On a New Class of Data Depths for Measuring Representativeness

Stephen M. S. Lee [1a]

[a]Department of Statistics and Actuarial Science, The University of Hong Kong

Abstract

Data depth provides a natural means to rank multivariate vectors with respect to an underlying multivariate distribution. The conventional notion of a depth function emphasizes a centre-outward ordering of data points. While useful for certain statistical applications, such emphasis has rendered most classical data depths insensitive to some distributional features, such as multimodality, of concern to other statistical applications. To get around the problem we introduce a new notion of data depth which seeks to rank data points according to their reepresentativeness, rather than centrality, with respect to an underlying distribution of interest. We propose a general device for defining such depth functions, based essentially on a choice of goodness-of-fit test statistic. Our device calls for a new interpretation of depth more akin to the concept of density than location. It copes particularly well with multivariate data exhibiting multimodality. In addition to providing depth values for individual data points, the new class of depth functions derived from goodness-of-fit tests also extends naturally to provide depth values for subsets of data points, a concept new to the data-depth literature. Applications of the new depth functions are demonstrated with both simulated and real data.

Keywords: centre-outward ordering, classification, data depth, goodness-of-fit tests, interpoint distance, multimodality, representativeness.

1. Introduction

Classical depth functions are formulated primarily to measure the "centrality" of a single point relative to a specified distribution function F or to a sample of observations $X_1, \ldots, X_n$ drawn from F. They provide a natural means to rank multivariate data with the deepest point being regarded as the "centre" of the distribution F. Insistence on the above centre-outward ordering property has, however, restricted to a certain extent the scope of depth-based inferences. For instance, when applying data depths to problems such as cluster analysis or classification, we often assume tacitly that a "deep" point with respect to a distribution must also be "representative" of that distribution, an assumption open to question if the distribution exhibits some nonstandard features such as multimodality. It is therefore desirable, at least for certain types of applications, to consider an alternative notion of depth as a measure of "representativeness", so that data points can be endowed with an ordering sufficiently responsive to the shape of the reference distribution. Recent years have seen significant moves in that direction, leading to new definitions of data depth such as the likelihood depth (Fraiman and Meloche, 1999), the kernelized spatial depth (Chen, Dang, Peng and Bart, 2009), the weighted halfspace depth (Hlubinka, Kotík and Vencálek, 2010), the local simplicial or halfspace depth (Agostinelli and Romanazzi, 2011) and a general method due to Paindaveine and van Bever (2012) for localising a global depth. Capitalising on a natural connection between goodness-of-fit tests and the extent to which a sample "represents" a null distribution F, we propose in this paper a new class of data depths useful for measuring representativeness of data points, or subsets of data points, with respect to multivariate distributions.

Supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 702508P)

[1]Corresponding author: Professor, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong. E-mail: smslee@hku.hk

2. Depth functions based on interpoint distances

2.1. General idea

Consider a random sample $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ drawn from a distribution F on the sample space $\mathcal{S}$. Let $T(\mathcal{X}_n, F)$ be a generic goodness-of-fit test statistic, large values of which indicate a lack of fit of the distribution F to the observed data $\mathcal{X}_n$, or in other words, $\mathcal{X}_n$ not sufficiently "representative" of the distribution F. This motivates our new formulation of data depth applicable to a subset of points, under which "depth" acquires a new meaning: representativeness.

For any collection of points $\{x_1, \ldots, x_n\} \subset \mathcal{S}$ and any distribution function F on $\mathcal{S}$, we define the depth of the pattern $\{x_1, \ldots, x_n\}$, with respect to F, to be

$$D(F, \{x_1, \ldots, x_n\}) = \eta(T(\{x_1, \ldots, x_n\}, F)|F),$$

for an arbitrary decreasing function $\eta(\cdot|F)$, possibly depending on F. As a canonical choice, we can set $\eta(t|F) = \mathbb{P}_F\left(T(\mathcal{X}_n, F) > t\right)$, in which case the depth function can also be interpreted as a p-value for testing whether the "sample" $\{x_1, \ldots, x_n\}$ has an underlying distribution F. Under our new formulation, a "deep" point pattern $\{x_1, \ldots, x_n\}$ relative to F can be viewed as a "sample" in no essential conflict with F, or one which is reasonably "representative" of F. The singleton case $n = 1$ reduces $D(F, \{x_1\})$ to a depth measure for a single point, as has traditionally been understood by a depth function. Without confusion we write $D(F, x) = D(F, \{x\})$ for the singleton case.

In what follows we apply the above construction to three types of goodness-of-fit tests derived from interpoint distances, calculated in terms of an arbitrary distance measure $\delta(\cdot, \cdot)$ defined on $\mathcal{S} \times \mathcal{S}$.

2.2. Nearest neighbours

Nearest neighbour tests (Schilling, 1986; Henze, 1988) are designed to handle general multivariate two-sample problems. Consider first testing whether the two samples $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and $\mathcal{Y}_m = \{Y_1, \ldots, Y_m\}$ come from the same distribution. Let $N_r(Z)$ be the rth nearest δ-neighbour in the combined sample $\mathcal{X}_n \cup \mathcal{Y}_m$ of the sample point $Z \in \mathcal{X}_n \cup \mathcal{Y}_m$. Then an unweighted version of the two-sample k-nearest neighbour test statistic can be written as

$$k^{-1}(m+n)^{-1}\left(\sum_{i=1}^{n}\sum_{r=1}^{k}\mathbf{1}\left\{N_r(X_i) \in \mathcal{X}_n\right\} + \sum_{j=1}^{m}\sum_{r=1}^{k}\mathbf{1}\left\{N_r(Y_j) \in \mathcal{Y}_m\right\}\right). \tag{1}$$

Define, for any fixed $x \in \mathcal{S}$ and any distribution F on $\mathcal{S}$, $\Lambda(t|x, F) = \mathbb{P}_F(\delta(x, X) \le t \mid X \sim F)$, $t \in \mathbb{R}$. By considering the limiting case where $m \to \infty$ and $k/m \to \gamma \in (0, 1)$, with $Y_1, Y_2, \ldots$ independently distributed under F, we derive from (1) a one-sample test statistic

$$T(\mathcal{X}_n, F) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\left\{\delta(X_i, X_j) \le \Lambda^{-1}(\gamma|X_i, F)\right\}. \tag{2}$$

Furthermore, by applying (2) to a point pair $\{x_1, x_2\}$ and letting $x_1, x_2 \to x$, we are led to a depth function of a single point x, given by $D(F, x) = \eta\left(\Lambda^{-1}(\gamma|x, F)\,\middle|\,F\right)$.

2.3. Energy tests

Aslan and Zech (2005) propose a two-sample energy test statistic which, on specialization to a one-sample problem, takes on the form

$$\begin{aligned} T(\mathcal{X}_n, F) \;=\; & \frac{1}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} R(\delta(X_i, X_j)) - \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{S}} R(\delta(X_i, y))\, dF(y) \\ & + \frac{1}{2}\int_{\mathcal{S}^2} R(\delta(y_1, y_2))\, d(F \otimes F)(y_1, y_2), \end{aligned} \tag{3}$$

for some decreasing energy function R defined on $[0, \infty)$. Thus, for a single $x \in \mathcal{S}$, (3) leads to the depth function

$$D(F, x) = \eta\left(\frac{1}{2} \int_{\mathcal{S}^2} R(\delta(y, z))\, d(F \otimes F)(y, z) - \int_{\mathcal{S}} R(\delta(x, y))\, dF(y) \,\middle|\, F \right). \tag{4}$$

In the case $\mathcal{S} = \mathbb{R}^d$, Fraiman and Meloche (1999) propose an affine invariant version of likelihood depth, which can be regarded as a special case of (4) if we set $\eta(t|F) = \frac{1}{2} \int_{\mathcal{S}^2} R(\delta(y, z))\, d(F \otimes F)(y, z) - t$, $R(t) = K(t/h)/h^d$ for some kernel function K and bandwidth h, and, with slight abuse of notation, $\delta(y, z) = (y - z)^{\mathrm{T}} \Sigma_F^{-1} (y - z)$, where Σ_F denotes the dispersion matrix of F.

2.4. Within-triplet distances

Based on within-triplet distances, Bartoszyński, Pearl and Lawrence (1997) propose a goodness-of-fit test statistic

$$T(\mathcal{X}_n, F) = \left[U_1^*,\ -U_1^* - U_3^*,\ U_3^* \right] A \left[U_1^*,\ -U_1^* - U_3^*,\ U_3^* \right]^{\mathrm{T}},$$

where, for Y_1, Y_2 iid from F, independent of $\mathcal{X}_n$,

$$U_1^* = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_F \left(\delta(Y_1, Y_2) < \min\{\delta(X_i, Y_1), \delta(X_i, Y_2)\} \mid X_i \right) - 1/3,$$

$$U_3^* = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_F \left(\delta(Y_1, Y_2) > \max\{\delta(X_i, Y_1), \delta(X_i, Y_2)\} \mid X_i \right) - 1/3,$$

and A is a positive semidefinite matrix chosen to render the test powerful against specific alternatives. Taking for simplicity a diagonal A and setting $n = 1$ in the above formulation, we obtain a depth function for a single point given by $D(F, x) = \eta\left(\sum_{j=1}^{3} w_j \nu_j(F, x)^2 \,\middle|\, F \right)$, where

$$\nu_1(F, x) = \mathbb{P}_F \left(\delta(Y_1, Y_2) < \min\{\delta(x, Y_1), \delta(x, Y_2)\} \right) - 1/3,$$

$$\nu_3(F, x) = \mathbb{P}_F \left(\delta(Y_1, Y_2) > \max\{\delta(x, Y_1), \delta(x, Y_2)\} \right) - 1/3,$$

and $\nu_2(F, x) = -\nu_1(F, x) - \nu_3(F, x)$, for some weights $w_1, w_2, w_3 \geq 0$.

2.5. Consistency of sample depth function

In applications where F is unavailable, we may consider using a sample depth function, that is a depth function calculated with respect to the empirical distribution $F_{\mathcal{Y}_m}$ of a random sample $\mathcal{Y}_m = (Y_1, \ldots, Y_m)$ drawn from F. We comment briefly on the conditions sufficient for consistency of sample depth functions of the interpoint-distance type, in the sense that $D(F_{\mathcal{Y}_m}, \{x_1, \ldots, x_n\})$ converges in probability to $D(F, \{x_1, \ldots, x_n\})$ as $m \to \infty$. In each case we assume η to be a continuous function.

Consistency of sample depth functions based on nearest neighbours follows from strong consistency of the sample γth quantile of the m distances $\delta(x, Y_1), \ldots, \delta(x, Y_m)$, which converges in probability to $\Lambda^{-1}(\gamma|x, F)$ for any $x \in \mathcal{S}$.

For consistency of sample depth functions based on energy tests, we may invoke the weak law of large numbers for U-statistics to show that, for any $x \in \mathcal{S}$,

$$m^{-1} \sum_{j=1}^{m} R(\delta(x, Y_j)) \to \int_{\mathcal{S}} R(\delta(x, y))\, dF(y) \ \text{ in probability}$$

and

$$m^{-2} \sum_{i=1}^{m} \sum_{j=1}^{m} R(\delta(Y_i, Y_j)) \to \int_{\mathcal{S}^2} R(\delta(y, z))\, d(F \otimes F)(y, z) \ \text{ in probability,}$$

provided that the limits exist. Similarly we can show that

$$m^{-2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{1}\left\{\delta(Y_i,Y_j)<\min\{\delta(x,Y_i),\delta(x,Y_j)\}\right\}-1/3\to\nu_1(F,x)\ \text{ in probability}$$

and so does $m^{-2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{1}\left\{\delta(Y_i,Y_j)>\max\{\delta(x,Y_i),\delta(x,Y_j)\}\right\}-1/3$ to $\nu_3(F,x)$, leading to consistency of the sample depth function based on within-triplet distances.

2.6. Choice of control parameters

A practical issue arising from depth-based applications concerns the setting of control parameters, that is γ for the nearest neighbour depth, h for the likelihood depth, and the weights (w_1,w_2,w_3) for the depth based on within-triplet distances. As a general approach we propose to set the control parameters for the sample depth $D(F_{\mathbf{y}_m},\cdot)$ by maximising a "correlation" measure

$$\frac{m^{-1}\sum_{i=1}^{m}D(F_{\mathbf{y}_m},Y_i)\varpi(Y_i)-m^{-1}\sum_{i=1}^{m}\varpi(Y_i)\int_{\mathcal{S}}D(F_{\mathbf{y}_m},x)\varpi(x)\,dx}{\sqrt{\int_{\mathcal{S}}D(F_{\mathbf{y}_m},x)^2\varpi(x)\,dx-\left(\int_{\mathcal{S}}D(F_{\mathbf{y}_m},x)\varpi(x)\,dx\right)^2}},\tag{5}$$

where ϖ denotes an arbitrary density function supported on $\mathcal{S}$. Note that if F is differentiable, then (5) may be viewed as an estimate of $\sqrt{\mathrm{Var}(F'(W))}\,\mathrm{Corr}(D(F,W),F'(W))$, for $W\sim\varpi$.

3. Application to economic data

To illustrate the practical relevance of a shift in emphasis from "centrality" to "representativeness", we compare the depth function based on within-triplet distances with the classical simplicial depth, calculated with respect to a set of bivariate observations on the life expectancy at birth and the gross national income (GNI) per capita of 162 countries for the year 2008. For the case of within-triplet distances, we set $(w_1,w_2,w_3)=(0.09,0.66,0.25)$, which maximises (5) over the simplex $\{(w_1,w_2,1-w_1-w_2):w_1,w_2\geq 0,w_1+w_2\leq 1\}$, with ϖ taken to be the uniform density function over the rectangle $[0,65]\times[40,90]$. Figure 1 displays the two depth functions calculated with respect to the data. The data points observed for the 162 countries, shown also in Figure 1, cluster in a crescent and do not exhibit clear unimodality. The simplicial depth identifies a

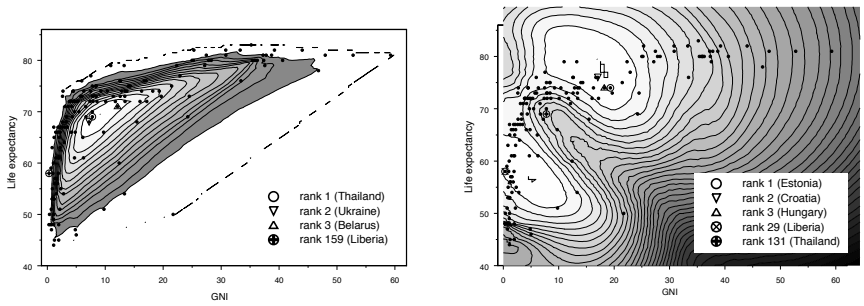

Figure 1: World Bank data — contour plots of simplicial depth (left) and within-triplet-distance depth (right), with respect to life expectancy and GNI indicators of 162 countries in 2008.

unique deep centre, near which can be located the three most "central" countries, namely Thailand, Ukraine and Belarus. However, their somewhat peripheral positions relative to the main data crescent casts doubt on their representativeness. On the other hand, the above three countries are ranked rather low (Thailand 131;

Ukraine 124; Belarus 111) based on within-triplet distances, and are surrounded by deep areas more representative of the entire dataset. According to within-triplet distances, the three deepest countries are Estonia, Croatia and Hungary, all of which lie on one side of the central dip.

We highlight on Figure 1 Thailand and Liberia, the two countries which register the most positive and negative differences between their two ranks obtained from the two depth functions. Thailand is ranked the deepest by simplicial depth but only 131st by within-triplet distances. Liberia, on the contrary, is ranked 29th by within-triplet distances but 159th by simplicial depth.

4. Application to supervised classification

Suppose that we have available two labelled training samples, $\mathcal{Y}^{[1]} = (Y_1^{[1]}, \ldots, Y_{n_1}^{[1]})$ and $\mathcal{Y}^{[2]} = (Y_1^{[2]}, \ldots, Y_{n_2}^{[2]})$, drawn respectively from two distinct distributions F_1 and F_2. We are interested in classifying a new set of data points $\{x_1, \ldots, x_n\}$, which are known to come from the same distribution, to one of the two distributions. Our extended notion of depth function $D(F, \{x_1, \ldots, x_n\})$ provides a natural procedure for classification. Denote by $F_{\mathcal{Y}^{[j]}}$ the empirical distribution of $\mathcal{Y}^{[j]}$, $j = 1, 2$. For each $j = 1, 2$ and $i = (i_1, \ldots, i_n) \in \{1, \ldots, n_j\}^n$, calculate the depth values $d_i^{[j]} = D(F_{\mathcal{Y}^{[j]}}, \{Y_{i_1}^{[j]}, \ldots, Y_{i_n}^{[j]}\})$. Then we classify $\{x_1, \ldots, x_n\}$ as coming from F_1 if

$$n_1^{-n}\mathrm{card}\left(\left\{i : d_i^{[1]} \le D(F_{\mathcal{Y}^{[1]}}, \{x_1, \ldots, x_n\})\right\}\right) > n_2^{-n}\mathrm{card}\left(\left\{i : d_i^{[2]} \le D(F_{\mathcal{Y}^{[2]}}, \{x_1, \ldots, x_n\})\right\}\right), \tag{6}$$

and from F_2 if the above inequality is reversed. The classification is inconclusive if the two sides of (6) are equal. Instead of maximising (5), we consider it more natural in the present context to set the control parameters for the depth functions by leave-n-out cross validation.

In the following numerical example we set $n_1 = n_2 = 50$ and consider the case of classifying a point pair, that is $n = 2$. We take F_1 to be the univariate normal mixture $0.2\,N(-2.5, 4) + 0.8\,N(2.5, 1)$ and F_2 to be $N(0, 3.24)$. The Bayes misclassification rate based on the uniform prior is found to be 0.1814, which can be obtained by evaluating the integral $2^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \min\{f_1(x)f_1(y), f_2(x)f_2(y)\}\,dx\,dy$, where f_j denotes the density of F_j, $j = 1, 2$. The energy test and the within-triplet distance test provide useful expressions for calculating the depth $D(F_{\mathcal{Y}^{[j]}}, \{x_1, x_2\})$ of a point-pair (x_1, x_2). A simulation study is carried out on the misclassification rates of these two depths. We consider for each depth function seven control parameter settings, under each of which the misclassification rate is estimated by leave-two-out cross validation and marked by the letter "**V**" in Figure 2. Minimising over the seven parameter settings for each depth function, the cross-validated

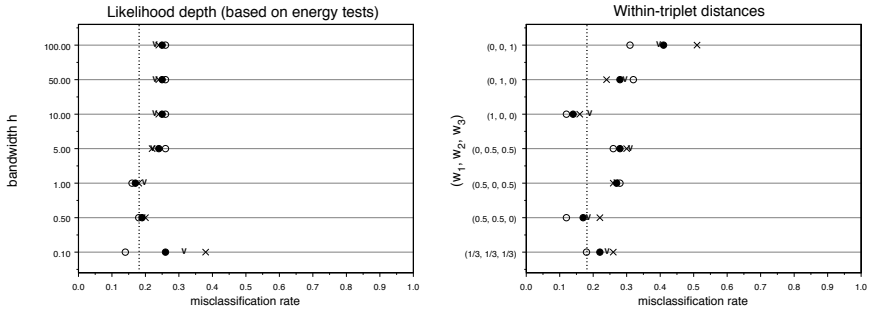

Figure 2: Misclassification rates: (i) Bayes rate (dotted vertical line); (ii) leave-two-out cross-validated estimates based on training data ("V"); (iii) F_1 misclassified as F_2 ("○") based on test sample of 50 point-pairs from F_1; (iv) F_2 misclassified as F_1 ("×") based on test sample of 50 point-pairs from F_2; (v) average of (iii) and (iv) ("●").

choices of the parameters are found to be $h = 0.5$ for the likelihood depth and $(w_1, w_2, w_3) = (0.5, 0.5, 0)$ for the depth based on within-triplet distances. Next we generate a test sample of 50 random point-pairs from

each of the two distributions and estimate the misclassification rates of the two depth functions, shown also in Figure 2. For both depth functions, at least one of the control parameter settings yields misclassification rates close to, or even lower than, the Bayes rate 0.1814. Cross validation is very effective in identifying the optimal, or nearly optimal, choice of the control parameter for each depth function.

References

Agostinelli, C. and Romanazzi, M. (2011). Local depth, *Journal of Statistical Planning and Inference*, **141**, 817–830.

Aslan, B. and Zech, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy, *Journal of Statistical Computation and Simulation*, **75**, 109–119.

Bartoszyński, R., Pearl, D. K. and Lawrence J. (1997). A multidimensional goodness-of-fit test based on interpoint distances, *Journal of the American Statistical Association*, **92**, 577–586.

Chen, Y., Dang, X., Peng, H. and Bart, H. L. J. (2009). Outlier detection with the kernelized spatial depth function, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 288–305.

Fraiman, R. and Meloche, J. (1999). Multivariate L-estimation (with discussion), *Test*, **8**, 255–317.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences, *The Annals of Statistics*, **16**, 772–783.

Hlubinka, D., Kotík, L. and Vencálek, O. (2010). Weighted halfspace depth, *Kybernetika*, **46**, 125–148.

Paindaveine, D. and van Bever, G. (2012). From depth to local depth: a focus on centrality, Working Papers ECARES 2012-047, ULB – Universite Libre de Bruxelles.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors, *Journal of the American Statistical Association*, **81**, 799–806.