Flexible asymmetric multivariate distributions based on two-piece univariate distributions

[redacted]

Flexible asymmetric multivariate distributions based on two-piece univariate distributions

Jonas Baillien · Irène Gijbels · Anneleen Verhasselt

Received: date / Revised: date

Abstract Classical symmetric distributions like the Gaussian are widely used. However, in reality data often display a lack of symmetry. Multiple distributions have been developed to specifically cope with asymmetric data. These can be grouped under the name "skewed distributions". In this paper we present a broad family of flexible multivariate skewed distributions for which statistical inference is a feasible task. The studied family of multivariate skewed distributions is derived by taking affine combinations of independent univariate distributions. These univariate distributions are members of a flexible family of asymmetric distributions and are an important basis for achieving statistical inference. Besides basic properties of the proposed distributions, also statistical inference based on a maximum likelihood approach is presented. We show that under some mild conditions, weak consistency and asymptotic normality of the maximum likelihood estimators hold. These results are backed up by a simulation study which confirms the developed theoretical results and some data examples to illustrate practical applicability.

Keywords Affine combination · Maximum likelihood estimation · Multivariate skew distribution

Jonas Baillien
Department of Mathematics and Leuven Statistics Research Center (LStat),
KU Leuven, Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Heverlee), Belgium
E-mail: jonas.baillien@kuleuven.be

Irène Gijbels
Department of Mathematics and Leuven Statistics Research Center (LStat),
KU Leuven, Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Heverlee), Belgium
E-mail: irene.gijbels@kuleuven.be

Anneleen Verhasselt Center for Statistics, Data Science Institute, Hasselt University,
Agoralaan-building D, B-3590 Diepenbeek, Belgium
E-mail: anneleen.verhasselt@uhasselt.be

1 Introduction

Multivariate distributions provide the necessary ingredients to model all sorts of events where multidimensional data occur. They have established their importance in economics, chemistry, biology, etc. The most prominently present multivariate distribution is the multivariate normal distribution, which is a member of the class of multivariate elliptical distributions. In general, the more widely used distributions tend to be multivariate elliptical extensions of their univariate counterparts, thereby mimicking the multivariate normal distribution. The general formulation of a multivariate elliptical distribution generated by a univariate density generator $\widetilde{f}$ is, according to Azzalini (2013),

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{\frac{d}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}k_d}\widetilde{f}\left((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \qquad \mathbf{x}\in\mathbb{R}^d. \tag{1}$$

In this, $\boldsymbol{\mu}\in\mathbb{R}^d$ is a location parameter, $\boldsymbol{\Sigma}\in\mathbb{R}^{d\times d}$ is a positive definite matrix and $\widetilde{f}$ is such that

$$k_d = \int_0^\infty s^{d-1}\widetilde{f}(s^2)ds < \infty.$$

However, being elliptical has one major drawback in the form of a severe degree of symmetry of the distribution which, in reality, is not always present.

To better model asymmetric data, many asymmetric (or skewed) distributions have been proposed in both univariate and multivariate settings. Examples of the latter are the multivariate split normal distribution (Villani and Larsson (2007)), the multivariate slash Laplace distribution (Punathumparambath (2012)), and the bivariate alpha-skew normal distribution (Louzada et al. (2017)), and the multivariate slash- and skew-slash Student's t-distributions (Tan et al. (2015)), among others. These distributions lack generality and a unified approach concerning statistical inference. However, an exception to this is the family of skew-elliptical distributions. A univariate skew-elliptical distribution has as density function

$$h(z;\xi,\sigma,\alpha) = 2\sigma^{-1}f\left(\sigma^{-1}(z-\xi)\right)G(\alpha\sigma^{-1}(z-\xi)) \qquad z\in\mathbb{R}.$$

In this, f is a symmetric unimodal density, G the cumulative distribution function of an absolutely continuous, symmetric (around zero) univariate random variable and $(\xi,\sigma,\alpha)\in\mathbb{R}\times\mathbb{R}^+\backslash\{0\}\times\mathbb{R}$ respectively a location, scale and skewing parameter.

In Azzalini and Dalla Valle (1996) the first multivariate extension was presented, the multivariate skew-normal distribution. In, among others, Azzalini and Capitanio (2003) and Azzalini (2013), this was generalised to the multivariate skew-elliptical distribution, which has density function of the form

$$h_d(\mathbf{z}) = 2f_d(\mathbf{z})G(w(\mathbf{z})) \qquad \mathbf{z}\in\mathbb{R}^d, \tag{2}$$

with f_d an elliptical density as in (1), and G is an absolutely continuous, symmetric around zero, cumulative distribution function. Further herein the

function $w : \mathbb{R}^d \to \mathbb{R}$ is such that $w(-\mathbf{z}) = -w(\mathbf{z})$, for all $\mathbf{z} \in \mathbb{R}^d$. There are ample of combinations of distributions that can be made via this construction. Within this family, the most popular member is the multivariate skew-normal distribution, obtained by taking f_d a standard multivariate normal density and G the univariate standard normal cumulative distribution function. If these are replaced with their Student's t-counterparts, one obtains the popular skew-t distribution. Further extensions of the family are possible, see for example Adcock and Azzalini (2020) for the extended skew-elliptical distributions which incorporate an extra parameter.

There are however a variety of other general multivariate skewing mechanisms available. In Ley and Paindaveine (2010) a transformation approach is proposed. For a diffeomorphism $\mathbf{H}$, a multivariate skewed distribution is obtained as the function $\mathbf{X} \to f_{\mathbf{H}}(\mathbf{X}) \left|\det\left(\nabla \mathbf{H}(\mathbf{X})\right)\right|$. Another class of distributions are the Transformation of Scale distributions developed in Jones (2010) (univariate) and Jones (2016) (bivariate). An advantage of these distributions is that they are closed under marginalisation (i.e. the marginals have the same distributions as the bivariate vector). In Transformation of Scale distributions skewness is introduced in the following way.

$$\widehat{f}(x, y) = 2g(W_1^{-1}(x), W_2^{-1}(y)),$$

where $g(\cdot, \cdot)$ is a continuous bivariate density function and W_j^{-1}, for $j = 1, 2$ is the inverse of an increasing (transformation) function W_j which has to satisfy certain properties. See Jones (2016) for more details. Besides skewed distributions obtained by transformations, be it in on the density or the distribution function, Arnold et al. (2006) constructed multivariate skewed distributions by employing the Rosenblatt construction. This idea was further extended in Abtahi and Towhidi (2013) by introducing the unified skew symmetric distribution. The density of a member of this family is given by

$$s_d(\mathbf{z}) = f(\mathbf{z}) p\left(F(z_1), F(z_2|z_1), \ldots, F(z_d|z_1, \ldots, z_{d-1})\right),$$

with $\mathbf{z} \in \mathbb{R}^d$, $f(\mathbf{z})$ the density function of a symmetric random vector $\mathbf{U} \in \mathbb{R}^d$ (with this, central symmetry is meant, i.e. $f(-\mathbf{x}) = f(\mathbf{x})$), $p(\cdot)$ a d-variate density function on $[0, 1]^d$ and $F(\cdot|z_1, \ldots, z_{i-1})$ the distribution of $U_i|U_1 = z_1, \ldots, U_{i-1} = z_{i-1}$.

A point of attention for multivariate distributions should be the tail behavior when the distribution shows clearly distinct behavior in different directions (marginals). This point is also mentioned in Babić et al. (2019). Skew-elliptical distributions have a single parameter to govern tail-behavior for all d dimensions, which can be too restrictive. Even though the skewing parameter does have an impact on the tail behavior, in a classical skew-t distribution, for example, it is still only regulated by the degrees of freedom. See also Jones (2008) and Balakrishnan and Captitanio (2008), among others. This problem is possibly shared with distributions obtained through transformations, depending on what transformation was used. Our goal is to provide a unified, tractable framework for statistical inference for the entire considered family with the added flexibility of allowing different types of behavior in different directions.

We start from the univariate quantile-based asymmetric (QBA) family of distributions recently studied in Gijbels et al. (2019). In its simplest form, the density function of a QBA-distribution is defined as

$$f_Z(z;\boldsymbol{\eta}) = 2\alpha(1-\alpha)\begin{cases} f\left(-(1-\alpha)z;\boldsymbol{\kappa}\right) & \text{if } z \leq 0 \\ f\left(\alpha z;\boldsymbol{\kappa}\right) & \text{if } z > 0, \end{cases} \tag{3}$$

with $\boldsymbol{\eta} = (\alpha, \boldsymbol{\kappa}^T)^T$. In this $f(.;\boldsymbol{\kappa})$ is a unimodal, symmetric (around zero) continuous density function. The interpretation of the elements contained in the parameter vector $\boldsymbol{\eta}$ is as follows. The parameter $\alpha \in (0,1)$ governs the skewness and $\boldsymbol{\kappa}$ are possible different parameters (excluding location or scale parameters) of f_Z. An example for the latter is the degrees of freedom parameter of a Student's t-distribution. Note that when $\alpha = 0.5$ then $f_Z = f$ everywhere and hence the density f_Z is symmetric. When α deviates from 0.5 one obtains a skewed distribution. This family of distributions falls in the category of two-piece or split-type distributions. Note that (3) does not incorporates a location or scale parameter. As made clear later on, including them would lead to identifiability problems for the multivariate extension. A vast literature is available on the approach of two-piece distributions, dating back as far as Fechner (1897). A recent review regarding two-piece distributions was provided by Wallis (2014). There are different ways of constructing two-piece distributions, i.e. different parametrisations are possible. See for example Rubio and Steel (2014). We opt to choose the particular parametrisation as in Gijbels et al. (2019), since it allows to provide statistical inference for *any* member of the resulting family of asymmetric multivariate distributions.

Applying the univariate skewing mechanism to a multivariate distribution is a common technique used to create multivariate skewed distributions. Examples can be found in Azzalini and Dalla Valle (1996) and Louzada et al. (2017). For two-piece distributions in general, this is proposed in Arellano-Valle et al. (2005) and Bauwens (2005). The downsides of such an approach are twofold, namely loss of tractability and flexibility. A different technique in obtaining multivariate distributions is the mechanism used in Villani and Larsson (2007) and generally exposed in Ferreira and Steel (2007). Fernández and Steel (1998) introduced the renowned Fernández-Steel skew distribution, and proposed the use of an affine combination of independent univariate skewed distributions to construct a multivariate skewed distribution.

Due to the general applicability of the affine combinations technique and its clear interpretation, and the close relation between the QBA-family and the Fernández-Steel distributions we opt for this technique to construct the proposed family of multivariate asymmetric distributions. In doing so, the added flexibility of different behavior in different directions is guaranteed. Ferreira and Steel (2007) introduced the multivariate skewing technique of affine combinations by taking, as the name suggests, affine combinations of (independent) univariate skewed distributions. For a random vector of independent univariate skewed distributions (in their example Fernández-Steel skewed distribution) $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_d)^T$, a flexible multivariate distribution is thus obtained as

a distribution of a random vector

$$\mathbf{S} = \mathbf{M}^T \boldsymbol{\epsilon} + \boldsymbol{\mu}.$$

In this, $\mathbf{M} \in \mathbb{R}^{d \times d}$ is the mixing matrix and $\boldsymbol{\mu} \in \mathbb{R}^d$ a location shift. In Ferreira and Steel (2007) the density function is provided, and conditions on $\mathbf{M}$ for the model to be identifiable are described. Expressions for moments are given. Inference is presented in a Bayesian context. What we will present is a similar approach, using the family of QBA-distributions for the independent univariate skewed components and an alternative set of identifiability conditions to obtain a family of multivariate skewed distributions. In doing so, the added flexibility of different behavior in different directions is guaranteed. Statistical inference results are developed for the maximum likelihood estimator (MLE). The literature of the linear combinations technique is expanded from a Bayesian setting in Ferreira and Steel (2007), to a frequentist setting with a general (unified) approach of obtaining maximum likelihood inference. Although it is in a specific (family) setting, these results can be extended to other families of distributions and provide a way to obtain statistical inference results.

The outline of the paper is as follows. In Section 2 the quantile-based asymmetric family of distributions is extended to create a family of flexible asymmetric multivariate distributions. Along with the formulation of the density function of the proposed family, probabilistic properties are also derived. A brief discussion on ways to measure asymmetry is included. In Section 3, the focus shifts towards the asymptotic distribution of maximum likelihood parameter estimates. This asymptotic behaviour is illustrated by a simulation study, of which results are presented in Section 4. Before ending the exposition with a short conclusion in Section 6, some real-data applications are presented in Section 5. Proofs of the main theoretical results are deferred to the Appendix. A brief explanation about the relation to independence component analysis, and proofs of the other theoretical results are given in the Supplementary Material. This material also includes an extension involving asymmetric Student's t-distributions. R codes for the practical use of the methodology are available via the GitHub platform at `https://github.com/Anonymous162222/LCQBA`. Furthermore, an R markdown document, guiding the user through some examples, is provided in the Supplementary Material.

2 Family of flexible asymmetric multivariate distributions and its probabilistic properties

2.1 Defining the family

Despite the wide array of available distributions that can be used in an affine combination, in what follows, we restrict ourselves to members of the QBA-family. We opt to use this type of distribution with flexibility in mind, i.e.

possibly different behavior with respect to asymmetry in the different directions. Starting from a family of univariate distributions which is in its own right flexible and contains the symmetric counterparts of its members, is beneficial for the obtained multivariate distributions.

Define $\mathbf{Z} = (Z_1, \ldots, Z_d)^T$ where Z_j, $j = 1, \ldots, d$, has a density function $f_{Z_j}(z_j; \boldsymbol{\eta}_j)$ as in (3) with $\boldsymbol{\eta}_j = (\alpha_j, \boldsymbol{\kappa}_j^T)^T$ and generated by a symmetric, unimodal continuous density f_j. Throughout the paper we assume that all generating densities f_j are continuously differentiable. Furthermore, we assume that the components of the random vector $\mathbf{Z}$ are independent. Therefore $\mathbf{Z}$ has a joint density

$$f_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\eta}) = \prod_{j=1}^{d} f_{Z_j}(z_j; \boldsymbol{\eta}_j),$$

in which $\mathbf{z} = (z_1, \ldots, z_d)^T \in \mathbb{R}^d$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^T, \ldots, \boldsymbol{\eta}_d^T)^T$. The proposed asymmetric multivariate density is defined as the density of a d-variate random vector $\mathbf{X}$

$$\mathbf{X} = \mathbf{A}^T\mathbf{Z} + \boldsymbol{\mu}_a, \tag{4}$$

in which $\boldsymbol{\mu}_a = (\mu_{a,1}, \ldots, \mu_{a,d})^T \in \mathbb{R}^d$ is a location shift and $\mathbf{A} \in \mathbb{R}^{d\times d}$, a non-singular matrix, governs the dependence structure. By introducing a location shift and scaling in (4), the need for a location and scale parameter in each of the components of $\mathbf{Z}$ is superfluous. Therefore, location and scale parameters are not included in (3). By the transformation formula for affine combinations of random variables, the joint density of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}; \mathbf{A}, \boldsymbol{\mu}_a, \boldsymbol{\eta}) = |\det(\mathbf{A})|^{-1} \prod_{j=1}^{d} f_{Z_j}((\mathbf{x} - \boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j}; \boldsymbol{\eta}_j), \tag{5}$$

where we introduced the notation that for any matrix $\mathbf{M}$, $\mathbf{M}_{\cdot,j}$ denotes the j-th column of $\mathbf{M}$ and $\mathbf{M}_{i,\cdot}$ the i-th row of $\mathbf{M}$.Ferreira and Steel (2007) considered construction (4) and densities of the form (5), using closely related two-piece distributions (with different parametrisations). Obviously, feasibility of statistical inference for the parameters $\mathbf{A}, \boldsymbol{\mu}_a$ and $\boldsymbol{\eta}$ in (5) heavily depends on the specific choice of the univariate two-piece distributions, and inference results for these.

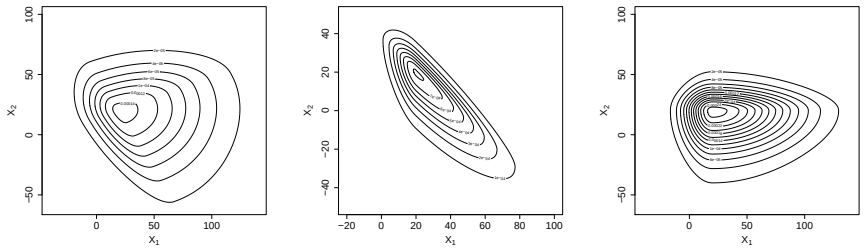
As an illustration of what this type of distribution looks like, consider the following example.

Example 1 Consider the following three models. For the first model, take as the first univariate component a QBA-normal distribution (f_{Z_1}) and as the second component a QBA-logistic distribution (f_{Z_2}) with the following parameters

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.25 \\ 0.65 \end{pmatrix}, \ \boldsymbol{\mu}_a = \begin{pmatrix} 20 \\ 20 \end{pmatrix}, \quad \text{and} \ \ \mathbf{A} = \begin{bmatrix} 12 & 4 \\ -5 & 8 \end{bmatrix}.$$

The second model is a variation of the first one wherein $\mathbf{A}$ is replaced with $\mathbf{B} = \begin{bmatrix} 7 & -6 \\ 0 & 3 \end{bmatrix}$. For the third model, the QBA-logistic component of the first

model is replaced with a QBA-Student's t-distribution with five degrees of freedom, and $\mathbf{A}$ is replaced with $\mathbf{C} = \begin{bmatrix} 12 & 0 \\ 0 & 8 \end{bmatrix}$. The contourplots of the densities of the resulting distributions are depicted in Figure 1. As can be seen, the mixing matrix can greatly impact the shape and scale of the resulting distribution. Note the change of main directions of the contours, when comparing the plots of Figure 1(a) and 1(b). This is due to the change of mixing matrix from $\mathbf{A}$ to $\mathbf{B}$. The contourplot in Figure 1(c) shows the benefit of combining different distributions. Note the heavier tails in the X_2 direction.



(a) *First model.* (b) *Second model.* (c) *Third model.*

Fig. 1: *Contour plots of the bivariate models of Example 1.*

2.2 Probabilistic properties of the family

Starting from the analytical expression for the density in (5), some of its basic properties can be derived. It is important to note that some of these properties, like moments and even the cumulative distribution function, may lack closed-form expressions.

2.2.1 Cumulative distribution function

The cumulative distribution function of any member of the proposed family is given by

$$F_{\mathbf{X}}(\mathbf{y}; \mathbf{A}, \boldsymbol{\mu}_a, \boldsymbol{\eta}) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_d} \prod_{j=1}^{d} \frac{1}{|\det(\mathbf{A})|} f_{Z_j}((\mathbf{x} - \boldsymbol{\mu}_a)^T (\mathbf{A}^{-1})_{\cdot,j}; \boldsymbol{\eta}_j) dx_1 \cdots dx_d. \quad (6)$$

In general, no analytical expression can be obtained for (6) due to the complexity of the integral. There are some specific cases where it is possible to derive a closed form expression, for example when only QBA-Laplace distributed random variables are used. In general however, numerical approximations of the cumulative distribution function are easy to obtain via Monte Carlo simulation. This is due to the ease with which the target distribution can be sampled,

as for each component of $\mathbf{Z}$ the quantile function is available in a formulation related to the quantile function of the underlying symmetrical density. To obtain a sample from $\mathbf{X}$, the same technique as the construction of (4) can be employed. Based on the sample, an approximation of the cumulative distribution function can then be obtained through its empirical counterpart.

2.2.2 Moments and characteristic function

From the linear combination of the components of $\mathbf{Z}$ and their independence, it is easy to see that the mean and variance of $\mathbf{X}$ are

$$\begin{aligned} E[\mathbf{X}] &= \mathbf{A}^T E[\mathbf{Z}] + \boldsymbol{\mu}_a, \\ \mathrm{Cov}(\mathbf{X}) &= \mathbf{A}^T \mathrm{Cov}(\mathbf{Z})\mathbf{A} = \mathbf{A}^T \mathrm{diag}(\mathrm{Var}(Z_1), \ldots, \mathrm{Var}(Z_d))\mathbf{A}, \end{aligned} \tag{7}$$

where diag is a diagonal matrix. The expressions for $E[Z_j]$ and $\mathrm{Var}(Z_j)$ are given by (see Gijbels et al. (2019))

$$E[Z_j] = \frac{1-2\alpha_j}{\alpha_j(1-\alpha_j)}\mu_{j,1},$$

and

$$\mathrm{Var}(Z_j) = \frac{(1-2\alpha_j)^2(\mu_{j,2} - \mu_{j,1}^2) + \alpha_j(1-\alpha_j)\mu_{j,2}}{\alpha_j^2(1-\alpha_j)^2},$$

with $\mu_{j,r} = 2\int_0^\infty s^r f_j(s)ds$, $r = 1, 2$. Other moments can be derived in a similar fashion. See Ferreira and Steel (2007) for similar moments expressions under different parametrisations.

The moments can also be calculated through the characteristic function. We make a distinction between the marginal characteristic functions and the joint characteristic function.

Proposition 1 *The marginal characteristic function $\varphi_{X_k}(t)$ of X_k is given by*

$$\varphi_{X_k}(t) = E\left[e^{itX_k}\right] = 2^d e^{it\mu_{a,k}} \prod_{j=1}^{d} \left(\alpha_j \varphi_j^+\left(\frac{-\mathbf{A}_{j,k}t}{1-\alpha_j}\right) + (1-\alpha_j)\varphi_j^+\left(\frac{\mathbf{A}_{j,k}t}{\alpha_j}\right)\right),$$

whereas the joint characteristic function $\varphi_{\mathbf{X}}(\mathbf{t})$ of $\mathbf{X}$ is given by

$$\varphi_{\mathbf{X}}(\mathbf{t}) = E\left[e^{i\mathbf{t}^T\mathbf{X}}\right] = 2^d e^{i\mathbf{t}^T\boldsymbol{\mu}_a} \prod_{j=1}^{d} \left(\alpha_j \varphi_j^+\left(\frac{-\mathbf{A}_{j,\cdot}\mathbf{t}}{1-\alpha_j}\right) + (1-\alpha_j)\varphi_j^+\left(\frac{\mathbf{A}_{j,\cdot}\mathbf{t}}{\alpha_j}\right)\right).$$

In both, $\varphi_j^+(t) = \int_0^\infty e^{its} f_j(s)ds$.

The proof is straightforward and given in the Supplementary Material.

Example 2 An easy to calculate joint characteristic function is when all Z_j are QBA-Laplace. This leads to

$$\varphi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}^T\boldsymbol{\mu}_a} \prod_{j=1}^{d} \frac{\alpha_j(1-\alpha_j)}{(\alpha_j - i\mathbf{A}_{j,\cdot}\mathbf{t})(1-\alpha_j + i\mathbf{A}_{j,\cdot}\mathbf{t})}.$$

This is different from the characteristic function of an elliptical multivariate Laplace distribution with skewing parameters $\mathbf{m}$ and scaling matrix $\boldsymbol{\Sigma}$, which is given by $\varphi_{\mathbf{X}}(\mathbf{t}) = \frac{1}{1+\frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t} - i\mathbf{m}^T\mathbf{t}}$ (see Kotz et al. (2001)). Hence, applying the affine transformation principle leads to a different, possibly non-elliptical, multivariate Laplace distribution.

2.2.3 Measures of asymmetry

Measures of asymmetry for a multivariate distribution can be characterized as a multivariate extension of skewness measures of a univariate distribution. For a univariate random variable X with mean μ and variance σ^2, skewness is defined as

$$\beta_1(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]. \tag{8}$$

There is no unique equivalent of (8) for a d-variate r.v. $\mathbf{X}$ with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. We briefly discuss three available measures of multivariate skewness, which resemble univariate skewness. In particular, we provide their expressions for the considered family of multivariate distributions. Denote $\mathbf{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X}-\boldsymbol{\mu})$, the standardized version of $\mathbf{X}$, where $\mathbf{X} = \mathbf{A}^T\mathbf{Z} + \boldsymbol{\mu}_a$. Recall from (7) that $\boldsymbol{\mu} = E[\mathbf{X}] = \mathbf{A}^T E[\mathbf{Z}] + \boldsymbol{\mu}_a$ and $\boldsymbol{\Sigma} = \mathbf{A}^T \mathrm{diag}(\mathrm{Var}(Z_1), \ldots, \mathrm{Var}(Z_d))\mathbf{A}$. We consider the following three measures of multivariate skewness.

Mardia's skewness index proposed in Mardia (1970). With $\mathbf{X}_1$ and $\mathbf{X_2}$ independent copies of $\mathbf{X}$

$$\beta_d(\mathbf{X}) = E[\left((\mathbf{X_1}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X_2}-\boldsymbol{\mu})\right)^3] = \sum_{j=1}^{d}\beta_1^2(Z_j). \tag{9}$$

In this, $\beta_1(Z_j)$ is the skewness as in (8) of the j-th component of $\mathbf{Z}$ as given in Gijbels et al. (2019). Due to rotational invariance of (9), it holds that $\beta_d(\mathbf{X}) = \beta_d(\mathbf{Y}) = \beta_d(\mathbf{Z})$.

Móri–Rohatgi–Székely measure proposed in Móri et al. (1994). This is a vector valued measure of asymmetry given by

$$\begin{aligned}\mathbf{s}(\mathbf{Y}) &= E\left[\left(\sum_{j=1}^{d} Y_j^2\right)\mathbf{Y}\right] \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T\right)_{j,i}^2\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T\right)_{\cdot,i} E\left[(Z_i - E[Z_i])^3\right].\end{aligned} \tag{10}$$

Kollo measure proposed in Kollo (2008). Like the Móri–Rohatgi–Székely measure, this is a vector valued measure of asymmetry that takes into account several extra terms. It is given by

$$\mathbf{b}(\mathbf{Y}) = E\left[\left(\sum_{j=1}^{d}\sum_{k=1}^{d} Y_j Y_k\right)\mathbf{Y}\right] \tag{11}$$
$$= \sum_{k=1}^{d}\left[\sum_{j=1}^{d}\sum_{i=1}^{d}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T\right)_{j,k}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T\right)_{i,k}\right]\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T\right)_{.,k} E\left[(Z_k - E[Z_k])^3\right].$$

All three measures of multivariate skewness are related in the sense that they are a combination of third order cumulants of $\mathbf{Y}$. Denote with $\mathbf{B}\otimes\mathbf{C}$ the Kronecker product of two matrices $\mathbf{B}\in\mathbb{R}^{n\times p}$ and $\mathbf{C}\in\mathbb{R}^{m\times q}$. The matrix of third order central moments of $\mathbf{Y}$ is in this case equal to that of the third order cumulants and given by

$$\mathbf{m}_3(\mathbf{Y}) = E[\mathbf{Y}\otimes\mathbf{Y}^T\otimes\mathbf{Y}]\in\mathbb{R}^{d^2\times d}.$$

The vectorization operator applied to this matrix leads to

$$\boldsymbol{\kappa}_3(\mathbf{Y}) = \mathrm{vec}(\mathbf{m}_3(\mathbf{Y}))\in\mathbb{R}^{d^3}.$$

The (i,j)-th element of $\mathbf{m}_3(\mathbf{Y})$ is given by $E[Y_j Y_{i-k^*d} Y_{k^*+1}]$ with $k^* = \{\max_{k\in\mathbb{N}} k | i - kd > 0\}$. In Jammalamadaka et al. (2020), it is noted that, with $\|\cdot\|$ the Euclidean norm of a vector,

$$\begin{aligned}\beta_d(\mathbf{Y}) &= \|\boldsymbol{\kappa}_3(\mathbf{Y})\|^2\\ \mathbf{s}(\mathbf{Y}) &= (\mathrm{vec}(\mathbf{I}_d)^T\otimes\mathbf{I}_d)\boldsymbol{\kappa}_3(\mathbf{Y})\\ \mathbf{b}(\mathbf{Y}) &= (\mathbf{1}_{d^2}^T\otimes\mathbf{I}_d)\boldsymbol{\kappa}_3(\mathbf{Y}),\end{aligned}$$

in which $\mathbf{1}_m$ is a vector of ones of dimension m, and $\mathbf{I}_d$ a d-dimensional identity matrix. It is also clear that the Móri–Rohatgi–Székely (MRS) measure $\mathbf{s}(\mathbf{Y})$ in (10) and the Kollo measure $\mathbf{b}(\mathbf{Y})$ in (11) are very similar, the only difference being that the Kollo measure takes into account extra third order cumulants. In our framework of linear combinations, the relation between the two measures is as follows

$$\mathbf{b}(\mathbf{Y}) =$$
$$\mathbf{s}(\mathbf{Y}) + \sum_{k=1}^{d}\sum_{j=1}^{d}\sum_{\substack{i=1\\ i\neq j}}^{d}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T\right)_{i,k}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T\right)_{j,k}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T\right)_{.,k} E\left[(Z_k - E[Z_k])^3\right].$$

The extra term in $\mathbf{b}(\mathbf{Y})$, when compared to $\mathbf{s}(\mathbf{Y})$, can cause a sign difference when comparing both measures, depending on the sign of elements of both $E\left[(\mathbf{Z} - E[\mathbf{Z}])^3\right]$ and $\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}^T$. If $\mathbf{A}$ is a diagonal matrix, both measures yield the same values.

Table 1: Measures of asymmetry for the different models in Example 1.

Model	First model	Second model	Third model
Mardia ($\beta_d(\mathbf{Y})$)	1.0440	1.0440	2.0044
MRS ($\mathbf{s}(\mathbf{Y})$)	$\begin{bmatrix} 0.9498 \\ -0.3768 \end{bmatrix}$	$\begin{bmatrix} 0.2652 \\ -0.9867 \end{bmatrix}$	$\begin{bmatrix} 0.6949 \\ -1.2335 \end{bmatrix}$
Kollo ($\mathbf{b}(\mathbf{Y})$)	$\begin{bmatrix} 1.1866 \\ 0.3813 \end{bmatrix}$	$\begin{bmatrix} -0.5305 \\ -1.2659 \end{bmatrix}$	$\begin{bmatrix} 0.6949 \\ -1.2335 \end{bmatrix}$

In Table 1 we list the values of these three measures of multivariate asymmetry for the three illustrative models in Example 1. This table illustrates the affine invariance of $\beta_d(\mathbf{Y})$ (see the first row of Table 1), and that $\mathbf{s}(\mathbf{Y})$ and $\mathbf{b}(\mathbf{Y})$ are equal if the mixing matrix is diagonal (see the last column). Note the remarkable higher (absolute) values of some skewness components in the third model.

3 Parameter estimation and asymptotic theory

A natural and efficient way of obtaining parameter estimates is through maximum likelihood estimation. Recall that the joint density of the random vector $\mathbf{X}$ is given by (5). Throughout the paper we assume that this model is correctly specified. For the moment, we restrict ourselves to densities f_j without additional parameters, i.e. for which $\boldsymbol{\kappa}_j$ is empty. Denote with $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\mu}_a^T, \mathrm{vec}(\mathbf{A})^T)^T \in \boldsymbol{\Theta} = [0,1]^d \times \mathbb{R}^{d(d+1)}$ the parameter vector of dimension $d^2 + 2d$. The parameters that need to be estimated are the $d \times d$ matrix $\mathbf{A}$, the d-vector $\boldsymbol{\mu}_a$ and the d-vector of skewing parameters $\boldsymbol{\alpha}$. Given a realization $\mathbf{x}^{(n)} = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$ of an i.i.d. sample $\mathbf{X}^{(n)} = (\mathbf{X}^1, \ldots, \mathbf{X}^n)$ of size n from $\mathbf{X}$, the log-likelihood function $\ell(\boldsymbol{\theta}, \mathbf{x})$ is

$$
\begin{aligned}
&\ell(\boldsymbol{\theta}, \mathbf{x}) \\
&= -\ln(|\det(\mathbf{A})|) + d\ln(2) + \sum_{j=1}^{d} \ln(\alpha_j(1-\alpha_j)) \\
&\quad + \sum_{j=1}^{d} \left[\mathbb{1}\left\{ (\mathbf{x}-\boldsymbol{\mu}_a)^T (\mathbf{A}^{-1})_{\cdot,j} \leq 0 \right\} \ln\left(f_j\left(-(1-\alpha_j)(\mathbf{x}-\boldsymbol{\mu}_a)^T (\mathbf{A}^{-1})_{\cdot,j} \right) \right) \right. \\
&\quad \left. + \mathbb{1}\left\{ (\mathbf{x}-\boldsymbol{\mu}_a)^T (\mathbf{A}^{-1})_{\cdot,j} > 0 \right\} \ln\left(f_j\left(\alpha_j (\mathbf{x}-\boldsymbol{\mu}_a)^T (\mathbf{A}^{-1})_{\cdot,j} \right) \right) \right], \qquad (12)
\end{aligned}
$$

where $\mathbb{1}\{B\}$ denotes the indicator function, i.e. $\mathbb{1}\{B\} = 1$ if B holds and 0 otherwise. The finite sample version of the log-likelihood function is given by

$$
\ell\left(\boldsymbol{\theta}, \mathbf{x}^{(n)}\right) = \sum_{i=1}^{n} \ell\left(\boldsymbol{\theta}, \mathbf{x}^i\right),
$$

and maximizing this log-likelihood with respect to $\boldsymbol{\theta}$ leads to the maximum likelihood estimator (MLE) for $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}_n^{\mathrm{ML}}$.

The log-likelihood is continuously differentiable with respect to $\boldsymbol{\alpha}$, but in general not with respect to elements of $\mathbf{A}$ or $\boldsymbol{\mu}_a$ whenever $(\mathbf{x}^i - \boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j} = 0$ for any $i = 1\ldots,n$. This occurs whenever the reference density f_j is not continuously differentiable. The latter is the case for, for example, the Laplace distribution at its mode. In addition, the second order derivative with respect to these parameters is non-continuous for an even larger selection of reference densities. This is for example the case when f_j is a normal density. Classical regularity conditions thus no longer apply. In the next section we formulate a set of conditions under which asymptotic theory for the MLE holds.

3.1 Identifiability of the model and consistency of the parameter estimator

A first issue that needs to be resolved with an eye on statistical inference is identifiability, so no two sets of parameters should lead to the same distribution. First we give necessary and sufficient conditions to ensure that the model is indeed identifiable. For example Allman et al. (2009) (on mixture models) and Beckmann and Smith (2004) (p140, on linear structure models) state that parameters of a random variable obtained from a combination of multiple univariate random variables is identifiable if it is unique up to a relabeling of the univariate random variables. Following this, we move away from the classical definition of identifiability and impose a slightly weaker one where we need (classical) identifiability up to a relabeling of the components of $\mathbf{Z}$ and the corresponding relabeling in both $\boldsymbol{\alpha}$ and the rows of $\mathbf{A}$.

Proposition 2 *Suppose* $\mathbf{X}$ *is generated according to* (4)*. Also assume that the vector of independent univariate random variables* $\mathbf{Z}$ *is such that all univariate densities* f_{Z_j}*, for* $j = 1,\ldots,d$*, are known up to their parameters. If the following conditions hold, the model with density function* (5) *is identifiable up to permutation of the independent components.*

(*I1*) *At most one component of* $\mathbf{Z}$ *can have a symmetric standard Gaussian distribution.*
(*I2*) *The diagonal elements of* $\mathbf{A}^{-1}$ *or* $\mathbf{A}$ *are strictly positive.*

For a square, invertible matrix, condition (*I2*) can always be satisfied (possibly after a permutation of the rows of $\mathbf{A}$ or columns of $\mathbf{A}^{-1}$ and a possible sign change) as shown in Proposition 3. However, other conditions on $\mathbf{A}$ or $\mathbf{A}^{-1}$ may also be imposed as long as they unambiguously fix the sign as mentioned above. For example, the condition

(*I2**) The first non-zero element in each column of $\mathbf{A}$ (or $\mathbf{A}^{-1}$) is strictly positive,

also suffices.

Proposition 3 *Let* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *be an invertible matrix. Then there exists a permutation of the rows of* $\mathbf{A}$ *such that every diagonal element of the permuted matrix is non-zero.*

Identifiability of the model is a key requirement in getting to statistical inference results. This is made clear in the following proposition, which states conditions under which the MLE is weakly consistent.

Proposition 4 *Let* $\mathbf{X}^{(n)}$ *be an i.i.d. sample from* $\mathbf{X}$ *with probability density as in* (5)*. Assume that the following assumptions hold:*

(*C1*) *Assumptions* $(I1)-(I2)$ *hold. In other words, the parameters are identifiable.*

(*C2*) *Let* $\boldsymbol{\Theta}_R = [-\mu_u, \mu_u]^d \times [\alpha_l, \alpha_u]^d \times [A_l, A_u]^{d^2}$*, with* $|\mu_u| < \infty$*,* $0 < \alpha_l < \alpha_u < 1$ *and* $-\infty < A_l < A_u < \infty$*, be a compact subset of* $\boldsymbol{\Theta}$*. Also assume that* $\boldsymbol{\theta}_0 \in \overset{\circ}{\boldsymbol{\Theta}}_R$*, with* $\overset{\circ}{\boldsymbol{\Theta}}_R$ *the interior of* $\boldsymbol{\Theta}_R$*.*

(*C3*) $\int_0^\infty |\ln f_j(s)|\, f_j(s) ds < \infty \ \forall j \in \{1, \ldots, d\}$*, where* $f_j(s)$ *are the underlying univariate symmetric densities.*

Then the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n^{ML}$ *is weakly consistent, i.e.* $\widehat{\boldsymbol{\theta}}_n^{ML} \xrightarrow{P} \boldsymbol{\theta}_0$ *for* $n \to \infty$*, with* $\boldsymbol{\theta}_0$ *the true parameter.*

3.2 Asymptotic normality

Before stating conditions under which asymptotic normality of the MLE holds, some matrix notations are introduced, needed in particular for providing expressions for the expected score and the Fisher information matrix. Denote with $\mathbf{A}_{-j;-i}$ the matrix $\mathbf{A}$ of which the j-th row and i-th column have been removed. Similarly $\mathbf{A}_{-j,-k;-i,-l}$ represents the matrix $\mathbf{A}$ in which the j- and k-th row and the i- and l-th column have been removed. In this, the order of the indices is of no importance as they are taken with respect to the original matrix $\mathbf{A}$. Should one start with a 2×2 matrix, $\mathbf{A}_{-1,-2;-1,-2}$ is defined as 1.

It is important to pay attention to the indexation of the rows and columns of the reduced matrices compared to the original one. For example, in $\mathbf{A}_{-j,-k;-i,-l}$ the r-th row has elements with row-index $r+2$ whereas the s-th column has elements with column-index $s+1$ provided that $r \geq k+2, k > j$ and $i < s \geq l+2$.

The next two well known results are also used. The first is the general result (see for example Zhang (2011), p12) that the determinant of a matrix $\mathbf{A}$ can be written as

$$\det(\mathbf{A}) = \sum_{l=1}^{d} (-1)^{l+k} \mathbf{A}_{k,l} \det(\mathbf{A}_{-k;-l}). \tag{13}$$

Using (13), the determinant of a reduced matrix is

$$\det(\mathbf{A}_{-k;-l}) = \sum_{\substack{i=1 \\ i \neq l}}^{d} (-1)^{i+j+\mathbb{1}\{j>k\}+\mathbb{1}\{i>l\}} \mathbf{A}_{j,i} \det(\mathbf{A}_{-j,-k;-i,-l}) \mathbb{1}\{j \neq k\}, \tag{14}$$

and

$$\frac{\partial}{\partial \mathbf{A}_{j,i}} \det(\mathbf{A}_{-k;-l}) = \begin{cases} 0 & \text{if } j = k \text{ or } i = l \\ (-1)^{i+j+\mathbb{1}\{j>k\}+\mathbb{1}\{i>l\}} \det(\mathbf{A}_{-j,-k;-i,-l}) & \text{otherwise.} \end{cases} \tag{15}$$

Note the added indicator functions in the exponent of -1. These follow from the previously made remark on the indexation of the reduced matrices.

The second result follows from the fact (see for example Zhang (2011), p13) that it is possible to express the inverse of a matrix in terms of its adjugate matrix (adj($\mathbf{A}$)) and determinant as

$$\text{adj}(\mathbf{A})_{k,l} = (-1)^{k+l} \det(\mathbf{A}_{-l;-k}) = \det(\mathbf{A})(\mathbf{A}^{-1})_{k,l}.$$

This makes that an elements of the inverse of a matrix $\mathbf{A}$ can be expressed as

$$(\mathbf{A}^{-1})_{k,l} = \frac{(-1)^{k+l} \det(\mathbf{A}_{-l;-k})}{\det(\mathbf{A})}. \tag{16}$$

For notational simplicity, also define

$$B_{h,j}^{k,l} = \sum_{\substack{i=1 \\ i \neq l}}^{d} \frac{(-1)^{i+j+k+l+\mathbb{1}\{j>k\}+\mathbb{1}\{i>l\}} \mathbf{A}_{h,i} \det(\mathbf{A}_{-j,-k;-i,-l})}{\det(\mathbf{A})}, \tag{17}$$

$$D_{k,l} = \frac{(-1)^{k+l+1} \det(\mathbf{A}_{-k;-l})}{\det(\mathbf{A})}. \tag{18}$$

Lemma 1 *Using the introduced notation and the above results on matrix algebra*

$$\frac{\partial}{\partial \mathbf{A}_{k,l}} \left[(\mathbf{x} - \boldsymbol{\mu}_a)^T (\mathbf{A}^{-1})_{\cdot,j} \right] = \begin{cases} D_{k,l} (\mathbf{x} - \boldsymbol{\mu}_a)^T (\mathbf{A}^{-1})_{\cdot,k} & \text{if } j = k \\ \sum_{\substack{h=1 \\ h \neq j}}^{d} B_{h,j}^{k,l} (\mathbf{x} - \boldsymbol{\mu}_a)^T (\mathbf{A}^{-1})_{\cdot,h} & \text{if } j \neq k. \end{cases} \tag{19}$$

The proof of Lemma 1 is given in the Supplementary material. This result is needed for both the expected score and the Fisher information matrix, where we require an expression for the left-hand side of (19). The following assumptions are needed:

(*N1*) $\gamma_{j,r} = \int_0^\infty s^{r-1} \frac{(f_j'(s))^2}{f_j(s)} ds < \infty \ \forall j \in \{1, \ldots, d\}$ and $r = 1, 2, 3$.
(*N2*) $\int_0^\infty s f_j'(s) ds = -\frac{1}{2}$ or $\lim_{s \to \infty} s f_j(s) = 0 \ \forall j \in \{1, \ldots, d\}$.

These assumptions (*N1*) and (*N2*) are quite mild. They, as well as Condition (*C3*), are satisfied for, for example, f_j standard normal, Student's-t, logistic or Laplace densities (see for example Gijbels et al. (2019), as well as Example 1). We formally state the results on the expected score and the Fisher information matrix in Propositions 5 and 6 respectively.

Proposition 5 *Suppose Assumption (N2) holds, then the expectation of the score vector for* $\mathbf{X}$ *with respect to the true underlying distribution is zero. i.e.*

$$E\left[\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}.$$

Proposition 6 *Suppose Assumptions (N1) and (N2) hold and denote by*

$$\mu_{j,r} = 2\int_0^\infty s^r f_j(s)ds \qquad \kappa_{j,1} = E_{Z_j}[Z_j] = \frac{1-2\alpha_j}{\alpha_j(1-\alpha_j)}\mu_{j,1}$$
$$\kappa_{j,2} = E_{Z_j}[Z_j^2] = \frac{(1-\alpha_j)^3+\alpha_j^3}{\alpha_j^2(1-\alpha_j)^2}\mu_{j,2},$$

for $j = 1,\ldots,d$ *and* $r = 1,2$. *Then the elements of the Fisher information matrix* $\mathbf{I}(\boldsymbol{\theta})_{i,j} = E\left[\left(\frac{\partial \ell(\boldsymbol{\theta};\mathbf{X})}{\partial \theta_i}\right)\left(\frac{\partial \ell(\boldsymbol{\theta};\mathbf{X})}{\partial \theta_j}\right)\right]$, $i,j = 1,\ldots,d$ *exist and are given by*

$$E\left[\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \alpha_k}\right)\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \alpha_l}\right)\right] = \begin{cases} \frac{2\left(\alpha_k^3+(1-\alpha_k)^3\right)\gamma_{k,3}-(1-2\alpha_k)^2}{\alpha_k^2(1-\alpha_k)^2} & \text{if } k=l \\ 0 & \text{if } k\neq l, \end{cases}$$

$$E\left[\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \alpha_k}\right)\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \mu_{a,l}}\right)\right] = -2(\mathbf{A}^{-1})_{l,k}\gamma_{k,2},$$

$$E\left[\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \alpha_k}\right)\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \mathbf{A}_{l,m}}\right)\right] = \begin{cases} \frac{D_{k,m}(1-2\alpha_k)(2\gamma_{k,3}-1)}{\alpha_k(1-\alpha_k)} & \text{if } k=l \\ 2\gamma_{k,2}\sum\limits_{\substack{h=1\\h\neq k}}^{d} B_{h,k}^{l,m}\kappa_{h,1} & \text{if } k\neq l, \end{cases}$$

$$E\left[\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \mu_{a,k}}\right)\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \mu_{a,l}}\right)\right] = \sum_{j=1}^{d} 2\alpha_j(1-\alpha_j)(\mathbf{A}^{-1})_{k,j}(\mathbf{A}^{-1})_{l,j}\gamma_{j,1},$$

$$E\left[\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \mu_{a,k}}\right)\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \mathbf{A}_{l,m}}\right)\right] = -2\sum_{\substack{j=1\\j\neq l}}^{d}\alpha_j(1-\alpha_j)(\mathbf{A}^{-1})_{k,j}\gamma_{j,1}\sum_{\substack{h=1\\h\neq j}}^{d} B_{h,j}^{l,m}\kappa_{h,1},$$

$$E\left[\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \mathbf{A}_{k,l}}\right)\left(\frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{X}\right)}{\partial \mathbf{A}_{r,s}}\right)\right]$$
$$= 2\sum_{\substack{j=1\\j\neq k,r}}^{d}\alpha_j(1-\alpha_j)\gamma_{j,1}\left[\sum_{\substack{m=1\\m\neq j}}^{d}\sum_{\substack{h=1\\h\neq j,m}}^{d} B_{m,j}^{k,l}\kappa_{m,1}B_{h,j}^{r,s}\kappa_{n,1} + \sum_{\substack{g=1\\g\neq j}}^{d} B_{g,j}^{k,l}B_{g,j}^{r,s}\kappa_{g,2}\right]$$
$$+\left[\sum_{\substack{q=1\\q\neq k}}^{d}\sum_{\substack{j=1\\j\neq q,r}}^{d} B_{j,q}^{k,l}B_{q,j}^{r,s}\right] + \begin{cases} D_{k,l}D_{k,s}(2\gamma_{k,3}-1) & \text{if } k=r \\ 0 & \text{if } k\neq r. \end{cases}$$

The proofs of both propositions are in the Supplementary Material.

Example 3 (Example 1 continued) For the first model introduced in Example 1, the values of the quantities $\gamma_{j,r}$, $\mu_{j,r}$ and $\kappa_{j,r}$ are given by

$$\begin{aligned}
&\gamma_{1,1} = \tfrac{1}{2} \quad \gamma_{2,1} = \tfrac{1}{6} \\
&\gamma_{1,2} = \tfrac{\sqrt{2}}{\sqrt{\pi}} \; \gamma_{2,2} = \tfrac{1}{6} + \tfrac{\ln(2)}{3} \\
&\gamma_{1,3} = \tfrac{3}{2} \quad \gamma_{2,3} = \tfrac{2}{3} + \tfrac{\pi^2}{18} \\
&\mu_{1,1} = \tfrac{\sqrt{2}}{\sqrt{\pi}} \; \mu_{2,1} = 2\ln(2) \quad \kappa_{1,1} = 2.1277 \quad \kappa_{2,1} = -2.8281 \\
&\mu_{1,2} = 1 \quad \mu_{2,2} = \tfrac{\pi^2}{3} \quad\quad \kappa_{1,2} = 12.4444 \; \kappa_{2,2} = 20.1818.
\end{aligned}$$

In this example, we have $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \mu_{a,1}, \mu_{a,2}, \mathbf{A}_{1,1}, \mathbf{A}_{2,1}, \mathbf{A}_{1,2}, \mathbf{A}_{2,2})^T$. The inverse of the Fisher information matrix becomes

$$\mathbf{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix}
0.4134 & -0.0000 & 42.2235 & 14.0745 & 13.2298 & 0.0000 & 4.4099 & 0.0000 \\
0.0000 & 0.4278 & -22.4351 & 35.8961 & 0.0000 & 2.8205 & 0.0000 & -4.5128 \\
42.2235 & -22.4351 & 6786.9789 & -917.4407 & 1286.1011 & -113.9967 & 554.4639 & 247.9866 \\
14.0745 & 35.8961 & -917.4407 & 5755.6372 & 791.2002 & 523.5225 & -395.1784 & -283.0692 \\
13.2298 & -0.0000 & 1286.1011 & 791.2002 & 583.4778 & 66.9827 & 24.1203 & 22.3276 \\
0.0000 & 2.8205 & -113.9967 & 523.5225 & 66.9827 & 132.5992 & -107.1723 & -25.5543 \\
4.4099 & 0.0000 & 554.4639 & -395.1784 & 24.1203 & -107.1723 & 280.6358 & -35.7241 \\
-0.0000 & -4.5128 & 247.9866 & -283.0692 & 22.3276 & -25.5543 & -35.7241 & 103.0884
\end{bmatrix}. \tag{20}$$

This matrix reveals that, for MLE, the hardest parameters to estimate in this model are $\mu_{a,1}$ and $\mu_{a,2}$. Consider as an example the parameter $\mathbf{A}_{2,2}$, then $I(\mathbf{A}_{2,2})^{-1} = 103.0884$. As is made clear in Theorem 1, this implies that for a sample of n observations, the asymptotic variance of $\mathbf{A}_{2,2}$ is given by $103.0884n^{-1}$.

We are now able to state the asymptotic normality result for the MLE. For completeness, in the following $(\mathbb{R}^d, \Omega, P)$ is a probability space. Denote with

$$\begin{aligned}
\boldsymbol{\Psi}_j(\mathbf{x};\boldsymbol{\theta}) &= \left[\frac{1}{2}\left(\frac{\partial^+ \ell(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_j} + \frac{\partial^- \ell(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_j}\right)\right] \\
\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}) &= \left(\boldsymbol{\Psi}_1(\mathbf{x};\boldsymbol{\theta}), \cdots, \boldsymbol{\Psi}_{d^2+2d}(\mathbf{x};\boldsymbol{\theta})\right)^T \\
\boldsymbol{\lambda}(\boldsymbol{\theta}) &= E\left[\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\theta})\right],
\end{aligned} \tag{21}$$

and

$$u(\mathbf{x};\boldsymbol{\theta},r) = \sup_{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\| < r} \|\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}^*) - \boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})\|.$$

In this, $\frac{\partial^+}{\partial\theta_j}$ and $\frac{\partial^-}{\partial\theta_j}$ denote the right-hand respectively left-hand derivative. The following two lemmas are needed. Their proofs are in the Supplementary Material.

Lemma 2 *$\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})$ as defined in (21) is measurable.*

Lemma 3 *Under Assumption (N2) and continuity of both $f_j(x)$ and $f_j'(x)$ on $\mathbb{R}\backslash\{0\}$, $\boldsymbol{\lambda}(\boldsymbol{\theta})$ is continuous in a neighborhood of $\boldsymbol{\theta}_0$.*

Theorem 1 *Suppose Assumptions* $(C1)-(C3)$ *and* $(N1)-(N2)$ *hold. Then the maximum likelihood estimator* $\widehat{\boldsymbol{\theta}}_n^{ML}$ *is asymptotically normally distributed with mean* $\mathbf{0}$ *and variance-covariance* $\mathbf{I}(\boldsymbol{\theta}_0)^{-1}$*, i.e.*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^{ML} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\right) \qquad \text{as } n \to \infty,$$

where $\mathbf{I}(\boldsymbol{\theta}_0)$ *is the Fisher information matrix with elements given in Proposition 6.*

3.3 Inclusion of other parameters

So far, it was assumed that $\boldsymbol{\kappa}$ is empty, i.e. f_j, $j = 1, \ldots, d$ does not come with extra parameters. In reality, that might not always be the case. Fortunately, when one or more of the univariate symmetric distributions f_j, $j = 1, \ldots, d$ used to generate the multivariate distribution comes with extra parameters, the obtained results can be directly extended. In this case the parameter vector becomes $\boldsymbol{\xi} = (\boldsymbol{\alpha}^T, \boldsymbol{\mu}_a^T, \mathrm{vec}(\mathbf{A})^T, \boldsymbol{\kappa}^T)^T$, with $\boldsymbol{\kappa} = (\boldsymbol{\kappa}_1^T, \ldots, \boldsymbol{\kappa}_d^T)^T$, where $\boldsymbol{\kappa}_j$ is the vector of additional parameters from f_j. The expression for the log-likelihood does not change much compared to (12) and is given by

$$\begin{aligned}
&\ell(\boldsymbol{\xi}, \mathbf{x}) = \\
&-\ln(|\det(\mathbf{A})|) + d\ln(2) + \sum_{j=1}^{d} \ln(\alpha_j(1-\alpha_j)) \\
&+\sum_{j=1}^{d} \Big[\mathbb{1}\left\{(\mathbf{x}-\boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j} \leq 0\right\} \ln\left(f_j\left(-(1-\alpha_j)(\mathbf{x}-\boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j}\right); \boldsymbol{\kappa}_j\right) \\
&+\mathbb{1}\left\{(\mathbf{x}-\boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j} > 0\right\} \ln\left(f_j\left(\alpha_j(\mathbf{x}-\boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j}\right); \boldsymbol{\kappa}_j\right)\Big]. \qquad (22)
\end{aligned}$$

This leads to the following

Theorem 2 *If the following conditions hold*

(E1) *Conditions* $(I1)-(I2)$ *hold.*

(E2) *Let* $\boldsymbol{\Xi}_R$ *be a compact subset of* $\boldsymbol{\Xi}$*, the parameter space of* $\boldsymbol{\xi}$*. Also assume that* $\boldsymbol{\xi}_0 \in \overset{\circ}{\boldsymbol{\Xi}}_R$*, with* $\overset{\circ}{\boldsymbol{\Xi}}_R$ *the interior of* $\boldsymbol{\Xi}_R$*.*

(E3) $\int_0^\infty |\ln f_j(s;\boldsymbol{\kappa}_j)|\, f_j(s;\boldsymbol{\kappa}_j)ds < \infty \ \forall j \in \{1, \ldots, d\}$ *and all* $\boldsymbol{\kappa}_j \in \mathcal{K}_j$*, their parameter space. In this* $f_j(s;\boldsymbol{\kappa}_j)$ *are the underlying univariate symmetric densities.*

(E4) $E\left[\frac{\partial \ell(\boldsymbol{\xi};\mathbf{X})}{\partial \boldsymbol{\xi}}\right]_{\boldsymbol{\xi}=\boldsymbol{\xi}_0} = \mathbf{0}.$

(E5) $E\left[\frac{\partial \ell(\boldsymbol{\xi};\mathbf{X})}{\partial \xi_i}\frac{\partial \ell(\boldsymbol{\xi};\mathbf{X})}{\partial \xi_j}\right]_{\boldsymbol{\xi}=\boldsymbol{\xi}_0} < \infty \qquad \forall i, j \in \{1, \ldots, d\},$

the maximum likelihood estimator

$$\widehat{\boldsymbol{\xi}}_n^{ML} = \left((\widehat{\boldsymbol{\alpha}}_n^{ML})^T, (\widehat{\boldsymbol{\mu}_a}_n^{ML})^T, (vec(\widehat{\mathbf{A}}_n^{ML}))^T, (\widehat{\boldsymbol{\kappa}}_n^{ML})^T\right)^T,$$

of the true parameter vector $\boldsymbol{\xi}_0$ *is asymptotically normally distributed with mean* $\mathbf{0}$ *and variance-covariance* $\mathbf{I}(\boldsymbol{\xi}_0)^{-1}$*, i.e.*

$$\sqrt{n}(\widehat{\boldsymbol{\xi}}_n^{ML} - \boldsymbol{\xi}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{I}(\boldsymbol{\xi}_0)^{-1}\right) \qquad \text{as } n \to \infty.$$

The proof of Theorem 2 is similar to that of Theorem 1 and therefore omitted. Note that $\mathbf{I}(\boldsymbol{\xi}_0)$ consists of $\mathbf{I}(\boldsymbol{\theta})$ extended with an additional block made up of the interactions between $\boldsymbol{\kappa}$ and $\boldsymbol{\theta}$. A classical example of a distribution involving other parameters is the Student's t-distribution. The extra parameters are the degrees of freedom, ν_j, involved with each different Student's t-distribution. For the Student's t-distribution, one can show that the conditions are satisfied. See the Supplementary Material. In a univariate setting, the details can be found in Gijbels et al. (2019).

4 Simulation study

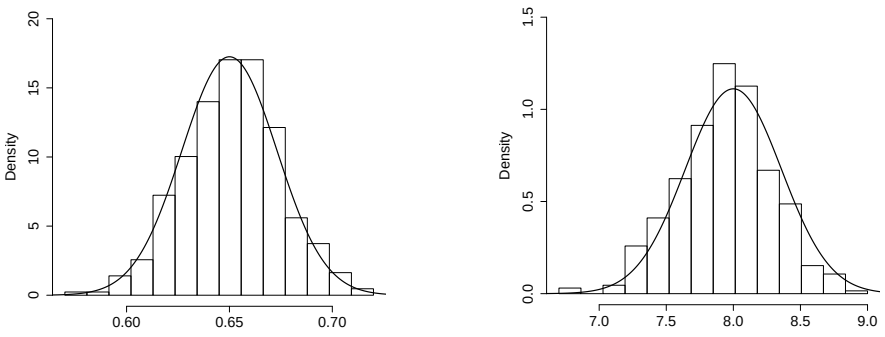
To estimate the parameters of the proposed distributions in (5), MLE is used. In order to maximize the log-likelihood, we rely on optimization software. Since the score functions can be discontinuous at certain points, even first order optimization algorithms might not be appropriate as the objective function may lack the necessary smoothness. For that reason, derivative free optimization is resorted to.

Several derivative free optimization routines are available in the `nloptr`-package (see Johnson (2018)). In order to chose an algorithm, the COBYLA-, NEWUOA-, BOBYQA- and Nelder-Mead-algorithms were taken into consideration. After extensive testing on models of different dimensionality and for different sample sizes, the BOBYQA-algorithm (Powell (2009)) was chosen to perform the fitting of the model. It showed the best and most consistent convergence results paired with a competitive computation time compared to its direct competitors. All computations are performed using the open-source software `R` and the therein available implementations.

As a warming up, 400 independent datasets are generated from the first model in Example 1 with sample size 800. Results concerning empirical bias (empirical bias$(\widehat{\theta}_j) = \frac{1}{400}\sum_{i=1}^{400}\widehat{\theta}_j^i - (\theta_0)_j$, with $\widehat{\theta}_j^i$ the parameter estimates based on the i-th realised dataset), estimated variance (estimated variance$(\widehat{\theta}_j) = \frac{1}{399}\sum_{i=1}^{400}(\widehat{\theta}_j^i - \frac{1}{400}\sum_{k=1}^{400}\widehat{\theta}_j^k)^2$) and asymptotic variance can be found in Table 2. Selected histograms for $\widehat{\alpha}_2$ and $\widehat{\mathbf{A}}_{2,2}$ are shown in Figure 2. The asymptotic variance is calculated using the expressions in Proposition 6, i.e. the asymptotic variance of the estimator $\widehat{\theta}_j$ is $n^{-1}\mathbf{I}(\boldsymbol{\theta}_0)^{-1}_{j,j}$. In this example the inverse of the Fisher information matrix is given by (20). Parameter estimates behave as expected under the developed theory. Asymptotic variance decreases at a n^{-1}-rate, empirical bias is approximately zero and the parameter estimates show clear normal behavior. These are thus the results expected under the presented asymptotic theory of Section 3.

Table 2: First model of Example 1: empirical bias, estimated variance and asymptotic variance of parameter estimates for sample size 800.

Sample size	Parameter	α_1	α_2	$\mu_{a,1}$	$\mu_{a,2}$	$\mathbf{A}_{1,1}$	$\mathbf{A}_{2,1}$	$\mathbf{A}_{1,2}$	$\mathbf{A}_{2,2}$
	Empirical bias	−0.0026	0.0013	−0.2691	−0.0361	−0.1653	0.0154	−0.0292	−0.0063
$n = 800$	Estimated variance	0.0006	0.0005	9.1021	7.6616	0.7676	0.1685	0.3514	0.1200
	Asymptotic variance	0.0005	0.0005	8.4837	7.1945	0.7293	0.1657	0.3508	0.1289



(a) $\widehat{\alpha}_2$. (b) $\widehat{\mathbf{A}}_{2,2}$.

Fig. 2: *First model of Example 1: histograms of selected parameter estimates for sample size 800. The solid curve indicates the asymptotic normal distribution of the corresponding parameter.*

4.1 Simulations

To investigate the finite sample performance of the MLE, several models are considered.

Model 1: a bivariate model consisting of a QBA-normal distribution (f_{Z_1}) and a QBA-Student's t-distribution (f_{Z_2}) with the following parameters

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.35 \\ 0.7 \end{pmatrix} \quad \boldsymbol{\mu}_a = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \mathbf{A} = \begin{bmatrix} 4 & 1 \\ -3 & 4 \end{bmatrix} \quad \nu_2 = 6\,.$$

A contourplot of this model is given in Figure 3 and the corresponding measures of asymmetry are listed in Table 3.

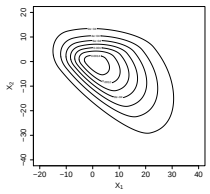


Table 3: Measures of asymmetry of Model 1.

$\beta_d(\mathbf{Y})$	$\mathbf{s}(\mathbf{Y})$	$\mathbf{b}(\mathbf{Y})$
8.1280	$\begin{bmatrix} 1.7366 \\ -2.2610 \end{bmatrix}$	$\begin{bmatrix} 0.9539 \\ -0.0103 \end{bmatrix}$

Fig. 3: Contourplot of Model 1.

Model 2: a six-dimensional model consisting of all components Z_j having QBA-Laplace distributions with the following parameters

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.2 \\ 0.24 \\ 0.28 \\ 0.32 \\ 0.36 \\ 0.4 \end{pmatrix} \quad \boldsymbol{\mu}_a = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} \quad \mathbf{A} = \begin{bmatrix} 10 & 0 & 5 & 0 & 1 & 0 \\ 0 & 10 & 1 & 0 & -4 & 2 \\ -5 & -1 & 10 & 0 & 6 & 0 \\ 0 & 0 & 0 & 10 & 0 & -2 \\ -1 & 4 & -6 & 0 & 10 & 0 \\ 0 & -2 & 0 & 2 & 0 & 10 \end{bmatrix}. \tag{23}$$

Measures of asymmetry for the second simulation model are given in Table 4. Note that the first components of both $\mathbf{s}(\mathbf{Y})$ and $\mathbf{b}(\mathbf{Y})$ show higher values.

Table 4: Measures of asymmetry of Model 2.

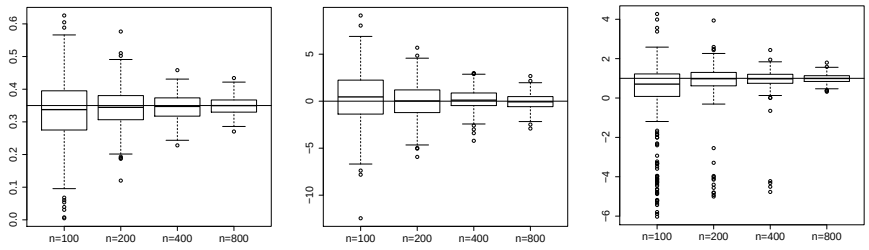
$\beta_d(\mathbf{Y})$	$\mathbf{s}(\mathbf{Y})$	$\mathbf{b}(\mathbf{Y})$
1.1011	$\begin{bmatrix} 0.5254 \\ 0.5665 \\ 0.6303 \\ 0.2338 \\ 0.2008 \\ 0.1087 \end{bmatrix}$	$\begin{bmatrix} 1.4178 \\ 0.4940 \\ 0.9030 \\ 0.1564 \\ 0.2899 \\ 0.1067 \end{bmatrix}$

For both models, sample sizes 100, 200, 400 and 800 are considered. In each of these settings, 400 independent random datasets are generated from which empirical bias and variance of the parameter estimates are computed. The approximate variance is compared to the corresponding theoretical value obtained from the asymptotic results presented in Section 3.

As for the optimization software, the following settings are used. The number of randomly generated starting values for the parameters is 40, the maximum number of iterations of the BOBYQA-algorithm is fixed at 35 000 to ensure convergence can take place. As a convergence criterion the first to occur between a relative change of 10^{-6} in the norm of the parameter values or an improvement of less than 10^{-9} in the log-likelihood is used. We summarize a selection of the results. A more detailed presentation of all simulation results can be requested from the authors.

Table 5: Model 1: empirical bias, estimated variance and asymptotic variance of parameter estimates for sample sizes 400 and 800.

Sample size	Parameter	α_1	α_2	$\mu_{a,1}$	$\mu_{a,2}$	$\mathbf{A}_{1,1}$	$\mathbf{A}_{2,1}$	$\mathbf{A}_{1,2}$	$\mathbf{A}_{2,2}$	ν_2
	Empirical bias	−0.0037	0.0016	−0.1796	0.0992	−0.0555	0.1160	−0.0688	−0.0989	1.0548
$n = 400$	Estimated variance	0.0015	0.0023	2.0624	1.2797	0.1165	0.5236	0.4074	0.2091	14.3817
	Asymptotic variance	0.0013	0.0010	1.8494	1.3707	0.1175	0.1078	0.1147	0.1194	2.9835
	Empirical bias	−0.0003	−0.0001	0.0334	−0.0530	−0.0160	0.0055	−0.0072	−0.0043	0.3072
$n = 800$	Estimated variance	0.0007	0.0006	1.0579	0.6803	0.0589	0.0534	0.0522	0.0625	2.1715
	Asymptotic variance	0.0006	0.0005	0.9247	0.6853	0.0587	0.0539	0.0573	0.0597	1.4917



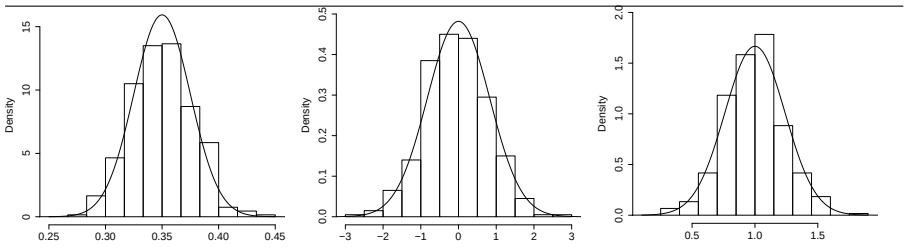
(a) $\widehat{\alpha}_1$. (b) $\widehat{\mu}_{a,2}$. (c) $\widehat{\mathbf{A}}_{1,2}$.

Fig. 4: *Model 1: boxplots of parameter estimates for sample size 100, 200, 400 and 800. The horizontal line represents the true parameter value.*

For Model 1, empirical bias, asymptotic variance and estimated variance are shown in Table 5. Boxplots of the estimates for $\widehat{\alpha}_1$, $\widehat{\mu}_{a,2}$ and $\widehat{\mathbf{A}}_{1,2}$ are shown in Figure 4. As can be seen, empirical bias is almost negligible for all parameters except the degrees of freedom of the Student's t-distribution. This is a problem inherent to the Student's t-distribution when sample sizes are small. The transition from $n = 400$ to $n = 800$ is an indication of the validity of the asymptotics. There is a further decrease in the empirical bias of the parameter estimates and estimated variances are much closer to their theoretical counterpart. This is most noticeable for the elements of the matrix $\mathbf{A}$. Although for $n = 400$ the variance of $\widehat{\mathbf{A}}_{1,1}$ is estimated excellent, all others are rather poorly estimated together with the degrees of freedom for the Student's t-distribution. The problem of this however, is that estimating degrees of freedom for Student's t-distributed random variables is hard in smaller samples. Even in the symmetric univariate case, the degrees of freedom parameter is often overestimated because heavy tails are hard to grasp from finite samples. For $n = 800$, all empirical variances except the one for the degrees of freedom approximate the theoretical variances well. Accuracy of variance estimates can thus be quite bad for smaller sample sizes. However, when sufficient data points are used, here 800, theory and reality are conform. This is nicely illustrated by Figure 5, where a histogram of the fitted parameters is plotted against the asymptotic distribution of the corresponding parameter.

Table 6: Model 2: empirical bias, estimated variance and asymptotic variance of parameter estimates for sample sizes 400 and 800.

Sample size	Parameter	α_3	α_5	$\mu_{a,3}$	$\mu_{a,5}$	$\mathbf{A}_{3,3}$	$\mathbf{A}_{3,5}$	$\mathbf{A}_{5,3}$	$\mathbf{A}_{5,5}$
$n = 400$	Empirical bias	0.0040	0.0000	0.4083	0.1977	−0.0683	−0.0432	0.0954	−0.1147
	Estimated variance	0.0009	0.0009	12.4362	11.3310	0.7867	0.3584	0.4238	0.4413
	Asymptotic variance	0.0005	0.0006	7.2579	6.5608	0.5511	0.2805	0.3076	0.4128
$n = 800$	Empirical bias	−0.0007	−0.0002	0.3041	0.0458	0.0055	−0.0359	0.0479	−0.0500
	Estimated variance	0.0003	0.0004	4.7258	5.0027	0.3147	0.1731	0.1831	0.2400
	Asymptotic variance	0.0003	0.0003	3.6289	3.2804	0.2756	0.1402	0.1538	0.2064



(a) $\widehat{\alpha}_1$. (b) $\widehat{\mu}_{a,2}$. (c) $\widehat{\mathbf{A}}_{1,2}$.

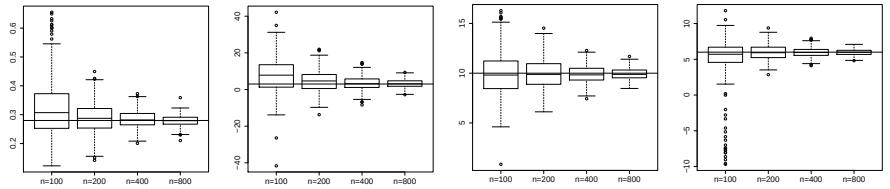
Fig. 5: *Model 1: histograms of parameter estimates for sample size 800. The solid curve indicates the asymptotic normal distribution of the corresponding parameter.*

The same conclusions can be drawn for Model 2. For selected parameter estimates, empirical bias, estimated and asymptotic variance can be found in Table 6. Boxplots of the same parameter estimates for all considered sample sizes are shown in Figure 6. The latter gradually center around the true parameter value as the sample size increases. The rate at which variance in parameter estimates drops corresponds to the desired n^{-1}-rate. Finite sample performance largely depends on the model considered, specifically the dimensionality of the model. Whereas for a bivariate model, 800 observations seem to suffice for the asymptotics to kick in, it is not enough for a six-dimensional model. The main reason for this is that the number of parameters, which is at least $d^2 + 2d$, so 48 in the six-dimensional model. It is natural that for a similar accuracy, a lot more observations are required.

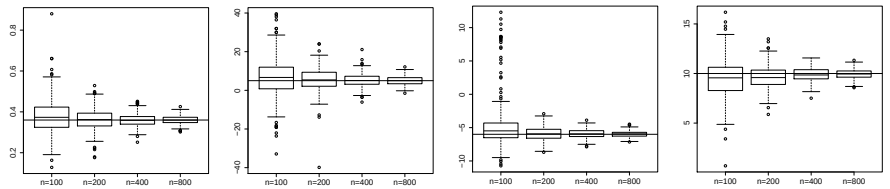
4.2 Impact of sample size and dimensionality

For practitioners, fitting a model should be possible within a reasonable timeframe. Of course, computing time is influenced by both the dimensionality of the problem and the sample size. To get a grasp at how these two factors impact computation time, two separate simulation cases have been studied. The first is aimed at exploring the impact of the sample size on the computation time. To this extent, a six-dimensional model consisting of only QBA-Laplace distributed univariate components with parameters given in (23) is used. For this model, 100 independent samples of size 2 000, 4 000, 6 000, 8 000 and 10 000 are generated. For each dataset 20 random starting points for the parameters are used. Boxplots of the resulting computation time (in seconds) are shown in Figure 7. The median computing time for sample size 2 000 is 410s whereas for sample size 10 000 it is 1793s. It thus seems that sample size has a linear impact on computation time, as could be expected. All simulations are run on a Dell Latitude 5590 with an Intel i5-8350U CPU clocked at 1.70GHz.

To assess the impact of the dimensionality d of the problem, a similar strategy is employed for 100 independent replicates with sample size 10 000



(a) $\widehat{\alpha}_3$. (b) $\widehat{\mu}_{a,3}$. (c) $\widehat{\mathbf{A}}_{3,3}$. (d) $\widehat{\mathbf{A}}_{3,5}$.



(e) $\widehat{\alpha}_5$. (f) $\widehat{\mu}_{a,5}$. (g) $\widehat{\mathbf{A}}_{5,3}$. (h) $\widehat{\mathbf{A}}_{5,5}$.

Fig. 6: *Model 2: boxplots of parameter estimates for sample size 100, 200, 400 and 800. The horizontal line represents the true parameter value.*

for models with dimensions 2, 4, 6, 8 and 10. As for the model on the impact of the sample size, for these models solely QBA-Laplace univariate components are used. For the skewing parameters $\boldsymbol{\alpha}$, d equally spaced values in the interval $[0.2, 0.4]$ are taken. The other parameter values are given by (24), with lower dimensional models as indicated by the dashed lines:

$$\boldsymbol{\mu}_a = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{pmatrix} \quad \mathbf{A} = \left[\begin{array}{cc:cc:cc:cc:cc} 20 & 0 & 5 & 0 & 1 & 0 & 2 & 0 & -3 & 0 \\ 0 & 20 & 1 & 0 & -4 & 2 & 0 & 1 & -2 & 3 \\ \hdashline -5 & -1 & 20 & 3 & 5 & 0 & 4 & 1 & 0 & -4 \\ 0 & 0 & -3 & 20 & 0 & -2 & -3 & 2 & -1 & 0 \\ \hdashline -1 & 4 & -5 & 0 & 20 & 0 & 0 & 0 & 0 & 2 \\ 0 & -2 & 0 & 2 & 0 & 20 & 1 & -3 & 2 & 0 \\ \hdashline -2 & 0 & -4 & 3 & 0 & -1 & 20 & 0 & 0 & 5 \\ 0 & -1 & -1 & -2 & 0 & 3 & 0 & 20 & 5 & -3 \\ \hdashline 3 & 2 & 0 & 1 & 0 & -2 & 0 & -5 & 20 & 0 \\ 0 & -3 & 4 & 0 & -2 & 0 & -5 & 3 & 0 & 20 \end{array}\right] \tag{24}$$

Each time, 20 random sets of starting values are used and the number of iterations is capped at 35 000 for dimensions 2, 4 and 6 and for dimensions 8 and 10 capped at 75 000 to make sure the algorithm can converge. Boxplots of the computation time can be found in Figure 8.

As expected, sample size impacts the computation time linearly. This is due to the log-likelihood being evaluated in more points, which doesn't change the complexity of the optimization. Dimensionality is a different story. As reported in Powell (2009), the BOBYQA-algorithm has a theoretical complexity of $\mathcal{O}(m^2)$, with m being the number of parameters. Since the number of parameters $(d^2 + 2d)$ of our model increases quadratically with the dimension d,

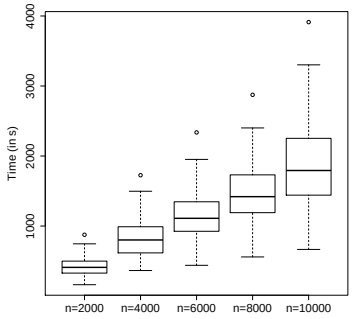


Fig. 7: *Impact of sample size on computation time.*

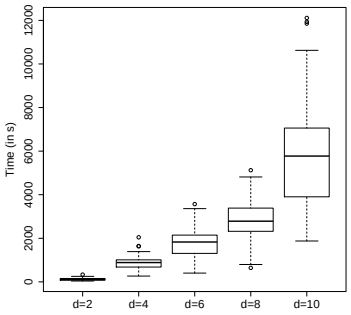


Fig. 8: *Impact of dimensionality on computation time.*

the complexity is expected to be of order $\mathcal{O}(d^4)$. The numerical results of this limited setting however, show a more quadratic behavior. It is also important to keep in mind that 20 different sets of starting values for the parameters are used.

5 Data applications

We present two data applications. As a benchmark, we use the current norm in multivariate asymmetric distributions: the skew-elliptical distributions (2), in particular the skew-normal and skew-t distribution. The comparison between our proposed distributions and the skew-elliptical ones is based on some goodness-of-fit criteria. For univariate goodness-of-fit, many test criteria are available. However, multivariate extensions of these goodness-of-fit tests are scarce. So in order to assess the goodness-of-fit, we rely on the Akaike's information criterion (AIC, Akaike (1974)) and a graphical goodness-of-fit diagnosis based on the depth-depth plot (DD-plot). The former criterion can be used to compare non-nested models which are fit using maximum likelihood estimation. Formally, the AIC is defined as

$$\mathrm{AIC}_n = -2\ell\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{ML}}; \mathbf{x}^{(n)}\right) + 2k,$$

in which $\ell\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{ML}}; \mathbf{x}^{(n)}\right)$ is the log-likelihood evaluated in the MLE and k the number of model parameters. When fitting multiple models using MLE, the model with the lowest AIC is considered to be the better model among the different candidate models.

The DD-plot as proposed in Liu et al. (1999), can be seen as a multivariate analogue of the well known quantile-quantile plot. As there is no unique way of defining quantiles in higher dimensions, instead statistical depth as defined in Zuo and Serfling (2000) is used. This provides an outward ordering of the data

based on some measure of centrality with respect to a distribution. In a way, statistical depth is thus an intuitive multivariate extension of quantiles. A DD-plot then compares the statistical depth of the data in the fitted distribution to that in its empirical distribution function. As an analytical expression for the depth of data in a certain distribution is in general not available, a numerical approximation is used. This approximation consists of calculating the depth of the data in a sufficiently large sample (here 10 000 observations are generated) from the fitted distribution. We then plot the depth of the data in itself against the depth of the data in the random sample to create the DD-plot. As a depth function, the halfspace depth (also called Tukey depth, Tukey (1975)) is used. This is defined in Zuo and Serfling (2000) as

$$D_H(\mathbf{X}|P) = \inf\left\{P(\mathcal{H}) : \mathcal{H} \text{ a closed halfspace, } \mathbf{X} \in \mathcal{H}\right\},$$

so the halfspace with the least probability mass containing $\mathbf{X}$. The sample version of the halfspace depth $D_H(\mathbf{z}, \mathbf{x}^{(n)})$ of a point $\mathbf{z}$ with respect to a sample $\mathbf{x}^{(n)}$ of size n with empirical distribution function P_n is given by (Struyf and Rousseeuw (1999))

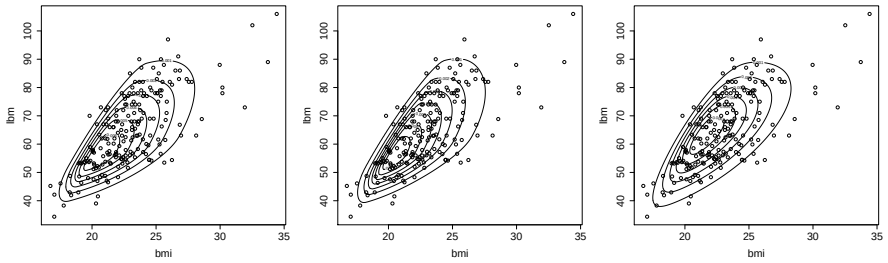
$$D_H(\mathbf{z}, \mathbf{x}^{(n)}, P_n) = \min_{\mathbf{u}\in\mathbb{R}^d, \|\mathbf{u}\|=1} \quad \frac{1}{n}\#\{i : \mathbf{u}^T\mathbf{x}^i \leq \mathbf{u}^T\mathbf{z}\}.$$

Halfspace depth is thus given by the minimal fraction of points of $\mathbf{X}^{(n)}$ contained in a halfspace that contains $\mathbf{z}$. Following this definition, it is clear that $D_H(\mathbf{z}, \mathbf{x}^{(n)})$ can only take on values in $[0, 0.5]$ and larger values imply that $\mathbf{z}$ lays closer to the center of the sample $\mathbf{x}^{(n)}$. If the fitted distribution provides a good fit, the point cloud of the DD-plot should be close to the 1:1 line.

As candidate models from the proposed family of distributions, we use all combinations of QBA-Laplace, QBA-normal, QBA-logistic and QBA-Student's t univariate components. As the ordering of these components is of no importance due to the relabeling problem explained in Section 3 and univariate components are allowed to be of the same type of distribution (e.g. two QBA-Laplace distributions). This is a combination with repetition but without ordering. Hence, for a d-dimensional dataset with m different options for the univariate components (in our setting $m = 4$), this leaves a total of $\binom{m+d-1}{d}$ possible models to fit. These are then compared to the benchmark using the above two criteria.

5.1 AIS-dataset

The first data example is often encountered in papers on multivariate asymmetric distributions, the AIS-dataset. The data, as depicted in Figure 9 concerns the body mass index (bmi) calculated as height (in cm) divided by squared mass (in kg) and lean body mass (lbm, expressed in kg), which is the body mass without fat mass, of 202 Australian athletes. The data is freely available in the `DAAG`-package in `R` and originates from Cook and Weisberg (1994).



(a) *Quantile based normal-normal.* (b) *Quantile based logistic-normal.* (c) *Skew-normal.*

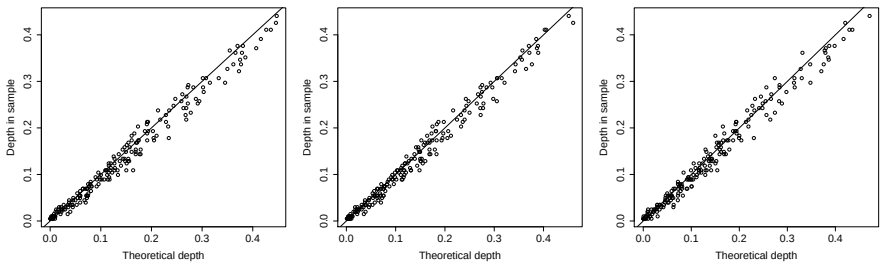
Fig. 9: *AIS-data with contour plots of some fitted models.*

Table 7: AIS-data. Fitted parameters for considered models. Standard deviations, based on the asymptotic normal distribution (Theorem 1) and the expression for the Fisher information matrix (Proposition 6), are between brackets.

		Quantile-based				Azzalini's bivariate	
		normal - normal	Student's t - normal	logistic - normal		skew-normal	skew-t
AIC		2432.1280	2429.0070	2426.718		2440.5220	2442.2150
	$\widehat{\alpha}_1$	0.2178 (0.0423)	0.2262 (0.0416)	0.2246 (0.0403)	$\widehat{\lambda}_1$	5.5153	5.2424
	$\widehat{\alpha}_2$	0.3020 (0.0480)	0.3002 (0.0479)	0.3005 (0.0479)	$\widehat{\lambda}_2$	−2.3022	−2.2349
	$\widehat{\mu}_{a,1}$	20.0532 (0.4256)	20.1088 (0.3789)	20.1003 (0.3737)	$\widehat{\beta}_1$	20.1355	20.1979
	$\widehat{\mu}_{a,2}$	54.6272 (2.2599)	54.4946 (2.1948)	54.5137 (2.1986)	$\widehat{\beta}_2$	61.7612	61.9651
	$\widehat{\mathbf{A}}_{1,1}$	0.7490 (0.1214)	0.6667 (0.1083)	0.4262 (0.0638)	$\widehat{\boldsymbol{\Omega}}_{1,1}$	16.116	14.8864
	$\widehat{\mathbf{A}}_{2,1}$	0.6493 (0.0978)	0.6274 (0.0927)	0.6305 (0.0929)	$\widehat{\boldsymbol{\Omega}}_{2,1}$	35.3676	32.6333
	$\widehat{\mathbf{A}}_{1,2}$	0.9029 (0.3755)	0.8514 (0.3107)	0.5402 (0.1974)	$\widehat{\boldsymbol{\Omega}}_{1,2}$	35.3676	32.6333
	$\widehat{\mathbf{A}}_{2,2}$	5.2251 (0.5445)	5.1827 (0.5445)	5.1894 (0.5443)	$\widehat{\boldsymbol{\Omega}}_{2,2}$	179.6722	171.7735
	$\widehat{\nu}$		7.3017 (3.5005)		$\widehat{\nu}$		51.0020

To this data the proposed distributions as well as a bivariate skew-normal and skew-t distributions are fitted. Fitted parameters for a bivariate QBA normal–normal, a QBA Student's t–normal, a QBA logistic–normal distribution, the bivariate skew-normal and the bivariate skew-t distribution are given in Table 7. The estimated standard errors between brackets are obtained from the asymptotic normality result established in Theorem 1, the expression for the elements in the Fisher information matrix provided in Proposition 6, and by substituting the parameters by their estimates.

DD-plots of three of the five fitted distributions are shown in Figure 10. The plots for the skew-t distribution are similar to these for the skew-normal and therefore not included. See also the estimated high value for the degrees of freedom ν in Table 7. A direct comparison, both visual from the DD-plots and based on AIC, between the quantile-based and skew-elliptical models reveals that they perform very similar. With only small differences in AIC and almost identical DD-plots, both types of distributions provide a good fit to the AIS-data. It terms of distribution itself, there are subtle differences between the



(a) *Quantile-based normal-normal.* (b) *Quantile based Student's t-normal.* (c) *Skew-normal.*

Fig. 10: *AIS-data. DD-plots for some fitted models.*

quantile-based and the skew-elliptical distributions as can be seen in Figure 9. The skew-normal model shows more elongated contours towards the lower left direction compared to the other two. It is also slightly more rounded at the top right of the data. Yet, despite the differences the quality of the fit is surprisingly similar.

As a brief mention, the best fitting QBA-distribution to the AIS-data is a QBA-logistic - QBA-normal one. In terms of DD-plot this gives similar results to the ones shown in Figure 10, but it has an AIC of 2426.718. The AIC of the other seven models not mentioned here are given in Table 8.

Table 8: AIS-data. AIC of seven fitted quantile-based models (not discussed).

		AIC
Quantile-based:	logistic - logistic	2434.9500
	normal - Laplace	2437.7540
	logistic - Laplace	2445.4860
	Laplace - Laplace	2458.8400
	Student's t - logistic	2429.1910
	Student's t - Laplace	2440.0490
	Student's t - Student's t	2431.4830

In this paper the focus is on a frequentist approach, allowing to establish statistical inference for the entire family. Of course any model in this context of linear combinations of QBA-distributions can also be fit using Bayesian estimation. To illustrate this we simply fit the QBA-logistic - QBA-normal model to the AIS data using Bayesian techniques. As the sample size is rather low, 4 MCMC chains are run, each consisting of a burn-in period of 20 000 iterations and a sampling period of 20 000 iterations. The final sample is then obtained by taking each fifth set of parameters from the sampling period. For this, the `rstan`-software package Stan Development Team (2021) is used.

Following Rubio and Steel (2015) priors are chosen to be vague priors, more specifically independent uniform priors. The range is based on the MLE

and the imposed restrictions on the parameter space. Taking these factors into account, the priors are $\mathcal{U}[0.05; 0.95]$ for both α_j, $j = 1, 2$; $\mathcal{U}[15; 35]$ for $\mu_{a,1}$; $\mathcal{U}[20; 110]$ for $\mu_{a,2}$; $\mathcal{U}[0; 10]$ for both $A_{1,1}$ and $A_{2,2}$; and finally $\mathcal{U}[-10; 10]$ for $A_{1,2}$ and $A_{2,1}$. As final estimate, the mean (with s.d.) and mode of the posterior distribution are reported.

A second set of priors is used based on the MLE reported in Table 7. This set consists of independent normal priors with mean the MLE rounded to two decimals and standard deviation twice that of the MLE, rounded to one decimal.

The resulting posterior distributions are depicted in Figures S.1 and S.2 in the Supplementary Material, for respectively the uniform and normal priors and the parameter estimates in Table 9. The prior distributions have an impact on the estimates and their precision. In particular the impact on the precision is very noticable from Figures S.1 and S.2. The Bayesian parameter estimates are of the same magnitude, but there are some striking differences between these and the MLE, mainly in the estimate for $A_{2,2}$, which is almost half as small as the MLE. This translates in a decent, but sub-optimal AIC-value, which is still on par with skew-normal and skew-t models, nevertheless, even though accurate MLE priors are used, the MCMC-algorithm still converges to a different optimum.

Table 9: AIS-data: Fitted parameters by a Bayesian approach using the previously mentioned sets of priors. For the mean estimator, the standard deviation is mentioned in brackets.

	Uniform priors		Normal priors	
Parameter	Mode estimator	Mean estimator	Mode estimator	Mean estimator
$\widehat{\alpha}_1$	0.2402	0.2392 (0.0383)	0.1833	0.1870 (0.0248)
$\widehat{\alpha}_2$	0.2916	0.3012 (0.0475)	0.3129	0.3161 (0.0377)
$\widehat{\mu}_{a,1}$	20.1332	20.2139 (0.3929)	19.9272	19.9316 (0.2831)
$\widehat{\mu}_{a,2}$	54.7715	55.1628 (2.2670)	55.4273	55.5103 (1.5965)
$\widehat{\mathbf{A}}_{1,1}$	0.7557	0.7802 (0.1108)	0.6314	0.6256 (0.0590)
$\widehat{\mathbf{A}}_{2,1}$	0.3911	0.3921 (0.0658)	0.4142	0.4195 (0.0496)
$\widehat{\mathbf{A}}_{1,2}$	0.7280	0.7711 (0.4443)	0.5593	0.5484 (0.1659)
$\widehat{\mathbf{A}}_{2,2}$	3.0055	3.0635 (0.3422)	3.2461	3.2271 (0.2875)
AIC	2440.939	2440.703	2442.359	2442.099

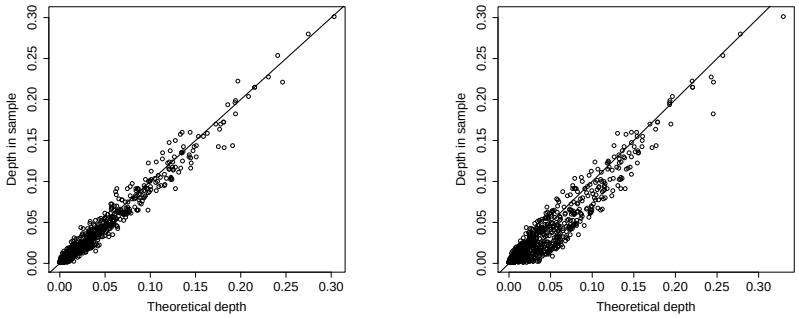
5.2 Pokémon data

In this data example, the base stats of 800 existing Pokémon up to generation 7 are used. The dataset is freely available from `https://www.kaggle.com/mlomuscio/pokemon`. The variables are: Hitpoints (HP), Attack, Defence, Special Attack, Special Defence and Speed. We thus have a six-dimensional dataset to which the proposed quantile-based distributions and both the skew-normal and skew-t distribution are fitted. For the quantile-based models, all

84 possibilities are fitted. In Table 10 the best 5 models, among the considered 84 models, according to AIC are presented and compared to the skew-normal and skew-t model. All 5 best quantile-based models have a lower AIC value than the skew-elliptical models. In fact, when we only take AIC into consideration, there are only 4 out of 84 quantile-based models that perform worse than the skew-t model and none that performs worse than the skew-normal model. A visual check of the fits is also provided in the form of DD-plots. These are only provided for the best fitting (based on AIC) quantile-based model (the quantile-based Laplace-Laplace-logistic-logistic-logistic-Student's t-model) and the skew-t model. The DD-plots can be found in Figure 11. Again, the quantile-based model provides a good fit to the data and clearly outperforms the skew-t model.

Table 10: Pokémon data. AIC for the 5 best performing quantile-based models, the skew-normal and the skew-t model.

Distribution	AIC
Quantile-based Laplace-Laplace-logistic-logistic-logistic-t	43867.86
Quantile-based Laplace-t-t-t-t-t	43875.64
Quantile-based t-t-t-t-t-t	43884.05
Quantile-based Laplace-logistic-t-t-t-t	43903.21
Quantile-based Laplace-Laplace-Laplace-logistic-logistic-t	43903.93
Skew-normal	44758.60
Skew-t	44397.44

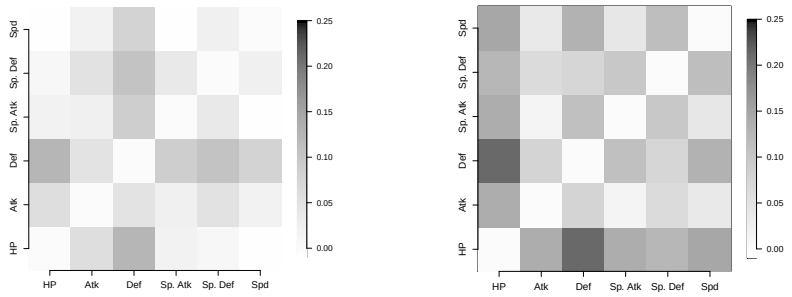


(a) *Best quantile-based model.* (b) *Skew-t model.*

Fig. 11: *Pokémon data. DD-plots for the best quantile-based and skew-elliptical distributions.*

Interesting to note is that Table 10 shows that the best performing quantile-based models all contain at least one or more Student's t-distributed compo-

nents. In all these models, the degrees of freedom are low for one component (or two when there are multiple components). The majority of the other components are light tailed distributions (or Student's t with high degrees of freedom). This provides an explanation for the better performance of our models as some, but not all variables in the data have heavy tails. The skew-elliptical distributions are less capable of capturing this different tail behavior and therefore perform worse.



(a) *Best quantile-based.* (b) *Skew-t.*

Fig. 12: *Pokémon data. Difference of the estimated correlation structure from the model and the empirical correlation structure.*

Another point of interest is the estimated dependence structure. For the quantile-based models the correlation is based on the covariance given in (7) using the estimated parameters. Figure 12 shows the heatmaps of the difference between the estimated and empirical correlation matrix for the best fitting quantile-based model and the skew-t model. For most variables correlation is estimated accurately. For the best fitting quantile-based model the correlation between defence and the other variables is less accurate whereas for the skew-t model, this is mainly the case for HP, but also in lesser extent for defence and special defence. In an attempt to represent Figure 12 with a single number that can be used to compare performances of models, one can consider the sum of squared differences between the estimated and the empirical correlation matrix

$$\sum_{i,j=1}^{6} \left(\widehat{\mathrm{Cor}}_{\mathrm{fitted}}(i,j) - \mathrm{Cor}_{\mathrm{emp}}(i,j) \right)^2 .$$

For the best fitting quantile-based model, this results in a value of 0.1047 whereas for the skew-t model it is 0.4209. This confirms that, overall, the correlation structure estimated by the quantile-based model deviates less from the empirical correlation than the estimated correlation structure from the skew-t model.

6 Conclusion and discussion

In this work, we study a family of asymmetric multivariate distributions based on an affine transformation of members of the quantile-based asymmetric family of distributions. The proposed family has an advantage over competing distributions in the form of added flexibility. This flexibility lies in the allowance of all types of distributions in the affine combination. This is contrary to other popular asymmetric multivariate distributions which rely on the skewing of a single elliptical multivariate distribution. We also show that under mild conditions, a maximum likelihood estimator is consistent and asymptotically normally distributed. A simulation study investigates the finite-sample performance of the MLE.

Asymptotic results for the maximum likelihood estimator for affine combinations of univariate random variables are, as far as we are aware, not published before. The results presented here, albeit restricted to the quantile-based asymmetric family of distributions, can readily be extended to incorporate other families of asymmetric univariate distributions. In doing this, a broad, general family of distributions is obtained. This is provided that statistical inference results for the univariate distributions exist. Other skewed distributions, like univariate skew-symmetric distributions can also be included as components for the linear combination. There is however, a trade-off to be made. The affine combination has great flexibility, but remains an affine combination. The dependency structure thereby imposed might be too simple to capture the dependency of the data in its full extent. Linear approximations generally provide decent results, but if the data is too complex, they might not suffice. So even though a good fit can be obtained, one has to reflect whether dependencies are modelled well enough.

It might be worthwhile to consider the QBA-family as margins together with a copula structure. Copulas provide a particular appealing flexible tool for constructing multivariate distributions, as they allow to combine, possibly in a dependent manner, marginals of a lower dimension (such as univariate ones). Rubio and Steel (2013), for example, use a Gaussian copula to model the dependence between two random variables. It is important however to go beyond Gaussian copulas, or more generally elliptical copulas, and general dimensions. There are ample of areas where such constructions are used. An example of this is in the construction of graphs, see for example Pircalabelu et al. (2017). The main challenge in such an approach in the context of asymmetric multivariate distributions lies again in providing theoretical support for statistical inference, in a unified manner, irrespectively of the specific lower dimensional asymmetric marginals and/or copula used. This is a topic of current research.

Appendix: Proofs of Propositions 2, 3 and 4, and of Theorem 1

Proof of Proposition 2

Suppose that $f(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\theta}^*)$ and that we know $\mathbf{Z}$ up to its parameters (e.g. Z_1 is of a QBA-logistic type etc.). We first prove that $\boldsymbol{\mu}_a$ is identifiable. By construction, f_{Z_j}, $j = 1, \ldots, d$ is unimodal with mode 0. Together with (5) this implies

$$\forall \mathbf{A}, \mathbf{A}^* \in \mathbb{R}^{d\times d}, \text{ non-singular} : \arg\max_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x};\boldsymbol{\theta}) = \arg\max_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x};\boldsymbol{\theta}^*) = \boldsymbol{\mu}_a.$$

Thus, $\boldsymbol{\mu}_a = \boldsymbol{\mu}_a^*$ and $|\det(\mathbf{A})| = |\det(\mathbf{A}^*)|$. Hence $\boldsymbol{\mu}_a$ is identifiable. Without loss of generality, we can assume that for the remainder of the proof, $\boldsymbol{\mu}_a = \mathbf{0}$.

The identifiability result we are aiming at is commonly referred to as uniqueness in the ICA-literature. In Eriksson and Koivunen (2004) necessary and sufficient conditions are provided for a noiseless ICA model ($\mathbf{X} = \mathbf{AZ}$) to be unique. These are

- There are no Gaussian sources. Or,
- If $\mathbf{A}$ has full column rank, there is at most one Gaussian source.

Since $\mathbf{A} \in \mathbb{R}^{d\times d}$ is non-singular, it has full column rank. If condition (*I1*) holds, the mixing matrix $\mathbf{A}$ is unique, i.e. identifiable up to a possible permutation and rescaling together with the accompanying permutation and rescaling of $\mathbf{Z}$. A location difference is not possible as $\mathbf{Z}$ does not contain a location parameter.

For, the scale ambiguity note that by (3)

$$\exists j = 1, \ldots, d : \alpha_j^* = 1 - \alpha_j \text{ and } (\mathbf{A}^*)^{-1}_{\cdot,j} = -(\mathbf{A}^{-1})_{\cdot,j} \Rightarrow f(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\theta}^*).$$

By restricting the sign of a single element of $(\mathbf{A}^{-1})_{\cdot,j}$ as in (*I2*), this problem can no longer occur. By

$$(\widetilde{\mathbf{I}}_j\mathbf{A})^{-1} = \mathbf{A}^{-1}(\widetilde{\mathbf{I}}_j)^{-1} = \mathbf{A}^{-1}\widetilde{\mathbf{I}}_j,$$

with $\widetilde{\mathbf{I}}_j \in \mathbb{R}^{d\times d}$ the identity matrix with -1 at $(\widetilde{I}_j)_{j,j}$, fixing the signs of the diagonal elements of $\mathbf{A}$ also suffices.

Since each of the Z_j's lacks a scaling parameter and none of the other parameters of Z_j affects the scaling in a linear way (otherwise it is considered a scaling parameter), any rescaling of $\mathbf{A}$ cannot be compensated by rescaling the parameters of $\mathbf{Z}$. Hence, $\mathbf{A}$ is identifiable up to a permutation. By the identifiability of each of the Z_j, also its parameters are uniquely determined up to the same possible permutation. Thus, $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ up to a possible permutation of $\mathbf{Z}$ and $\mathbf{A}$. Therefore the model is identifiable.

Proof of Proposition 3

We employ a proof by induction on the dimension of the matrix. For $d = 2$ this is trivial as $\mathbf{A}$ is invertible and thus has a non-zero determinant. Suppose the statement holds for any invertible $(d-1) \times (d-1)$-matrix. Consider the matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & E \end{bmatrix} \in \mathbb{R}^{d \times d},$$

with $\mathbf{B} \in \mathbb{R}^{(d-1)\times(d-1)}$, $\mathbf{C}, \mathbf{D}^T \in \mathbb{R}^{d-1}$ and $E \in \mathbb{R}$. Since $\mathbf{A}$ is invertible, it must hold that

$$\det(\mathbf{A}) = \det(\mathbf{B})E + \sum_{j=1}^{d-1}(-1)^{d+j} C_j \det((\mathbf{B}^*)_j) \neq 0, \tag{25}$$

where $(\mathbf{B}^*)_j = \begin{bmatrix} (\mathbf{B})_{-j,\cdot} \\ \mathbf{D} \end{bmatrix}$, so the $(d-1) \times (d-1)$-matrix where the j-th row of $\mathbf{B}$ is omitted and $\mathbf{D}$ is added. Now consider the following two cases.

1. $\det(\mathbf{B}) \neq 0$ and $E \neq 0$. By induction, the statement holds for $\mathbf{A}$.
2. $\{\det(\mathbf{B}) \neq 0$ and $E = 0\}$ or $\det(\mathbf{B}) = 0$. In this case, by (25), $\exists j \in \{1, \ldots, d-1\}$ such that $\det((\mathbf{B}^*)_j) \neq 0$ and $C_j \neq 0$. By swapping the j-th row of $\mathbf{A}$ with $(\mathbf{D}, E)$, the resulting matrix falls into case 1. This holds because the element replacing E is non-zero and the new matrix that takes the place of $\mathbf{B}$ is invertible as it is a row permutation of $(\mathbf{B}^*)_j$, thus conserving the non-zero determinant. Hence, the statement holds.

This concludes the proof as the above two cases contain all possible configurations of $\mathbf{A}$.

Proof of Proposition 4

The proof is largely based on similar arguments concerning the consistency of the maximum likelihood estimator for the univariate quantile-based asymmetric family of distributions: Theorem 3.3 in Gijbels et al. (2019), which in term uses Theorem 2.5 of Newey and McFadden (1994). The latter theorem states that under the following conditions (i) to (iv) the maximum likelihood estimator is weakly consistent, i.e. $\widehat{\boldsymbol{\theta}}_n^{\mathrm{ML}} \xrightarrow{P} \boldsymbol{\theta}_0$ for $n \to \infty$.

(i) If $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ then $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) \neq f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}_0)$.
(ii) The true parameter $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, with $\boldsymbol{\Theta}$ a parameter space which is compact.
(iii) The log-likelihood function $\ell(\boldsymbol{\theta}; \mathbf{x})$ is continuous at each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
(iv) It holds that $E[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\ell(\boldsymbol{\theta}; \mathbf{X})\|] < \infty$, where $\|.\|$ is the Euclidean norm.

Condition (i) is fulfilled by Proposition 2, in which the identifiability of the parameters is guaranteed by assumption (*C1*). Conditions (ii) and (iii) follow from respectively Assumption (*C2*) and the continuity of both the natural

logarithm and f_{Z_j}. So only condition (iv) remains to be checked. From (5) and (12), we have that

$$\begin{aligned} E\left[\|\ell\left(\boldsymbol{\theta};\mathbf{X}\right)\|\right] &= E\left[\left|-\ln\left|\det(\mathbf{A})\right| + \sum_{j=1}^{d} \ln f_{Z_j}((\mathbf{X}-\boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j};\boldsymbol{\eta}_j)\right|\right] \\ &\leq E\left[\left|\ln\left|\det(\mathbf{A})\right|\right| + \sum_{j=1}^{d} \left|f_{Z_j}((\mathbf{X}-\boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j};\boldsymbol{\eta}_j)\right|\right] \\ &= \left|\ln\left|\det(\mathbf{A})\right|\right| + \sum_{j=1}^{d} E\left[\left|f_{Z_j}((\mathbf{X}-\boldsymbol{\mu}_a)^T(\mathbf{A}^{-1})_{\cdot,j};\boldsymbol{\eta}_j)\right|\right] \\ &< \infty, \end{aligned}$$

where boundedness follows from the invertibility of $\mathbf{A}$ and Assumption (*C3*), as proven in Theorem 3.3 of Gijbels et al. (2019). Since the inequality holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_R$, condition (iv) is satisfied and consistency of the maximum likelihood estimator holds.

Proof of Theorem 1

The proof is largely based on Theorem 3 in Huber (1967), which handles asymptotic normality of maximum likelihood estimators for non-differentiable likelihood functions when consistency has been established.

Since consistency is shown in Proposition 4, only the following four conditions from Huber (1967) need to be fulfilled for the theorem to hold

I) For each fixed $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})$ is Ω-measurable and $\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})$ is separable. (See Assumptions A-1 p222 of Huber (1967).)
II) There exists a $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ for which $\boldsymbol{\lambda}(\boldsymbol{\theta}_0) = \mathbf{0}$.
III) There are strictly positive numbers a, b, c, r_0 such that
  i) $\|\boldsymbol{\lambda}(\boldsymbol{\theta})\| \geq a\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|$ for $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq r_0$.
  ii) $E[u(\mathbf{X};\boldsymbol{\theta},r)] \leq br$ for $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + r \leq r_0 \quad , \quad r \geq 0$.
  iii) $E[(u(\mathbf{X};\boldsymbol{\theta},r))^2] \leq cr$ for $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + r \leq r_0 \quad , \quad r \geq 0$.
IV) The expectation $E[\|\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\theta})\|^2]$ is finite.

These conditions are checked in a similar way as in the proof of Theorem 3.4 in Gijbels et al. (2019), which is already quite general. We start with condition I). By Lemma 2 $\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})$ is measurable. That $\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})$ is separable holds under the stated assumptions. Indeed, each of the component functions $\boldsymbol{\Psi}_j(\mathbf{x};\boldsymbol{\theta})$, for $j = 1,\ldots,d^2+2d$, is separable, and this is a finite number of functions. That each component function is separable follows from its continuity, except on a set with probability measure zero. Condition II) is met by Proposition 5, whereas for condition IV) we have by the definition of the Euclidean norm

$$E\left[\|\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\theta})\|^2\right] = E\left[\operatorname{trace}\left(\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\theta})\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\theta})^T\right)\right] = \operatorname{trace}\left(\mathbf{I}(\boldsymbol{\theta}_0)\right) < \infty,$$

where the finiteness follows from Proposition 6.

Remains to look into condition III). The key property in this is continuity of $\boldsymbol{\lambda}(\boldsymbol{\theta})$ in a neighborhood of $\boldsymbol{\theta}_0$, which holds by Lemma 3. The proof can be completed similarly as in Gijbels et al. (2019). For details, the reader is referred to that paper.

Acknowledgements The authors thank the anonymous reviewers for their valuable comments that led to an improvements of the work. The first and second author gratefully acknowledge support from the Research Fund KU Leuven [C16/20/002 project]. The third author was supported by Special Research Fund (Bijzonder Onderzoeksfonds) of Hasselt University [BOF14NI06].

References

Abtahi, A. and Towhidi M. (2013). The new unified representation of multivariate skewed distributions. *Statistics*, 47(1), 126–140.

Adcock, C. and Azzalini, A. (2020). A selective overview of skew-elliptical and related distributions and of their applications. *Symmetry*, 12(1).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Arellano-Valle, R. B., Gómez, H. W., and Quintana, F. A. (2005). Statistical inference for a general class of asymmetric distributions. *Journal of Statistical Planning and Inference*, 128(2):427–443.

Arnold, B. C., Castillo, E. and Sarabia, J. M. (2006) Families of Multivariate Distributions Involving the Rosenblatt Construction. *Journal of the American Statistical Association*, 101(476), 1652–1662.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178.

Azzalini, A. (2013). *The Skew-Normal and Related Families.* Institute of Mathematical Statistics Monographs. Cambridge University Press.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society, Series B*, 65(2): 367–389.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

Babić, S., Ley, C., and Veredas, D. (2019). Comparison and classification of flexible distributions for multivariate skew and heavy-tailed data. *Symmetry*, 11(10):1216.

Balakrishnan, N. and Captitanio, A. (2008). Discussion: The t family and their close and distant relations. *Journal of The Korean Statistical Society*, 37:305–307.

Bauwens, L. (2005). A new class of multivariate skew densities , with application to GARCH models. *Journal of Business & Economic Statistics*, 23(3):346–354.

Beckmann, C. F. and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2):137–152.

Bonhomme, S. and Robin, J.-M. (2009). Consistent noisy independent component analysis. *Journal of Econometrics*, 149(1):12 – 25.

Cardoso, J. F. and Souloumiac, A. (1993). Blind beamforming for non-gaussian signals. *IEEE Proceedings F - Radar and Signal Processing*, 140(6):362–370.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York (N.Y.).

Eriksson, J. and Koivunen, V. (2004). Identifiability, separability, and uniqueness of linear ica models. *IEEE Signal Processing Letters*, 11(7):601–604.

Fechner, G. (1897). *Kollektivmasslehre*. Engelmann.

Fernández, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Ferreira, J. T. A. S. and Steel, M. F. J. (2007). A new class of skewed multivariate distributions with application in regression analysis. *Statistica Sinica*, 17:505–529.

Gijbels, I., Karim, R., and Verhasselt, A. (2019). On quantile-based asymmetric family of distributions: Properties and inference. *International Statistical Review*, 87(3):471–504.

Gouriéroux, C., Monfort, A., and Renne, J. (2017). Statistical inference for independent component analysis: Application to structural VAR models. *Journal of Econometrics*, 196(1):111–126.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233, Berkeley, California. University of California Press.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *ICA by Maximum Likelihood Estimation*, chapter 9, pages 203–219. John Wiley & Sons, Ltd.

Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2012). On asymptotics of ICA estimators and their performance indices. *arXiv preprint arXiv:1212.3953*.

Jammalamadaka, S. R., Taufer, E., and Terdik, G. H. (2020). On multivariate skewness and kurtosis. *Sankhya A*, pages 1–38.

Johnson, S. G. (2018). The NLopt nonlinear-optimization package. url: http://ab-initio.mit.edu/nlopt.

Jones, M. C. (2008). The t family and their close and distant relations. *Journal of The Korean Statistical Society*, 37:293–302.

Jones, M. C. (2010). Distributions generated by transformation of scale using an extended Cauchy-Schl"omilch transformation. *Sankhya A*, 72:359–375.

Jones, M. C. (2016). On bivariate transformation of scale distributions. *Communications in Statistics-Theory and Methods*, 45(3), 577–588.

Kollo, T. (2008). Multivariate skewness and kurtosis measures with an application in ica. *Journal of Multivariate Analysis*, 99(10):2328–2338.

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001). *The Laplace Distribution and Generalizations*. Springer, New York.

Ley, C. and Paindaveine, D. (2010). Multivariate skewing mechanisms: a unified perspective based on the transformation approach. *Statistics and Probability Letters*, 80(23-24), pp.1685. Elsevier.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics*, 27(3):783–858.

Louzada, F., Ara, A., and Fernandes, G. (2017). The bivariate alpha-skew-normal distribution. *Communications in Statistics - Theory and Methods*, 46(14):7147–7156.

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

Miettinen, J., Taskinen, S., Nordhausen, K., and Oja, H. (2015). Fourth moments and independent component analysis. *Statistical Science*, 30(3):372–390.

Móri, T. F., Rohatgi, V. K., and Székely, G. (1994). On multivariate skewness and kurtosis. *Theory of Probability & Its Applications*, 38(3):547–551.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111 – 2245. Elsevier.

Ollila, E. (2010). The deflation-based fastica estimator: Statistical analysis revisited. *IEEE Transactions on Signal Processing*, 58(3):1527–1541.

Pircalabelu, E., Claeskens, G. and Gijbels, I. (2017). Copula directed acyclic graphs. *Statistics and Computing*, 27(1): 55–78.

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Technical Report, Department of Applied Mathematics and Theoretical Physics*.

Punathumparambath, B. (2012). The multivariate asymmetric slash Laplace distribution and its applications. *Statistica*, 72(2).

Rubio, F. J. and Steel, M. F. J. (2013). Bayesian inference for $P(X < Y)$ using assymetric dependent distibutions. *Bayesian Analysis*, 8(1):44–62.

Rubio, F. J. and Steel, M. F. J. (2014). Inference in two-piece location-scale models with Jeffreys priors. *Bayesian Analysis*, 9(1):1–22.

Rubio, F. J. and Steel, M. F. J. (2015). Bayesian modelling of skewness and kurtosis with Two-Piece Scale and shape distributions. *Electronic Journal of Statistics*, 9(2):1884–1912.

Stan Development Team (2021). RStan: the R interface to Stan. *R package* version 2.21.3, https://mc-stan.org/

Struyf, A. J. and Rousseeuw, P. J. (1999). Halfspace depth and regression depth characterize the empirical distribution. *Journal of Multivariate Analysis*, 69(1):135 – 153.

Tan, F., Tang, Y. and Peng H. (2015). The multivariate slash and skew-slash student t distributions. *Journal of Statistical Distributions and Applications*, 2(1):1–22

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531.

Villani, M. and Larsson, R. (2007). The multivariate split normal distribution and asymmetric principal components analysis. *Communications in Statistics—Theory and Methods*, 35(6):1123–1140.

Wallis, K. F. (2014). The two-piece normal, binormal, or double Gaussian distribution: Its origin and rediscoveries. *Statistical Science*, 29(1):106–112.

Zhang, F. (2011). *Matrix Theory: Basic Results and Techniques*. Springer Science & Business Media, New York, NY, 2nd edition.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482.