Statistical Depth Functions

In the multivariate and functional setting

Satvik Saha

Supervised by
Dr. Anirvan Chakraborty

Master's Thesis

Department of Mathematics and Statistics
Indian Institute of Science Education and Research, Kolkata

March 2024

Abstract

Dedication

Declaration

Acknowledgments

Contents

Chapter 1

Introduction

1.1 Centrality vs Density

1.2 Nonparametric procedures

Chapter 2

Multivariate Data

It is desirable for a depth function $D\colon \mathbb{R}^d \times \mathscr{F} \to \mathbb{R}$ to satisfy the following properties, described by Zuo and Serfling (2000).

P1. *Affine invariance.* For any random vector $\boldsymbol{X}$ in $\mathbb{R}^d$, any $d \times d$ nonsingular matrix A, and any d-vector $\boldsymbol{b}$,

$$D(A\boldsymbol{x} + \boldsymbol{b}, F_{A\boldsymbol{X}+\boldsymbol{b}}) \;=\; D(\boldsymbol{x}, F_{\boldsymbol{X}}). \tag{2.0.1}$$

This makes $D(\boldsymbol{X}, F_{\boldsymbol{X}})$ independent of the choice of coordinate system.

P2. *Maximality at center.* For any $F \in \mathscr{F}$ having 'center' $\boldsymbol{\theta}$,

$$D(\boldsymbol{\theta}, F) \;=\; \sup_{\boldsymbol{x}\in\mathbb{R}^d} D(\boldsymbol{x}, F). \tag{2.0.2}$$

This means that the deepest point coincides with some center of symmetry of the distribution F.

P3. *Monotonicity relative to deepest point.* For any $F \in \mathscr{F}$ having deepest point $\boldsymbol{\theta}$ and for $\alpha \in [0, 1]$,

$$D(\boldsymbol{x}, F) \;\leq\; D(\boldsymbol{\theta} + \alpha(\boldsymbol{x} - \boldsymbol{\theta}), F). \tag{2.0.3}$$

Thus, $D(\,\cdot\,, F)$ monotonically decreases along any ray pointing away from the deepest point.

P4. *Vanishing at infinity.* For any $F \in \mathscr{F}$,

$$D(\boldsymbol{x}, F) \to 0 \quad \text{as} \quad \|\boldsymbol{x}\| \to \infty. \tag{2.0.4}$$

Furthermore, we demand that D be non-negative and bounded. Thus, we may assume hereon that D only takes values in $[0, 1]$.

The notion of a 'center' of a distribution in **P2** is typically described in terms of symmetry. We say that a random vector $\boldsymbol{X}$ is *centrally symmetric* about $\boldsymbol{\theta} \in \mathbb{R}^d$

if $\boldsymbol{X} - \boldsymbol{\theta} \overset{d}{=} \boldsymbol{\theta} - \boldsymbol{X}$. Similarly, we say that $\boldsymbol{X}$ is *angularly symmetric* about $\boldsymbol{\theta}$ if $(\boldsymbol{X} - \boldsymbol{\theta})/\|\boldsymbol{X} - \boldsymbol{\theta}\|$ is centrally symmetric about $\mathbf{0}$. An even more restrictive notion of symmetry is *spherical symmetry*, where we demand that $U(\boldsymbol{X} - \boldsymbol{\theta}) \overset{d}{=} \boldsymbol{X} - \boldsymbol{\theta}$ for every orthonormal matrix U. *Elliptical symmetry* requires that $V\boldsymbol{X}$ is spherically symmetric about $\boldsymbol{\theta}$ for some nonsingular matrix V. Finally, the weakest notions of symmetry discussed here is *halfspace symmetry*, where we impose $P(\boldsymbol{X} \in H) \geq 1/2$ for every closed halfspace in $\mathbb{R}^d$ containing $\boldsymbol{\theta}$. Thus, the symmetries in decreasing order of strength are $S > E > C > A > H$.

2.1 Depth contours

The following definitions are adapted from Liu et al., 1999.

Definition 2.1.1. The contour of depth t is the set $\{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}, F) = t\}$.

Definition 2.1.2. The region enclosed by the contour of depth t is the set

$$R_F(t) = \{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}, F) > t\}. \tag{2.1.1}$$

Definition 2.1.3. The p-th central region is the set

$$C_F(p) = \bigcap_t \{R_F(t) : P_F(R_F(t)) \geq p\}. \tag{2.1.2}$$

Definition 2.1.4. The p-th level contour, or center-outward contour surface, is the set $Q_F(p) = \partial C_F(p)$.

Example 2.1.5. Consider $\mathcal{U}(B^d)$, i.e. the uniform distribution on the unit ball in $\mathbb{R}^d$. While there are no proper density contours to speak of, halfspace depth contours are concentric spheres centered at the origin, the deepest point. This illustrates how depth contours are more suited to indicating centrality than density contours.

2.2 Depth-Depth plots

Definition 2.2.1 (DD plot)**.** Let F, G be two distributions on $\mathbb{R}^d$, and let D be a depth function. The Depth-Depth plot, also known as the DD plot, of F and G is given by

$$\mathrm{DD}(F, G) = \{(D(\boldsymbol{z}, F), D(\boldsymbol{z}, G)) : \boldsymbol{z} \in \mathbb{R}^d\}. \tag{2.2.1}$$

Remark. The above definition generalizes naturally to involve more than two distributions on $\mathbb{R}^d$.

When the depth function D only takes values in $[0, 1]$, the DD plot is a subset of $[0, 1]^2$ and hence easily visualized. Clearly when $F = G$, the corresponding DD

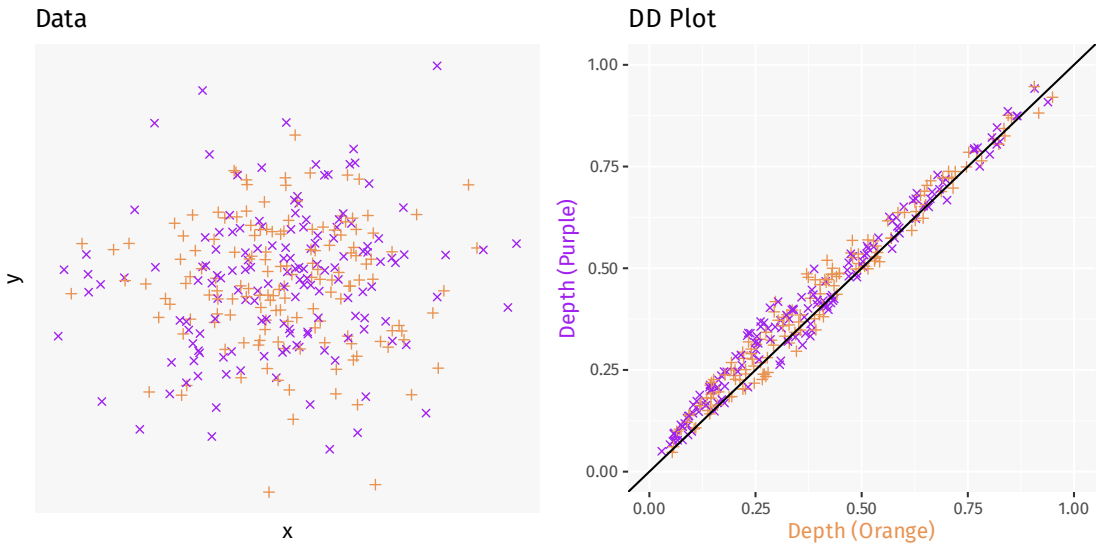


Figure 2.1: Empirical DD plot using spatial depth, where both underlying distributions (bivariate normal) are identical. Observe how the points in the DD plot stay close to the diagonal black line.

plot is confined to the diagonal $\{(t,t) : t \in [0,1]\}$. However, when $d \geq 2$ and F, G are absolutely continuous, $\mathrm{DD}(F,G)$ has non-zero area (Lebesgue measure) when $F \neq G$. Assuming that D is affine invariant, Liu et al. (1999) propose this area as an affine invariant measure of the discrepancy between F and G.

If the distributions F, G are unknown, we may use data samples $\mathscr{D}_F = \{\boldsymbol{X}_i\}$ and $\mathscr{D}_G = \{\boldsymbol{Y}_j\}$ where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \overset{\text{iid}}{\sim} G$, then construct empirical distributions $\hat{F}_n, \hat{G}_m$. With this, we may examine the empirical DD plot

$$\mathrm{DD}(\hat{F}_n, \hat{G}_n) \;=\; \{(D(\boldsymbol{z}, \hat{F}_n), D(\boldsymbol{z}, \hat{G}_m)) : \boldsymbol{z} \in \mathscr{D}_F \cup \mathscr{D}_G\}. \tag{2.2.2}$$

DD plots can be used as a diagnostic tool to detect differences in location and scale between two multivariate distributions.

1. If $F = G$, the points in $\mathrm{DD}(\hat{F}_n, \hat{G}_m)$ stay close to the diagonal. See Figure 2.1.
2. If the same point $\boldsymbol{z}_0$ achieves maximum depths with respect to both distributions F and G, this indicates that $\boldsymbol{z}_0$ is their common center. See Figure 2.2.
3. Suppose that F and G have the same center. If the points in $\mathrm{DD}(\hat{F}_n, \hat{G}_m)$ arch above the diagonal, i.e. the bulk of points are deeper in G than in F, this indicates that F has a greater spread than G. See Figure 2.3a.

Liu et al. (1999) also demonstrate the use of DD plots to detect differences in skewness and kurtosis. This tool is especially convenient since the DD plot is always two dimensional regardless of the dimension d of the sample points.

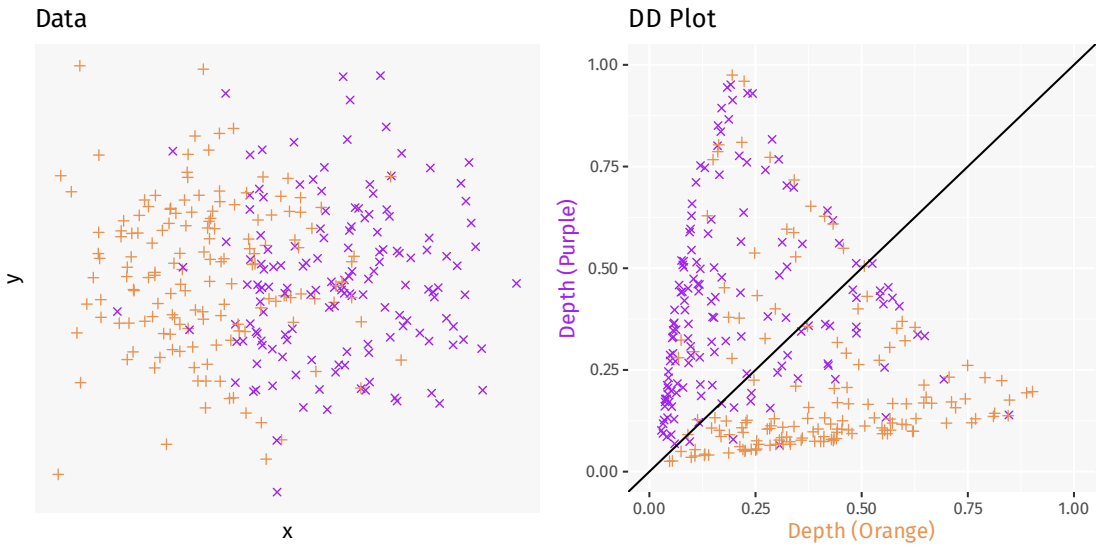


Figure 2.2: Empirical DD plot using spatial depth, where both underlying distributions (bivariate normal) differ only in location. Observe how most of the orange points fall in the lower triangle, while the purple ones fall in the upper triangle. The deepest point with respect to the orange distribution has fairly low depth with respect to the purple one, and vice versa.

2.3 Testing

We are mainly interested in the two sample homogeneity test. Given samples from F and G, we wish to test the null hypothesis $H_0 : F = G$ against an alternate hypothesis that F and G differ in location or scale.

When F, G are distributions on $\mathbb{R}$, rank based tests such as the Wilcoxon rank-sum test or the Siegel-Tukey test are readily available. A very useful tool in this setting is the probability integral transform.

Proposition 2.3.1. *Let $\boldsymbol{X} \sim F$, and let the distribution F be continuous. Then, $F(\boldsymbol{X}) \sim \mathcal{U}[0, 1]$.*

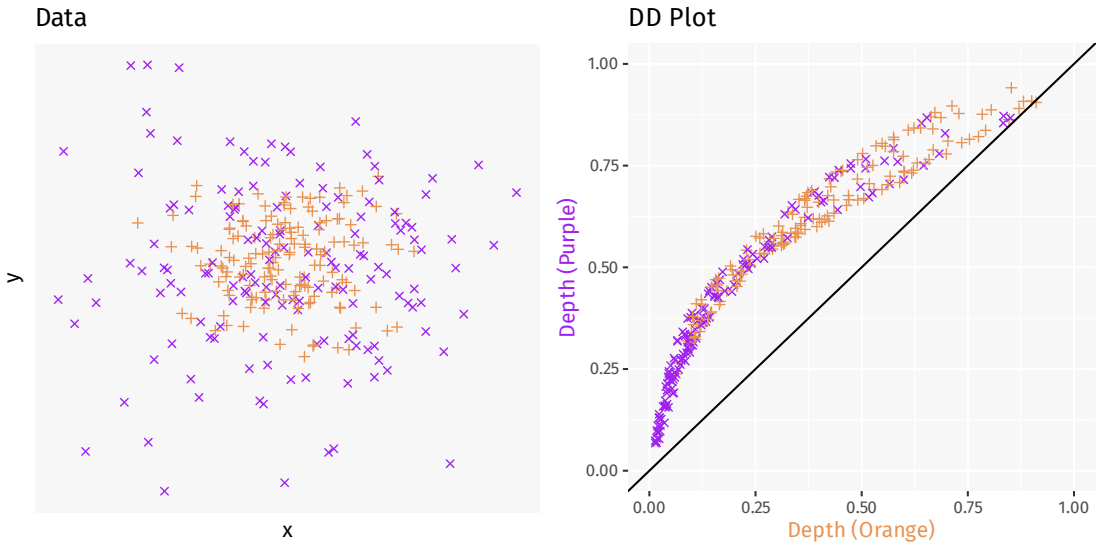
Since $F(\boldsymbol{X}_j)$ has the same rank within $\{F(\boldsymbol{X}_i)\}$ as does $\boldsymbol{X}_j$ within $\{\boldsymbol{X}_i\}$, the above result is the key towards establishing many distribution-free tests and procedures.

In the multivariate setting, Liu and Singh (1993) use the following depth based analogue.

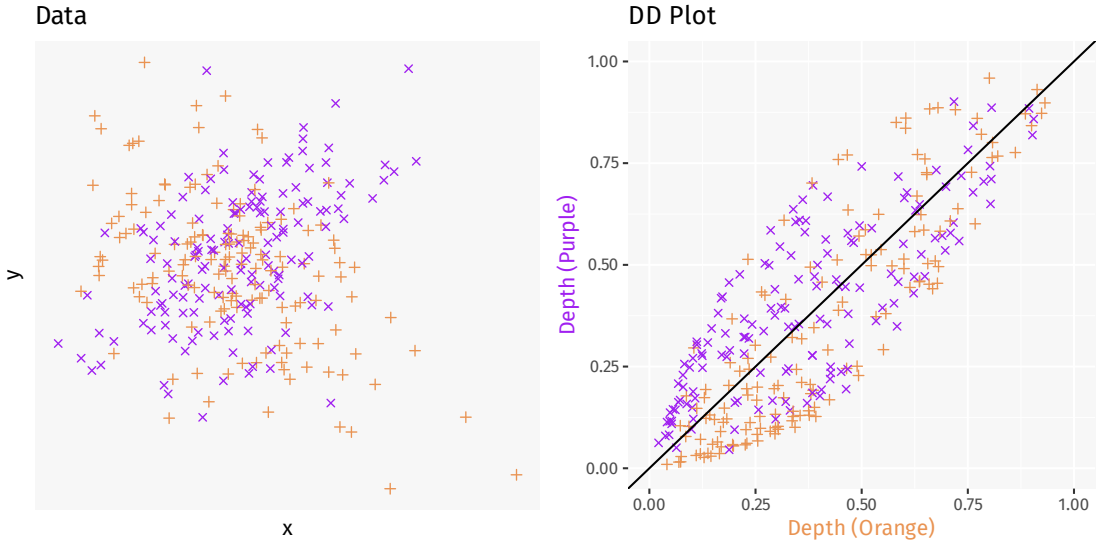
Definition 2.3.2. Denote

$$R(\boldsymbol{z}, F) = P(D(\boldsymbol{X}, F) \leq D(\boldsymbol{z}, F) \mid \boldsymbol{X} \sim F). \tag{2.3.1}$$

Note that in the empirical setting, $R(\boldsymbol{z}, \hat{F}_n)$ is simply the proportion of sample points $\{\boldsymbol{X}_i\}$ which are deeper in F than $\boldsymbol{z}$.



(a) $\Sigma_+ = \mathbb{I}_2, \quad \Sigma_\times = 4\mathbb{I}_2,$



(b) $\Sigma_+ = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \quad \Sigma_\times = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$

Figure 2.3: Empirical DD plot using spatial depth, where both underlying distributions (bivariate normal) differ only in scale. In **(a)**, observe how the points remain in the upper triangle in the DD plot. In **(b)**, observe how there are more orange points in the lower triangle, and more purple points in the upper triangle in the DD plot, especially in the region close to the origin.

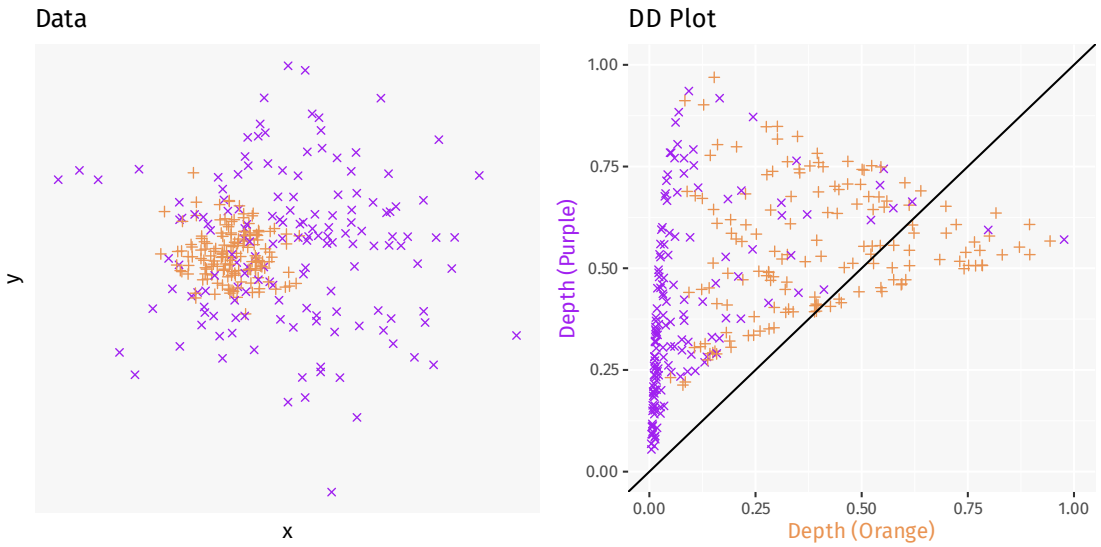


Figure 2.4: Empirical DD plot using spatial depth, where both underlying distributions (bivariate normal) differ in both location and scale. Observe that there is a clear separation between the orange and purple points in the DD plot, although not about the diagonal line.

Proposition 2.3.3 (Liu and Singh, 1993)**.** *Let $\boldsymbol{X} \sim F$, and let the distribution of $D(\boldsymbol{X}, F)$ be continuous. Then, $R(\boldsymbol{X}, F) \sim \mathcal{U}[0,1]$.*

Definition 2.3.4. Denote the quality index

$$Q(F,G) = P(D(\boldsymbol{X},F) \leq D(\boldsymbol{Y},F) \mid \boldsymbol{X} \sim F,\, \boldsymbol{Y} \sim G). \tag{2.3.2}$$

Note that $Q(F,G)$ and $Q(G,F)$ are not necessarily the same. We may also write

$$Q(F,G) = \mathbb{E}_{\boldsymbol{Y}\sim G}[R(\boldsymbol{Y},F)]. \tag{2.3.3}$$

It is clear that $Q(F,G) = 1/2$ when $F = G$. It can be shown under special circumstances that $Q(F,G) < 1/2$ if F, G differ in terms of location or scale. This will form the basis of our testing scheme, with $H_0 : F = G$ versus $H_A : Q(F,G) < 1/2$.

Here, we restrict our attention to elliptical distributions on $\mathbb{R}^d$.

Definition 2.3.5 (Elliptical distributions)**.** We say that a distribution is elliptical if it has a density of the form

$$f(\boldsymbol{x}) = c\,|\Sigma|^{-1/2}\, h\left((\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{2.3.4}$$

for some non-increasing function h. This is denoted by $\mathrm{Ell}(h; \boldsymbol{\mu}, \Sigma)$.

The quality index obeys the following properties.

Proposition 2.3.6 (Liu and Singh, 1993)**.** *Let* $F \sim \text{Ell}(h; \boldsymbol{\mu}_1, \Sigma_1)$ *and* $G \sim \text{Ell}(h; \boldsymbol{\mu}_2, \Sigma_2)$ *where* $\Sigma_1 - \Sigma_2$ *is positive definite. Further suppose that* $D(\cdot, F)$ *has the affine invariance and monotonicity properties. Then,* $Q(F, G) \leq 1/2$ *decreases monotonically as* $\boldsymbol{\mu}_2$ *is moved away from* $\boldsymbol{\mu}_1$ *along any line.*

Proposition 2.3.7 (Liu and Singh, 1993)**.** *Let* $F \sim \text{Ell}(h; \boldsymbol{\mu}, \Sigma_1)$ *and* $G \sim \text{Ell}(h; \boldsymbol{\mu}, \Sigma_2)$ *where* $\Sigma_1 - \Sigma_2$ *is positive definite. Consider Huber's contamination of the form*

$$G_\alpha = (1 - \alpha)F + \alpha G \tag{2.3.5}$$

where $0 \leq \alpha \leq 1$*. Then,* $Q(F, G_\alpha)$ *decreases monotonically as* α *increases.*

This motivates a modified Wilcoxon rank-sum test in the multivariate setting, using the quality index $Q(F, G)$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$, and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \overset{\text{iid}}{\sim} G$. Since $R(\cdot, F), Q(F, \cdot)$ depend on $D(\cdot, F)$, the latter has to be approximated using $D(\cdot, \hat{F}_{n_0})$, where $\hat{F}_{n_0}$ is based on a (fairly large) additional sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n_0} \overset{\text{iid}}{\sim} F$, with $n_0 \gg n, m$. With this, we compute

$$R(\cdot, \hat{F}_{n_0}) = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{1}(D(\boldsymbol{Z}_i, \hat{F}_{n_0}) \leq D(\cdot, \hat{F}_{n_0})). \tag{2.3.6}$$

Assign ranks $1, \ldots, n+m$ to the arranged values $R(\boldsymbol{X}_i, \hat{F}_{n_0}), R(\boldsymbol{Y}_j, \hat{F}_{n_0})$ (ascending order), and define W to be the sum of ranks of the $R(\boldsymbol{Y}_j, \hat{F}_{n_0})$. If necessary, break ties at random. Under the null hypothesis $F = G$, it is clear that W has the same distribution as the sum of m numbers drawn without replacement from $\{1, \ldots, n + m\}$. Under the alternate hypothesis $Q(F, G) < 1/2$, the ranks of $R(\boldsymbol{Y}_j, \hat{F}_{n_0})$ will tend to be lower on average, making W smaller.

Theorem 2.3.8 (Liu and Singh, 1993)**.** *Let* $H_{n,m}$ *be the distribution of the sum of* m *numbers drawn randomly without replacement from* $\{1, \ldots, n + m\}$*. Suppose that* F *admits a density function* f*. Under the null hypothesis* $F = G$*, we have* $W \sim H_{n,m}$*.*

It is also possible to approximate $Q(F, G)$ more directly via $Q(\hat{F}_n, \hat{G}_m)$ and perform our test this way. This sidesteps the need for the 'reference' sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n_0} \overset{\text{iid}}{\sim} F$. Note that

$$Q(\hat{F}_n, \hat{G}_m) = \frac{1}{m} \sum_{j=1}^{m} R(\boldsymbol{Y}_j, \hat{F}_n) = \frac{1}{nm} \sum_{i,j} \mathbf{1}(D(\boldsymbol{X}_i, \hat{F}_n) \leq D(\boldsymbol{Y}_j, \hat{F}_n)). \tag{2.3.7}$$

This estimate is indeed consistent under mild assumptions.

Theorem 2.3.9 (Liu and Singh, 1993)**.** *Suppose that the distribution of* $D(\boldsymbol{Y}, F)$ *is continuous where* $\boldsymbol{Y} \sim G$*, and that*

$$\sup_{\boldsymbol{z} \in \mathbb{R}^d} |D(\boldsymbol{z}, \hat{F}_n) - D(\boldsymbol{z}, F)| \xrightarrow{a.s.} 0. \tag{2.3.8}$$

Then, $Q(\hat{F}_n, \hat{G}_n) \xrightarrow{a.s.} Q(F, G)$ *as* $\min\{n, m\} \to \infty$*.*

This allows us to determine the asymptotic null distribution of $Q(\hat{F}_n, \hat{G}_m)$.

Theorem 2.3.10 (Liu and Singh, 1993)**.** *Let F be absolutely continuous, such that $\mathbb{E}_{\boldsymbol{X}\sim F} \|\boldsymbol{X}\|^4 < \infty$. Using Mahalanobis depth to define Q, we have*

$$S(\hat{F}_n, \hat{G}_m) = \left[\frac{1}{12}\left(\frac{1}{n} + \frac{1}{m}\right)\right]^{-1/2}\left[Q(\hat{F}_n, \hat{G}_m) - \frac{1}{2}\right] \xrightarrow{d} \mathcal{N}(0,1) \tag{2.3.9}$$

as $\min\{n, m\} \to \infty$, under the null hypothesis $F = G$.

Observe that given two samples, we have a choice between using $Q(\hat{F}_n, \hat{G}_m)$ or $Q(\hat{G}_m, \hat{F}_n)$. Shi et al. (2023) propose a weighted combination of the form

$$W^{\alpha}_{n,m} = \alpha S(\hat{F}_n, \hat{G}_m)^2 + (1-\alpha) S(\hat{G}_m, \hat{F}_n)^2 \tag{2.3.10}$$

for $\alpha \in [0, 1]$, or a maximum

$$M_{n,m} = \max\{S(\hat{F}_n, \hat{G}_m)^2, S(\hat{G}_m, \hat{F}_n)^2\}. \tag{2.3.11}$$

Under similar assumptions, they show that both $W^{\alpha}_{n,m} \xrightarrow{d} \chi^2_1$ and $M_{n,m} \xrightarrow{d} \chi^2_1$ as $\min\{n, m\} \to \infty$ and n/m converges to a positive constant, under the null hypothesis $F = G$.

2.4 Classification

The k-class classification task involves assigning an observation $\boldsymbol{x}$ to one of k populations, described by distributions F_i for $1 \le i \le k$. The populations may also be associated with prior probabilities π_i.

Definition 2.4.1 (Classifier)**.** A classifier is a map $\hat{\imath}\colon \mathbb{R}^d \to \{1, \dots, k\}$.

Example 2.4.2 (Bayes classifier)**.** Suppose that the population densities f_i for each $1 \le i \le k$ are known. The Bayes classifier assigns $\boldsymbol{x}$ to the $\hat{\imath}_B$-th population where

$$\hat{\imath}_B(\boldsymbol{x}) \;=\; \operatorname*{arg\,max}_{1\le i\le k} \; \pi_i f_i(\boldsymbol{x}). \tag{2.4.1}$$

One way of measuring the performance of a classifier (given the population distributions and their priors) is by measuring its average misclassification rate.

Definition 2.4.3 (Average misclassification rate)**.** The average misclassification rate of a classifier $\hat{\imath}$ is given by

$$\Delta(\hat{\imath}) \;=\; \sum_{i=1}^{k} \pi_i P(\hat{\imath}(\boldsymbol{X}) \neq i \mid \boldsymbol{X} \sim F_i). \tag{2.4.2}$$

Proposition 2.4.4. *The Bayes classifier has the lowest possible average misclassification rate. This is known as the optimal Bayes risk, denoted Δ_B.*

The simplest depth based classifier is the maximum depth classifier (Ghosh & Chaudhuri, 2005).

Example 2.4.5 (Maximum depth classifier)**.** Suppose that the prior probabilities π_i are equal. The maximum depth classifier $\hat{\imath}_D$ for a choice of depth function D is described by

$$\hat{\imath}_D(\boldsymbol{x}) \;=\; \underset{1\le i\le k}{\arg\max}\; D(\boldsymbol{x}, F_i). \tag{2.4.3}$$

In practice, instead of having direct access to the population distributions F_i, we have typically deal with labeled training data

$$\mathscr{D} \;=\; \{(\boldsymbol{x}_{ij}, i)\} \subset \mathbb{R}^d \times \{1, \dots, k\}, \tag{2.4.4}$$

where $\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i} \overset{\text{iid}}{\sim} F_i$ for each $1 \le i \le k$. The empirical maximum depth classifier simply replaces the population distributions F_i with their empirical counterparts $\hat{F}_i$ determined by $\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i}$. Thus, it is given by

$$\hat{\imath}_D(\boldsymbol{x}) \;=\; \underset{1\le i\le k}{\arg\max}\; D(\boldsymbol{x}, \hat{F}_i). \tag{2.4.5}$$

Under certain restrictions, this classifier becomes asymptotically optimal in the following sense.

Theorem 2.4.6 (Ghosh and Chaudhuri, 2005)**.** *Suppose that the population density functions f_i are elliptically symmetric, with $f_i(\boldsymbol{x}) = g(\boldsymbol{x} - \boldsymbol{\mu}_i)$ for parameters $\boldsymbol{\mu}_i$ and a density function g such that $g(k\boldsymbol{x}) \le g(\boldsymbol{x})$ for every $\boldsymbol{x}$ and $k > 1$. Further suppose that the priors on the populations are equal, and the depth function D is one of HD, SD, MJD, PD. Then, $\Delta(\hat{\imath}_D) \to \Delta_B$ as $\min\{n_1, \dots, n_k\} \to \infty$.*

Note that this result deals with elliptic population densities differing only in location. Relax this assumption, and instead suppose that $f_i \sim \mathrm{Ell}(h_i; \boldsymbol{\mu}_i, \Sigma)$, i.e.

$$f_i(\boldsymbol{x}) = c_i |\Sigma|^{-1/2} h_i\left((\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right) \tag{2.4.6}$$

for strictly decreasing h_i, and that the depths can be expressed as $D(\,\cdot\,, F_i) = l_i(f_i(\,\cdot\,))$ for strictly increasing functions l_i. It follows that the Bayes decision rule can be reformulated as

$$\pi_i f_i(\boldsymbol{x}) > \pi_j f_j(\boldsymbol{x}) \iff D(\boldsymbol{x}, F_i) > r_{ij}(D(\boldsymbol{x}, F_j)) \tag{2.4.7}$$

for some real increasing function r_{ij}. Using this observation, the DD classifier (Li et al., 2012) picks separating functions r_{ij} which best classify the training data $\mathscr{D}$.

Definition 2.4.7 (Empirical misclassification rate)**.** The empirical misclassification rate of a classifier $\hat{\imath}$, with respect to data $\mathscr{D}$, is given by

$$\hat{\Delta}(\hat{\imath}) \;=\; \sum_{i=1}^{k} \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} \mathbf{1}(\hat{\imath}(\boldsymbol{x}_{ij}) \ne i). \tag{2.4.8}$$

Definition 2.4.8 (DD classifier)**.** Suppose that $k = 2$, that D is a depth function, and that $r\colon [0,1] \to [0,1]$ is an increasing function. The DD classifier $\hat{\imath}_{D,r}$ is given by

$$\hat{\imath}_{D,r}(\boldsymbol{x}) \;=\; \begin{cases} 1, & \text{if } D(\boldsymbol{x}, F_2) \le r(D(\boldsymbol{x}, F_1)), \\ 2, & \text{if } D(\boldsymbol{x}, F_2) > r(D(\boldsymbol{x}, F_1)). \end{cases} \tag{2.4.9}$$

The empirical DD classifier $\hat{\imath}_{D,\hat{r}}$ replaces F_i by their empirical counterparts $\hat{F}_i$. Here, the separating curve $\hat{r}$ is chosen from a family Γ so as to minimize the empirical misclassification rate, i.e.

$$\hat{r} = \arg\min_{r \in \Gamma} \hat{\Delta}(\hat{\imath}_{D,r}). \tag{2.4.10}$$

Remark. The maximum depth classifier $\hat{\imath}_D$ is simply the DD classifier $\hat{\imath}_{D,\mathrm{id}}$, where $\mathrm{id}(x) = x$. Figure 2.4 clearly illustrates how this choice of separating function may not always be appropriate.

Li et al. (2012) show that under certain restrictions, the empirical DD classifier is asymptotically equivalent to the Bayes rule. We give one such instance below.

Lemma 2.4.9. *Suppose that the following conditions hold.*

1. Γ *is the class of polynomial functions on* $[0,1]$.
2. *The depth functions* $D(\,\cdot\,, F_i)$ *are continuous.*
3. *As* $\min\{n_1, n_2\} \to \infty$*, we have for each* $i \in \{1,2\}$*,*

$$\sup_{\boldsymbol{z} \in \mathbb{R}^d} |D(\boldsymbol{z}, \hat{F}_i) - D(\boldsymbol{z}, F_i)| \xrightarrow{a.s.} 0. \tag{2.4.11}$$

4. *The distributions* F_i *are elliptical and satisfy for all* $\delta \in \mathbb{R}$

$$P(D(\boldsymbol{Z}, F_i) = \delta \mid \boldsymbol{Z} \sim F_i) = 0. \tag{2.4.12}$$

Then, $\Delta(\hat{\imath}_{D,\hat{r}}) \to \Delta_B$ *as* $\min\{n_1, n_2\} \to \infty$.

In all the depth based classifiers we have seen so far, the classification rule depends on the observation $\boldsymbol{x}$ only through the depths $D(\boldsymbol{x}, F_i)$. Thus, we are motivated to define the following transformation from $\mathbb{R}^d$ to a depth feature space.

Definition 2.4.10. The depth feature vector $\boldsymbol{x}^D$ of an observation $\boldsymbol{x}$, with respect to the population distributions F_i and a choice of depth function D, is defined as

$$\boldsymbol{x}^D \;=\; (D(\boldsymbol{x}, F_1), \ldots, D(\boldsymbol{x}, F_k))\,. \tag{2.4.13}$$

Remark. The graph

$$\mathrm{DD}(F_1, \ldots, F_k) = \{\boldsymbol{x}^D : \boldsymbol{x} \in \mathbb{R}^d\} \tag{2.4.14}$$

is the analogue of the DD plot, with k distributions.

Assuming that the depth function D only takes values in $[0,1]$, the map $\boldsymbol{x} \mapsto \boldsymbol{x}^D$ takes values in $[0,1]^k$, regardless of the dimensionality of the original vector $\boldsymbol{x}$. With this, the maximum depth classification rule can be expressed as

$$\hat{\imath}_D(\boldsymbol{x}) = i \iff \boldsymbol{x}^D \in R_i^D = \{\boldsymbol{y} \in [0,1]^k : y_i = \max_j y_j\}. \tag{2.4.15}$$

Indeed, any partition of the unit cube $[0,1]^k$ into k decision regions R_i^D gives rise to a depth based classifier. The DD classifier achieves this by using an increasing separating function r to partition $[0,1]^2$. Furthermore, $r \in \Gamma$ is chosen so as to best separate the training data $\mathscr{D}$ transformed into the depth feature space. However, we can in principle use the transformed training data

$$\mathscr{D}^D = \{(\boldsymbol{x}_{ij}^D, i)\} \subset [0,1]^k \times \{1, \ldots, k\} \tag{2.4.16}$$

along with any multivariate classification algorithm (LDA, QDA, kNN, GLM, etc) to devise suitable decision regions. This is the basis of the DD^G classifier (Cuesta-Albertos et al., 2017).

2.5 Clustering

The unsupervised clustering grouping a collection of observations, such that points within the same group are more similar to each other than those from different groups.

Definition 2.5.1 (Clustering)**.** Given observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$, a clustering assignment is a choice of a partition $I_1, \ldots, I_K$ of $\{1, \ldots, N\}$.

With this notation, the k-th cluster consists of the points $\{\boldsymbol{x}_i\}_{i \in I_k}$. A good cluster assignment is one that maximizes similarity within clusters, as well as dissimilarity between clusters. Thus, the problem of clustering can be framed as the optimization of some objective function which combines these notions of similarity and dissimilarity. A simple algorithm such as the K-means clustering seeks to minimize

$$\{I_1, \ldots, I_K\} \mapsto \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2, \tag{2.5.1}$$

the average sum of square distances between each point and its cluster mean

$$\boldsymbol{\mu}_k = \frac{1}{|I_k|} \sum_{i \in I_k} \boldsymbol{x}_i = \arg\min_{\boldsymbol{\mu} \in \mathbb{R}^d} \sum_{i \in I_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2. \tag{2.5.2}$$

Jörnsten (2004) proposes a depth based approach to this problem, by examining the depth of a point within its cluster, relative to its depth within the best competing cluster.

In this section, we will abbreviate $D_k(\boldsymbol{x}) = D(\boldsymbol{x}, \hat{F}_{I_k})$, i.e. the empirical depth of $\boldsymbol{x}$ with respect to the points in the k-th cluster. Jörnsten (2004) chooses L_1 depth, the empirical version of spatial depth.

Definition 2.5.2. The within cluster depth of $\boldsymbol{x}_i$ is $D_i^w = D_k(\boldsymbol{x}_i)$, where $i \in I_k$.

To deal with dissimilarity between clusters, we represent each cluster by its L_1-median.

Definition 2.5.3 (L_1-median)**.** The L_1-median of the k-th cluster is given by

$$\boldsymbol{\theta}_k = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \sum_{i\in I_k} \|\boldsymbol{x}_i - \boldsymbol{\theta}\|. \tag{2.5.3}$$

Definition 2.5.4. The between cluster depth of $\boldsymbol{x}_i$ is $D_i^b = D_\ell(\boldsymbol{x}_i)$, where

$$\ell = \underset{k\colon i\notin I_k}{\arg\min} \|\boldsymbol{x}_i - \boldsymbol{\theta}_k\|. \tag{2.5.4}$$

In other words, the between cluster depth of $\boldsymbol{x}_i$ is its depth within the best competing cluster.

Definition 2.5.5 (Relative depth)**.** The relative depth of $\boldsymbol{x}_i$ is $\mathrm{ReD}_i = D_i^w - D_i^b$.

A point $\boldsymbol{x}_i$ is *well clustered* if ReD_i is very high, i.e. it is deep within its own cluster, and has low depth with respect to its next best competing cluster. Thus, to obtain a good clustering, we may choose to maximize the objective function

$$\{I_1, \ldots, I_K\} \mapsto \frac{1}{N} \sum_{k=1}^{K} \sum_{i\in I_k} \mathrm{ReD}_i, \tag{2.5.5}$$

which is simply the average relative depth. This maximization can be achieved iteratively, starting with a random cluster assignment and reassigning a subset of observations with low ReD_i to their nearest competing clusters. The reassignment is accepted if the objective function increases, and the process is repeated. Jörnsten (2004) also suggests the use of simulated annealing to overcome the problem of getting trapped in local maxima. Here, the reassignment is accepted with some probability $P(\beta, \delta)$ where δ is the change in the objective function value, even if the objective function decreases at that step. $P(\beta, \delta)$ is chosen to decrease with increasing β and δ. The tuning parameter β can be increased every iteration so that the probability of accepting poorer clustering assignments drops to zero eventually.

Another notion of similarity and dissimilarity involves *silhouette width*.

Definition 2.5.6 (Silhouette width)**.** Denote the average distance of $\boldsymbol{z}$ from points in the k-th cluster not equal to $\boldsymbol{z}$ by

$$\bar{d}_k(\boldsymbol{z}) = \frac{1}{|\{i \in I_k\colon \boldsymbol{x}_i \neq \boldsymbol{z}\}|} \sum_{\substack{i\in I_k \\ \boldsymbol{x}_i\neq\boldsymbol{z}}} \|\boldsymbol{x}_i - \boldsymbol{z}\|. \tag{2.5.6}$$

The silhouette width of $\boldsymbol{x}_i$ where $i \in I_k$ is given by

$$\mathrm{Sil}_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \qquad a_i = \bar{d}_k(\boldsymbol{x}_i), \quad b_i = \min_{\ell\neq k} \bar{d}_\ell(\boldsymbol{x}_i). \tag{2.5.7}$$

It has been observed that the silhouette width is greatly affected by differences in scale between clusters, while the relative depth is not. An objective function of the form

$$\{I_1, \ldots, I_K\} \mapsto \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} (1 - \lambda) \operatorname{Sil}_i + \lambda \operatorname{ReD}_i \tag{2.5.8}$$

may be used to combine both notions. Here, $\lambda \in [0, 1]$ controls the influence of the relative depth. It seems that small values of λ encourages equal scale clusters, while large values of λ allows unequal scale clusters. Thus, λ may be tuned accordingly to favour these different kinds of clustering assignments.

2.6 Outlier detection

Chapter 3

Functional Data

3.1 Classification

3.2 Clustering

3.3 Outlier detection

3.4 Partially Observed Functional Data

Chapter 4

Local Depth Functions

4.1 Regression using Local Depth Regions

Chapter 5

Conclusion

Bibliography

Agostinelli, C., & Romanazzi, M. (2011). Local depth. *Journal of Statistical Planning and Inference*, *141*(2), 817–830.

Chakraborty, A., & Chaudhuri, P. (2014). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, *42*(3), 1203–1231.

Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, *45*(1), 223–256.

Cuesta-Albertos, J. A., Febrero-Bande, M., & Oviedo de la Fuente, M. (2017). The DD^G-classifier in the functional setting. *TEST*, *26*(1), 119–142.

Cuesta-Albertos, J. A., & Nieto-Reyes, A. (2008). The Tukey and the random Tukey depths characterize discrete distributions. *Journal of Multivariate Analysis*, *99*(10), 2304–2311.

Dai, W., & Genton, M. G. (2018). An outlyingness matrix for multivariate functional data classification. *Statistica Sinica*, *28*(4), 2435–2454.

Elías, A., Jiménez, R., Paganoni, A. M., & Sangalli, L. M. (2023). Integrated depths for partially observed functional data. *Journal of Computational and Graphical Statistics*, *32*(2), 341–352.

Elías, A., Jiménez, R., & Shang, H. L. (2022). On projection methods for functional time series forecasting. *Journal of Multivariate Analysis*, *189*, 104890.

Elías, A., Jiménez, R., & Shang, H. L. (2023). Depth-based reconstruction method for incomplete functional data. *Computational Statistics*, *38*(3), 1507–1535.

Fraiman, R., Liu, R. Y., & Meloche, J. (1997). Multivariate Density Estimation by Probing Depth. *Lecture Notes-Monograph Series*, *31*, 415–430.

Ghosh, A. K., & Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, *32*(2), 327–350.

Gijbels, I., & Nagy, S. (2017). On a General Definition of Depth for Functional Data. *Statistical Science*, *32*(4), 630–639.

Jörnsten, R. (2004). Clustering and classification based on the L_1 data depth [Special Issue on Multivariate Methods in Genomic Data Analysis]. *Journal of Multivariate Analysis*, *90*(1), 67–89.

Kneip, A., & Liebl, D. (2020). On the optimal reconstruction of partially observed functional data. *The Annals of Statistics*, *28*(3), 1692–1717.

Kong, L., & Zuo, Y. (2010). Smooth depth contours characterize the underlying distribution. *Journal of Multivariate Analysis*, *101*(9), 2222–2226.

Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *77*(4), 777–801.

Li, J., Cuesta-Albertos, J. A., & Liu, R. Y. (2012). DD-Classifier: Nonparametric Classification Procedure Based on DD-Plot. *Journal of the American Statistical Association*, *107*(498), 737–753.

Liu, R. Y. (1990). On a Notion of Data Depth Based on Random Simplices. *The Annals of Statistics*, *18*(1), 405–414.

Liu, R. Y., Parelius, J. M., & Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *The Annals of Statistics*, *27*(3), 783–858.

Liu, R. Y., & Singh, K. (1993). A Quality Index Based on Data Depth and Multivariate Rank Tests. *Journal of the American Statistical Association*, *88*, 252–260.

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, *104*(486), 718–734.

Mosler, K., & Mozharovskyi, P. (2022). Choosing Among Notions of Multivariate Depth Statistics. *Statistical Science*, *37*(3), 348–368.

Nagy, S. (2017). Monotonicity properties of spatial depth. *Statistics & Probability Letters*, *129*, 373–378.

Nagy, S. (2021). Halfspace depth does not characterize probability distributions. *Statistical Papers*, *62*(3), 1135–1139.

Nieto-Reyes, A., & Battey, H. (2016). A Topologically Valid Definition of Depth for Functional Data. *Statistical Science*, *31*(1), 61–79.

Shi, X., Zhang, Y., & Fu, Y. (2023). Two-sample tests based on data depth. *Entropy*, *25*(2).

Villani, C. (2003). *Topics in optimal transportation*. American Mathematical Society.

Zuo, Y., & Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, *28*(2), 461–482.