The Multivariate **g**-and-**h** Distribution

Christopher Field
Department of Mathematics and Statistics
Dalhousie University
Halifax, Nova Scotia
Canada B3H 3J5
(*field@mathstat.dal.ca*)

Marc G. Genton
Department of Statistics
Texas A&M University
College Station, TX 77843-3143
(*genton@stat.tamu.edu*)

In this article we consider a generalization of the univariate g-and-h distribution to the multivariate situation with the aim of providing a flexible family of multivariate distributions that incorporate skewness and kurtosis. The approach is to modify the underlying random variables and their quantiles, directly giving rise to a family of distributions in which the quantiles rather than the densities are the foci of attention. Using the ideas of multivariate quantiles, we show how to fit multivariate data to our multivariate **g**-and-**h** distribution. This provides a more flexible family than the skew-normal and skew-elliptical distributions when quantiles are of principal interest. Unlike those families, the distribution of quadratic forms from the multivariate **g**-and-**h** distribution depends on the underlying skewness. We illustrate our methods on Australian athletes data, as well as on some wind speed data from the northwest Pacific.

KEY WORDS: Kurtosis; Multivariate; Quantiles; Shape; Skewness; Transformation.

1. INTRODUCTION

The univariate and multivariate normal distributions have played a central role in statistics. Nevertheless, there are many natural phenomena that do not follow the normal law, calling for multivariate nonnormal distributions. In this article, we use data from extreme events as motivating examples. In the univariate case, Dupuis and Field (2004) fit extreme oceanic wind speeds with the g-and-h distribution. In the multivariate situation, Genton and Thompson (2003) were interested in finding multivariate distributions that fit their bivariate extreme sea level data. Departures from normality can be achieved by introducing skewness and different kurtosis in the distribution. Two main approaches are available for this purpose.

The first approach consists of modifying the probability densities. It was first developed by Azzalini (1985) for the univariate normal distribution and by Azzalini and Dalla Valle (1996) for the multivariate normal distribution, yielding the so-called "skew-normal" distribution. Extensions to skew-elliptical distributions have been proposed by Azzalini and Capitanio (1999), Branco and Dey (2001), Sahu, Dey, and Branco (2003), and Wang and Genton (2006) and include, for instance, skew-t, skew-Cauchy, and skew-slash distributions. Most of these distributions are particular types of generalized skew-elliptical distributions recently introduced by Genton and Loperfido (2005), defined as the product of a multivariate elliptical density with a skewing function π,

$$2|\mathbf{\Omega}|^{-1/2}\cdot f\big[\mathbf{\Omega}^{-1/2}(\mathbf{y}-\boldsymbol{\xi})\big]\cdot\pi\big[\mathbf{\Omega}^{-1/2}(\mathbf{y}-\boldsymbol{\xi})\big],\qquad \mathbf{y}\in\mathbb{R}^d, \tag{1}$$

where $\boldsymbol{\xi}\in\mathbb{R}^d$ and $\mathbf{\Omega}\in\mathbb{R}^{d\times d}$ are location and scale parameters, $0\le\pi(\mathbf{y})\le 1$, and $\pi(-\mathbf{y})=1-\pi(\mathbf{y})$. These distributions are attractive because their properties are very similar to those of the normal distribution. However, one possible drawback is that the distribution of their quadratic forms is independent of the skewing function π (see, e.g., Genton, He, and Liu 2001; Loperfido 2001; Wang, Boyer, and Genton 2004a; Genton and Loperfido 2005). This invariance property also occurs with more general classes of multivariate distributions, such as skew-symmetric distributions (Wang et al. 2004b), as well as flexible skew-symmetric distributions (Ma and Genton 2004). This can be an issue for statistical inference based on quadratic forms of the observations. For instance, it implies that the distribution of the length of centered vectors does not depend on the skewness. A recent account of the properties and applications of those multivariate skewed distributions was given in the book edited by Genton (2004).

The second approach consists of modifying the random variables themselves, and thus also their quantiles. This approach was suggested by Tukey in 1977 and had been discussed by Hoaglin and Peters (1979) and Hoaglin (1983) in the univariate setting. Basically, a standard normal random variable Z is transformed to $Y=\tau_{g,h}(Z)$, where

$$\tau_{g,h}(Z)=\left(\frac{\exp(gZ)-1}{g}\right)\exp\left(\frac{h}{2}Z^2\right), \tag{2}$$

with $h\ge 0$. The resulting random variable Y is said to have a g-and-h distribution, where g is a real constant controlling the skewness and h is a nonnegative real constant controlling the kurtosis, or elongation. Note that the distribution of $\tau_{0,0}(Z)=Z$ is the normal one, the distribution of $\tau_{0,h}(Z)$ has heavier tails than the normal (increasingly with h), and $\tau_{g,0}(Z)$ coincides with the lognormal distribution. Thus this family of distributions encompasses a considerable variety of distribution shapes. The quantiles of $Y=\tau_{g,h}(Z)$ can be easily computed, because $\tau_{g,h}$ increases monotonically in Z and is therefore a bijective transformation. Thus the uth quantile of the distribution of Y is simply $\tau_{g,h}(z_u)$ where z_u is the uth quantile of the standard normal distribution.

In this article, we extend the g-and-h distribution to the multivariate setting. This requires us to define multivariate quantiles, and we make use of the recent proposals by Chaudhuri (1996) and Chakraborty (2001). In particular, we show that multivariate quantiles defined with the l_1- and l_2-norm are appropriate

© 2006 American Statistical Association and
the American Society for Quality
TECHNOMETRICS, FEBRUARY 2006, VOL. 48, NO. 1
DOI 10.1198/004017005000000562

for multivariate **g**-and-**h** transformations. The article is structured as follows. In Section 2 we define the multivariate **g**-and-**h** distribution based on appropriate multivariate quantiles and illustrate the various distributional shapes resulting from quantile transformations. In Section 3 we discuss the properties of the multivariate **g**-and-**h** distribution. In Section 4 we describe a fitting procedure based on quantiles and present two applications. We conclude in Section 5.

2. MULTIVARIATE **g**-AND-**h** DISTRIBUTION

A random vector $\mathbf{Y} \in \mathbb{R}^d$ is said to have a standard multivariate **g**-and-**h** distribution, where $\mathbf{g} = (g_1, \dots, g_d)^T \in \mathbb{R}^d$ controls the skewness and $\mathbf{h} = (h_1, \dots, h_d)^T \in \mathbb{R}^d_+$ controls the kurtosis, if it can be represented as

$$\mathbf{Y} = \left(\tau_{g_1,h_1}(Z_1), \dots, \tau_{g_d,h_d}(Z_d)\right)^T = \boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}), \tag{3}$$

where $\mathbf{Z} = (Z_1, \dots, Z_d)^T \sim \mathrm{N}_d(\mathbf{0}, \mathbf{I}_d)$ has a standard multivariate normal distribution and the univariate function $\tau_{g,h}$ is defined by (2). It is worth commenting on our choice of componentwise transformation rather than transformation of the vector $\mathbf{Z}$. One of the principles used in the original choice of the g transformation in one dimension was that the transformation is approximately linear for small g. By choosing the componentwise approach, we are able to maintain this property, which is not true of any natural vector transformation.

To define the general multivariate **g**-and-**h** distribution, we let $\boldsymbol{\Sigma}$ be an arbitrary covariance matrix and $\boldsymbol{\mu}$ be an arbitrary location. Then the general multivariate **g**-and-**h** distribution can be represented as

$$\mathbf{Y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) + \boldsymbol{\mu}. \tag{4}$$

We note that because $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}$ is a nonlinear function, we can make the affine transformation before or after applying $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}$. But because $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}$ is set up for standardized variables, it makes more sense to apply the affine transformation after transforming $\mathbf{Z}$. This also gives a definition consistent with the one-dimensional case; that is, (3) reduces to (2) when $d = 1$.

The next step is to consider the quantiles of the multivariate **g**-and-**h** distribution. To do this, we first need the concept of a multivariate quantile. In this article we use the idea of quantiles defined in terms of norm minimization as proposed by Chaudhuri (1996) and extended by Chakraborty (2001) to be affine-equivariant. As noted at the end of this section, it is possible to use other multivariate quantiles as long as they are affine-equivariant. Serfling (2002) provided a critical review of various versions of multivariate quantile functions.

For the moment, assume that we have the idea of a multivariate quantile in direction $\mathbf{u}$ (defined later) based on the l_p-norm for $1 \le p < \infty$, for the standard normal random vector $\mathbf{Z}$, say $\mathbf{q}_\mathbf{Z}(\mathbf{u})$. The corresponding quantiles of the general multivariate **g**-and-**h** distribution, $\mathbf{Y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) + \boldsymbol{\mu}$, are given by

$$\mathbf{q}_\mathbf{Y}(\mathbf{u}) = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{q}_\mathbf{Z}(\tilde{\mathbf{u}})) + \boldsymbol{\mu}, \tag{5}$$

where $\tilde{\mathbf{u}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{u}\|\mathbf{u}\|_r / \|\boldsymbol{\Sigma}^{-1/2}\mathbf{u}\|_r$ and $1/p + 1/r = 1$ with the convention that for l_1, $r = \infty$ gives the maximum norm. Note also that (5) is consistent with the one-dimensional case.

The next step is to give details of the multivariate quantiles of Chaudhuri (1996). He noted that in one dimension, for any α with $0 < \alpha < 1$ and $u = 2\alpha - 1$, the αth quantile q, for a sample $y_1, \dots, y_n$, can be obtained by minimizing

$$\operatorname*{argmin}_{q \in \mathbb{R}} \sum_{i=1}^{n} |y_i - q| - nuq.$$

Chaudhuri then indexed the multivariate quantiles in $\mathbb{R}^d$ by elements of the unit open ball $\{\mathbf{u} | \mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\|_p < 1\}$ and defined the geometric quantile $\mathbf{q}$ corresponding to $\mathbf{u}$, for a multivariate sample $\mathbf{y}_1, \dots, \mathbf{y}_n$ in $\mathbb{R}^d$, as

$$\operatorname*{argmin}_{\mathbf{q} \in \mathbb{R}^d} \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{q}\|_p - n \sum_{j=1}^{d} u_j q_j, \tag{6}$$

where $\|\cdot\|_p$ denotes the l_p-norm for $1 \le p < \infty$. In our approach we focus on the l_1-norm and the l_2-norm.

We now find the l_1-quantile, $\mathbf{q}_\mathbf{Z}$, for the standard multivariate normal distribution. Note that with $p = 1$, $\|\mathbf{y}_i - \mathbf{q}\|_1 = \sum_{j=1}^{d} |y_{ij} - q_j|$, (6) can be written as

$$\operatorname*{argmin}_{\mathbf{q} \in \mathbb{R}^d} \sum_{j=1}^{d} \left(\sum_{i=1}^{n} |y_{ij} - q_j| - nu_j q_j \right). \tag{7}$$

From (7), we note that we can do the minimization separately in each dimension, so that q_j is the $(u_j + 1)/2$th quantile for the jth component of the data. Hence the l_1-quantile corresponding to $\mathbf{u}$ for the standard multivariate **g**-and-**h** distribution has jth component $\tau_{g_j,h_j}(z_{(u_j+1)/2})$, where z_u is the uth quantile of the standard normal distribution. The quantiles of the general multivariate **g**-and-**h** distribution are given by (5). For the l_2 case, the quantile for $\mathbf{Z}$ in direction $\mathbf{u}$ is given by $\gamma^{-1}(\|\mathbf{u}\|_2)\mathbf{u}/\|\mathbf{u}\|_2$, where

$$\gamma(t) = \sum_{j=0}^{\infty} \frac{\Gamma((d+1)/2 + j)}{\Gamma(d/2 + j) j! 2^{j-1/2}} (2j - t^2) t^{2j-1} e^{-t^2/2}$$

and d is the dimension of $\mathbf{Z}$ (B. Chakraborty, personal communication). As can be seen, the quantiles for directions of a fixed length are circular. We note that the l_1-quantiles build in some robustness and give the usual quantiles in one dimension but are essentially computed componentwise in higher dimensions. In contrast, the l_2-quantiles incorporate the dependence among the variables but give less common quantiles in one dimension. Figure 1 depicts the contour lines (5–95%) of the bivariate standard normal l_1-quantiles (left) and l_2-quantiles (right). These contour lines can then be modified in various ways with the **g**-and-**h** transformation (Fig. 2).

To fit the multivariate **g**-and-**h** to data, we need to compute the quantiles for sets of data. It is important that our computed quantiles are affine-equivariant. To achieve this, we use a technique similar to that proposed by Chakraborty (2001). In his method, he chose a data-driven coordinate system based on $d + 1$ of the observations, with the first observation serving as the origin, and denoted the transformation by $Y(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the subset of indices of the $d + 1$ chosen observations. He then transformed the data to get $\mathbf{x}_i = (Y(\boldsymbol{\alpha}))^{-1}\mathbf{y}_i$ and computed the quantiles for $\mathbf{x}$ in transformed direction $\tilde{\mathbf{u}}$, say $\tilde{\mathbf{q}}(\tilde{\mathbf{u}})$, then transformed back to get the quantile $Y(\boldsymbol{\alpha})\tilde{\mathbf{q}}(\tilde{\mathbf{u}})$ in direction $\mathbf{u}$. As shown by Chakraborty, this quantile is affine-equivariant. We

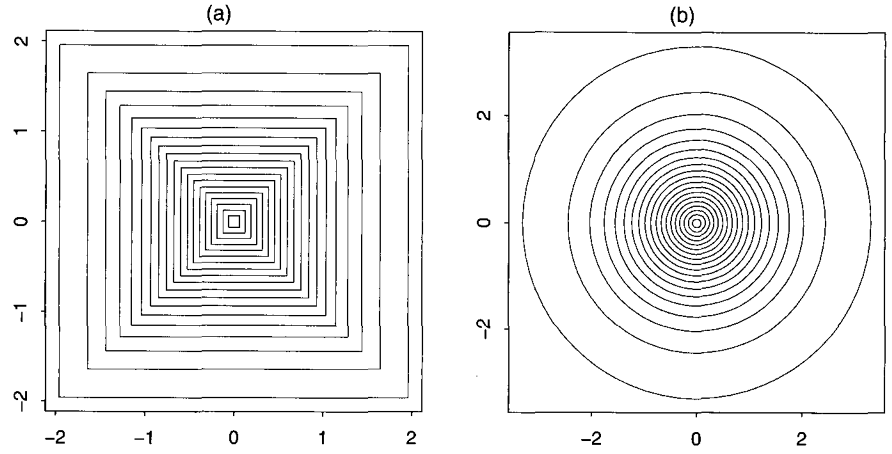

Figure 1. Bivariate Standard Normal l_1-Quantile (a) and l_2-Quantile (b) Contour Lines (5–95%).

have modified his technique by using the square root of the minimum covariance determinant (MCD) estimator (Rousseeuw 1985), $\hat{\mathbf{\Sigma}}_{\mathrm{MCD}}^{1/2}$ in place of $Y(\boldsymbol{\alpha})$. By choosing the MCD, we have affine equivariance of $\hat{\mathbf{\Sigma}}_{\mathrm{MCD}}^{1/2}$. This property is not automatically true for other choices of affine-equivariant covariance estimates. Given the affine equivariance of $\hat{\mathbf{\Sigma}}_{\mathrm{MCD}}^{1/2}$, Chakraborty's proof of the affine equivariance of these quantiles (Chakraborty 2001, thm. 2.1 proof, p. 396) goes through directly. To summarize, we compute affine-equivariant quantiles as follows:

1. Compute $\mathbf{x}_i = \hat{\mathbf{\Sigma}}_{\mathrm{MCD}}^{-1/2}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{\mathrm{MCD}})$, where $\hat{\mathbf{\Sigma}}_{\mathrm{MCD}}$ and $\hat{\boldsymbol{\mu}}_{\mathrm{MCD}}$ are the estimated MCD covariance and mean of the data, $\mathbf{y}$.
2. Compute the quantile (either l_1 or l_2) in direction $\tilde{\mathbf{u}} = \hat{\mathbf{\Sigma}}_{\mathrm{MCD}}^{-1/2}\mathbf{u}\|\mathbf{u}\|_r/\|\hat{\mathbf{\Sigma}}_{\mathrm{MCD}}^{-1/2}\mathbf{u}\|_r$ on the data $\mathbf{x}$, say $\tilde{\mathbf{q}}(\tilde{\mathbf{u}})$, where $1/p + 1/r = 1$, with the convention that for l_1, $r = \infty$ gives the maximum norm.
3. $\mathbf{q}(\mathbf{u}) = \hat{\mathbf{\Sigma}}_{\mathrm{MCD}}^{1/2}\tilde{\mathbf{q}}(\tilde{\mathbf{u}}) + \hat{\boldsymbol{\mu}}_{\mathrm{MCD}}$.

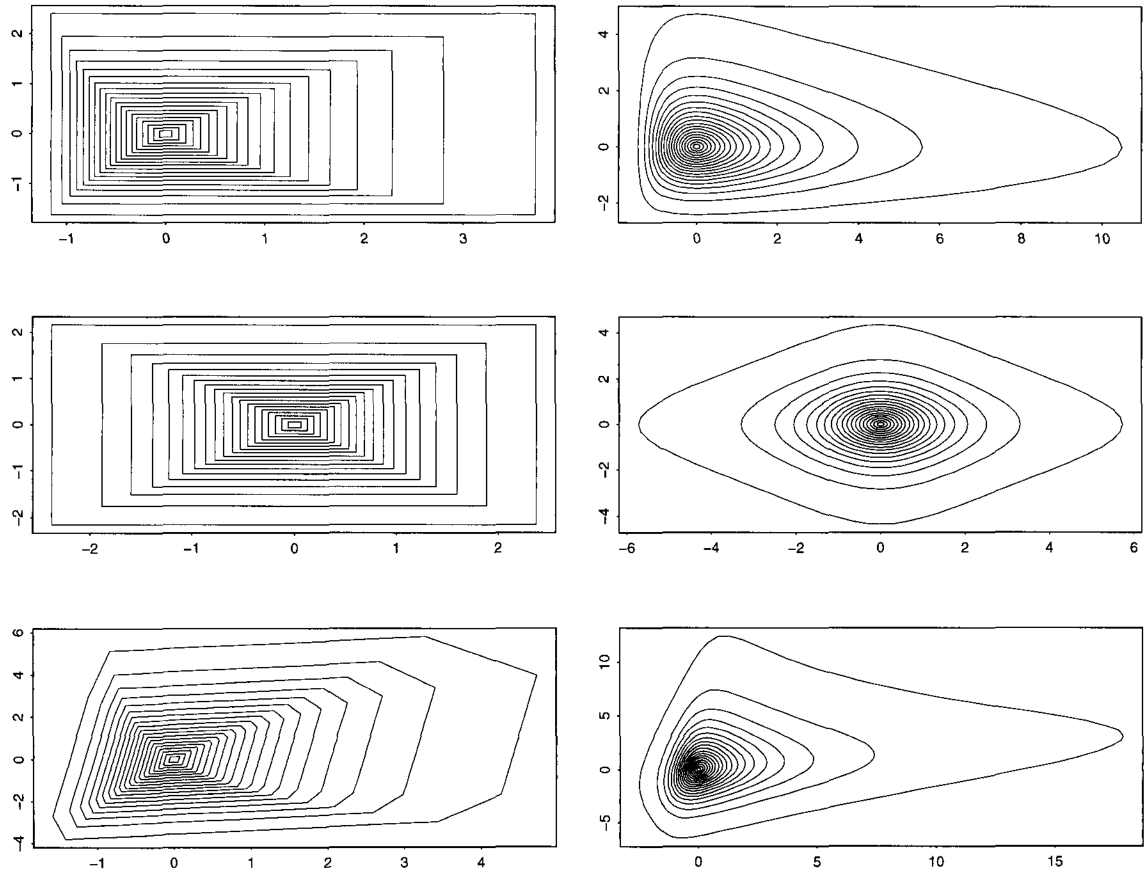

*Figure 2. Some Bivariate Standard Normal l_1-Quantile (left column) and l_2-Quantile (right column) Contour Lines (5–95%) After **g**-and-**h** Transformation: $g_1 = .6$, $g_2 = .2$, $h_1 = h_2 = 0$ (first row); $g_1 = g_2 = 0$, $h_1 = .1$, $h_2 = .05$ (second row); $g_1 = .6$, $g_2 = .2$, $h_1 = .1$, $h_2 = .05$ With Scale Matrix Premultiplication (third row).*

Note that although we have used the multivariate quantiles proposed by Chaudhuri (1996) and Chakraborty (2001), we could have used other affine-equivariant quantiles, such as those summarized by Serfling (2002). For instance, the quantiles defined on the basis of an appropriate depth function by means of the depth inner region are affine-equivariant and nested, particularly for the half-space depth and simplicial depth. The only drawback to this approach is computational, in that only very limited algorithms are available, especially in three and higher dimensions.

3. PROPERTIES

We begin by briefly discussing the form of the density and the moments of the **g**-and-**h** distribution, although we note that these are of secondary importance, because the motivation was to model the quantiles directly rather than the density. In particular, where the interest is in modeling tail behavior, the quantiles provide the natural scale for the tails, and a model of the quantiles as provided in the **g**-and-**h** distribution is more natural than working through the density.

For the univariate case with $g = 0$, the density is given by

$$f_h(y) = \frac{\phi(\tau_{0,h}^{-1}(y))}{\tau_{0,h}'(\tau_{0,h}^{-1}(y))},$$

where ϕ is the standard normal density (see Morgenthaler and Tukey 2000). There is no closed form of the density for $g \neq 0$, but the value of the distribution function can be evaluated numerically by inverting the quantile function. The moments in the univariate case have been given by Martinez and Iglewicz (1984), where the kth moment exists for $h < 1/k$. These are computed by integrating $(\tau_{g,h}(Z))^k$ with respect to the normal density.

We now turn to the multivariate case. From the definition (3), it is clear that the components of $\mathbf{Y}$ are independent, so that the joint density is simply the product of the component densities. This has a closed form when all of the $g_i = 0$. As noted earlier, the univariate density function can be evaluated numerically, and hence we can obtain the joint density. To model the dependent case, we can simply premultiply $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})$ by a matrix $\boldsymbol{\Sigma}^{1/2}$ to achieve a desired correlation structure as in (4). The marginal moments of $\mathbf{Y}$ come directly from the univariate moments. The moments of $\mathbf{Y}$ in (4) can then be obtained in the usual fashion, that is

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\Sigma}^{1/2}\mathrm{E}(\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})) + \boldsymbol{\mu} \tag{8}$$

and

$$\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Sigma}^{1/2}\,\mathrm{var}(\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}))(\boldsymbol{\Sigma}^{1/2})^T. \tag{9}$$

Because $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}$ is nonlinear, we use the delta method to approximate its expectation and covariance matrix in the foregoing expressions. Note that $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}$ operates componentwise on $\mathbf{Z}$, and thus only univariate approximations need to be computed. A first-order approximation yields $\mathrm{E}(\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})) \approx \mathrm{E}(\mathbf{Z}) = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})) \approx \mathrm{var}(\mathbf{Z}) = \mathbf{I}_d$. We further compute an approximation of order four, yielding $\mathrm{E}(\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})) \approx \mathbf{d}$ and $\mathrm{var}(\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})) \approx \mathbf{D}^2$, with $\mathbf{D}$ a diagonal matrix, where for $i = 1, \dots, d$,

$$d_i = \frac{1}{8}g_i^3 + \frac{3}{4}g_i h_i + \frac{1}{2}g_i \tag{10}$$

and

$$D_{ii}^2 = 1 + \frac{1}{6}g_i^6 + 2g_i^4 h_i + \frac{11}{12}g_i^4 + 6g_i^2 h_i^2 + \frac{11}{2}g_i^2 h_i + \frac{3}{2}g_i^2 + \frac{15}{4}h_i^2 + 3h_i. \tag{11}$$

Note that although $\mathbf{D}^2$ is diagonal, the full matrix $\mathrm{var}(\mathbf{Y})$ is affected by $\mathbf{g}$ and $\mathbf{h}$ in (9).

As noted in Section 1 there have been a number of proposals to modify multivariate densities to allow variation in both skewness and kurtosis. In carrying out inference in multivariate settings, many of the statistics are based on quadratic forms. For the class of various skew-normal and skew-elliptical distributions, the density of the quadratic form is independent of the skewing function π in (1) (see Genton and Loperfido 2005). Although this may be a nice mathematical property, it does raise the question as to whether these distributions can really provide good models for broad classes of multivariate data. The motivation in developing alternatives to multivariate normal or t-distributions is to have densities for data not fit well by the standard elliptical distributions. This invariance property for the quadratic forms suggests that standard inferential methods might be misleading when applied to the family of skew-elliptical distributions, although the same property is beneficial for inference from nonrandom samples (see, e.g., Genton and Loperfido 2005). In particular, the invariance implies that the distribution of the length of centered vectors remains the same across the skew-elliptical family. Comparing the diagrams in Figures 1 and 2 would certainly show that the distributions of the lengths does not remain the same for the **g**-and-**h** distribution. Indeed, Figure 3 depicts the histograms of

$$\log\big((\mathbf{y}_i - \bar{\mathbf{y}})^T(\mathbf{y}_i - \bar{\mathbf{y}})\big), \qquad i = 1, \dots, 1{,}000, \tag{12}$$

where $\mathbf{y}_1, \dots, \mathbf{y}_{1{,}000}$ are iid bivariate realizations from the **g**-and-**h** transformation (3) and $\bar{\mathbf{y}}$ denotes the sample mean. Figure 3(a) is for $g_1 = g_2 = h_1 = h_2 = 0$, and Figure 3(b) is for $g_1 = .6$, $g_2 = .2$, $h_1 = .1$, $h_2 = .05$. The distribution of the lengths clearly changes across the multivariate **g**-and-**h** distribution family.

We argue that having the distribution of quadratic forms vary across the **g**-and-**h** distribution indicates that this family is more flexible for fitting multivariate data than the skew-elliptical family, in the sense that it will enable us to get better inference for a wide class of multivariate data. To carry out inference, it will be necessary to say something about the distribution of quadratic forms. For our family, we recommend using parametric bootstrap methodology. We generate data from the multivariate **g**-and-**h** distribution with the estimated parameters and use the distribution of the quadratic form over the bootstrap samples as the basis for inference.

4. APPLICATIONS

In this section we start by describing a fitting procedure of the multivariate **g**-and-**h** distribution to data by means of empirical

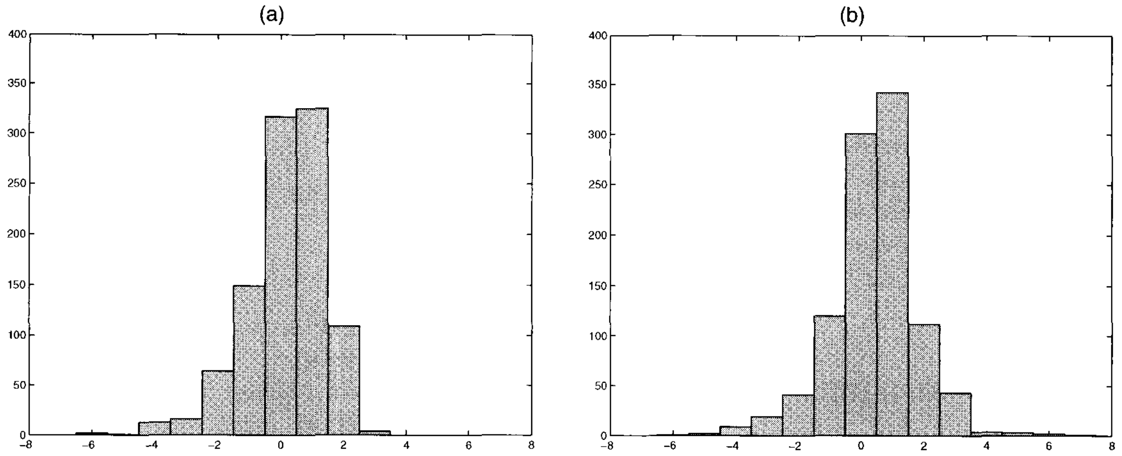

Figure 3. Histograms of the Logarithm of the Length of 1,000 Bivariate Realizations From the **g**-*and*-**h** *Transformation. (a) $g_1 = g_2 = h_1 = h_2 = 0$; (b) $g_1 = .6$, $g_2 = .2$, $h_1 = .1$, $h_2 = .05$.*

quantiles. We then present two applications of the multivariate **g**-and-**h** distribution. The first application is aimed at modeling the skewed bivariate distribution of body mass index (BMI) and lean body mass (LBM) measured on Australian athletes (see Cook and Weisberg 1994). The second illustrates the use of our methodology to model maxima of wind speed in the northwest Pacific.

4.1 Fitting With Quantiles

Consider a sample $\mathbf{y}_1, \dots, \mathbf{y}_n$, where each observation belongs to $\mathbb{R}^d$. To fit the multivariate **g**-and-**h** distribution to this sample, we suggest first estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (4) with the MCD estimator, which corresponds to the case where $\mathbf{g} = \mathbf{h} = \mathbf{0}$. Next we estimate the parameters **g** and **h** by matching the theoretical quantiles of the multivariate **g**-and-**h** distribution to the empirical quantiles from the data. We then update the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using (8) and (9) and the approximations in (10) and (11). This yields the following algorithm for a set $\mathcal{U}$ of vectors on the unit open ball:

Step 0. Initialize $\hat{\mathbf{g}} = \hat{\mathbf{h}} = \mathbf{0}$.

Step 1. Compute $\hat{\boldsymbol{\mu}}_{\text{MCD}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{MCD}}$ on $\mathbf{y}_1, \dots, \mathbf{y}_n$, and set $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_{\text{MCD}}$, $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\text{MCD}}$.

Step 2. For fixed $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, compute $\hat{\mathbf{g}}$ and $\hat{\mathbf{h}}$ with the minimization

$$(\hat{\mathbf{g}}, \hat{\mathbf{h}}) = \operatorname*{argmin}_{(\mathbf{g},\mathbf{h})} \sum_{\mathbf{u}\in\mathcal{U}} \left\| \hat{\mathbf{q}}_{\mathbf{y}}(\mathbf{u}) - \hat{\boldsymbol{\Sigma}}^{1/2} \tau_{\mathbf{g},\mathbf{h}}(\tilde{\mathbf{q}}_{\mathbf{z}}(\tilde{\mathbf{u}})) - \hat{\boldsymbol{\mu}} \right\|^2,$$

where the empirical quantile $\hat{\mathbf{q}}_{\mathbf{y}}(\mathbf{u})$ is computed as described at the end of Section 2.

Step 3. For fixed $\hat{\mathbf{g}}$ and $\hat{\mathbf{h}}$, update $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ with

$$\hat{\boldsymbol{\Sigma}} \to \hat{\boldsymbol{\Sigma}}_{\text{MCD}}^{1/2} \mathbf{D}^{-2} (\hat{\boldsymbol{\Sigma}}_{\text{MCD}}^{1/2})^T \tag{13}$$

and

$$\hat{\boldsymbol{\mu}} \to \hat{\boldsymbol{\mu}}_{\text{MCD}} - \hat{\boldsymbol{\Sigma}}^{1/2} \mathbf{d}. \tag{14}$$

Step 4. Repeat Steps 2 and 3 until convergence.

The choice of the set $\mathcal{U}$ should be based on the objective of the analysis. In most cases when using quantiles, the interest lies in the tail of the distribution. Therefore, we suggest emphasizing quantiles at the levels above 90%; see the applications in Sections 4.2 and 4.3, as well as the discussion on computational issues in Section 5.

4.2 Australian Athletes Data

We consider the bivariate sample $\mathbf{y}_1, \dots, \mathbf{y}_n$ of BMI/LBM measured on $n = 202$ Australian athletes by the Australian Institute of Sport. The MCD estimates are

$$\hat{\boldsymbol{\mu}}_{\text{MCD}} = \begin{pmatrix} 22.52 \\ 64.14 \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_{\text{MCD}} = \begin{pmatrix} 6.13 & 24.86 \\ 24.86 & 189.64 \end{pmatrix}. \tag{15}$$

Figure 4 presents a scatterplot of the data $\mathbf{y}_1, \dots, \mathbf{y}_n$ and some empirical bivariate l_1-quantiles [panel (a)] computed in 32 directions equispaced on the edges of a square for the levels 15%, 30%, 45%, 60%, 75%, 90%, 92%, 94%, 96%, and 98%. The levels 90–98% are used to capture the tail behavior of the data. We see that the BMI variable on the horizontal axis exhibits skewness on the right. Next we fit those empirical l_1-quantiles to the theoretical ones of the **g**-and-**h** distribution with the algorithm described in Section 4.1, yielding the estimates listed in the first row of Table 1.

Figure 4(b) shows the fitted l_1-quantiles from the **g**-and-**h** distribution with $\hat{\mathbf{g}}$ and $\hat{\mathbf{h}}$ from Table 1. Notice how our algorithm tries to achieve a good fit in both the center and the tails of the distribution.

Figure 5 presents some empirical bivariate l_2-quantiles [panel (a)] computed in 32 directions given by $2\pi/j$, $j = 1, \dots, 32$, for the levels 15%, 30%, 45%, 60%, 75%, 90%, 92%, 94%, 96%, and 98%. We fit those empirical l_2-quantiles to the theoretical ones of the **g**-and-**h** distribution with the algorithm described in Section 4.1, yielding the estimates listed in the second row of Table 1, and the fitted l_2-quantiles in Figure 5(b). Azzalini and Dalla Valle (1996) fitted a bivariate skew-normal distribution to the BMI/LBM data to account for skewness.

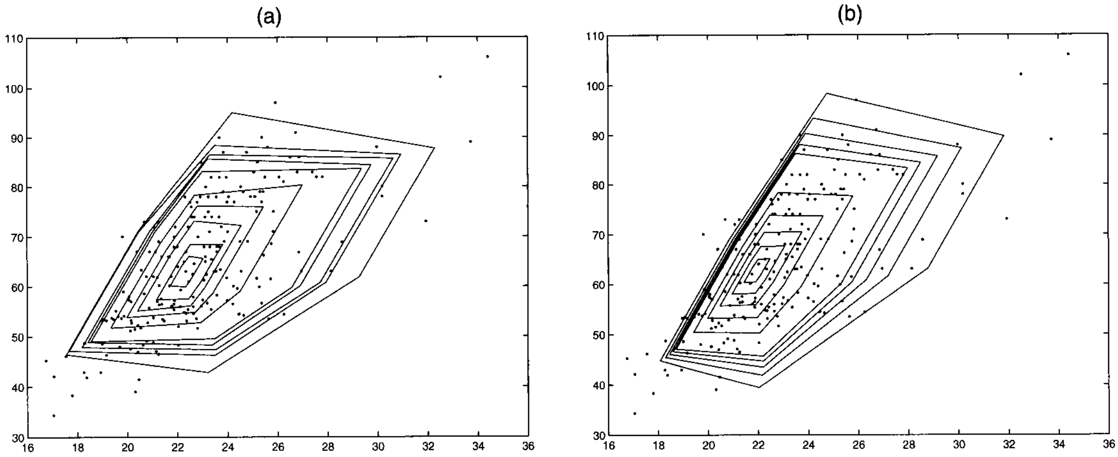

Figure 4. Scatterplot of the BMI/LBM Data With (a) Empirical l_1-Quantiles and (b) Fitted l_1-Quantiles With the ***g**-and-**h*** *Transformation.*

Their fit also indicated stronger skewness on the BMI than on the LBM, and thus is consistent with our **g**-and-**h** fit.

The **g**-and-**h** parameters in Table 1 fitted with l_1 or l_2 quantiles are somewhat different. However, recall from Section 2 that l_1 quantiles, although simple to compute, are calculated componentwise. Therefore, l_1 quantiles do not provide information about the joint distribution, whereas l_2 quantiles do. For this reason, we focus on l_2 quantiles in the next section.

4.3 Wind Speed Data

As an example in higher dimensions, we use wind speed data collected from moored buoys in the northwest Pacific. The data were provided by the Comprehensive Ocean-Atmosphere Data Sets (COADS) and are available at *http://www.cdc.noaa.gov/coads/products.html*. The raw data provided courtesy of COADS consisted of hourly wind speeds (m/sec) from moored buoys in the range of latitude 45–55 N and longitude 220–230 E for the years 1987–1997. For a number of the buoys, there were large gaps in the data series caused by the buoy being out of service or the data being of low quality. To have fairly complete data, we chose to fit the data to four buoys with data from 1996 and 1997. In wind speed data, there is considerable interest in the maximum wind speed over fixed time periods. We have used weekly maxima, leading to a dataset with $d = 4$ dimensions and sample size of $n = 92$.

We first consider fitting the data on four buoys with the multivariate **g**-and-**h** distribution. We compute empirical l_2-quantiles in the directions $2\pi/j, j = 1, \dots, 32$, for the levels 15%, 30%, 45%, 60%, 75%, 90%, 92%, 94%, 96%, and 98%. We fit those empirical l_2-quantiles to theoretical ones from the **g**-and-**h** distribution using the algorithm in Section 4.1, yielding

$$\hat{\mathbf{g}} = (-.006, .007, .017, -.032)^T \quad \text{and} \quad \hat{\mathbf{h}} = (.055, .076, 0, .723)^T. \tag{16}$$

We see that this fit indicates essentially no skewness for buoys 1 and 2, small right skewness for buoy 3, and left skewness for buoy 4 (compare with Fig. 2). Also, the third buoy does not have heavier tails than the normal distribution, but buoys 1 and 2 have slightly heavier tails and buoy 4 has significantly heavier tails.

We now look at comparing the winds for April–December of 1996 with the corresponding months of 1997. We are interested in comparing the winds from the two years to see if there is any evidence of a change. Each observation consists of a four-dimensional vector, and one approach is to compute vector differences and use the squared length as a measure of magnitude for each vector difference. If the years are similar, then this total wind difference should be small. To carry out the test with the **g**-and-**h** distribution on the 39 vectors of differences of the weekly wind maxima, we start by computing empirical l_2-quantiles in the directions $2\pi/j, j = 1, \dots, 32$, for the levels 15%, 30%, 45%, 60%, 75%, 90%, 92%, 94%, 96%, and 98%. We fit those empirical l_2-quantiles to theoretical ones from the **g**-and-**h** distribution, yielding the following estimates of **g** and **h** for the differences:

$$\hat{\mathbf{g}} = (-.125, -.017, -.156, .012)^T \quad \text{and} \quad \hat{\mathbf{h}} = (0, .042, 0, 0)^T, \tag{17}$$

with location and scale parameters

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} .677 \\ -.158 \\ -.602 \\ -.558 \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 17.863 & .462 & 5.074 & 2.556 \\ .462 & 13.511 & 7.707 & 9.791 \\ 5.074 & 7.707 & 10.174 & 7.781 \\ 2.556 & 9.791 & 7.781 & 11.002 \end{pmatrix}. \tag{18}$$

Table 1. Fitted Parameters of the Multivariate ***g**-and-**h*** *Distribution to the BMI/LBM Data With l_1- and l_2-Quantiles*

Quantile	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{12}$	$\hat{\Sigma}_{22}$	$\hat{g}_1$	$\hat{g}_2$	$\hat{h}_1$	$\hat{h}_2$
l_1	21.93	62.66	4.53	22.90	182.72	.569	.146	.029	0
l_2	22.42	63.86	5.94	23.78	180.84	.074	.033	.004	.016

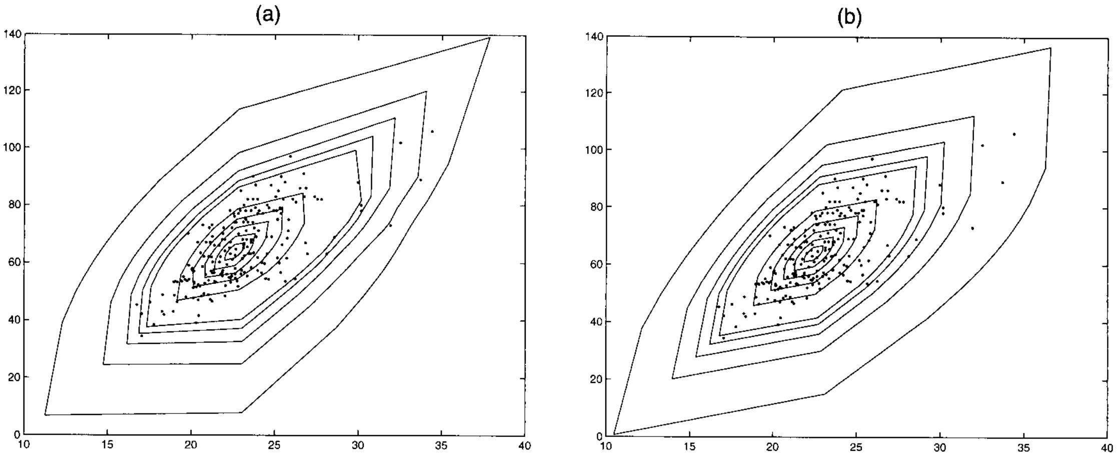

*Figure 5. Scatterplot of the BMI/LBM Data With (a) Empirical l_2-Quantiles and (b) Fitted l_2-Quantiles With the **g**-and-**h** Transformation.*

To test whether our observed quadratic form gives evidence of a difference, we carry out a parametric bootstrap to estimate the distribution of the quadratic form assuming that the differences have a mean of 0. Specifically, we simulate 999 samples each of size 39, from the **g**-and-**h** distribution with the parameters (17), $\hat{\boldsymbol{\Sigma}}$ computed in (18), and $\boldsymbol{\mu} = \mathbf{0}$ under the null hypothesis. For each sample, we compute the mean of the squared length of each difference and compare them with the corresponding value on the original data. This gives a p value of 54.7%, and thus we cannot reject the null hypothesis of no difference. In contrast, if we would use a skew-elliptical distribution to model the data, then the distribution of the quadratic form under the hypothesis of a zero mean for the differences would not depend on the skewness.

5. CONCLUSION

In this article, we have introduced the multivariate **g**-and-**h** distribution as a flexible family for multivariate data. In many settings it is the quantiles that are of interest, and by modeling them directly, we provide a natural method of fitting the distribution to data. One issue in higher dimensions is that the number of directions in which to compute the quantiles grows exponentially. If we chose eight directions in two dimensions, then we would need 8^{d-1} in d dimensions. This introduces computational issues that can be addressed by using ideas from partial factorial designs. For instance, it might be sensible to require that in any plane for two of the axis, we have eight directions; we then would only need $\binom{d}{2} \cdot 8$ directions.

ACKNOWLEDGMENTS

Part of this research was carried out in the fall of 2001 while the second author was visiting the first at Dalhousie University, whose hospitality is gratefully acknowledged. The authors thank Wade Blanchard for processing the wind speed data, as well as the editors and referees for helpful comments on the manuscript. The work of Genton was supported in part by National Science Foundation grant DMS-05-04896.

[Received October 2002. Revised December 2004.]

REFERENCES

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Azzalini, A., and Capitanio, A. (1999), "Statistical Applications of the Multivariate Skew Normal Distribution," *Journal of the Royal Statistical Society*, Ser. B, 61, 579–602.

Azzalini, A., and Dalla Valle, A. (1996), "The Multivariate Skew-Normal Distribution," *Biometrika*, 83, 715–726.

Branco, M. D., and Dey, D. K. (2001), "A General Class of Multivariate Skew-Elliptical Distributions," *Journal of Multivariate Analysis*, 79, 99–113.

Chakraborty, B. (2001), "On Affine Equivariant Multivariate Quantiles," *Annals of the Institute of Statistical Mathematics*, 53, 380–403.

Chaudhuri, P. (1996), "On a Geometric Notion of Quantiles for Multivariate Data," *Journal of the American Statistical Association*, 91, 862–872.

Cook, R. D., and Weisberg, S. (1994), *An Introduction to Regression Graphics*, New York: Wiley.

Dupuis, D., and Field, C. A. (2004), "Large Windspeeds, Modeling and Outlier Detection," *Journal of Agricultural, Biological & Environmental Statistics*, 9, 1–17.

Field, C. A. (2004), "Using the *gh* Distribution to Model Extreme Wind Speeds," *Journal of Statistical Planning and Inference*, 122, 15–22.

Genton, M. G. (ed.) (2004), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Boca Raton, FL, Chapman & Hall/CRC.

Genton, M. G., He, L., and Liu, X. (2001), "Moments of Skew-Normal Random Vectors and Their Quadratic Forms," *Statistics Probability Letters*, 51, 319–325.

Genton, M. G., and Loperfido, N. (2005), "Generalized Skew-Elliptical Distributions and Their Quadratic Forms," *Annals of the Institute of Statistical Mathematics*, 57, 389–401.

Genton, M. G., and Thompson, K. R. (2003), "Skew-Elliptical Time Series With Application to Flooding Risk," in *Time Series Analysis and Applications to Geophysical Systems*, eds. D. R. Brillinger, E. A. Anderson, and F. P. Schoenberg, New York: Springer-Verlag, pp. 169–186.

Hoaglin, D. C. (1983), "Summarizing Shape Numerically: The *g*-and-*h* Distributions," in *Exploring Data, Tables, Trends and Shapes*, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: Wiley, pp. 461–513.

Hoaglin, D. C., and Peters, S. C. (1979), "Software for Exploring Distributional Shapes," in *Proceedings of Computer Science and Statistics: 12th Annual Symposium on the Interface*, Ontario, Canada: University of Waterloo, pp. 418–443.

Loperfido, N. (2001), "Quadratic Forms of Skew-Normal Random Vectors," *Statistics and Probability Letters*, 54, 381–387.

Ma, Y., and Genton, M. G. (2004), "A Flexible Class of Skew-Symmetric Distributions," *Scandinavian Journal of Statistics*, 31, 459–468.

Martinez, J., and Iglewicz, B. (1984), "Some Properties of the Tukey *g* and *h* Family of Distributions," *Communications in Statistics, Part A—Theory and Methods*, 13, 353–369.

Morgenthaler, S., and Tukey, J. W. (2000), "Fitting Quantiles: Doubling, *HR*, *HQ*, and *HHH* Distributions," *Journal of Computational and Graphical Statistics*, 9, 180–195.

Rousseeuw, P. J. (1985), "Multivariate Estimation With High Breakdown Point," in *Mathematical Statistics and Applications*, Vol. B, eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht: Reidel, pp. 283–297.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003), "A New Class of Multivariate Skew Distributions With Applications to Bayesian Regression Models," *Canadian Journal of Statistics*, 31, 129–150.

Serfling, R. (2002), "Quantile Functions for Multivariate Analysis: Approaches and Applications," Special Issue: Frontier Research in Theoretical Statistics, 2000 (Eindhoven), *Statistica Neerlandica*, 56, 214–232.

Wang, J., Boyer, J., and Genton, M. G. (2004a), "A Note on an Equivalence Between Chi-Square and Generalized Skew-Normal Distributions," *Statistics and Probability Letters*, 66, 395–398.

——— (2004b), "A Skew-Symmetric Representation of Multivariate Distributions," *Statistica Sinica*, 14, 1259–1270.

Wang, J., and Genton, M. G. (2006), "The Multivariate Skew-Slash Distribution," *Journal of Statistical Planning and Inference*, 136, 209–220.