The multivariate leptokurtic-normal distribution and its application in model-based clustering

Luca BAGNATO[1], Antonio PUNZO[2] and Maria G. ZOIA[1]*

[1]*Dipartimento di Discipline Matematiche, Finanza Matematica e Econometria, Università Cattolica del Sacro Cuore, Milano, Italy*
[2]*Dipartimento di Economia e Impresa, Università di Catania, Milano, Italy*

Key words and phrases: Elliptical distributions; EM algorithm; mixture models; orthogonal polynomials.

MSC 2010: Primary 60E05; secondary 62H30

Abstract: This article proposes the elliptical multivariate leptokurtic-normal (MLN) distribution to fit data with excess kurtosis. The MLN distribution is a multivariate Gram–Charlier expansion of the multivariate normal (MN) distribution and has a closed-form representation characterized by one additional parameter denoting the excess kurtosis. It is obtained from the elliptical representation of the MN distribution, by reshaping its generating variate with the associated orthogonal polynomials. The strength of this approach for obtaining the MLN distribution lies in its general applicability as it can be applied to any multivariate elliptical law to get a suitable distribution to fit data. Maximum likelihood is discussed as a parameter estimation technique for the MLN distribution. Mixtures of MLN distributions are also proposed for robust model-based clustering. An EM algorithm is presented to obtain estimates of the mixture parameters. Benchmark real data are used to show the usefulness of mixtures of MLN distributions. *The Canadian Journal of Statistics* 45: 95–119; 2017 © 2016 Statistical Society of Canada

Résumé: Cet article porte sur le rôle joué par une distribution leptokurtique dérivée de la distribution normale multivariée (NM) selon une expansion de type Gram-Charlier. Les auteurs proposent la distribution normale leptokurtique multivariée (NLM), une distribution qui s'exprime sous forme analytique et dont un paramètre capture le surplus d'applatissement. Ils obtiennent la distribution NLM en modifiant la représentation elliptique de la NM avec des polynômes orthogonaux. La force de cette approche réside dans sa polyvalence puisqu'elle peut être appliquée à toute distribution elliptique dans le but d'obtenir une distribution adéquate pour les données. Les auteurs examinent la méthode du maximum de vraisemblance en vue de l'estimation des paramètres d'intérêt. Ils suggèrent également les mélanges de NLM pour une analyse de regroupement robuste et présentent un algorithme espérance-maximisation pour l'estimation des paramètres du mélange par le maximum de vraisemblance. Ils montrent finalement l'efficacité des mélanges de NLM en les appliquant à des données réelles. *La revue canadienne de statistique* 45: 95–119; 2017 © 2016 Société statistique du Canada

1. INTRODUCTION

Statistical inference dealing with continuous multivariate data is commonly focused on elliptical distributions (Cambanis, Huang, & Simons, 1981); among them the normal one is the most widely used because of computational and theoretical convenience. However for many applied problems like those inherent to financial asset returns and stock indexes the kurtosis of the normal distribution is lower than required (Szegö, 2004).

Additional supporting information may be found in the online version of this article at the publisher's website.
* *Author to whom correspondence may be addressed.*
E-mail: maria.zoia@unicatt.it

Unlike the concepts of location, spread, and skewness, the meaning of kurtosis is a topic of considerable debate (see, e.g., Balanda & MacGillivray, 1988, 1990; Wang & Zhou, 2012) and we can only say that the statistical concept behind it is concerned with the curvature, the amount of peakedness, and the tail weight of a distribution (Arevalillo & Navarro, 2012). The classical notion of univariate kurtosis is moment-based and given by the standardized fourth central moment. A natural multivariate extension is

$$\operatorname{Kurt}(\boldsymbol{X}) = E\left\{\left[(\boldsymbol{X}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{X}-\boldsymbol{\mu})\right]^2\right\},$$

where $\boldsymbol{X}$ is a random vector having mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (Mardia, 1970). Hereafter we will use this notion of kurtosis.

Several multivariate distributions have been proposed to account for the leptokurtic nature of empirical data (see, e.g., Akgiray & Booth, 1988; Mittnik, Rachev, & Kim, 1998; Szegö, 2004; and the references therein). Typically the multivariate normal (MN) distribution is embedded in a larger model with one or more additional parameters denoting the deviation from normality in terms of kurtosis. Significant examples are the multivariate power exponential (MPE) distribution (Gómez, Gómez-Viilegas, & Marin, 1998) and the multivariate t (Mt) distribution (see, e.g., Lange, Little, & Taylor, 1989; Kotz & Nadarajah, 2004); both models have only one additional parameter. The MPE distribution also allows for light tails. A further alternative is the so-called multivariate contaminated normal (MCN) distribution, a two-component normal mixture in which the component with a larger prior probability represents the "good" observations (bulk of the data), and the other, with the same mean and an inflated covariance matrix, represents the "bad" observations (Aitkin & Wilson, 1980). This model requires two additional parameters (the proportion of good observations and the inflation parameter) with respect to the MN distribution. Note that all of these models can be seen as scale mixtures of normal distributions by choosing convenient scaling variables (see Lange, Little, & Taylor, 1989 for the Mt; Gómez-Sánchez-Manzano, Gómez-Villegas, & Marín, 2008 for the leptokurtic versions of the MPE; and Punzo & McNicholas, 2014 for the MCN).

However the kurtosis of multivariate data can be different from that implied by the models discussed above. Motivated by this consideration we propose the multivariate leptokurtic-normal (MLN) distribution which is the multivariate Gram–Charlier expansion of the MN distribution. In detail, the MLN distribution is obtained by reshaping the generating variate of its elliptical representation (Cambanis, Huang, & Simons, 1981) with the associated orthogonal polynomials (Zoia, 2010) which, unlike the Hermite ones, are not common in the literature. The result is a distribution characterized by one additional parameter corresponding to the excess kurtosis with respect to the original MN distribution. As confirmed by the results reported in the present work the MLN distribution appears suitable in fitting several real data sets showing different levels of excess kurtosis. It is worth noticing that the approach proposed here to obtain the MLN distribution can be easily extended to other multivariate elliptical distributions. When it is applied to leptokurtic distributions the resulting Gram–Charlier-like expansions can be used to adjust kurtoses greater than those treatable with the MLN distribution. This issue is going to be investigated by the authors.

The article is organized as follows. Some preliminary results are given in Section 2 about elliptical distributions and the polynomial reshaping method proposed by Zoia (2010). Section 3 presents two of the main contributions of the work namely the MLN distribution and its genesis; moreover maximum likelihood is described to estimate the parameters of the proposed distribution. Section 4 illustrates the use of the MLN distribution in robust clustering based on mixture models, which is a further proposal of the present article, and presents an EM algorithm to fit mixtures of MLN distributions. Section 5 investigates the performance of these mixtures, in comparison with mixtures of some well-established multivariate elliptically countered distributions, on benchmark real data. Conclusions as well as avenues for further research are given in Section 6. To increase the readability details are deferred to the Appendix.

2. PRELIMINARY RESULTS

2.1. Elliptical Distributions

According to Cambanis, Huang, & Simons (1981) a d-variate continuous random vector $\boldsymbol{X}$, with mean $\boldsymbol{\mu}$, has an elliptical distribution if, and only if, it can be written as

$$\boldsymbol{X} = \boldsymbol{\mu} + R\boldsymbol{\Lambda}\boldsymbol{U}, \tag{1}$$

where $\boldsymbol{U}$ is a d-variate random vector uniformly distributed on the unit hypersphere with $d-1$ topological dimensions $\left\{\boldsymbol{u} \in \mathbb{R}^d : ||\boldsymbol{u}|| = 1\right\}$, R is a non-negative random variable—called a generating variate—stochastically independent of $\boldsymbol{U}$, and $\boldsymbol{\Lambda}$ is a $d \times d$ matrix such that

$$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' = \boldsymbol{\Sigma} = \frac{d}{E\left(R^2\right)} \text{Var}\left(\boldsymbol{X}\right).$$

The density of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, g_R\right) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} g_R\left(\left(\boldsymbol{x}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right), \tag{2}$$

with

$$g_R\left(t\right) = \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{d/2}} \sqrt{t}^{-(d-1)} f_R\left(\sqrt{t}\right), \quad t > 0, \tag{3}$$

where $t = \left(\boldsymbol{x}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)$ and $f_R\left(\cdot\right)$ is the density of R. In (1), $\boldsymbol{U}$ produces elliptically countered surfaces, whereas R determines shape and, in particular, kurtosis of the resulting distribution. It can be proved (Gómez, Gómez-Villegas, & Marín, 2003, Theorem 4 (iv)) that

$$\text{Kurt}(\boldsymbol{X}) = d^2 \frac{E\left(R^4\right)}{\left[E\left(R^2\right)\right]^2}. \tag{4}$$

The factor $E(R^4)/[E(R^2)]^2$, that appears in Equation (4), is the kurtosis of a univariate symmetric random variable Z with mean 0 such that $|Z| = R$ (see Gómez, Gómez-Villegas, & Marín, 2003 for details).

As a summary example of the concepts illustrated in this section consider a random variable R having a χ distribution with d degrees of freedom, in symbols $R \sim \chi\left(d\right)$. In such a case, based on (1), $\boldsymbol{X}$ has a MN distribution, that is, $\boldsymbol{X} \sim N_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. Moreover being $E\left(R^2\right) = d$ and based on (4) it results that

$$\text{Kurt}(\boldsymbol{X}) = E\left(R^4\right) = d\left(d+2\right). \tag{5}$$

This example will be the building block for the genesis of the distribution we propose in Section 3.

2.2. Orthogonal Polynomials in Reshaping Distributions

Zoia (2010) proposed a methodological approach, based on orthogonal polynomials, to reshape an ad hoc parent bell-shaped distribution in order to account for features like skewness and possibly severe kurtosis, which are commonly encountered in real data. In detail specific moments of a parent bell-shaped distribution are modified by using a linear combination of orthogonal polynomials whose coefficients depend on the moments of the parent distribution. The number and the order of the polynomials involved in the linear combination depend on the number of moments to modify of the parent distribution, whereas the parameters of the linear combination depend on the extent to which the moments of the parent distribution must be adapted. In the following we summarize the key aspects of this approach.

Given a density $f(\cdot)$ with finite moments m_j, $j \in \mathbb{N}$ we can determine a system of polynomials $p_n(x) = \sum_{j=0}^{n} a_j x^j$ with the orthogonal property

$$\int_{-\infty}^{\infty} p_n(x) p_m(x) f(x) dx = \int_{-\infty}^{\infty} x^n p_m(x) f(x) dx = \begin{cases} \gamma_n & \text{for } m = n, \\ 0 & \text{for } m \neq n. \end{cases} \tag{6}$$

The condition (6) determines $p_n(\cdot)$, $n \in \mathbb{N}$, up to a constant factor. The parameters a_j, $j = 0, \ldots, n-1$, are given by:

$$a_j = \frac{M_{n+1,j+1}}{M_{n+1,n+1}}, \tag{7}$$

where $M_{n+1,i}$ is the cofactor of the element in position $(n+1, i)$ of the moment matrix

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \cdots & m_n \\ \vdots & \vdots & \vdots & & \vdots \\ m_{n-1} & m_n & m_{n+1} & \cdots & m_{2n-1} \\ 1 & x & x^2 & \cdots & x^n \end{bmatrix}.$$

Besides $p_0(x) = 1$ for all $x \in \mathbb{R}$ and monic polynomials are adopted as it is usually the case.

The trinomial

$$q_n(x; \alpha, \beta) = 1 + \frac{\alpha}{\gamma_{2n-1}} p_{2n-1}(x) + \frac{\beta}{\gamma_{2n}} p_{2n}(x), \tag{8}$$

where α and β are two real numbers, plays the role of a polynomial shape adapter as it can be used to modify the $2n-1$ and $2n$ order moments of $f(\cdot)$ of a quantity equal to α and β, respectively. The Gram–Charlier expansion of $f(\cdot)$, namely

$$\varphi(x; \alpha, \beta) = q_n(x; \alpha, \beta) f(x),$$

where α and β are restricted so that $q_n(\cdot; \alpha, \beta)$ is non-negative, is a density whose moments μ_j, up to the $2n$-th order, can be expressed in terms of the moments m_j of $f(\cdot)$ as follows:

$$\mu_j = \begin{cases} m_j & \text{for } j < 2n-1, \\ m_j + \alpha & \text{for } j = 2n-1, \\ m_j + \beta & \text{for } j = 2n. \end{cases}$$

When $n = 2$ the density $\varphi(\cdot; \alpha, \beta)$ proves able to account for skewness and kurtosis of an amount equal to α and β, respectively, with respect to $f(\cdot)$.

In the aforementioned paper of Zoia (2010) the focus was on the univariate normal distribution whose related orthogonal polynomials are the Hermite ones

$$H_n(x) = (-1)^n e^{\frac{x^2}{2}} \frac{\partial^n e^{-\frac{x^2}{2}}}{\partial x^n} = x^n - \binom{n}{2} x^{n-2} + 3 \binom{n}{4} x^{n-4} + \ldots, \qquad n \in \mathbb{N},$$

whose coefficients, according to (7), can be computed from the (even) moments of the Gaussian law.

If we are interested in modifying only the fourth moment of a normal density $\phi(x; \mu, \sigma^2)$ then the Gram–Charlier expansion (Jondeau & Rockinger, 2001) is

$$\begin{aligned} \varphi(x; \mu, \sigma^2, \beta) &= q_2(x; 0, \beta)\phi(x; \mu, \sigma^2) \\ &= \left[1 + \frac{\beta}{24} H_4\left(\frac{x-\mu}{\sigma}\right)\right] \phi(x; \mu, \sigma^2) \\ &= \left\{1 + \frac{\beta}{24}\left[\left(\frac{x-\mu}{\sigma}\right)^4 - 6\left(\frac{x-\mu}{\sigma}\right)^2 + 3\right]\right\} \phi(x; \mu, \sigma^2), \end{aligned} \tag{9}$$

where β must be lower than 4 to assure the non-negativeness of $\varphi(\cdot; \mu, \sigma^2, \beta)$ and lower than 2.4 when the unimodality of $\varphi(\cdot; \mu, \sigma^2, \beta)$ is required (see Zoia, 2010).

The same Gram–Charlier-like expansion could be obtained by reshaping a χ_1^2 with its related second order orthogonal polynomial, upon noting that the normal distribution is an elliptical distribution with a χ_1 distribution as generating variate. This latter approach will be developed in the next section to get the Gram–Charlier-like expansion of a MN law.

3. THE MLN DISTRIBUTION

In this section we illustrate two of the main proposals of the present work: the MLN distribution and the way the MLN distribution is generated by reshaping the MN distribution. A similar approach can be applied to reshape any multivariate elliptical distribution in order to obtain a more general elliptical distribution with the desired amount of excess kurtosis. Note that a first attempt to use orthogonal polynomials in the multivariate context has been made in Faliva, Zoia, & Poti (2016). Nevertheless the authors define multivariate distributions in the time series context and they use a pairwise approach: multivariate dependencies are defined on successive pairs of variables only (see formula (4.24) of the cited paper).

3.1. The Model

We start by introducing the following definition.

Definition 1 *A d-variate random vector $\widetilde{\boldsymbol{X}}$ follows a leptokurtic normal distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and excess kurtosis β, in symbols $\widetilde{\boldsymbol{X}} \sim LN_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta)$, if its density is given by*

$$f_{\widetilde{\boldsymbol{X}}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta) = q(t; \beta)\, \phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{for } \boldsymbol{x} \in \mathbb{R}^d, \tag{10}$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of a d-variate normal random vector $\boldsymbol{X}$ with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and $q(t; \beta)$ is defined as follows:

$$q(t; \beta) = 1 + \frac{\beta}{8d(d+2)}\left[t^2 - 2(d+2)t + d(d+2)\right]; \quad t = (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{11}$$

The excess kurtosis β must satisfy the constraint $\beta \in \left[0, \min\left(4d, 4d(d+2)/5\right)\right]$ so that (10) is a unimodal elliptical distribution.

The kurtosis of $\widetilde{\boldsymbol{X}} \sim LN_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta\right)$ is

$$\operatorname{Kurt}(\widetilde{\boldsymbol{X}}) = \operatorname{Kurt}(\boldsymbol{X}) + \beta = d(d+2) + \beta, \tag{12}$$

where, advantageously, β directly represents the excess kurtosis.

The following theorem explains the genesis of the MLN distribution.

Theorem 1. *Let $\boldsymbol{X}$ be a d-variate normal random vector $\boldsymbol{X} \sim N_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ with stochastic representation $\boldsymbol{X} = \boldsymbol{\mu} + R\boldsymbol{\Lambda}\boldsymbol{U}$. Then the random vector $\widetilde{\boldsymbol{X}}$ with stochastic representation*

$$\widetilde{\boldsymbol{X}} = \boldsymbol{\mu} + \widetilde{R}\boldsymbol{\Lambda}\boldsymbol{U}$$

follows a multivariate leptokurtic normal distribution, that is,

$$\widetilde{\boldsymbol{X}} \sim LN_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta\right),$$

where $\widetilde{R}^2$ is obtained by reshaping the density $f_{R^2}(r)$ of R^2 with the second-order polynomial

$$q\left(r; \beta\right) = 1 + \frac{\beta}{8d(d+2)}\left[r^2 - 2(d+2)r + d(d+2)\right].$$

Proof. According to (5) in order to increase the kurtosis of $\boldsymbol{X} \sim N_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ we should modify the fourth moment of $R \sim \chi\left(d\right)$, namely the second moment of $R^2 \sim \chi^2\left(d\right)$. Hence the density $f_{R^2}\left(\cdot\right)$ of R^2 can be reshaped as

$$f_{\widetilde{R}^2}(r; \beta) = q(r; \beta) f_{R^2}\left(r\right), \quad r \geq 0, \tag{13}$$

where $q\left(\cdot; \beta\right)$ is a second order polynomial specified as

$$q\left(r; \beta\right) = 1 + \frac{\beta}{\gamma_2} p_2(r) = 1 + \frac{\beta}{\gamma_2}(r^2 + \vartheta_1 r + \vartheta_0), \tag{14}$$

with ϑ_0 and ϑ_1 being parameters. Given the moments

$$m_j = 2^j \frac{\Gamma\left(j + \frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}, \quad j \in \mathbb{N},$$

of $\chi^2(d)$, Equation (7) can be used to determine ϑ_0 and ϑ_1 and the resulting polynomial $p_2(\cdot)$ turns out to be

$$p_2(r) = r^2 - 2(d+2)r + d(d+2), \quad r \geq 0. \tag{15}$$

Note that $p_2(\cdot)$ reduces to the fourth order Hermite polynomial when $d = 1$. Further algebra yields the normalization factor

$$\gamma_2 = \int r^2 p_2(r) f_{R^2}(r) dr = 8d(d+2).$$

Accordingly the polynomial shape modifier $q(\cdot;\beta)$ in (14), suitable to augment the second order moment of R^2 of a quantity β, is given by

$$q(r;\beta) = 1 + \frac{\beta}{8d(d+2)}\left[r^2 - 2(d+2)r + d(d+2)\right], \qquad r \geq 0. \tag{16}$$

Some computations prove that the second moment of $\widetilde{R}^2$ is

$$\mathrm{E}\left[\left(\widetilde{R}^2\right)^2\right] = \int r^2 f_{\widetilde{R}^2}(r)dr = d(d+2) + \beta. \tag{17}$$

This, together with (5), proves (12). The density of the generating variate $\widetilde{R}$ of $\widetilde{\boldsymbol{X}}$ can be obtained from $f_{R^2}(\cdot;\beta)$ as follows:

$$\begin{aligned} f_{\widetilde{R}}(r;\beta) &= 2rf_{\widetilde{R}^2}(r^2;\beta) \\ &= 2rf_{R^2}(r^2)q(r^2;\beta) \\ &= f_R(r)q(r^2;\beta), \qquad r \geq 0, \end{aligned} \tag{18}$$

where $f_R(\cdot)$ is the density of R. Now by replacing (18) in (3) and considering (2) it results that

$$\begin{aligned} f_{\widetilde{\boldsymbol{X}}}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\beta) &= |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{d/2}}\left[(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{d-1}{2}} \\ &\quad \times f_{\widetilde{R}}\left(\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})};\beta\right) \\ &= (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &\quad \times q\left((\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu});\beta\right) \\ &= \phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\,q\left((\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu});\beta\right), \qquad \boldsymbol{x}\in\mathbb{R}^d, \end{aligned}$$

which is the density of $\widetilde{\boldsymbol{X}} \sim LN_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$. ■

The following corollary proves that the parameter β must satisfy some constraints in order for $f_{\widetilde{\boldsymbol{X}}}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$ to be positive and unimodal.

Corollary 1 *Let $\widetilde{\boldsymbol{X}} \sim LN_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$. The parameter β, representing the excess kurtosis with respect to a parent standard multinormal distribution, must satisfy the following constraint*

$$\beta \in \left[0, \min\left(4d, 4d(d+2)/5\right)\right], \tag{19}$$

so that $\widetilde{\boldsymbol{X}}$ has a positive and unimodal multivariate distribution.

Proof. The constraint $\beta \in [0, \min(4d, 4d(d+2)/5)]$ is the intersection of two constraints:

(i) $\beta \in [0, 4d]$, which assures that $f_{\widetilde{\boldsymbol{X}}}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$ is a positive elliptical density;
(ii) $\beta \in [0, 4d(d+2)/5]$, which guarantees that $f_{\widetilde{\boldsymbol{X}}}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$ is unimodal.

Some considerations about items (i) and (ii) are given below.

(i) From the stochastic representation in Theorem 1 $\widetilde{\boldsymbol{X}}$ is elliptically distributed if $\widetilde{R}$ is a generating variate (i.e., it is a positive random variable). This happens when the density in (18) is well-defined, that is, when the polynomial shape modifier $q(\cdot;\beta)$ in (16) is positive. Now upon noting that $q(r;\beta)$ is limited from below and that its minimum occurs at $r = d+2$ simple computations lead to the conclusion that $f_{\widetilde{R}}(\cdot;\beta)$ is always positive when β is lower or equal to $4d$.

(ii) As far as unimodality is concerned, according to Gómez, Gómez-Villegas, & Marín (2003, Theorem 7, point (iv)), all the d variates of $\widetilde{\boldsymbol{X}}$ have excess kurtosis β_j, $j = 1,\ldots,d$ equal to $3\beta/[d(d+2)]$. Now in the univariate context the unimodality is guaranteed when the excess kurtosis parameter is equal or lower than 2.4 (Zoia, 2010). Assuming unimodality of all the marginals, or analogously imposing constraint (ii), means assuring the unimodality of $f_{\widetilde{\boldsymbol{X}}}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$. Note that, when $d > 3$, constraint (ii) is satisfied by constraint (i) because the latter turns out to be more stringent than the former.

■

3.2. Maximum Likelihood Estimation

Several estimators of the parameters of the MLN distribution may be considered. Among them maximum likelihood (ML) estimators are most attractive because of their large sample or asymptotic properties.

Given a random sample $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ from $\boldsymbol{X} \sim LN_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$ the ML estimation method is based on the maximization of the log-likelihood function, say $l(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$. If $\beta = 0$ then the log-likelihood function of the MN distribution is obtained. It is possible to show that $l(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$ is not concave. Nevertheless with the same arguments in Zhang & Liang (2010) we can conclude that $l(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$ has a global maximum as it is continuous. Details about the maximization of $l(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$ are given in Appendix D whereas the performance of the algorithm we use to maximize is evaluated via simulations in Appendix F.

4. MIXTURES OF MLN DISTRIBUTIONS

In this section we propose mixtures of MLN distributions as an application of the proposed model. Our mixtures constitute an alternative to other mixtures of multivariate elliptically countered distributions.

4.1. The Model

For a d-variate random vector $\boldsymbol{X}$ a finite mixture of MLN distributions can be written as

$$p(\boldsymbol{x};\boldsymbol{\vartheta}) = \sum_{j=1}^{k} \pi_j f\left(\boldsymbol{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j,\beta_j\right), \tag{20}$$

where π_j is the mixing proportion of the jth component, with $\pi_j > 0$ and $\sum_{j=1}^{k}\pi_j = 1$, f is defined as in (10), and $\boldsymbol{\vartheta}$ contains all the parameters of the mixture. As a special case when $\beta_j = 0$ for each $j = 1,\ldots,k$ we obtain classical mixtures of MN distributions.

4.2. An EM Algorithm for Maximum Likelihood Estimation

Let $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ be a random sample from model (20). To find ML estimates for the parameters of this model, we adopt the classical EM algorithm (Dempster, Laird, & Rubin, 1977), which is a natural approach for ML estimation when data are incomplete. In our case, the source of incompleteness, the classical one in the use of mixture models arises from the fact that for each observation we do not know its component membership; this source is governed by an indicator vector $z_i = (z_{i1}, \ldots, z_{ik})'$, where $z_{i1} = 1$ if $\boldsymbol{x}_i$ comes from component j and $z_{ij} = 0$ otherwise. The values of z_{ij} are used for the definition of the following complete-data log-likelihood

$$l_c(\boldsymbol{\vartheta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \ln\left(\pi_j\right) + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \ln\left[f\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \beta_j\right)\right]. \tag{21}$$

The EM algorithm iterates between two steps, one E-step and one M-step, until convergence.

The E-step on the $(q+1)$th iteration requires the calculation of

$$E_{\boldsymbol{\vartheta}^{(q)}}\left[Z_{ij}\,|\boldsymbol{x}_i\right] = z_{ij}^{(q)} = \frac{\pi_j^{(q)} f\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j^{(q)}, \boldsymbol{\Sigma}_j^{(q)}, \beta_j^{(q)}\right)}{p\left(\boldsymbol{x}_i; \boldsymbol{\vartheta}^{(q)}\right)},$$

which corresponds to the posterior probability that the unlabelled observation $\boldsymbol{x}_i$ belongs to the jth component of the mixture, using the current fit $\boldsymbol{\vartheta}^{(q)}$ for $\boldsymbol{\vartheta}$. Then by substituting z_{ij} with $z_{ij}^{(q)}$ in (21) we obtain the conditional expectation of the complete-data log-likelihood, say $Q(\boldsymbol{\vartheta}) = Q_1(\boldsymbol{\pi}) + Q_2(\boldsymbol{\psi})$, where the two terms on the right-hand side are ordered as the two terms on the right-hand side of (21), being $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ and $\boldsymbol{\psi} = \boldsymbol{\vartheta} \setminus \boldsymbol{\pi}$.

The M-step on the $(q+1)$th iteration of the EM algorithm requires the calculation of $\boldsymbol{\vartheta}^{(q+1)}$ as the value of $\boldsymbol{\vartheta}$ that maximizes $Q(\boldsymbol{\vartheta})$. As $Q_1(\boldsymbol{\pi})$ and $Q_2(\boldsymbol{\psi})$ have zero cross-derivatives they can be maximized separately. Maximizing $Q_1(\boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$, subject to the constraints on those parameters, yields

$$\pi_j^{(q+1)} = \sum_{i=1}^{n} z_{ij}^{(q)} \Big/ n \quad \text{for } j = 1, \ldots, k.$$

Maximizing $Q_2(\boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$ is equivalent to independently maximizing each of the k weighted log-likelihood functions

$$Q_{2j}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \beta_j\right) = \sum_{i=1}^{n} z_{ij}^{(q)} \ln\left[f\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \beta_j\right)\right] \tag{22}$$

with respect to $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$, and β_j, $j = 1, \ldots, k$. Details about the maximization of Q_{2j} are given in Appendix D.

5. RESULTS

In this section we investigate the behaviour of the proposed mixture model through benchmark real data. We further provide a comparison with (unconstrained) finite mixtures of some well-established multivariate elliptically contoured distributions. In detail we compare:

(i) mixtures of MN distributions (abbreviated by MNMs hereafter),
(ii) mixtures of multivariate t distributions (MtMs) as proposed by Peel & McLachlan (2000),
(iii) mixtures of multivariate contaminated normal distributions (MCNMs) as proposed by Punzo & McNicholas (2016),
(iv) mixtures of multivariate power exponential distributions (MPEMs) as proposed by Zhang & Liang (2010),
(v) mixtures of multivariate leptokurtic-normal distributions (MLNMs) as given in (20).

Apart from MNMs each mixture component has an additional parameter, say β_j for uniformity with MLNMs, governing the tail weight. For M*t*Ms $\beta_j > 0$ coincides with the degrees of freedom: lower values of β_j are related to heavier tails. However the kurtosis which is $d(d+2)+6/\left(\beta_j-4\right)$ is only defined for $\beta_j > 4$ and can assume any value in the interval $(d(d+2), \infty)$, with the extreme values being assumed for $\beta_j \rightarrow \infty$ and $\beta_j \rightarrow 4^+$, respectively (see, e.g., Zografos, 2008). For MCNMs $\beta_j > 1$ is defined as one of the two parameters governing the tail weight; it is the inflation parameter denoting the degree of outlierness with higher values related to heavier tails (see Maruotti & Punzo, 1998; Punzo & Maruotti, 2016; Punzo & McNicholas, 2016 for details). Also in this case the kurtosis which is given in Equation (G.11) can assume any value in the interval $(d(d+2), \infty)$; details are given in Appendix G. For MPEMs, $\beta_j > 0$ is a shape parameter; heavy tails are obtained for $\beta_j < 1$, light tails are obtained for $\beta_j > 1$, whereas normal tails are obtained for $\beta_j = 1$. The kurtosis is

$$\frac{d\,\Gamma\left(\frac{d}{2\beta_j}\right)\Gamma\left(\frac{d+4}{2\beta_j}\right)}{\Gamma^2\left(\frac{d+2}{2\beta_j}\right)},$$

and can assume any value in $\left(\frac{d(d+2)^2}{d+4}, \infty\right)$, with the extreme values being assumed for $\beta_j \rightarrow \infty$ (d-variate uniform distribution) and $\beta_j \rightarrow 0^+$, respectively (see Gómez, Gómez-Villegas, & Marin, 1988 for details).

MNMs are fitted via the `gpcm()` function of the R-package **mixture** (Browne, ElSherbiny, & McNicholas, 2015), M*t*Ms are fitted via the `teigen()` function of the R-package **teigen** (Andrews & McNicholas, 2016), MCNMs are fitted via the `CNmixt()` function of the R-package **ContaminatedMixt** (Punzo, Mazza, & McNicholas, 2015), MPEMs are fitted via the `mpe()` function of an R package available at http://onlinelibrary.wiley.com/doi/10.1111/biom.12351/suppinfo (Dang, Browne, & McNicholas, 2015), whereas a specific R code, available as Supplementary Material for Review, has been implemented to fit MLNMs. To allow for a direct comparison of the competing models, all the algorithms are initialized by providing the initial quantities $z_i^{(0)}$, $i = 1, \ldots, n$: nine times using a random initialization and once with a k-means initialization (as implemented by the `kmeans()` function for R). The solution maximizing the observed-data log-likelihood among these 10 runs is then selected; see Dang et al. (2016). For alternative initialization strategies, see Bagnato & Punzo (2013).

We use the Bayesian information criterion (BIC; Schwarz, 1978) for model selection (Appendix B), a stopping criterion based on the Aitken acceleration (Aitken, 1926) to determine convergence (Appendix A), and the adjusted Rand index (ARI; Hubert & Arabie, 1985) for clustering/classification assessment (Appendix C).

5.1. Assessing the Impact of Background Noise

A sensitivity study is here described to compare how background noise can affect the classification performance of the competing mixture models. This study is based on the `students` data set introduced by Ingrassia, Minotti, & Punzo (2014) and available in the **flexCWM** package for R (Mazza, Punzo, & Ingrassia, 2015). The data come from a survey of $n = 270$ students attending a statistics course at the Department of Economics and Business of the University of Catania in the academic year 2011/2012. Although the questionnaire included seven items, the following analysis only concerns, for illustrative purposes, the variables HEIGHT (height of the respondent, measured in centimeters) and HEIGHT.F (height of respondent's father, measured in centimeters). Moreover there are $k = 2$ groups of respondents with respect to gender: 119 males and 151 females. On these data we fit MNMs, M*t*Ms, MCNMs, MPEMs, and MLNMs, for $k \in \{1, 2, 3\}$. Table 1 shows the obtained results. For all the considered models $k = 2$ components are selected by

TABLE 1: `students` data. Number of components, BIC, and ARI for the model selected by BIC

	k	BIC	ARI
MNM	2	−3,601.953	0.913
MtM	2	−3,614.346	0.913
MCNM	2	−3,624.348	0.913
MPEM	2	−3,612.447	0.899
MLNM	2	−3,613.150	0.913

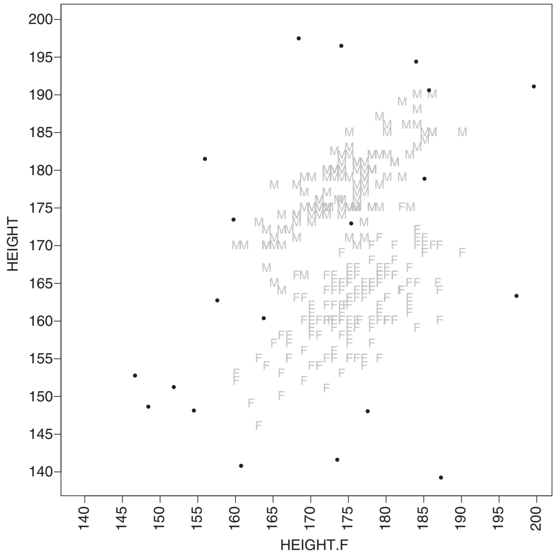

FIGURE 1: `students` data with noise. Example of scatter plot (M denotes male and F female). Background uniform noise points are denoted by •.

the BIC. The best model according to the BIC is the MNM, whereas the worst is the MCNM. According to these results the data do not seem to need components with heavy tails and, as such, the MCNM is the most penalized approach being the least parsimonious. In terms of classification performance, apart from the selected MPEM, the models provide an ARI value of 0.913 denoting a good behaviour.

Now we modify the original data by including 20 noisy points generated from a uniform distribution over a square centered on the bivariate mean (174.963, 168.652) of the observations and with side of length 60 (centimeters). This square contains the original data. To make the conclusions about this study more stable, we repeat this generation one hundred times, so obtaining one hundred different data sets with noise. Figure 1 shows an example of a modified data set with bullets denoting uniform noise points. On each replication, MNMs, MtMs, MCNMs, MPEMs, and MLNMs, are fitted for $k \in \{1, 2, 3\}$.

Table 2 shows the obtained results. The first part of this table displays the number of times each value of k is selected by the BIC. The selected MNM is almost always (95 times out of 100) characterized by three components; by looking at the corresponding ARI value the additional third component seems to be attempting to model the background noise. Apart from a sporadic case where a single component ($k = 1$) is selected for the MtM and the MPEM the selected

TABLE 2: `students` data with noise. Number of times each value of $k \in \{1, 2, 3\}$ is selected by the BIC, and average values of BIC and ARI, over 100 replications of the noisy points

	Best value of k			Averages values	
	1	2	3	BIC	ARI
MNM	0	5	95	−4,064.826	0.905
M*t*M	1	99	0	−4,060.367	0.895
MCNM	0	100	0	−4,061.456	0.895
MPEM	1	99	0	−4,076.971	0.885
MLNM	0	100	0	−4,060.044	0.913

TABLE 3: `geyser2` data set. Best BIC values, and associated value of k, for the fitted mixtures

	MNM	M*t*M	MCN	MPEM	MLNM
k	5	4	3	3	4
BIC	−1,113.080	−1,118.659	−1,139.531	−1,145.911	−1,115.995

remaining models have two components. The rest of Table 2 shows the averages values, over the 100 replications, of BIC and ARI values. Note that each ARI value evaluates the agreement between the predicted classification from the model selected via BIC with respect to the true group labels; this is done with respect to the original observations only without considering the added noisy points. The best average BIC (−4,060.044), as well as the best classification performance (ARI $= 0.921$), are related to the MLNM. From the opposite side the worst performer, both in terms of average BIC and ARI values, is the MPEM (BIC $= -4{,}076.971$ and ARI $= 0.885$). Interestingly by comparing the average ARI values in Table 2 with the ARI values in Table 1, we can note the classification from the MLNM is the only one to be not affected (in average) by the added noisy points.

5.2. Real Data: Old Faithful Geyser

This analysis considers the Old Faithful Geyser data set, which contains $n = 272$ observations of eruption length (see, e.g., Azzalini & Bowman, 1990). In line with (García-Escudero, Gordaliza, & Matrán (2003) and Fritz, García-Escudero, & Mayo-Iscar (2012), a bivariate data set can be constructed considering the eruption lengths and the corresponding previous eruption lengths (see Figure 2). This data set, named `geyser2`, accompanies the **tclust** package (Fritz, García-Escudero, & Mayo-Iscar, 2012) for R.

From Figure 2 it is possible to note the presence of three main clusters plus six "short followed by short" eruptions (data points in the bottom-left corner). As well-motivated by García-Escudero et al. (2008) in situations like these it is not obvious if a small group of tightly joined outliers should be considered as a proper cluster instead of a contamination phenomenon, and this issue is largely debated in the literature. Instead of giving personal considerations about the topic we limit ourselves to evaluate the behaviour of MNMs, M*t*Ms, MCNMs, MPEMs, and MLNMs, on these data; the models are fitted for $k \in \{1, 2, 3, 4, 5, 6\}$.

Table 3 compares the best BIC value, and the associated value of k, for each of the competing models. The best model is the MNM with $k = 5$ components, whereas the worst is the MPEM

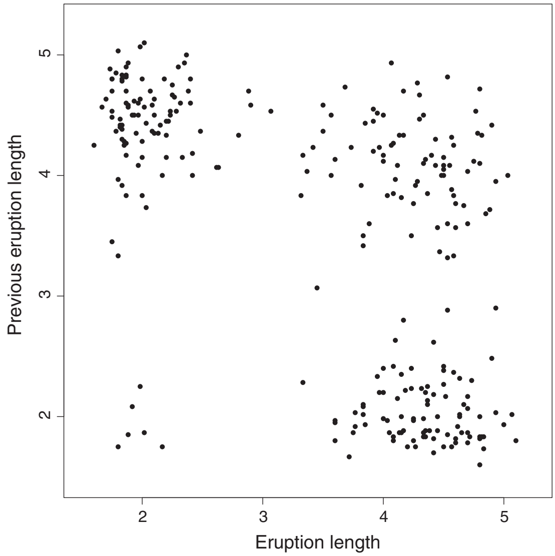

FIGURE 2: `geyser2` data set. Scatter plot.

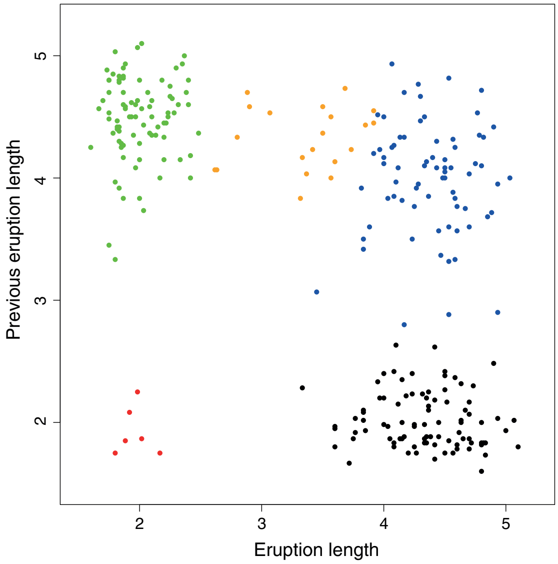

FIGURE 3: `geyser2` data set. Clustering results from the MNM selected by the BIC ($k = 5$).

with $k = 3$ components. However the clustering provided by the former model (see Figure 3) is not as expected: the orange group seems to be composed by two well-separated subgroups, one of them being very overlapped with the blue group. Motivated by these results we look for a different model. The second best MNM, having BIC $= -1{,}128.001$, has $k = 3$ components; compared with the BIC values in Table 3, this MNM is no more the best one. So the overall second best model is the MLNM with $k = 4$ components (see Figure 4b for the obtained clustering). For the selected MLNM, the estimates of the excess kurtosis for the four components are $\widehat{\beta}_1 = 9.768 \cdot 10^{-8}$ (refer to the black bullets in Figure 4b), $\widehat{\beta}_2 = 1.282 \cdot 10^{-6}$ (red bullets), $\widehat{\beta}_3 = 2.339$ (green bullets), and $\widehat{\beta}_4 = 0.972$ (blue bullets); therefore it seems that two of the obtained clusters need heavier

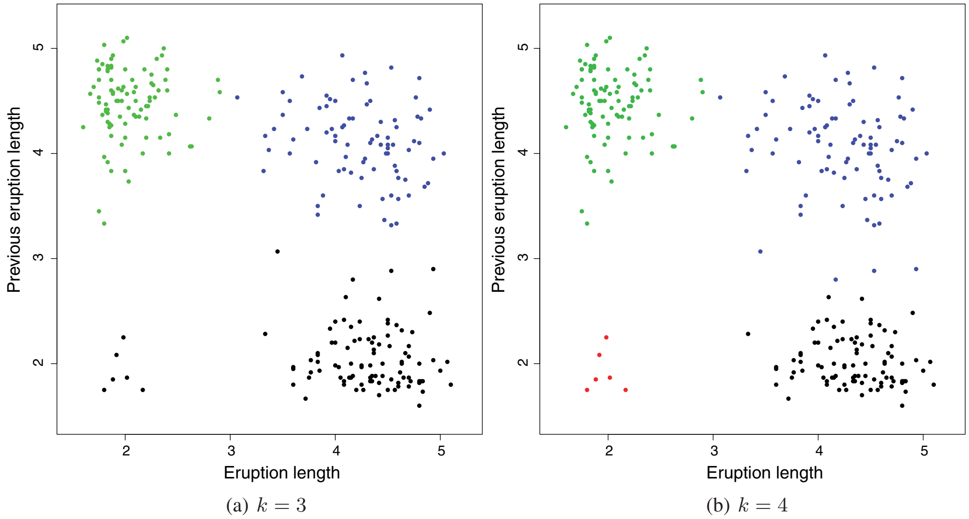

FIGURE 4: geyser2 data set. Clustering results for some MLNMs.

tails than the normal ones. For completeness Figure 4a displays the clustering results obtained for the MLNM with $k = 3$ components (BIC $= -1, 119.837$). As we can note by the green bullets in Figure 4a the small cluster on the bottom-left corner is captured by the tail of the MLN distribution located on the cluster on the bottom-right corner; this is confirmed by the estimated excess kurtosis of such a component which is almost 6.4, which is the maximum excess kurtosis the MLN distribution can reach in the bivariate case.

5.3. Real Data: Australian Institute of Sports

We also evaluate the performance of MLNMs on the ais (Australian Institute of Sports) data set (Cook & Weisberg, 1994), a real benchmark data set which is often used (in some or all of its variables) for illustration in the model-based clustering literature (see, e.g., Galimberti and Soffritti, 2014; Morris et al., 2014; Murray, Browne, & McNicholas, 2014; Tortora et al., 2015; Azzalini et al., 2016; Dang et al., 2016). The data set contains measurements on $n = 202$ athletes, subdivided in $k = 2$ groups (100 female and 102 male), and is available in the R-packages **alr3** (Weisberg, 2011) and **sn** (Azzalini, 2016). A subset of $d = 6$ variables is analyzed here for illustrative purposes: height in centimeters (Ht), weight in kilograms (Wt), lean body mass (LBM), red cell count (RCC), white cell count (WCC), and Hematocrit (Hc). The scatter plot of the data, with labelling based on the two groups, is shown in Figure 5.

On these data MNMs, MtMs, MCNMs, MPEMs, and MLNMs are fitted for $k \in \{1, 2, 3\}$. Table 4 compares the clustering performance of the five mixture models fitted on these data; here the predicted classifications from the selected model (using the BIC) are compared to the true class labels in each case.

All of the selected models recognize the presence of two groups. The best (greatest) BIC value is obtained for the MLNM (BIC $= -5{,}857.493$), whereas the (worst) BIC value is obtained for the MNM (BIC $= -5885.869$); hence the MLNM with two components is the best, in terms of BIC, among the fitted models. The best ARI value is obtained for the MLNM too (ARI $= 0.811$), and it corresponds to ten misclassified observations, whereas the worst is obtained for the MNM (ARI $= 0.724$) and it corresponds to fifteen misclassifications. This latter result could motivate the need for leptokurtic mixture components, and this conjecture is also corroborated by the estimated values of β_1 and β_2 for the MLNM which are 17.127 and 0.404, respectively.

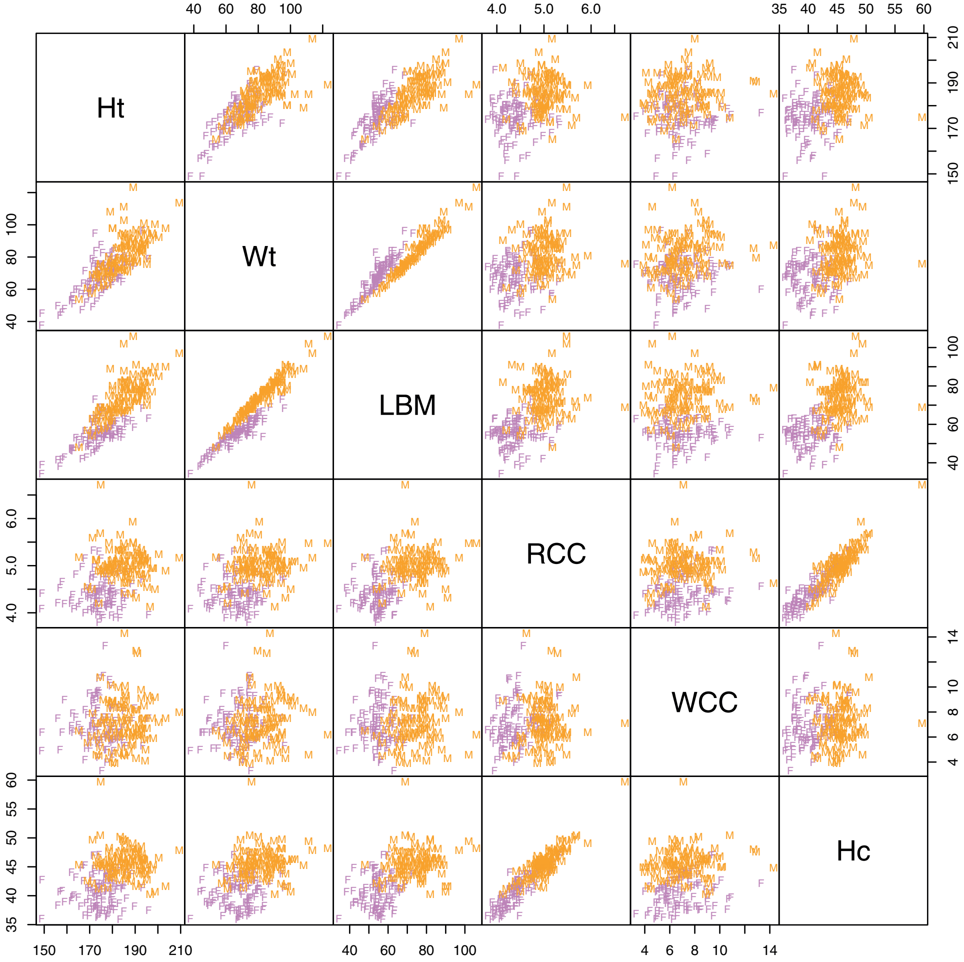

FIGURE 5: ais data. Scatter plot (M denotes male and F female).

TABLE 4: ais data set. For each selected mixture model, the BIC, the number of components, and the ARI are given

Model	k	BIC	ARI
MNM	2	−5,885.869	0.724
MtM	2	−5,860.737	0.775
MCNM	2	−5,864.429	0.758
MPEM	2	−5,865.062	0.775
MLNM	2	−5,857.493	0.811

6. CONCLUSIONS

In this article we have proposed the MLN distribution which is a multivariate Gram–Charlier expansion of the normal law. The MLN distribution is obtained from the elliptical representation of the MN distribution by reshaping its generating variate with its related orthogonal polynomials.

The MLN distribution has a closed form representation and depends on one additional parameter which represents the excess kurtosis.

It is worth noticing that our approach to obtain the MLN distribution can be easily extended to any other elliptical distribution for modifying some moments of interest. In fact being the generating variate a scalar variable also in a multivariate context it can be easily reshaped to get a polynomially modified distribution. This latter turns out to be characterized by some additional parameters whose number is equal to the number of the moments to change for the parent distribution. When the focus is on kurtosis (namely on the fourth moment), and the parent distribution is the normal law, this approach leads to the MLN distribution.

Maximum likelihood estimation has been illustrated for the parameters of the MLN distribution. Furthermore an application of the MLN distribution to robust clustering has been provided by introducing mixtures of MLN distributions; an EM algorithm has been also described to obtain maximum likelihood estimates for the mixture parameters. The analyses of Section 5 on real data have shown how our mixture represents a possible alternative to other existing mixtures of multivariate elliptically countered distributions.

Future work will focus on the following avenues.

- Ad hoc conditions for the identifiability of finite mixtures of MLN distributions need to be determined. In these terms, note that the sufficient conditions given in Holzmann, Munk, & Gneiting (2006), for the identifiability of finite mixtures of elliptical distributions, do not apply in our context.
- Our mixture model implies elliptically contoured distributions for each cluster which, under specific empirical settings, could be rather restrictive. This is justified by the fact that non-elliptical distributions can be approximated quite well by a mixture of several basic elliptical distributions like the normal one (Titterington, Smith, & Makov, 1985, p. 24 and McLachlan & Peel, 2000, p. 1). As this can be very helpful for modelling purposes it can be misleading when dealing with clustering/classification applications as one cluster may be represented by more than one mixture component just because it has, in fact, a non-elliptical distribution. A first possible route to continue to use our approach also in the presence of conditional non-elliptical distributions for each cluster, consists in considering transformations so as to make the components as elliptical as possible (Schork & Schork, 1988; Zhu & Melnykov, 2016). Although such a treatment is very convenient to use the achievement of joint ellipticality is rarely satisfied and the transformed variables become more difficult to be interpreted. Instead of applying transformations we could extend our MLN distribution, via the method of Zoia (2010), with the aim of jointly modifying kurtosis and skewness of the MN distribution; the resulting model could be used to define a mixture having the obtained distributions as components. Examples of competing approaches in this directions are: mixtures of multivariate skew-*t* distributions (see, e.g., Lin, 2010; Lee & McLachlan, 2014), mixtures of multivariate *t* distributions with the Box–Cox transformation (Lo & Gottardo, 2012), and mixtures of generalized hyperbolic distributions (Browne & McNicholas, 2015).
- In the fashion of Banfield & Raftery (1993) and Celeux & Govaert (1995) for mixtures of MN distributions, Andrews & McNicholas (2012) for mixtures of multivariate *t* distributions, Punzo & McNicholas (2016) for MCNMs, and Dang, Browne, & McNicholas (2015) for MPEMs, mixtures of MLN distributions could be made more flexible and parsimonious by imposing constraints on the eigen-decomposed component matrices $\boldsymbol{\Sigma}_j$, $j = 1, \ldots, k$. However the resulting family of parsimonious models will present a difficult parameter estimation problem because none of the parameter estimates will be available in closed form, and such a problem may be exacerbated for constrained models. An analogous problem is shared by MPEMs (Dang, Browne, & McNicholas, 2015). To deal with this issue borrowing the idea applied to MPEMs by Dang, Browne, & McNicholas (2015) we could make use of MM algorithms

(Hunter & Lange, 2000) and accelerated line search algorithms on the Stiefel manifold (Absil, Mahony, & Sepulchre, 2009 and Browne & McNicholas, 2014a,b). This should allow for the estimation of a wide range of constrained models.

- In the fashion of Ghahramani & Hinton (1997), McLachlan & Peel (2000, Chapter 8), McLachlan, Peel, & Bean (2003), and McNicholas & Murphy (2008) for mixtures of MN distributions; McLachlan, Bean, & Ben-Tovim Jones (2007) and Andrews & McNicholas (2011) for mixtures of multivariate t distributions; and Punzo & McNicholas (2014) for MCNMs, parsimony, but also dimension reduction, could be obtained by exploiting local factor analyzers.

APPENDIX

A - CONVERGENCE CRITERION

A stopping criterion based on Aitken's acceleration (Aitken, 1926) is used to determine convergence of the EM algorithm illustrated in Section 4.2. The commonly used stopping rules can yield convergence earlier than the Aitken stopping criterion, resulting in estimates that might not be close to the ML estimates. The Aitken acceleration at iteration q is

$$a^{(q)} = \frac{l^{\text{new}} - l^{(q)}}{l^{(q)} - l^{(q-1)}},$$

where $l^{(q)}$ is the (observed-data) log-likelihood value from iteration q. An asymptotic—with respect to the iteration number—estimate of the log-likelihood at iteration $q+1$ can be computed via

$$l_A^{\text{new}} = l^{(q)} + \frac{1}{1 - a^{(q)}} \left(l^{\text{new}} - l^{(q)} \right);$$

cf. Böhning et al. (1994). Convergence is assumed to have been reached when $l_A^{\text{new}} - l^{(q)} < \epsilon$, provided that this difference is positive (cf. Lindsay, 1995, McNicholas et al., 2010; and Subedi et al., 2013, 2015). We use $\epsilon = 0.005$ in the analyses of Section 5.

B - MODEL SELECTION

In model-based clustering applications it is common to fit a mixture model for a range of values of the number of components k. After that the "best" value for k is chosen based on some likelihood-based criterion. Note that such a choice does not necessarily correspond to optimal clustering; for the alternative use of likelihood-ratio tests, see Punzo, Browne, & McNicholas (2016).

The Bayesian information criterion (BIC; Schwarz, 1978) is commonly used for mixture model selection. It is intended to provide a measure of the weight of evidence favouring one model over another, or Bayes factor (Weakliem, 1999). Even though the regularity properties needed for the development of the BIC are not satisfied by mixture models (Keribin, 1998, 2000), it has been used extensively (see, e.g., Dasgupta & Raftery, 1998; Fraley & Raftery, 2002) and performs well in practice. The BIC can be computed as

$$\text{BIC} = 2l(\widehat{\boldsymbol{\vartheta}}) - m \ln n,$$

where $l(\widehat{\boldsymbol{\vartheta}})$ is the maximized (observed-data) log-likelihood, m is the number of free parameters, and n is the sample size. Note that Bayes factors can be used to compare models that are not nested, and the BIC approximation thereto holds when models are not nested (cf. Raftery, 1995).

C - PERFORMANCE ASSESSMENT

The adjusted Rand index (ARI; Hubert & Arabie, 1985) is the method used for determining the performance of the chosen model by comparing predicted classifications to true group labels, when known. The ARI corrects the Rand index (Rand, 1971) to account for chance when calculating the agreement between true labels and estimated classifications. An ARI of 1 corresponds to perfect agreement, and the expected value of the ARI is 0 under random classification. Steinley (2004) provides a thorough evaluation of the ARI.

D - MAXIMIZATION OF THE WEIGHTED LOG-LIKELIHOOD FUNCTION

The weighted log-likelihood function related to the MLN distribution $f\left(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\beta\right)$ is

$$\widetilde{l}\left(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta\right)=\sum_{i=1}^{n} w_i \ln\left[f\left(\boldsymbol{x}_i;\boldsymbol{\mu},\boldsymbol{\Sigma},\beta\right)\right], \tag{D.1}$$

where $w_i \geq 0$. If $w_i = 1$, $i = 1,\ldots,n$, then (D.1) coincides with $l\left(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta\right)$.

Maximizing (D.1) with respect to $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and β, is a constrained problem due to $\boldsymbol{\Sigma}$ and β. To make the maximization problem unconstrained, we apply the following reparameterization for $\boldsymbol{\Sigma}$ and β. According to the Cholesky decomposition we write

$$\boldsymbol{\Sigma}=\boldsymbol{\Omega}'\boldsymbol{\Omega},$$

where $\boldsymbol{\Omega}$ is an upper triangular matrix. As concerns β we write

$$\beta=\beta_{\max}\frac{\exp(\gamma)}{1+\exp(\gamma)},$$

where $\beta_{\max}=\min\left(4d, 4d(d+2)/5\right)$ is the maximum value for β as defined in Corollary 1.

According to the above parametrization the function in (D.1) can be re-written as

$$\begin{aligned}\widetilde{l}\left(\boldsymbol{\mu},\boldsymbol{\Omega},\gamma\right)=&-\frac{1}{2}\sum_{i=1}^{n} w_i \ln\left|\boldsymbol{\Omega}'\boldsymbol{\Omega}\right|-\frac{d\widetilde{w}}{2}\ln\left(2\pi\right)-\frac{1}{2}\sum_{i=1}^{n} w_i\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)\\&+\sum_{i=1}^{n} w_i\ln\left\{1+\frac{\beta_{\max}}{8d(d+2)}\frac{\exp(\gamma)}{1+\exp(\gamma)}\left[\left(\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)\right)^2\right.\right.\\&\left.\left.-2(d+2)\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)+d(d+2)\right]\right\},\end{aligned} \tag{D.2}$$

where $\widetilde{w}=\sum_{i=1}^{n} w_i$. By applying vector derivatives of (D.2) with respect to $\boldsymbol{\mu}$ we obtain the equation

$$\frac{\partial\widetilde{l}\left(\boldsymbol{\mu},\boldsymbol{\Omega},\gamma\right)}{\partial\boldsymbol{\mu}}=\sum_{i=1}^{n} w_i\left(1-\nu_i\right)\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right), \tag{D.3}$$

with

$$\nu_i=\frac{\beta_{\max}}{2d(d+2)}\frac{\exp(\gamma)}{1+\exp(\gamma)}\frac{\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)-(d+2)}{q\left(\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right);\beta_{\max}\frac{\exp(\gamma)}{1+\exp(\gamma)}\right)}.$$

By computing the derivative of (D.2) with respect to $\boldsymbol{\Omega}$ we obtain

$$\frac{\partial \widetilde{l}\,(\boldsymbol{\mu},\boldsymbol{\Omega},\gamma)}{\partial \boldsymbol{\Omega}} = -\widetilde{w}\left(\boldsymbol{\Omega}'\right)^{-1} + \sum_{i=1}^{n} w_i\,(1-\nu_i)\left(\boldsymbol{\Omega}'\right)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\,(\boldsymbol{x}_i-\boldsymbol{\mu})'\,\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}; \tag{D.4}$$

the underlying algebra is given in Appendix E. Finally the derivative of (D.2) with respect to γ is

$$\frac{\partial \widetilde{l}\,(\boldsymbol{\mu},\boldsymbol{\Omega},\gamma)}{\partial \gamma} = \frac{1}{1+\exp(\gamma)}\sum_{i=1}^{n} w_i\left[1-\frac{1}{q\left((\boldsymbol{x}_i-\boldsymbol{\mu})'\,\boldsymbol{\Omega}^{-1}\,(\boldsymbol{\Omega}')^{-1}\,(\boldsymbol{x}_i-\boldsymbol{\mu})\,;\,\beta_{\max}\frac{\exp(\gamma)}{1+\exp(\gamma)}\right)}\right]. \tag{D.5}$$

The value of $\boldsymbol{\theta}=(\boldsymbol{\mu},\boldsymbol{\Omega},\gamma)$ that maximizes $\widetilde{l}$ is the maximum weighted likelihood estimate $\widehat{\boldsymbol{\theta}}$ and it is obtained by setting the first derivatives (D.3)–(D.5) equal to 0.

Operationally unconstrained maximization of (D.2) with respect to $\boldsymbol{\mu}$, $\boldsymbol{\Omega}$, and γ is obtained by the general-purpose optimizer `optim()` for R, included in the **stats** package. The BFGS method/algorithm, passed to `optim()` via the argument `method`, is used for maximization. It is an iterative unconstrained variable-metric/quasi-Newton method solving an approximate version of the Newton equations (to seek a zero of the gradient) by using an approximation of the inverse Hessian. A line search is applied to the resulting search direction, and a new trial solution found. The inverse Hessian is updated at each iteration by the Broyden–Fletcher–Goldfarb–Shanno formula giving the method its acronym (see, e.g., Nash, 2014, for further details). However by default the algorithm uses a numerical approximation for the gradient too. This is done using a first-order forward difference scheme. This involves m evaluations of the objective function $\widetilde{l}$, where m is the dimension of the parameter vector (number of free parameters). This greatly increases the computational effort. Motivated by this consideration we pass the analytical first derivatives (D.3)–(D.5) to `optim()` via its argument `gr`, and this significantly helps the computation. Finally we use the weighted moments to initialize the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Omega}$, and γ, with weights given by $w_i,\ldots,w_n$. Some simulations showing the behaviour of the BFGS algorithm are illustrated in Appendix F.

E - DERIVATIVE WITH RESPECT TO $\boldsymbol{\Omega}$

With reference to (D.4), and according to Equation (11) in Lütkepohl, (1996)., p. 182), we have

$$\frac{\partial \ln\left|\boldsymbol{\Omega}'\boldsymbol{\Omega}\right|}{\partial \boldsymbol{\Omega}} = 2\left(\boldsymbol{\Omega}'\right)^{-1}. \tag{E.6}$$

Using formula (19) in Faliva and Zoia (2008, p. 18), formula (16) in Lütkepohl, (1996), p. 177), and formula (1) in Lütkepohl, (1996) (2004, p. 198), together with the properties of the vec($\cdot$) operator versus the Kronecker product $\otimes$, we have

$$\begin{aligned}\operatorname{vec}\left[\frac{\partial \boldsymbol{v}_i'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\boldsymbol{v}_i}{\partial \boldsymbol{\Omega}}\right]' &= \frac{\partial \boldsymbol{v}_i'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\boldsymbol{v}_i}{\partial \operatorname{vec}(\boldsymbol{\Omega})'}\\ &= \frac{\partial \boldsymbol{v}_i'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\boldsymbol{v}_i}{\partial \operatorname{vec}\left(\boldsymbol{\Omega}^{-1}\right)'}\,\frac{\operatorname{vec}\left(\boldsymbol{\Omega}^{-1}\right)}{\operatorname{vec}(\boldsymbol{\Omega})'}\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{vec}\left[\frac{\partial \boldsymbol{v}_i' \boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\boldsymbol{v}_i}{\partial \boldsymbol{\Omega}^{-1}}\right]' \frac{\mathrm{vec}\left(\boldsymbol{\Omega}^{-1}\right)}{\mathrm{vec}\left(\boldsymbol{\Omega}\right)'} \\
&= -2\mathrm{vec}\left(\boldsymbol{v}_i\boldsymbol{v}_i'\boldsymbol{\Omega}^{-1}\right)'\left(\boldsymbol{\Omega}'\right)^{-1} \otimes \boldsymbol{\Omega}^{-1} \\
&= -2\mathrm{vec}\left[\left(\boldsymbol{\Omega}'\right)^{-1}\boldsymbol{v}_i\boldsymbol{v}_i'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\right]', \qquad \text{(E.7)}
\end{aligned}
$$

where $\boldsymbol{v}_i = \boldsymbol{x}_i - \boldsymbol{\mu}$. From (E.7) it follows that

$$
\frac{\partial \boldsymbol{v}_i' \boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}\boldsymbol{v}_i}{\partial \boldsymbol{\Omega}} = -2\left(\boldsymbol{\Omega}'\right)^{-1}\boldsymbol{v}_i\boldsymbol{v}_i'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}'\right)^{-1}. \qquad \text{(E.8)}
$$

Formulas (E.6) and (E.8) provide the basic derivatives needed to obtain (D.4).

F - PERFORMANCE OF THE BFGS ALGORITHM

In order to evaluate the behaviour of the BFGS algorithm in maximizing the log-likelihood function $\widetilde{l}(\boldsymbol{\mu}, \boldsymbol{\Omega}, \gamma)$ given in (D.2) we perform a simulation study. We consider four scenarios depending on d ($d = 2, 3, 4, 5$). For each scenario we generate 100 samples of size $n = 1{,}000$ from a single MLN distribution with a zero mean vector, an identity covariance matrix, and consider 17 different values of β spanned on a grid of equidistant points on the parameter space $[0, \beta_{\max}]$ (we recall that the maximum value for β depends on d). This set of values for β can be considered as a further simulation factor.

The subtables in Table 5 show the behaviour of the BFGS algorithm in the mentioned scenarios. In each subtable the true value of β is reported in the first column, followed by the average number of times the log-likelihood function is evaluated (`fn`), the average number of times the gradient function is evaluated (`gr`; which can be considered as the number of iterations of the BFGS algorithm), and the average (over 100 replications) of the estimated values of β ($\widehat{E}(\widehat{\beta})$).

Before commenting the obtained results it is important to underline that the algorithm reached convergence, and showed a monotonic behaviour, for all the values of d and β considered. As regards the number of times the log-likelihood function and the gradient function are evaluated it seems to increase near to the boundary of the β-parameter space, especially as β approaches $\beta_{\max}$. Finally the average of the estimated values of β appears close to the true one regardless of: (1) the true value of β (i.e., point in the parameter space), and (2) the considered scenario (dimension d).

G - KURTOSIS OF THE MCN DISTRIBUTION

A d-variate continuous random vector $\boldsymbol{X}$ is said to have a contaminated normal distribution with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, proportion of good points $\alpha \in [0.5, 1)$, and inflation parameter $\beta > 1$, if its density is given by

$$
f_{\boldsymbol{X}}\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \beta\right) = \alpha\phi\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) + (1-\alpha)\phi\left(\boldsymbol{x}; \boldsymbol{\mu}, \beta\boldsymbol{\Sigma}\right). \qquad \text{(G.9)}
$$

The covariance matrix of $\boldsymbol{X}$ is

$$
\mathrm{Var}(\boldsymbol{X}) = \left[\alpha + (1-\alpha)\beta\right]\boldsymbol{\Sigma}; \qquad \text{(G.10)}
$$

TABLE 5: Behaviour of the BFGS algorithm at the varying of d and β (over 100 replications)

(a) $[d=2]$				(b) $[d=3]$			
β	fn	gr	$\widehat{E}(\widehat{\beta})$	β	fn	gr	$\widehat{E}(\widehat{\beta})$
0	93.01	23.31	0.09	0	82.41	19.99	0.13
0.4	65.15	16.25	0.39	0.75	71.79	16.50	0.76
0.8	49.69	12.08	0.77	1.5	71.83	16.22	1.52
1.2	51.18	12.01	1.15	2.25	72.01	16.32	2.17
1.6	50.83	11.74	1.60	3	72.93	16.29	2.91
2	50.38	11.94	1.99	3.75	73.22	16.29	3.76
2.4	51.38	11.83	2.38	4.5	72.57	16.34	4.54
2.8	50.89	11.67	2.82	5.25	71.70	16.49	5.32
3.2	51.26	11.93	3.19	6	73.49	16.57	6.02
3.6	52.19	11.98	3.52	6.75	73.04	16.42	6.71
4	53.13	12.34	3.98	7.5	74.56	16.43	7.47
4.4	53.82	12.34	4.43	8.25	75.34	16.34	8.27
4.8	54.79	12.67	4.86	9	78.96	16.67	8.99
5.2	59.99	14.26	5.19	9.75	82.94	17.57	9.79
5.6	69.84	17.79	5.62	10.5	78.50	16.87	10.47
6	127.91	34.78	6.03	11.25	113.44	25.69	11.30
6.4	206.87	57.38	6.28	12	202.87	51.01	11.98

(c) $[d=4]$				(d) $[d=5]$			
β	fn	gr	$\widehat{E}(\widehat{\beta})$	β	fn	gr	$\widehat{E}(\widehat{\beta})$
0	100.53	25.78	0.12	0	119.95	31.68	0.18
1	97.59	22.34	0.95	1.25	131.76	29.83	1.21
2	101.09	22.56	2.00	2.5	133.04	30.08	2.40
3	100.35	22.34	2.97	3.75	133.62	30.00	3.86
4	100.91	22.22	3.95	5	133.41	29.96	4.92
5	101.00	22.50	4.98	6.25	133.39	29.75	6.23
6	99.88	22.15	6.06	7.5	134.80	29.85	7.60
7	101.35	22.31	6.94	8.75	135.08	29.49	8.67
8	100.89	22.12	7.98	10	134.35	29.32	10.14
9	100.23	22.06	9.06	11.25	137.04	29.29	11.34
10	102.82	22.26	10.01	12.5	136.42	29.20	12.47
11	102.91	22.12	10.97	13.75	136.15	29.10	13.82
12	104.30	22.06	12.06	15	138.84	29.20	15.05
13	106.11	22.62	13.08	16.25	140.33	29.40	16.25
14	106.74	22.56	14.02	17.5	143.67	29.73	17.64
15	136.59	30.01	15.02	18.75	173.08	37.23	18.87
16	186.68	44.86	15.98	20	191.21	49.36	19.98

cf. Punzo & McNicholas (2016). Based on (G.10),

$$\begin{aligned}\mathrm{Kurt}(\boldsymbol{X}) &= E\left\{\left[(\boldsymbol{X}-\boldsymbol{\mu})'\left\{[\alpha+(1-\alpha)\beta]\,\boldsymbol{\Sigma}\right\}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right]^2\right\}\\ &= d(d+2)\frac{\alpha+(1-\alpha)\beta^2}{[\alpha+(1-\alpha)\beta]^2}\\ &= d(d+2)\,c(\alpha,\beta). \qquad \text{(G.11)}\end{aligned}$$

It is interesting to note that Kurt($\boldsymbol{X}$) is equal to the kurtosis of a d-variate normal random vector, that is, $d(d+2)$, multiplied by a factor $c(\alpha, \beta)$ which depends on the parameters α and β of the MCN distribution. The minimum of $c(\alpha, \beta)$ is 1 and it is reached when both $\alpha \to 1^-$ and $\beta \to 1^+$. Note that, fixed α, the limit of $c(\alpha, \beta)$ for $\beta \to \infty$ is equal to $\alpha/(1-\alpha)$. Nevertheless an upper bound for $c(\alpha, \beta)$ does not exist as $c(\alpha, \beta) \to \infty$ when $\alpha \to 1^-$ and $\beta \to \infty$. Summarizing, Kurt($\boldsymbol{X}$) $\geq d(d+2)$.

BIBLIOGRAPHY

Absil, P. A., Mahony, R., & Sepulchre, R. (2009). Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ.

Aitken, A. C. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh*, 45(1), 14–22.

Aitkin, M. & Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. *Technometrics*, 22(3), 325–331.

Akgiray, V. & Booth, G. G. (1988). The stable-law model of stock returns. *Journal of Business & Economic Statistics*, 6(1), 51–57.

Andrews, J. L. & McNicholas, P. D. (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing*, 21(3), 361–373.

Andrews, J. L. & McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing*, 22(5), 1021–1029.

Andrews, J. L. & McNicholas, P. D. (2016). **teigen**: Model-based clustering and classification with the multivariate t-distribution. R package version 2.1.1 available at http://CRAN.R-project.org/package=teigen

Arevalillo, J. M. & Navarro, H. (2012). A study of the effect of kurtosis on discriminant analysis under elliptical populations. *Journal of Multivariate Analysis*, 107, 53–63.

Azzalini, A. (2016). R package **sn**: The skew-normal and skew-t distributions. R package version 1.4-0 available at http://CRAN.R-project.org/package=sn

Azzalini, A. & Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, 39(3), 357–365.

Azzalini, A., Browne, R. P., Genton, M. G., & McNicholas, P. D. (2016). On nomenclature for, and the relative merits of, two formulations of skew distributions. *Statistics & Probability Letters*, 110, 201–206.

Bagnato, L. & Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. *Computational Statistics*, 28(4), 1571–1597.

Balanda, K. P. & MacGillivray, H. L. (1988). Kurtosis: A critical review. *The American Statistician*, 42(2), 111–119.

Balanda, K. P. & MacGillivray, H. L. (1990). Kurtosis and spread. *Canadian Journal of Statistics*, 18(1), 17–30.

Banfield, J. D. & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3), 803–821.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., & Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2), 373–388.

Browne, R. P., ElSherbiny, A., & McNicholas, P. D. (2015). **mixture**: Mixture models for clustering and classification. R package version 1.4 available at https://cran.r-project.org/package=mixture

Browne, R. P. & McNicholas, P. D. (2014a). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification*, 8(2), 217–226.

Browne, R. P. & McNicholas, P. D. (2014b). Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*, 24(2), 203–210.

Browne, R. P. & McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2), 176–198.

Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3), 368–385.

Celeux, G. & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5), 781–793.

Cook, R. D. & Weisberg, S. (1994). *An Introduction to Regression Graphics*. John Wiley & Sons, New York.

Dang, U. J., Browne, R. P., & McNicholas, P. D. (2015). Mixtures of multivariate power exponential distributions. *Biometrics*, 71(4), 1081–1089.

Dang, U. J., Punzo A., McNicholas, P. D., Ingrassia, S., & Browne, R. P. (2016). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, 34(1).

Dasgupta, A. & Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441), 294–302.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1), 1–38.

Faliva, M., Zoia, M. G., & Poti, V. (2016). Orthogonal polynomials for tailoring density functions to excess kurtosis, asymmetry and dependence. *Communications in Statistics: Theory and Methods*, 45(1), 49–62.

Faliva, M. & Zoia, M. G. (2008). *Dynamic model analysis: Advanced matrix methods and unit-root econometrics representation theorems*. Springer-Verlag, Berlin Heidelberg.

Fraley, C. & Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611–631.

Fritz, H., García-Escudero, L. A., & Mayo-Iscar, A. (2012). **tclust**: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12), 1–26.

Galimberti, G. & Soffritti, G. (2014). A multivariate linear regression analysis using finite mixtures of *t* distributions. *Computational Statistics & Data Analysis*, 71, 138–150.

García-Escudero, L. A., Gordaliza, A., & Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2), 434–449

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3), 1324–1345.

Ghahramani, Z. & Hinton, G. E. (1997). The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, Toronto.

Gómez, E., Gómez-Viilegas, M. A., & Marin, J. M. (1998). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics—Theory and Methods*, 27(3), 589–600.

Gómez, E., Gómez-Villegas, M. A., & Marín, J. M. (2003). A survey on continuous elliptical vector distributions. *Revista Matemática Complutense*, 16(1), 345–361.

Gómez-Sánchez-Manzano, E., Gómez-Villegas, M. A., & Marín, J. M. (2008). Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications. *Communications in Statistics - Theory and Methods*, 37(6), 972–985.

Holzmann, H., Munk, A., & Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics*, 33(4), 753–763.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218.

Hunter, D. R. & Lange, K. (2000). Rejoinder to discussion of "optimization transfer using surrogate objective functions". *Journal of Computational and Graphical Statistics*, 9(1), 52–59.

Ingrassia, S., Minotti, S. C., & Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, 71, 159–182.

Jondeau, E. & Rockinger, M. (2001). Gram-Charlier densities. *Journal of Economic Dynamics and Control*, 25(10), 1457–1483.

Keribin, C. (1998). Estimation consistante de l'ordre de modèles de mélange. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 326(2), 243–248.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 62, 49–66.

Kotz, S. & Nadarajah, S. (2004). *Multivariate t-Distributions and Their Applications*. Cambridge University Press, Cambridge.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association*, 84(408), 881–896.

Lee, S. X. & McLachlan, G. J. (2014). Finite mixtures of multivariate skew *t*-distributions: Some recent and new results. *Statistics and Computing*, 24(2), 181–202.

Lin, T. I. (2010). Robust mixture modeling using multivariate skew *t* distributions. *Statistics and Computing*, 20(3), 343–356.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*, volume 5. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, California.

Lo, K. & Gottardo, R. (2012). Flexible mixture modeling via the multivariate *t* distribution with the Box-Cox transformation: an alternative to the skew-*t* distribution. *Statistics and Computing*, 22(1), 33–52.

Lütkepohl, H. (1996). *Handbook of Matrices*. John Wiley & Sons, New York.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530.

Maruotti, A. and Punzo, A. (2016). Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. *Computational Statistics & Data Analysis*. DOI: 10.1016/j.csda.2016.05.024.

Mazza, A., Punzo, A., & Ingrassia, S. (2015). **flexCWM**: Flexible cluster-weighted modeling. R package version 1.5 available at http://CRAN.R-project.org/package=flexCWM

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons, New York.

McLachlan, G. J., Peel, D., & Bean, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3), 379–388.

McLachlan, G. J., Bean, R. W. and Ben-Tovim Jones, L. (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate *t*-distribution. *Computational Statistics & Data Analysis*, 51(11), 5327–5338.

McNicholas, P. D. & Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3), 285–296.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., & Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics & Data Analysis*, 54(3), 711–723.

Mittnik, S., Rachev, S. T., & Kim, J.-R. (1998). Chi-square-type distributions for heavy-tailed variates. *Econometric Theory*, 14(03), 339–354.

Morris, K., McNicholas, P. D., Punzo, A., & Browne, R. P. (2014). Robust asymmetric clustering. arXiv.org e-print 1402.6744, available at:http://arxiv.org/abs/1402.6744

Murray, P. M., Browne, R. P., & McNicholas, P. D. (2014). Mixtures of skew-*t* factor analyzers. *Computational Statistics & Data Analysis*, 77, 326–335.

Nash, J. C. (2014). On best practice optimization methods in R. *Journal of Statistical Software*, 60(2), 1–14.

Peel, D. & McLachlan, G. J. (2000). Robust mixture modelling using the *t* distribution. *Statistics and Computing*, 10(4), 339–348.

Punzo, A., Browne, R. P., & McNicholas, P. D. (2016). Hypothesis testing for mixture model selection. *Journal of Statistical Computation and Simulation*, 86(14), 2797–2818.

Punzo, A. & Maruotti, A. (2016). Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model. *Journal of Computational and Graphical Statistics*, 25(4), 1097–1116.

Punzo, A., Mazza, A., & McNicholas, P. D. (2015). **ContaminatedMixt**: Model-Based Clustering and Classification with the Multivariate Contaminated Normal Distribution. R package version 1.0 available at https://cran.r-project.org/package=ContaminatedMixt

Punzo, A. & McNicholas, P. D. (2014). Robust high-dimensional modeling with the contaminated Gaussian distribution. arXiv.org e-print 1408.2128, available at: http://arxiv.org/abs/1408.2128

Punzo, A. & McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, 58(6), 1506–1537.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–164.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850.

Schork, N. J. & Schork, M. A. (1988). Skewness and mixtures of normal distributions. *Communications in Statistics-Theory and Methods*, 17(11), 3951–3969.

Schwarz, G. (1978). Estimating the Dimenson of a Model. *The Annals of Statistics*, 6, 461–464.

Steinley, D. (2004). Properties of the Hubert-Arable adjusted Rand index. *Psychological Methods*, 9(3), 386–396.

Subedi, S. & Punzo, A. and Ingrassia, S. and McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, 7(1), 5–40.

Subedi, S. & Punzo, A. and Ingrassia, S. and McNicholas, P. D. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications*, 24(4), 623–649.

Szegö, G. (2004). *Risk measures for the 21st century*. The Wiley Finance Series. Wiley, New York.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

Tortora, C. & Franczak, B. C. and Browne, R. P. and McNicholas, P. D. (2015). A mixture of coalesced generalized hyperbolic distributions. arXiv.org e-print 1403.2332, available at: http://arxiv.org/abs/1403.2332.

Wang, J. & Zhou, W. (2012). A generalized multivariate kurtosis ordering and its applications. *Journal of Multivariate Analysis*, 107, 169–180.

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3), 359–397.

Weisberg, S. (2011). **alr3**: Data to accompany applied linear regression 3rd edition. R package version 2.0.5 available at http://CRAN.R-project.org/package=alr3

Zhang, J. & Liang, F. (2010). Robust clustering using exponential power mixtures. *Biometrics*, 66(4), 1078–1086.

Zhu, X. and Melnykov, V. (2016). Manly transformation in finite mixture modeling. *Computational Statistics & Data Analysis*. DOI: 10.1016/j.csda.2016.01.015.

Zografos, K. (2008). On Mardia's and Song's measures of kurtosis in elliptical distributions. *Journal of Multivariate Analysis*, 99(5), 858–879.

Zoia, M. G. (2010). Tailoring the Gaussian law for excess kurtosis and skewness by Hermite polynomials. *Communications in Statistics-Theory and Methods*, 39(1), 52–64.

Received 24 November 2015
Accepted 3 October 2016