Semi-Parametric Estimation for Multivariate Skew-Elliptical Distributions

Jing Huang
ESMT European School of Management and Technology

Abstract

We investigate a class of skew-elliptical distributions which generalize the skewed distributions proposed by Azzalini and Dalla Valle (1996). These distributions are generated by elliptical contours plus an unknown distribution function which assigns weights or densities to these contours. Decomposing such a skewed distribution into the two constituent parts allows us to estimate them separately, with both parametric and non-parametric methods. For the one and two dimensional cases, we use simulated data to illustrate the estimation algorithm. In a parametric framework and under the assumption that the density assigning function has a closed form, our approach leads to a faster estimation process than traditional log likelihood estimation (MLE) through direct optimization or through expectation maximization (EM). Finally, we show the results from fitting our model to a real data set on body mass measurements of Australian athletes.
1 Introduction

A class of skewed distributions was proposed by Azzalini and Dalla Valle (1996). They use a conditional elliptical distribution to create skewness in the marginal distribution. Suppose, for example, that \((x, y)\) come from a two-dimensional normal distribution with parameters \(\mu = [0, 0]\) and \(\Sigma = [1, \delta; \delta, 1]\). Then the conditional distribution \(f(y|x > 0)\) follows a skewed distribution with parameter \(\lambda(\delta)\), which has density

\[
f(y, \lambda) = 2\phi(y)\Phi(\lambda y),
\]

where \(\phi\) and \(\Phi\) denote the \(N(0, 1)\) density and distribution function. Gupta (2003) constructed a multivariate skewed t-distribution by defining \(Y_j = X_j / \sqrt{W_j / v}\), where \(X = (X_1, \ldots, X_p)^T\) follows a skewed normal distribution and \(W \sim \chi^2_v\) is independent of \(X\). Fang and Zhang (1990) suggested a class of elliptical distributions given by \(x = Ru(k)\), where \(R > 0\) follows \(g(r)\) and \(u^T u = 1\) is a random variable uniformly distributed on the unit sphere. Since the density function of \(F(r)\) may not exist, they assumed that the elliptically distributed random variable \(x\) has characteristic function \(\phi(x)\). The unit sphere can be changed to an elliptical contour by a change of shape \(\Sigma\), together with a location parameter \(\mu: x = \mu + R\Sigma u_k\). A general notation for the elliptical contour is \(EC_k(\mu, \Sigma, \phi)\). Based on this property, Branco and Dey (2001) derived a general class of multivariate skewed-elliptical distributions by using a generalized density function \(g(r)\) instead of the normal distribution or the t distribution. Other parametric forms of skewed distributions are the skew-Cauchy distribution proposed by Arnold and Beaver (2000), and the extended skew-exponential power distribution suggested by Salinas et al. (2007), with a combined expression for both common ways of defining skewed distributions. Dey and Liu (2005) studied the density of different kinds of combinations for skewed distributed random variables.

For the non-parametric case, Genton and Loperfido (2005) suggested using a general skew function \(\pi(\Sigma^{-1/2}(x - \mu))\) to replace the cdf part for those definitions in Azzalini and Dalla Valle (1996) or Branco and Dey (2001). In Ma et al. (2005), a semiparametric estimation method is given; they prove that the estimator is locally efficient if the model that is used to estimate the parameters contains the true model. They also assume that the general unknown function \(\pi_0(x)\) has the form of \(\Phi\{m(x)\}\), where \(\Phi\) is the normal cdf and \(m(x)\) is an odd function. They used an odd polynomial \(P_K(x)\) of order \(K\) in order to estimate \(m(x)\), hoping that \(\Phi\{P_K(x)\}\) is close to \(\pi_0(x)\).

In this paper, we develop an approach for estimating the parameters \(\Sigma\) and \(\mu\) of the elliptical contour without knowing the explicit form of the distribution of \(g(r)\) defined according to Fang and Zhang (1990, p.59). In the later part of this paper, we also assume that the density assigning function is known, in order to compare our estimation approach with direct maximization of the likelihood and with the expectation-maximization algorithm. The elliptical contour appears in many common definitions, as for example the multivariate normal or the multivariate t distribution. In fact, any distribution function where the random variable \(X\) appears in the form of \((X - \mu)\Sigma^{-1}(X - \mu)\) is an elliptical distribution, and any skewed distribution defined on it can be estimated using the method discussed in this paper. We focus here on the one dimensional case, and make the single assumption that this unknown function follows an elliptical distribution. The definition of elliptical distribution is not unique, but all definitions are equivalent. One definition given by Fang and Zhang (1990) is as follows:

**Definition 1** If a \(n \times 1\) random vector follows an elliptical distribution, then it has a stochastic representation \(R \cdot u^{(n)}\), where \(R \geq 0\) is independent of \(u^n\), and \(u^n\) follows a uniform distribution on the sphere \(|u^{(n)}|^2 = 1\).

Thus an elliptical distribution has two independent parts, \(R \geq 0\) and \(u^{(n)}\). Here \(u^{(n)}\) is a sphere, and under some transformation (discussed in the next section) \(u^{(n)}\) can be an elliptical contour. Fig-
Figure 1 shows an example of skew distribution generated from the normal distribution. The colored lines between the two graphs indicate the correspondence between densities and elliptical contours.

In this paper, we make three contributions. First, we show how a multivariate distribution (skewed or non-skewed) can be decomposed into two parts: a $n$-dimensional elliptical sphere, and a univariate function $g(r)$. Second, we develop a non-parametric approach for estimating the function $g(r)$. Third, we add skewness to those elliptical spheres which satisfy the definition in Azzalini and Dalla Valle (1996). In the one-dimensional and two-dimensional cases, we give the moment conditions which can be used for parametric estimation.

2 Skewed elliptical contour

In this section we discuss the skewed elliptical contour in the $R^n$ case, which is the basic element for the skewed elliptical distribution. There are two steps in reaching the skewed elliptical contour from a uniformly distributed ball with radius $r$: $||u^n||^2 = r^2$. In the first step we follow Azzalini and Dalla Valle (1996) and calculate the conditional density of $u_{n-1}$ given $u_1 > 0$, where $[u_{n-1}; u_1]$ is a partition of $u^n$. In the second step we apply the linear transformation $x = T \cdot u_{n-1}$ to change the shape from a half ball to a skewed elliptical contour.

Let $u^{n-1}$ and $u^1$ be random variables uniformly distributed on the $R^{n-1}$ ball with radius $r$ and on the $R^1$ ball with radius $r$, respectively. Then the relationship between $u_{n-1}$, $u_1$ and $u^{n-1}$, $u^1$ is

$$
\begin{bmatrix}
  u_{n-1} \\
  u_1
\end{bmatrix} \overset{d}{\rightarrow} \begin{bmatrix}
  d_{n-1} u^{n-1} \\
  d_1 u^1
\end{bmatrix}
$$

where $(d_{n-1}^2, d_1^2) \sim B\left(\frac{1}{2}, \frac{n-1}{2}\right)$. So $r = R \cdot d_{n-1}$ has density:

$$
g(r) = \frac{2 \Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{n-1}{2}\right)} R^{2-n} r^{n-2} (R^2 - r^2)^{-\frac{1}{2}}.
$$

(1)

Fang and Zhang (1990) gave another relationship between the multivariate density $f_n(x)$ and $g(r)$:

$$
g(r) = \frac{2 \pi^{n/2}}{\Gamma(n/2)} r^{n-1} f_n(r^2).
$$

(2)
In our case, \( f_{n-1}(u_{n-1}) \) is the multivariate density for \( u_{n-1} \), thus
\[
f^R_{n-1}(u_{n-1}) = \frac{\Gamma(\frac{n}{2})}{\pi^{n/2}R^{2-n}}(R^2 - ||u_{n-1}||^2)^{-\frac{3}{2}}. \tag{3}
\]

Now we can proceed to the second step. Let \( [x_{n-1}; x_1] \) be a partition of \( x \), where \( x_{n-1} \in \mathbb{R}^n \) and \( x_1 \in \mathbb{R} \). Furthermore, \( x \) and \( u^n \) have the linear relationship
\[
x^{'}\Sigma^{-1}x = ||u^n||^2,
\]
where \( \Sigma \) is a positive semi-definite matrix. We need to find out the conditional density of \( x_{n-1} \) given \( x_1 > 0 \). Without loss of generality we assume that \( x \) is standardized, so the diagonal elements of \( \Sigma \) are all equal to one and \( \Sigma \) can be written as
\[
\Sigma = \begin{bmatrix} \Sigma_{11} & \delta \\ \delta' & 1 \end{bmatrix}.
\]
Let \( \Sigma_{112} = \Sigma_{11} - \delta \delta' \) and \( \Sigma_{112} = A'_{112}A_{112} \). Then \( \Sigma \) can be decomposed as
\[
\Sigma = \begin{bmatrix} I & \delta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \Sigma_{112} \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \delta' & 1 \end{bmatrix}.
\]
Thus
\[
\begin{bmatrix} x_{n-1} \\ x_1 \end{bmatrix} = \begin{bmatrix} A'_{112}u_{n-1} + \delta u_1 \\ u_1 \end{bmatrix}.
\]
\tag{4}

Then the conditional density function can be written as
\[
f^R(x_{n-1}|x_1 > 0) = f(A'_{112}u_{n-1} + \delta u_1|u_1 > 0).
\]

Since \( ||u_{n-1}||^2 + ||u_1||^2 = R^2 \) and \( x_{n-1} = A'_{112}u_{n-1} + \delta u_1 \), we can solve for \( u_1 \) and obtain
\[
u_1 = B(x_{n-1}) \pm \sqrt{B(x_{n-1})^2 - A \cdot (C(x_{n-1}) - R^2)}, \tag{5}
\]
where
\[
A = \delta'_{112} \Sigma_{112} - 1, \quad B(x_{n-1}) = \delta'_{112} x_{n-1}, \quad C(x_{n-1}) = x_{n-1}'_{1} \Sigma^{-1}_{112} x_{n-1}.
\]
Inserting (5) into (4), we can now solve for \( u_{n-1} \) as a function of \( x_{n-1} \). Let \( |J| = |\frac{\partial(u_{n-1})}{\partial x_{n-1}}| \) be the determinant of the Jacobian. From (5) and (3) it follows that the marginal density of \( x_{n-1} \) is
\[
f^R(x_{n-1}) = \frac{\Gamma(\frac{n}{2})}{\pi^{n/2}R^{2-n}}||u_{n-1}||^{-1}|J|. \tag{6}
\]
Denote the conditional density by \( f^R \). The whole density is determined by two parts: \( g(r) \) which is the density of \( R \), and, for each \( R \), the \( f^R \) defined on it. The whole density function can be written as
\[
f(x_{n-1}|x_1 > 0) = \int_0^\infty f^R(x_{n-1})g(R)dR. \tag{7}
\]
By using the properties from Fang and Zhang (1990, p.68-70), after some trivial calculations (see Appendix), we can derive the whole density function (7):
\[
f(x_{n-1}|x_1 > 0) = 2F^*(\lambda(x_{n-1} - \mu))f_1(x) \tag{8}
\]
where \( b = (x_{n-1} - \mu)'\Sigma^{-1}(x_{n-1} - \mu), \lambda = \delta'_{112} / \sqrt{1 - \delta'_{112} \delta} \) and
\[
F^*_b(x) = \frac{\int^x_{-\infty} f_n(x^2 + b)dx}{\int_{-\infty}^\infty f_n(x^2 + b)dx}
\]
is a cdf function, \( f_1(\cdot) \) is the one dimensional and \( f_n(\cdot) \) is the n-dimensional density in (2). This shows that our model is equivalent to the definition in Azzalini and Dalla Valle (1996).

In real life applications, \( g(r) \) is usually unknown, so we need to estimate it. If we know \( \Sigma_{11} \), let \( r = \sqrt{x_{n-1}^T\Sigma_{11}x_{n-1}} \), then we can estimate the density of \( r \) non-parametrically. Denote the estimated density function as \( \hat{g}_{n-1}(r) \), an estimation of the true density of the radius of the \((R - 1)\)-dimensional sphere. For deriving properties of conditional densities we need the density (1), therefore we have to find the link between \( \hat{g}_{n-1}(r) \) and \( g(r) \). We can again decompose \( u^{n-1} \) into \( u_{11}^{n-1} \) and \( u_{n_2}^{n-1} \), then \( u_1 \) and \( u_{11}^{n-1} \) should have the same marginal density. Denote the densities of \( u_1 \) and \( u_{11}^{n-1} \) on the \( R^n \) sphere and on the \( R^{n-1} \) sphere as \( f_1 \) and \( f_{11}^{n-1} \) respectively, which can be easily derived from equation (3). Then their relationship can be derived from the equality of the marginals:

\[
\int f_1^{n-1}\hat{g}_{n-1}(r)dr = \int f_1^{n}g(r)dr. \tag{9}
\]

In the following two sections we illustrate this structure in the one and two dimensional cases. We discuss in detail a semi-parametric estimation method who can be extended to the high dimensional case in a straightforward manner.

3 Skewed elliptical structure in the one dimensional case

3.1 Conditional density

When \( n = 2 \) and

\[
\Sigma = \begin{bmatrix}
1 & \delta \\
\delta & 1
\end{bmatrix},
\]

we have \( \Sigma_{11,2} = 1 - \delta^2 \) and \( A_{11,2} = \sqrt{1 - \delta^2} \). Denote \( x_{n-1} = x \) and \( u_{n-1} = u \). Then (5) becomes

\[
u_1 = \delta x \pm \sqrt{\delta^2 x^2 - (x^2 - (1 - \delta^2)R^2)}.
\tag{10}
\]

Using (4) we obtain

\[
|J| = \left| \frac{\partial u}{\partial x} \right| = \frac{\delta x \pm \sqrt{(1 - \delta^2)(R^2 - x^2)}}{\sqrt{R^2 - x^2}}.
\tag{11}
\]

Inserting (10) and (11) into (6), we obtain the density function

\[
f^R(x) = \frac{1}{\pi \sqrt{R^2 - x^2}}. \tag{12}
\]

We can write \([u, u_1]\) as a function of \([x, x_1]\). The condition \( x_1 > 0 \), or equivalently \( u_1 > 0 \), implies ruling out some areas where \( x \) can take values. Without loss of generality we assume \( \delta > 0 \), so if \( x < 0 \), from (10), the "\( \pm \)" can only take "\( + \)" and \( x \geq -\sqrt{1 - \delta^2}R \). If \( x > 0 \), when \( x \leq \sqrt{1 - \delta^2}R \), the "\( \pm \)" can only be "\( + \)". When \( x > \sqrt{1 - \delta^2}R \), the "\( \pm \)" can be both signs. The conditional density is then

\[
f^R(x) = \begin{cases}
\frac{1}{\pi \sqrt{R^2 - x^2}} & \text{if } -\sqrt{1 - \delta^2}R \leq x \leq \sqrt{1 - \delta^2}R \\
\frac{2}{\pi \sqrt{R^2 - x^2}} & \text{if } \sqrt{1 - \delta^2}R < x \leq R.
\end{cases}
\tag{13}
\]

If we also consider a location parameter \( \mu \), the conditional density becomes

\[
f^R(x) = \begin{cases}
\frac{1}{\pi \sqrt{R^2 - (x - \mu)^2}} & \text{if } -\sqrt{1 - \delta^2}R + \mu \leq x \leq \sqrt{1 - \delta^2}R + \mu \\
\frac{2}{\pi \sqrt{R^2 - (x - \mu)^2}} & \text{if } \sqrt{1 - \delta^2}R + \mu < x \leq R + \mu.
\end{cases}
\]

5
For each $R$, this density is weighted with probability $g(r)$, so the full density function $f(x)$ is $\sum f'^n(x) g(r_i) \Delta r$, or more specifically:

$$
f(x) = \begin{cases} 
\int_{-\mu}^{\infty} \frac{2g(r)}{\rho \sqrt{r^2 - (x - \mu)^2}} dr + \int_{\mu}^{\infty} \frac{g(r)}{\rho \sqrt{r^2 - (x - \mu)^2}} dr & \text{if } x \geq \mu \\
\int_{-\mu}^{\infty} \frac{g(r)}{\rho \sqrt{r^2 - (x - \mu)^2}} dr & \text{if } x < \mu.
\end{cases} $$  

\((14)\)

As an example, let us consider the normal distribution. Replace $x_1$ by $y$, let $\mu = 0$ without loss of generality, and recall that the two-dimensional normal distribution has density

$$
\frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix} \Sigma^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \right).
$$  

\((15)\)

Let $u = \sin(\theta)$, $u_1 = \cos(\theta)$, and $\beta = \arcsin(\sqrt{1 - \delta^2})/2$. Then we can write the relationship between $[x, y]$ and $[u, u_1]$ as:

$$
\begin{cases}
  x = R \cdot \sin(\theta - \beta) \\
y = R \cdot \sin(\theta + \beta)
\end{cases}
$$  

\((16)\)

From (16) and (15) it follows that

$$
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix} \Sigma^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \right) dxdy = \int_{0}^{2\pi} \int_{0}^{\infty} \frac{1}{2\pi \sqrt{1 - \rho^2}} |J| \cdot \exp\left(-\frac{1}{2} r^2 \right) drd\theta = \int_{0}^{\infty} r \cdot \exp\left(-\frac{1}{2} r^2 \right) dr,
$$

where $|J|$ is the Jacobian of the transformation from $dxdy$ to $drd\theta$. Comparing with (14), the $g(r)$ density in the case of the normal distribution is

$$
g(r) = r \cdot \exp\left(-\frac{1}{2} r^2 \right).$$  

\((17)\)

Thus the two-dimensional normal distribution can be fully determined by the elliptical contours (14) and the density (17). The same result for $g(r)$ can be obtained from (2). For the convenience of calculations, we therefore use the polar system (16) in the two dimensional case where $\sin(2\beta) = \sqrt{1 - \delta^2}$.

### 3.2 Distribution function

The goal of this paper is to estimate $g(r)$ so that we can then estimate $\mu$ and $\delta$. From (13) we see that as $\delta$ decreases ($\beta$ increases), the density moves from the interval $[-R, -R \cdot \sqrt{1 - \delta^2}]$ to $[R \cdot \sqrt{1 - \delta^2}, R]$. Since $\beta$ must have a symmetric distribution when $\delta = 0$ and $\beta = \pi/4$, the skewness derives from the movement of these densities. So by moving them back, the original symmetric distribution can be recovered. Note that we assume $\mu = 0$. Suppose that when $\delta = 0$, the density function of $y$ is $\hat{f}(x)$. Then for any $\xi \geq 0$, we must have

$$f(x + \xi) - \hat{f}(x + \xi) = \hat{f}(x - \xi) - f(x - \xi)
$$

$$\Rightarrow \quad \hat{f}(x + \xi) = \frac{1}{2} (f(x + \xi) + f(x - \xi)).$$

Following Li and Racine (2007), we can use the kernel density $\hat{k}(x,X)$ to estimate $f(x)$, then we can estimate the symmetric density $\hat{f}(y)$. Note that for some distributions (e.g. the multivariate normal), $x$ and $y$ are independent when $\delta = 0$, so $f_a(R^2) = f_x(\sqrt{R^2 - a^2}) \cdot f_y(\sqrt{a^2})$ for any $0 \leq a \leq c$. 

6
In our setting \( f_x = f_y = f \), so we can estimate \( f_x \) and then \( g(c) \) by giving \( a \) (e.g. \( a = 0 \)) in (2). The accuracy of this estimation depends on the accuracy of the kernel estimation. But for other distributions (e.g. the multivariate \( t \)), \( x \) and \( y \) are not independent, so we need to check those more general cases.

Since (14) has a singular point which can raise difficulties in estimation, we calculate the distribution function instead of the density function. On the left side of general cases.

\[
\int_{\mathbb{R}} \frac{1}{\pi} \arcsin \left( \frac{x - \mu}{r} \right) g(r) dr + \int_{\mathbb{R}} \frac{1}{\pi} 2\beta \cdot g(r) dr,
\]

where \( \mu - R\sin(2\beta) = x \). On the right side of \( \mu \), the distribution function is:

\[
1 - G(x) = \int_{\mathbb{R}} g(r) dr - \int_{\mathbb{R}} \frac{1}{\pi} \arcsin \left( \frac{x - \mu}{r} \right) g(r) dr - \int_{\mathbb{R}} \frac{1}{\pi} \arcsin \left( \frac{\bar{x} - \mu}{r} \right) g(r) dr - \int_{\mathbb{R}} \frac{1}{\pi} 2\beta \cdot g(r) dr
\]

where \( \bar{R} + \mu = x \) and \( R\sin(2\beta) + \mu = x \). From (18) or (19) we can easily see that, if \( x = \mu \),

\[
G(\mu) = \frac{2\beta}{\pi}.
\]

Since \( \delta \) is a one-to-one function of \( \beta \), this provides a fast estimation method of \( \delta \) if we knew \( \mu \), or of \( \mu \) if we knew \( \delta \), under the condition that we have a good estimate of \( G(\mu) \). Note also that when \( f(x) \) is given, \( \mu \) and \( \delta \) can be determined by each other, so there is one less parameter to estimate. From now on, we only focus on \( \mu \).

### 3.3 Moments of \( g(r) \)

**Proposition 1.** Let \( f(x) \) be the density function defined in (14). Then the following relationship stands:

\[
\int_{-\infty}^{\infty} (x - \mu)^n f(x) dx = W_n \cdot \int_{0}^{\infty} r^n g(r) dr,
\]

where

\[
W_n = \frac{\sin^{n-1}(2\beta) + \sin^{n-1}(-2\beta) \cdot \cos(2\beta) + \frac{n-1}{n} W_{n-2}}{n\pi}, \quad \text{if } n > 2
\]

and

\[
W_n = \frac{1}{\pi} \int_{-2\beta}^{2\beta} \sin^n \theta d\theta + \frac{2}{\pi} \int_{2\beta}^{\pi/2} \sin^n \theta d\theta.
\]

### 3.4 Parameter estimation

We know that \( f(x) \) can be uniquely determined by \( g(r) \). In this section, we provide a way to estimate the function \( g(r) \). Assume we know \( \mu \), then the density function of \( |x - \mu| \) is \( g_{n-1}(\cdot) \) in (9). Since in this section \( n = 2 \), we write \( g_{n-1}(\cdot) \) as \( g_1(\cdot) \). According to the result we derived above, in the one-dimensional case (9) becomes:

\[
g_1(x) = \int_{x}^{\infty} \frac{g(R)}{\pi \sqrt{R^2 - x^2}} dR.
\]

In the one-dimensional case, we write \( g(r) \) as \( g_2(r) \) to differentiate it from \( g_1(r) \). Figure 2 shows the good fitting performance of the estimated \( \hat{g}(r) \) compared to the true \( g(r) \), for the normal and \( t_3 \) distributions.

After obtaining \( g(r) \) from the pre-assumed value of \( \mu \), we need to check whether this value of \( \mu \) is correctly chosen. We insert the estimated \( \hat{g}(r) \) back to compute the density or distribution functions and compare the result with the empirical or kernel densities. Since we do not have the closed form of
Figure 2: The estimated $\hat{g}_2(R)$ versus the true $g_2(R)$ for the normal and $t_3$ distributions, with 10000 and 50000 simulated data.

$\hat{g}(r)$ and computing the density involves integrating, using $\hat{g}(r)$ in the integrals can lead to significant inaccuracy. Therefore we can use the density function in (8). In the one-dimensional case it becomes

$$f(x) = 2F_b^*(\lambda(x - \mu))f_1(x)$$  \hspace{1cm} (22)

where $b = (x - \mu)^2$, $\lambda = \delta / \sqrt{1 - \delta^2}$ and

$$F_b^*(x) = \frac{\int_{-\infty}^{x} f_2(t^2 + b)dt}{\int_{-\infty}^{\infty} f_2(t^2 + b)dt}$$

is a cdf function. In the case of the skewed normal density, $f_1(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$ and $f_2(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$ is the standard two-dimensional density function. Note that $f_1(x)$ is the one-dimensional standard normal density function since the diagonal elements of our $\Sigma$ are all ones. In section 5 we will briefly discuss the case when they are not.

Thus our algorithm for estimating $\mu$ is as follows:

a. pre-assign a starting value for $\mu$, and center the data $\tilde{x} = x - \mu$.

b. estimate $g_1(\tilde{x})$ using $\hat{g}_1(\tilde{x}) = k(|\tilde{x}|)$, where $k(\cdot)$ is kernel density or other non-parametric density function.

c. calculate $\hat{g}(r)$ according to (21).

d. calculate the distance $\hat{d} = \int |k(x) - \hat{f}(x)|dx$ using kernel or other non-parametric estimator and (22); write $\hat{d}$ as a function of $\mu$: $\hat{d} = D(\mu)$.

e. solve the minimization problem $\min_{\mu} D(\mu)$

The optimization process in step (e) is not the only possible choice and maybe not the optimal choice either. We can also apply other criteria here, for example maximizing the likelihood, minimizing the mean squared error (MSE), etc. We do not discuss these criteria further here, since our purpose is mainly to introduce a new model.
3.5 A simulation study

In order to investigate the performance of our estimation approach, we ran a small simulation study. We generated data \((x_i, y_i)\) with \(i = 1 \cdots 10000\), from the two-dimensional \(t\) and normal distributions:

\[
t \sim t([0, \mu], \left[ \begin{array}{cc} 1 & \delta \\ \delta & 1 \end{array} \right]) \quad \text{or} \quad N([0, \mu], \left[ \begin{array}{cc} 1 & \delta \\ \delta & 1 \end{array} \right]).
\]

We retained \(y_i\) when \(x_i \geq 0\), and applied the estimation algorithm outlined in the previous section. We repeated the simulation 50 times in each scenario. To assess the estimation performance, we computed the bias and mean squared error (MSE) of our estimates as follows:

\[
\text{bias } \mu = \frac{1}{50} \sum_{n=1}^{50} (\hat{\mu}_n - \mu),
\]

and

\[
\text{MSE } \mu = \frac{1}{50} \sum_{n=1}^{50} (\hat{\mu}_n - \mu)^2,
\]

where \(\hat{\mu}_n\) is the estimated value for the true \(\mu\). Similarly, we computed the bias and MSE for \(\hat{\rho}\).

Table 1 shows the bias and MSE values in different simulation scenarios.

<table>
<thead>
<tr>
<th>Model</th>
<th>True (\mu)</th>
<th>True (\delta)</th>
<th>Bias (\mu)</th>
<th>Bias (\delta)</th>
<th>MSE (\mu)</th>
<th>MSE (\delta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N)</td>
<td>2.8</td>
<td>0.1</td>
<td>-0.007</td>
<td>0.005</td>
<td>0.008</td>
<td>0.012</td>
</tr>
<tr>
<td>(N)</td>
<td>2.9</td>
<td>0.3</td>
<td>-0.09</td>
<td>0.097</td>
<td>0.022</td>
<td>0.022</td>
</tr>
<tr>
<td>(N)</td>
<td>1.3</td>
<td>0.7</td>
<td>-0.080</td>
<td>0.072</td>
<td>0.008</td>
<td>0.006</td>
</tr>
<tr>
<td>(N)</td>
<td>3.0</td>
<td>0.9</td>
<td>0.059</td>
<td>-0.0429</td>
<td>1e-4</td>
<td>5e-5</td>
</tr>
<tr>
<td>(t_3)</td>
<td>2.8</td>
<td>0.1</td>
<td>-0.027</td>
<td>0.030</td>
<td>0.014</td>
<td>0.017</td>
</tr>
<tr>
<td>(t_3)</td>
<td>2.9</td>
<td>0.3</td>
<td>-0.032</td>
<td>-0.034</td>
<td>0.0192</td>
<td>0.0238</td>
</tr>
<tr>
<td>(t_3)</td>
<td>1.3</td>
<td>0.7</td>
<td>0.009</td>
<td>-0.0105</td>
<td>0.003</td>
<td>0.0023</td>
</tr>
<tr>
<td>(t_3)</td>
<td>3.0</td>
<td>0.9</td>
<td>0.006</td>
<td>-0.004</td>
<td>-0.0007</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

Table 1: Estimation performance with data generated from the \(t\) distribution with 3 and 5 degrees of freedom and from the normal distribution

3.6 Comparison with other parametric models

In this section we compare our method with other popular parametric estimation methods. We assume that the data follows a skew-normal distribution. Direct maximization of the likelihood function does not generally work for this type of skewed data. According to Azzalini and Capitanio (1999), the biggest problem with likelihood maximization is that the expected Fisher information matrix becomes singular at \(\delta = 0\). So the expectation-maximization method (EM) is proposed instead in order to compute the MLE. To use EM in this setting, we assume that the whole data is \((x_1, x_2)\), but we only observe \(x_2\). Remember that \((x_1, x_2)\) has a two-dimensional half normal distribution: if \((y_1, y_2)\) follows a bivariate normal distribution, then \((y_1, |y_2|)\) follows a bivariate half normal distribution. So \((x_1, x_2)\) has density function given by

\[
f(x_1, x_2) = \frac{1}{\pi^{1/2}} \exp\left(-\frac{1}{2}(x_1 - \mu, x_2)^t \Sigma^{-1} (x_1 - \mu, x_2)^t\right)
\]

\[
= \frac{1}{\pi \sqrt{1 - \delta^2}} \exp\left(-\frac{1}{2(1 - \delta^2)} ((x_1 - \mu)^2 - \rho (x_1 - \mu)x_2 + x_2^2)\right),
\]

where \(\Sigma = \left[ \begin{array}{cc} 1 & \delta \\ \delta & 1 \end{array} \right]\).
The log-likelihood function for \((x_1, x_2)\) is
\[
\mathbb{L}(\mu, \hat{\delta}|x_1, x_2) = -n \cdot \log(\pi) - \frac{n}{2} \log(1 - \hat{\delta}^2) - \frac{1}{2(1 - \hat{\delta}^2)} \sum_{i=1}^{n} (x_{1i} - \mu)^2 \\
+ \frac{\hat{\delta}}{(1 - \hat{\delta}^2)} \sum_{i=1}^{n} (x_{1i} - \mu)x_{2i} - \frac{1}{2(1 - \hat{\delta}^2)} \sum_{i=1}^{n} x_{2i}^2.
\]

Then
\[
\mathbb{E}[\mathbb{L}(\mu, \hat{\delta}|x_1, x_2)|x_1, \hat{\mu}, \hat{\delta}] = -n \cdot \log(\pi) - \frac{n}{2} \log(1 - \hat{\delta}^2) - \frac{1}{2(1 - \hat{\delta}^2)} \sum_{i=1}^{n} (x_{1i} - \mu)^2 \\
+ \frac{\hat{\delta}}{(1 - \hat{\delta}^2)} \sum_{i=1}^{n} (x_{1i} - \mu)E(x_2|x_{1i}) - \frac{1}{2(1 - \hat{\delta}^2)} \sum_{i=1}^{n} E(x_2^2|x_{1i}).
\]

Note that \(f(x_2|x_1)\) has the density function
\[
f(x_2|x_1) = \frac{1}{\sqrt{2\pi(1 - \hat{\delta}^2)}} \exp\left(\frac{1}{2} \left(\frac{x_2}{\sqrt{1 - \hat{\delta}^2}} - \frac{\hat{\delta}(x_1 - \mu)}{\sqrt{1 - \hat{\delta}^2}}\right)^2\right), \Phi(\hat{\lambda}(x_1 - \mu)),
\]

so we shall have
\[
E(x_2|x_1) = \sqrt{1 - \hat{\delta}^2} \frac{A_1}{A_0} + \hat{\delta}(x_1 - \hat{\mu}),
\]
\[
E(x_2^2|x_1) = (1 - \hat{\delta}^2) \frac{A_1}{A_0} + 2\hat{\delta} \sqrt{1 - \hat{\delta}^2}(x_1 - \mu) \frac{A_1}{A_0} + \hat{\delta}^2(x_1 - \hat{\mu})^2,
\]

where
\[
A_j = \int_{-\hat{\lambda}(x_1 - \mu)}^{\hat{\lambda}(x_1 - \mu)} \rho_j \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} r^2\right) dr,
\]

and \(\hat{\lambda} = \hat{\delta} / \sqrt{1 - \hat{\delta}^2}\). Then the two steps of the EM algorithm are as follows:

- **E-step**: compute \(\mathbb{E}[\mathbb{L}(\mu, \hat{\delta}|x_1, x_2)|x_1, \hat{\mu}, \hat{\delta}]\) using (23) and (24).
- **M-step**: update \((\mu, \hat{\delta})\) by maximizing \(\mathbb{E}[\mathbb{L}(\mu, \hat{\delta}|x_1, x_2)|x_1, \hat{\mu}, \hat{\delta}]\).

The biggest drawback of EM is that the iterations require a lot of calculation time until convergence. Here we introduce a fast parametric estimation method based on the moment condition in Theorem 1. Consider the first two moment equations:
\[
\int_{-\infty}^{\infty} x f(x)dx - \mu = \frac{2\hat{\delta}}{2} \int_{0}^{\infty} r g(r)dr,
\]
\[
\int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu \int_{-\infty}^{\infty} x f(x)dx + \mu^2 = \frac{1}{2} \int_{0}^{\infty} r^2 g(r)dr,
\]

where \(\int_{-\infty}^{\infty} x^2 f(x)dx\) and \(\int_{-\infty}^{\infty} x f(x)dx\) can be estimated by \(\frac{1}{2} \sum_{i=1}^{n} x_i^2\) and \(\frac{1}{n} \sum_{i=1}^{n} x_i\) respectively. In the parametric setting we have \(g(r) = r \cdot \exp(-\frac{1}{2} r^2)\), so we can solve for \(\mu\) from (26), and then for \(\hat{\delta}\) from (25). Using higher moments is not encouraged since the tail has too much weight and we only have a limited number of data points.

Table 2 shows the results of performance comparisons between different methods. The results show that the MLE has the worst performance; the bias of \(\hat{\delta}\) is almost equal to the true \(\delta\) in each scenario, indicating that the MLE of \(\hat{\delta}\) is almost equal to \(\delta\) in each scenario. The EM algorithms performs quite well, but the calculation time is tremendous. Our estimation approach using moment conditions (SE) performs better than both MLE and EM, since it both requires small computation times and leads to accurate estimates. Note that the computation time for our estimation method is almost independent of data size (the only difference is in estimating the mean of \(x\) and \(x^2\)), while the computation time required by EM increases with increasing data size. Therefore, our method is to be preferred particularly when the data is large.
Table 2: Performance comparisons. SE stands for "skewed elliptical" method based on (25) and (26). EM stands for the "expectation maximization" algorithm and MLE stands for "numerical maximum likelihood estimation". The data generation processes are the same as in Table 1. For each algorithm in each scenario the simulation has been carried out 20 times.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Time (seconds)</th>
<th>Bias µ</th>
<th>Bias δ</th>
<th>MSE µ</th>
<th>MSE δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 1000, µ = 2.8, δ = 0.1</td>
<td>SE</td>
<td>0.427</td>
<td>-0.036</td>
<td>0.037</td>
<td>0.020</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1523</td>
<td>-0.056</td>
<td>-0.080</td>
<td>0.020</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>31.53</td>
<td>0.058</td>
<td>-0.080</td>
<td>0.008</td>
<td>0.014</td>
</tr>
<tr>
<td>n = 10000, µ = 2.8, δ = 0.1</td>
<td>SE</td>
<td>0.472</td>
<td>0.010</td>
<td>-0.010</td>
<td>0.004</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>16802</td>
<td>-0.017</td>
<td>0.022</td>
<td>0.004</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>267.01</td>
<td>0.082</td>
<td>-0.100</td>
<td>0.007</td>
<td>0.010</td>
</tr>
<tr>
<td>n = 1000, µ = 2.9, δ = 0.3</td>
<td>SE</td>
<td>0.392</td>
<td>0.009</td>
<td>-0.021</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1103</td>
<td>0.024</td>
<td>-0.040</td>
<td>0.016</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>25.77</td>
<td>0.231</td>
<td>-0.300</td>
<td>0.055</td>
<td>0.090</td>
</tr>
<tr>
<td>n = 10000, µ = 2.9, δ = 0.3</td>
<td>SE</td>
<td>0.431</td>
<td>-0.004</td>
<td>0.005</td>
<td>0.013</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>7700</td>
<td>-0.003</td>
<td>-0.003</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>252.56</td>
<td>0.240</td>
<td>-0.300</td>
<td>0.057</td>
<td>0.090</td>
</tr>
</tbody>
</table>

Table 3: Performance comparisons with the MLEs of the generalized skew-elliptical distribution. For each scenario, the simulation was conducted 20 times.

The results in Table 3 show that the MLE of GSE has some limitations. The true data generating process needs to have known closed form of the first part of (27). Mis-specification of the first part can lead to estimation failure. For the simulation, we assume that the first part of the density in (27) is normal. When the true density is $t_3$, the estimated values are quite far from the true parameter.
values, while our approach based on Fourier series can still lead to accurate estimates albeit with a large cost of computation time.

4 Skewed elliptical structure in the two dimensional case

In this section we provide the specific density function for the two-dimensional case, and briefly discuss parameter estimation. We have $n = 3$, and let $x_{n-1} = [x, y]$ and

\[
\Sigma = \begin{bmatrix}
1 & \rho & \delta_1 \\
\rho & 1 & \delta_2 \\
\delta_1 & \delta_2 & 1
\end{bmatrix}.
\]

According to (3), after simple calculations we obtain the density function as

\[
f^R(x, y) = \frac{1}{2\pi R \sqrt{(1-\rho^2)R^2 - (x^2 - 2\rho xy + y^2)}}.
\]

Here $B(x_{n-1})$ and $C(x_{n-1})$ become

\[
\begin{align*}
B(x, y) &= (\delta_1 - \rho \delta_2)x + (\delta_2 - \rho \delta_1)y \\
C(x, y) &= \frac{1}{k}[(1 - \delta_2^2)x^2 + 2(\delta_1 \delta_2 - \rho)xy + (1 - \delta_1^2)y^2],
\end{align*}
\]

where $k = (1 - \delta_1^2)(1 - \delta_2^2) - (\rho - \delta_1 \delta_2)^2$. With a similar argument as in the one-dimensional case, we conclude that the conditional density is given by

\[
f^R(x, y|u_1 > 0) = \begin{cases} 
\frac{2}{2\pi R \sqrt{(1-\rho^2)R^2 - (x^2 - 2\rho xy + y^2)}} & \text{if } B(x, y) > 0 \text{ and } C(x, y) > R^2 \\
\frac{1}{2\pi R \sqrt{(1-\rho^2)R^2 - (x^2 - 2\rho xy + y^2)}} & \text{otherwise.}
\end{cases}
\]

Figure 3: The graph shows the areas where the conditional density is defined. The black line is $B(x, y)$, the red elliptical contour is $C(x, y) = R^2$, and the blue elliptical contour is $x^2 - 2\rho xy + y^2 = (1-\rho^2)R^2$. If $B(x, y) > 0$, the conditional density moves the density from area II to I (areas between the red curve and the blue curve) and vice versa.
Figure (3) shows the area where a skewed elliptical contour is defined. We can see that, as in the one-dimensional case, the skewness is caused by the "transport" of the density. In the two-dimensional case the density is moved to area A from area B symmetrically to the origin. If there are location parameters, say \( \mu_1 \) for \( x \) and \( \mu_2 \) for \( y \), then the density transport is symmetric to \([\mu_1, \mu_2]\).

This source of skewness suggests recovering the symmetric density by moving the density back to the origin or the center. Once symmetry is recovered, we can estimate \( \rho \) with the contour.

### 4.1 Marginal density and moment conditions

From (28) and (29) we can derive the marginal density of \( x \) on each two-dimensional skewed sphere with radius \( R \) (assume that \( x \) is centered by \( x = x - \mu_1 \)):

If \( \delta_1 < 0 \), then

\[
f^{(R)}(x) = \begin{cases} 
\frac{1}{R} & \text{if } m_x \leq x < m_y \\
\frac{1}{R} \cdot \frac{\arcsin(\text{upper}(x)) - \arcsin(\text{lower}(x))}{2R} & \text{if } m_y \leq x < m_x \\
\frac{1}{R} \cdot \frac{\arcsin(\text{upper}(x)) - \arcsin(\text{lower}(x))}{2R} + \frac{\arcsin(\text{lower}(x)) + \pi/2}{2R} & \text{if } m_x \leq x < M_x \text{ and } \text{sign}(\rho)(\rho \delta_2)^2 < \text{sign}(\rho)(\delta_1 - \rho \delta_2)^2 \\
\frac{1}{R} \cdot \frac{\arcsin(\text{upper}(x)) - \arcsin(\text{lower}(x))}{2R} + \frac{\arcsin(\text{lower}(x)) + \pi/2}{2R} & \text{if } m_x \leq x < M_x \text{ and } \text{sign}(\rho)(\rho \delta_2)^2 \geq \text{sign}(\rho)(\delta_1 - \rho \delta_2)^2 \end{cases}
\]

If \( \delta_1 > 0 \), then

\[
f^{(R)}(x) = \begin{cases} 
\frac{1}{R} & \text{if } M_x \leq x < m_x \\
\frac{1}{R} + \frac{\arcsin(\text{upper}(x)) - \arcsin(\text{lower}(x))}{2R} & \text{if } M_x \leq x < M_y \\
\frac{1}{R} + \frac{\arcsin(\text{upper}(x)) - \arcsin(\text{lower}(x))}{2R} + \frac{\arcsin(\text{lower}(x)) + \pi/2}{2R} & \text{if } m_x \leq x < M_x \text{ and } \text{sign}(\rho)(\rho \delta_2)^2 < \text{sign}(\rho)(\delta_1 - \rho \delta_2)^2 \\
\frac{1}{R} + \frac{\arcsin(\text{upper}(x)) - \arcsin(\text{lower}(x))}{2R} + \frac{\arcsin(\text{lower}(x)) + \pi/2}{2R} & \text{if } m_x \leq x < M_x \text{ and } \text{sign}(\rho)(\rho \delta_2)^2 \geq \text{sign}(\rho)(\delta_1 - \rho \delta_2)^2 \end{cases}
\]

where \( M_x = R, m_x = -R, M_y = R \sqrt{1 - \delta_1^2}, m_y = -M_x, I = (\delta_2 - \rho \delta_1)^2/(\delta_1^2 + \delta_2^2 - 2 \rho \delta_1 \delta_2), M_x = R \sqrt{I}, M_y = -M_x \) and

\[
\begin{align*}
\text{upper}(x) &= (-\frac{\delta_2 - \rho \delta_1}{1 - \delta_1^2} x - \rho x + \frac{k}{2} (R^2 - \frac{x^2}{1 - \delta_1^2}) - \sqrt{(1 - \rho^2)(R^2 - x^2)})/\sqrt{(1 - \rho^2)(R^2 - x^2)} \\
\text{lower}(x) &= (-\frac{\delta_2 - \rho \delta_1}{1 - \delta_1^2} x - \rho x - \frac{k}{2} (R^2 - \frac{x^2}{1 - \delta_1^2}) + \sqrt{(1 - \rho^2)(R^2 - x^2)})/\sqrt{(1 - \rho^2)(R^2 - x^2)}.
\end{align*}
\]

To derive the moment function, we first introduce two propositions:

**Proposition 2.** If \( \delta_1 < 0 \), then the following equation holds:

\[
M_n(x) = \int_{-\infty}^{\infty} (x - \mu)^n (\int_{-\infty}^{\infty} f^{(R)}(x) g(R) dR) dx = \int_{-\infty}^{\infty} (\int_{-\infty}^{R} f^{(R)}(x) g(R) dR)dx.
\]

where \( h(x) \) is a linear function of \( x \) which can be either \((x - \mu)/\sqrt{1 - \delta_1^2}\) or \( \mu - x \) depending on whether \( x > \mu \) or \( x \leq \mu \).

The proof of this proposition is relatively straightforward: \( f^{(R)}(x) \) can be defined to be equal to zero when \( R < x \), so that \( \int_{-\infty}^{\infty} f^{(R)}(x) g(R) dR \) becomes \( \int_{-\infty}^{\mu} f^{(R)}(x) g(R) dR \). Also note that \( f^{(R)}(x) = 0 \) when \( x \) is outside the interval \([-R, \sqrt{1 - \delta_1^2}], \) so that we can write \( \int_{-\infty}^{\mu} (x - \mu)^n f^{(R)}(x) dx \) as \( \int_{-\infty}^{\mu} (x - \mu)^n f^{(R)}(x) dx \), which completes the proof.

**Proposition 3.** For densities on spheres with different radius, we have

\[
R \cdot f^{(R)}(Rx) = f^{(1)}(x).
\]

13
This proposition follows directly from (30).

With these two propositions, we have the following moments conditions ($\delta_1 < 0$):

$$
\int_0^\infty (x - \mu)^n f_{\text{margin}}(x) dx = \int_0^\infty \left( \int_{\tilde{R}(\tilde{x})}^{\infty} R^p f_{\tilde{R}}(\tilde{x}) d\tilde{x} \right) g(R) dR
= \int_{-1}^{\infty} \tilde{x} f_{\tilde{R}}(\tilde{x}) d\tilde{x} \int_0^\infty R^p g(R) dR
= M_n^{p+1} \int_0^\infty R^p g(R) dR,
$$

(31)

where $\tilde{x} = x - \mu_1$ and $M_n^{p+1}$ is the $n$-th moment of the random variable uniformly distributed on the sphere with radius one. This gives us the relationship between the moments of the random variable and the moments of $g(R)$. As in the previous section, if we know the closed form of $g(R)$, we can estimate the parameters through moment conditions.

### 4.2 Parameter estimation

Similar to subsection 3.4 we first need to estimate the function $g(r)$. According to (3) and (9) we have

$$
\int_1^\infty \frac{1}{\pi \sqrt{R^2 - x^2}} g^2_2(R) dR = \int_1^\infty \frac{1}{2R} g(R) dR.
$$

(32)

Then $g_2(R)$ can be estimated non-parametrically if we know $\mu_1$, $\mu_2$ and $\rho$, and assume that we have data $(x_i, y_i), i = 1, \cdots , n$.

- a. center the data: $(\hat{x}_i, \hat{y}_i) = (x_i - \mu_1, y_i - \mu_2)$.
- b. calculate the radius: $r_i = \sqrt{(\hat{x}_i^2 - 2\rho \hat{x}_i \hat{y}_i + \hat{y}_i^2)/(1 - \rho^2)}$
- c. calculate $g_2(R)$ in (32) non-parametrically (kernel or empirical method) using $r_i$, denote the estimate by $\hat{g}_2(R)$.

Once we have $\hat{g}_2(R)$, we can estimate the left hand side of (32), denote it by $\hat{H}(x)$. Then $g(R)$ can be estimated by

$$
g(R) = -2R \frac{d^2 \hat{H}(R)}{dR^2}.
$$

\[\text{Figure 4: The left plot gives } \hat{g}(R) \text{ versus the true } g(R) \text{ in the case of the normal distribution, while the right plot shows the case of the } t_3 \text{ distribution.}\]

There is a close relationship between $\mu_1$, $\mu_2$ and $\rho$ by which we can estimate $\rho$ through $\mu_1$ and $\mu_2$. We discussed how the skewness is caused by density transportation. Figure 5 shows the contours for the two-dimensional $t$ and normal densities with parametric forms $f_t(x'|\Sigma^{-1}x')$ and $f_\mu(x'|\Sigma^{-1}x')$ respectively. The red points represent the elliptical contour $x'|\Sigma^{-1}x' = 1$ whose shape is determined by the value of $\delta$. We see that the $t$ and normal distributions produce exactly the same shape of elliptical contours, even though they are generated by different functions $g(r)$. Therefore the estimation of $\rho$ is independent of the $g(r)$ function. In other words, without knowing the explicit form of $g(r)$ we can still estimate $\rho$. According to our theory, if $x, y$ follows a skew-elliptical distribution, then $x, y$ follows a conditional distribution. Suppose we also observe the third variable $z$, and the underlying 3-dimensional density is $f(x, y, z)$, then $x, y$ follows $f(x, y|z > 0)$. Apparently $2\mu_1 - x, 2\mu_2 - y$ follows $f(x, y|z \leq z)$ and $x, y \cup 2\mu_1 - x, 2\mu_2 - y$ follow the marginal density

$$
\int_{-\infty}^{\infty} f(x, y, z) dz.
$$
Denote \( \tilde{x}, \tilde{y} = \{ x, y \} \cup \{ 2\mu_1 - x, 2\mu_2 - y \} \). Since \( \{ \tilde{x}, \tilde{y} \} \) follow the non-skew marginal density, \( \rho \) can be easily estimated with the correlation of \( \{ \tilde{x}, \tilde{y} \} \). In other words, if

\[
\begin{bmatrix}
\tilde{x} \\
\tilde{y}
\end{bmatrix} = \begin{bmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{bmatrix},
\]

then \( \hat{\rho} \) is a good estimator of \( \rho \). If we replace \( \rho \) by \( \hat{\rho} \) in the process of estimating \( \hat{g}(R) \), we only need \( \mu_1 \) and \( \mu_2 \) to estimate \( \hat{g}(R) \). So from now on we write \( \hat{g}(R) = \hat{g}(R, \mu_1, \mu_2) \).

After deriving \( \hat{g}(R, \mu_1, \mu_2) \), we can recover the whole density of \( x, y \) by using (29). As in the one-dimensional case, in order to increase the accuracy, we use the difference of density at \((x, y)\) and \((2\mu_1 - x, 2\mu_2 - y)\), the absolute value of which, by (29) and the discussion of density transportation, should be equal to

\[
d(x, y) = \int_{B(x, y)} \frac{2\hat{g}(R, \mu_1, \mu_2)}{2\pi R \sqrt{1 - \hat{\rho}^2} R^2 - ((x - \mu_1)^2 - 2\hat{\rho}(x - \mu_1)(y - \mu_2) + (y - \mu_2)^2)} dR
\]

where

\[
\begin{align*}
\text{lb}(x, y) &= \sqrt{((x - \mu_1)^2 - 2\hat{\rho}(x - \mu_1)(y - \mu_2) + (y - \mu_2)^2)/(1 - \hat{\rho}^2)} \\
\text{ub}(x, y) &= \sqrt{C(x - \mu_1, y - \mu_2)} \quad (C(x, y) \text{ is defined in (28)).}
\end{align*}
\]

We can take different pairs \( x, y \) and compare \( d(x, y) \). Note that \( d(x, y) \) is a function of \( \mu_1, \mu_2, \delta_1, \delta_2 \), so we can write \( d(x, y) \) as \( d(x, y; \mu_1, \mu_2, \delta_1, \delta_2) \). One criterion of choosing parameters could be minimizing:

\[
D(\mu_1, \mu_2, \delta_1, \delta_2) = \int \int d(x, y; \mu_1, \mu_2, \delta_1, \delta_2) dx dy.
\]

In practice, we can choose sufficiently many pairs \( x, y \) and let \( D(\mu_1, \mu_2, \delta_1, \delta_2) \) be the summation of all \( d(x, y) \). The choice of criterion is again not unique; this can be, for example, minimizing the distance between the estimated density, maximizing the likelihood function, etc. In the next section we give estimation results from a simulated data set in both a semi-parametric and a parametric framework.

We generate a data set of \( n \) observations \( (x_i, y_i, z_i), i = 1 \cdots n \) from either the normal or the \( t_3 \) distribution, with \( |\mu_1| = 1, \mu_2 = 0.3, \mu_3 = 0 \) and

\[
\Sigma = \begin{bmatrix} 1 & 0.6 & 0.3 \\ 0.6 & 1 & -0.5 \\ 0.3 & -0.5 & 1 \end{bmatrix}.
\]

Then we only keep \( (x_i, y_i) \) where \( z_i \geq 0 \), so we have about \( \frac{n}{2} \) observations left. First we estimate the function \( g(r) \) non-parametrically. The results are shown in Tables 4 and 5.

Parametric estimation is much more straightforward. Just as in the one-dimensional case, we can use moment conditions to solve for all the parameters. Note that since we have all the information on \( g(R) \), we also know its characteristic function, so we can choose \( n \) in (31) as a non-integer number. However, the empirical estimation of the higher moment condition is not very accurate, so it is better not to use values for \( n \) larger than 3. The results of parametric estimation are shown in Table 6.
This can be estimated by adding, as we mentioned above, by letting
where \( \int \int \) to form the equation system:

\[
\int \int f^{-1}(x,y)dx\,dy = \text{cov}(x,y)
\]

It is easy to check that if \( f(x,y) \) is the whole density, the following equation stands:

\[
\int \int f^{-1}(x,y)dx\,dy = \int \int f^{-1}(x,y)\int_{0}^{R^{2}} g(R)\,dR.
\]  

This can be estimated by \( \frac{1}{2} \sum_{i=1}^{n} x_i y_i \), quite accurately. So we can choose the first and second moments and the covariance condition (34) to form the equation system:

\[
\begin{align*}
\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_1)(y_i - \mu_2) &= \int \int f^{-1}(x,y)dx\,dy \int_{0}^{R^{2}} g(R)\,dR \\
\frac{1}{n} \sum_{i=1}^{n} x_i y_i &= \int \int f^{-1}(x,y)\int_{0}^{R^{2}} g(R)\,dR \\
\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_2)^2 &= \int \int f^{-1}(x,y)dx\,dy \int_{0}^{R^{2}} g(R)\,dR \\
\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_2)^2 &= \int \int f^{-1}(x,y)\int_{0}^{R^{2}} g(R)\,dR \\
\frac{1}{n} \sum_{i=1}^{n} x_i (y_i - \mu_2) &= \int \int f^{-1}(x,y)dx\,dy \int_{0}^{R^{2}} g(R)\,dR \\
\frac{1}{n} \sum_{i=1}^{n} y_i (x_i - \mu_1) &= \int \int f^{-1}(x,y)\int_{0}^{R^{2}} g(R)\,dR
\end{align*}
\]

These five equations can then be used to solve for the five parameters. Further moment conditions can also be added, as we mentioned above, by letting \( n \) be non-integer. The estimation results are showed in Table 6.
In this section we show the results of applying our model to the analysis of a dataset of body mass index measurements on Australian athletes. So far, we have assumed that the diagonal elements of $\Sigma$ are all one, but in reality this is not often the case. Therefore, we first discuss how the model changes when the variance is not one.

### 5.1 Adding variance parameters

With small modifications to $\Sigma$ in subsection 3.1, it becomes:

$$\Sigma = \begin{bmatrix} \Delta^2 & \delta \\ \delta & 1 \end{bmatrix} = \begin{bmatrix} \Delta & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & \delta/\Delta \\ \delta/\Delta & 1 \end{bmatrix} \cdot \begin{bmatrix} \Delta & 0 \\ 0 & 1 \end{bmatrix}. \tag{35}$$

Thus, if a univariate random variable $x$ is generated by a skew-elliptical distribution as defined in previous sections with $\mu$ and $\Sigma$ as in (35), then $(x - \mu)/\Delta + \mu$ follows another skew-elliptical distribution with the same density of radius $g(r)$, the same $\mu$ and new $\tilde{\Sigma}$ given by:

$$\begin{bmatrix} 1 & \delta/\Delta \\ \delta/\Delta & 1 \end{bmatrix}.$$

Unfortunately, since we do not know $\mu$, we can not perform the standardization by subtracting the mean and dividing by the standard deviation. We need to treat $\Delta$ as a new parameter. After some straightforward calculations, the first two moment conditions (25) and (26) become:

$$\begin{align*}
\int_{-\infty}^{\infty} (x - \mu)f(x)dx &= \frac{2\delta}{\pi} \int_{0}^{\infty} r \cdot g(r)dr \\
\int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx &= \frac{\Delta^2}{2} \int_{0}^{\infty} r^2 \cdot g(r)dr.\tag{36}
\end{align*}$$

In parametric settings, the function $g(r)$ is already known. In non-parametric settings, $g(r)$ is not known and we can use the same method described in subsection 3.4 to estimate $g(r)$. We can then estimate $\mu$, $\delta$, and $\Delta$ as follows:

a. write $\delta$ and $\Delta$ as functions of $\mu$ according to (36).

b. write the density function $\tilde{f}$ as a function of $\mu$. In parametric settings, we can use (7) to calculate the density. In non-parametric settings, we can use the estimated $\hat{g}(r)$ and (14). By using (14), we need to perform a transformation from $x$ to $x/\Delta$.

c. compute the difference between the estimated density and the kernel density $k(x)$. This is given by $d(\mu) = \int [k(x) - \tilde{f}(x)]dx$.

d. solve for $\mu$ by minimizing $d(\mu)$ and then solve for $\delta$ and $\Delta$ from the moment conditions.

We conduct two simulations studies to show how our model performs with simulated data. We generated 10000 data points $\{x_i\}, i = 1 \cdots 10000$ with $\mu = 10$, $\delta = 4.5$, and $\Delta = 5$ assuming both the normal and the $t_3$ distributions, and we estimated the parameters using both the parametric and the non-parametric methods described above. This scenario was repeated 50 times, and the resulting bias and mean squared errors (MSE) are reported in Tables 7 and 8.

In the non-parametric framework, there is a large bias for parameters $\delta$ and $\Delta$. 

### Table 6: Parametric estimation result assuming $g(r)$ is known.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Computing time</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\delta_1$</th>
<th>$\delta_2$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>1648s (total)</td>
<td>true</td>
<td>-1</td>
<td>3</td>
<td>0.3</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bias</td>
<td>-5.3e-4</td>
<td>3.9e-4</td>
<td>8e-4</td>
<td>-1.2e-4</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>1.2e-5</td>
<td>7.1e-6</td>
<td>2.1e-5</td>
<td>6.3e-6</td>
<td>1.1e-5</td>
</tr>
<tr>
<td>$t_3$</td>
<td>2931s (average)</td>
<td>true</td>
<td>-1</td>
<td>3</td>
<td>0.3</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bias</td>
<td>0.0183</td>
<td>0.0095</td>
<td>0.0244</td>
<td>0.0117</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0042</td>
<td>0.0022</td>
<td>0.0039</td>
<td>0.0018</td>
<td>4.6e-4</td>
</tr>
</tbody>
</table>
In the two-dimensional case, the decomposition becomes:

\[
\begin{bmatrix}
\Delta_1^2 & \rho & \delta_1 \\
\rho & \Delta_2^2 & \delta_2 \\
\delta_1 & \delta_2 & 1
\end{bmatrix}
= 
\begin{bmatrix}
\Delta_1 & 0 & 0 \\
0 & \Delta_2 & 0 \\
0 & 0 & 1
\end{bmatrix}
\cdot 
\begin{bmatrix}
1 & \frac{\rho}{\Delta_1} & \frac{\delta_1}{\Delta_1} \\
\frac{\rho}{\Delta_2} & 1 & \frac{\delta_2}{\Delta_2} \\
\frac{\delta_1}{\Delta_1} & \frac{\delta_2}{\Delta_2} & 1
\end{bmatrix}
\cdot 
\begin{bmatrix}
\Delta_1 & 0 & 0 \\
0 & \Delta_2 & 0 \\
0 & 0 & 1
\end{bmatrix}.
\]

The estimation process is similar to that discussed above.

### 5.2 Estimation with real data

We next illustrate the use of our model for the analysis of a dataset containing the body mass index (BMI) of Australian athletes. The data was reported by Cook and Weisberg (1994), and it includes BMI measurements for 102 male and 100 female athletes. First, we assume that the BMI variable has a skew-normal distribution or a skew-t distribution. Table 9 gives the parameter estimates in the different cases.

<table>
<thead>
<tr>
<th>Model</th>
<th>True $\mu$</th>
<th>True $\delta$</th>
<th>True $\Delta$</th>
<th>Bias $\mu$</th>
<th>Bias $\delta$</th>
<th>Bias $\Delta$</th>
<th>MSE $\mu$</th>
<th>MSE $\delta$</th>
<th>MSE $\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>10</td>
<td>4.5</td>
<td>5</td>
<td>-0.025</td>
<td>0.003</td>
<td>0.004</td>
<td>0.002</td>
<td>4e-4</td>
<td>4e-4</td>
</tr>
<tr>
<td>$t_3$</td>
<td>10</td>
<td>4.5</td>
<td>5</td>
<td>-0.042</td>
<td>0.002</td>
<td>7e-4</td>
<td>0.003</td>
<td>0.001</td>
<td>6e-4</td>
</tr>
</tbody>
</table>

**Table 8:** Non-parametric estimation result with $\Delta$

<table>
<thead>
<tr>
<th>Model</th>
<th>True $\mu$</th>
<th>True $\delta$</th>
<th>True $\Delta$</th>
<th>Bias $\mu$</th>
<th>Bias $\delta$</th>
<th>Bias $\Delta$</th>
<th>Bias $\Delta/\Delta$</th>
<th>MSE $\mu$</th>
<th>MSE $\delta$</th>
<th>MSE $\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>10</td>
<td>4.5</td>
<td>5</td>
<td>0.321</td>
<td>0.113</td>
<td>-2.301</td>
<td>-2.405</td>
<td>0.052</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>$t_3$</td>
<td>10</td>
<td>4.5</td>
<td>5</td>
<td>-0.042</td>
<td>0.002</td>
<td>7e-4</td>
<td>0.003</td>
<td>0.001</td>
<td>6e-4</td>
<td></td>
</tr>
</tbody>
</table>

**Table 9:** Estimation results for the Australian athletes BMI data

Figure 6 shows that neither the skewed normal nor the skewed-t distributions can capture the right tail of the BMI variable very well. We next check if the fit improves when no assumption is made on a specific density for $g(r)$. Figure 7 shows that the BMI data can be fitted much better under a non-parametric framework. The difference between the density from our model estimation and the true density is mainly due to the accuracy of the kernel estimation. With a better estimate of the non-parametric density function, our model would better predict the potential $g(r)$ functions.

### 6 Conclusions

In this paper, we decomposed the traditional skewed elliptical distribution into component structures given by the elliptical contours and the density assigning function. The decomposition allows us to estimate these two components separately, resulting in a faster and more accurate estimation process than both direct likelihood maximization and the expectation-maximization algorithm, if we assume a parametric form. In a semi-parametric framework when we assume that the assigning function is unknown, our model still provides accurate estimators.

We also discuss the extension of our approach to higher dimension settings. Since the density assigning functions are always one-dimensional, independent of the dimensions of the skewed data, the assigning function is always easy to estimate no matter how many dimensions we have. This density function can take any form, making our model very flexible and general.

Another important application of our model is its use in the symmetric case. If we let the skew parameters be zero, our model becomes a generalized class of elliptical distributions, including the multivariate normal
and t-distribution. The decomposition introduced in our paper provides thus another useful perspective for the analysis of this class of elliptical distributions.

Appendices

A Proof of (8)

First, recall two lemmas from Fang and Zhang (1990, p.68-69).

**Lemma 1:** Suppose that nonnegative random variables $R$ and $S$ are independent, $R$ has cdf $F$, and $S$ is absolutely continuous with density $g$. Let $T = RS$, then

$$P(R \leq r | T = t) = \frac{1}{h(t)} \int_0^r r^{-1} g(t/r) dF(r)$$

where

$$h(t) = \int_0^\infty r^{-1} g(t/r) dF(r).$$

**Lemma 2:** Under the notation of Lemma 1, let $R_2^2 = R(a^2)$ for simplicity. We have:

$$P(R_2^2 \leq \rho) = \frac{\int_0^\rho (r^2 + a^2)^{\frac{1}{2}} (r^2 - a^2)^{1/2} r^{-(n-2)} dF(r)}{\int_0^\rho (r^2 - a^2)^{1/2} r^{-(n-2)} dF(r)}$$

(37)

for $\rho \geq 0$, $a > 0$ and $F(a) < 1$.

Now we start to prove (8) with **Lemma 1** and **Lemma 2**. The density function of $z$ can be calculated as:

$$f(z) = Pr(x_{n-1} = z | x_1 > 0) = \frac{Pr(x_{n-1} = z, x_1 > 0)}{Pr(x_1 > 0)} = 2Pr(x_1 > 0 | x_{n-1} = z)f_{n-1}(z).$$

(38)

To calculate $Pr(x_1 > 0 | x_{n-1} = z)$ we need to know the form of $f(x_1 | x_{n-1} = z)$. In our skewed copula formula, $a^2 = (z - \mu)^T \Sigma_{11}^{-1} (z - \mu), m = 1$. Denote the denominator in Lemma 2 by $C_m$, the conditional density is given by taking the derivative of (37):

$$f(x_1 | x_{n-1} = z) = \frac{1}{C_{m} \sqrt{1 - \delta^T \Omega_{m}^{-1} \delta}} \frac{\pi^2}{(2)^{m}} h(x_m^2)$$

where

$$h(x_m^2) = f_m \left( \frac{(x_1 - \delta^T \Sigma_{11}^{-1} (z - \mu))^2}{1 - \delta^T \Sigma_{11}^{-1} \delta} + (z - \mu)^T \Sigma_{11}^{-1} (z - \mu) \right)$$
Figure 7: Australian athletes BMI data: fitting distributions adding ∆ and assuming that \(g(r)\) is unknown

\[
C_m = \int_a^\infty (r^2 - a^2)^{(\frac{5}{2} - 1)} r^{-(n-1)} g_n(r) dr
= \int_a^\infty (r^2 - a^2)^{\frac{1}{2}} r^{-n} g_n(r) dr
= \int_a^\infty (r^2 - a^2)^{\frac{1}{2}} r^{-n} \frac{2\pi^2}{\Gamma(\frac{3}{2})} r^m f_n(r) dr
= \frac{\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} \int_0^\infty f_n(x^2 + a^2) dx
= \frac{\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} \int_0^\infty f_n(x^2 + (z - \mu)^T \Sigma^{-1} (z - \mu)) dx.
\]

It follows that

\[
P(x_1 > 0|x_{n-1} = z) = \frac{1}{2\sqrt{1 - \delta^T \Sigma_{11}^{-1} \delta}} \int_0^\infty h(x^2) dx
\]

\[
= \frac{1}{2\sqrt{1 - \delta^T \Sigma_{11}^{-1} \delta}} \int_0^\infty f_n(x^2 + (z - \mu)^T \Sigma_{11}^{-1} (z - \mu)) dx
\]

\[
= \frac{1}{2\sqrt{1 - \delta^T \Sigma_{11}^{-1} \delta}} \int_\delta \Sigma_{11}^{-1} (y - \mu)^2 \Sigma_{11}^{-1} (z - \mu) dy
\]

(let \( t = \frac{(x - \delta^T \Sigma_{11}^{-1} (y - \mu))^2}{1 - \delta^T \Sigma_{11}^{-1} \delta} \) and write \( x = t \))

\[
= \frac{1}{2\sqrt{1 - \delta^T \Sigma_{11}^{-1} \delta}} \int_0^\infty \delta^T \Sigma_{11}^{-1} (y - \mu)^2 \Sigma_{11}^{-1} (z - \mu) dy
\]

\[
= \frac{1}{2\sqrt{1 - \delta^T \Sigma_{11}^{-1} \delta}} \int_0^\delta \Sigma_{11}^{-1} (y - \mu)^2 \Sigma_{11}^{-1} (z - \mu) dy
\]

\[
= \frac{1}{2\sqrt{1 - \delta^T \Sigma_{11}^{-1} \delta}} \int_0^\delta \Sigma_{11}^{-1} (y - \mu)^2 \Sigma_{11}^{-1} (z - \mu) dy.
\]

Inserting (39) into (38) now leads to (8).
B Proof of Proposition 1

Proof: Let \( x - \mu = r \cdot \sin(\theta) \). Then

\[
\int_{-\infty}^{\infty} (x - \mu)^n f(x)dx = \int_{-\infty}^{\infty} (x - \mu)^n \int_{0}^{\infty} f_r(x, r, \beta)g(r)drdx
\]

\[
= \frac{1}{n} \int_{-2\beta}^{2\beta} \sin^n \theta d\theta \int_{0}^{\infty} r^n g(r)dr + \frac{2}{n} \int_{-2\beta}^{2\beta} \sin^n \theta d\theta \int_{0}^{\infty} r^n g(r)dr
\]

\[
E_{n}^1 = \frac{-1}{n} \int_{-2\beta}^{2\beta} \sin^n \theta d\theta d\cos \theta
\]

\[
= -\frac{1}{n} \left[ \sin^{n-1} \theta \cos \theta \right]_{-2\beta}^{2\beta} - (n-1) \int_{-2\beta}^{2\beta} \cos^2 \theta \sin^{n-2} \theta d\theta
\]

\[
= -\frac{1}{n} \sin^{n-1} \theta \cos \theta \right]_{-2\beta}^{2\beta} + \frac{n-1}{n} \int_{-2\beta}^{2\beta} \sin^{n-2} \theta d\theta
\]

\[
\Rightarrow E_{n}^1 = \frac{A_1}{n} + \frac{n-1}{n} E_{n}^2.
\]

Similarly,

\[
E_{n}^2 = \frac{\sin^{n-1}(2\beta) \cos(2\beta) + n-1}{n} E_{n}^2
\]

\[
\Rightarrow W_n = E_{n}^1 + E_{n}^2 = \frac{\sin^{n-1}(2\beta) + \sin^{n-1}(-2\beta)}{n} \cos(2\beta) + \frac{n-1}{n} W_{n-2}.
\]

References


