaudus*) based on scale mixtures of skew-normal distributions. *Fisheries Research*, 147:137–144, 2013.
- J. E. Contreras-Reyes, R. B. Arellano-Valle, and T. M. Canales. Comparing growth curves with asymmetric heavy-tailed errors: Application to southern blue whiting (*micromesistius australis*). *Fisheries Research*, 159:88–94, 2014.
- J. E. Contreras-Reyes, T. M. Canales, and P. M. Rojas. Influence of climate variability on anchovy reproductive timing off northern Chile. *Journal of Marine Systems*, 164:67–75, 2016.
- T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley & Son, Inc., New York, NY, USA, 2006.
- G. E. Crooks. Inequalities between the Jensen-Shannon and Jeffreys divergences. *Tech. Note*, 4, 2008.
- L. Cubillos and G. Claramunt. Length-structured analysis of the reproductive season of anchovy and common sardine off central southern Chile. *Marine Biology*, 156:1673–1680, 2009.
- M. M. De Queiroz, R. W. Silva, and R. H. Loschi. Shannon entropy and Kullback–Leibler divergence in multivariate log fundamental skew-normal and related distributions. *Canadian Journal of Statistics*, 44:219–237, 2016.
- A. Dembo, T. M. Cover, and J. A. Thomas. Information Theoretic Inequalities. *IEEE Transactions on Information Theory*, 37:1501–1518, 1991.
- H. Dette, C. Ley, and F. J. Rubio. Natural (non-) informative priors for skew-symmetric distributions. *Scandinavian Journal of Statistics*, in press, 2017.
- J. A. Domínguez-Molina and A. Rocha-Arteaga. On the infinite divisibility of some skewed symmetric distributions. *Statistics & Probability Letters*, 77:644–648, 2007.
- J. L. Durrieu, J. Thiran, and F. Kelly. Lower and upper bounds for approximation of the Kullback–Leibler divergence between Gaussian mixture models. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4833–4836, 2012.

- M. Eling. Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics*, 51:239–248, 2012.
- T. Eltoft, A. Doulgeris, and S. N. Anfinsen. Analysis of textured PolSAR data by Shannon entropy. *IEEE Transactions on Geoscience and Remote Sensing. (IGARSS 2012)*, pages 1449–1452, 2012.
- K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*, volume 36. Chapman and Hall, Ltd., London, UK, 1990.
- H. Finner. A generalization of Hölder’s inequality and some probability inequalities. *Annals of Probability*, 20:1893–1901, 1992.
- C. Flecher, P. Naveau, and D. Allard. Estimating the Closed Skew-Normal distributions parameters using weighted moments. *Statistics & Probability Letters*, 79:1977–1984, 2009.
- C. Flecher, D. Allard, and P. Naveau. Truncated skew-normal distributions: moments, estimation by weighted moments and application to climatic data. *Metron*, 68:265–279, 2010.
- A. C. Frery, A. Nascimento, and R. Cintra. Information theory and image understanding: An application to polarimetric SAR imagery. *Chilean Journal of Statistics*, 2:81–100, 2011.
- S. Frühwirth-Schnatter and S. Pyne. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew- t distributions. *Biostatistics*, 11:317–336, 2010.
- J. H. Gao and B. Zhang. Estimation of seismic wavelets based on the multivariate scale mixture of Gaussians model. *Entropy*, 12:14–33, 2009.
- M. G. Genton. *Skew-elliptical distributions and their applications: A journey beyond normality*. Chapman & Hall/CRC, Boca Raton, FL, USA, 2004.
- M. G. Genton, L. He, and X. Liu. Moments of skew-normal random vectors and their quadratic forms. *Statistics & Probability Letters*, 51:319–325, 2001.
- T. Ghizzoni, G. Roth, and R. Rudari. Multivariate skew- t approach to the design of accumulation risk scenarios for the flooding hazard. *Advances in Water Resources*, 33:1243–1255, 2010.
- T. Gneiting, H. Sevcikova, and D. B. Percival. Estimators of fractal dimension: Assessing the smoothness of time series and spatial data. *Statistical Science*, 27:247–277, 2012.
- L. G. Godoi, M. D. Branco, and F. Ruggeri. Concentration function for the skew-normal and skew- t distributions, with application in robust Bayesian analysis. *Brazilian Journal of Probability and Statistics*, 31:373–393, 2017.
- L. Golshani and E. Pasha. Rényi entropy rate for Gaussian processes. *Information Sciences*, 180:1486–1491, 2010.
- M. A. Gómez-Villegas, P. Main, H. Navarro, and R. Susi. Assessing the effect of kurtosis deviations from Gaussianity on conditional distributions. *Applied Mathematics and Computation*, 219:10499–10505, 2013.

- G. González-Farías, J. Domínguez-Molina, and A. Gupta. Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference*, 126:521–534, 2004.
- I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Elsevier, Amsterdam, 7th edition, 2007.
- R. M. Gray. *Entropy and information theory*. Springer-Verlag, New York, USA, 1990.
- A. K. Gupta. Multivariate skew- t distribution. *Statistics*, 37:359–363, 2003.
- A. K. Gupta, F. C. Chang, and W. J. Huang. Some skew-symmetric models. *Random Operators and Stochastic Equations*, 10:133–140, 2002.
- M. Gupta and S. Srivastava. Parametric Bayesian estimation of differential entropy and relative entropy. *Entropy*, 12:818–843, 2010.
- R. C. Gupta and N. Brown. Reliability studies of the skew-normal distribution and its application to a strength-stress model. *Communications in Statistics - Theory and Methods*, 30:2427–2445, 2001.
- N. Henze. A probabilistic representation of the ‘skew-normal’ distribution. *Scandinavian Journal of Statistics*, 13:271–275, 1986.
- M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck. On entropy approximation for Gaussian mixture random vectors. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188, 2008.
- A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley & Sons, Inc., New York, NY, USA, 2001.
- IFOP. Instituto de Fomento Pesquero, Valparaíso, Chile. 2010. URL <http://www.ifop.cl/>.
- A. Jamalizadeh, R. Pourmousa, and N. Balakrishnan. Truncated and limited skew-normal and skew- t distributions: properties and an illustration. *Communications in Statistics - Theory and Methods*, 38:2653–2668, 2009.
- E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70:939–952, 1982.
- H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A*, 186:453–461, 1946.
- R. Jenssen, K. E. Hild, D. Erdogmus, J. C. Principe, and T. Eltoft. Clustering using Renyi’s entropy. *IEEE Proceeding International Joint Conference on Neural Networks (IJCNN)*, 1:523–528, 2003.
- K. K. Jose and S. R. Naik. A class of asymmetric pathway distributions and an entropy interpretation. *Physica A*, 387:6943–6951, 2008.
- A. Kawabata, H. Yamaguchi, S. Kubota, and M. Nakagami. Growth and fatness of 1975–2002 year classes of Japanese sardine in the Pacific waters around northern Japan. *Fisheries Science*, 77:291–299, 2011.

- H. M. Kim and B. K. Mallick. Moments of random vectors with skew t distribution and their quadratic forms. *Statistics & Probability Letters*, 63:417–423. Corrigendum (2009) 79, 2098–2099, 2003.
- S. Kullback. *Information theory and statistics*. Dover Edition, Gloucester, UK, 1978.
- S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.
- D. Kundu, N. Balakrishnan, and A. Jamalizadeh. Generalized multivariate Birnbaum–Saunders distributions and related inferential issues. *Journal of Multivariate Analysis*, 116: 230–244, 2013.
- M. Kupperman. *Further applications of information theory to multivariate analysis and statistical inference*. The George Washington University, Washington, DC, USA, p. 270, 1957.
- E. D. Le Cren. The length–weight relationship and seasonal cycle in gonad weight and condition in the perch (*perca fluviatilis*). *Journal of Animal Ecology*, 20:201–219, 1951.
- S. Lee and G. J. McLachlan. EMMIXuskew: An R package for fitting mixtures of multivariate skew t -distributions via the EM algorithm. *Journal of Statistical Software*, 55:1–22, 2013.
- S. Lee, M. G. Genton, and R. B. Arellano-Valle. Perturbation of numerical confidential data via skew- t distributions. *Management Science*, 56:318–333, 2010.
- J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151, 1991.
- T. I. Lin. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 100:257–265, 2009.
- C. S. Liu. Nonsymmetric entropy and maximum nonsymmetric entropy principle. *Chaos, Solitons & Fractals*, 40:2469–2474, 2009.
- R. López-Ruiz, H. L. Mancini, and X. Calbet. A statistical measure of complexity. *Physics Letters A*, 209:321–326, 1995.
- Y. Ma and M. G. Genton. A flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics*, 31:459–468, 2004.
- K. Madani, V. Kachurka, C. Sabourin, V. Amarger, V. Golovko, and L. Rossi. A human-like visual-attention-based artificial vision system for wildland firefighting assistance. *Applied Intelligence*, in press, 2017.
- P. Main, J. M. Arevalillo, and H. Navarro. Local Effect of Asymmetry Deviations from Gaussianity Using Information-Based Measures. In *Proceedings of the 2nd International Electronic Conference on Entropy and Its Applications*, Santa Barbara, CA, USA, November 2015:15–30, 2016.
- K. V. Mardia. *Mardia’s test of multinormality*, volume 5. Wiley, Inc., New York, USA, 1985.

- E. H. Martínez, H. Varela, H. W. Gómez, and H. Bolfarine. A note on the likelihood and moments of the skew-normal distribution. *SORT-Statistics and Operations Research Transactions*, 32:57–66, 2008.
- G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2000.
- S. Nadarajah and S. Kotz. Skewed distributions generated by the normal kernel. *Statistics & Probability Letters*, 65:269–277, 2003.
- F. Nielsen and K. Sun. Guaranteed bounds on the kullback–leibler divergence of univariate mixtures. *IEEE Signal Processing Letters*, 23:1543–1546, 2016.
- NOAA. Climate Prediction Center. Cold & Warm Episodes by Season. National Oceanographic and Atmospheric Administration (NOAA), USA. 2015. URL http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml.
- R. Nock and F. Nielsen. A closed-form expression for the Sharma–Mittal entropy of exponential families. *Journal of Physics A: Mathematical and Theoretical*, 45:032003, 2012.
- J. T. Ormerod. *Skew-Normal Variational Approximations for Bayesian Inference*. Technical Report CRGTR-93-1, School of Mathematics and Statistics, University of Sydney, Sydney, Australia, 2011. URL <http://www.maths.usyd.edu.au/u/jormerod/JTOpapers/snva.pdf>.
- D. B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27:1075–1090, 1956.
- C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic organization of DNA nucleotides. *Physical Review E*, 49:1685–1689, 1994.
- R. Piessens, E. deDoncker Kapenga, C. Uberhuber, and D. Kahaner. *Quadpack: a Subroutine Package for Automatic Integration*. Springer Verlag, Berlin Heidelberg, 1983.
- C. A. Pires and A. Hannachi. Independent Subspace Analysis of the Sea Surface Temperature Variability: Non-Gaussian Sources and Sensitivity to Sampling and Dimensionality. *Complexity*, 2017:3076810, 2017.
- C. A. Pires and A. F. Ribeiro. Separation of the atmospheric variability into non-Gaussian multidimensional sources by projection pursuit techniques. *Climate Dynamics*, 48:821–850, 2017.
- M. Pourahmadi. Skew-Normal ARMA Models with Nonlinear Heteroscedastic Predictors. *Communications in Statistics - Theory and Methods*, 36:1803–1819, 2007.
- M. O. Prates, V. H. Lachos, and C. Cabral. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, 54:1–20, 2013.
- S. Qian and T. Cuffney. To threshold or not to threshold? That’s the question. *Ecological Indicators*, 15:1–9, 2012.
- P. Quelle, F. González, M. Ruiz, X. Valeiras, O. Gutierrez, E. Rodriguez-Marin, and J. Mejuto. An approach to age and growth of south Atlantic swordfish (*xiphias gladius*) Stock. *Collective Volume of Scientific Papers ICCAT*, 70:1927–1944, 2014.

- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL <http://www.R-project.org>.
- A. Rényi. *Probability theory*. North-Holland, Amsterdam, 1970.
- J. Rezaie, J. Eidsvik, and T. Mukerji. Value of information analysis and Bayesian inversion for closed skew-normal distributions: Applications to seismic amplitude variation with offset data. *Geophysics*, 79:R151–R163, 2014.
- R. H. Roa-Ureta. A likelihood-based model of fish growth with multiple length frequency data. *Journal of Agricultural, Biological, and Environmental Statistics*, 15:416–429, 2010.
- M. Salicrú, M. L. Menéndez, L. Pardo, and D. Morales. On the applications of divergence type measures in testing statistical hypothesis. *Journal of Multivariate Analysis*, 51:372–391, 1994.
- P. Sánchez-Moreno, S. Zozor, and J. S. Dehesa. Upper bounds on Shannon and Rényi entropies for central potentials. *Journal of Mathematical Physics*, 52:022105, 2011.
- P. Sánchez-Moreno, J. C. Angulo, and J. S. Dehesa. A generalized complexity measure based on Rényi entropy. *The European Physical Journal D*, 68:212, 2014.
- SEREMI. Indices de Calidad del Aire, Santiago de Chile. Seremi de Salud, Gobierno de Chile. 2006. URL <http://www.seremisaludrm.cl/sitio/pag/aire/indexjs3aireindgasesdemo-prueba.asp>.
- C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.
- C. Silva and A. Quiroz. Optimization of the atmospheric pollution monitoring network at Santiago de Chile. *Atmospheric Environment*, 37:2337–2345, 2003.
- K. S. Song. Rényi information, loglikelihood and an intrinsic distribution measure. *Journal of Statistical Planning and Inference*, 93:51–69, 2001.
- SSN. Servicio Sismológico Nacional of Chile, Universidad de Chile, Santiago de Chile. 2011. URL <http://ssn.dgf.uchile.cl/>.
- M. Stehlík. Distributions of exact tests in the exponential family. *Metrika*, 57:145–164, 2003.
- M. Stehlík. Decompositions of information divergences: Recent development, open problems and applications. In *Proceedings of the 9th International Conference on Mathematical Problems in Engineering, Aerospace and Sciences: ICNPAA 2012, Vienna, Austria, 10–14 July 2012*, 1493:972–976, 2012.
- M. Stehlík, L. Střelec, and M. Thulin. On robust testing for normality in chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 130:98–108, 2014.
- M. Stehlík, J. Somorčík, L. Střelec, and J. Antoch. Approximation of information divergences for statistical learning with applications. *Mathematica Slovaca*, in press, 2017.

- C. L. Sun, S. P. Wang, and S. Z. Yeh. Age and growth of the swordfish (*xiphias gladius l.*) in the waters around Taiwan determined from anal-fin rays. *Fishery Bulletin*, 100:822–835, 2002.
- H. Tong. *Non-linear time series modelling in population biology: A preliminary case study*. Springer Berlin Heidelberg, Germany, 1990.
- H. Tong and K. S. Lim. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B*, 42:245–292, 1980.
- M. Tumminello, F. Lillo, and R. N. Mantegna. Kullback–Leibler distance as a measure of the information filtered from multivariate data. *Physical Review E*, 76:031123, 2007.
- A. Ullah. Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference*, 40:137–162, 1996.
- J. van der Meer, J. J. Beukema, and R. Dekker. Population dynamics of two marine polychaetes: the relative role of density dependence, predation, and winter conditions. *ICES Journal of Marine Science*, 57:1488–1494, 2000.
- I. Vidal, P. Iglesias, M. D. Branco, and R. B. Arellano-Valle. Bayesian Sensitivity Analysis and Model Comparison for Skew Elliptical Models. *Journal of Statistical Planning and Inference*, 136:3435–3457, 2006.
- Z. Wang, B. Zhang, and J. Gao. The residual phase estimation of a seismic wavelet using a Rényi divergence-based criterion. *Journal of Applied Geophysics*, 106:96–105, 2014.
- C. S. Withers and S. Nadarajah. Negentropy as a function of cumulants. *Information Sciences*, 271:31–44, 2014.
- T. Yamano. A statistical measure of complexity with nonextensive entropy. *Physica A*, 340:348–354, 2004.
- E. Yáñez, S. Hormazábal, C. Silva, A. Montecinos, M. A. Barbieri, A. Valdenegro, A. Órdenes, and F. Gómez. Coupling between the environment and the pelagic resources exploited of northern Chile: ecosystem indicators and a conceptual model. *Latin American Journal of Aquatic Research*, 36:159–181, 2008.
- A. Youssef, C. Delpha, and D. Diallo. Enhancement of incipient fault detection and estimation using the multivariate Kullback-Leibler Divergence. *IEEE Signal Processing Conference (EUSIPCO)*, 24th European:1408–1412, 2016.
- K. Zografos and S. Nadarajah. Expressions for Rényi and Shannon entropies for multivariate distributions. *Statistics & Probability Letters*, 71:71–84, 2005.

Appendix A

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

Proof of Proposition 2. Let $\mathbf{Z}_c = \Gamma\mathbf{Z}_0$ and $W = \gamma^T \mathbf{Z}_c$, where $\Gamma \in \mathbb{R}^{k \times k}$ is an orthogonal matrix such that $\Gamma\bar{\eta} = \|\bar{\eta}\|\gamma$ and the vector $\gamma \in \mathbb{R}^k$ is such that $\gamma^T \bar{\eta} = \|\bar{\eta}\|$ and $\|\gamma\| = 1$. Note that $W = \gamma^T \mathbf{Z}_c = \gamma^T \mathbf{Z}_0$ and $S_c = \|\mathbf{Z}_c\|^2 = \|\mathbf{Z}_0\|^2 = S$ since $\Gamma\Gamma^T = I_k$. Thus, considering also that $\mathbf{Z}_0 = \Gamma^T \mathbf{Z}_c$ and the absolute value of the determinant of Γ^T equals 1, the Jacobian method yields $f_{\mathbf{Z}_c}(\mathbf{z}) = 2h^{(k)}(s)F(\|\bar{\eta}\|w, h_s^{(1)})$, where $s = \|\mathbf{z}^T \mathbf{z}\|^2$ and $w = \gamma^T \mathbf{z}$, i.e., $\mathbf{Z}_c \sim SE_k(\mathbf{0}, I_k, \|\bar{\eta}\|\gamma, h^{(k+1)})$, and by Proposition 4.1 in Arellano-Valle and Genton (2010), we have $W = \gamma^T \mathbf{Z}_c \sim SE_1(0, 1, \|\bar{\eta}\|, h^{(k+1)})$. On the other hand, since $\mathbf{Z}_c \stackrel{d}{=} \bar{\delta}|X_0| + (I_k - \bar{\delta}\bar{\delta}^T)^{1/2}\mathbf{X}$ (see e.g. Arellano-Valle and Azzalini, 2006, Arellano-Valle et al., 2006), where $\bar{\delta} = \bar{\eta}/\sqrt{1 + \|\bar{\eta}\|^2}$ and

$$\begin{pmatrix} \mathbf{X} \\ X_0 \end{pmatrix} \sim EC_{k+1} \left(\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} I_k & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix}, h^{(k+1)} \right),$$

we have $W \stackrel{d}{=} \|\bar{\delta}\||X_0| + \sqrt{1 - \|\bar{\delta}\|^2}X_1$, where $X_1 = \gamma^T \mathbf{X} \sim EC_1(0, 1, h^{(1)})$ is distributed as the first component of \mathbf{X} . Thus, since we can assume without loss of generality that γ is the first column of Γ , we find $\mathbf{Z}_c \stackrel{d}{=} \gamma W + (I_k - \gamma\gamma^T)^{1/2}\mathbf{X} = (W, X_2, \dots, X_k)^T$, where X_2, \dots, X_k are the last $k-1$ components of \mathbf{X} . Consider now the transformation $U_1 = W$, $U_j = X_j$, $2 \leq j \leq k-1$, and $R = (W^2 + \sum_{j=2}^k X_j^2)^{1/2}$. This transformation has two inverses given by $w = u_1$, $x_j = u_j$, $2 \leq j \leq k-1$, and $x_k = \pm(r^2 - \sum_{j=1}^{k-1} u_j^2)^{1/2}$. The corresponding Jacobians are $J_1 = r/(r^2 - \sum_{j=1}^{k-1} u_j^2)^{1/2}$ and $J_2 = -J_1$. Thus, we obtain for $k \geq 2$ that

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

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

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

we have

$$\begin{aligned} f_{U_1, R}(u_1, r) &= 4r^{k-2} \left(1 - \frac{u_1^2}{r^2}\right)^{\left(\frac{k-1}{2}-1\right)} h^{(k)}(r^2) F(\|\tilde{\eta}\| u_1; h_{r^2}^{(1)}) \\ &\quad \times \int_{\{w_2, \dots, w_{k-1} : 0 < \sum_{j=2}^{k-1} w_j^2 < 1\}} \left(1 - \sum_{j=2}^{k-1} w_j^2\right)^{-1/2} dw_2 \cdots dw_{k-1} \\ &= \frac{4\pi^{k/2-1}}{\Gamma(\frac{k}{2})} r^{k-2} \left(1 - \frac{u_1^2}{r^2}\right)^{\left(\frac{k-1}{2}-1\right)} h^{(k)}(r^2) F(\|\tilde{\eta}\| u_1; h_{r^2}^{(1)}). \end{aligned}$$

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

Proof of Proposition 3. Let $\mathbf{Z}_0 = \Omega^{-1/2}(\mathbf{Z} - \xi)$. By Lemmas 1 and 3 and by (2.5), we have

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

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

Proof of Proposition 5. To compute the integral $\int [f(\mathbf{z})]^\alpha d\mathbf{z}$, we use the change of variables $\Omega_\alpha = \alpha^{-1}\Omega$ and $\mathbf{Z}_0 = \Omega_\alpha^{-1/2}(\mathbf{Z} - \mu)$, $\mathbf{Z}_0 \sim SN_d(\mathbf{0}, I_d, \tilde{\eta})$, $\tilde{\eta} = \Omega_\alpha^{1/2}\eta$. We shall use the fact that $|\Omega_\alpha| = \alpha^{-d}|\Omega|$ for k -dimensional matrices (Nock and Nielsen, 2012). Then, according to Lemma 3, the integral $\int [f(\mathbf{z})]^\alpha d\mathbf{z}$ should be rewritten in terms of an expected value with respect to a standardized normal density as

$$\begin{aligned} \int_{\mathbb{R}^k} [f(\mathbf{z})]^\alpha d\mathbf{z} &= \frac{2^\alpha}{|\Omega|^{d/2}} |\Omega_\alpha|^{1/2} (2\pi)^{(1-\alpha)d/2} E\{[\Phi_1(\tilde{\eta}^\top \mathbf{Z}_0)]^\alpha\} \\ &= \frac{2^\alpha}{\alpha^{d/2}} (2\pi)^{(1-\alpha)d/2} |\Omega|^{(1-\alpha)/2} E\{[\Phi_1(W)]^\alpha\}. \end{aligned}$$

where $W \sim SN_1(\mathbf{0}, \|\tilde{\eta}\|^2, \|\tilde{\eta}\|)$ with $\|\tilde{\eta}\| = \tilde{\eta}^\top \tilde{\eta}$ (Arellano-Valle et al., 2013, Contreras-Reyes and Arellano-Valle, 2012), i.e., the expected value $E\{[\Phi_1(\tilde{\eta}^\top \mathbf{Z}_0)]^\alpha\}$ is reduced from

d dimensions to one dimension (Arellano-Valle et al., 2013, Contreras-Reyes, 2014). By Lemma 2 and setting $\mu = \mathbf{0}$, $\Omega = \|\tilde{\eta}\|^2$, $\mathbf{D} = \|\tilde{\eta}\|$, $r = \alpha$, $\mathbf{A} = s = h(w) = 1$; we obtain $\tilde{\mathbf{A}} = I_{\alpha+1}$ and $\tilde{\mathbf{D}} = (\mathbf{1}_\alpha, \|\tilde{\eta}\|)^\top$. Therefore, the expected value of the integral is reduced to

$$E\{\Phi_1(W)^\alpha\} = \frac{\Phi_{\alpha+1}(\mathbf{0}; \mathbf{0}, I_{\alpha+1} + \|\tilde{\eta}\|^2 \tilde{\mathbf{D}}^\top \tilde{\mathbf{D}})}{\Phi_1(0; 0, 1 + \|\tilde{\eta}\|^4)}. \quad \square$$

Proof of Corollary 1.

- (i) Follows from (1.8) and Proposition 5.
- (ii) See Proposition 3.
- (iii) Right side: see Contreras-Reyes and Arellano-Valle (2012). Left side: consider the nonsymmetrical entropy of Liu (2009) given by

$$S(\mathbf{u}) = - \int_{\mathbb{R}^k} f(\mathbf{u}) \log[\beta(\mathbf{u}) f(\mathbf{u})] d\mathbf{u}, \quad (\text{A.1})$$

where $f(\mathbf{u})$ is the probability density function of a normal variable \mathbf{u} . By choosing $\beta(\mathbf{u}) = 2\Phi_1(\eta^\top \Omega^{-1/2}(\mathbf{u} - \mu))$, $\mathbf{u} = \mathbf{Z}_N$, it follows that $E\{\log[\beta(\mathbf{Z}_N)]\} = \log(2) + \Phi_1(0) = (1/2)\ln(4e)$ (see Proposition 4 of Azzalini and Dalla-Valle, 1996). Then, as $E\{\log[\beta(\mathbf{Z})]\} \leq 2E\{\log[\beta(\mathbf{Z}_N)]\}$, the result is obtained.

- (iv) Follows from properties (i), (ii) and (1.6). \square

Proof of Lemma 4.

- (i) See e.g. Arellano-Valle and Genton (2005) and Arellano-Valle and Azzalini (2006).
- (ii) It is straightforward from part (i).
- (iii) It comes from (ii) and the well-known fact that $E\{(\mathbf{Z} - \mathbf{a})^\top \mathbf{B}(\mathbf{Z} - \mathbf{a})\} = \text{tr}\{\mathbf{B}E(\mathbf{Z}\mathbf{Z}^\top)\} - 2\mathbf{a}^\top \mathbf{B}E(\mathbf{Z}) + \mathbf{a}^\top \mathbf{B}\mathbf{a}$, see also Genton et al. (2001).
- (iv) It follows from the part (i) of this lemma that $\tilde{\eta}^\top(\mathbf{Z} - \tilde{\xi}) \stackrel{d}{=} \tilde{\eta}^\top(\xi - \tilde{\xi}) + \tilde{\eta}^\top \delta |U_0| + \tilde{\eta}^\top \mathbf{U}$. Since $\tilde{\eta}^\top \mathbf{U} \sim N_1(0, \tilde{\eta}^\top \Omega \tilde{\eta} - (\tilde{\eta}^\top \delta \delta)^\top)$, which is independent of U_0 , then we can write $\tilde{\eta}^\top \mathbf{U} = \sqrt{1 - \delta_0^2} U_1$, where $\delta_0 = \tilde{\eta}^\top \delta / \sqrt{\tilde{\eta}^\top \Omega \tilde{\eta}}$ and $U_1 \sim N_1(0, 1)$ and is independent of U_0 . Hence, we obtain $\tilde{\eta}^\top(\mathbf{Z} - \tilde{\xi}) \stackrel{d}{=} \tilde{\eta}^\top(\xi - \tilde{\xi}) + \sqrt{\tilde{\eta}^\top \Omega \tilde{\eta}} Z_0$, where $Z_0 = \tilde{\delta} |U_0| + \sqrt{1 - \delta_0^2} U_1$. Since $Z_0 \sim SN_1(0, 1, \eta_0)$, where $\eta_0 = \delta_0 / \sqrt{1 - \delta_0^2} =$

$\tilde{\eta}^\top \delta / \sqrt{\tilde{\eta}^\top \Omega \tilde{\eta} - (\tilde{\eta}^\top \delta)^2}$, the proof follows. \square

Proof of Lemma 5. By (2.4) we have for the logarithm of the pdf of $\mathbf{Y} \sim SN_k(\xi_2, \Omega_2, \eta_2)$ that

$$\begin{aligned} \log f_{\mathbf{Y}}(\mathbf{x}) &= \log \phi_k(\mathbf{x}; \xi_2, \Omega_2) + \log[2\Phi\{\eta_2^\top(\mathbf{x} - \xi_2)\}] \\ &= -\frac{1}{2} \left\{ \log\{(2\pi)^k |\Omega_2|\} + (\mathbf{x} - \xi_2)^\top \Omega_2^{-1} (\mathbf{x} - \xi_2) \right\} + \log[2\Phi\{\eta_2^\top(\mathbf{x} - \xi_2)\}]. \end{aligned}$$

Thus, since by (1.12) $CH(\mathbf{X}, \mathbf{Y}) = -E[\log f_{\mathbf{Y}}(\mathbf{X})]$, we have by applying the Lemma 4(iii) with \mathbf{Z} replaced by \mathbf{X} , $\mathbf{a} = \xi_2$ and $\mathbf{B} = \Omega_2^{-1}$ that

$$\begin{aligned} CH(\mathbf{X}, \mathbf{Y}) &= \frac{1}{2} \left\{ k \log(2\pi) + \log |\Omega_2| + E\{(\mathbf{X} - \xi_2)^\top \Omega_2^{-1} (\mathbf{X} - \xi_2)\} \right\} \\ &\quad - E[\log\{2\Phi(\eta_2^\top(\mathbf{X} - \xi_2))\}] \\ &= \frac{1}{2} \left\{ k \log(2\pi) + \log |\Omega_2| + \text{tr}(\Omega_2^{-1} \Omega_1) + (\xi_1 - \xi_2)^\top \Omega_2^{-1} (\xi_1 - \xi_2) \right. \\ &\quad \left. + 2\sqrt{\frac{2}{\pi}} (\xi_1 - \xi_2)^\top \Omega_2^{-1} \delta_1 \right\} - E[\log\{2\Phi(\eta_2^\top(\mathbf{X} - \xi_2))\}] \\ &= \frac{1}{2} \left\{ k \log(2\pi) + \log |\Omega_2| + \text{tr}(\Omega_2^{-1} \Omega_1) + (\xi_1 - \xi_2)^\top \Omega_2^{-1} (\xi_1 - \xi_2) \right\} \\ &\quad + \sqrt{\frac{2}{\pi}} (\xi_1 - \xi_2)^\top \Omega_2^{-1} \delta_1 - E[\log\{2\Phi(\eta_2^\top(\mathbf{X} - \xi_2))\}]. \end{aligned}$$

From Lemma 4(iii) we find that the random variable $\eta_2^\top(\mathbf{X} - \xi_2)$ has the same distribution of W_{21} in (3.9). Thus, the proof follows by noting that

$$\begin{aligned} CH(\mathbf{X}_0, \mathbf{Y}_0) &= -E[\log \phi_k(\mathbf{X}_0; \xi_2, \Omega_2)] \\ &= \frac{1}{2} \left\{ k \log(2\pi) + \log |\Omega_2| + \text{tr}(\Omega_2^{-1} \Omega_1) + (\xi_1 - \xi_2)^\top \Omega_2^{-1} (\xi_1 - \xi_2) \right\}. \square \end{aligned}$$

Proof of Proposition 6. Note first by Lemma 5 that $H(\mathbf{X}) = CH(\mathbf{X}, \mathbf{X})$ is given by

$$\begin{aligned} H(\mathbf{X}) &= \frac{1}{2} \{k + k \log(2\pi) + \log |\Omega_1|\} - E[\log\{2\Phi(\eta_1^\top(\mathbf{X} - \xi_1))\}] \\ &= H(\mathbf{X}_0) - E[\log\{2\Phi(\eta_1^\top(\mathbf{X} - \xi_1))\}], \end{aligned}$$

where $H(\mathbf{X}_0) = CH(\mathbf{X}_0, \mathbf{X}_0) = \frac{1}{2} \{k + k \log(2\pi) + \log |\Omega_1|\}$ and by the property (iii) of the Lemma 4 we have $\eta_1^\top(\mathbf{X} - \xi_1) \stackrel{d}{=} W_{11}$. Thus, the proof follows from the fact that

$$K(\mathbf{X}, \mathbf{Y}) = CH(\mathbf{X}, \mathbf{Y}) - H(\mathbf{X}). \quad \square$$

Proof of Proposition 8. Considering the first entropy of (4.7) and (2.25), we get

$$\begin{aligned} H\left(\frac{X+Y}{2}\right) &= -\int_{-\infty}^{\infty} \phi(x)\Phi(\eta_1x) \log\{\phi(x)[\Phi(\eta_1x) + \Phi(\eta_2x)]\}dx, \\ &\quad -\int_{-\infty}^{\infty} \phi(x)\Phi(\eta_2x) \log\{\phi(x)[\Phi(\eta_1x) + \Phi(\eta_2x)]\}dx, \\ &= -\frac{1}{2}E[\log \phi(x)] - \frac{1}{2}E[\log \phi(y)] - \frac{1}{2}E[\log\{\Phi(\eta_1x) + \Phi(\eta_2x)\}] \\ &\quad - \frac{1}{2}E[\log\{\Phi(\eta_1y) + \Phi(\eta_2y)\}]. \end{aligned}$$

For the first expected value of the last expression, we have $E[\log \phi(x)] = -(1/2)E[\log(2\pi) + X^2] = -(1/2)[\log(2\pi) + 1] = -H(U) = E[\log \phi(y)]$. Then,

$$H\left(\frac{X+Y}{2}\right) = H(U) - \frac{1}{2}E[\log\{\Phi(\eta_1x) + \Phi(\eta_2x)\}] - \frac{1}{2}E[\log\{\Phi(\eta_1y) + \Phi(\eta_2y)\}].$$

Finally, by replacing (3) (case $k = 1$) and the last result in (4.7), the JS distance is obtained.

□

Proof of Proposition 9.

- (i) For any finite mixture $f(\mathbf{x}; \tilde{\boldsymbol{\theta}}, \boldsymbol{\pi}) = \sum_{i=1}^m \pi_i f(\mathbf{x}; \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i$ is the associated parameter set of each i -th component $f(\mathbf{x}; \boldsymbol{\theta}_i)$, $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_m)$, $\pi_i \geq 0$, $\sum_{i=1}^m \pi_i = 1$, and $\mathbf{X} \in \mathbb{R}^k$ not necessarily normal with non-zero location vector and dispersion matrix Λ . Then,

$$\sum_{i=1}^m \pi_i H[\mathbf{X}; \boldsymbol{\theta}_i] \leq H[\mathbf{X}; \tilde{\boldsymbol{\theta}}] \leq \frac{1}{2} \ln\{(2\pi e)^k |\Lambda|\}. \quad (\text{A.2})$$

For a proof of (A.2), see pp. 27 and 663 of Cover and Thomas (2006). Basically, the fact that $g(t) = -\log t$ is a concave function ($-g(t)$ is convex) allows the use of Jensen's inequality. Then, considering the location vector (2.11) and dispersion matrix (2.12), we have the inequalities $H[\mathbf{Y}; \tilde{\boldsymbol{\theta}}] \leq A_{upper}$ and $H[\mathbf{Y}; \tilde{\boldsymbol{\theta}}] \geq \sum_{i=1}^m \pi_i H[\mathbf{Y}; \boldsymbol{\theta}_i]$. Considering the Proposition 3 and the condition $\sum_{i=1}^m \pi_i = 1$, we prove the left side of the inequality.

- (ii) Left side: by the property of log concavity for skew-normal densities (Gupta and Brown, 2001) and employing Jensen's inequality Cover and Thomas (2006), the proof is analogous to Theorem 2 of Huber et al. (2008). Right side: see Theorem 3 of Huber et al. (2008). □

Proof of Lemma 6. Consider the Proposition 1 (B1) of Bennett (1986). Let $p \geq 1$, then for a α th-order, $1 < \alpha \leq p$, we have

$$\begin{aligned}
 f(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \pi)^\alpha &= \left[\sum_{i=1}^m \pi_i f(\mathbf{y}; \boldsymbol{\theta}_i) \right]^\alpha \\
 &\stackrel{(B1)}{\geq} \left(\sum_{i=1}^m f(\mathbf{y}; \boldsymbol{\theta}_i)^p \right)^{\frac{\alpha}{p}-1} \left[\sum_{i=1}^{m-1} \left\{ i^{1-\frac{\alpha}{p}} \left(\sum_{k=1}^i \pi_k \right)^\alpha \left(f(\mathbf{y}; \boldsymbol{\theta}_i)^p - f(\mathbf{y}; \boldsymbol{\theta}_{i+1})^p \right) \right\} \right. \\
 &\quad \left. + m^{1-\frac{\alpha}{p}} \left(\sum_{k=1}^m \pi_k \right)^\alpha f(\mathbf{y}; \boldsymbol{\theta}_m)^p \right]. \tag{A.3}
 \end{aligned}$$

By choosing $p = \alpha$ related to condition (iii) of Proposition 1 of Bennett (1986) in (A.3), the following equality holds

$$f(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \pi)^\alpha \geq f(\mathbf{y}; \boldsymbol{\theta}_m)^\alpha + \sum_{i=1}^{m-1} \left\{ \left(\sum_{k=1}^i \pi_k \right)^\alpha \left(f(\mathbf{y}; \boldsymbol{\theta}_i)^\alpha - f(\mathbf{y}; \boldsymbol{\theta}_{i+1})^\alpha \right) \right\}. \tag{A.4}$$

The conditions (i), (ii) and (iv) of Proposition 1 of Bennett (1986) can not be accomplished given the Rényi entropy conditions of (1.6), thus the equality in (A.4) is not accomplished. Finally, integrating both sides of (A.4) the result is obtained. \square

Proof of Lemma 7. Stirling's approximation (3.16) yields

$$U_\alpha(\tilde{\alpha}|Q) \approx e^{\alpha H(\tilde{\alpha}|Q)}, \tag{A.5}$$

as $\alpha \rightarrow \infty$. By replacing (A.5) in (3.14), we obtain

$$\begin{aligned}
 \int_{\mathbb{R}^k} f(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \pi)^\alpha d\mathbf{y} &\approx \int_{\mathbb{R}^k} \sum_{k_i \in \mathcal{A}} e^{\alpha H(\tilde{\alpha}|Q)} d\mathbf{y}, \\
 &= \int_{\mathbb{R}^k} \sum_{k_i \in \mathcal{A}} \exp \left\{ -\alpha \sum_{i=1}^m \alpha_i \log \left(\frac{\alpha_i}{Q_i(\mathbf{y})} \right) \right\} d\mathbf{y}, \\
 &= \int_{\mathbb{R}^k} \sum_{k_i \in \mathcal{A}} \exp \left\{ \sum_{i=1}^m \ln \left(\frac{\alpha_i}{Q_i(\mathbf{y})} \right)^{-\alpha \alpha_i} \right\} d\mathbf{y}, \\
 &= \int_{\mathbb{R}^k} \sum_{k_i \in \mathcal{A}} \prod_{i=1}^m \left(\frac{\alpha_i}{Q_i(\mathbf{y})} \right)^{-k_i} d\mathbf{y}, \\
 &= \sum_{k_i \in \mathcal{A}} \left[\prod_{i=1}^m \alpha_i^{-k_i} \right] \left[\int_{\mathbb{R}^k} \prod_{i=1}^m Q_i(\mathbf{y})^{k_i} d\mathbf{y} \right].
 \end{aligned}$$

Finally, by replacing $Q_i(\mathbf{y}) = \pi_i f(\mathbf{y}; \boldsymbol{\theta}_i)$ in the last expression, the result is obtained. \square

Proof of Lemma 8. From (3.14), we have

$$\int_{\mathbb{R}^k} f(\mathbf{y}; \tilde{\boldsymbol{\theta}}, \pi)^\alpha d\mathbf{y} = \sum_{k_i \in \mathcal{A}} \frac{\alpha!}{k_1! \cdots k_m!} \left(\prod_{i=1}^m \pi_i^{k_i} \right) \int_{\mathbb{R}^k} \prod_{i=1}^m f(\mathbf{y}; \boldsymbol{\theta}_i)^{k_i} d\mathbf{y}. \quad (\text{A.6})$$

(i) For the integral of (A.6), we applied the GH inequality (Finner, 1992) for m products to obtain

$$\int_{\mathbb{R}^k} \prod_{i=1}^m f(\mathbf{y}; \boldsymbol{\theta}_i)^{k_i} d\mathbf{y} \stackrel{(\text{GH})}{\leq} \prod_{i=1}^m \left(\int_{\mathbb{R}^k} f(\mathbf{y}; \boldsymbol{\theta}_i)^{q_i} d\mathbf{y} \right)^{1/p_i} \quad (\text{A.7})$$

$$\stackrel{(1.6)}{=} \exp \left(\sum_{i=1}^m \left(\frac{1-q_i}{p_i} \right) R_{q_i}[\mathbf{Y}; \boldsymbol{\theta}_i] \right), \quad (\text{A.8})$$

with $q_i = p_i k_i$, $p_i > 1$, $k_i \in \mathcal{A}$, and $\sum_{i=1}^m (1/p_i) = 1$, $i = 1, \dots, m$. Then, the last term of (A.8) can be upper bounded using Chebyshev's inequality for sums (see formula 11.115 of Gradshteyn and Ryzhik, 2007) as

$$\begin{aligned} \sum_{i=1}^m \left(\frac{1-q_i}{p_i} \right) R_{q_i}[\mathbf{Y}; \boldsymbol{\theta}_i] &\leq \frac{1}{m} \left(\sum_{i=1}^m \frac{1-q_i}{p_i} \right) \left(\sum_{i=1}^m R_{q_i}[\mathbf{Y}; \boldsymbol{\theta}_i] \right) \\ &= \frac{1-\alpha}{m} \sum_{i=1}^m R_{q_i}[\mathbf{Y}; \boldsymbol{\theta}_i] \\ &\leq \frac{1-\alpha}{m} \sum_{i=1}^m R_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]. \end{aligned} \quad (\text{A.9})$$

In (A.9), the property $R_{q_i}[\mathbf{Y}; \boldsymbol{\theta}_i] \leq R_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]$, given that $q_i > k_i$, $i = 1, \dots, m$, is used. Finally, the result is straightforward from (A.8) and (A.9).

(ii) By choosing $p_i = \alpha/k_i$, we have that $\sum_{i=1}^m (1/p_i) = \sum_{i=1}^m (k_i/\alpha) = 1$ and $k_i < \alpha$. This implies that $1 < \alpha/k_i = p_i$. Then, we obtain from the inequality (A.7):

$$\begin{aligned} \prod_{i=1}^m \left(\int_{\mathbb{R}^k} f(\mathbf{y}; \boldsymbol{\theta}_i)^{q_i} d\mathbf{y} \right)^{1/p_i} &= \prod_{i=1}^m P_{q_i}[\mathbf{Y}; \boldsymbol{\theta}_i]^{1/p_i} \\ &\leq \left(\prod_{i=1}^m P_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]^{k_i} \right)^{1/\alpha} \end{aligned}$$

$$\stackrel{(AG)}{\leq} \left(\frac{1}{m} \sum_{i=1}^m P_{k_i}[\mathbf{Y}; \boldsymbol{\theta}_i]^{k_i} \right)^{m/\alpha}, \quad (\text{A.10})$$

where in (A.10), the Arithmetic-geometric (AG) inequality (see formula 11.116 of Gradshteyn and Ryzhik, 2007) was applied. Applying (A.10) in (A.7), the result is obtained. \square

Proof of Lemma 10:

- (i) See e.g. Kim and Mallick (2003).
- (ii) It is straightforward from the representation (2.14) (see Theorem 2.(a)-(b) of (Kim and Mallick, 2003) for details).
- (iii) The computation of the quadratic form comes from parts (i), (ii) and the fact

$$E[\mathbf{z}^\top \mathbf{A} \mathbf{z}] = K_2(\mathbf{Z}) \text{tr}(\mathbf{A} \boldsymbol{\Omega}) + \boldsymbol{\xi}^\top \mathbf{A} \boldsymbol{\xi} + 2K_1(\mathbf{Z}) \boldsymbol{\xi}^\top \mathbf{A} \boldsymbol{\delta}$$

(see Theorem 3.(a) of (Kim and Mallick, 2003)). For a proof of Property (iii), see Azzalini and Capitanio (2003) and Contreras-Reyes and Arellano-Valle (2012). \square

Proof of Lemma 11: By (2.13) we have for the logarithm of the pdf of $\mathbf{Y} \sim SN_k(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \eta_2)$ that

$$\begin{aligned} \log f_{\mathbf{Y}}(\mathbf{x}) &= \log \{B_k(v_2)\} - \frac{1}{2} \log |\boldsymbol{\Omega}_2| - \left(\frac{v_2 + k}{2} \right) \log \left(1 + \frac{\mathbf{Q}_2}{v_2} \right) \\ &\quad + \log \left\{ 2T \left(\sqrt{\frac{v_2 + k}{v_2 + \mathbf{Q}_2}} \boldsymbol{\eta}_2^\top (\mathbf{x} - \boldsymbol{\xi}_2); v_2 + k \right) \right\}. \end{aligned}$$

On the another hand, by the Taylor expansion series of $g(x) = \log(1 + x)$ around zero, $x > 0$, we have

$$\log(1 + x) = \sum_{k=1}^{\infty} g^{(k)}(0) \frac{x^k}{k!},$$

where $g^{(k)}(0)$ is the k -derivative of $g(x)$ around zero. Now is possible to obtain the following approximation for the calculus of the expected value of $1 + \mathbf{Q}_2/v_2$ with respect to the pdf $f_{\mathbf{X}}(\mathbf{x})$

$$\log \left(1 + \frac{u}{n} \right) = \frac{u}{n} + \mathcal{O}(n^{-2}), \quad (\text{A.11})$$

for any $u > -n$, and replacing $x = u/n$ when $n \rightarrow \infty$, i.e., $x \rightarrow 0$. This mean that exist a constant $\alpha > 0$ such that $|\log(1 + u/n) - u/n| \leq \alpha/n^2$. This yields

$$E \left[\log \left(1 + \frac{\mathbf{Q}_2}{v_2} \right) \right] \approx \frac{E[\mathbf{Q}_2]}{v_2},$$

as $v_2 \rightarrow \infty$. Thus, since by (1.12), $CH(\mathbf{X}, \mathbf{Y}) = -E[\log f_{\mathbf{Y}}(\mathbf{X})]$, by applying (2.15) and Lemma 10(ii) with \mathbf{Z} replaced by \mathbf{X} and $\mathbf{a} = \xi_2$ the proof follows. \square

Proof of Proposition 11. By (2.20), $\phi_k(\mathbf{y}; \mu, \Omega) = |\Omega|^{-1/2} \phi_k(\Omega^{-1/2}(\mathbf{y} - \mu))$, where $\phi_k(\mathbf{z})$ is the probability density function of $N_k(\mathbf{0}, I_k)$. Then, as in (5), to compute the integral $\int_{\mathbb{R}^k} [f(\mathbf{z})]^\alpha d\mathbf{z}$ we use the change of variables $\Omega_\alpha = \alpha^{-1} \Omega$ and $\mathbf{Z}_0 = \Omega_\alpha^{-1/2}(\mathbf{Z} - \mu)$. In this case, $\mathbf{Z}_0 \sim ESN_k(\mathbf{0}, I_k, \tilde{\eta}, \tau)$ with $\tilde{\eta} = \Omega_\alpha^{1/2} \eta$. We shall use the fact that $|\Omega_\alpha| = \alpha^{-k} |\Omega|$ for k -dimensional matrices (Nock and Nielsen, 2012). Then, according to Lemma 3, the integral $\int_{\mathbb{R}^k} [f(\mathbf{z})]^\alpha d\mathbf{z}$ should be rewritten in terms of an expected value with respect to a standardized normal density as

$$\begin{aligned} \int_{\mathbb{R}^k} [f(\mathbf{z})]^\alpha d\mathbf{z} &= \frac{1}{[\Phi_1(\tau)]^\alpha} |\Omega|^{-\frac{\alpha}{2}} |\Omega_\alpha|^{1/2} (2\pi)^{(1-\alpha)\frac{k}{2}} E\{[\Phi_1(\tilde{\eta}^\top \mathbf{z}_0 + \tilde{\tau})]^\alpha\} \\ &= \frac{1}{[\Phi_1(\tau)]^\alpha} \alpha^{-k} (2\pi)^{(1-\alpha)k/2} |\Omega|^{(1-\alpha)/2} E\{[\Phi_1(W)]^\alpha\}. \end{aligned}$$

where $W = \tilde{\eta}^\top \mathbf{Z}_0 + \tilde{\tau} \sim ESN_1(\tilde{\tau}, \|\tilde{\eta}\|^2, \|\tilde{\eta}\|, \tau)$ with $\|\tilde{\eta}\| = \tilde{\eta}^\top \tilde{\eta}$ (Arellano-Valle et al., 2013, Contreras-Reyes and Arellano-Valle, 2012), i.e., the expected value $E\{[\Phi_1(\tilde{\eta}^\top \mathbf{z}_0 + \tilde{\tau})]^\alpha\}$ is reduced from k dimensions to one dimension (Arellano-Valle et al., 2013, Contreras-Reyes, 2014). \square

Proof of Corollary 6.

(i) From Proposition 11, we obtain directly

$$\begin{aligned} R_\alpha(\mathbf{Z}) &= \frac{1}{1-\alpha} \{\log[\psi_{\alpha,k}(\Omega)] - \alpha \log[2\Phi_1(\tau)] + \log[E\{[\Phi_1(W)]^\alpha\}]\}, \\ &= R_\alpha(\mathbf{Z}_N) + \frac{\alpha}{1-\alpha} \log \left[\frac{1}{\Phi_1(\tau)} \right] + \frac{1}{1-\alpha} \log[E\{[\Phi_1(W)]^\alpha\}]. \end{aligned}$$

(ii) Considering Jensen's inequality, we obtain $E\{[\Phi_1(W)]^\alpha\} \geq [\Phi_1(E[W])]^\alpha$. Then, (ii) is straightforward from (2.21).

(iii) By (1.6), it follows that

$$H(\mathbf{Z}) = -E \left\{ \log \left[\phi_k(\mathbf{Z}_0) \frac{\Phi_1(\tilde{\eta}^\top \mathbf{Z}_0 + \tilde{\tau})}{\Phi_1(\tau)} \right] \right\} = H(\mathbf{Z}_0) - E \left\{ \log \left[\frac{\Phi_1(W)}{\Phi_1(\tau)} \right] \right\},$$

where, $\mathbf{Z}_0 = \Omega^{-1/2}(\mathbf{Z} - \mu) \sim ESN_k(\mathbf{0}, I_k, \tilde{\eta}, \tau)$ and $W = \tilde{\eta}^\top \mathbf{Z}_0 + \tilde{\tau} \sim ESN_1(\tilde{\tau}, \|\tilde{\eta}\|^2, \|\tilde{\eta}\|, \tau)$.

(iv) Right side: by Cover and Thomas (2006), for any density $g(\mathbf{x})$ of a random vector $\mathbf{x} \in \mathbb{R}^k$ (not necessary normal) with zero mean and variance $\Omega = E[\mathbf{X}\mathbf{X}^\top]$, the Shannon entropy of \mathbf{X} is maximized under normality as $H(\mathbf{X}) \leq (1/2)\log[(2\pi e)^k |\Omega|]$. Then, the result is obtained from (2.22). Left side: as in (iii), by choosing $\beta(\mathbf{u}) = \Phi_1(\eta^\top \Omega^{-1/2}(\mathbf{u} - \mu) + \tilde{\tau})/\Phi_1(\tau)$ in the nonsymmetrical entropy (A.1), it follows that

$$E\{\log \beta(\mathbf{Z}_N)\} = \Phi_1 \left(\frac{\tilde{\tau}}{\sqrt{1 + \|\eta\|}} \right) - \log[\Phi_1(\tau)]$$

(see Proposition 4 of Azzalini and Dalla-Valle, 1996). Then, as $E\{\log[\beta(\mathbf{Z})]\} \leq E\{\log[\beta(\mathbf{Z}_N)]\}/\Phi_1(\tau)$, the result is obtained.

(v) Follows from properties (i), (iii) and (1.6). \square

Proof of Proposition 13. By (2.28), it follows that

$$\int_a^b [g(w)]^\alpha dw = \frac{1}{([F(z)]_a^b)^\alpha} \int_a^b [f(w)]^\alpha dw$$

and, by Proposition 5, the integral $\int_a^b [f(w)]^\alpha dw$ should be rewritten in terms of an expected value as

$$\int_a^b [f(w)]^\alpha dw = \psi_{\alpha,1}(\sigma^2) E\{[\Phi_1(u)]^\alpha | a_0 < u \leq b_0\},$$

where $U \sim SN_1(0, \tilde{\eta}^2, \tilde{\eta})$, $\tilde{\eta}^2 = \omega \eta^2 / \alpha$, $a_0 = \eta(a - \mu) / \omega$ and $b_0 = \eta(b - \mu) / \omega$. Again, by Lemma 2 and setting $\mu = 0$, $J = \tilde{\eta}^2$, $r = \alpha$, $d = s = \mathbf{A} = h(u) = 1$; we obtain $\tilde{\mathbf{A}} = I_{\alpha+1}$, $\tilde{\mathbf{D}} = (\mathbf{1}_\alpha, \tilde{\eta})^\top$ and $\tilde{\mathbf{\Omega}} = I_{\alpha+1} + \tilde{\eta}^2 \tilde{\mathbf{D}}^\top \tilde{\mathbf{D}}$. Then, the expected value is

$$E\{[\Phi_1(u)]^\alpha | a_0 < u \leq b_0\} = 2\Phi_{\alpha+1}(\mathbf{0}; \mathbf{0}, \tilde{\mathbf{\Omega}})[H(v)]_{a_0}^{b_0},$$

where $H(v)$ is the cumulative density function of a closed skew-normal variable $V \sim CSN_{1,2}(0, \tilde{\eta}^2, \tilde{\mathbf{B}}, \mathbf{0}, I_2)$ with $\tilde{\mathbf{B}} = (1, \tilde{\eta})^\top$ (see Proposition 3 of Flecher et al., 2010). \square

Appendix B

```
#####  
### The following measures can be computed using  
### skewtools (0.1.2) package of Contreras-Reyes (2012):  
### 1. Shannon entropy for multivariate and univariate SN & ST,  
### 2. Mutual information for multivariate and univariate SN & ST,  
### 3. KL divergence for multivariate and univariate SN.  
  
### For other measures, we presented the main R codes:  
  
##### SECTION 4.1.2 #####  
  
##### Univariate SN Renyi entropy #####  
  
Renyi.entropy <- function(alpha, xi, sigma2, lambda){  
  Re.normal = 0.5*log(sigma2*2*pi) + 0.5*log(alpha)/(alpha-1)  
  eta.tilde = sqrt(sigma2)*lambda  
  eta.norm = eta.tilde^2  
  D.tilde = c(rep(1,alpha),eta.norm)  
  Omega.tilde = diag(alpha+1) + eta.norm^2*(D.tilde%*%t(D.tilde))  
  omega = 1 + eta.norm^4  
  
  normL1 = pmnorm(rep(0, alpha+1), mean = rep(0, alpha+1),  
                  varcov = Omega.tilde)[1]  
  normL2 = pnorm(0, 0, sqrt(omega))  
  normL = normL1 / normL2  
  Neg = log((2^alpha)*normL)/(alpha-1)
```

```
Re.normal - Neg
}

##### Multivariate SN Renyi entropy #####

library(mnormt)

MRenyi.entropy <- function(alpha, xi, sigma2, lambda){
d = dim(sigma2)[1]
Re.normal = 0.5*log(det(sigma2)*(2*pi)^d) + 0.5*log(alpha)/(alpha-1)

eta.norm = as.numeric(t(lambda) %*% sigma2 %*% lambda)
D.tilde = c(rep(1,alpha), eta.norm)
Omega.tilde = diag(alpha+1) + eta.norm^2*(D.tilde%*%t(D.tilde))
omega = 1 + eta.norm^4

normL1 = pmnorm(rep(0, alpha+1), mean = rep(0, alpha+1),
                varcov = Omega.tilde)[1]
normL2 = pnorm(0, 0, sqrt(omega))
normL = normL1 / normL2

Neg = log((2^alpha)*normL)/(alpha-1)

Re.normal - Neg
}

##### SECTION 4.1.3 #####

##### Asymptotic ST KL divergence #####

library(skewtools)
library(sn)

## st density
```

```

theta1 = c(0,1,2,5) # Example
theta2 = c(0,2,3,4)

## Numerical

dst1 <- function(x, location, scale, shape, df){
  z <- (x - location)/scale
  pdf <- dt(z, df = df)
  cdf <- pt(shape * z * sqrt((df + 1)/(z^2 + df)), df = df + 1)
  2 * pdf * cdf/scale
}

CE.st.I <- function(x){ # Cross-entropy of ST
  fx <- dst1(x, location=theta1[1], scale=theta1[2],
             shape=theta1[3], df=theta1[4])
  fy <- dst1(x, location=theta2[1], scale=theta2[2],
             shape=theta2[3], df=theta2[4])
  -fx*log(fy)
}

integrate(CE.st.I, lower=-Inf, upper=Inf, subdivisions=100)$value

## Asymptotic

k = 1
K_m <- function(m,nu) (nu/2)^(m/2)*gamma((nu-m)/2)/gamma(nu/2)
Bk_nu <- gamma((theta2[4]+k)/2) / (gamma(theta2[4]/2)*(theta2[4]*pi)^(k/2))

delta <- theta1[2]*theta1[3] / sqrt(1+theta1[2]*theta1[3]^2)
EQ <- K_m(2,theta1[4])*theta1[2]/theta2[2] +
      (theta1[1]-theta2[1])^2/theta2[2] +
      2*K_m(1,theta1[4])*(theta1[1]-theta2[1])*delta/theta2[2]
A <- (theta2[4]+k)*EQ / (2*theta2[4])

Elog <- function(x) {
  Q = (x-theta2[1]) / theta2[2]

```

```
eta2 = theta2[3]* sqrt((theta2[4]+k)/(theta2[4]+Q^2))
fx <- dst1(x, location=theta1[1], scale=theta1[2],
          shape=theta1[3], df=theta1[4])
log(2*pt(eta2*(x-theta2[1]), df=theta2[4]+k))*fx
}

EV = integrate(Elog, lower=-Inf, upper=Inf, subdivisions=100)$value
-log(Bk_nu) + 0.5*log(theta2[2]) + A - EV

##### Gamma function

B.aprox <- function(nu,k) (2*pi)^(-k/2)
B <- function(nu,k) gamma((nu+k)/2)/(gamma(nu/2)*(nu*pi)^(k/2))

##### T-Student

DKL.T <- function(theta1, theta2){
A1=0.5*log(theta2[1]/1)
A2=0.5*((theta2[2]+1)/theta2[2])*(theta1[2]/(theta1[2]-2))*
      (theta1[1]/theta2[1])
A3=0.5*(theta1[2]+1)*(digamma(0.5*(theta1[2]+1))-digamma(0.5*theta1[2]))
A1 + A2 - A3
}

##### SECTION 4.1.4 #####

##### JS SN divergence #####

JS.sn <- function(eta1, eta2){

eta1 = abs(eta1); eta2 = abs(eta2)

Emint <- function(a) {
f1 <- function(x) 2*dnorm(x)*pnorm(a*x)*log(2*pnorm(a*x))
try(2*integrate(f1, lower=0, upper=Inf, subdivisions=100)$value, TRUE)
```

```

}

Emint2 <- function(a,b) {
f2 <- function(x) 2*dnorm(x)*pnorm(a*x)*log(pnorm(a*x)+pnorm(b*x))
try(2*integrate(f2, lower=0, upper=Inf, subdivisions=100)$value, TRUE)
}

JS = 0.5*(Emint(eta1) - Emint2(eta1,eta2) + Emint(eta2) - Emint2(eta2,eta1))
JS = as.numeric(JS)
return(JS)
}

##### SECTION 4.2 #####

##### Multivariate FMSN Shannon entropy bounds ###

Us.entropy <- function(model){

Emint <- function(a) {
  f1 <- function(x) 2 * pnorm(a * x) * log(2 * pnorm(a * x)) * dnorm(x)
  2*integrate(f1, 0, 100, subdivisions = 100)$value
}

Emint2 <- function(a,b) {
  f1 <- function(x) 2 * pnorm(a * x) * log(2 * pnorm(a * x)) * dnorm(x)
  f2 <- function(x) 2 * pnorm(b * x) * log(2 * pnorm(b * x)) * dnorm(x)
  f <- function(x) f1(x)*f2(x)
  2*integrate(f, 0, 100, subdivisions = 100)$value
}

P = model$pii
m = length(P)

Alower <- Bupper <- Blower <- rep(0, m)
VY = 0

```

```
for(i in 1:m){
Xi = model$mu[[i]]
Omega = model$sigma2[[i]]
Eta = model$shape[[i]]
Delta <- as.numeric(sqrt(1 + Omega*Eta^2)^(-1)) * Omega*Eta
N.Eta = as.numeric(sqrt(Omega*Eta^2))

EY <- P[i]*(Xi + sqrt(2/pi)*Delta)
Mu <- Xi + sqrt(2/pi)*Delta - EY
VY <- VY + P[i]*(Omega - (2/pi)*(Delta^2) + Mu^2)

Alower[i] = P[i]*Emint(N.Eta)
Bupper[i] = P[i]*log(P[i])

Blower.aux <- rep(0, m)

for(s in 1:m) {
Omega.aux = model$sigma2[[s]]
Eta.aux = model$shape[[s]]
N.Eta2 = as.numeric(sqrt(Omega.aux * Eta^2))
Blower.aux[s] = P[s]*Emint2(N.Eta, N.Eta2)
}

Blower[i] = P[i]*log(sum(Blower.aux))
}

Aupper.new = 0.5*log(VY*2*pi*exp(1))
Alower.new = Aupper.new - sum(Alower)
Bupper.new = Alower.new - sum(Bupper)
Blower.new = -sum(Blower)

return(list(Aupper = Aupper.new, Alower = Alower.new,
            Bupper = Bupper.new, Blower = Blower.new))
}

##### Multivariate FMSN Renyi entropy bounds ###
```

```

Ur.entropy <- function(alpha, model){

  p = model$pii
  xi = model$mu
  sigma2 = model$sigma2
  lambda = model$shape

  delta = sigma2*lambda / sqrt(1 + lambda^2*sigma2)
  EY = sum(p*(xi + sqrt(2/pi)*delta))
  mu = xi + sqrt(2/pi)*delta - EY
  varY = sum(p*(sigma2 - (2/pi)*delta^2 + mu^2))
  b = 3*alpha - 1
  F = 0.5*log(pi*b/(alpha-1)) + log(b/log(2*alpha))/(alpha-1)
      + log(gamma(alpha/(alpha-1))) - log(gamma(b/(2*(alpha-1))))
  RE.up = 0.5*log(varY) + F

  m = length(p)
  RE.low = exp((1-alpha) * Renyi.entropy(alpha, xi[m], sigma2[m], lambda[m]))

  for(i in 1:(m-1)){
    pi.new = sum(p[1:i])^alpha
    RE.low = RE.low +
    pi.new * (exp((1-alpha) * Renyi.entropy(alpha, xi[i], sigma2[i], lambda[i]))
    - exp((1-alpha) * Renyi.entropy(alpha, xi[i+1], sigma2[i+1], lambda[i+1])))
  }

  RE.low = log(RE.low)/(1 - alpha)

  return(list(RE.up = RE.up, RE.low = RE.low))
}

##### SECTION 4.3 #####

##### SN (Skew-Normal) FUNCTIONS #####

```

Moments for Normal distribution (Henze, 1986)

```
momN <- function(m){
K=0
x=length(which(m %% 2 != 0))
if(x==0) K=factorial(m)/((2^(m/2))*factorial(m/2))
K
}
```

Coefficients $a_k(m)$ for SN moments (Martinez et al., 2008)

```
coef.SN <- function(t){
if(t==1) A <- matrix(1,1,1)
if(t > 1){
A <- matrix(NA,t,t)
A[1,1] = 1
for(k in 2:t){
A[k,1] = (2*k-1)*A[k-1,1]
for(m in 2:k) A[k,m] = 2*(k-1)*(A[k-1,m] + A[k-1,m-1])
A[k,k] = 2*(k-1)*A[k-1,k-1]
}
}
return(A)
}
```

Moments for SN distribution:

```
momSN <- function(tau, m, method=1){
# method=1: Martinez et al.
# method=2: Henze
# method=3: Numerical integration
is.even <- function(x) x %% 2 == 0
is.odd <- function(x) x %% 2 != 0
if(is.even(m)) mu = factorial(m)/((2^(m/2))*factorial(m/2))
if(is.odd(m)) {
b = sqrt(2/pi)
```

```

k = (m+1)/2
if(method==1){
  A = coef.SN(m)
  mu = 0
  for(i in 1:k) mu = mu + A[k,i]*tau^(2*i-1)
  mu = mu*b*(1+tau^2)^(-m/2)
}
if(method==2){
  mu = 0
  for(i in 0:(k-1)){
    mu = mu + factorial(i)*(2*tau)^(2*i)/(factorial(2*i+1)*factorial(k-i-1))
  }
  mu = mu*b*tau*factorial(2*k-1)*(1+tau^2)^(-m/2)/2^(k-1)
}
if(method==3){
  sn.mod <- function(u, tau, k){
    u^k*dnorm(u)*pnorm(tau*u)
  }
  mu = 2*integrate(sn.mod, lower=-Inf, upper=Inf, tau=tau, k=k)$value
}
}
mu
}

```

Cumulant function:

```

KP <- function(m){
  K=rep(0,m)
  for(i in 1:m){
    if(i<=4) K[i]=zeta(i,0) # require 'sn' library
    if(i>=5){
      K[i]=(i-2)*K[i-2]
      x=length(which(i %% 2 != 0))
      if(x==0){ # i is even
        for(p in 1:(0.5*i)){
          if(K[p]==K[i-p]) aux=choose(i-2,p-1)*K[p]*K[i-p]

```

```
if(K[p]!=K[i-p]) aux=2*choose(i-2,p-1)*K[p]*K[i-p]
K[i]=K[i]+aux
}
}
if(x==1){ # i is odd
for(p in 1:(0.5*(i-1))) K[i]=K[i]+2*choose(i-2,p-1)*K[p]*K[i-p]
}
K[i]=-K[i]
}
}
round(K,6)
}

#### SN negentropy:

neg.SN <- function(tau, t){
b = sqrt(2/pi)
delta.tau = tau/sqrt(1+tau^2)
Exp.SN = 0
for(i in 1:t) Exp.SN = Exp.SN + KP(i)[i]*momSN(tau,i)*(tau^i)/factorial(i)
N = 0.5*log(1-(b*delta.tau)^2) + Exp.SN
H = 0.5*log(2*pi*exp(1)) - Exp.SN
list(Entropy=H, Negentropy=N, K=Exp.SN)
}

##### MSN (Modified Skew-Normal) FUNCTIONS #####

### Integral xi_k(tau) for odd MSN moments:

xi <- function(k, tau){
b = sqrt(2/pi)
msn.mod <- function(x, tau, k) {
exp(-x/2)*pnorm(tau*sqrt(x)/sqrt(1+x))*x^k/(gamma(k+1)*2^(k+1))
}
xi0 = integrate(msn.mod, lower=0, upper=Inf, tau=tau, k=k)
xi0$value
```

```
}
```

```
### MSN moments of Z_tau:
```

```
momMSN <- function(tau,k){
  b = sqrt(2/pi)
  is.even <- function(x) x %% 2 == 0
  is.odd <- function(x) x %% 2 != 0
  if(is.even(k)) mu = momN(k)
  if(is.odd(k)) mu = b*factorial(k-1)*2^(k-1)*(2*xi(k,tau) - 1)
  mu
}
```

```
### MSN moments of Z^{*}_tau for U:
```

```
momMSN1 <- function(tau,k){
  msn.mod <- function(u, tau, k){
    u^k*(1-u^2)^(-3/2)*dnorm(u/sqrt(1-u^2))*pnorm(tau*u)
  }
  mu = 2*integrate(msn.mod, lower=-1, upper=1, tau=tau, k=k)$value
  mu
}
```

```
#### MSN negentropy:
```

```
neg.MSN <- function(tau, t){
  b = sqrt(2/pi)
  Exp.MSN = 0
  for(i in 1:t) {
    Exp.MSN = Exp.MSN + KP(i)[i]*momMSN1(tau,i)*(tau^i)/factorial(i)
  }
  xi0 = xi(0, tau)
  N = 0.5*log(1-momMSN(tau,1)^2) + Exp.MSN
  H = 0.5*log(2*pi*exp(1)) - Exp.MSN
  list(Entropy=H, Negentropy=N, K=Exp.MSN)
}
```

