# Depth functions in nonparametric multivariate inference

**Article** · November 2006

DOI: 10.1090/dimacs/072/01



1 author:

Robert Serfling
University of Texas at Dallas

DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

# Depth Functions in Nonparametric Multivariate Inference

Robert Serfling

*This paper is dedicated to Regina Liu, who opened up the view of "depth functions" as a broad and general approach.*

Abstract. Depth functions, as an emerging methodology in nonparametric multivariate inference, are reviewed in brief. The special relationships among *depth*, *outlyingness*, *centered rank*, and *quantile* functions are indicated.

## 1. Summary

In passing from univariate to *multivariate* statistical analysis, especially for the purpose of *nonparametric* approaches, various issues and special considerations come into play. We examine some of these in Section 2 and address the question

> Where do *depth functions* fit into nonparametric multivariate inference?

Section 3 gives an overview of depth functions with emphasis on their connections with outlyingness functions. We consider depth functions defined not only on the *observation space*, with orientation to *nonparametric multivariate description*, but also as defined on the *parameter space*, for example on the multivariate space of "regression fits" in univariate multiple regression. Section 4 examines quantile and centered rank functions as entities closely related to each other and connects them with depth and outlyingness functions. Some brief historical notes are provided in Section 5 and a concluding remark in Section 6.

## 2. Nonparametric Multivariate Analysis

To capture the setting for considering depth functions, let us examine some key perspectives that are relevant to the choice of a procedure in nonparametric multivariate analysis.

Use of normal model versus nonparametric description. *Parametric* modeling of multivariate data enjoys few tractable models other than the normal,

---

2000 *Mathematics Subject Classification.* Primary 62G05; Secondary 62G20, 62H05.
*Key words and phrases.* Multivariate analysis, Nonparametric methods, Depth functions, Quantile functions, Outlyingness functions, Centered rank functions.
The author gratefully acknowledges support by NSF Grant DMS-0103698.

which thus occupies the central position along with comprehensive treatment and wide application. Therefore, among the alternatives to reliance on normal models, the *semiparametric* and *nonparametric* approaches are even more significant in the multivariate case than in the univariate case. For the multivariate case, however, such methods are far from fully developed. No single definitive nonparametric methodology has emerged.

Coordinatewise versus geometric methods. The classical approach to nonparametric inference for data in $\mathbb{R}^d$ utilizes $d$-vectors of univariate statistics. Examples include vectors of *means*, or of *medians*, or of *sign statistics*. This indeed yields worthwhile procedures but fails to take into account the *geometric features* often inherent in an important way in multivariate data. For example, the vector of coordinatewise medians fails to properly represent the "center" of a point cloud in $\mathbb{R}^d$ and even can be located outside the convex hull of the data. In recent years, therefore, *multidimensional* notions of standard univariate concepts have been increasingly pursued. For the *mean*, this is immediate and unequivocal. For notions of *symmetry*, *median*, *sign*, *rank*, and *quantile*, however, in each case there are competing versions that differ critically in both appeal and characteristics.

Center-outward ordering versus linear ordering. The *linear order* on $\mathbb{R}$ induces *ordering* and *ranking* for observations $X_i$ and *quantiles* for corresponding cdf's $F$. Generalization to $\mathbb{R}^d$ for $d \geq 2$ is thwarted, however, by the absence of such an order in this case. As a compensation, however, it is convenient and natural to *orient to a "center"*, which can be defined in a variety of ways. This leads naturally to the use of *center-outward* orderings of points and description in terms of *nested contours*.

Density contours versus outlyingness contours. *Contours* in $\mathbb{R}^d$ are equivalence classes of points $\boldsymbol{x}$ sharing equal values of some function $g(\boldsymbol{x})$. For $g(\cdot)$ a *probability density function*, the contours have interpretability merely in a *local* sense, that is, they characterize the amount of probability mass in a neighborhood of a point. Points in this type of equivalence class need not, however, be equivalent in *outlyingness*. To characterize the outlyingness of points, and to organize points into contours of equal outlyingness, we need $g(\cdot)$ to be an *outlyingness function* in some sense. Contours in multivariate analysis are thus used for two distinctly different purposes: *probability density description* and *outlier identification*. The type of contour to be used depends on which of these is the relevant purpose.

Curse(s) of dimensionality. Dimensionality affects the handling of data in several ways, unfortunately all adverse.

*Distribution of probability and sample.* The fraction of unit $d$-cube in the unit sphere diminishes from 1 for $d = 1$ to 0.0001 for $d = 10$. Over half the probability mass of a 10-variate normal distribution is in the low density tail region. Thus in high dimension a sample falls mostly in the tails, leaving empty most "local" neighborhoods.

*Dimensionality.* Typically, data in $\mathbb{R}^d$ has structure of lesser dimension than the nominal $d$.

*Computional complexity.* The feasibility of computation is a function of the size $n$ and and dimension $d$ of the data, with practical limitations arising in the case of higher $d$.

*Visualization.* Vizualization of data in $\mathbb{R}^d$, and of functions of points in $\mathbb{R}^d$, relies on *contouring* for $d = 3$, lower dimensional *slices* for $d = 4, 5, 6$, and *projection pursuit* for higher $d$.

Choice of approach. For *univariate* statistical analysis, various distinctive methodological approaches have been developed, including

*parametric modeling*,
*robust parametric methods*,
*nonparametric density estimation*,
*empirical likelihood methods*,
*sign and rank* methods*,
*order statistic* methods*,
*quantile* methods*, and
*outlyingness* function methods*.

All of these have multivariate extensions. For each of the latter four, indicated by $^*$, the extensions typically are *ad hoc*, are *various*, and involve sophisticated formulations derived from quite different conceptual orientations. Thus, in the multivariate case, *nonparametric* inference entails selection of an approach from an array of possibilities much more diverse and complex than in the univariate case.

**The role of depth functions**. It turns out that the four methodological choices indicated by $^*$ above may coherently be brought together and interpreted as a single nonparametric methodology, via *depth functions*. These extend to the multivariate setting in a unified way the univariate methods of signs and ranks, order statistics, quantiles, and outlyingness measures. In particular, they provide a basis for generating outlyingness contours, for taking into account the geometry of the data, and for defining a "center" and a corresponding center-outward ordering of points. Depth-based nonparametric measures of *location*, *dispersion*, *skewness*, and *kurtosis* in the multivariate setting can be formulated. Indeed, Liu, Parelius and Singh (1999) stress the importance of *visualizing particular features* of higher dimensional distributions via *one-dimensional curves*. For these and other purposes, *depth functions* offer very effective ways to extract relevant information from data.

## 3. Depth and Outlyingness Functions

Let us try to gain perspective on the following questions: What are *depth functions*? What *properties* are desirable? What are *not* depth functions? What *roles* do they play in applications? What *computational burdens* are involved? How are depth functions connected with multivariate notions of *order statistic*, *rank*, *outlyingness* and *quantile*?

**Definition and purpose**. Associated with a given distribution $P$ on $\mathbb{R}^d$, a *depth function* is designed to provide a $P$-based *center-outward ordering* (and thus a *ranking*) of points $\boldsymbol{x}$ in $\mathbb{R}^d$. *High* depth corresponds to "centrality", *low* depth to "outlyingness". The "center" consists of the point(s) that *globally maximize* depth. Multimodality features of $P$ are ignored.

**Examples**. The seminal example of depth function is the *halfspace depth* (Tukey, 1975): for $\boldsymbol{x} \in \mathbb{R}^d$,

$$D(\boldsymbol{x}, P) = \inf\{P(H) : \boldsymbol{x} \in H \text{ closed halfspace}\},$$

the minimal probability attached to any closed halfspace with $\boldsymbol{x}$ on the boundary. Competing depth functions began with the *simplicial depth* (Liu, 1988): for $\boldsymbol{x} \in \mathbb{R}^d$,

$$D(\boldsymbol{x}, P) = P(\boldsymbol{x} \in S[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}]),$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}$ represent independent observations from $P$ and $S[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}]$ denotes the simplex in $\mathbb{R}^d$ with vertices $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}$, that is, the set of points in $\mathbb{R}^d$ that are convex combinations of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}$. For each fixed $\boldsymbol{x}$, the sample version is a *U-statistic* based on the kernel

$$h(\boldsymbol{x}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}) = \mathbf{1}\{\boldsymbol{x} \in S[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}]\}. \tag{3.1}$$

Other depth functions of particular interest include the *majority depth* (Singh, 1991; Liu and Singh, 1993), the *projection depth* (Liu, 1992; Zuo, 2003), the *Mahalanobis depth* (Liu and Singh, 1993), the *spatial depth* (Chaudhuri, 1996), the *zonoid depth* (Koshevoy and Mosler, 1997) the *$L^p$ depth* (see [**62**]), and the *simplicial volume depth* (based on Oja, 1983; see [**62**]). For illustration, several views of the halfspace and spatial depth functions are provided in Figure 1.

**But the likelihood is not a depth**. It does not in general measure centrality or outlyingness. Its interpretation has no global perspective. It is sensitive to multimodality. The point of maximality is not interpretable as a "center".

**Desirable properties of depth functions**. Effective use of depth functions calls upon a variety of properties. Without details, we list those which are especially desirable and useful. Most of them are satisfied by the leading depth functions (see [**29**], [**62**], [**37**]).

- *Affine invariance.* $D(\boldsymbol{x}, P)$ is independent of the coordinate system.
- *Maximality at center.* If $P$ is symmetric about $\boldsymbol{\theta}$ in some sense, then $D(\boldsymbol{x}, P)$ is maximal at this point.
- *Symmetry.* If $P$ is symmetric about $\boldsymbol{\theta}$ in some sense, then so is $D(\boldsymbol{x}, P)$.
- *Decreasing along rays.* The depth $D(\boldsymbol{x}, P)$ decreases along each ray from the deepest point.
- *Vanishing at infinity.* $D(\boldsymbol{x}, P) \to 0$, $\|\boldsymbol{x}\| \to \infty$.
- *Continuity of $D(\boldsymbol{x}, P)$ as a function of $\boldsymbol{x}$.* Or merely *upper semicontinuity.*
- *Continuity as of $D(\boldsymbol{x}, P)$ a functional of $P$.*
- *Quasi-concavity as a function of $\boldsymbol{x}$.* The set $\{\boldsymbol{x}\colon D(\boldsymbol{x}, P) \geq c\}$ is convex for each real $c$.

**Central regions and volume functional**. With the $\alpha$ *depth inner region* given by $\{\boldsymbol{x} : D(\boldsymbol{x}, P) \geq \alpha\}$, the *pth central region* $C_P(p)$ is that inner region which has probability weight $p$. A corresponding *volume functional* is defined by

$$v_P(p) = \text{volume of } C_P(p).$$

Central regions and the volume functional are instrumental in making application of depth functions. Ideally, and typically, depth-based central regions $\{C_P(p)\}$ are *affine equivariant*, *nested*, *connected*, and *compact*.

**What we compute from a depth function**. Briefly, we suggest how depth functions become used by listing the typical quantities computed from them.

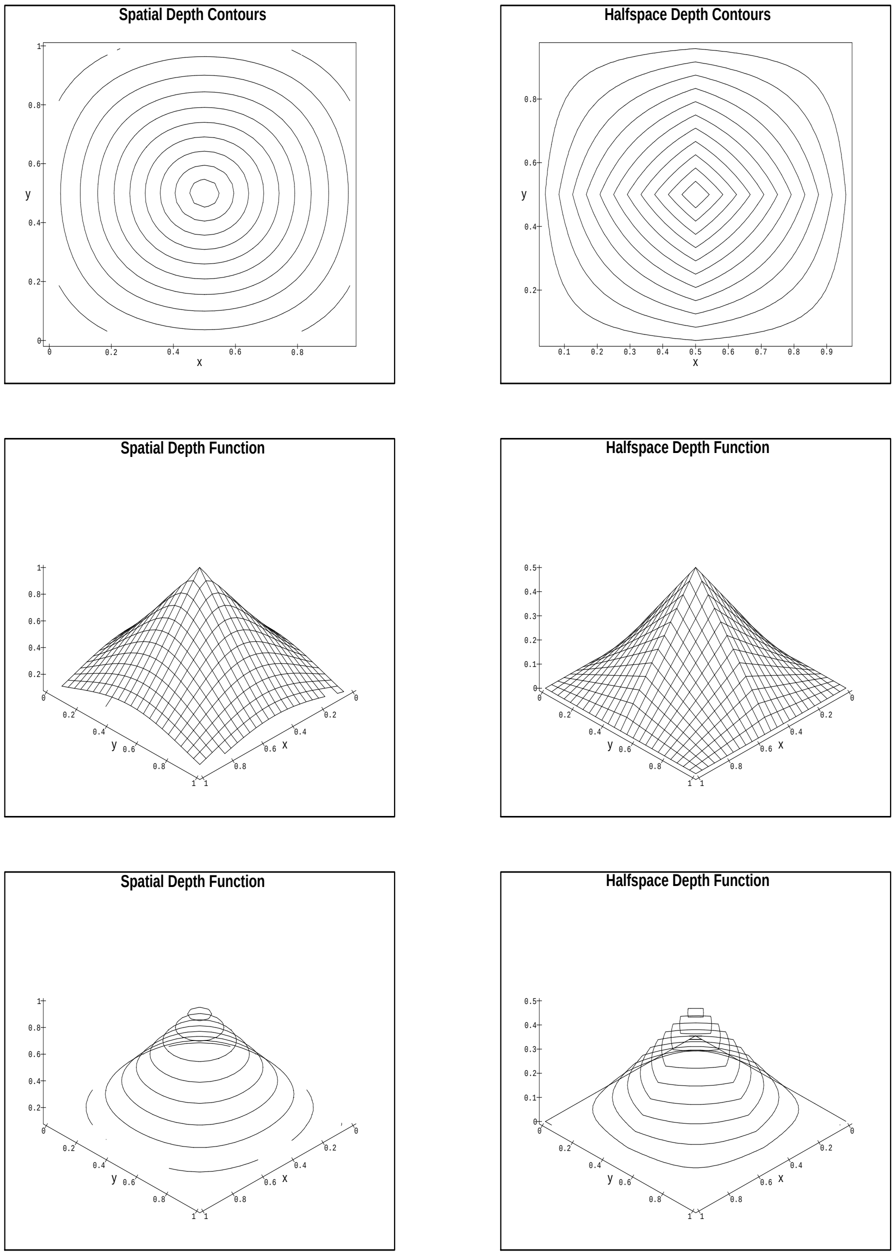


FIGURE 1. Views of Spatial and Halfspace Depth Functions, for $F$ Uniform on Unit Square.

- *Contours*. Boundaries of $p$th central regions for various choices of $p$.
- *Depth-based order statistics*. Ordering of data by depth value, center-outward: $\boldsymbol{X}_{[1]}, \cdots, \boldsymbol{X}_{[n]}$.
- *Depth-weighted location functionals*.

$$\sum c_{i:n} \boldsymbol{X}_{[i]}, \text{ or } \frac{\int_{\mathbb{R}^d} \boldsymbol{x}\, W(D(\boldsymbol{x}, P))\, dP(\boldsymbol{x})}{\int_{\mathbb{R}^d} W(D(\boldsymbol{x}), P))\, dP(\boldsymbol{x})}.$$

- *Depth-weighted scatter matrix functionals*. (Analogously defined.)
- *Volume functional*, $v_P(\cdot)$.
- *Scale curves*. Plot of $v_P(p)$ as a function of $p$.
- *Skewness functionals*. Scaled difference of two location functionals.
- *Kurtosis functionals*. Via transformation of scale curve.

**Desirable theoretical results for sample versions**. Supporting theory for application of depth functions includes results such as *almost sure uniform consistency*, i.e.,

$$\|D(\boldsymbol{x}, \hat{P}_n) - D(\boldsymbol{x}, P)\|_\infty \xrightarrow{a.s.} 0, \ n \to \infty$$

(see [**62**, Appendix A] for discussion), *weak convergence* of $n^{1/2}[D(\boldsymbol{x}, \hat{P}_n) - D(\boldsymbol{x}, P)]$ (e.g., [**33**]), *convergence of sample central regions* (e.g., [**42**]), and convergence of related functionals.

**Depth functions via outlyingness functions**. As noted earlier, "depth" is equivalent to "outlyingness" in an *inverse* sense. Each of these has its own appeal and role, just as cdf's and quantile functions are equivalent but have differing roles. Given an outlyingness function $O(\boldsymbol{x}, P)$ with range $[0, \infty)$, an associated depth function is defined by

$$D(\boldsymbol{x}, P) = \frac{1}{1 + O(\boldsymbol{x}, P)}, \tag{3.2}$$

or, for $O(\boldsymbol{x}, P)$ bounded, by

$$D(\boldsymbol{x}, P) = 1 - O(\boldsymbol{x}, P)/\sup O(\cdot, P). \tag{3.3}$$

The study of outlyingness has a long tradition. Let us now examine how depth functions, including some familiar cases, may be introduced from the conceptual standpoint of outlyingness.

**Formulation of outlyingness functions for location**. Outlyingness of a point $\boldsymbol{x}$ relative to a data set $X = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ in $\mathbb{R}^d$ can be defined in various ways. Below we sketch three approaches toward construction of such functions and for several examples indicate associated familiar depth functions.

Projection pursuit approach. Given any *univariate* outlyingness function $O_{1n}(x, X)$, a corresponding *d-dimensional extension* is defined by

$$O_{dn}(\boldsymbol{x}, X) = \sup_{\|\boldsymbol{u}\|=1} O_{1n}(\boldsymbol{u}'\boldsymbol{x}, \boldsymbol{u}'X), \tag{3.4}$$

i.e., the maximum outlyingness of $\boldsymbol{x}$ within any one-dimensional projection of the data points. As a *general class of examples*, for given univariate location and scale statistics $\mu(X)$ and $\sigma(X)$, respectively, defined on $X$, and given "score" function $\psi$

that measures (signed) deviation from 0 in $\mathbb{R}$, a corresponding univariate outlyingness function is defined by

$$O_{1n}(x, X) = \left| \mu \left( \psi \left( \frac{X_1 - x}{\sigma(X)} \right), \ldots, \psi \left( \frac{X_n - x}{\sigma(X)} \right) \right) \right|.$$

In particular, with $\mu(\cdot)$ the *mean* and $\psi(\cdot)$ the *sign function*, we have, independently of the choice of $\sigma(\cdot)$,

$$O_{1n}(x, X) = \left| n^{-1} \sum_{1}^{n} \operatorname{sign}(X_i - x) \right|,$$

which generates via (3.4) the *halfspace depth*. Or, with $\mu(\cdot)$ the *median*, i.e.,

$$\operatorname{Med}(X) = \operatorname{median}\{X_i, 1 \le i \le n\},$$

$\sigma(\cdot)$ the *MAD*, given by

$$\operatorname{MAD}(X) = \operatorname{median}\{|X_i - \operatorname{Med}(X)|, 1 \le i \le n\},$$

and $\psi(\cdot)$ the *identity function*, we have the classical univariate location outlyingness measure of Mosteller and Tukey (1977),

$$O_{1n}(x, X) = \left| \frac{x - \operatorname{Med}(X)}{\operatorname{MAD}(X)} \right| = \left| \operatorname{median} \left\{ \frac{X_1 - x}{\operatorname{MAD}(X)}, \ldots, \frac{X_n - x}{\operatorname{MAD}(X)} \right\} \right|,$$

which generates via (3.4) the *projection depth*.

Distance approach. Let $h(\boldsymbol{x}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ be a nonnegative function which measures in some sense the *distance* of $\boldsymbol{x}$ from the set of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ in $\mathbb{R}^d$, and define

$$O_{dn}(\boldsymbol{x}, X) = \binom{n}{k}^{-1} \sum h(\boldsymbol{x}; \boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_k}),$$

i.e., the average distance of $\boldsymbol{x}$ from point subsets of size $k$ drawn from the data set. This leads to depth functions using (3.2) for unbounded $h$ and (3.3) for $h$ in $[0, 1]$. For example, with $h(\boldsymbol{x}; \boldsymbol{x}_1) = \|\boldsymbol{x} - \boldsymbol{x}_1\|_p$ using the $L^p$ norm on $\mathbb{R}^d$, we have

$$O_{dn}(\boldsymbol{x}, X) = n^{-1} \sum_{i=1}^{n} \|\boldsymbol{x} - \boldsymbol{X}_i\|_p,$$

i.e., the average $L^p$ distance of $\boldsymbol{x}$ from points in the data set, yielding via (3.2) the $L^p$ depth. Or, with $h$ given by (3.1), we obtain via (3.3) a variant of the *simplicial depth*. Other depths obtained in similar fashion are the *simplicial volume*, *majority*, and *Mahalanobis depths*. See [**62**] for discussion.

Quantile function approach. Let $Q_{\hat{P}_n}(\cdot)$ be a sample multivariate quantile function defined over $\boldsymbol{u}$ in the $d$-dimensional unit ball, and define

$$O_{dn}(\boldsymbol{x}, X) = \|Q_{\hat{P}_n}^{-1}(\boldsymbol{x}\|,$$

i.e., the magnitude of the centered rank of $\boldsymbol{x}$ in the data set (we elaborate on quantile and centered rank functions later). Using the spatial quantile function [**9**], we obtain the *spatial depth* [**57**], [**48**].

**Parameter estimation via outlyingness: location**. We mention two ways to use outlyingness functions for location parameter estimation.

Minimize outlyingness, or maximize depth. For *location* estimation, the *parameter space* and the *data space* coincide. Thus a depth or outlyingness function on the data space may very naturally be viewed as defined on the parameter space. In this case, a natural location estimator is given by choosing the parameter value with minimal outlyingness (or maximal depth). That is, minimize

$$O_{dn}(\boldsymbol{\theta}, X) = \sup_{\|\boldsymbol{u}\|=1} O_{1n}(\boldsymbol{u}'\boldsymbol{\theta}, \boldsymbol{u}'X),$$

which we may regard as a *minimax* approach. The projection pursuit examples discussed above yield, respectively, the *halfspace median* and the *projection median* as location estimators. Modifications of these estimators designed to attain higher efficiency relative to the mean in the normal model and at the same time to attain the optimal breakdown point 1/2 have been developed by Zhang (2002) and Zuo (2003), respectively.

In the univariate case, with for example the outlyingness function

$$O_{1n}(\theta, X) = \left| n^{-1} \sum_{1}^{n} \psi(X_i - \theta) \right|,$$

this approach reduces to classical *M-estimation*: $\hat{\theta}$ is the solution of

$$\sum_{1}^{n} \psi(X_i - \hat{\theta}) = 0.$$

Outlyingness-downweighted means. A natural modification of the usual sample mean estimator of location, designed to achieve greater robustness, is to downweight the more outlying observations. Thus, for a real weight function $w(\cdot)$, define

$$w_i = w(O_{dn}(\boldsymbol{X_i}, X)),$$

and take

$$\hat{\boldsymbol{\theta}}(X) = \frac{\sum_1^n w_i \boldsymbol{X}_i}{\sum_1^n w_i}.$$

This approach has been developed by Mosteller and Tukey (1977) in the univariate case and extended to the multivariate case by Stahel (1981) and Donoho (1982).

**Parameter estimation via outlyingness: dispersion**. We consider analogues of the location case.

Minimize outlyingness, or maximize depth. Again use the projection pursuit approach via

$$O_{dn}(\boldsymbol{C}, X) = \sup_{\|\boldsymbol{u}\|=1} O_{1n}(\sqrt{\boldsymbol{u}'\boldsymbol{C}\boldsymbol{u}}, \boldsymbol{u}'X),$$

where $\boldsymbol{C}$ is a covariance matrix, with either

$$O_{1n}(\sigma, X) = \left| \mu\left( \frac{|X_1 - m(\hat{P}_n)|}{\sigma}, \ldots, \frac{|X_n - m(\hat{P}_n)|}{\sigma} \right) \right|$$

and $\mu(\cdot)$ and $m(\cdot)$ univariate location statistics, or

$$O_{1n}(\sigma, X) = \left| \mu\left( \frac{|X_i - X_j|}{\sigma}, 1 \le i < j \le n \right) \right|,$$

the latter eliminating the need for $m(\cdot)$. In any case, *minimize outlyingness* to obtain a *maximum dispersion depth estimator*. See [**60**] for relevant development.

OUTLYINGNESS-DOWNWEIGHTED MEANS. Use

$$\hat{\boldsymbol{\Sigma}}(X) = \frac{\sum_1^n w_i(\boldsymbol{X}_i - \hat{\boldsymbol{\theta}}(X))(\boldsymbol{X}_i - \hat{\boldsymbol{\theta}}(X))'}{\sum_1^n w_i},$$

with $w_i$'s and $\hat{\boldsymbol{\theta}}(X)$ exactly as above for *location* estimation.

**Extended notions of depth function**. Above we have seen outlyingness (and implicitly depth) functions defined on parameter spaces instead of data spaces. Here we take a quick look at the various extended notions of depth function that have appeared.

DATA DEPTH ON CIRCLES AND SPHERES (Liu and Singh, 1992).

REGRESSION DEPTH (Rousseeuw and Hubert, 1999; Bern and Eppstein, 2002). Extends notions of halfspace ("location") depth and simplicial depth to define depth notions for fitted regression lines and corresponding notions of "deepest regression line". For both univariate and multivariate cases of response variable.

TANGENT DEPTH (Mizera, 2002). Encompasses notions of halfspace depth ("location depth") and "regression depth" within a general framework: "tangent depth" defined with respect to "gradient probability fields" and equipped with a differential calculus.

GENERALIZED FORMS OF TUKEY DEPTH (Zhang, J., 2002). Defines a class of outlyingness functions that generalize the one associated with halfspace depth, and selects favorable members.

LOCATION-SCALE DEPTH (Mizera and Müller, 2004). Applies the "tangent depth" to the classical univariate location-scale problem through a "location-scale depth" defined on a relevant bivariate parameter space (on the Klein disk).

**Depth-based statistical procedures in action**. Depth-based methods are providing strong and in some cases *especially natural* or *advantageous*, competitors to standard approaches for exploratory data analysis, multi-sample inference, regression, classification, clustering, discrimination, directional analysis, and multivariate density estimation, for example. Contexts include monitoring of aviation safety data, industrial quality control, measurement of economic disparity and concentration, social choice and voting, and game-theoretic analysis of competition. A variety of depth-based nonparametric multivariate statistical methods have been developed. The range, flexibility, and potential of the depth function approach can be indicated by a few examples.

BAGPLOTS, SUNBURST PLOTS. These extend univariate boxplots to dimensions 2 and 3, using contours along with rays to outlying points. They answer questions like: *where is the "middle half" of the data?* (See [**29**], [**46**].)

DD, PP, QQ PLOTS, ETC. One may compare two samples by a plot of depth values of the combined samples, or of the volumes of the sample central regions versus each other, or of the respective depth-based quantiles versus each other. (See [**29**], [**32**].)

COMPARISON OF SEVERAL DISTRIBUTIONS. One may plot several scale curves in a single exhibit, or, alternatively, several kurtosis curves in a single exhibit. (See [**29**].)

Nonparametric description of multivariate distributions. Depth-based versions of the basic descriptive measures, *location*, *spread*, *asymmetry*, and *kurtosis*, are being developed and explored. (See [**29**], [**49**], [**58**].)

Tests of multivariate symmetry. For example, to test *spherical symmetry*, plot the fraction of data within the smallest sphere containing the $p$th sample central region. For *central symmetry*, plot the fraction of data within the intersection of the $p$th sample central region and its reflection. (See [**29**].)

Diagnosis of nonnormality. Use a trimmed depth-weighted scatter matrix. Or a kurtosis curve.

Outlier identification. (Discussed above.)

$P$-values in hypothesis testing via bootstrap and data depth. (See [**31**].)

Statistical process control procedures and depth-based quality indices. One can use depth-based ranks for monitoring and thresholding with multivariate data via univariate quality control procedures. The outlyingness of a new observation can be evaluated relative to an in-control reference point cloud. (See [**30**].)

Multivariate density estimation by probing depth. (See [**14**].)

Depth-based classification and clustering. (See [**10**], [**23**].)

**Computational burden**. Depth-functions and depth-based procedures are presenting challenging new problems in computational geometry. Many results on complexity have been established, more are in progress. In the bivariate case, for example, all *halfspace contours*, *the halfspace depthh bagplot*, and all *halfspace data depths* can be computed in $O(n^2)$ time (see [**35**]), and the *deepest regression line* in $O(n \log^2 n)$ time (see [**56**]). For some further references, see [**53**].

## 4. Quantile and Centered Rank Functions

The term "quantile" in the multivariate case has become used rather too loosely. Thus we ask:

> How to formulate *multivariate quantile functions*?
>
> What *properties* are desirable?
>
> What are the *interrelations* among multivariate notions of *order statistic*, *rank*, *depth*, *outlyingness*, and *quantile*?

Let us try to gain perspective on these questions.

**Some basic ideas**. In the univariate case, quantiles represent *boundary points* that demark specified lower and upper fractions of the population. Each point $x \in \mathbb{R}$ has a *quantile interpretation*: it may be written as $F^{-1}(p)$ for some $p \in (0, 1)$. For extension to $\mathbb{R}^d$, $d > 1$, it is convenient and natural to *orient to center*, as something of a compensation for lack of an order. For *quantile-based* inference in $\mathbb{R}^d$, for "center" one should adopt some notion of multidimensional *median* $M$. The center $M$ then serves as the *starting point* for developing a *median-oriented* formulation of multivariate quantiles.

**Formulation for the univariate case**. The median $M$ is given by $F^{-1}(1/2)$. The "$p$th central region" may be defined by the closed interval

$$\left[F^{-1}\left(\frac{1-p}{2}\right),\ F^{-1}\left(1-\frac{1-p}{2}\right)\right],$$

which has probability weight $p$. (This particular choice equalizes tail probabilities.) For $p = 1/2$, this gives an "interquartile region" whose width is the usual IQR. As $p \to 0$, this reduces to the median $M$. Each $x \in \mathbb{R}$ has a *quantile interpretation*: a boundary point of some $p$th central region. In this sense, the associated $p$ indicates the "outlyingness" of $x$. The $p$th central regions are *nested intervals* ordered by *probability weight*, $0 \le p < 1$, or, from another point of view, ordered and indexed by an *outlyingness parameter*.

**Equivalent univariate formulation**. A "median-oriented quantile function" $Q_F(u)$, with $u = 2p - 1$ and median $M = Q_F(0)$, is defined by

$$Q_F(u) = F^{-1}\left(\frac{1+u}{2}\right),\ -1 < u < 1,$$

with sign of $u$ corresponding to direction from $M$. The quantile function $Q_F(\cdot)$ has *inverse*

$$Q_F^{-1}(x) = 2F(x) - 1,\ x \in \mathbb{R},$$

which is recognized to be the usual *centered rank function*. Its magnitude $|Q_F^{-1}(x)|$ $= |2F(x) - 1|$ in a natural way measures the *outlyingness* of $x$ relative to the distribution $F$. Since $x$ satisfies $x = Q_F(Q_F^{-1}(x))$, we may think of the quantiles $Q_F(u)$ as indexed by a directional *outlyingness* parameter $u$ whose magnitude $|u|$ measures outlyingness numerically, with "central" and "extreme" quantiles $Q_F(u)$ corresponding to $|u|$ close to 0 and 1, respectively. Also (in the univariate case), $|u|$ gives the *probability weight* of the central region bounded by $Q_F(\pm u)$.

**Multivariate formulation**. A "median-oriented quantile function" $Q_F(\boldsymbol{u})$ is defined, for $\boldsymbol{u}$ in the unit ball $\mathbb{B}^{d-1}(\mathbf{0})$, with $M_F = Q_F(\mathbf{0})$ a version of $d$-dimensional median. The quantile function $Q_F(\cdot)$ has an *inverse* $Q_F^{-1}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, satisfying $\boldsymbol{x} = Q_F(Q_F^{-1}(\boldsymbol{x}))$. We may interpret this inverse as a *centered rank function* in $\mathbb{R}^d$. Its magnitude $\|Q_F^{-1}(\boldsymbol{x})\|$ in $[0, 1)$ measures the *outlyingness* of $\boldsymbol{x}$ relative to $F$. Thus, again, $\|\boldsymbol{u}\|$ measures outlyingness of $Q_F(\boldsymbol{u})$, with "central" and "extreme" corresponding to $\|\boldsymbol{u}\|$ close to 0 and 1, respectively. But, in the multivariate case, $\boldsymbol{u}$ need *not* be the direction of $Q_F(\boldsymbol{u})$ from $M_F$, nor need $\|\boldsymbol{u}\|$ be the probability weight of the central region bounded by $Q_F(\boldsymbol{u})$ for fixed $\|\boldsymbol{u}\|$. We also associate with $Q_F(\cdot)$ a *depth function*: $D(x, F) = 1 - \|Q_F^{-1}(\boldsymbol{x})\|$. In sum: $Q_F^{-1}(\boldsymbol{x})$ is interpreted as a *centered rank function* whose magnitude is an *outlyingness function* that generates a *depth function*.

**Special case: depth-based quantile functions**. Let $D(\cdot, F)$ be a given depth function maximized at a point $M_F$ interpreted as "median". For convenience, let $F$ be continuous with support $\mathbb{R}^d$. For $\boldsymbol{x} \in \mathbb{R}^d$, let $p$ index the corresponding central region with $\boldsymbol{x}$ on its boundary, and put $\boldsymbol{u} = p\boldsymbol{v}$, for $\boldsymbol{v}$ the unit vector toward $\boldsymbol{x}$ from $M_F$. Setting $Q_F(\boldsymbol{u}) = \boldsymbol{x}$, with $Q_F(\mathbf{0}) = M_F$, the points $\boldsymbol{x} \in \mathbb{R}^d$ generate a quantile function $Q_F(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{B}^{d-1}(\mathbf{0})$. Here the outlyingness parameter $\|\boldsymbol{u}\|$ also gives the *probability weight* of the central region with $Q_F(\boldsymbol{u})$ on its boundary, as in the univariate case. Here $\boldsymbol{u}$ gives the direction toward $Q_F(\boldsymbol{u})$ from the median

$M_F$, also as in the univariate case. The contours of the depth $\|1 - Q_F^{-1}(\boldsymbol{x})\|$ agree with those of $D(x, F)$. For (as typical) $D(\cdot, F)$ *affine invariant*, $Q_F(\cdot)$ is *affine equivariant*: for $\boldsymbol{b} \in \mathbb{R}^d$ and nonsingular $d \times d$ $\boldsymbol{A}$,

$$Q_{\boldsymbol{AX}+\boldsymbol{b}}\left(\frac{\boldsymbol{Au}}{\|\boldsymbol{Au}\|}\|\boldsymbol{u}\|\right) = \boldsymbol{A}Q_{\boldsymbol{X}}(\boldsymbol{u}) + \boldsymbol{b},\ \boldsymbol{u} \in \mathbb{B}^{d-1}. \tag{4.1}$$

**Special case: the spatial quantile function**. The *spatial* quantile function (Chaudhuri, 1996) may be represented as the solution $\boldsymbol{x} = Q_F(\boldsymbol{u})$ of the equation

$$-E\left\{\frac{\boldsymbol{X} - \boldsymbol{x}}{\|\boldsymbol{X} - \boldsymbol{x}\|}\right\} = \boldsymbol{u},$$

for $\boldsymbol{u} \in \mathbb{B}^{d-1}(\boldsymbol{0})$. The case $M_F = Q_F(\boldsymbol{0})$ is the well-known *spatial median.* The inverse function, $Q_F^{-1}(\boldsymbol{x}) = -E\{(\boldsymbol{X} - \boldsymbol{x})/\|\boldsymbol{X} - \boldsymbol{x}\|\}$, has been treated by Möttönen and Oja (1995) as the "spatial centered rank function". Here $r = \|\boldsymbol{u}\|$ represents the outlyingness of $Q_F(\boldsymbol{u})$ but *not* the probability weight of the central region $\{Q_F(\boldsymbol{t}) : \|\boldsymbol{t}\| \leq r\}$ demarked by $\{Q_F(\boldsymbol{u}), \|\boldsymbol{u}\} = r\}$. Here $\boldsymbol{u}$ is *not* the direction from $Q_F(\boldsymbol{u})$ to $M_F$, but rather the *expected direction* from $Q_F(\boldsymbol{u})$ to a random $\boldsymbol{X}$. The equivariance (4.1) holds only for $\boldsymbol{A}$ proportional to an *orthogonal* matrix. This quantile function has attractive *robustness* and *computational* properties, however, and serves as a basis for useful *nonparametric multivariate descriptive measures* [**49**]. The corresponding depth function,

$$D(\boldsymbol{x}, F) = 1 - \left\|E\left\{\frac{\boldsymbol{X} - \boldsymbol{x}}{\|\boldsymbol{X} - \boldsymbol{x}\|}\right\}\right\|,$$

was first formulated (in a different way) by Vardi and Zhang (2000) and is also discussed in [**48**].

## 5. Some brief history on depth and related notions

Here we very briefly discuss a few significant milestones in the development of the notion of depth function and related concepts as tools in nonparametric multivariate analysis.

Depth. The idea of using a depth function to generate contours in higher dimensional space as analogues of rank and order statistics in univariate inference was introduced by Tukey (1975), formulating the *halfspace depth* on the data space. The properties and performance of this depth function have been much studied [**13**], [**12**], [**42**], [**34**], [**59**], [**44**], [**33**], and it remains a leading version with interesting variants also under study [**60**].

The introduction by Liu (1988, 1992) of a strong competitor, the *simplicial depth*, opened up the potential of "depth functions" as a new methodology both powerful and broad in nonparametric multivariate inference. Besides extensive study of the simplicial depth, this spawned the formulation of other novel depth functions. Among these, the *spatial depth* of Vardi and Zhang (2000) importantly links with the classical *spatial median* (see, e.g., [**52**]), the *spatial quantile function* of Chaudhuri (1966) and the *spatial centered rank function* of Oja and Möttönen (1995).

An innovative reformulation of depth for the setting of the univariate multiple regression model was developed by Rousseeuw and Hubert (1999), who introduced *regression depth* defined on the *parameter space of regression fits*. Mizera (2002) characterized and generalized this approach so as to encompass a very broad range of parametric fitting problems treated via depth functions on parameter space. See Mizera and Müller (2004) for application to the classical univariate location-scale problem through a "location-scale depth" defined on the relevant bivariate parameter space.

Liu, Parelius, and Singh (1999) give a broad overview of depth functions along with a variety of depth-based methodological tools. A review of depth functions from a conceptual standpoint, emphasizing structure and properties, is given by Zuo and Serfling (2000). The first monograph on depth functions, Mosler (2002), emphasizes a particular depth, the *zonoid depth*, but through a treatment having broader appeal. The recent DIMACS workshop (2003) reported in this volume brought researchers together to discuss computational geometry problems arising in connection with depth functions.

The notion of depth has interesting antecedents. In fact, the *halfspace depth* is a special case of the "index functions" (Small, 1987) used in economic game theory by Hotelling (1929) and Chamberlin (1937). Also, the bivariate halfspace depth of a point is equivalent to the "sign test statistic" of Hodges (1955) for the null hypothesis specifying that point as "center" (see also Hill, 1960). Gnanadesikan and Kettenring (1972) develop robust covariance estimates employing a trimmed average of squared Euclidean distance over observations in a central region based on a Euclidean distance center-outward ordering with a robust location estimate as center (see also Gnanadesikan, 1997). We also note that the *Mahalanobis depth* (see, e.g., [**29**], [**62**]) has long been in use in various statistical procedures due to its intuitive appeal and mathematical tractability (despite the restriction to elliptical contours that it imposes).

Outlyingness, ranks, and quantiles. The interconnections among depth, outlyingness, ranks, and quantiles in multivariate inference have been noted earlier. Of course, *outlyingness* has a long history, even in the multivariate case (see, e.g., Barnett and Lewis, 1994). Connections with depth notions are especially prominent in Donoho and Gasko (1992) and Zhang (2002). *Multivariate ranks* have been developed as a completely different line of investigation (e.g., [**4**], [**5**], [**6**], [**17**], [**18**], [**19**], [**39**], [**40**], [**41**]). Likewise, notions of *multivariate quantile function* have been pursued quite separately (e.g., [**2**], [**9**], [**24**], [**3**], [**7**]; see [**47**] for a brief review).

## 6. Concluding Remark

Multivariate *depth*, *outlyingness*, *centered rank*, and *quantile* functions are closely interrelated and essentially equivalent. Specifically, quantile and centered rank functions may be formlulated as mutually inverse, and the magnitude of a centered rank function defines an outlyingness function having an associated depth functions as inverse and whose contours define a quantile function. These four entities differ, of course, in type of appeal and domain of application. Key questions are:

- Are these entities *computable*?
- Can we *interest the computational geometers*?
- Can we assess the behavior of *sample versions*?
- Can we *conceptualize* important extensions?
- Can we *apply* the depth approach *effectively*?

With respect to these questions, much has been done with much more to do.

## 7. Acknowledgment

The author offers very hearty thanks to G. L. Thompson, R. Randles, H. Oja, and anonymous other commentators for very thoughtful and helpful remarks. Also, support by NSF Grants DMS-9705209 and DMS-0103698 is greatly appreciated.

## References

[1] Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd edition. John Wiley & Sons, New York.

[2] Breckling, J. and Chambers, R. (1988). $M$-quantiles. *Biometrika* **75** 761–771.

[3] Breckling, J., Kokic, P. and Lübke, O. (2001). A note on multivariate $M$-quantiles. *Statistics & Probability Letters* **55** 39–44.

[4] Brown, B. M. and Hettmansperger, T. P. (1987). Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society, Series B* **49** 301–310.

[5] Brown, B. M. and Hettmansperger, T. P. (1989). An affine invariant bivariate version of the sign test. *Journal of the Royal Statistical Society, Series B* **51** 117–125.

[6] Brown, B. M. and Hettmansperger, T. P. (1989). On certain bivariate sign tests and medians. *Journal of the American Statistical Association* **87** 127–135.

[7] Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics* **53** 380–403.

[8] Chamberlin, E. (1937). *The Theory of Monopolistic Competition*. Harvard University Press.

[9] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association* **91** 862–872.

[10] Christmann, A. (2002). Classification based on the SVM and on regression depth. In *Statistical data analysis based on the L1 norm and related methods*, Y. Dodge (ed.), pp. 341–352. Birkhäuser.

[11] Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Department of Statistics, Harvard University.

[12] Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estinates based on halfspace depth and projected outlyingness. *Annals of Statistics* **20** 1803–1827.

[13] Eddy, W. F. (1985). Ordering of multivariate data. In *Computer Science and Statistics: The Interface* (L. Billard, ed.), pp. 25–30, North-Holland, Amsterdam.

[14] Fraiman, R., Liu, R. Y. and Meloche, J. (1997). Multivariate density estimation by probing depth. In *$L_1$-Statistical Procedures and Related Topics* (Y. Dodge, ed.), pp. 415–430, IMS Lecture Notes — Monograph Series, Volume 31, Hayward, California.

[15] Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd edition. Wiley, New York.

[16] Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biommetrics* **28** 81–124/

[17] Hettmansperger, T. P., Nyblom, J. and Oja, H. (1992). On multivariate notions of sign and rank. In *$L_1$-Statistical Analysis and Related Methods* (Y. Dodge, ed.), pp. 267–278, North-Holland, Amsterdam.

[18] Hettmansperger, T. P., Nyblom, J. and Oja, H. (1994). Affine invariant multivariate one-sample sign tests. *Journal of the Royal Statistical Society, Series B* **56** 221–234.

[19] Hettmansperger, T. P., Oja, H. and Visuri, S. (1999). Discussion to Liu, Parelius and Singh (1999). *Annals of Statistics* **27** 845–854.

[20] Hill, B. M. (1960). A relationship between Hodges' bivariate sign test and a non-parametric test of Daniels. *Annals of Mathematical Statistics* **31** 1190–1192 (Correction **32** 619).

[21] Hodges, Jr., J. L. (1955). A bivariate sign test. *Annals of Mathematical Statistics* **26** 523–527.

[22] Hotelling, H. (1929). Stability in competition. *Econometrics Journal* **39** 41–57.

[23] Jörnsten, R. (2004). Clustering and classification based on the $\mathrm{L}_1$ data depth. *Journal of Multivariate Analysis* **90** 67–89.

[24] Koltchinskii, V. (1997). $M$-estimation, convexity and quantiles. *Annals of Statistics* **25** 435–477.

[25] Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *Annals of Statistics* **25** 1998–2017.

[26] Liu, R. Y. (1988). On a notion of simplicial depth. *Proceedings of the National Academy of Science USA* **85** 1732–1734.
[27] Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Annals of Statistics* **18** 405–414.
[28] Liu, R. Y. (1992). Data depth and multivariate rank tests. In *$L_1$-Statistics and Related Methods* (Y. Dodge, ed.), pp. 279–294, North-Holland, Amsterdam.
[29] Liu, R. Y., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). *Annals of Statistics* **27** 783–858.
[30] Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association* **88** 252–260.
[31] Liu, R. Y. and Singh, K. (1997). Notions of limiting $P$ values based on data depth and bootstrap. *Journal of the American Statistical Association* **92** 266–277.
[32] Marden, J. I. (1998). Bivariate QQ-plots and spider web plots. *Statistica Sinica* **8** 813–826.
[33] Massé, J.-C. (2004). Asymptotics for the Tukey depth process. Application to a multivariate trimmed mean. *Bernoulli*, in press.
[34] Massé, J. C. and Theodorescu, R. (1994). Halfplane trimming for bivariate distributions. *Journal of Multivariate Analysis* **48** 188–202.
[35] Miller, K., Ramaswami, S., Rousseeuw, P., Sellares, T., Souvaine, D., Steinu, I., and Struyf, A. (2001). Fast implementation of depth contours using topological sweep. In *Proceedings of the Twelfth ACM-SIAM Symposium on Discrete Algorithms*, Washington, D.C.
[36] Mizera, I. (2002). On depth and deep points: a calculus. *Annals of Statistics* **30** 1681–1736.
[37] Mosler, K. (2002). *Multivariate Dispersion, Central Regions and Depth*. Springer, New York.
[38] Mosteller, C. F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, Mass.
[39] Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics* **5** 201–213.
[40] Möttönen, J., Oja, H., and Tienari, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *Annals of Statistics* **25** 542–552.
[41] Möttönen, J., Hettmansperger, T. P., Oja, H. and Tienari, J. (1998), On the efficiency of affine invariant multivariate rank tests. *Journal of Multivariate Analysis* **66** 118–132.
[42] Nolan, D. (1992). Asymptotics for multivariate trimming. *Stochastic Processes and Their Applications* **42** 157–169.
[43] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters* **1** 327–333.
[44] Romanazzi, M. (2001). Influence function of halfspace depth. *Journal of Multivariate Analysis* **77** 138–161.
[45] Rousseeuw, P. J. and Hubert, M. (1999). Regression depth (with discussion). *Journal of the American Statistical Association***94** 388–433.
[46] Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician* **53** 383–387.
[47] Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica* **56** 214–232.
[48] Serfling, R. (2002). A depth function and a scale curve based on spatial quantiles. In *Statistical Data Analysis Based On the $L_1$-Norm and Related Methods* (Y. Dodge, ed.), pp. 25-38. Birkhaüser.
[49] Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference* **123** 259–278
[50] Singh, K. (1991). A notion of majority depth. Preprint.
[51] Small, C. G. (1987). Measures of centrality for multivariate and directional distributions. *Canadian Journal of Statistics* **15** 31–39.
[52] Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Institute Review* **58** 263–277.
[53] Souvaine, D. (2003). Computational geometry and statistical depth measures. ICORS 2003.
[54] Stahel, W. A. (1981). Breakdown of covariance estimators. Research Report 31, Fachgruppe für Statistik, ETH, Zürich.
[55] Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver 1974* (R. D. James, ed.), **2** 523–531.

[56] van Kreveld, M., Mitchell, J. S. B., Roosseeuw, P. J., Sharir, M., Snoeyink, J. and Speckmann, B. (1999). Efficient algorithms for maximum regression depth. *Proceedings of 15th Annual ACM Symposium on Computational Geometry* 31–40.

[57] Vardi, Y. and Zhang, C.-H. (2000). The multivariate $L_1$-median and associated data depth. *Proceedings of National Academy of Science USA* **97** 1423–1426.

[58] Wang, J. and Serfling, R. (2005). Nonparametric multivariate kurtosis and tailweight measures. *Journal of Nonparametric Statistics*, to appear.

[59] Yeh, A. B. and Singh, K. (1997). Balanced confidence regions based on Tukey's depth and the bootstrap. *Journal of the Royal Statistical Society*, *Series B* **59** 639–652.

[60] Zhang, J. (2002). Some extensions of Tukey's depth function. *Journal of Multivariate Analysis* **82** 134–165.

[61] Zuo, Y. (2003). Projection-based depth functions and associated medians. *Annals of Statistics* **31** 1460–1490.

[62] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics* **28** 461–482.

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75083 USA

*E-mail address*: serfling@utdallas.edu

*URL*: www.utdallas.edu/∼serfling