

Depth Advances

Mario Romanazzi and Claudio Agostinelli

Abstract Data depth is a rapidly growing area in nonparametric statistics, especially suited for the analysis of multidimensional data. This paper covers influence functions and robustness, depth-based dispersion measures and a generalization of the basic notion of depth function, called local depth, able to cope with multimodal data.

Key words: Influence function, Local depth, Multivariate dispersion, Simplicial depth, Nonparametric statistical methods.

1 Introduction

Data depth provides a coherent framework for nonparametric analysis of multivariate data and is currently registering several new methodological achievements. Let $X = (X_1, \dots, X_p)^T$ be a random vector with distribution F and let $F_{A,b}$ denote the distribution of $AX + b$. A depth function $d(\cdot; F) : \mathbb{R}^p \rightarrow \mathbb{R}^+ \cup \{0_p\}$ should meet several requirements (Liu, 1990; Zuo and Serfling, 2000):

1. affine invariance, i. e., $d(x; F) = d(Ax + b; F_{A,b})$;
2. if F is centro-symmetric about $c \in \mathbb{R}^p$, then $d(c; F) \geq d(x; F)$ for all $x \in \mathbb{R}^p$;
3. ray-monotonicity: if $\theta \in \mathbb{R}^p$ satisfies $d(\theta; F) \geq d(x; F)$ for all $x \in \mathbb{R}^p$, then $d(\theta + \lambda_1 u; F) \geq d(\theta + \lambda_2 u; F)$ for any p -vector u and $0 \leq \lambda_1 < \lambda_2$;
4. when the norm $\|x\|$ is large enough, the rank of x is negligible, i. e., $\lim_{\|x\| \rightarrow \infty} d(x; F) = 0$.

Mario Romanazzi

Department of Statistics, Ca' Foscari University, Venice, e-mail: romanaz@unive.it

Claudio Agostinelli

Department of Statistics, Ca' Foscari University, Venice, e-mail: claudio@unive.it

The main result of data depth is a ranking of the points of Euclidean space with respect to F which describes the degree of centrality of x , highest values corresponding to the center of the distribution, lowest values corresponding to the tails. Several depth-based functionals summarizing the features of F have been suggested. The most important one is the deepest point $\theta(F) = \arg \max_x d(x; F)$ describing the location of F . In general, for a given depth value $d \geq 0$, one can consider the depth regions $D(d; F) = \{x \in \mathbb{R}^p : d(x; F) \geq d\}$ whose boundary is the d -contour of the depth function. When properties (1) to (4) hold, the depth regions form a family of nested affine equivariant subsets of $\mathbb{R}^p$ whose features convey valuable information about F . For example, the spread of F can be described by the scale curve (Liu et al, 1999) $\gamma(d) = \{(p(d), \text{vol}(D(d; F))), d \geq 0\}$, where $p(d)$ is the probability mass enclosed by $D(d; F)$. In a similar way, the correlation curve (Romanazzi, 2004) summarizes the degree of interdependence of the components of F . The D–D plot is another graphical display for pairwise comparison of observed versus theoretical depth distribution and it can be used for informal inference and data check, similarly to a Q–Q plot. The reader is referred to Liu et al (1999) for a comprehensive coverage of the methodology.

In this work more recent developments are presented. The influence function and breakdown point for the simplicial depth and its maximizer are presented in Section 2, while depth-based dispersion functionals are discussed in Section 3. The final topic is local depth illustrated in Section 4. Results are available for several geometrical depth functions based on simple structures, like Tukey’s halfspace depth, Liu’s simplicial depth and Oja’s simplicial volume depth, but here we concentrate on simplicial depth.

Definition 1. The simplicial depth function (Liu, 1990) $d_S(\cdot; F) \equiv d_S(\cdot)$ is the probability that a random simplex $S_{p+1} = S(x_1, \dots, x_{p+1})$ covers x ,

$$d_S(x) = P_F(S_{p+1} : x \in S_{p+1}).$$

Full details and results for other geometrical depths are given in the references.

2 Robustness

The Influence Function (IF) (Hampel, 1974) is the main tool to investigate local sensitivity of a statistical functional to an infinitesimal perturbation of F . In this section the IF of depth values $d(x; F)$, $x \in \mathbb{R}^p$, and depth maximizer $\theta(F)$ are reviewed. Proofs and illustrations of the results can be found in Chen (1995), Romanazzi (2001, 2008) and Chen and Tyler (2002).

For $0 \leq k \leq p+1$, let A_k be the event that a random sample of $p+1$ elements from $F(\varepsilon, z) = (1-\varepsilon)F + \varepsilon\delta_z$ includes exactly k observations from δ_z and let $p_k(\varepsilon)$ be the corresponding probability

$$p_k(\varepsilon) \equiv P(A_k) = \binom{p+1}{k} \varepsilon^k (1-\varepsilon)^{p+1-k}.$$

Moreover, let $d_S(x|A_k, z)$ be the simplicial depth of x conditional on A_k , that is, evaluated from random simplices with $0 \leq k \leq p+1$ vertices coincident with z . The perturbed simplicial depth value turns out to be $d_S(x; F(\varepsilon, z)) = \sum_{k=0}^{p+1} p_k(\varepsilon) d_S(x|A_k, z)$ and the IF is (Romanazzi, 2008)

$$IF(z; d_S(x; F)) = (p+1)(d_S(x|A_1, z) - d_S(x; F)). \quad (1)$$

The IF of simplicial depth is a bounded function of $x \in \mathbb{R}^p$, linearly dependent on the dimension of the space. This suggests that the simplicial ranks could be exposed to the risk of distortion by outliers in high dimensions. However, the rank of the center of a symmetric distribution seems to be remarkably stable.

Example 1. Let F be an absolutely continuous distribution centro-symmetric about the origin 0_p . The origin is the deepest point and $d_S(0_p) = 2^{-p}$. Moreover,

$$IF(z; d_S(0_p; F)) = \begin{cases} (p+1)(1-2^{-p}), & z = 0_p, \text{ closed simplices} \\ -(p+1)/2^p, & z = 0_p, \text{ open simplices} \\ 0, & z \neq 0_p. \end{cases}$$

Therefore, the simplicial rank of the center should be highly resistant to perturbations, for all p .

Let us now pass to depth maximizers. Suppose F to be absolutely continuous and centro-symmetric about the origin. Under the same hypotheses, the perturbed simplicial median is the maximizer of

$$d_S(x; F(\varepsilon, z)) = \begin{cases} (1-\varepsilon)^{p+1}(d_S(x; F) + (p+1)\frac{\varepsilon}{1-\varepsilon}d_S(x|A_1, z)), & x \neq z, \\ (1-\varepsilon)^{p+1}(d_S(x; F) + \frac{1-(1-\varepsilon)^{p+1}}{(1-\varepsilon)^{p+1}}), & x = z. \end{cases}$$

When F is absolutely continuous and spherically symmetric about the origin, the perturbed simplicial median can be represented as $\theta_S(\varepsilon, z) = t(\varepsilon, z)u_z$, with $0 \leq t(\varepsilon, z) \leq \|z\|$, which shows that it belongs to the closed segment joining the origin and the perturbing point z . However, a complete characterization of $t(\varepsilon, z)$ and IF are still lacking. We end with a result on the breakdown point. The finite breakdown point (Donoho and Gasko, 1992) of the empirical simplicial median satisfies

$$\lim_{n \rightarrow \infty} \varepsilon^*(\theta_S(\hat{F}_n)) \leq \frac{2^p \sup_x d_S(x; F)}{p+1+2^p \sup_x d_S(x; F)} \leq \frac{1}{p+2},$$

almost surely (Chen, 1995, Proposition 5).

3 Multivariate dispersion

The information on F provided by a depth function is by no means confined to location but also includes dispersion, as confirmed by the scale-curve (Liu et al, 1999) and the generalized quantile curves (Serfling, 2002). Hereafter, we will show that some well-known multivariate dispersion coefficients are the Lebesgue integral of a depth function. For any simplex $S_{p+1} = S(x_1, \dots, x_{p+1})$, we write V for the $(p+1) \times p$ matrix whose rows are the vertices $x_1, \dots, x_{p+1}$ of the simplex and we write 1_{p+1} for the $(p+1)$ -vector with unit elements. We recall that the volume of the simplex is a function of the determinant of the augmented matrix $(1_{p+1}; V)$, i.e. (Sommerville, 1958, p. 124)

$$\text{vol}(S_{p+1}) = (p!)^{-1} |\det(1_{p+1}; V)|.$$

Moreover, for any non singular affine transformation $x \rightarrow Ax + b$

$$\text{vol}(S(Ax_1 + b, \dots, Ax_{p+1} + b)) = |\det A| \text{vol}(S(x_1, \dots, x_{p+1})). \quad (2)$$

Standard dispersion orderings in the univariate case are based on the comparison of interquantile intervals. For random numbers X and Y with cumulative distribution functions F and G , X is said to be less dispersed than Y if $F^{-1}(\beta) - F^{-1}(\alpha) \leq G^{-1}(\beta) - G^{-1}(\alpha)$ for all $0 < \alpha < \beta < 1$ (Bickel and Lehmann, 1979). A multivariate generalization of the above formula to the multivariate setting based on the volume of simplices was suggested by Oja (1983) and is reproduced below.

Definition 2. Suppose there is a function $h : \mathbb{R}^p \rightarrow \mathbb{R}^p$ such that $Y = h(X)$. The p -variate random vector $X \sim F$ is less dispersed than the p -variate random vector $Y \sim G$ if for any set of $p+1$ points $x_1, \dots, x_{p+1}$ of $\mathbb{R}^p$

$$\text{vol}(S(x_1, \dots, x_{p+1})) \leq \text{vol}(S(h(x_1), \dots, h(x_{p+1}))).$$

We write $X \leq_\Delta Y$ (or $F \leq_\Delta G$) to mean that Definition 2 holds. Let $\mathcal{F}$ be a suitable family of probability distributions on $\mathbb{R}^p$ including the absolutely continuous distributions. A sensible dispersion measure is required to be a Δ -monotone function and it must be translation invariant and satisfy a type of scale equivariance with respect to linear transformations (Oja, 1983, Definition 4.2).

Definition 3. A function $\psi : \mathcal{F} \rightarrow \mathbb{R}^+ \cup \{0\}$ is a dispersion measure if

- i. for $F, G \in \mathcal{F}$, $F \leq_\Delta G$ implies $\psi(F) \leq \psi(G)$;
- ii. a non singular affine transformation $Y = AX + b$ with distribution $F_{A,b}$ satisfies $\psi(F_{A,b}) = |\det A| \psi(F)$.

Two well-known examples of dispersion measures based on the volume of random simplices are Gini's multivariate coefficient $\psi^{(1)}(F) = E_{x_1, \dots, x_{p+1}} \{\text{vol}(S_{p+1})\}$ and Wilks' generalized variance (Wilks, 1960) $\psi^{(2)}(F) = (p+1)^{-1} E_{x_1, \dots, x_{p+1}} \{(\text{vol}(S_{p+1}))^2\} =$

$\det(\text{Cov}(X))$. We now illustrate the relations between the previous coefficients and the depth functions.

Proposition 1. For $F \in \mathcal{F}$,

$$\psi^{(1)}(F) = \int_{\mathbb{R}^p} d_S(x; F) dx.$$

Proof. The result is a particular application of Robbins' theorem (Robbins, 1944).

Recall that for continuous univariate distributions $d_S(x; F) = 2F(x)(1 - F(x))$ and $2 \int_{\mathbb{R}} F(x)(1 - F(x)) dx$ is a well-known computational formula of Gini's mean difference. The next example deals with the multivariate normal distribution.

Example 2. Let $X \sim N(0_p, I_p)$. According to Proposition 1,

$$\psi^{(1)}(X) = \int_{\mathbb{R}^p} d_S(x; F) dx = 4p^{-1}(p+1) \frac{\pi^{p/2}}{\Gamma(p/2)} \int_{-\infty}^{\infty} \phi^{p+1}(t) dt,$$

where $\Gamma(\cdot)$ is the gamma function and $\phi(\cdot)$ denotes the probability density function of the standard normal distribution (see Efron, 1965, p. 343, for further details). Hence $\psi^{(1)}(X)$ only depends on dimension p . The results for Wilks' generalized variance are similar.

4 Local depth

The monotonicity property is the reason why a depth function always exhibits a unique center, even for data showing multiple centers. It was recently suggested that one way to obtain a ranking system able to recognize multimodality is to constrain the depth function to use information from a neighbourhood surrounding any given point, instead of using information from the whole space. The ranks of different points will be comparable provided the neighbourhood size is kept constant. More theoretical background can be found in Agostinelli and Romanazzi (2007, 2008). We illustrate the general idea with the local versions of simplicial depth. For any simplex S_{p+1} we denote with $t(S_{p+1})$ a scalar measure of its size, like diameter, perimeter or volume.

Definition 4. For any $\tau > 0$, the simplicial local depth function is given by

$$ld_S(x; \tau) = P_F(S_{p+1} : x \in S_{p+1} \cap t(S_{p+1}) \leq \tau). \quad (3)$$

A discussion of the computational problems of simplicial (local) depth with large samples and/or high dimensions can be found in Agostinelli and Romanazzi (2008). An R package `localdepth` for computing local and global depth is available at CRAN.

Example 3. In the scalar case, (3) is equivalent to

$$ld_S(x; \tau) = P_F((\min\{X_1, X_2\} \leq x \leq \max\{X_1, X_2\}) \cap (\max\{X_1, X_2\} - \min\{X_1, X_2\} \leq \tau)) ,$$

that is, the probability that a random interval with length not greater than τ covers x . In the case of absolutely continuous distributions on the real line the expression becomes

$$ld_S(x; \tau) = \int_{x-\tau}^x (F(y+\tau) - F(x)) f(y) dy + \int_x^{x+\tau} (F(x) - F(y-\tau)) f(y) dy. \quad (4)$$

Local depth functions should meet the general properties of depth functions mentioned in Section 1, except monotonicity property. In particular, when the reference distribution is centro-symmetric and unimodal about $c \in \mathbb{R}^p$, local depth should agree with global depth.

Proposition 2. Fix $x \in \mathbb{R}^p$ and let $P_F(\{x\})$ be the probability of the singleton x according to the distribution function F .

- i. If $t(S_{p+1}) = \text{vol}(S_{p+1})$, $ld_S(\cdot; \tau)$ is affine invariant whereas, if $t(S_{p+1}) = \text{diam}(S_{p+1})$, $ld_S(\cdot; \tau)$ is translation and rotation invariant;
- ii. $0 < \tau_1 < \tau_2$ implies $ld_S(x; \tau_1) \leq ld_S(x; \tau_2)$;
- iii. for $\tau > 0$, $P_F(\{x\}) \leq ld_S(x; \tau) \leq d_S(x)$;
- iv. $\lim_{\tau \rightarrow \infty} ld_S(x; \tau) = d_S(x)$;
- v. if $t(S_{p+1}) = \text{diam}(S_{p+1})$, $\lim_{\tau \rightarrow 0^+} ld_S(x; \tau) = P_F(\{x\})$.

Proof. Proofs are straightforward and are omitted.

Local depth $ld_S(\cdot; \tau)$ is a family of functions indexed by the values of τ , the parameter that defines the neighbourhood size and the degree of smoothing. To be effective, the choice of τ must avoid the opposite risks of over-smoothing – potentially important information filtered out – and under-smoothing – uninteresting point-to-point variation shown up. We suggest to set τ equal to a low-order percentile of the distribution of simplex sizes (e. g., diameters or volumes). Percentile orders ranging from 10% to 30% prove effective in most situations.

Local depth, like global depth, implicitly defines a (vector) location parameter

$$\theta_\tau(F) = \arg \max_x ld_S(x; \tau) \quad (5)$$

whose components are the partial centers of F , for a given τ . The following proposition describes the behaviour of (5) for simplicial local depth in the scalar case, when F is unimodal.

Proposition 3. Let X have an absolutely continuous distribution with cumulative distribution function F and density function f strictly positive for all $x \in \mathbb{R}$. Suppose the distribution to be unimodal about $c \in \mathbb{R}$ with median $x_{0.5}$.

- i. If the distribution is symmetric, $\theta_\tau(F) = \theta(F) = c$ for all $\tau > 0$;

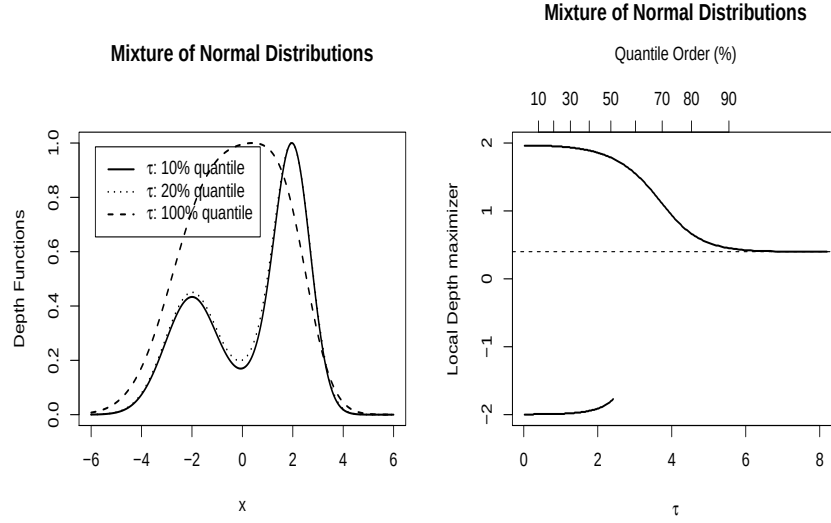


Fig. 1 Mixture of normal distributions. Left: simplicial global and local depth (functions are normalized with respect to maximum value). Right: local maxima versus τ parameter; horizontal line marks the median.

- ii. if the distribution is asymmetric, simplicial local depth admits a unique maximizer $\theta_\tau(F)$ and $c < \theta_\tau(F) < \min\{c + \tau, x_{0.5}\}$ or $\max\{c - \tau, x_{0.5}\} < \theta_\tau(F) < c$, according to whether the distribution is positive or negative asymmetric;
- iii. if $0 < \tau_1 < \tau_2$ and the distribution is asymmetric, $\theta_{\tau_1}(F) < \theta_{\tau_2}(F)$ or $\theta_{\tau_1}(F) > \theta_{\tau_2}(F)$ according to whether the distribution is positive or negative asymmetric.

Proof. Agostinelli and Romanazzi (2007).

According to Proposition 3, in the unimodal case the maximizer of simplicial local depth is intermediate between the mode and the median and it converges to the mode or to the median according to whether $\tau \rightarrow 0^+$ or $\tau \rightarrow \infty$.

Example 4. Let $X \sim 0.5N(\mu_1 = -2, \sigma_1 = 3/2) + 0.5N(\mu_2 = 2, \sigma_1 = 1)$. The simplicial depth function is increasing for $x < x_{0.5} = 0.4$ and decreasing for $x > x_{0.5}$. The local version, with τ equal to 10% quantile of interval lengths has two local maxima $\theta_\tau^{(1)} \simeq -1.994$ and $\theta_\tau^{(2)} \simeq 1.962$, near to the modes. The plot of the (normalized) depth functions is shown in the left panel of Figure 1. While the results do depend on the value of τ , both the number and the locations of the local maxima stay almost constant for a wide range of the smoothing parameter, up to 52% quantile order (see the right panel of Figure 1). Convergence to the maximum value of the global depth, i.e., the median of the distribution, is observable starting from 90% quantile order.

In the multivariate case, the behaviour of local depth is more complex and the details can be found in Agostinelli and Romanazzi (2008).

References

- Agostinelli C, Romanazzi M (2007) Local depth of univariate distributions. Submitted for publication
- Agostinelli C, Romanazzi M (2008) Multivariate local depth. Submitted for publication
- Bickel PJ, Lehmann EL (1979) Descriptive statistics for nonparametric models (IV. Spread). In: Jureckova J (ed) *Contributions to Statistics*, Academia, Prague, pp 33–40
- Chen Z (1995) Bounds for the breakdown point of the simplicial median. *Journal of Multivariate Analysis* 55:1–13
- Chen Z, Tyler DE (2002) The influence function and maximum bias of Tukey median. *The Annals of Statistics* 30:1737–1759
- Donoho D, Gasko M (1992) Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics* 20:1803–1827
- Efron B (1965) The convex hull of a random set of points. *Biometrika* 52:331–343
- Hampel F (1974) The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69:383–393
- Liu RY (1990) On a notion of data depth based on random simplices. *The Annals of Statistics* 18:405–414
- Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference. *The Annals of Statistics* 27:783–858
- Oja H (1983) Descriptive statistics for multivariate distributions. *Statistics and Probability Letters* 1:327–332
- Robbins HE (1944) On the measure of a random set. I. *Annals of Mathematical Statistics* 15:70–74
- Romanazzi M (2001) Influence function of halfspace depth. *Journal of Multivariate Analysis* 77:138–161
- Romanazzi M (2004) Data depth and correlation. *Allgemeines Statistisches Archiv* 88:191–214
- Romanazzi M (2008) A note on simplicial depth function. *Advances in Statistical Analysis* 92:235–253
- Serfling R (2002) Generalized quantile processes based on multivariate depth functions, with applications in nonparametric multivariate analysis. *Journal of Multivariate Analysis* 83:232–247
- Sommerville DMY (1958) *An introduction to the geometry of n dimensions*. Dover Publications, New York
- Wilks SS (1960) Multidimensional statistical scatter. In: Olkin I (ed) *Contributions to Probability and Statistics in Honor of H. Hotelling*, Stanford University Press, pp 486–503
- Zuo Y, Serfling R (2000) General notions of statistical depth function. *The Annals of Statistics* 28:461–482