The Annals of Statistics
1999, Vol. 27, No. 3, 783–858

SPECIAL INVITED PAPER

MULTIVARIATE ANALYSIS BY DATA DEPTH: DESCRIPTIVE STATISTICS, GRAPHICS AND INFERENCE[1]

BY REGINA Y. LIU, JESSE M. PARELIUS AND KESAR SINGH

Rutgers University, New York Times Electronic Media Company, Rutgers University

A data depth can be used to measure the "depth" or "outlyingness" of a given multivariate sample with respect to its underlying distribution. This leads to a natural center-outward ordering of the sample points. Based on this ordering, quantitative and graphical methods are introduced for analyzing multivariate distributional characteristics such as location, scale, bias, skewness and kurtosis, as well as for comparing inference methods. All graphs are one-dimensional curves in the plane and can be easily visualized and interpreted. A "sunburst plot" is presented as a bivariate generalization of the box-plot. *DD*-(depth versus depth) plots are proposed and examined as graphical inference tools. Some new diagnostic tools for checking multivariate normality are introduced. One of them monitors the exact rate of growth of the maximum deviation from the mean, while the others examine the ratio of the overall dispersion to the dispersion of a certain central region. The affine invariance property of a data depth also leads to appropriate invariance properties for the proposed statistics and methods.

1. Introduction. Multivariate analysis plays a role of ever-increasing importance in statistics. Most statistical experiments are multivariate by nature, and large scale multivariate datasets are now made tractable by recent explosive advances in computer technology. However, classical multivariate analysis relies heavily on the assumption of normality or near-normality, which is often difficult to justify in practice.

The goal of this paper is to develop a general nonparametric multivariate methodology based on the concept of data depth. This methodology provides a systematic nonparametric approach for defining quantitative and graphical multivariate distributional characteristics and inference methods. A commonly adopted method for obtaining multivariate distributional characteristics has been a straightforward extension of the moment approach in the univariate case. More specifically, the location, scale, skewness and kurtosis are defined, respectively, in terms of the first, second, third and fourth moments. This leads to matrix or vector forms of outputs which are hard to

Received October 1997; revised October 1998.

[1] Supported in part by NSF Grants DMS-90-22126 and DMS-96-24940.

AMS 1991 *subject classifications*. 62H05, 62J20, 62-09.

Key words and phrases. Multivariate descriptive statistics, multivariate ordering, multivariate normality, data depth, depth ordering, depth-*L*-statistics, location, scale, bias, skewness, kurtosis, sunburst plots, *DD*-plots.

grasp conceptually and graphically. Worse still, this approach would not even be applicable if the moments do not exist. In our approach, these characteristics and their corresponding descriptive statistics are defined as functionals of data depth. They can be displayed as simple graphs on the plane and be easily visualized. Since these graphs are all based on an analysis of the contours derived from the data depth used, they convey a more intuitive picture of the distributional properties. Furthermore, our approach is *moment-free* if the underlying data depth is, which in fact is the case for almost all data depths we consider in this paper (cf. Section 2).

Our depth-based methodology may be viewed in part as a multivariate generalization of standard univariate rank methods, in the sense that they are both based on the idea of *ranks*. In the univariate setting, our methodology reduces to the standard univariate rank method, albeit without the distinction between positive and negative directions. There is, however, a marked difference in the two ways of ranking: the ranking in the univariate case is a *linear ranking* from the smallest to the largest, while our multivariate one is a *center-outward ranking* induced by depth. In this context, we note that the subject of multivariate ordering has attracted considerable interest over the years. For a survey of work in the area up to 1976 we refer to Barnett (1976). We should also point out that there exist many nonparametric multivariate approaches which in essence apply univariate nonparametric methods to analyze the multivariate observations componentwise [see, for example, Chapter 6 of Hettmansperger (1984)]. However, they often have difficulties in cases of dependence between component variates.

The paper is organized as follows. In Section 2 we present some background material and basic definitions, beginning with the general concept and examples of data depth. The notion of ordering according to depth, its affine invariance, and the resulting quantiles are then discussed at length. We introduce a bivariate generalization of the boxplot, which we call the *sunburst plot* [cf. Figure 5(a, b)]. We also review the Lorenz curve, which will serve later to interpret the descriptive statistics proposed in the paper. In Sections 3 through 6, we define several parameters according to depth, which can characterize a multivariate distribution in terms of its location, scale, skewness and kurtosis. In each section, graphs based on simulated data corresponding to the relevant descriptive statistics are displayed. Regardless of the dimension of the underlying distribution, the graphs are always one-dimensional curves in the plane and can be easily visualized and interpreted. We also show how the affine invariance property of a data depth leads to a corresponding invariance property for each proposed method. In Section 7, we introduce the *DD*-plot as a simple graphical tool for comparing two given samples or their underlying distributions. Different patterns of the *DD*-plot are associated with differences in location, scale, skewness or kurtosis between the two distributions. In Section 8, we establish some properties of multivariate normal distributions, and explain how to use them as diagnostic tools for checking normality. These properties include an almost surely asymptotic bound for the maximum of a multivariate normal sample (Theorem 8.2), and a fixed constant multiplier (Theorem 8.1) between the overall dispersion matrix and the dispersion matrix derived from a central region (Definition 2.3). Finally, in Section 9, we discuss aspects of our methodology such as computational feasibility, graphical presentability, conceptual and theoretical tractabilities and other ramifications of data depth. Most proofs are deferred to the Appendix.

One immediate application of the proposed scale parameter in Section 4 is the comparison of estimation errors of different estimators when they are used to estimate the same parameter. In the context of accounting for estimation error, we also introduce several definitions of bias in Section 4 and view the combination of the newly defined sale and bias as a generalization of the univariate mean-square-error. In the situation where only a sample is available for computing the estimators, we may use the bootstrap procedure to approximate the sampling distributions of the estimators, and then derive from these distributions the scale and bias for the estimators. Specific examples are graphically illustrated in Figure 9. Again, the comparison there can be easily interpreted visually.

In view of the vast literature on multivariate analysis, we cannot describe adequately in this paper all the important developments on the subject. Only the most closely related methods have been mentioned in each section. We refer to Anderson (1984) for a general reference on classical multivariate analysis and the Gnanadesikan (1997) (and references therein) for surveys of multivariate data analysis tools. We would also like to call the reader's attention to the systematic and insightful treatment of univariate descriptive statistics for nonparametric models in Bickel and Lehmann (1975a, b, 1976, 1979).

2. Data depth and background material.

Let F be a probability distribution in $\mathbb{R}^d$, $d \geq 1$. Throughout the paper, unless stated otherwise, we assume that F is absolutely continuous and also that $\{X_1, \dots, X_n\}$ is a random sample from F. Each sample point X_i is viewed as a $d \times 1$ column vector.

2.1. *Data depth and ordering/ranking multivariate observations.*

A data depth is a way of measuring how deep (or central) a given point $x \in \mathbb{R}^d$ is w.r.t. F or w.r.t. a given data cloud $\{X_1, \dots, X_n\}$. Some useful examples of data depth are:

1. The *Mahalanobis depth* ($M_h D$) [Mahalanobis (1936)] at x w.r.t. F is defined to be

$$M_h D(F; x) = \left[1 + (x - \mu_F)\Sigma_F^{-1}(x - \mu_F)\right]^{-1},$$

where μ_F and Σ_F are the mean vector and dispersion matrix of F, respectively. The sample version of $M_h D$ is obtained by replacing μ_F and Σ_F with their sample estimates.

2. The *half-space depth* (HD) [Hodges (1955), Tukey (1975)] at x w.r.t. F is defined to be

$$HD(F;x) = \inf_H \{P(H)\colon H \text{ is a closed half-space in } \mathbb{R}^d \text{ and } x \in H\}.$$

The sample version of $HD(F;x)$ is $HD(F_n;x)$. Here F_n denotes the empirical distribution of the sample $\{X_1,\dots,X_n\}$. The half-space depth is sometimes also referred to as the Tukey depth in the literature, for example, Liu and Singh (1993, 1997) and Yeh and Singh (1997).

3. The *convex hull peeling depth* (CD) [Barnett (1976)] at the sample point X_k w.r.t the data set $\{X_1,\dots,X_n\}$ is simply the *level of the convex layer* X_k belongs to. A convex layer is defined as follows. Construct the smallest convex hull which encloses all sample points $\{X_1,\dots,X_n\}$. The sample points on the perimeter are designated the first convex layer and removed. The convex hull of the remaining points is constructed; these points on the perimeter are the second convex layer. The process is repeated, and a sequence of nested convex layers is formed. The higher layer a point belongs to, the deeper the point is within the data cloud. Note that several other versions of "convex peeling" exist; see Eddy (1982) and Tukey's approach described in Huber (1972). Although only the simple convex hull peeling of Barnett (1976) is specified here, the methods we develop in this paper apply to all variations of convex peeling.

4. The *Oja depth* (OD) [Oja (1983)] at x w.r.t. F is defined to be

$$OD(F;x) = \left[1 + E_F\{\text{volume}(S[x, X_1,\dots,X_d])\}\right]^{-1},$$

where $S[x, X_1,\dots,X_d]$ is the closed simplex with vertices x, and d random observations $X_1,\dots,X_d$ from F. Obviously $OD(F_n;x) = \binom{n}{d}^{-1}[1 + \sum_*\{\text{volume}(S[x, X_{i_1},\dots,X_{i_d}])\}]^{-1}$ is the sample version of $OD(F;x)$. Here $*$ indicates all d-plets $(i_1,\dots,i_d)$ such that $1 \le i_1 \le \cdots \le i_d \le n$.

5. The *simplicial depth* (SD) [Liu (1990)] at x w.r.t. F is defined to be

$$SD(F;x) = P_F\{x \in S[X_1,\dots,X_{d+1}]\}.$$

Here $S[X_1,\dots,X_{d+1}]$ is a closed simplex formed by $(d+1)$ random observations from F. The sample version of $SD(F;x)$ is obtained by replacing F in $SD(F;x)$ by F_n, or alternatively, by computing the fraction of the sample random simplices containing the point x. In other words,

$$SD(F_n;x) = \binom{n}{d+1}^{-1} \sum_* I_{(x \in S[X_{i_1},\dots,X_{i_{d+1}}])},$$

where $I_{(\cdot)}$ is the indicator function.

6. The *majority depth* (M_jD) [Singh (1991)] of x w.r.t. F is defined to be

$$M_jD(F;x) = P_F\{x \text{ is in a major side determined by } (X_1,\dots,X_d)\}.$$

Here a *major side* is the half-space bounded by the hyperplane containing $(X_1, \dots, X_d)$ which has probability ≥ 0.5. The sample version of $M_jD(F; x)$ is $M_jD(F_n; x)$.

7. The *likelihood depth* (LD) [Fraiman and Meloche (1996)] of x w.r.t. F is simply its probability density, that is, $LD(F; x) = f(x)$, and the empirical version can be any consistent density estimate at x, for example, the kernel density estimate.

Henceforth, $D(\cdot\,;\cdot\,)$ or $D_F(\cdot)$ will be used to indicate any of the above-mentioned depths unless specified otherwise. The value of $D(F; x)$ may vary with the notion of data depth, but for each notion of depth, a larger value of $D(F; x)$ always implies a deeper (or more central) x w.r.t. F. When there is no possibility of confusion, we may omit the underlying distribution F and simply use $D(\cdot)$ and $D_n(\cdot)$ to denote, respectively, $D(F; \cdot\,)$ and $D(F_n; \cdot\,)$.

Given a notion of data depth, one can compute the depths of all the sample points $\{X_1, \dots, X_n\}$ and order them according to decreasing depth values. This gives a ranking of the sample points from the center outward. Let $X_{[i]}$ denote the sample point associated with the ith highest depth value. We view $X_{[1]}, \dots, X_{[n]}$ as the order statistics, with $X_{[1]}$ being the *deepest* or the *most central* point or simply the *center*, and $X_{[n]}$ the most outlying point. The implication is that *a larger rank is associated with a more outlying position w.r.t. the data cloud.* These order statistics induced by a data depth are different from the usual order statistics on the real line, since the latter are ordered from the smallest sample point to the largest, while the former start from the *middle* sample point and move outwards in all directions. In this article, only the depth-induced order statistics are studied. They will be referred to as *depth order statistics* (denoted by *DO*-statistics), and their ordering or ranking as *depth ordering* or *depth ranking*. When ties occur in the ordering, the corresponding sample points are viewed as depth-equivalent, and the set of these points is termed a *depth-equivalence class* (*de*-class for short). In the particular case where there is more than one sample point with the highest depth value, we refer to their average as the deepest point, for convenience. The notation ν_n is used to denote the deepest point w.r.t. the sample, and ν_F w.r.t. the underlying distribution F. For simplicity, the following method is used to assign ranks to points belonging to the same *de*-class: if $x_{i_1}, x_{i_2}, \dots, x_{i_k}$ all belong to the same *de*-class where $i_1 < i_2 < \cdots < i_k$ and there are exactly j sample points with higher depth values, then we assign $x_{i_1}, x_{i_2}, \dots, x_{i_k}$ to be $x_{[j+1]}, x_{[j+2]}, \dots, x_{[j+k]}$, in that order.

Some multivariate rankings which preserve the directions of the data can be derived from the multivariate quantile processes proposed in Einmahl and Mason (1992), Chaudhuri (1996) and Koltchinskii (1997). These rankings certainly retain more information from the original data than the simple center-outward ranking by depth. A comparison study on these different multivariate rankings should be worthwhile. In this paper, our objective is to show that even the simple depth ranking can have far-reaching applications.

To proceed further, we need the following set of notations and definitions.

DEFINITION 2.1. The set $\{x \in \mathbb{R}^d\colon D(x) = t\}$ is called the *level set* or *contour* of depth t.

DEFINITION 2.2. The set $\{x \in \mathbb{R}^d\colon D(x) > t\}$ is referred to as the *region enclosed by the contour of depth* t, and denoted by $R(t)$.

DEFINITION 2.3. The set

$$C_p = \bigcap_t \{R(t) \colon P_F(R(t)) \geq p\}. \tag{2.1}$$

is referred to as the *pth central region*. In other words, C_p is the smallest region enclosed by depth contours to amass probability p. The boundary of C_p is referred to as the *pth level contour*, and is denoted by $Q(p)$ or $Q_F(p)$ (when we need to stress that F is the underlying distribution). In fact, if f is nonzero everywhere, $Q_F(p)$ is the contour of $\{x \in \mathbb{R}^d\colon D(x) = t_p\}$ where $P\{x \in \mathbb{R}^d\colon D(x) \geq t_p\} = p$. To distinguish it from univariate quantiles, we shall refer to $Q_F(p)$, $0 \leq p \leq 1$ as the *center-outward quantile surface*.

If F is absolutely continuous and its density function f is nonzero everywhere, then we have

$$C_p = R(t_p),$$

where $R(t_p)$ is characterized by the requirement that $P_F(R(t_p)) = p$. In theory, the empirical versions of C_p and $Q_F(p)$ should be defined by replacing F and $D(\cdot)$ in (2.1) with F_n and $D_n(\cdot)$, respectively. However, for computational and graphical convenience, we shall focus only on the $D_n(\cdot)$ values computed on the sample $\{X_1, \dots, X_n\}$ and view the convex hull containing the most central fraction p sample points as the sample estimate of C_p. More precisely, we set

$$C_{n,p} = \text{convex hull}\{X_{[1]}, \dots, X_{[\lceil np \rceil]}\},$$

where $\lceil np \rceil = np$ if np is an integer, and (1 + the integer part of np) otherwise. This simplification can also be justified by the fact that C_p is typically a convex region. In practice, the convexity is not crucial for the interpretation of the methods proposed in this paper. We call $C_{n,p}$ the *sample pth central hull*, and its boundary, denoted by $Q_n(p)$, the *sample pth level contour* or the *empirical pth center-outward quantile surface*. Here, $Q_n(p)$ may be viewed as an estimate of the quantile $Q_F(p)$. It is clear that if there are multiple points in a *de*-class, these points should belong to the same sample pth level contour for some p. A remark on the breakdown property of the bootstrap version of $Q_n(\cdot)$ is given in Section 4 of Singh (1998).

Note that, except for (4) and (7), all the above-mentioned depths are *affine invariant*. The affine invariance ensures that the depth value remains the same after the data are transformed by any affine transformation. That is, if a data point X is transformed to $\mathbf{A}X + b$, with a nonsingular $d \times d$ matrix $\mathbf{A}$ and a $d \times 1$ constant vector b, and if F_X and $F_{\mathbf{A}X+b}$ denote, respectively, the c.d.f.'s for the datum before and after the transformation, then

$$D(F_{\mathbf{A}X+b}; \mathbf{A}x + b) = D(F_X; x). \tag{2.2}$$

Although the depth values defined by (4) and (7) change by a factor of the determinant of **A** under the above transformation, the ordering induced by *OD* or *LD* remains affine invariant. Therefore, we conclude that based on any notion of depth in (1) to (7) *the deepest point is affine invariant*. Some properties related to this observation are discussed further in Proposition 3.1.

Some general properties of several of the above depths can be found in Liu and Singh (1993). The Mahalanobis depth studies the elliptical structure of a multivariate distribution as in classical multivariate analysis, and its properties have been fully developed. Needless to say, the definition of Mahalanobis depth depends on the existence of the second moments, and it is not a moment-free approach. Some asymptotics and properties of the contours of the half-space depth are given in Nolan (1992) and Donoho and Gasko (1992). The convex hull peeling approach is intuitively appealing, but its use is somewhat limited by the lack of an associated distribution theory. A discussion of convex peeling in the context of partial ordering of multivariate observations is given in Barnett (1976), and some asymptotic distribution theory is in Eddy (1982). The Oja depth is studied extensively in Oja (1983). For simplicial depth, Liu (1990), Dümbgen (1992) and Arcones, Chen and Giné (1994) cover some basic properties and a range of asymptotics, including the strong uniform convergence of $D(F_n; \cdot)$ to $D(F; \cdot)$. This convergence allows us to approximate $D(F; \cdot)$ by $D(F_n; \cdot)$ when F is unknown, an approximation which we use often to justify our methods (see, e.g., *DD*-plots in Section 7). Rousseeuw and Ruts (1996), Ruts and Rousseeuw (1996), and Rousseeuw and Struyf (1997) address the issues of computing the halfspace and simplicial depths. He and Wang (1997) investigate the convergence of depth contours formed by several notions of depth.

2.2. *Depth contours and sunburst plots.* To illustrate depth ordering and its ramifications, we use the simplicial and Mahalanobis depths to order sample points and graph some representative contours. Applying simplicial depth ordering to a sample of 500 points drawn from the bivariate standard normal distribution, we obtain in Figure 1 the sample pth level contours for $p = 0.25$, 0.5, 0.75 and 0.9. The contours are nested within one another. As the p-value increases, the contour expands. The deepest point in the sample is marked by a cross. Figure 2 shows the same set of contours for a sample from the standard bivariate exponential distribution (i.e., with independent margins and with marginal mean 1). Applying Mahalanobis depth ordering to the same normal and exponential samples leads to the contours displayed in Figures 3 and 4, respectively. As the p-value increases, the contours there also expand from the center outward. How the contour expands in terms of speed, direction and rate of change actually motivates our definitions of scale (dispersion), skewness and kurtosis in Sections 4 to 6.

There is little difference between the contours plotted in Figures 1 and 3, since the underlying sample is normal and symmetric. However, there is a noticeable difference between the two sets of contours in Figures 2 and 4 for the asymmetric exponential sample. The contours in Figure 2 are fanning out

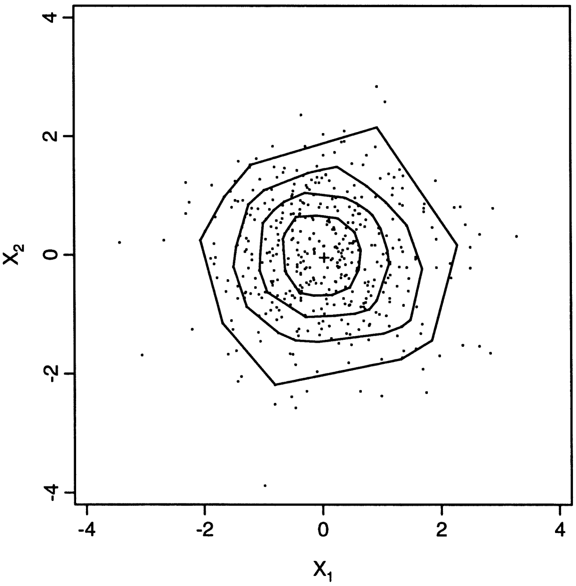


FIG. 1. *Normal contours by SD.*

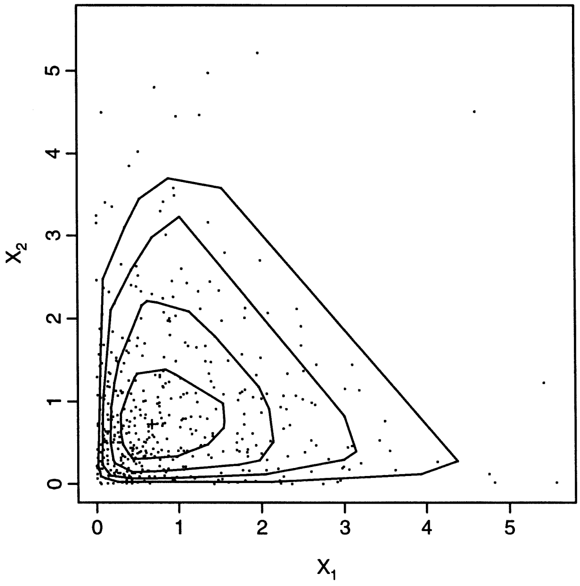


FIG. 2. *Exponential contours by SD.*

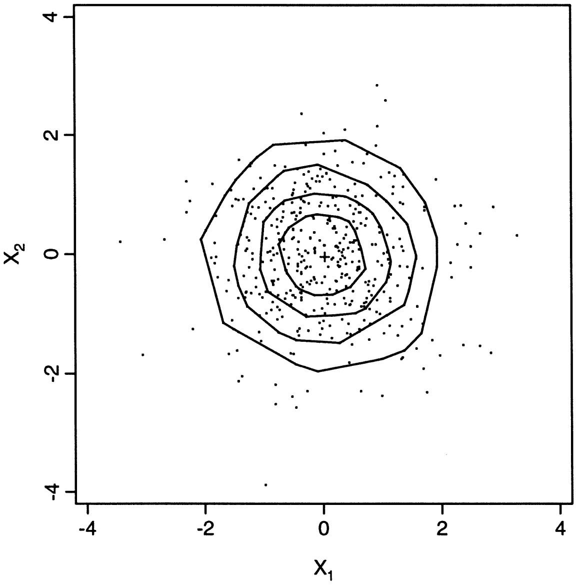


FIG. 3. *Normal contours by MD.*

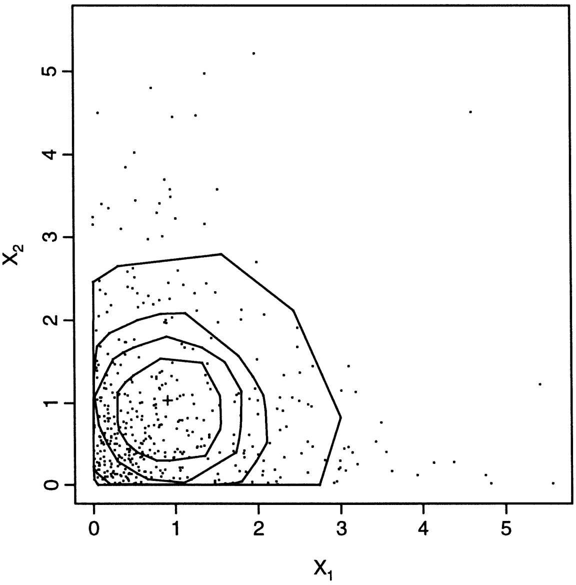


FIG. 4. *Exponential contours by MD.*

up right, reflecting the probabilistic geometry of the exponential distribution, while those in Figure 4 expand in a somewhat symmetric manner. This can be explained by the intrinsic difference between the two depths used. As the Mahalanobis depth measures the quadratic distance from each sample point to the sample mean, the contours expand outward from the sample mean, following only the distance change and ignoring the asymmetric nature of the exponential sample. On the other hand, the simplicial depth is defined to measure the relative position of a point w.r.t. to a distribution and thus to capture the underlying probabilistic geometry. Similar explanations hold for the half-space depth. The graphs of contours formed by the half-space depth are similar to those formed by the simplicial depth, and they are omitted.

The center-outward ordering induced by a data depth immediately gives rise to a simple graphic technique for presenting bivariate data sets. This technique can be viewed as a generalization of the univariate box-plot or box-and-whiskers plot. The plot resembles the sun with its rays radiating in all directions, and is thus named the "sunburst plot." For a given sample, we apply a data depth to identify its center and the central 50% sample points. We mark the center and draw the contour to enclose the 50% central hull. The rays in the plot are obtained by joining the sample points outside of the 50% central hull to the center, keeping only the segments outside the contour. The center, the central region and the rays obtained this way can be regarded as the analogues of the median, the interquartile range and the whiskers in the box-plot. Examples based on the samples used in Figures 1 and 2 are given in Figure 5(a, b).

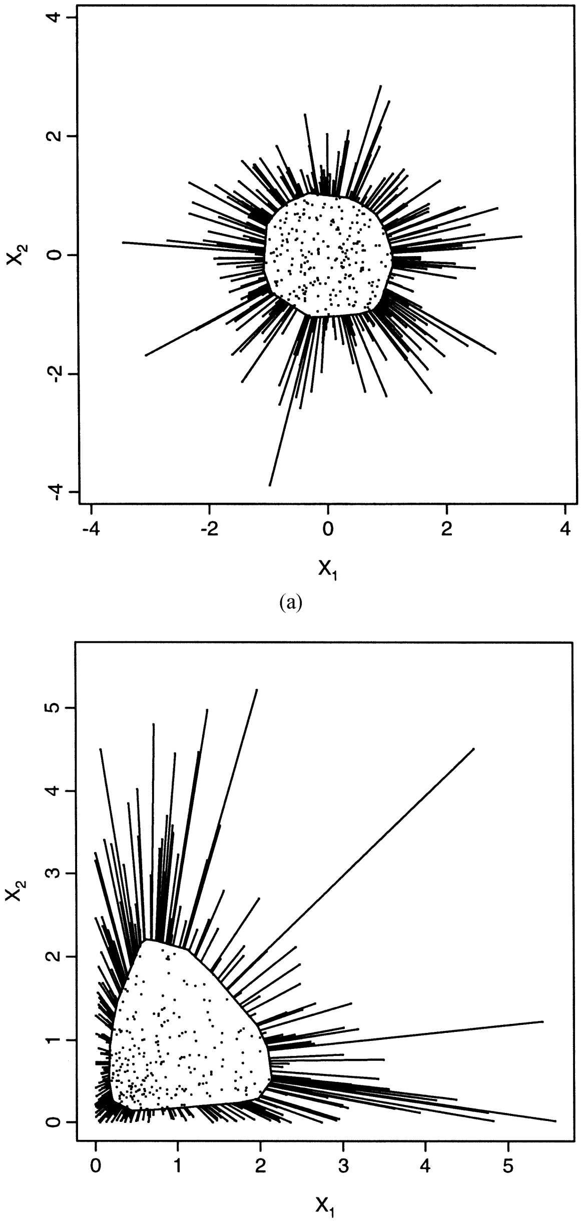
The above examples show that both the sunburst plot and the contours plot provide a quick and informative overview of the shape, concentration, spread and skewness of the underlying distribution for a given sample.

The sunburst plot is also investigated independently in Rousseeuw and Ruts (1997), where it is called *bagplot*. The idea of "fence" in the univariate box-plot is also incorporated into the bagplot.

2.3. *The Lorenz curve.* The Lorenz curve, introduced in Lorenz (1905), has been used to measure the inequality or concentration of a wealth distribution. Generally, it is defined to be the plot of $(F(x), \Omega(x))$, where F is the c.d.f. and $\Omega(x) = (1/\mu)\int_{-\infty}^{x} v\, dF(v)$. Here $\mu = \int_{-\infty}^{\infty} s\, dF(s)$. Alternatively [Gastwirth (1971)], the Lorenz curve can be defined in terms of the inverse of the probability distribution function, that is,

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)\, dt, \qquad 0 \le p \le 1, \tag{2.3}$$

where $F^{-1}(t) = \inf_v\{v \in \mathbb{R}\colon F(v) \ge t\}$. The area between the Lorenz curve and the line $L(p) = p$ is called the area of concentration, or the degree of inequality in the context of quantifying wealth distribution. The Lorenz curve will be used in later sections as a way to standardize properly a proposed descriptive statistic for specific interpretations. As such, it is more appropri-



(a)

(b)

FIG. 5. *Sunburst plot.* (*a*) *Normal sample.* (*b*) *Exponential sample.*

ately defined as follows. Let X be a nonnegative random variable in $\mathbb{R}$. Then the Lorenz curve is

$$L(p) = p \times \frac{E(X|X \le F^{-1}(p))}{E(X)}. \tag{2.4}$$

It has the following properties:

1. $L(p)$ is nondecreasing in p;
2. $L(p) \le p$;
3. And $L(p) = p$ if $F(\cdot)$ is degenerate.

3. Location.

3.1. *Median/Center*. Given a notion of data depth, there is a natural choice of location parameter for the underlying distribution, namely the deepest point or the average of the deepest points if there is more than one. For the same distribution, different notions of depth may lead to different deepest points, which can all be reasonable candidates for *multivariate medians*, as they are generally referred to [see, for example, Rousseeuw and Leroy (1987)]. They may, however, be quite different if the underlying distribution is asymmetric. An illustrative example is given later at the end of this section.

Viewing the deepest point as the sample median, we state below a general property regarding its distributional symmetry and unbiasedness.

PROPOSITION 3.1. *If the population distribution is symmetric, then the distribution of any affine invariant sample median* (*the deepest point*) *is also symmetric about the population center of symmetry.*

PROOF. Without loss of generality, we assume the population distribution is symmetric about the origin $\mathbf{0}$. Let $\mathbf{X}$ denote the given random sample from this distribution, that is, $\mathbf{X} = \{X_1, \dots, X_n\}$. Let $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$ denote the sample median derived from the sample $\mathbf{X}$. The affine invariance property of the median [cf. (2.2)] immediately implies that $\hat{\boldsymbol{\theta}}_n(-\mathbf{X}) = -\hat{\boldsymbol{\theta}}_n(\mathbf{X})$, where $-\mathbf{X} \equiv \{-X_1, \dots, -X_n\}$. Following the symmetry of the population distribution, we know that $\mathbf{X}$ and $-\mathbf{X}$ are identically distributed. This in turn implies that $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$ is symmetric about $\mathbf{0}$.

REMARK 3.1. Note that Proposition 3.1 holds even for symmetric distributions which have no moments, such as the Cauchy distribution. Pushing from the above result of symmetry of $\hat{\boldsymbol{\theta}}_n$ further to claim its unbiasedness in the sense of moment, namely showing $E\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$, we would need to require the existence of $E\hat{\boldsymbol{\theta}}_n$. This expectation exists whenever the underlying population has the first moment, following the arguments below. Note that the sample medians derived from all the depths described in Section 2 lie in the convex hull formed by the data cloud $\mathbf{X}$. This means that the sample median $\hat{\boldsymbol{\theta}}_n$ can be expressed as a convex combination of the sample points in **X**. Consequently,

$$|\hat{\theta}_{n,j}| \le \Sigma_i |X_{ij}|,$$

where $\hat{\theta}_{n,j}$ indicate the jth component of the vector $\hat{\boldsymbol{\theta}}$ and X_{ij} the jth component of the ith sample point.

The remark here immediately implies that *the half-space median*, *the simplicial median and all deepest points derived from the depths listed in Section* 2.1 *are unbiased estimators for the mean of a multivariate normal distribution*.

3.2. *Depth L-statistics*. As in the univariate case, most of the data points do not directly influence any of the medians described earlier, except for the Mahalanobis depth with respect to which the deepest point turns out to be the mean. This suggests concepts of location which are intermediate between the mean and the median, such as the trimmed means. In analogy with the concept of univariate L-statistics, that is, linear combinations of order statistics, we define here *depth-L*-statistics, (*DL*-statistics). The *DL*-statistics are robust location statistics which are designed to reduce or eliminate the weights of data at outlying positions. All the location parameters we define here, be it the median or *DL*-statistics, inherit the affine invariance property of the data depth from which they are derived.

Multivariate data, when ranked according to a data depth, may have a large number of ties. In defining *DL*-statistics, we follow the principle that data points which are in the same *de*-class, assuming the same depth value, must receive equal weights. Let $\omega(t)$ be a weight function on $0 \le t \le 1$, that is, $\omega(t) \ge 0$ and $\int_0^1 \omega(t)\,dt = 1$. Taking into account robustness, we tacitly assume that $\omega(t)$ is nonincreasing.

Given the *DO*-statistics $X_{[1]}, \dots, X_{[n]}$, we denote the stochastic process associated with these statistics by

$$\xi_n(t) = \begin{cases} X_{[i]}, & \text{for } \dfrac{i-1}{n} < t \le \dfrac{i}{n}, \\ X_{[1]}, & \text{at } t = 0. \end{cases} \tag{3.1}$$

Let $\bar{\xi}_n(t)$ be the average of $\xi_n(t)$ over the *de*-class that it belongs to. The *DL*-statistic based on the weight function $\omega(t)$ is defined as

$$DL_n = \int \bar{\xi}_n(t)\,\omega(t)\,dt. \tag{3.2}$$

The case of α-depth-trimmed mean uses the weight function $\omega(t) = (1/(1-\alpha))I_{([0 \le t \le 1-\alpha])}$, that is, $\omega(t) = (1/(1-\alpha))$ on $[0, 1-\alpha]$ and 0 otherwise. If there are no ties and if $n\alpha$ is an integer, then the α-depth-trimmed mean is $\Sigma_{i=1}^{n(1-\alpha)} X_{[i]}\{1/[n(1-\alpha)]\}$. This is, in the usual form of univariate L-statistics, $\Sigma_{i=1}^{n} X_{[i]}\omega_i$, with $\Sigma_{i=1}^{n}\omega_i = 1$. The special case of $\alpha = 0$ yields the mean as the resulting *DL*-statistic, and the limiting case of $\alpha = 1$ yields the median associated with the particular data depth used in ordering the data.

Before we present an illustrative example, we show in Proposition 3.2 that an alternative definition to the above DL_n is

$$DL_n = \int_0^1 \xi_n(t)\,\bar{\omega}(t)\,dt, \tag{3.3}$$

where $\bar{\omega}(\cdot)$ is defined as follows. Suppose that $X_{[i+1]}, \dots, X_{[i+\ell]}$ belong to the same *de*-class. Then $\bar{\omega}(t) = (\ell/n)^{-1}\int_{i/n}^{(i+\ell)/n}\omega(s)\,ds$, for all $t \in (i/n, (i+\ell)/n]$. In other words, $\bar{\omega}(t)$ is derived by averaging (w.r.t. Lebesgue-measure) $\omega(t)$ over the range of each *de*-class.

PROPOSITION 3.2. $\int_0^1 \bar{\xi}_n(t)\omega(t)\,dt = \int_0^1 \xi_n(t)\bar{\omega}(t)\,dt$.

This follows readily from the fact that the weight received by X_i in either expression is $(1/\ell)\int_{j/n}^{(j+\ell)/n}\omega(s)\,ds$, if we assume that X_i belongs to a *de*-class which contains ℓ members, and that the deeper classes contain altogether j members.

EXAMPLE 3.1. Assume that the *DO*-statistics for a sample of size $n = 8$ are (X_3), (X_1, X_2, X_7, X_8), (X_4, X_5, X_6). In other words, $X_{[1]} = X_3$, $X_{[2]} = X_1$, $X_{[3]} = X_2$, $X_{[4]} = X_7$, $X_{[5]} = X_8$, $X_{[6]} = X_4$, $X_{[7]} = X_5$ and $X_{[8]} = X_6$. To compute the 10% trimmed mean, we observe that

$$\bar{\xi}_n(t) = \begin{cases} X_3, & \text{for } 0 \le t \le \frac{1}{8}, \\ \frac{1}{4}(X_1 + X_2 + X_7 + X_8), & \text{for } \frac{1}{8} < t \le \frac{5}{8}, \\ \frac{1}{3}(X_4 + X_5 + X_6), & \text{for } \frac{5}{8} < t \le 1. \end{cases}$$

In view of (3.2) and (3.3) for DL_n, the 10% trimmed mean can be expressed as

$$\begin{aligned} &\left[\frac{1}{8}X_3\frac{1}{0.9} + \frac{4}{8}\frac{1}{4}(X_1 + X_2 + X_7 + X_8)\frac{1}{0.9}\right. \\ &\qquad \left. + \left(0.9 - \frac{5}{8}\right)\frac{1}{3}(X_4 + X_5 + X_6)\frac{1}{0.9}\right] \\ &\quad = \frac{1}{7.2}[X_3] + \frac{1}{7.2}[X_1 + X_2 + X_7 + X_8] + \frac{2.2}{3}\frac{1}{7.2}(X_4 + X_5 + X_6). \end{aligned}$$

Note that

$$\bar{\omega}(t) = \begin{cases} \dfrac{8}{7.2}, & \text{for } 0 \le t \le \dfrac{1}{8}, \\ \dfrac{8}{7.2}, & \text{for } \dfrac{1}{8} < t \le \dfrac{5}{8}, \\ \dfrac{2.2}{3}\dfrac{8}{7.2}, & \text{for } \dfrac{5}{8} < t \le 1. \end{cases}$$

The total weight is equal to 1. The outermost three data received reduced weights compared to the five inner data.

There is actually a simple recipe for computing a depth-trimmed mean: let n be the sample size and α be the trimming proportion. Move in decreasing depth order along the *DO*-statistics, and assign weight $1/n\alpha$ to each data point until the sum of the weights just crosses the threshold 1. Assume that this point belongs to the jth *de*-class. Weights can then be reassigned as follows. The weight $1/n\alpha$ is retained for all the data points belonging to classes up to the $(j-1)$th class, but the remaining weight (so that the total weight adds up to 1) is divided uniformly over all the points in the jth class. The points in higher classes are assigned zero weight. The weighted mean based on this weight distribution is the α-depth-trimmed mean.

We now turn to the population version of DL_n, which we denote by DL_F. Clearly,

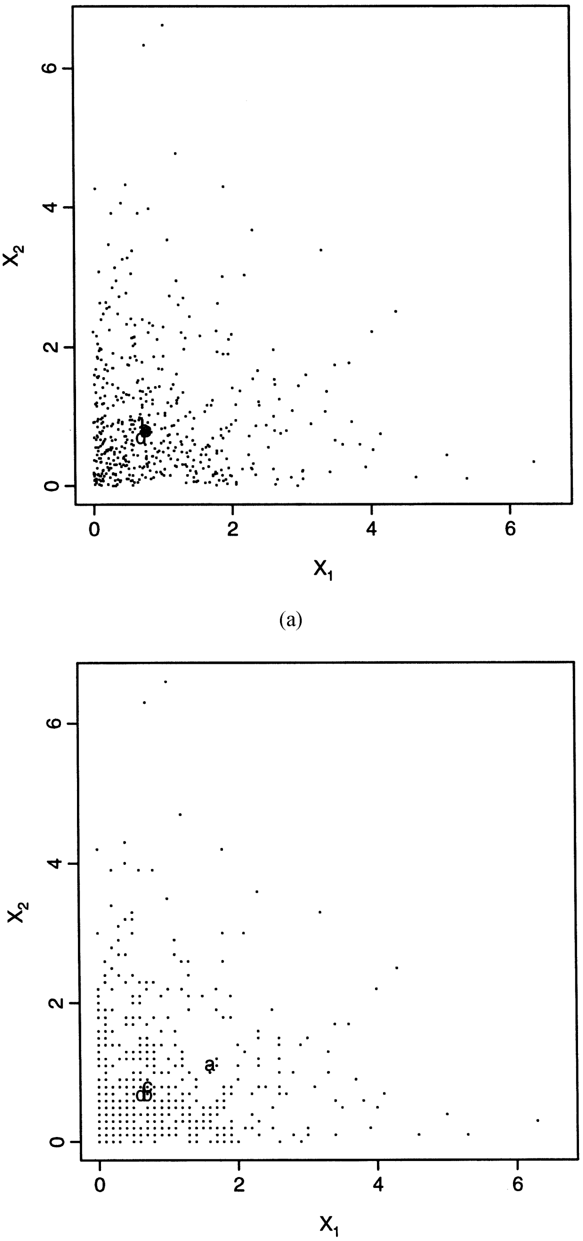
$$DL_F = \int_0^1 \overline{Q}_F(t)\,\omega(t)\,dt, \tag{3.4}$$

where $\overline{Q}_F(t)$ is the population counterpart of $\bar{\xi}_n(t)$. More specifically, $Q_F(t)$ is the tth center-outward quantile defined in Definition 2.3, and $\overline{Q}_F(t)$ is the mean along $Q_F(t)$. Alternatively, if the distribution of $D_F(X)$ is continuous, then

$$DL_F = \int_{\mathbb{R}^d} x\,\omega\big(\tilde{R}(x)\big)\,dF(x). \tag{3.5}$$

Here $\tilde{R}(x) = P_F\{I_{(D_F(X) \geq D_F(x))}\}$, namely the probability of the central region enclosed by the level set to which x belongs. Note that if $D_F(X)$ is continuous, $\tilde{R}(X)$ is uniformly distributed [cf. Liu and Singh (1993)]. We can view $\overline{Q}_F(t)$ as the conditional mean of the population random variable X, conditional on $\tilde{R}(X) = t$. It can be shown that for a symmetric distribution, $\overline{Q}_F(t) = \theta$, the center of symmetry, and thus $DL_F = \theta$. Furthermore, if the chosen data depth is affine invariant, then the linear transformation $(AX + b)$ on the population X transforms DL_F to $(A(DL_F) + b)$.

We conclude this section by observing an odd property of peeling-related trimmed means and in particular, of the convex hull peeling median. Suppose that the bivariate data are on polygonal contours centered at θ, and the data to the left of θ are much more dense than those to the upper right of θ, say with the ratio 9 to 1. Suppose that the convex hull peeling median is θ to begin with. Now if we drag the halves of the contours which are to the left of θ further left, the peeling median still remains at θ. This phenomenon can arise in practice, as in the following example. We begin with a sample of 500 observations from the standard bivariate exponential distribution. The plot of this sample is shown in Figure 6(a). The spots marked (a) to (d) (which lie virtually on top of one another in Figure 6(a), but (a) is clearly separated from the rest in Figure 6(b)) are, respectively, the medians identified by convex hull peeling, simplicial depth, half-space depth and componentwise medians. All four medians are quite close to each other. Now, if we round all data off to



(a)

(b)

FIG. 6. *Deepest points by four depths.* (*a*) *Before rounding.* (*b*) *After rounding.*

their first decimal place, then many points have common coordinate values or become collinear subsets, as seen in Figure 6(b). The rounding yields convex hulls with a larger number of points from the dense side of the data cloud (near the origin), since data which differed sufficiently to be on different hulls before the rounding are now constrained to one. Consequently, this rounding off pushes the convex-hull-peeling median in the up-right direction (see Figure 6(b)), to the extent that the resulting median seems far away from the central mass of the data set. This phenomenon reflects the fact that each round of peeling removes far more points from the dense side of the data cloud than from the sparse side. The plots in Figure 6(b) is also a good example to show that different data depths may yield different deepest points for the same dataset.

4. Scale or dispersion.

There are two common approaches to quantify the scale or dispersion of a multivariate distribution: as a matrix or as a scalar. Naturally, it is easier to grasp the magnitude of the scale by a scalar than by a matrix. However, as seen in the definition of the covariance matrix in classical multivariate analysis, a matrix scale can reveal other information, such as the orientation of the probability mass distribution and the variations of individual variates or covariates.

4.1. *Matrix form of scale dispersion.*

Define

$$\mathbf{S}_n(t) = \begin{cases} (X_{[i]} - \nu_n)(X_{[i]} - \nu_n)', & \text{for } \dfrac{i-1}{n} < t \le \dfrac{i}{n}, \\ \mathbf{O}, & \text{for } t = 0. \end{cases} \tag{4.1}$$

Here ν_n is the deepest sample point, and $\mathbf{O}$ is the zero matrix. Following the same idea of *DL*-statistics described in Section 3, and we can define a general weighted scale matrix as follows: for a given weight function $\omega(t)$, $0 \le t \le 1$,

$$\mathbf{S}_n = \int_0^1 \overline{\mathbf{S}}_n(t)\,\omega(t)\,dt. \tag{4.2}$$

We recall that the integral of a matrix is the matrix of the integrals of its entries. The overline "$^-$" indicates the averaging over all $X_{[i]}$'s which belong to the same *de*-class, as described in Section 2. The proposition below is similar to Proposition 3.1 and provides an alternative definition to (4.2).

Proposition 4.1.

$$\mathbf{S}_n = \int_0^1 \mathbf{S}_n(t)\,\overline{\omega}(t)\,dt. \tag{4.3}$$

Note that if $\omega(t) = 1$ for $0 \le t \le 1$, and if $\nu_n = \overline{X}_n$, then $\mathbf{S}_n$ is the classical sample dispersion matrix. If $\omega(t) = 1/(1-\alpha)$ on $[0, 1-\alpha]$, then $\mathbf{S}_n$ is the α-trimmed sample dispersion matrix.

As usual, we can interpret loosely the entries of $\mathbf{S}_n$ as variations and covariations, without requiring the existence of the corresponding moments. We can also imitate the definition of the so-called generalized sample variance in classical multivariate analysis by taking the determinant of our sample scale matrix $\mathbf{S}_n$ and using the resulting *single* numerical value to describe the variation expressed by $\mathbf{S}_n$. This determinant will be called the *generalized sample scale.*

It is obvious that the scale matrix $\mathbf{S}_n$ is equivariant under affine transformations. In other words, if $\mathbf{S}_Y$ denotes the sample scale matrix of the transformed data Y_i's such that $Y_i = \mathbf{A}X_i + b$, for $i = 1, \dots, n$, where $\mathbf{A}$ is a $d \times d$ nonsingular matrix and b is a $d \times 1$ constant vector, then $\mathbf{S}_Y = \mathbf{A}\mathbf{S}_X\mathbf{A}'$, where $\mathbf{S}_X$ is the sample scale matrix of the X_i's.

4.2. *Scalar form of scale/dispersion.* A different measure of scale or dispersion of a distribution can be defined by keeping track of how the pth central region C_p expands as p increases. This is a distributional property which can be characterized easily by the speed with which the data depth decreases. Recall that, for a given data depth $D(\cdot\,;\cdot)$, the level sets $\{x\colon D(F; x) = c\}$ form nested contours as the level c decreases. Thus we can define a scale curve by taking the plot of p versus $S(p)$, where

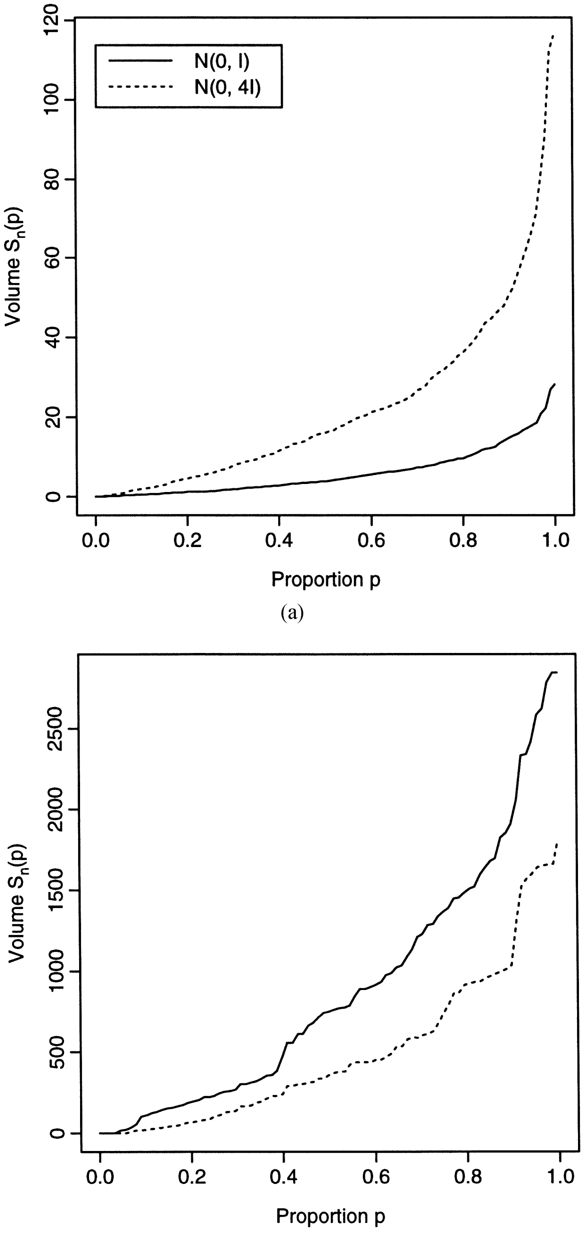
$$S(p) = \text{volume}\{C_p\}. \tag{4.4}$$

Here C_p is the pth central region. The sample scale curve, $S_n(p)$, is simply the volume of the convex hull containing $\lceil np \rceil$ most central points, that is,

$$S_n(p) = \text{volume}\{C_{n,p}\}. \tag{4.5}$$

Clearly, the faster growing $S(p)$ or $S_n(p)$ is associated with a larger scale of the distribution. To some extent, our definition of scale in terms of $S(p)$ is in the same spirit as the *spread* defined in Bickel and Lehmann (1979). That is, C_p is the central region amassing probability p, which expands as p grows. For $p_1 < p_2$, the difference $S(p_2) - S(p_1)$ reflects the central probability increment speed relative to the central-region expansion from C_{p_1} to C_{p_2}. Following this line of interpretation, we may view the distribution G as more *spread out* than F if for $p_1 < p_2$, the volume expansion $S(p_2) - S(p_1)$ under G is larger than that under F. In other words, if the scale curve of G is consistently above the scale curve of F, then G has a larger scale than F.

Figure 7(a) shows two sample scale curves, each based on a random sample of size $n = 100$, from two bivariate distributions, namely the standard bivariate normal and the normal distribution with enlarged covariance matrix $4\mathbf{I} = \binom{4\ \ 0}{0\ \ 4}$. The standard normal has a smaller scale and should be able to enclose $\lceil np \rceil$ observations by a smaller pth central region. Consequently, its scale curve, plotted as the solid curve in Figure 7(a), is consistently below the dashed scale curve for the distribution with the enlarged covariance matrix. The value of $S_n(0.5)$ indicates the area of the convex region that amasses the central 50% probability. This value is 4.354 for the



(a)

(b)

FIG. 7. *Scale curve.* (*a*) *Normal case.* (*b*) *Test scores.*

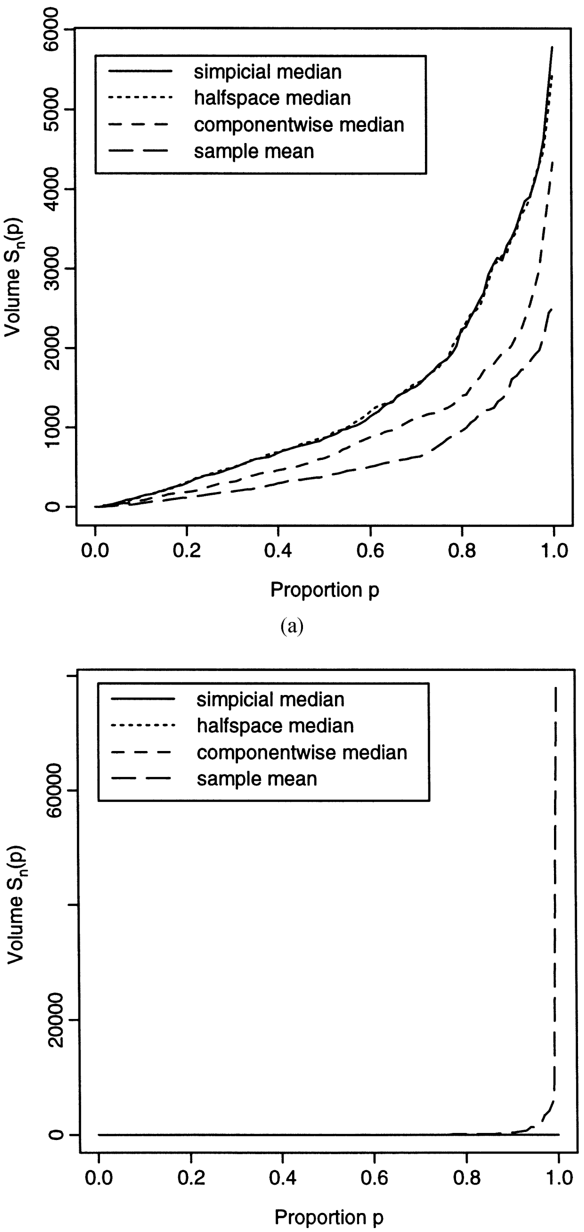
standard normal, and 17.416 for the other. This confirms that the central probability of the second distribution is much more dispersed, and that a larger region is needed to collect the same fixed amount of probability.

Two additional scale curves computed from the test data set given on page 3 of Mardia, Kent and Bibby (1979) are presented in Figure 7(b). The solid one corresponds to the test scores for mechanics and vectors taken in closed book format and the dashed one to algebra and analysis in open book format. The fact that the dashed curve lies below the solid one suggests that there is less variation in open book tests scores.

Not only does the scale curve described above allow us to visualize the scale of a multidimensional distribution via a simple one-dimensional curve, but it also gives us a tool for quantifying the evolution of the scale of a distribution as the distribution spreads out. We will return to this point later.

An application to comparing different estimators. One immediate application of scale curves is to compare the efficiencies of alternative estimators for the same parameter. We demonstrate this through the following example. Assume that we are interested in estimating the mean μ of a bivariate normal distribution, which happens to be the center as well as the median. Therefore, it is natural to consider the following four estimators: the sample mean $\overline{X}_n$, the sample componentwise median M_n^c, the sample simplicial median M_n^s and the sample half-space median M_n^H. The sample mean contains in each component the average of each component, and the sample componentwise median contains the sample median of each component. Both the simplicial and half-space medians are simply the deepest sample points based on the simplicial and half-space depths. If the sampling distribution of each estimator is known, we may generate a random sample from the distribution and plot the sample scale curves for comparison. Equivalently, we may draw K samples from the known population distribution to obtain K realizations of each estimator, namely, K of $\overline{X}_n$, K of M_n^c and so on. For each estimator, we plot its scale curve based on its K sample estimates. Figure 8(a) presents a set of simulation results, with $K = 500$ and sample size $n = 100$. The lowest and the second lowest curves correspond to the sample mean and the sample componentwise median. The other two curves, with the simplicial median case indicated by the solid one, are hardly distinguishable. These plots imply that the simplicial and half-space medians are less efficient than the sample mean, which is to be expected. That the componentwise median outperforms the simplicial and half-space medians is probably due to the lack of dependence between the two component variables.

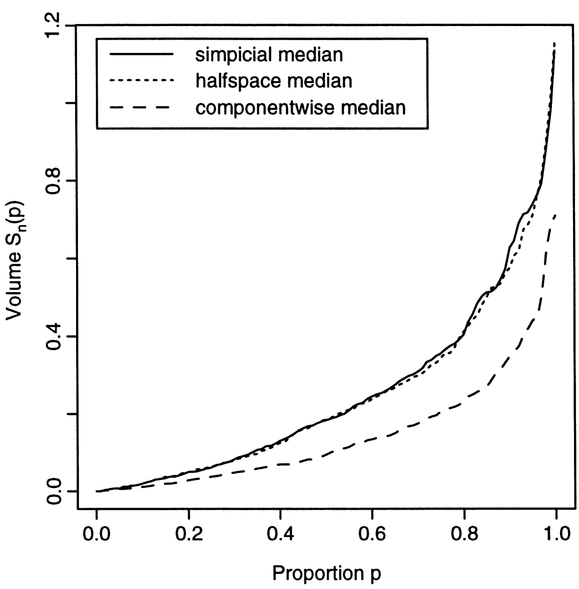
Figure 8(b) displays the exact same set of scale curves as in Figure 8(a), except that each component of the underlying bivariate distribution is now Cauchy with parameter 1. It is worth noting that the sample mean, represented by the dashed curve, is far worse than the other three, since it is so much higher in terms of scale that the other three curves are collapsed into one flat line throughout. To put the other three curves in proper perspective, we plot them on Figure 8(c). Note the sharp difference in the values along the vertical axes in Figure 8(b, c). This confirms that the sample mean is defi-



(a)

(b)

FIG. 8. *Scale curve comparison*: (*a*) *Four estimators of the normal mean*. (*b*) *Four estimators of the Cauchy center*.



(c)

FIG. 8. *(Continued). (c) Three estimators (excluding the mean) of the Cauchy center.*

nitely not appropriate for estimating the Cauchy location parameter, since the mean does not exist. The interpretation of the three curves in Figure 8(c) is similar to that of the curves in Figure 8(a).

In the above scale comparison of different estimators, if the underlying population distribution is unknown, we may rely on the bootstrap procedure to generate enough samples for each estimator, more specifically, to generate K bootstrap samples, each with sample size n, by sampling with replacement from $\{X_1, \dots, X_n\}$. From each bootstrap sample, compute the four different estimates. Repeat this procedure on all K bootstrap samples to obtain K sample points for each estimator. The K points are then used to plot the scale curve for its corresponding estimator. Figure 9 is the bootstrap version of the four scale curves in Figure 8(a). The same pattern of behaviors are more or less retained, although the curves now zigzag more due to some replicates stemming from the bootstrap procedure.

It is important to note that in order to have a meaningful comparison, the same notion of data depth should be employed at the last stage for determining the level sets used in plotting all the scale curves. This ensures that the level sets start from the same center point, and thus provides a legitimate ground for comparison. figures 8 and 9 are all carried out with level contours determined by the simplicial depth ordering.

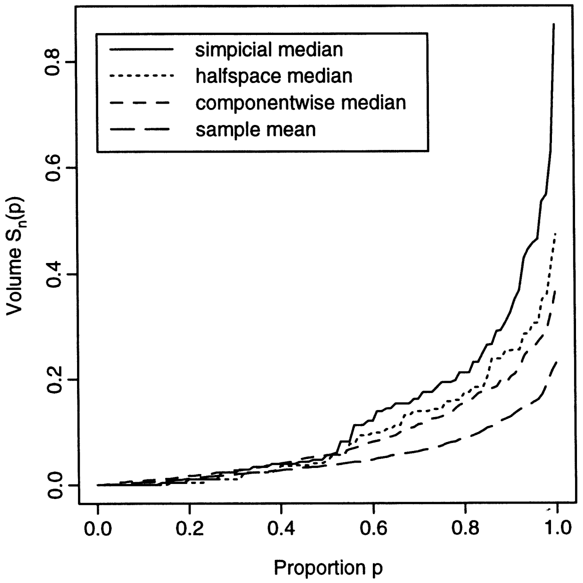


FIG. 9. *Bootstrap scale curve comparison: Four estimators of the normal mean.*

4.3. *Measures of bias.* In addition to the scale or the variance of an estimator, the bias is also an important component for measuring estimation error. We outline below some possible approaches to quantify the bias of a multivariate estimator.

The first approach is motivated by the same reasoning behind our definition of scale curve. Let θ_n be an estimator of the parameter θ. Mimicking the procedure for plotting its scale curve, we first obtain K sample estimates from the population distribution, and denote them by $\theta_n(1), \dots, \theta_n(K)$. We apply a specific data depth to these K points, and identify the deepest point $\tilde{\theta}$. We expand the central region outward from $\tilde{\theta}$ until it encloses θ, and denote this central region by $C(\tilde{\theta})$. Next, we shift the region $C(\tilde{\theta})$ along the line connecting θ and $\tilde{\theta}$ until it is centered at θ. This shifted region is denoted by $\tilde{C}(\theta)$. The *bias* of the estimator θ_n is defined to be *the volume of the intersection of* $\tilde{C}(\theta)$ *and* $C(\tilde{\theta})$. The population version of this bias is defined similarly with the true sampling distribution of θ_n. Note that in the univariate case, this notion of bias is in fact the absolute value of the usual bias. Consider the example of estimating the bivariate normal mean by either the simplicial or the half-space median. In either case, the bias computed from a sample of size 100 is nearly zero. This should be expected, since both the simplicial and the half-space medians are unbiased estimators for the bivariate normal mean (cf. Proposition 3.1 and Remark 3.1). Two other examples with more noticeable biases are also considered. The first is on the estimation of the bivariate mean ratio. Let W, X and Y be independent random variables from exponential distributions with means μ_W, μ_X and μ_Y, respectively. The parameter of interest is the vector $(\mu_W/\mu_X, \mu_X/\mu_Y)$ and the proposed estimator is $(\overline{W}/\overline{X}, \overline{X}/\overline{Y})$, where the bar indicates a sample mean. The sample size n here is 5, and all means are assumed to be 1. The obtained bias is 0.0146. The second example is on the estimation of the fourth moment of a bivariate normal random vector. Let (Z_1, Z_2) be a standard bivariate normal random vector. The parameter of interest is $(E(Z_1^4), E(Z_2^4))$, and the estimator is the vector of the componentwise fourth sample moments. The sample size n is 5. In this case, the obtained bias is 29.6. The nonzero bias in the last two examples clearly reflects the fact that the two estimators are not unbiased estimators for their population counterparts.

Again, if the underlying population is unknown and only a sample is available which yields the estimate θ_n, we may employ the bootstrap procedure to obtain many, say K, estimates based on the given sample. We denote them by $\theta_n^*(1), \dots, \theta_n^*(K)$ and repeat the same procedure as above for getting the bias by treating θ_n as the true parameter. The role of $\tilde{\theta}$ is then assigned to the center point of $\theta_n^*(i)$'s.

Another perfectly legitimate measure of bias is simply the volume of $C(\tilde{\theta})$, defined in the first approach. This generalization of bias reflects better the multidimensional nature of the estimate, since $\tilde{\theta}$ may be viewed as one realization of a whole class of estimators in the same level contour surrounding θ. This definition can be made scale-free by considering the probability content of $C(\tilde{\theta})$ instead of the volume.

One can also try to make this bias measure scale-free by taking the bias as the probability mass contained in $C(\tilde{\theta})$. In fact, this may be viewed as a generalization of the so-called *median bias* [cf. page 6, Lehmann (1991)]. Following this definition, the four simulated samples considered earlier yield the median bias 0, 0, 0.024, and 0.132, respectively.

4.4. *Bias + scale.* To account for the estimation error of an estimator, we may consider simultaneously the scale and the bias and view the sum of scale and bias as a multivariate generalization of the *mean-square-error* in the univariate statistics. Although several definitions of bias have been discussed earlier, the one using the volume of $C(\tilde{\theta})$ as a bias measure seems to be more in line with the construction of the scale curve. This would seem to suggest that the bias + scale curve be the plot of $(\text{volume}(C(\tilde{\theta})) + \text{volume}(C_{n,p}))$, $0 \le p \le 1$. However, the overlap of the two central regions, $C(\tilde{\theta})$ and $C_{n,p}$, is being used to compute both the bias and the scale. As a result the curve will begin at a nonzero level. Removing this overlap leads to the following proposal of a simultaneous measure of scale and bias. We begin the construction of the bias + scale curve for the estimator θ_n as if we were to compute the bias in the first approach. We start with many, say K, sample estimates, that is, $\theta_n(1), \dots, \theta_n(K)$. We then identify the deepest point and denote it by $\tilde{\theta}$. For a given p value, $0 \le p \le 1$, we obtain the pth central region centered at

$\tilde{\theta}$ which encloses $\lceil np \rceil$ of $\theta_n(i)$'s, and denote this region by $C_p(\tilde{\theta})$. We shift $C_p(\tilde{\theta})$ along the line connecting $\tilde{\theta}$ and the true parameter θ until $C_p(\tilde{\theta})$ is centered at θ. *The bias + scale at level p is defined to be the volume of the smallest convex hull containing the union of* $C_p(\tilde{\theta})$ *and its shifted region centered at* θ. The dashed curve in Figure 10 is the *bias + scale* curve for the simplicial median when it is used as an estimator for the bivariate normal mean. Again, as expected, the bias is nearly zero (cf. Proposition 3.1). The *bias + scale* curves for the cases of estimating exponential mean ratios and estimating normal fourth moments are shown as dashed curves in Figure 11(a, b). The bias in either case is clearly not negligible.

REMARK 4.1. Remarkably, our definitions of bias happens to be very well behaved with respect to dimensionality. The bias defined either way quickly becomes insignificant as the dimension of the distribution rises. More precisely, we first note that the vector $(\tilde{\theta} - \theta)$ typically has length $O(n^{-1})$. Consequently, the volume of $C(\tilde{\theta})$, as well as the other bias measures, are all $O(n^{-d})$, where d is the dimensionality of the distribution. This desirable property is also confirmed by our simulation results with a moderately large sample size, even in the case $d = 2$. This observation suggests that more attention be given to the scale in measuring the estimation error of a multivariate estimator.

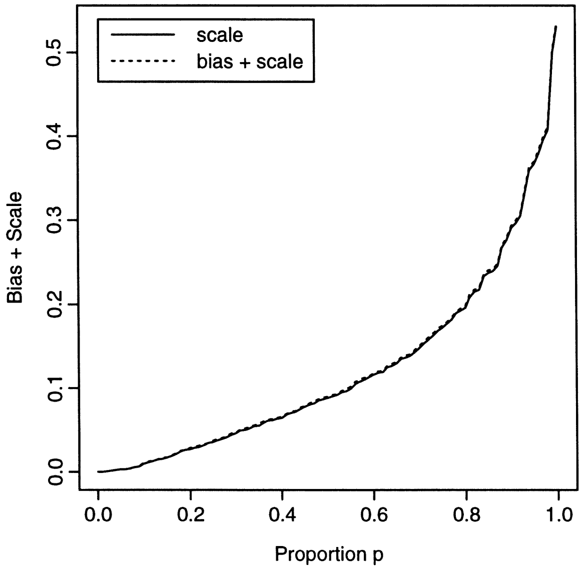
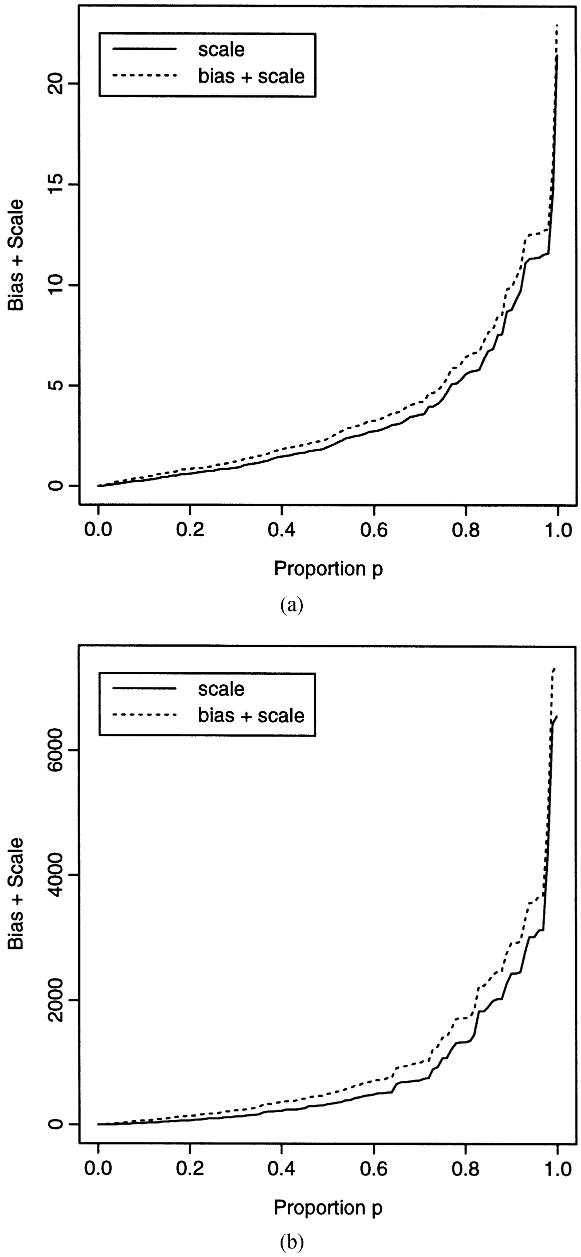


FIG. 10. *Estimating normal mean by simplicial median.*



(a)

(b)

FIG. 11. (*a*) *Estimating exponential mean ratio.* (*b*) *Estimating fourth normal moment.*

5. Skewness. Skewness is a measure of deviation from symmetry. For a multidimensional distribution, we consider the following four types of symmetry:

A. *Spherical symmetry*. The distribution of the random variable X is said to be spherically symmetric about the point c if the distributions of $(X - c)$ and $\mathbf{U}(X - c)$ are identical, for any orthonormal matrix $\mathbf{U}$.
B. *Elliptical symmetry*. The distribution of the random variable X is said to be elliptically symmetric about the point c if there exists a nonsingular matrix $\mathbf{V}$ such that $\mathbf{V}X$ is spherically symmetric about c.
C. *Antipodal symmetry*. The distribution of the random variable X is said to be antipodally symmetric about the point c if the distributions of $(X - c)$ and $-(X - c)$ are identical.
D. *Angular symmetry*. The distribution of the random variable X is said to be angularly symmetric about the point c if, conditional on $X \neq c$, the distributions of $(X - c)/\|X - c\|$ and $-(X - c)/\|X - c\|$ are identical.

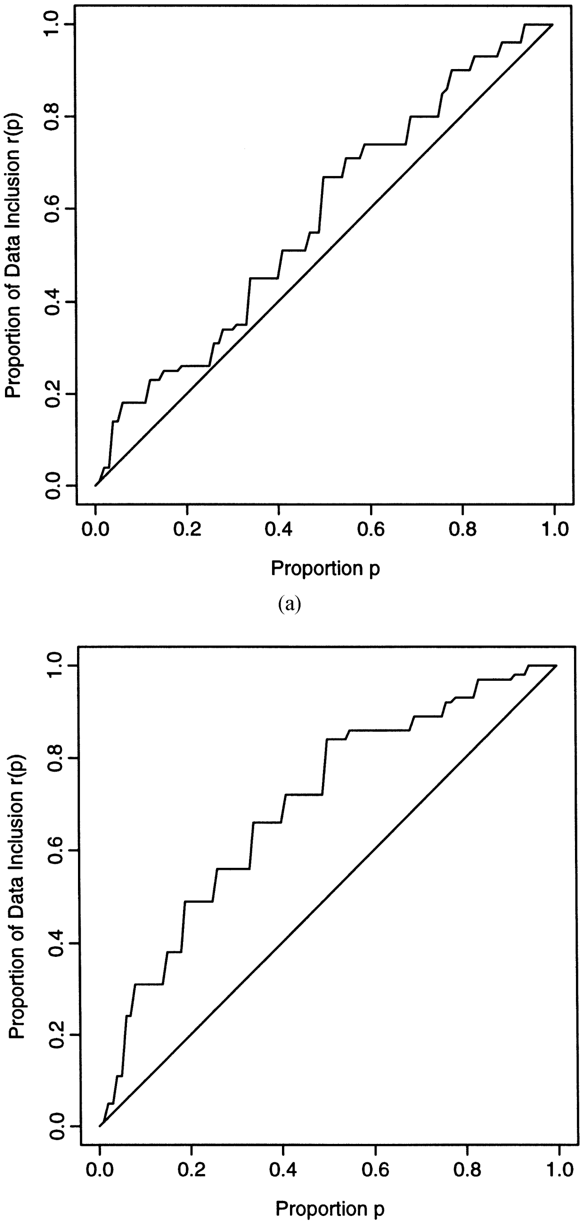
The preceding four notions of symmetry are increasingly less restrictive. That is, spherically symmetric distributions are elliptically symmetric, elliptically symmetric distributions are antipodally symmetric and antipodally symmetric distributions are angularly symmetric.

A detailed investigation of elliptical symmetry can be found in Beran (1979). The recent paper by Beran and Millar (1997) contains an extensive study of many multivariate symmetry models (where antipodal symmetry is called *simple symmetry*). Some qualitative *directional* measure of skewness are explored in Avérous and Meste (1997). Our focus is on an overall *quantitative* measure of skewness of various types, which seems simpler and more practical in comparison.

For each type of symmetry in (A) to (D), we can define a measure of skewness as the deviation from that particular symmetry. Thus, there are four measures of skewness.

A. Skewness as departure from spherical symmetry. For each pth central hull $C_{n,p}$, $0 \leq p \leq 1$, we find the smallest sphere containing $C_{n,p}$ and determine the fraction of the data within that sphere. This fraction is plotted with respect to its level p. In principle, if the underlying distribution is spherically symmetric, the resulting plot should be the diagonal line from $(0, 0)$ to $(1, 1)$. *The area of the gap between the plot and the diagonal line, denoted by* Δ_n, *is thus a measure of skewness due to lack of spherical symmetry.*

The zigzag plot in Figure 12(a) is the proposed spherical skewness for a standard bivariate normal distribution based on a sample of size 100. The plot closely follows the diagonal line, and indicates that the standard normal is spherically symmetric. The sample Δ_n value in this case is 0.0779, nearly zero. The plot in Figure 12(b) is for a bivariate normal distribution whose first component variable is a standard univariate normal and the second component is the sum of the first component variable and another independent standard univariate normal variable. The plot arches away from the diagonal



(a)

(b)

FIG. 12. *Spherical skewness.* (*a*) *Spherically symmetric case* ($\Delta_n = 0.0779$). (*b*) *Spherically asymmetric case* ($\Delta_n = 0.197$).

in the middle, and suggests a nonspherically symmetric distribution. This is indeed in the case, since the distribution is elliptical but not spherical. The sample Δ_n here is 0.197.

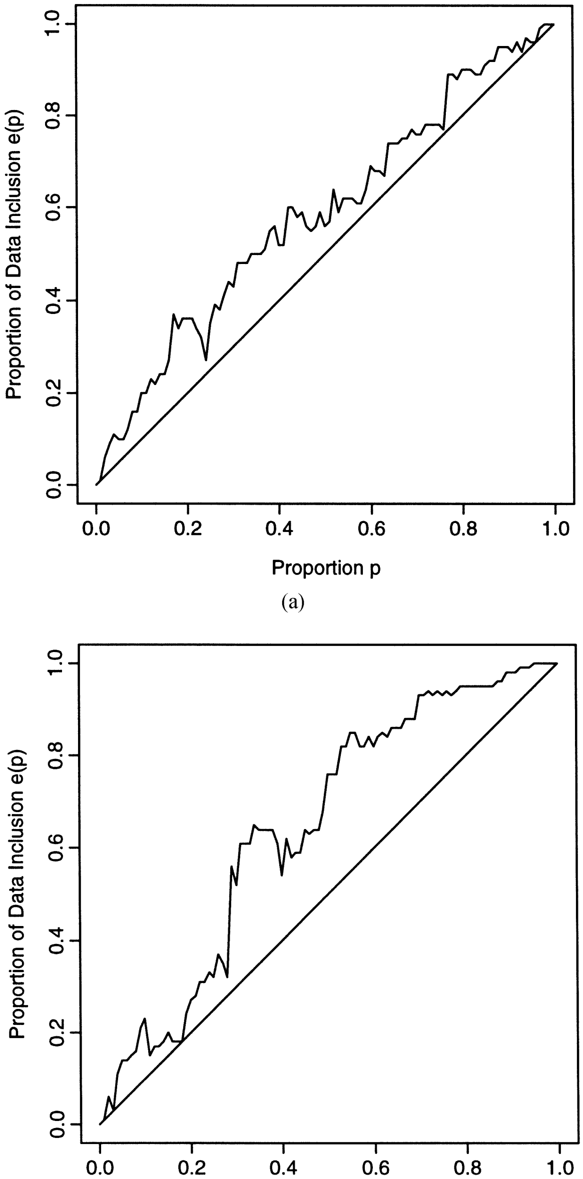
B. Skewness as departure from elliptical symmetry. For each pth central hull $C_{n,p}$, $0 \le p \le 1$, we first apply the transformation $\mathbf{S}_{n,p}^{-1/2}$ to each data point, where $\mathbf{S}_{n,p}$ is the scale matrix [cf. (4.1)] derived from the data points inside the level set $C_{n,p}$. We then proceed as if we were to measure the skewness as the departure from spherical symmetry on the transformed data, namely by plotting the fraction of the transformed data falling inside the smallest sphere containing the transformed level set. Again, we expect the plot to hug the diagonal line closely under elliptical symmetry.

Figure 13(a) is the elliptical skewness plot from the standard bivariate normal sample examined in Figure 12(a). The plot follows closely and often touches the diagonal, with $\Delta_n = 0.0852$. The elliptical skewness plot for the sample in Figure 12(b) shows a similar result with $\Delta_n = 0.0848$. These observations confirm that the two underlying distributions are elliptically symmetric. Figure 13(b) presents the plot based on a sample from the standard bivariate exponential distribution. It shows that the exponential distribution is not elliptical, since the plot deviates substantially from the diagonal line, not touching it in the upper two-thirds range of p. The Δ_n here is 0.1464.

C. Skewness as departure from antipodal symmetry. Since antipodal symmetry means that the distribution coincides exactly with its image under reflection about the point of symmetry, we define the corresponding skewness as the degree of nonoverlapping between the central region and its reflection image. That is, for each pth central hull $C_{n,p}$ we calculate *the fraction of the data points falling inside the intersection of $C_{n,p}$ and its reflection*. Under antipodal symmetry, this fraction is exactly the same as the level of the central region since the two regions coincide. The graph of this fraction versus the level p should therefore be the diagonal line from $(0,0)$ to $(1,1)$.

We obtain the plots for the samples drawn from five bivarite distributions whose component variables are independent and have the following marginal distribution: (a) standard normal, (b) Cauchy(1), (c) uniform $[-0.5, 0.5]$, (d) exponential (1) and (e) gamma (5) (5 is the shape parameter). The first three distributions are antipodally symmetric, and their plots, as expected, hug the diagonal line closely. Since the three plots are similar, we present only the one from Cauchy distribution in Figure 14(a). The distributions in (d) and (e) are antipodally asymmetric and their plots deviate significantly from the diagonal line, especially towards the upper right. Figure 14(b) is the plot based on the sample from the exponential distribution in (d).

D. Skewness as departure from angular symmetry. To obtain a measure of skewness as deviation from angular symmetry, we proceed as follows. We apply a specific data depth to identify the deepest point. We calculate the half-space depth of the deepest point w.r.t. only the data points within each central region C_p, $0 \le p \le 1$. We then plot the value of ($\frac{1}{2}$, the half-space



(a)

(b)

FIG. 13. *Elliptical skewness. (a) Elliptically symmetric case* ($\Delta_n = 0.0852$). *(b) Elliptically asymmetric case* ($\Delta_n = 0.1464$).

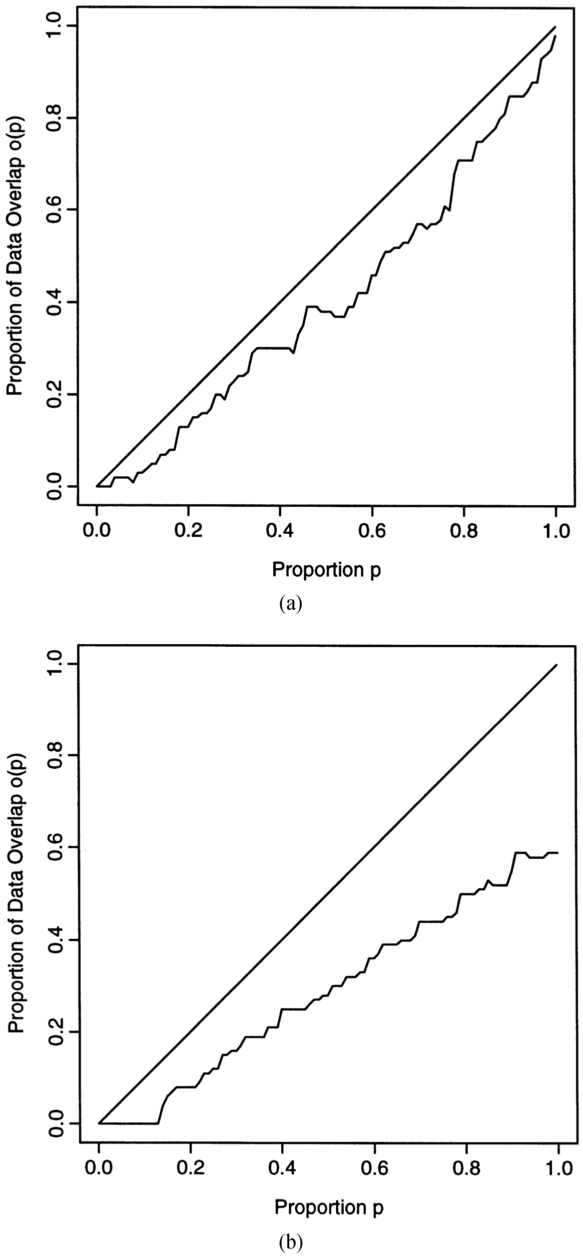
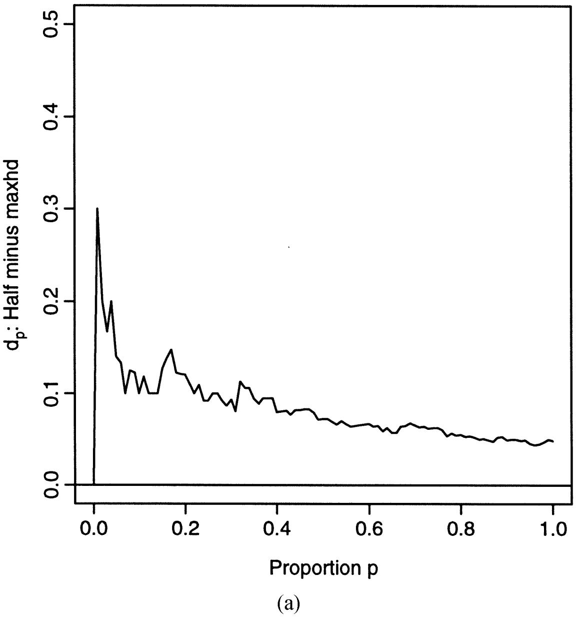


FIG. 14. *Antipodal skewness. (a) Antipodally symmetric case. (b) Antipodally asymmetric case.*

depth at the deepest point) w.r.t. level p for $0 \leq p \leq 1$. The deviation of the plot from the horizontal line at value zero is the desired measure of skewness due to lack of angular symmetry. To see this, we observe that the half-space depth value in essence identifies the pair of half-spaces whose probability contents differ the most. Under angular symmetry, all pairs of half-spaces joining at the deepest point have equal probability mass $\frac{1}{2}$, which implies that the deepest point has the half-space depth value exactly $\frac{1}{2}$ and it is also the center of angular symmetry. Therefore, the deviation of the half-space depth at the deepest point from the value $\frac{1}{2}$ is a measure of the departure from angular symmetry of the empirical distribution determined by the sample points within each level set. We obtain angular skewness plots for samples of size 100 from the following distributions: (a) the standard bivariate normal; (b) the standard bivariate normal with its right-hand side compressed to its half, that is, the distribution of (X_1, X_2) is given by $X_1 = Y_1/2$ and $X_2 = Y_2/2$ if $Y_1 > 0$, and $X_1 = Y_1$ and $X_2 = Y_2$ if $Y_1 \leq 0$, where Y_1 and Y_2 are two independent univariate standard normal random variables; (c) the two component variables are independent Cauchy (1); (d) the two component variables are independent exponential with mean 1; and (e) the two component variables are independent chi-square with mean 1. All graphs begin to stabilize after some initial fluctuation in the range of small p values. This is expected since the half-space depth there is an estimate based on only a small sample, and thus highly unstable. For practical purposes, we make use only of the graph in the range of $p \geq 0.4$. The first three distributions are angularly symmetric. Their graphs stay below 0.1 after $p > 0.4$ and edge closer to zero if the sample size increases, as shown in Figure 15(a), which is the angular skewness plot from a sample drawn from the compressed normal distribution in (b). The distributions (d) and (e) are angularly asymmetric, and the plots from their samples maintain a substantial gap throughout, as seen in Figure 15(b), which is the angular skewness plot with the underlying distribution (e).

REMARK 5.1. A careful examination of the graphs in Figure 15(a, b) reveals that their maximum half-space depths never quite reach $\frac{1}{2}$, even though the three distributions there are angularly symmetric. This is in part due to the slow convergence of the sample half-space depth, which is particularly acute for a small or medium sample size. It is also a consequence of having restricted ourselves to identifying the deepest of the sample points instead of the overall deepest point on the space $\mathbb{R}^2$. The overall deepest point is more efficient and should provide a faster convergence to the true half-space depth. However, it is highly impractical, if not impossible, to search the infinitely many points in space to locate the “deepest” point. A practical compromise here would be simply to include some natural efficient location estimators, such as the sample mean, in our sample during the search for the deepest point. This inclusion should be particularly helpful in distinguishing plots when the sample size is small.



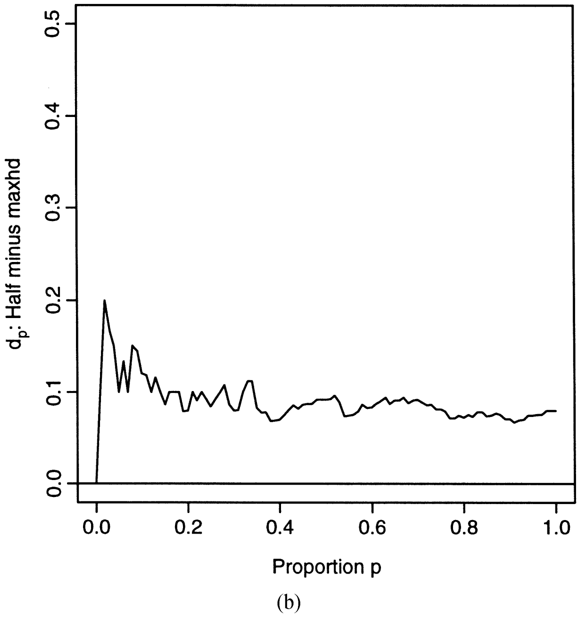
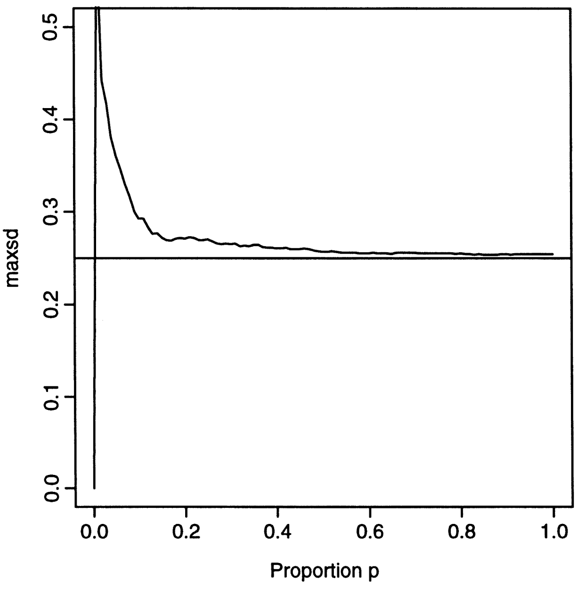
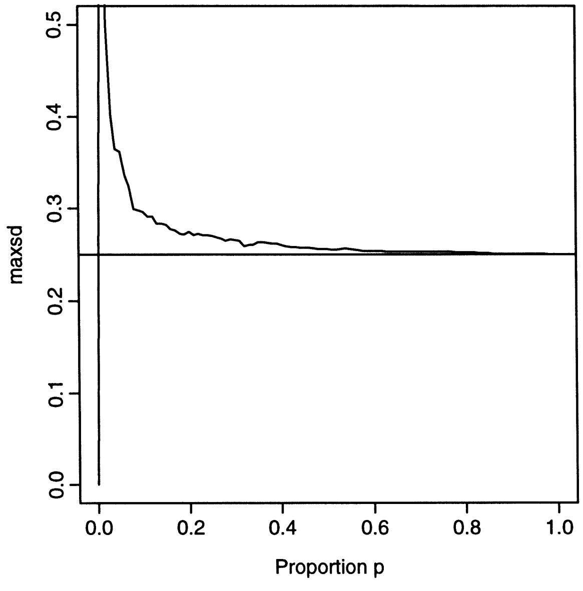


FIG. 15. *Angular skewness.* (*a*) *Angularly symmetric case.* (*b*) *Angularly asymmetric case.*



(c)



(d)

FIG. 15. *(Continued). (c) Angularly symmetric case. (d) Angularly asymmetric case.*

Remark 5.2. For the above reasons, the angular skewness plots based on the half-space depth in some cases may appear to be less than decisive. In such a case, one can consider an alternative angular skewness plot which plots the simplicial depth value at the "deepest" point. It has been shown in Liu (1990) that the simplicial depth at the point of angular symmetry is 2^{-d}, where d is the dimension of the distribution. Therefore, one can easily plot the simplicial depth of the deepest point within each pth central region. Taking into account the discussion in Remark 5.1, we may also treat the sample mean as one of the sample points in computing the highest depth value to enhance the results. Note that the computer program we use here [cf. Rousseeuw and Ruts (1996)] treats a boundary point of a triangle as being contained in the triangle. This leads to a systematic overcount of the triangles containing each sample point. This fact does not affect the ordering of the sample points, since their simplicial depth values are all equally inflated. This inflation in simplicial depth values in the finite sample calculation explains why all graphs in the alternative angular skewness plots for angularly symmetry distributions (except for the initial zero value for the central region which contains only two points) are always above 2^{-d}, and seem to converge from above to 2^{-d} toward the end. An illustrative example is given in Figure 15(c), which shows the angular skewness based on plotting the highest simplicial depth for the same sample used in Figure 15(a). Note that the graph here never quite reaches 0.25. On the other hand, the graph in Figure 15(d) which plots the highest simplicial depth for the sample used in Figure 15(b) does reach 0.25 fairly quickly.

6. Kurtosis, relative-spread, heavy tailedness. These three expressions are synonymous in this section. On the real line, kurtosis is defined to be the ratio of the fourth central moment to the square of the second central moment, see, for example, Section 3.31 of Kendall, Stuart and Ord (1987). It is interpreted as an inverse of the "peakedness" of a distribution or as a measure of the overall spread relative to the spread in the tails. Unless the concept is viewed relative to the scale, it can be quite confusing. Consider the following two univariate distributions: $\mathscr{N}(0, 1/100)$ and $U[-1, 1]$. Looking at the two densities, the normal curve seems obviously more peaked. However, it is the uniform distribution which is relatively more peaked after equalizing the scales. In fact, the kurtosis is 1.8 for a uniform distribution and 3 for a normal distribution.

We describe below four approaches for defining notions of kurtosis to measure heavy tailedness of a multidimensional distribution. Each approach is illustrated by four bivariate distributions, whose components are: (i) uniform, (ii) normal, (iii) double exponential and (iv) Cauchy. The first two approaches give very straightforward curves. The others give more complicated sets of curves, but they do provide more information about tail probability behaviors. Again, all our approaches are moment-free if the underlying depth is. Furthermore, these concepts are invariant under location and scale changes, that is, the plots are the same for $\mathbf{X}$ and $(\mathbf{AX} + b)$ if $\mathbf{A}$ is a nonsingular matrix and b is a vector.

6.1. *Kurtosis in the form of a Lorenz curve of sample Mahalanobis distances.* Given a sample $\{X_1, \dots, X_n\}$, we compute their Mahalanobis distances from the deepest point μ_n determined by a specific data depth

$$Z_i = (X_i - \mu_n)'\mathbf{S}_n^{-1}(X_i - \mu_n), \qquad i = 1, \dots, n.$$

Here $\mathbf{S}_n$ is either a dispersion matrix defined in (4.2) or the classical sample covariance matrix. Consider the Lorenz curve (cf. Section 2.3) of $\{Z_1, \dots, Z_n\}$. The area between this curve and the diagonal line is a measure of kurtosis, since the curve for a distribution with a higher kurtosis falls further below the diagonal line and gives a larger area in between. More precisely, the Lorenz curve of the Z_i's can be constructed as either L_p or L_p^* expressed in (6.1) and (6.2),

$$L_p = \frac{\Sigma_1^{[np]} Z_i}{\Sigma_1^n Z_i}, \tag{6.1}$$

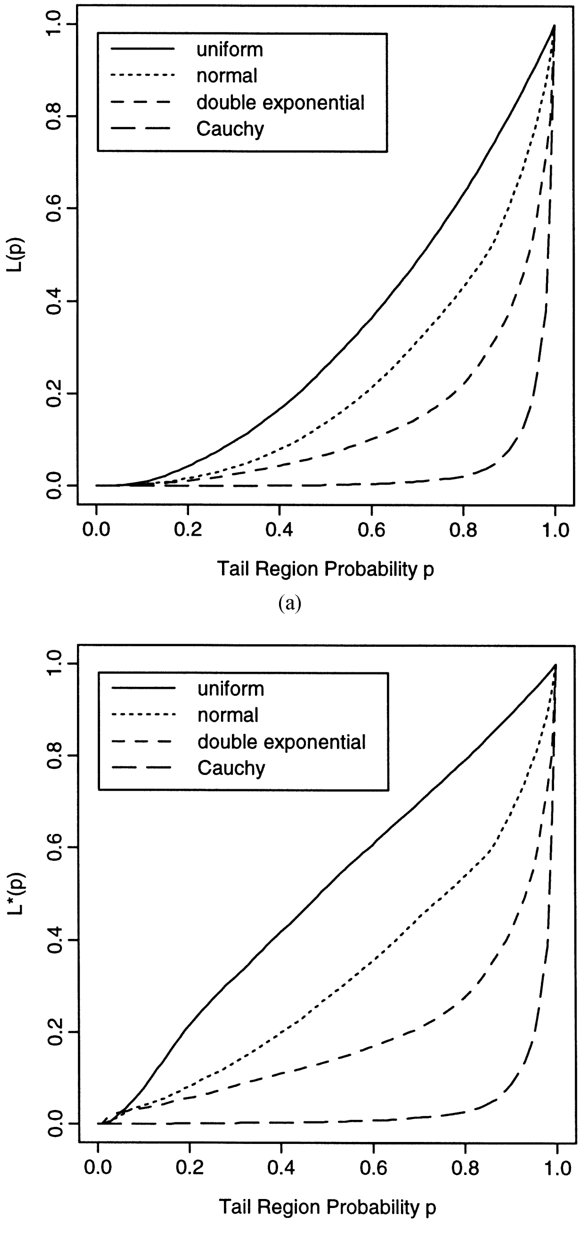
$$L_p^* = \frac{\Sigma_1^{[np]} Z_i / [np]}{\Sigma_i^n Z_i / n}. \tag{6.2}$$

The definition of L_p follows that of the Lorenz curve in (2.4), which is the proportion of the mean confined to the central hull $C_{n,p}$ to the overall mean. Figure 16(a) displays the L_p's for samples from the four distributions (i)–(iv). The L_p of the uniform sample is the closest to the diagonal line. The other three, in the order listed, move gradually further away from the diagonal line. The area in between and thus the kurtosis increases accordingly. The visual effect of the four curves in Figure 16(a) can be enhanced by modifying L_p to L_p^*, which leads to a more pronounced separation of the curves. This effect is shown clearly in the resulting graphs of L_p^*'s in Figure 16(b). Note that $L_p \approx pL_p^*$, and L_p^* is the proportion of the subtotal in $C_{n,p}$ to the grand total. It is worth noting that both the L_p and the L_p^* curves are completely free of location and scale.

6.2. *Kurtosis and a Lorenz curve with density function as wealth.* Let $f(\cdot)$ be the probability density function of the random variable X on $\mathbb{R}^d$. Define a new random variable $T = f(X)$, and denote its c.d.f. by $H(\cdot)$. The kurtosis of the distribution $F(\cdot)$ in terms of the Lorenz curve associated with a positive random variable in (2.3) can be measured as follows. Define

$$L(p) = \frac{1}{E(T)} \int_0^{H^{-1}(p)} t\, dH(t) \quad \text{for } 0 \le p \le 1. \tag{6.3}$$

Note that $E(T) = \int f(x)\, dF(x)$, and $\int_0^{H^{-1}(p)} t\, dH(t) = \int_{C_p^*(LD)} f(x)\, dF(x)$ since $\{T \le H^{-1}(p)\} = \{x \in C_p^*(LD)\}$. Here $C_p^*(LD)$ is the region of the $100p\%$ smallest likelihood depth $f(x)$, and may be viewed as the *pth tail region*. In



(a)

(b)

FIG. 16. (*a*) *Kurtosis using* L_p [*Definition* (6.1)]. (*b*) *Kurtosis using* L_p^* [*Definition* (6.2)].

our usual notations for depth central regions, $C_p^*(LD)$ is simply the complement of the $(1-p)$th central region determined by the likelihood depth. Thus, it is obvious that (6.3) is equivalent to

$$L(p) = \int_{C_p^*(LD)} f(x)\,dF(x) \bigg/ \int f(x)\,dF(x), \tag{6.4}$$

for $0 \le p \le 1$. The right-hand side of (6.4) can be viewed as a Lorenz curve with $f(x)$ as the wealth of the point x. If we plot $L(p)$ w.r.t. p. the area between $L(p)$ and the diagonal line is a measure of the degree of heavy tailedness and hence of kurtosis.

To motivate this concept of kurtosis, we recall from (2.4) that the measure of concentration of the distribution $H(\cdot)$ can be expressed as $L(p) = p \times (E(T|T \le H^{-1}(p)))/E(T)$. Since $\{T \le H^{-1}(p)\} = \{x \in C_p^*(LD)\}$,

$$L(p) = p \times \frac{E\big(f(X)|X \in C_p^*(LD)\big)}{E(f(x))}. \tag{6.5}$$

Without the factor p, the right-hand side of (6.5) can be viewed as the ratio of the overall scale to the conditional scale of the pth tail region. It is reasonable to view $[\int f dF]^{-1}$ as the overall scalar scale since $[\int f dF] =$ constant $\times |\Sigma|^{-1/2}$ within the same location and scale family. Similarly, $[\int_{C_p^*(LD)} f(x)\,dF(x)/p]^{-1}$ can be viewed as the conditional scalar scale for the pth tail region where the $100p\%$ probability is more sparsely distributed.

Another way of interpreting this kurtosis measure is to view the uniform distribution as a model case which divides the wealth of density equally to 100% of the population and the compare the other distributions with the model case. A heavy tailed distribution has a sizable fraction of its population with much lower wealth density than the modal points do. In other words, it spreads a small amount of tail probability (i.e., wealth in the context of wealth distribution) over a large domain (i.e., portion of the population). This results in an overall lowering of the density function in the tail region. The integral of the density in the tail zone is divided by the integral of the density over the entire range to make the function $L(p)$ free of scale.

The kurtosis in terms of $L(p)$ is plotted in Figure 17 for the distributions (i)–(iv). The conclusion drawn from them is similar to the one drawn from Figures 16(a, b).

6.3. *Shrinkage plots.* Let s, $0 < s < 1$, be a predetermined shrinkage level. Consider the pth central hull $C_{n,p}$. We shrink its boundary towards it center (the deepest point) by a factor of s. We refer to the region enclosed by this shrunken boundary as the shrunken hull, and denote it by $C_{n,p}^s$. We then count the number of data points enclosed in $C_{n,p}^s$. Let $a_s(p)$ (or simply $a(p)$ when s is fixed) denote this fraction of data points. The shrinkage plot is the plot of $a(p)$ versus p. For the same value of p and s, we can compare the kurtosis of two distributions by comparing their shrinkage plots. For a fixed shrinkage level, a distribution with heavier tails tends to have higher $a(p)$,

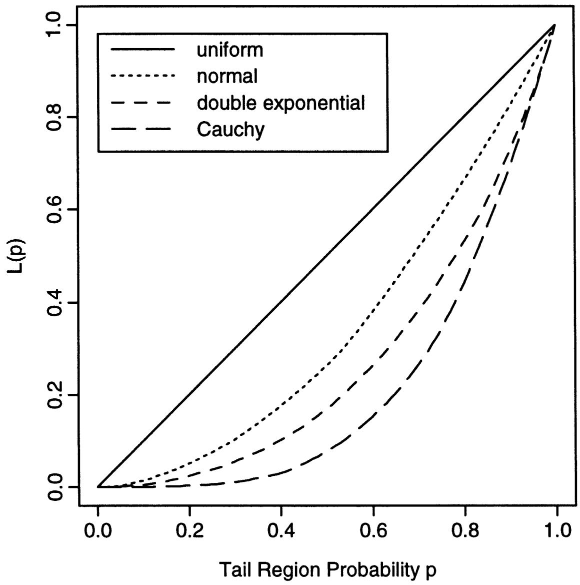


FIG. 17. *Kurtosis using* $L(p)$ [*Definition* (6.5)].

especially in the range of larger p values. This corresponds to the fact that a higher concentration of probability mass in the tail region makes it harder to lose a substantial number of data points by shrinking. Evidently, a more compelling conclusion can be drawn if this shrinkage plot comparison is carried out with many different values of s, say $s = 0.25, 0.5, 0.75$. However, we find that the common features of the plots usually do not vary greatly. Therefore, we present our simulated plots only for $s = .5$ and draw conclusions from there. Figure 18 contains the shrinkage plots for the four distributions displayed in Figure 16. The horizontal axis represents the pth central hull $C_{n,p}$, and the vertical axis represents $a(p)$. Except for a few minor crisscrosses in the range of small p, overall the four plots show clearly very different speeds of increase for $a(p)$. From low to high, they correspond to the four distributions (i) to (iv) in that order. This finding is consistent with the ones derived from Figures 16(a, b) and 17.

To standardize the magnitude of the difference in the plots above, we can force them to have the same range on $[0,1]$ by converting them into the following Lorenz curves. Consider a fixed $C_{n,p}$. After the s-shrinkage, we let $V(s)$ denote the loss of area or volume (in fraction of the total area or volume of $C_{n,p}$) and $l(s)$ the loss of inclusion of data (in fraction of the total number of data in $C_{n,p}$). Now we plot $V(s)$ on the horizontal axis and $l(s)$ on the vertical axis. Note that $V(s)$ and $l(s)$ increase from 0 to 1 as s decreases from 1 to 0, and also that $l(s)$ increases as $V(s)$ increases. For a unimodal distributions, $l(s) \leq V(s)$. This clearly generates a Lorenz curve. The area

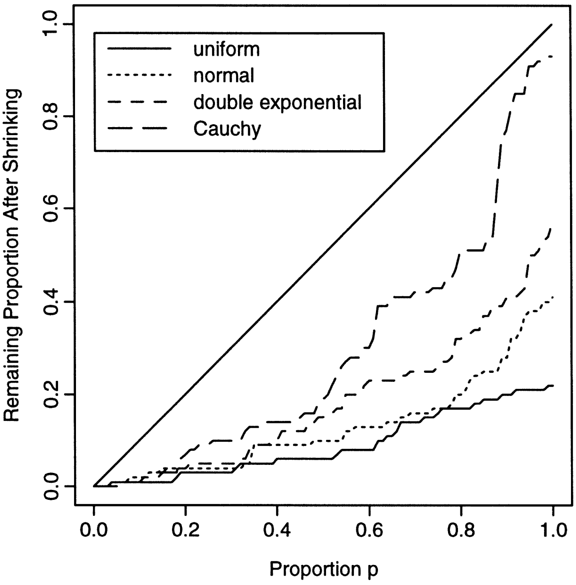


FIG. 18. *Kurtosis as shrinkage plots.*

between the curve and the diagonal line can be viewed as a measure of kurtosis at the pth central region. A plot of this kurtosis value for the range of p from 0 to 1 should give a complete picture of the kurtosis measure of a distribution.

In the Appendix, we provide a proof of the affine invariance of the above shrinkage plots.

6.4. *Fan plots.* Consider the pth central region $C_{n,p}$. For a given t, $0 \le t \le 1$, form the convex hull of the $100t\%$ central-most data points in $C_{n,p}$ and denote it by $C_{n,p}(t)$. Let $b_p(t) = \{\text{volume } (C_{n,p}(t))/\text{volume } (C_{n,p})\}$. The plot of $b_p(t)$ versus t for a fixed p is a measure of heavy tailedness in the pth central region of the distribution. For a fixed value of p, the value of $b_p(t)$ ranges from 0 to 1. Repeat this procedure for a set of values of p. The collection of these plots resembles the shape of a fan. For a distribution with heavier tails $b_p(t)$ tends to be smaller, and thus its fan plot is more spread out. Figures 19(a)–(d) show the fan plots for the four distributions. They are consistent with all earlier results in this section.

7. Data-depth plots. In this section we focus on graphical comparisons of two multivariate distributions based on data-depth plots of their samples. Specifically, we consider the *DD*-plot which plots the depth values of the combined sample under the two corresponding empirical distributions. If the two given distributions are identical, then these plots are segments of the

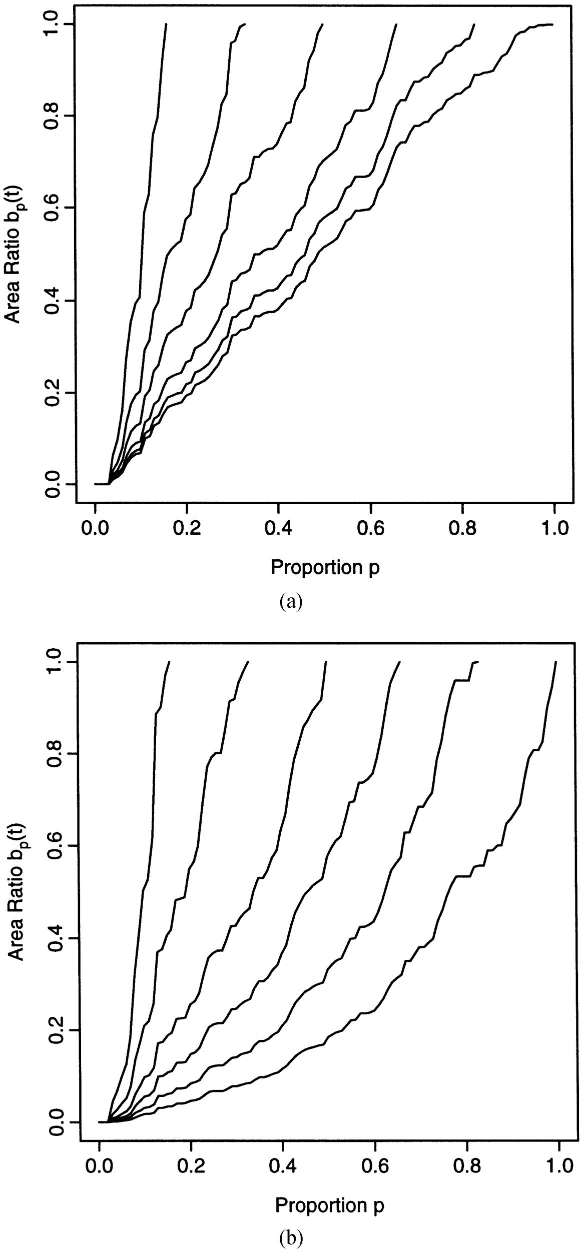
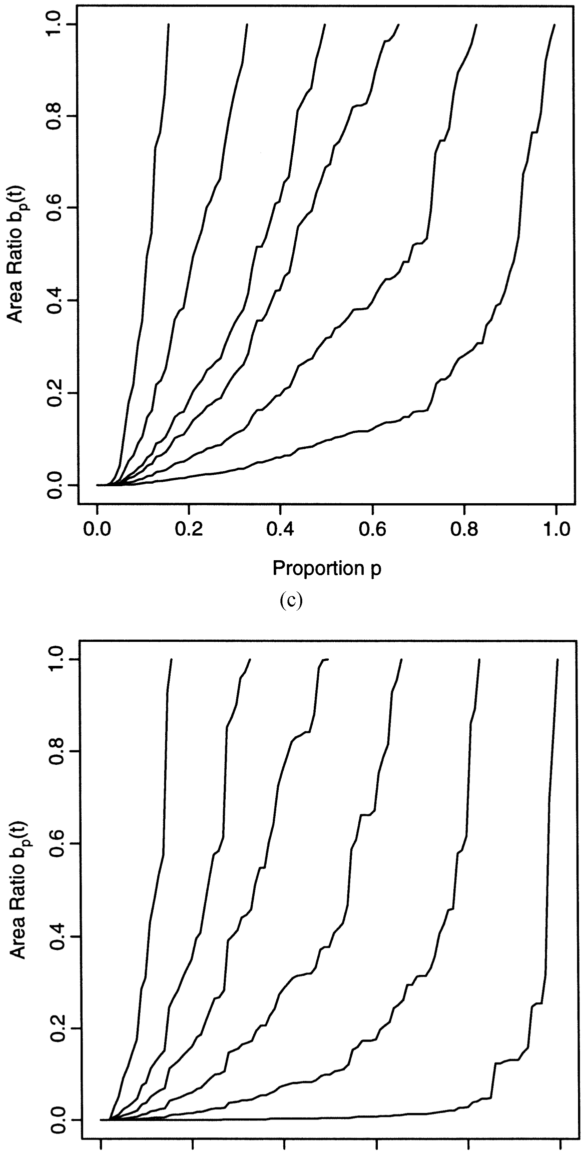


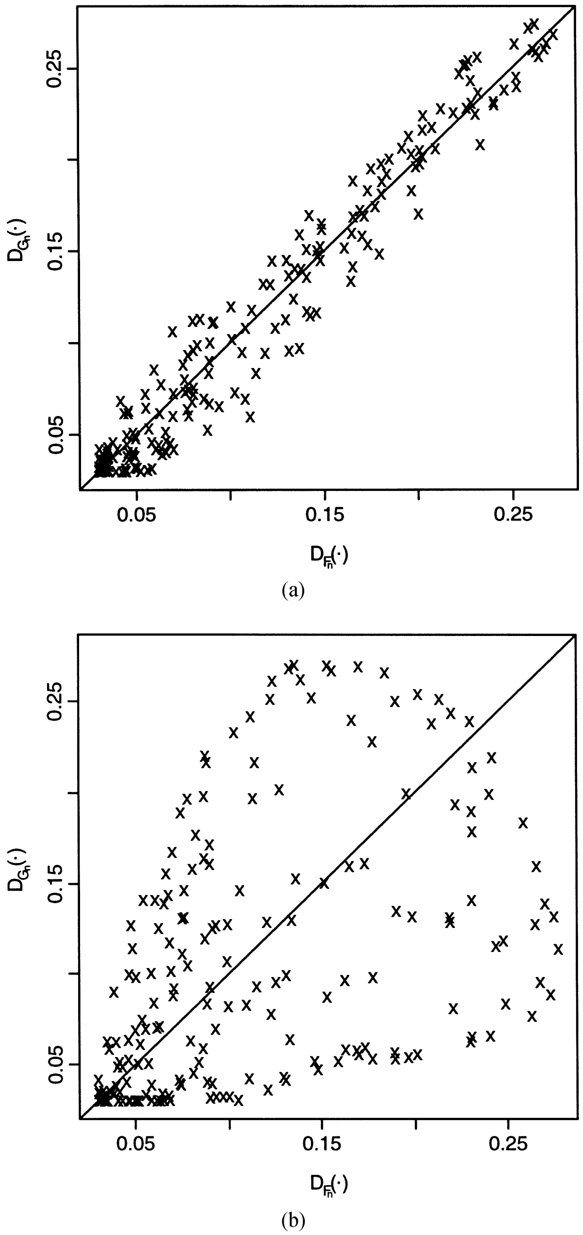
FIG. 19. *Kurtosis as fan plots.* (*a*) *Uniform.* (*b*) *Normal.*



(c)

(d)

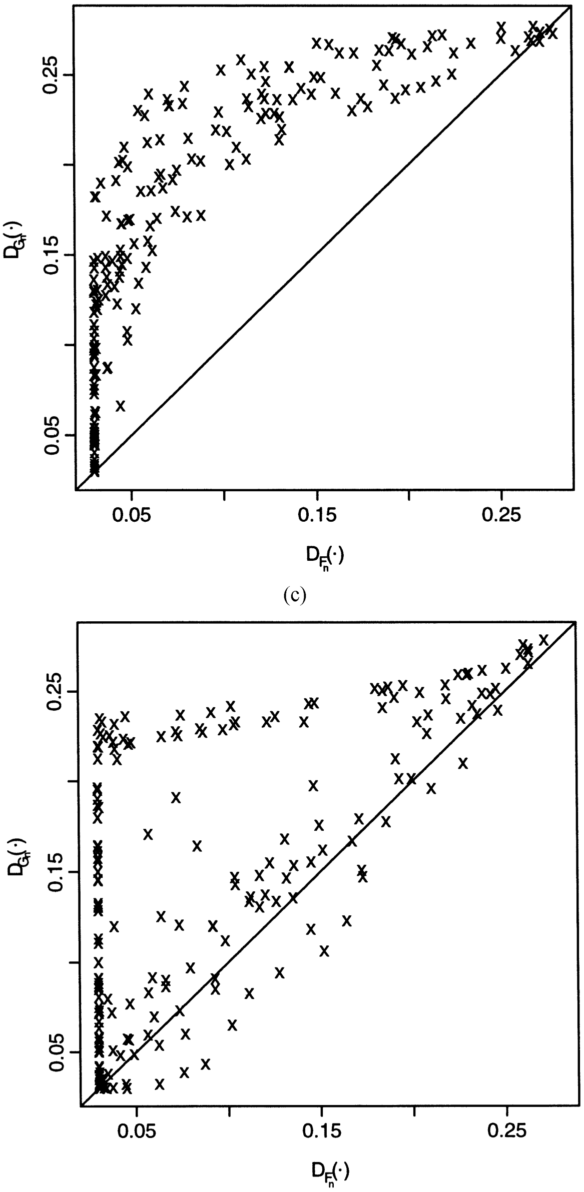
FIG. 19. (*Continued*). (*c*) *Double exponential.* (*d*) *Cauchy.*



(a)

(b)

FIG. 20. *DD-plot.* (*a*) *Identical distributions.* (*b*) *Location shift.*



(c)

(d)

FIG. 20. *(Continued). (c) Scale difference. (d) Skewness difference.*

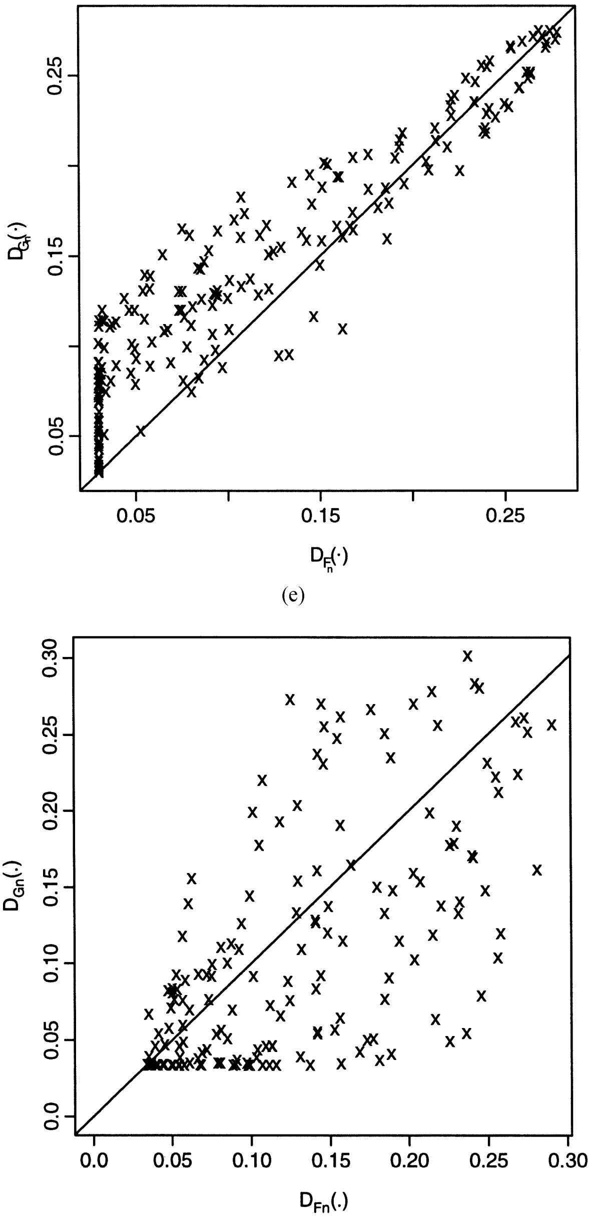
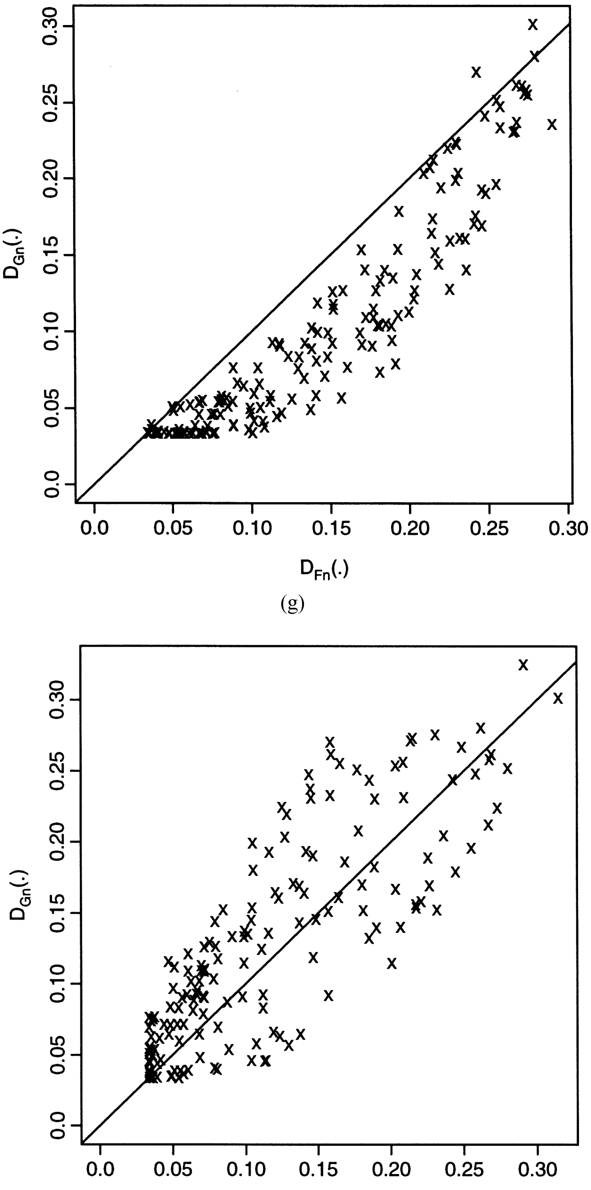


FIG. 20. (*Continued*). (*e*) *Kurtosis difference.* (*f*) *Raw test scores.*



(g)

(h)

FIG. 20. (*Continued*). (*g*) *Scale adjusted test scores.* (*h*) *Location adjusted test scores.*

diagonal line from $(0, 0)$ to $(1, 1)$ in $\mathbb{R}^2$. Plots which deviate from this straight line indicate differences between the two distributions. We observe that different deviation patterns in the plot correspond to different types of variations between the distributions, for example location shifts or scale increases.

Let F and G be two distributions on $\mathbb{R}^d$ and $D(\cdot)$ be an affine-invariant depth. We define $DD(F, G)$ as

$$DD(F,G) = \{(D_F(x), D_G(x)) \text{ for all } x \in \mathbb{R}^d\}. \tag{7.1}$$

When F and G are fixed, we denote $DD(F, G)$ simply by DD. Since DD is a subset of $\mathbb{R}^2$, its plot can be easily visualized. We call it the DD-plot. It is affine invariant if the underlying depth is. When the distributions are unknown, we may construct the empirical version of a DD-plot. If the underlying distribution F is unknown, with a given dataset $\{X_1, \dots, X_n\}$, we may determine whether F is some specified distribution, say G by examining the following DD-plot:

$$DD(F_n, G) = \left\{(D_{F_n}(x), D_G(x)) \text{ for all } x \in \{X_1, \dots, X_n\}\right\}.$$

If F and G are the population distributions for the samples $\{X_1, \dots, X_n\}(\equiv \mathbf{X})$ and $\{Y_1, \dots, Y_m\}(\equiv \mathbf{Y})$, then the DD-plot below can be used to determine whether or not the two distributions are identical:

$$DD(F_n, G_m) = \left\{(D_{F_n}(x), D_{G_m}(x)),\ x \in \{\mathbf{X} \cup \mathbf{Y}\}\right\}. \tag{7.2}$$

We observe that if $d = 1$, then the Lebesgue measure of DD is zero when $F \neq G$. However, if $d \geq 2$ and if F and G are both absolutely continuous, then DD is a region with a nonzero area. *The area of this region can serve as an affine-invariant measure of the discrepancy between F and G.*

If the two distributions are identical, the DD-plot in (7.2) should be concentrated along the diagonal line, as seen in Figure 20a where the two samples are drawn from the same standard bivariate exponential distribution. Different patterns of deviations from the diagonal line in the DD-plots are indications of differences in specific characteristics of F and G. Several examples are given here to identify these characteristics. The depth measure used in Figures (a–h) is simplicial depth.

Consider first the case when $G(\cdot) = F(\cdot - \theta)$, that is, G is a location shift of F. In this case, the DD-plot exhibits a noticeable departure from the diagonal line from $(0, 0)$ to $(0.25, 0.25)$, in such a symmetric manner as if the diagonal is the regression line and the DD-plot is the scatter plot. The departure here usually takes the form of pulling down from the point $(0.25, 0.25)$ to $(0, 0)$, leaving the upper right corner empty and spreading out the points around the midrange of the diagonal line, as if fitting a heart-shaped leaf on the diagonal pointing at $(0, 0)$. An example of this can be seen in Figure 20(b), which is the DD-plot with one sample from the standard bivariate normal and the other sample with a mean shift to $(0.5, 0)$. From the theoretical point of view, we note that if the two samples have the same center (deepest point) then this center achieves the highest depth within each individual sample. In other words, *a DD-plot with a common maximum point for both coordinates indicates a common center* (*i.e., no location shift*) *for the two underlying samples*. This is clearly not the case in Figure 20(b), and the resulting heart-shaped plot indicates a location difference in the two samples.

In order to bring out scale differences, the center of the samples should be equalized first by subtracting from the data their respective centers. Suppose that F and G have the same center, but F is more spread out than G. Then the points in DD tend to arch above the diagonal line in the shape of an early half moon as seen in Figure 20(c). This is a DD-plot for two bivariate normal samples where F has an enlarged scale.

To bring out skewness and kurtosis-associated differences in DD-plots, both location and scale should be equalized first. To equalize the scales, the data should be transformed to $S_X^{-1/2}X$ and $S_Y^{-1/2}Y$, where S_X and S_Y are dispersion matrices of the central 50% of the data [cf. (2.1)]. Once F and G have the same location and (central) spread, a difference in skewness manifests itself in the DD-plot in the form seen in Figure 20(d). As in Figure 20(c), the plot arches up above the diagonal line. However, unlike Figure 20(c), the arch is not symmetric as a half-moon shape. Rather, it spreads out more toward the lower left corner. The skewness difference comes from the two bivariate chi-square distributions, with different degrees of freedom 1 and 5.

Finally, suppose F and G have the same center and (central) spread and both are more or less symmetric. If F has higher kurtosis than G, then the lower part of the DD-plot shifts to one side of the diagonal line, although the upper part still points straight toward $(1, 1)$. Figure 20(e) provides such an example, with the first sample from standard bivariate normal and the other from the bivariate distribution whose components are independent Cauchy (1).

As an application, we present three DD-plots for the test score data set discussed in Section 4. Figure 20(f) is the DD-plot of the original data set. The plot deviates significantly from the diagonal line and does not have a common maximum point. This strongly suggests a possible location difference of the two sets of test scores. After centering both sets of test scores on $(0, 0)'$ (by subtracting the corresponding deepest point from each observation), we obtain the DD-plot in Figure 20(g). The half-moon shape of the plot hanging below the diagonal line indicates a scale difference between the two sets of scores, with the larger scale for closed-book scores. This conclusion was independently obtained by the earlier scale plot in Figure 7(b). Finally, the DD-plot for the standardized data (multiplying each datum by the square root inverse of the corresponding central 50% covariance matrix) is given in Figure 20(h). This plot shows minor asymmetry along the diagonal line, which seems to be an indication of a possible difference in skewness and/or kurtosis.

8. Data-depth based estimation of a dispersion matrix and diagnostics of nonnormality. In this section, two results related to multivariate normality are derived. They provide some simple diagnostic tools for checking normality. The first result establishes a relationship between the overall dispersion matrix and the dispersion of a smaller central region determined by a data depth. The second result provides an almost sure bound for the maximum deviation from the mean in a multivariate normal sample. We now proceed to discuss the first result.

The estimation of the dispersion matrix $\mathbf{\Sigma} = E((X - \mu)(X - \mu)')$ plays an important role in multivariate analysis. One of the important functions of Σ is in standardizing or sphericizing the observed data $\mathbf{X}$, that is, in defining $Z = \mathbf{\Sigma}^{-1/2}(X - \mu)$ so that Z has mean vector $\mathbf{0}$ and dispersion matrix $\mathbf{I}$, the identity matrix. Our objective in estimating $\mathbf{\Sigma}$ is to obtain a sample version of Z, namely $Z^* = \hat{\mathbf{\Sigma}}^{-1/2}(X - \overline{X})$, where $\hat{\mathbf{\Sigma}}$ is the estimated dispersion matrix.

The most natural estimator of Σ is of course the sample variance–covariance matrix $\mathbf{S}_n$, where $\mathbf{S}_n = (1/n)\Sigma_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})'$. This estimator, however, is highly sensitive to extreme observations. It puts in question whether the dispersion matrix restricted to a central region C_p can give rise to a consistent estimator of $\mathbf{\Sigma}$. The answer turns out to be indeed that it cannot, if nothing else is specified about the population distribution. If the tail portion of a population is unspecified, the ranges of the values of the elements in $\mathbf{\Sigma}$ are unbounded.

For the rest of this section we shall restrict ourselves to the family of distributions commonly known as elliptical distributions. The density function of a member of this family can be expressed as [cf. page 34 of Muirhead (1982)]

$$f(x) = \frac{c_d}{|\mathbf{\Sigma}_0|^{1/2}} h\big((X - \mu)'\mathbf{\Sigma}_0^{-1}(X - \mu)\big), \qquad x \in \mathbb{R}^d, \tag{8.1}$$

where c_d is a constant depending on d and the function h, μ is the center of the distribution, $\mathbf{\Sigma}_0$ is a positive definite matrix, and $h(\cdot)$ is a nonnegative function with $\int_0^\infty t^{(k/2-1)}h(t)\,dt < \infty$. The notation "$|*|$" stands for the determinant of $*$. It is clear that $\mathbf{\Sigma}_0 = c\mathbf{\Sigma}$, for some constant c.

Let $\mathbf{\Sigma}(p)$ denote the dispersion matrix when the population is conditioned on the central region C_p. We shall show that $\mathbf{\Sigma}(p)$ and $\mathbf{\Sigma}$ are related by a constant multiplier if the underlying population is elliptical and the depth used is affine invariant. It should be noted here that the contours of the density function of an elliptical distribution agree with the depth contours determined by an affine invariant data depth [cf. Liu and Singh (1993)].

THEOREM 8.1. *Let p be a fixed value between* 0 *and* 1. *If the underlying distribution F satisfies* (8.1), *then*

$$\mathbf{\Sigma}(p) = \eta_p \mathbf{\Sigma}, \tag{8.2}$$

where η_p is a scalar depending on $h(\cdot)$, p and d. If $R^2 = (X - \mu)' \Sigma_0^{-1}(X - \mu)$, and ξ_p stands for the pth quantile of R^2, then

$$\eta_p = \frac{E\left(R^2 | R^2 \le \xi_p\right)}{E(R^2)}. \tag{8.3}$$

The density function of R^2 is given by

$$g(r^2) = \frac{c_d \pi^{d/2}}{\Gamma(d/2)} (r^2)^{d/2-1} h(r^2). \tag{8.4}$$

The constant c_d is the same as the one in (8.1).

Note that the density in (8.4) can be found on page 37 of Muirhead (1982).

Example 8.1. Let $p = 0.5$. Consider the case where $d = 2$ and $h(r^2) = \exp(-r^2/2)$, namely the case of bivariate normal distribution. Here $c_d = (2\pi)^{-d/2}$, and the distribution of R^2 turns out to be exponential with mean $= 2$. Thus the formula in (8.4) gives

$$\begin{aligned} \eta_p &= 2\int_0^{\ln 2} y e^{-y}\, dy \Big/ \int_0^{\infty} y e^{-y}\, dy \\ &= (1 - \ln 2) \approx 0.31. \end{aligned} \tag{8.5}$$

The constant (0.31) derived in the above example can be utilized as a diagnostic tool for checking bivariate normality. More precisely, we compute the sample version of $\Sigma(0.5)$ and compare it with the sample dispersion matrix $\hat{\Sigma}$ to see if $\hat{\Sigma}(0.5)$ is close to $0.31 \times \hat{\Sigma}$. Our simulations give ratios of 0.314 for normal, 0.274 for exponential, and 0.014 for Cauchy distributions. The results are supportive of our claim and they give a clear distinction between normal and nonnormal cases.

A closer examination of the distribution of R^2 in (8.4) provides yet another simple method for checking the normality. In the bivariate normal case with $\Sigma = \mathbf{I}$, the median of R^2 is $2 \ln 2$, and thus

$$\text{area}(C_{0.5}) = \pi(2 \ln 2) \approx 4.355.$$

For a general Σ, we have instead

$$\text{area}(C_{0.5}) = \pi(2 \ln 2)|\Sigma|^{1/2} \approx 4.355|\Sigma|^{1/2}, \tag{8.6}$$

which again, for diagnostic purposes, can be used for normality checking. This can be viewed as a bivariate generalization of the simple univariate normality checking, which examines whether or not the interquartile range of the standardized sample is 1.25.

In implementing the checking procedure in (8.6), we need to obtain a good estimate of area($C_{0.5}$) from the given sample. Recall from Section 2.1 the definitions of the pth central region C_p and its sample version $C_{n,p}$, the pth central hull. Our definition of $C_{n,p}$ was motivated by the intended probability mass inclusion. However, the area, and not the probability inclusion, is the main focus in (8.6). Using area($C_{n,p}$) as an estimate for area(C_p) often does not achieve the desired level of accuracy, since our forming the convex hull $C_{n,p}$ has trimmed off most of the smooth surface area of C_p. Consequently, $C_{n,p}$ consistently underestimates C_p in terms of area. Thus, we propose to estimate area(C_p) by area($C^I_{n,p}$), where

$$(8.7) \quad C^I_{n,p} = \text{the central hull containing } \lceil np \rceil \text{ sample points in its interior.}$$

Our simulations show that the ratios of area(C_p) to $|\Sigma|^{1/2}$ are 4.74, 4.73 and 4.39 for bivarite normal samples with sizes 50, 100 and 500, respectively. The same simulations yield the ratios ≈ 3.5 for bivariate exponential samples, 0.5 for bivariate uniform samples, and 15.1 for bivariate Cauchy samples. In all three nonnormal cases, the ratio is clearly far from 4.355.

We summarize the differences in the roles of the two estimates for C_p. If the area or the volume of the region is the main object, then $C^I_{n,p}$ performs better. In terms of capturing the probability mass inclusion in C_p, $C_{n,p}$ performs quite adequately. It is easier to determine $C_{n,p}$, since it does not require the extra sequential search needed in determining $C^I_{n,p}$.

Turning to the issue of outliers in model checking, we now establish an almost sure bound for the maximum deviation from the mean for a normal random sample. This bound can be used to verify the normality of a given sample. It is in the form of an exact growth rate with no unspecified constants. This point is elaborated after the statement of the exact bound in Theorem 8.2.

THEOREM 8.2. *Let $\{X_1, \dots, X_n\}$ be a sequence of i.i.d. random variables following the d-dimensional normal distribution $N(\mu, \Sigma)$, where Σ is a $d \times d$ positive definite matrix. Then,* $\max_{1 \le i \le n} \|X_i - \mu\|$ *grows at the exact rate of* $\{2\lambda_{\max}(\log n + (m/2)\log\log n)\}^{1/2}$, *almost surely, where $\lambda_{\max}$ is the maximum eigenvalue of Σ and m is the multiplicity of $\lambda_{\max}$.*

The proof of this theorem is quite involved and is presented in Hüsler, Liu and Singh (1999).

REMARK 8.1. By "grows at the exact rate," we mean the following:

$$P\Big(\max_{1 \le i \le n} \|X_i - \mu\| \ge \{2\lambda_{\max}(\log n + s\log\log n)\}^{1/2}, \text{i.o.}\Big) = \begin{cases} 1, & \text{if } s \le \frac{m}{2}, \\ 0, & \text{if } s > \frac{m}{2}. \end{cases}$$

Clearly, $\lambda_{\max} = 1$ and $m = d$ if $\boldsymbol{\Sigma} = \mathbf{I}$. In this case, it is interesting to note that the dimension of the data appears only in the secondary term involving $(\log\log n)$ in the rate of growth.

Before putting this result directly to use, we recommend that the data be standardized first. This is in order to have $\boldsymbol{\Sigma} = \mathbf{I}$, since it may be difficult to determine the multiplicity of $\lambda_{\max}$ from $\hat{\boldsymbol{\Sigma}}$. In the process of standardizing the data, we should use a robust version of $\hat{\boldsymbol{\Sigma}}$. For instance, $\{\hat{\boldsymbol{\Sigma}}(0.5)/0.31\}$ should be a reasonable choice for such a purpose in the bivariate case, following the result in (8.5). Figure 21 contains the plot (looking like a step function) of the maximum of a random normal sample from $\mathbb{R}^2$ as n grows to 5000, against the growth curve (appearing in a small-dot curve) expressed in Theorem 8.2. The two plots seem to match closely. Raising the dimension of the distribution to 10, the plot is given in Figure 22. In both figures, the lower dashed curve is the growth curve in Theorem 8.2 without the secondary $\log\log n$ term. The lack of fit of the lower dashed curve shows the crucial role of the secondary term, especially in high dimension cases. Finally, the same plot for a bivariate exponential sample is presented in Figure 23, where the lack of fit of the growth curve intended for the normal case is evident.

9. Concluding remarks.

9.1. *Computation and graphics.* Many numerical and graphical simulations have been presented in this paper. In general, analyzing and presenting multivariate observations require more sophisticated algorithms than in the

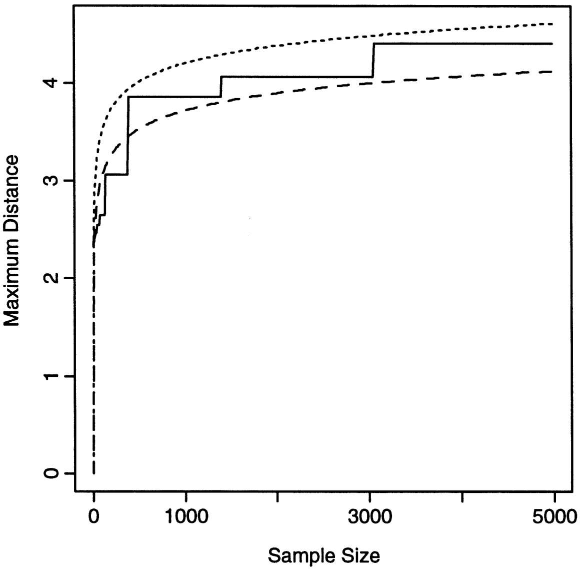


FIG. 21. *Maxima of normal samples* ($\mathbb{R}^2$).

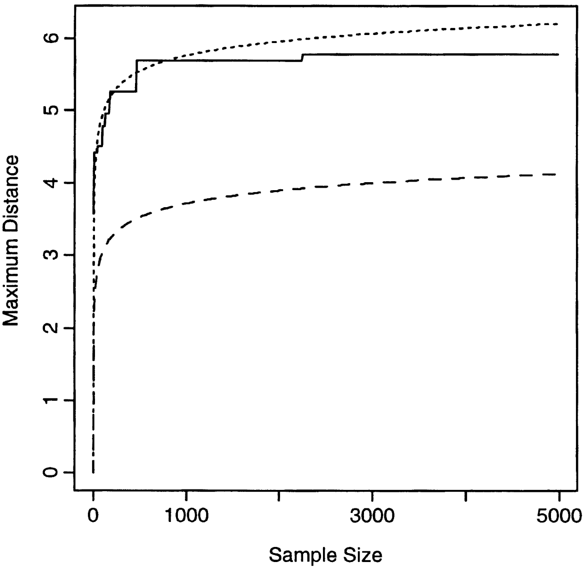


FIG. 22. *Maxima of normal samples* ($\mathbb{R}^{10}$).

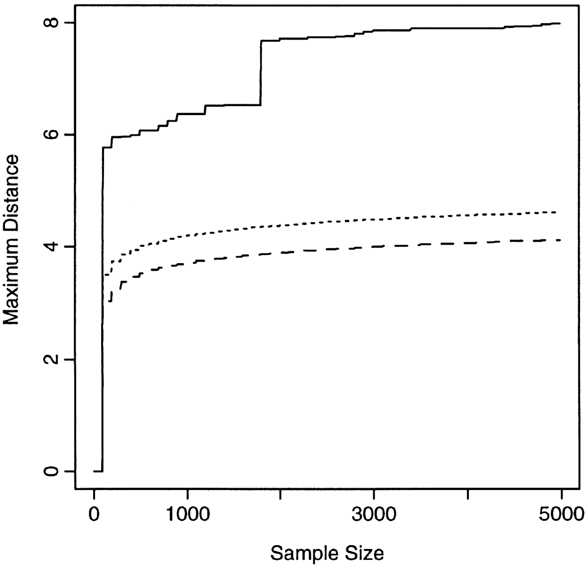


FIG. 23. *Maxima of exponential samples* ($\mathbb{R}^2$).

univariate case. Since the contour plot in Figure 2 shows that the simplicial ordering captures more of the probabilistic structure of the underlying distribution than the Mahalanobis ordering, the simplicial ordering has been used throughout for all simulations, except in a few specific cases. In principle, the sample simplicial depth at a given point in any dimensional space can be calculated in a straightforward manner by solving a system of linear equations. More precisely, we can determine whether or not a point is inside a random simplex by checking if the point can be expressed as a convex combination of the vertices of the simplex. However, this routine checking is tedious and time-consuming. Instead, we utilize a FORTRAN program provided in Rousseeuw and Ruts (1996) to calculate the bivariate simplicial depth. This program effectively reduces the number of operations to $O(n \log n)$. It also contains a subroutine for calculating the half-space depth, which we use in Section 5 for measuring angular skewness. Some three-dimensional simulation results are available, but they are omitted since they would only serve to confirm the results already presented. Clearly, it would be desirable to have efficient computing algorithms for the higher dimensional cases.

All simulations were run using S-Language on a Sun Workstation SPARC 5. The computer programs used and their detailed instructions are available in Parelius (1997).

Many interesting graphical methods have been proposed to present features of multivariate data sets, for example, Andrews (1972), Chernoff (1973), Kleiner and Hartigan (1981), and Wegman (1990). Friedman and Rafsky (1979, 1981), Easton and McCulloch (1990) and Marden (1998) have introduced different multivariate versions of the quantile–quantile plot and use them as diagnostic tools for comparing two multivariate samples or checking multivariate normality. A comparison of these qq-plots to our DD-plots could be worthwhile.

9.2. *Ramifications of data depth.* Besides providing a new set of parameters for multivariate analysis, the depth ordering concept has many theoretical and practical ramifications. For example, multivariate sign or rank tests based on the Oja depth are studied in a series of papers: Brown and Hettmansperger (1989), Hettmansperger, Nyblom and Oja (1992) and Hettmansperger and Oja (1994). Liu (1992) and Liu and Singh (1993) develop a quality index and several multivariate rank tests based on the general concept of depth ranking. Liu (1995) applies some of these tests to develop nonparametric multivariate control charts. These control charting techniques are then used in Cheng, Liu and Luxhoj (1999) to develop a monitoring scheme with thresholding systems for the analysis of multivariate aviation safety data. With the help of bootstrap methods, a general methodology based on data depth is developed for constructing confidence regions [see Yeh and Singh (1997)], and for determining P-values in testing hypotheses [see Liu and Singh (1997)]. Using the likelihood depth, Fraiman, Liu and Meloche (1997) provide a multivariate density estimate with an improved convergence rate. Rousseeuw and Hubert (1999) extend the half-space and simplicial depth rankings to develop robust regression methods.

We have seen that depth-based multivariate methods can be completely nonparametric or even moment free. Clearly, defining the distributional characteristics and the corresponding descriptive statistics is only a first step. Properties such as consistency and other asymptotics of the descriptive statistics proposed in this paper are yet to be fully investigated. Many questions still need to be addressed, for example, how the proposed statistics can be applied to making inferences, and which notion of depth should be more suitable for what analysis purpose. It may also be worthwhile to investigate how resampling methods can be utilized in the depth-based methodology to help achieve better graphical presentations.

APPENDIX

PROOF OF THE AFFINE INVARIANCE PROPERTY OF THE SHRINKAGE PLOT IN SECTION 6.2. Let Ω be a convex hull on $\mathbb{R}^d$ and θ a given point in the interior of Ω. Consider the linear transformation $x \to (y = a + Bx)$, where a is a $d \times 1$ vector and B is a $d \times d$ nonsingular matrix. The point $\theta^* = a + B\theta$ is in the interior of the transformed hull $\Omega^* = a + B\Omega$. For any point $x_0 \in \mathbb{R}^d$, with $x_0 \neq \theta$, a line segment passing through x_0 and θ which falls completely inside the convex hull Ω can be expressed as

$$L = \{x \in \mathbb{R}^d\colon x = \theta + c(x_0 - \theta) \text{ for all } c \text{ between some numbers } c_1 \text{ and } c_2\} \tag{A.1}$$

After the transformation, the line segment L becomes

$$L^* = \{x^* \in \mathbb{R}^d\colon x^* = \theta^* + c(x_0^* - \theta^*) \text{ for all } c \text{ between } c_1 \text{ and } c_2\}. \tag{A.2}$$

Here $x_0^* = a + Bx_0$. For a given shrinkage proportion p, the shrunk version of L, denoted by L_p, can be expressed as (A.1) with the constant c now falling between pc_1 and pc_2. Similarly, the shrunk version of L^*, denoted by L_p^*, can be expressed as (A.2) with the constant c again falling between pc_1 and pc_2.

The invariance property of the shrinkage plot follows from the equivalence of the following four statements for a given data point x_i inside Ω: (1) x_i lies on L_p; (2) $x = \theta + c(x_0 - \theta)$ for some c with $pc_1 \leq c \leq pc_2$; (3) $a + Bx_i = \theta^* + c(x_0^* - \theta^*)$ and (4) $a + Bx_i$ lies on L_p^*.

PROOF OF THEOREM 8.1. We begin by noting that on $\mathbb{R}^d$, the surface area of a sphere with radius r is given by

$$A(d, r) = \frac{2\pi^{d/2} r^{d-1}}{\Gamma(d/2)}, \qquad r \geq 1. \tag{A.3}$$

Let $Y = \mathbf{\Sigma}_0^{-1/2}X$. Recall that X has the elliptical density

$$f(x) = \frac{c_d}{|\mathbf{\Sigma}_0|^{1/2}} h\big((x-\mu)'\mathbf{\Sigma}_0^{-1}(x-\mu)\big).$$

The density of Y is then $l(y) = c_d h(y'y)$. By definition, $R^2 = (X-\mu)'\mathbf{\Sigma}_0^{-1}(X-\mu) = Y'Y$. Since the density of Y is spherical, we immediately obtain

$$P(R \in (r, r+\Delta r)) = A(d,r)c_d h(r^2)\,\Delta r + o(\Delta r).$$

This implies that the density function of R is $A(d,r)c_d h(r^2)$. Substituting $A(d,r)$ with the formula in (A.3), the density of R becomes

$$\frac{2\pi^{d/2}r^{d-1}}{\Gamma(d/2)} c_d h(r^2).$$

In particular, the density function of R^2 is

$$g(r^2) = c_d \frac{\pi^{d/2}(r^2)^{d/2-1}}{\Gamma(d/2)} h(r^2).$$

Note that for the elliptical distribution with the density function $f(x)$ given earlier, the overall dispersion matrix $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \frac{ER^2}{d}\mathbf{\Sigma}_0. \tag{A.4}$$

This can be seen as follows. Let $\mathbf{\Sigma}_Y$ denote the dispersion matrix of Y. Since Y is spherical, $\mathbf{\Sigma}_Y = a\mathbf{I}$, for some constant $a > 0$. Hence $ER^2 = E(Y'Y) = ad$ and $a = ER^2/d$, which implies (A.4).

It is clear from the statement of the theorem that $\mathbf{\Sigma}(p)$ is the dispersion of a distribution restricted to C_p. This distribution is also elliptical, with density function

$$f^*(x) = \frac{c_d}{p|\mathbf{\Sigma}_0|^{1/2}} h^*\big((x-\mu)'\mathbf{\Sigma}_0^{-1}(x-\mu)\big),$$

where $h^*(t) = h(t)$ when $t \le$ (the pth quantile of R^2), and $h^*(t) = 0$ otherwise. This is because the contour of the central region C_p agrees with the density contour of the elliptical distribution, provided the depth is affine invariant [see Liu and Singh (1993)]. Arguments which led to (A.4) immediately show that $\mathbf{\Sigma}(p) = ((E(R^2|R^2 \le \xi_p))/d)\mathbf{\Sigma}_0$, or

$$\mathbf{\Sigma}(p) = \eta_p \mathbf{\Sigma},$$

where $\eta_p = E(R^2|R^2 \le \xi_p)/ER^2$, with $\xi_p =$ (the pth quantile of $\mathbb{R}^2$).

REFERENCES

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Andrews, D. (1972). Plots of high-dimensional data. *Biometrics* **28** 125–136.

Arcones, M., Chen, Z. and Gine, E. (1994). Estimators related to U-processes with applications to multivariate medians: asymptotic normality. *Ann. Statist.* **22** 1460–1477.

AVÉROUS, J. and MESTE, M. (1997). Skewness for multivariate distributions: two approaches. *Ann. Statist.* **25** 1984–1997.

BARNETT, V. (1976). The ordering of multivariate data. *J. Roy. Statist. Soc.* Ser. *A* **139** 319–354.

BERAN, R. (1979). Testing for ellipsoidal symmetry of a multivariate density. *Ann. Statist.* **7** 150–162.

BERAN, R. and MILLAR, P. (1997). Multivariate symmetry models. In *Festschrift for Lucien Le Cam* 13–42. (L. Le Cam, E. Torgersen and G. Yang, eds.) Springer, New York.

BICKEL, P. and LEHMANN, E. (1975a). Descriptive statistics for nonparametric models I. Introduction. *Ann. Statist.* **3** 1038–1044.

BICKEL, P. and LEHMANN, E. (1975b). Descriptive statistics for nonparametric models II. Location. *Ann. Statist.* **3** 1045–1069.

BICKEL, P. and LEHMANN, E. (1976). Descriptive statistics for nonparametric models III. Dispersion. *Ann. Statist.* **4** 1139–1158.

BICKEL, P. and LEHMANN, E. (1979). Descriptive statistics for nonparametric models IV. Spread. In *Contributions to Statistics*, *Hájek Memorial Volume* (J. Jurecková, ed.) 33–40. Reidel, London.

BROWN, B. and HETTMANSPERGER, T. (1989). The affine invariant bivariate version of the sign test. *J. Roy. Statist. Soc. B* **51** 117–125.

CHAUDHURI, P. (1996). On a geometric notion of multivariate data. *J. Amer. Statist. Assoc.* **90** 862–872.

CHENG, A., LIU, R. and LUXHOJ, J. (1999). Monitoring multivariate aviation safety data: control charts and threshold systems. *IIE Transactions*. To appear

CHERNOFF, H. (1973). The use of faces to represent points in k-dimensional graphically. *J. Amer. Statist. Assoc.* **68** 361–368.

DONOHO, D. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827.

DÜMBGEN, L. (1992). Limit theorems for simplicial depth. *Statist. Probab. Lett.* **14** 119–128.

EASTON, G. and MCCULLOCH, R. (1990). A multivariate generalization of quantile–quantile plots. *J. Amer. Statist. Assoc.* **85** 376–386.

EDDY, W. (1982). Convex hull peeling. In *COMPSTAT* (H. Caussinus et al., eds.) 42–47. Physica, Vienna.

EINMAHL, J. and MASON, D. (1992). Generalized quantile process. *Ann. Statist.* **20** 1062–1078.

FRAIMAN, R., LIU, R. and MELOCHE, J. (1997). Multivariate density estimation by probing depth. In *L_1-Statistical Procedures and Related Topics* 415–430. IMS, Hayward, CA.

FRAIMAN, R. and MELOCHE, J. (1996). Multivariate L-estimation. Preprint.

FRIEDMAN, J. and RAFSKY, L. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717.

FRIEDMAN, J. and RAFSKY, L. (1981). Graphics for the multivariate two-sample problem (with comments). *J. Amer. Statist. Assoc.* **76** 277–295.

GASTWIRTH, J. (1971). A general definition of the Lorenz curve. *Econometrica* **39** 1037–1039.

GNANADESIKAN, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd ed. Wiley, New York.

HE, X. and WANG, G. (1997). Convergence of depth contours for multivariate datasets. *Ann. Statist.* **25** 495–504.

HETTMANSPERGER, T. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

HETTMANSPERGER, T., NYBLOM, J. and OJA, H. (1992). On multivariate notions of sign and rank. In *L-1 Statistical and Related Methods* (Y. Dodge, ed.) 267–278. North-Holland, Amsterdam.

HETTMANSPERGER, T. and OJA, H. (1994). Affine invariant multivariate multisample sign tests. *J. Roy. Statist. Soc. Ser. B* **56** 235–249.

HODGES, J. (1955). A bivariate sign test. *Ann. Math. Statist.* **26** 523–527.

HUBER, P. (1972). Robust statistics: a review. *Ann. Math. Statist.* **43** 1041–1067.

HÜSLER, J., LIU, R. and SINGH, K. (1999). A formula for the tail probability of a multivariate normal distribution and its applications. Preprint.

KENDALL, K., STUART, A. and ORD, J. K. (1987). *Kendall's Advanced Theory of Statistics* **1**. Oxford Univ. Press.

KLEINER, B. and HARTIGAN, J. (1981). Representing points in many dimensions by trees and castles (with comments). *J. Amer. Statist. Assoc.* **76** 260–276.

KOLTCHINSKII, V. (1997). M-estimator, convexity and quantiles. *Ann. Statist.* **25** 435–477.

LEHMANN, E. (1991). *Theory of Point Estimation.* Wadsworth and Brooks/Cole, Belmont, CA.

LIU, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414.

LIU, R. (1992). Data depth and multivariate rank tests. In *L-1 Statistics and Related Methods* (Y. Dodge, ed.) 279–294. North-Holland, Amsterdam.

LIU, R. (1995). Control charts for multivariate processes. *J. Amer. Statist. Assoc.* **90** 1380–1388.

LIU, R. and SINGH, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.* **88** 257–260.

LIU, R. and SINGH, K. (1997). Notions of limiting P-values based on data depth and bootstrap. *J. Amer. Statist. Assoc.* **91** 266–277.

LORENZ, M. (1905). Methods of measuring the concentration of wealth. *J. Amer. Statist. Assoc.* **9** 209–219.

MAHALANOBIS, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Acad. Sci. India* **12** 49–55.

MARDEN, J. (1998). Bivariate qq-plot. *Statist. Sinica* **8** 813–826.

MARDIA, K., KENT, J. and BIBBY, J. (1979). *Multivariate Analysis.* Academic Press, New York.

MUIRHEAD, R. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

NOLAN, D. (1992). Asymptotics for multivariate trimming. *Stochastic Process. Appl.* **42** 157–169.

OJA, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1** 327–332.

PARELIUS, J. (1997). Multivariate analysis based on data depth. Ph.D. dissertation. Dept. Statistics, Rutgers Univ., New Jersey.

ROUSSEEUW, P. and HUBERT, M. (1999). Regression depth. (with discussion). *J. Amer. Statist. Assoc.* **4**, 388–433.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

ROUSSEEUW, P. and RUTS, I. (1996). AS 307: bivariate location depth. *Appl. Statist.* **45** 516–526.

ROUSSEEUW, P. and RUTS, I. (1997). The bagplot: a bivariate box-and-whiskers plot. Preprint.

ROUSSEEUW, P. and STRUYF, A. (1998). Computing location depth and regression depth in higher dimensions. *Statist. Comput.* **8**, 193–203.

RUTS, I. and ROUSSEEUW, P. (1996). Computing depth contours of bivariate point clouds. *Computational Statistics and Data Analysis* **23** 153–168.

SINGH, K. (1991). Majority depth. Unpublished manuscript.

SINGH, K. (1998). Breakdown theory for bootstrap quantiles. *Ann. Statist.* **26** 1719–1732.

TUKEY, J. (1975). Mathematics and picturing data. In *Proceedings of the 1975 International Congress of Mathematics* **2** 523–531.

WEGMAN, E. (1990). Hyperdimensional data analysis using parallel coordinates. *J. Amer. Statist. Assoc.* **85** 664–675.

YEH, A. and SINGH, K. (1997). Balanced confidence sets based on the Tukey depth. *J. Roy. Statist. Soc. Ser. B* **3** 639–652.

R. Y. LIU
K. SINGH
DEPARTMENT OF STATISTICS
HILL CENTER
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854-8019
E-MAIL: rliu@stat.rutgers.edu
kern@stat.rutgers.edu

J. M. PARELIUS
THE NEW YORK TIMES
ELECTRONIC MEDIA COMPANY
NEW YORK, NEW YORK 10036

DISCUSSION

WILLIAM F. EDDY[1]

Carnegie Mellon University

Liu, Parelius, and Singh (henceforth, LPS) have taken some simple ideas and pushed them a long way to provide some useful tools for exploratory multivariate data analysis. They have written an interesting paper and I congratulate them. I also applaud the editors for their willingness to publish a nonstandard contribution to the *Annals*.

I am enthusiastic about what LPS have done and I encourage them to continue this work. I also encourage the interested reader to do what I did after reading LPS; I reread Schervish (1987) and the associated discussion. If you haven't yet read that paper, its author describes it as "a thoroughly biased and narrow look at the development of multivariate analysis." I find it an excellent overview of the scope of multivariate analysis. My rereading provided some context for my thinking about LPS.

Exploratory data analysis. The work in LPS, although couched somewhat in the language of mathematical statistics, really belongs in a branch of exploratory data analysis. Since Tukey (1962, 1977), and Mosteller and Tukey (1977) this field has really languished. Parametric modeling and inference, especially Bayesian, has made dramatic strides; our journals are filled with developments. The bootstrap (which the present authors have made contributions to), Markov chain Monte Carlo, and so on are prime topics for academic research. Nonparametric modeling has made equally dramatic strides with advances in smoothing, CART, generalized additive models, and so on. The reality of applied statistical work, namely, exploration of data (and development of models) has made almost no progress at all, except the peripheral progress gleaned from the improvements I've mentioned. We still use the box plots that Tukey gave us more than 20 years ago. The sunburst and contour plots in this paper are a nice generalization. We can quibble about details:

1. I don't want so many whiskers (rays).
2. I would prefer more contours.
3. I would like outliers distinguished in some way.

But the spirit is right.

Computing. In the days since Tukey introduced us to exploratory data analysis, the world has undergone major changes brought about by the

[1] Supported in part by NSF Grant DMS 97-05034.

integrated circuit revolution and the resulting ubiquity of computational devices. Nowhere in this paper is there a hint that this has happened. The computations needed for this article (with the exception of the simulations) could easily be done (by John Tukey sitting in the back of the lecture hall) by hand.

Even more important than the increase in computation power has been the attendant increase in the size of data sets. No longer, very often, does one find a data set with a few hundred or even a few thousand observations. The data set I am currently studying contains 120 billion individual numbers (there are no images); what constitutes a multivariate observation in this data set depends on one's point of view. The tools in LPS, in their present form, will simply not work on data sets this large. We need even more advanced tools, especially dynamical ones. We need to be able to elucidate conditional effects.

Graphics. The graphics in this paper are a throwback to the days of the pencil-and-paper graphs that Tukey liked to draw. After that there were pen-plotters; this let us make graphs that looked just like the ones he drew by hand. Only with the computer revolution mentioned above have bit-mapped graphics become commonplace. Where are they in this paper? We have much more spatial resolution available. Why not use it?

In the very same issue of *Statistical Science* as Schervish (1987) is an article on "Dynamic Graphics for Data Analysis" by Becker, Cleveland and Wilks (1987). Twelve years ago in my comment on that article I looked forward to a ten-year "Golden Age of Graphics." I continue to look forward to it and I hope that LPS will look at Becker, Cleveland and Wilks and think about how they could add dynamics to their graphics specifically to reveal interdependence among the variables.

Dependence. Finally, my criticism. LPS say they want to develop a general nonparametric methodology for multivariate analysis. However, I could not find the most important notion of multivariate analysis, *dependence*, seriously addressed in this paper. We have location, scale, skewness and kurtosis. We have sunburst plots which can show bivariate dependence, but what else is there? I can imagine a scatterplot matrix with sunburst plots of each pair of variables. But each element of the scatterplot matrix is actually a projection of all the dimensions into the two that are plotted. What about a dynamic conditional scatterplot matrix of sunburst plots where the dynamics lets us change the conditioning event?

I greatly hope that LPS will continue work on these ideas. I particularly hope they will work on dynamic and genuinely multidimensional generalizations of these ideas. And I hope that twelve years from now their latest publication will not be confined to the monochrome, static pages of this journal but will rather take full advantage of the integration of computing, television and dynamic graphical methodology!

REFERENCES

BECKKER, R. A., CLEVELAND, W. S. and WILKS, A. R. (1987). Dynamic graphics for data analysis (with discussion). *Statist. Sci.* **2** 353–395.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, MA.

SCHERVISH, M. J. (1987). Multivariate analysis (with discussion). *Statist. Sci.* **2** 396–433.

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67. (Corr: V33 p812)

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213-3890
E-MAIL: bill@stat.cmu.edu

DISCUSSION

KEITH A. BAGGERLY AND DAVID W. SCOTT[1]

Rice University

The authors have demonstrated a wide range of potential applications of data-depth ideas. Since in one dimension the data-depth and rank are closely related, the authors have succeeded in developing a novel line of multivariate extensions to nonparametric statistics. In this brief comment, we explore two aspects of interest in any multivariate extension of rank-based procedures. The first issue is computation and the second is the definition of what constitutes a central fraction of the data.

Algorithms for many multivariate point-oriented procedures are combinatorial in nature and difficult to solve exactly in feasible time. The algorithms provided by Rouseeuew and Ruts (1996) for computing the simplicial and halfspace depths at a given point dodge the combinatorial problem in the bivariate case by making use of an efficient ordering of the data values, but the method becomes noticeably more complex in dimensions higher than two. The computational complexity is $O(n^{d-1} \log n)$ for d-dimensional data. Many of the techniques described herein involve computing the areas of nested convex hulls containing fixed fractions of the data. The problem of finding the convex hull of a set of data has already attracted the interest of computer scientists in two and three dimensions, but again the computational complexity increases exponentially with dimension, adding another $O(n^{d-1} \log n)$ term. The computational burden may be further exacerbated in the inferential setting, as the only means of establishing acceptable deviations from a specified model would seem to involve computation for multiple samples; bootstrapped p-values. Clearly, there are many research opportunities created by the authors' line of inquiry.

[1] Supported in part by NSF Grant DMS 96-26187.

With multivariate data, a common goal is to understand the structure among the variables and any grouping among the points. We are interested in many of the same questions as the authors, Are the data normal? for example, and wish to explore the power of these techniques in many settings. We recall that the simple idea of replacing each data coordinate by its rank within that variable before performing multivariate techniques can radically change the structure of the data, for example, by making adjacent two widely separated clusters.

In Figure 1, we examine the simplicial depth contours for 200 variates drawn from a bivariate mixture of two normals, $f(x) = 0.6 * N(0, 0.25 * I) + 0.4 * N(1, 0.25 * I)$. Half-space depth contours are similar. We also provide contours from a kernel density estimate using a bivariate normal product kernel. The density estimate clearly shows the multimodal structure of the data. The authors include such density estimates in their hierarchy of data-depth measures (likelihood depth) but evidently different measures can give quite different types of information about the data at hand. The near convexity of the simplicial depth contours severely limits the interpretability of the corresponding contours for multimodal data.

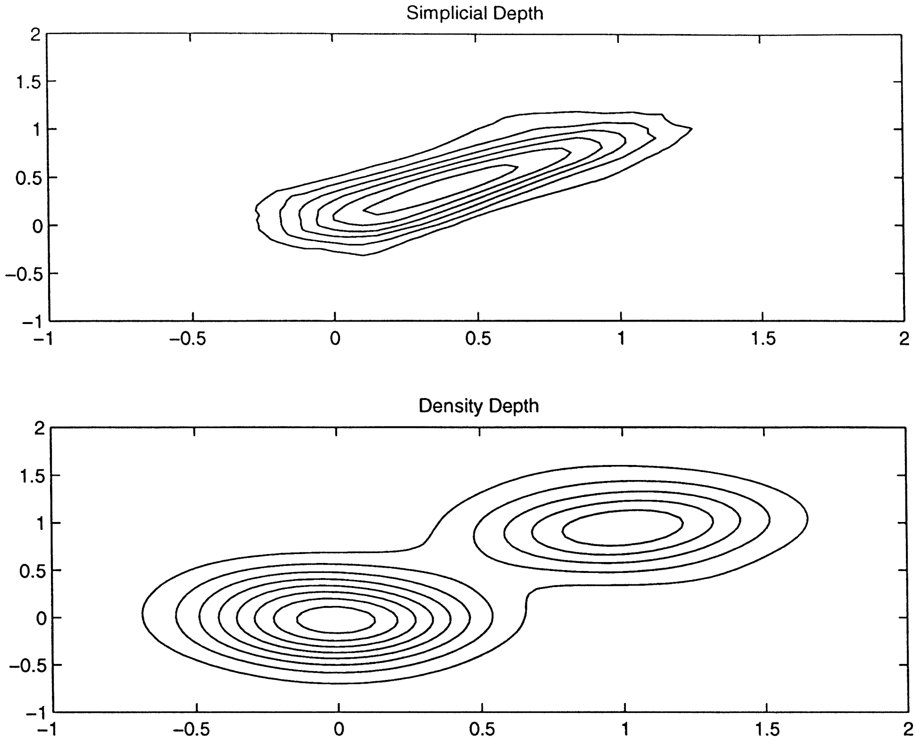


FIG. 1. *Depth contours for a bivariate mixture density using* (*a*) *simplicial depth and* (*b*) *density or likelihood depth.*

Among the mean-median-mode trio, the mean and mode are well defined beyond one dimension. The authors provide a specific description of a multivariate median point as the highest data-depth contour. However, the median is less a measure of center than a device for splitting the data in half. Thus we are more attracted to the notion of a multivariate median as a true density contour, specifically, a high-density contour capturing 50% of the probability mass. The median contour may be connected or, as in the multimodal case, consist of several density shells. (Recall that the multivariate mode is a set of all local modes.)

Extending the notion of ordering to multivariate data is an exciting challenge. As there is no unique way of doing so, different definitions will be appropriate for different classes of data. The authors' proposals seem limited to two or three dimensions and to unimodal data. This paper challenges us to think harder about these issues.

Department of Statistics MS-138
Rice University
6100 Main Street
Houston, Texas 77005-1892
E-mail: scottdw@stat.rice.edu

DISCUSSION

T. P. Hettmansperger, H. Oja and S. Visuri

Penn State University, University of Jyväskylä and Tampere University of Technology

The authors are to be congratulated for providing a sweeping introduction to multivariate descriptive statistics. This paper should inspire many investigations extending these methods into valuable inference tools. The unifying themes are data depth and corresponding sample characteristics. The development contrasts with the traditional approach based on moments. The result is a set of graphical displays rather than matrix displays that yield information on location, scale, skewness, and tail weight. They are much easier to interpret and, hence, more user friendly. Computer power now makes it possible to generate displays for reasonably sized data sets.

Rather than attempt a review of the current methods, below we restrict attention to samples that come from elliptical distributions. This is, of course, a subclass of the general class of underlying distributions considered by the authors. The class of elliptical models, while not as general, still provides a wide nonparametric class of distributions.

Our goal is to illustrate several of the data displays suggested by Liu, Parelius and Singh in the context of elliptical models. This allows for a discussion and elucidation of some of the more interesting suggestions made in the paper. In particular, we will consider sunburst plots, *DD*-plots, scale and kurtosis plots, and introduce *PP*-plots for scale and tail weight comparisons.

We suppose that our data comes from an elliptical distribution. Hence, the density is of the form

$$f(\mathbf{x}) = f(x_1, \dots, x_p) = |\det \Sigma|^{-1/2} f_0\left[(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

where Σ is a positive definite covariance matrix and $\boldsymbol{\mu}$ is a location vector. The covariance matrix Σ can be expressed in terms of its eigenvalue decomposition as follows:

$$\Sigma = \lambda U C U^T,$$

where λ^p is the generalized variance, the orthogonal matrix U contains the eigenvectors and C is the diagonal matrix of standardized eigenvalues ($|\det(C)| = 1$). As in Bensmail and Celeux (1996), we use the terms *scale*, *shape* and *orientation* for items λ, C and U. If $\mathbf{z}$ comes from a spherical distribution with the location vector $\mathbf{0}$ and covariance matrix I, then $\mathbf{y} = UC^{1/2}\lambda^{1/2}\mathbf{z} + \boldsymbol{\mu}$ is elliptically symmetric with the location vector $\boldsymbol{\mu}$, scale λ, shape C and orientation U.

Our plan is to first define a multivariate centered rank vector. This vector, in many ways, represents an extension of the idea of a univariate rank. In addition, it has certain nice affine equivariance properties. We only provide a sketch here; see Hettmansperger, Möttönen and Oja (1998) or Oja (1999) for details. We then consider the rank covariance matrix, RCM. Visuri, Koivunen and Oja (1999) show that if the standardized eigenvalues and the eigenvectors of the covariance matrix Σ are $c_1 > \cdots > c_p$ and $\mathbf{u}_1, \dots, \mathbf{u}_p$, respectively, then $c_1^{-1} < \cdots < c_p^{-1}$ and $\mathbf{u}_1, \dots, \mathbf{u}_p$ are the standardized eigenvalues and the eigenvectors for the theoretical RCM. The sample RCM is more robust than the sample covariance matrix and, hence, provides a robust estimate of the underlying shape and orientation for the elliptical distribution. This, along with a robust estimate of Wilk's generalized variance, can be used to robustly estimate Σ. However, here we use only the standardized eigenvalues and the eigenvectors to define a robust version of depth.

We next sketch the construction of the rank vector and corresponding sample RCM. We begin with p-dimensional data $\mathbf{x}_1, \dots, \mathbf{x}_n$. The volume of the p-variate simplex determined by $\mathbf{x}$ and p observation vectors with indices $i_1 < \cdots < i_p$ is

$$V(\mathbf{x}, \mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_p}) = \frac{1}{p!} \operatorname{abs}\left\{\det\begin{pmatrix} 1 & \cdots & 1 & 1 \\ \mathbf{x}_{i_1} & \cdots & \mathbf{x}_{i_p} & \mathbf{x} \end{pmatrix}\right\}.$$

We introduce the criterion function and its gradient

$$D(\boldsymbol{\mu}) = \operatorname{ave}\left\{V(\boldsymbol{\mu}, \mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_p})\right\} \quad \text{and} \quad \mathbf{R}(\boldsymbol{\mu}) = p!\,\nabla D(\boldsymbol{\mu}),$$

where the average is taken over all choices of subscripts $i_1 < \cdots < i_p$ from $1, \dots, n$. In the univariate case $R(x_i)$ reduces to the centered rank of x_i among $x_1, \dots, x_n$. This multivariate extension $\mathbf{R}(\mathbf{x}_i)$ retains many of the features of a centered rank and enjoys the following equivariance property: if the observations $\mathbf{x}_i$ are transformed to $\mathbf{x}_i^* = A\mathbf{x}_i + \mathbf{b}$ where A is nonsingular, then $\mathbf{R}^*(\mathbf{x}_i^*) = A^*\mathbf{R}(\mathbf{x})$ where $A^* = \text{abs}\{\det(A)\}(A^{-1})^T$ and $\mathbf{R}^*$ is the rank function calculated from the $\mathbf{x}_i^*$ observations. When A is orthogonal, $A^* = A$.

The sample RCM is simply

$$\text{RCM} = \text{ave}\left\{\mathbf{R}(\mathbf{x}_i)\mathbf{R}(\mathbf{x}_i)^T\right\}.$$

Let U be the matrix whose columns are $\mathbf{u}_1, \dots, \mathbf{u}_p$, the eigenvectors of the sample RCM, and let $D = \text{diag}\{d_1, \dots, d_p\}$, $d_1 < \cdots < d_p$, be a diagonal matrix with the standardized eigenvalues of RCM as diagonal entries. U and D^{-1} are then robust estimates of the eigenvectors (orientation) and the standardized eigenvalues (shape) of the covariance matrix. Finally, let $W = UDU^T$.

The Oja (1983) multivariate sample median $\hat{\boldsymbol{\mu}}$ solves $\mathbf{R}(\boldsymbol{\mu}) = \mathbf{0}$. Define the depth of $\mathbf{x}$ relative to $\mathbf{x}_1, \dots, \mathbf{x}_n$ by the Mahalanobis-type distance from the Oja median,

$$d(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T W(\mathbf{x} - \hat{\boldsymbol{\mu}}).$$

Note here that we retain $d(\mathbf{x})$ as the measure of depth rather than $(1 + d(\mathbf{x}))^{-1}$ suggested by the authors. Hence, large values of $d(\mathbf{x})$ suggest that $\mathbf{x}$ is at or beyond the convex hull of the data and small values suggest that $\mathbf{x}$ is near the Oja median at the center of the data.

We will illustrate the plots with data drawn from the following models with $\mathbf{X}$ distributed as $N(\mathbf{0}, I_2)$ and sample size $n = 50$:

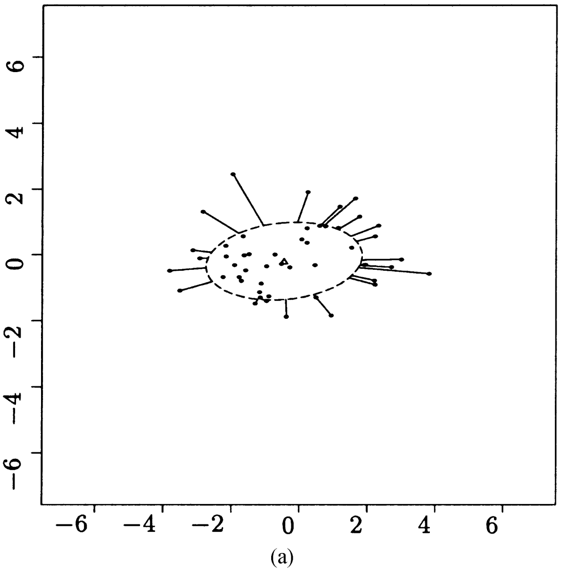
$$\text{(A)}: \quad \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\mathbf{X},$$

$$\text{(B)}: \quad \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\mathbf{X} + \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$\text{(C)}: \quad \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\mathbf{X} + \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$\text{(D)}: \quad (1 - B)\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\mathbf{X} + B\sqrt{10}\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\mathbf{X}, \text{ with } B \sim \text{Bin}(1, 0.05).$$

We first consider the sunburst plot. This extends the boxplot and is valuable in assessing the location, scale, shape and orientation of the sample as well as identifying outliers. In the case of an elliptical model, we seek an ellipse that includes fifty percent of the data, reflects the proper orientation and shape and is not affected by outliers. In Figure 1, we show the sunburst



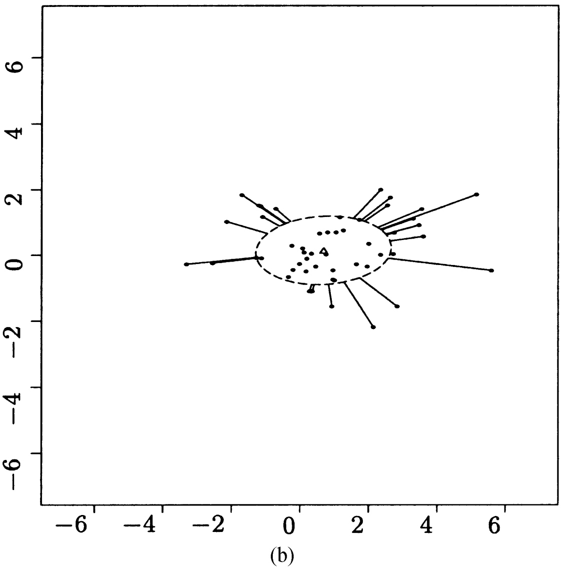
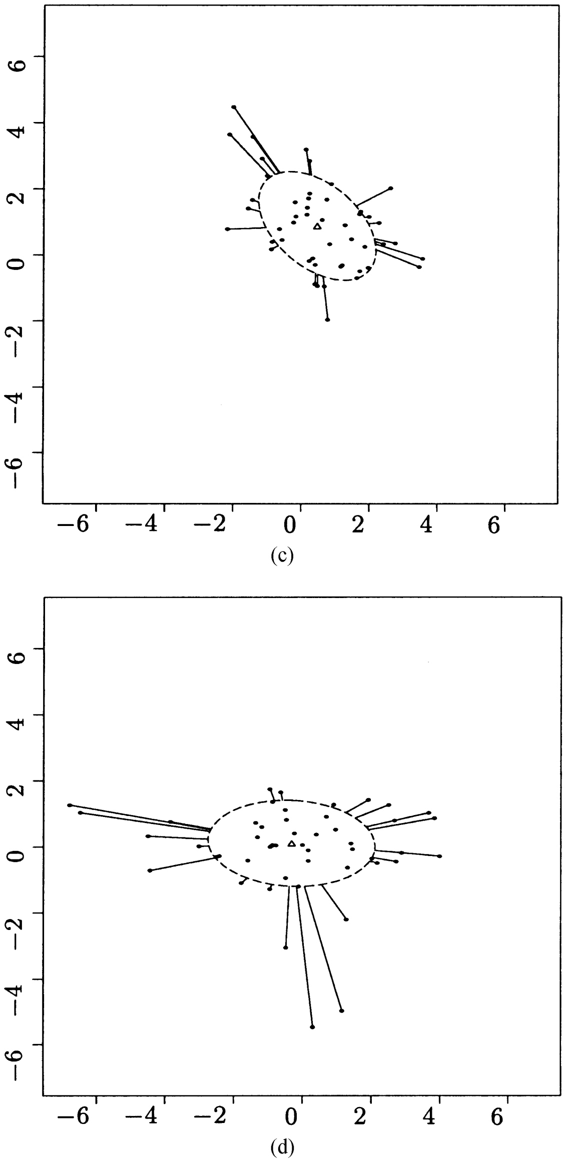


FIG. 1. *Sunburst plots.* (*a*) (A), (*b*) (B), (*c*) (C) *and* (*d*) (D).

plots for the four situations in (A)–(D) above. In each plot, the center of the ellipse is the sample Oja median and the shape and the orientation are given by the RCM.

Note especially in the contaminated normal case (D) that the rays highlight the extreme observations. Further, compare (A) and (D) and notice that



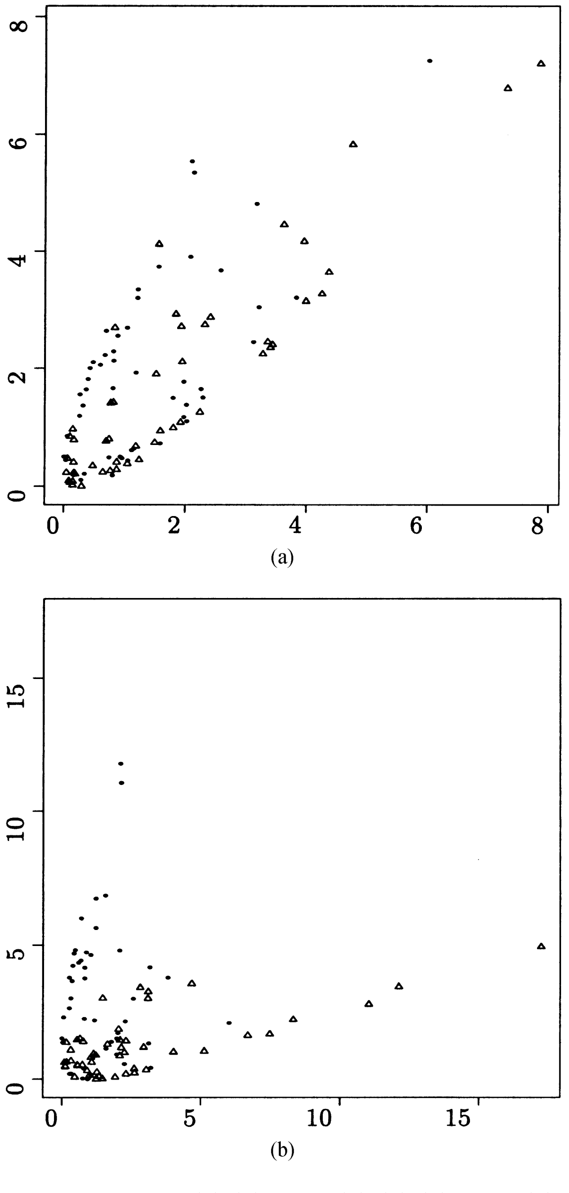
(c)

(d)

FIG. 1. (*Continued*)

the contamination did not change the ellipse by very much. This is a reflection of the robustness of the RCM approach.

Next, in Figure 2 we show the *DD*-plots for comparisons of (A) to (B), (C) and (D). In our *DD*-plots we use $d(\mathbf{x})$ rather than $(1 + d(\mathbf{x}))^{-1}$ since the latter compresses outliers into the lower left corner of the plot near the origin. Our plots correspond to some of the Figure 20(a–h) in the paper.

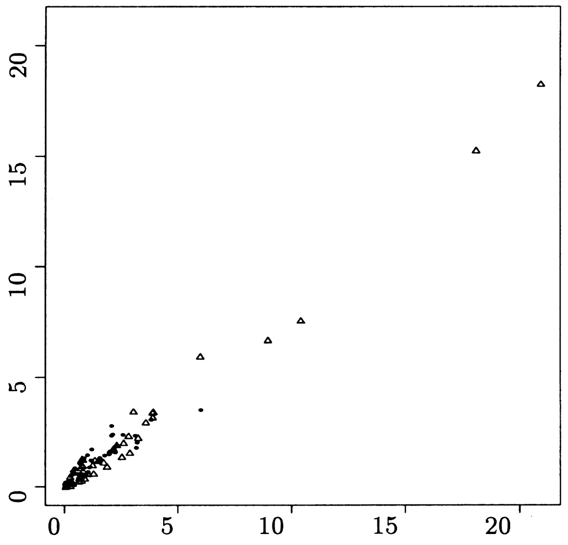


(a)

(b)

FIG. 2. *DD-plots.* (*a*) (A) *versus* (B), (*b*) (A) *versus* (C).

The shape and distribution in these plots reflect differences in location in (A) versus (B), differences in location and orientation in (A) versus (C) and effects of contamination in (A) versus (D). In cases where there is a high concentration of points near the origin, a logarithmic scale may be more revealing but we do not pursue that here.



(c)

FIG. 2. (*Continued*). (*c*) (A) *versus* (D).

The scale plot is, in the bivariate case, a plot of the area of the ellipse enclosing a proportion p of the data versus the proportion p. A more rapid increase in the plot indicates larger underlying scale. In Figure 3 we compare (A) to (D) on both the natural scale and on a logarithmic scale. Only scale differences λ are revealed since the scale plot does not depend on the location μ, shape C or orientation U.

The log scale facilitates comparison of scale near the centers. Compare these plots to Figure 7(a, b) in the paper. The other nice application discussed by the authors is for the comparison of scatter of the multivariate estimates of location; see Figure 8(a, b, c) in the paper. The comparison based on ellipses would be quite natural here since, typically, the estimators will have multivariate normal limiting distributions.

Another way to compare scales for two distributions is to look at a *PP*-plot of the elliptical areas for the two samples. Essentially, it is a plot of the empirical cdf's of the elliptical areas determined by the data in each sample. Figure 3 shows a *PP*-scale plot of (A) versus (D).

Note that beyond 0.5 the empirical cdf's of the elliptical areas, $\hat{F}_A(u)$'> $\hat{F}_D(u)$', indicating that (D) has more scatter or larger scale than (A). The area under the curve could provide a measure and, hence, in the elliptical case, an asymptotically distribution-free test for scale differences. The test statistic then is the Mann–Whitney–Wilcoxon U-statistic calculated from the depths. In the univariate case, this corresponds to a rank test based on magnitudes of the centered observations. In the comparison in Figure 4, the observed p-value (one-sided test) is 0.22.

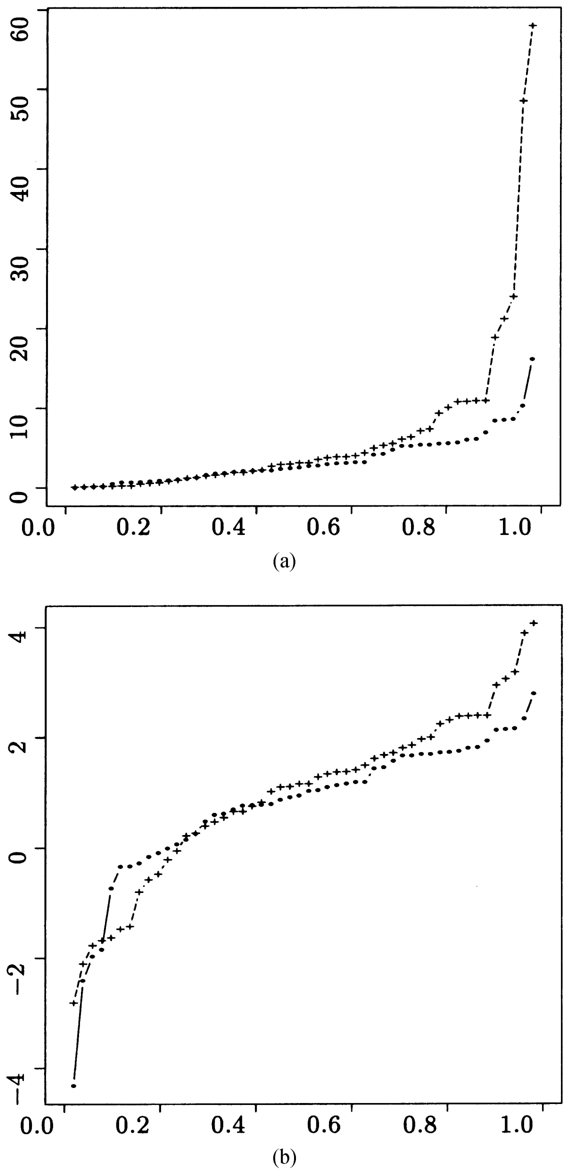


FIG. 3. *Scale plots.* (*a*) (A) *versus* (D), (*b*) (A) *versus* (D) (*logarithm scale*).

Finally, we consider a simple plot for comparing tailweight or kurtosis across two samples. First we must standardize the scales in some way. We simply standardize the depths by dividing the depths by their respective medians of depths in the two samples. Then both samples have median depth equal to 1. Now an S-shaped curve in the standardized *PP* scale plot

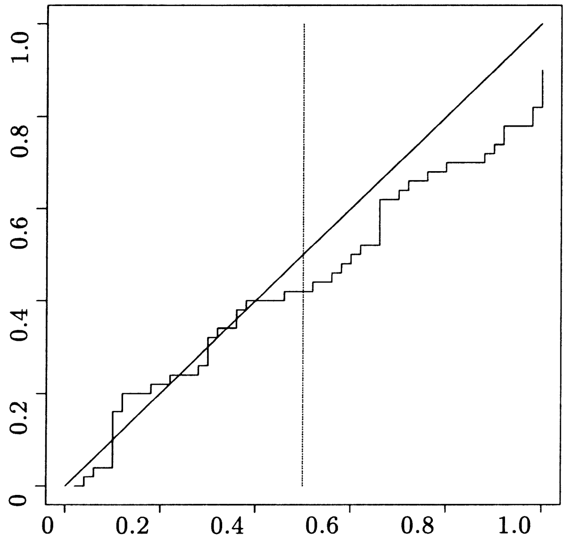


FIG. 4. *PP-scale plot* (A) *versus* (D).

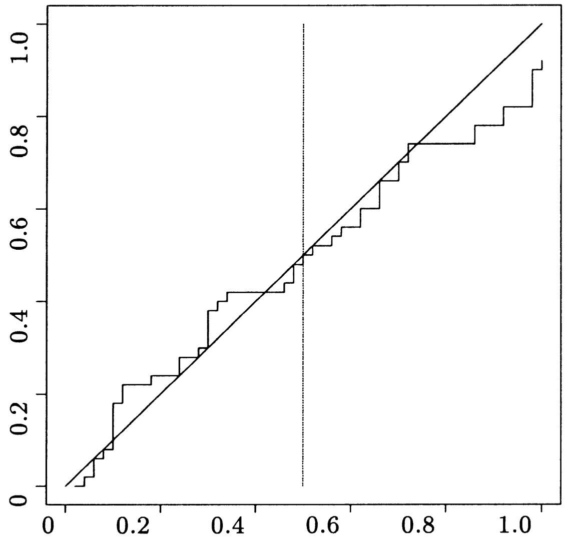


FIG. 5. *PP kurtosis plot* (A) *versus* (D).

indicates difference in kurtosis; see Figure 5 which compares (A) to (D). The increased kurtosis due to contamination can now be seen. An asymptotically distribution-free test for comparing the kurtosis in elliptic cases could be constructed using the difference between the areas under the curve for lower values ($0 < p < 0.5$) and above the curve for upper values ($0.5 < p < 1$). The observed p-value (one-sided test) for this comparison is now 0.07.

The paper contains many more interesting plots for other features such as skewness. We look forward to the development of associated inferences.

REFERENCES

BENSMAIL, H. and CELEUX, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *J. Amer. Statist. Assoc.* **91** 1743–1749.

HETTMANSPERGER, T. P., MÖTTÖNEN, J. and OJA, H. (1998). Affine invariant multivariate rank tests for several samples. *Statist. Sinica* **8** 785–800.

OJA, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: a review. *Scand. J. Statist.* (invited paper).

VISURI, S., KOIVUNEN, V. and OJA, H. (1999). Sign and rank covariance matrices. Conditionally accepted to the *J. Statist. Plann. Inference.*

T. P. HETTMANSPERGER
DEPARTMENT OF STATISTICS
PENN STATE UNIVERSITY
317 CLASSROOM BLDG.
UNIVERSITY PARK, PENNSYLVANIA 16802-2111
E-MAIL: tph@stat.psu.edu

H. OJA
UNIVERSITY OF JYVÄSKYLÄ

S. VISURI
TAMPERE UNIVERSITY OF TECHNOLOGY

REJOINDER

REGINA Y. LIU AND KESAR SINGH

Rutgers University

We thank all discussants for their interesting and encouraging comments. Since the three discussions focus on largely different aspects of our paper, we shall respond to each of them separately.

Discussion by W. F. Eddy. The Schervish paper (1987) recommended by Eddy is indeed a useful reference. It provides a good survey of developments in multivariate analysis from the 1960s to the 1980s, concentrating mostly however on normal distributions. As for Eddy's other comments:

Exploratory data analysis. We agree that our work has a strong exploratory data analysis flavor, although we certainly think that it has a solid theoretical foundation and belongs definitely to "standard" nonparametric statistics! In fact, by selecting a suitable notion of depth, we also recover many traditional approaches. Thus classical multivariate analysis corresponds to the notion of Mahalanobis depth and the general density estimation approach to the likelihood depth.

We are gratified that Eddy views the sunburst plot as a nice and "in the right spirit" generalization of the boxplot. We note only that:

1. The *many whiskers* (*rays*) may seem redundant, but their concentration reflects the directions of probability mass concentration. This can be useful in characterizing the underlying multivariate distribution and may be viewed as an advantage of the sunburst plot over the univariate boxplot.
2. More contours would admittedly provide more information about the distribution, but may obscure the simplicity of the idea of the boxplot.
3. As already indicated in our paper, the bagplot in Rousseeuw and Ruts (1997) is the same as our sunburst plot with an additional "fence" built in to detect outliers. This fence is a generalization of the fence in the univariate boxplot.

Computing and graphics. We share with Eddy his enthusiasm for computer graphics and believe that dynamic graphics will be a powerful aid to multivariate analysis. Although dynamic graphics have not been treated as yet in the present paper, it may be worth pointing out that our plots made full use of state-of-the-art algorithms for computing data depths and could hardly be done "by hand." In fact, the best algorithms so far for the computation of the two- or three-dimensional half-space or simplicial depths still require the order of computations $O(n^{d-1} \log n)$ [Rousseeuw and Ruts (1996) and Rousseeuw and Struyf (1997)]. In plotting scale curves, the computation of the area or volume of the sample pth central region (which is the convex hull of the $100p\%$ deepest points) adds another layer of complexity. Some problems, such as large data sets of the size of 120 billion, may be more an issue of computational power or of applying a suitable data reduction method to make the data set more manageable. But others, such as designing better algorithms, may lead to new and interesting questions of both a theoretical and practical nature. We view the improvement of computational feasibility as one of the most important research directions in the next stage of the development of the theory, one which will hopefully attract the attention of statisticians and computer scientists alike. In this context, we note that the computational aspects of various geometric notions of data depth, particularly the simplicial and half-space depths, have and continue to generate much interest in computer science. See, for example, Gil, Steiger and Wigderson (1992), Cheng and Ouyan (1998) and Johnson, Kwok and Ng (1998).

We believe that one of the main achievements of our paper is actually to have provided a way of visualizing multivariate distributional characteristics by *one-dimensional* curves. It is the very simplicity of such objects which makes them powerful as a general tool for the practicing statistician. For example, the scale curves described in Section 4.2 have been applied with much success in Cheng, Liu and Luxhoj (1999) to provide a clear ranking of ten air carriers in terms of the degree of consistency of their multiple performance measures collected by the Federal Aviation Administration. For some special purposes, it may be really necessary to visualize the data depth or sunburst plot of a ten-dimensional distribution or to layer our fan plots for the purpose of comparing several multidimensional distributions. Computer graphical tools such as the *pan and zoom*, *rotation* and other methods discussed in the article Becker, Cleveland and Wilks (1987) may then come in handy.

Dependence. In this paper, we have introduced basic descriptive statistics based on data depth. Our approach is both exploratory (geometric) and probabilistic, but it is only the first step in developing a broader and more flexible alternative to classical, normal-based multivariate analysis. Clearly, dependence is one of the most important steps further down the road. Others include general inference methods and regression methods. It should be mentioned that Rousseeuw and Hubert (1999) and Teng (1999) have recently introduced very promising regression methods based on depth, so that the prospects for rapid progress are quite real.

Discussion by K. A. Baggerly and D. W. Scott. As Baggerly and Scott noted and as is also apparent from our response to Eddy, our approach requires at this moment considerable computational power. This difficulty should lessen over time with the advent of ever better computers and softwares. More important, we share with Baggerly and Scott the opinion that the computational aspects of the data depth approach should be viewed as a research opportunity, in statistics as well as in related fields such as computer science and combinatorics. On the other hand, we would like to stress that the approach is not limited to two or three dimensions or to just unimodal data, as may be inferred from Baggerly and Scott's last paragraph. From a conceptual viewpoint, the computation of, say, simplicial depth in any dimension is actually easy, since it requires only the solution of a system of linear equations.

Baggerly and Scott also observe that there is no unique way of extending ordering to the multivariate setting, and we readily agree. This is why we felt it was important for our data depth approach to be applicable to as many notions of depth as possible. For a specific purpose of a given statistical analysis, a certain notion of depth may be more suitable than others. For example, the Mahalanobis depth captures well the mean of an elliptical distribution, the simplicial and half-space depths are more adept at identifying a central (or median) point of a general distribution, while the likelihood depth (i.e., density estimation) is more desirable for identifying the modes. If a "center" for a distribution is required, as in statistical process control, then either the simplicial or the half-space depth would be more appropriate than density estimates. This is supported by the nearly convex contour plot in (a) of Figure 1. Note that the notion of a center is very different from the notion of a mode. The center of a distribution is well understood conceptually even when the distribution has multiple modes. Of course, the method of density estimation is particularly well suited to detecting modes or separating mixtures, as amply shown in the book of Scott (1992). In this novel field of computer-aided statistical analysis, this may be yet another argument for exploring simultaneously as many approaches as possible. In the context of Bayesian analysis, we believe that the likelihood depth can be particularly useful for defining P-values in testing hypotheses using a posterior distribution, and we plan to explore this topic further.

Discussion by T. P. Hettmansperger, H. Oja and S. Visuri. We appreciate very much the strong endorsement of our approach by Hettmansperger, Oja and Visuri. Indeed, the results of most of our proposals are graphs with easy interpretations. To us, this was in fact the main motivation for this research. The best illustration is perhaps the scale curve. It tells a simple and yet rather complete story of scale in terms of some positive numbers which are just the volumes of the growing central regions. We are very excited to see that Hettmansperger, Oja and Visuri (HOV) have carried our proposals to the next stage of inference. Their illustrations of our various plots in the elliptical setting together with many P-values demonstrate well the potential of the data depth approach in multivariate analysis.

Among all notions of depth, the Mahalanobis depth is clearly the most suitable for the analysis of elliptical models. HOV have replaced the center and the dispersion matrix in the Mahalanobis depth with their robust estimators and produced more robust outcomes. This provides a robust alternative to the standard analysis of elliptical models. A minor clarification here: the expression $d(\mathbf{x})$ in HOV is the distance of $\mathbf{x}$ to the center and, in our framework, a smaller $d(\mathbf{x})$ value is actually associated with a deeper point. In order to be consistent with our notion that a higher depth value indicates a deeper position with respect to the underlying distribution, we chose in our paper to define depth by $(1 + d(\mathbf{x}))^{-1}$ instead of $d(\mathbf{x})$. Obviously, the results of the analysis will be the same with either choice.

One of our current projects is to develop formal inference based on our proposals. We are naturally very encouraged by the successes shown in the P-values obtained by HOV. We also hope to add some asymptotics and bootstrap-related developments to provide a fuller range of depth-based inference. The idea of *pp*-plots for scale and kurtosis comparisons introduced in HOV is very nice indeed! It seems to us, however, that the orientations (as well as the locations) of the distributions may need to be aligned before plotting the *pp*-plot; consider two populations which are elliptical and quite elongated, and assume that they are orthogonal in terms of orientation but otherwise identical. It appears then that the *pp*-plot may not provide an accurate picture of the scale comparison, since it will imply that one distribution has a higher scale than the other. This concern should be easily taken care of, and we expect this type of plot to become soon a standard tool in multivariate data analysis.

REFERENCES

BECKER, R. A., CLEVELAND, W. S. and WILKS, A. R. (1987). Dynamic graphics for data analysis (with discussion). *Statist. Sci.* **2** 353–395.

CHENG, A., LIU, R. and LUXHOJ, J. (1999). Monitoring multivariate processes: control charts, culpability indices, consistency curves and threshold systems. Preprint.

CHENG, A. and OUYANG, M. (1998). On algorithms for computing simplicial depth. Preprint.

GIL, J., STEIGER, W. AND WIGDERSON, A. (1992). Geometric medians. *Discrete Math.* **108** 37–51.

JOHNSON, T., KWOK, I. and NG, R. (1998). Fast computation of 2-dimensional depth contours. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*.

ROUSSEEUW, P. and HUBERT, M. (1999). Regression depth (with discussion). *J. Amer. Statist. Assoc.* **94** 388–433.

ROUSSEEUW, P. and RUTS, I. (1996). A5 307: bivariate location depth. *Appl. Statist.* **45** 516–526.

ROUSSEEUW, P. and STRUYF, A. (1998). Computing location depth and regression depth in higher dimensions. *Statist. Comput.* **8** 193–203.

SCHERVISH, M. J. (1987). Multivariate analysis (with discussion). *Statist. Sci.* **2** 396–433.

SCOTT, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

TENG, J. (1999). New methodology in regression and multivariate quality control via data depth. Ph.D. thesis. Dept. Statistics, Rutgers Univ. To appear.

DEPARTMENT OF STATISTICS
HILL CENTER
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854-8019
E-MAIL: rliu@stat.rutgers.edu
kesar@stat.rutgers.edu