Chapter 1

Introduction

A *depth function* is a map $D\colon \mathscr{X} \times \mathscr{F} \to [0,1]$ which quantifies the *centrality* of a point $\boldsymbol{x} \in \mathscr{X}$, with respect to a distribution $F \in \mathscr{F}$, as a number from $[0,1]$ (Gijbels & Nagy, 2017; Liu et al., 1999; Mosler & Mozharovskyi, 2022; Zuo & Serfling, 2000). For fixed F, this induces a *center-outwards* ordering on the space $\mathscr{X}$, which may be the real line $\mathbb{R}$, a finite-dimensional space like $\mathbb{R}^d$, even infinite dimensional spaces like ℓ^2 or function spaces like $L^2[a,b]$.

The introduction of depth functions is motivated by a desire to lift the notions of univariate *centers*, *quantiles*, *ranks*, and *order statistics* to the multivariate setting and beyond (Mosler & Mozharovskyi, 2022; Zuo & Serfling, 2000). Indeed, once one has a suitable depth function $D(\,\cdot\,,\cdot\,)$ in hand, this may be accomplished as follows: the center $\boldsymbol{\theta}$ of a distribution F ought to be its deepest point, via $\boldsymbol{\theta} = \arg\max_{\boldsymbol{x}\in\mathscr{X}} D(\boldsymbol{x}, F)$. Instead of a p-th quantile, we may deal with a p-th quantile surface which consists of all points $\boldsymbol{x}$ such that the probability of $\boldsymbol{X} \sim F$ being deeper than $\boldsymbol{x}$ is p, via

$$R(\boldsymbol{x}, F) = P_{\boldsymbol{X}\sim F}(D(\boldsymbol{X}, F) \le D(\boldsymbol{x}, F)) = p. \tag{1.0.1}$$

Note the similarities between the 'rank' map $R(\boldsymbol{x}, F)$ and the cumulative distribution function $F(x) = P_{X\sim F}(X \le x)$; this will be explored further in our discussion of multivariate testing (see Section 2.4). With this, the rank of a point $\boldsymbol{x}_j$ within a sample $\{\boldsymbol{x}_i\}_{i=1}^n$ may be defined as the proportion of sample points deeper than it, via

$$R(\boldsymbol{x}_j, \hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(D(\boldsymbol{x}_i, \hat{F}_n) \le D(\boldsymbol{x}_j, \hat{F}_n)). \tag{1.0.2}$$

Finally, the center-outward order induced by $D(\,\cdot\,, \hat{F}_n)$ on $\mathscr{X}$ naturally lets us order our sample as $\boldsymbol{x}_{[1]}, \ldots, \boldsymbol{x}_{[n]}$, where $D(\boldsymbol{x}_{[1]}, \hat{F}_n) \le \cdots \le D(\boldsymbol{x}_{[n]}, \hat{F}_n)$. This allows us to extend a number of powerful tools in univariate nonparametric inference to these more general domains. For instance, the median is a useful, robust measure of location in the univariate setting; the *spatial median* of a distribution F, the point $\boldsymbol{\theta} \in \mathbb{R}^d$ satisfying

$$\mathbb{E}_{\boldsymbol{X}\sim F}\left[\frac{\boldsymbol{\theta} - \boldsymbol{X}}{\|\boldsymbol{\theta} - \boldsymbol{X}\|}\right] = \mathbf{0}, \tag{1.0.3}$$

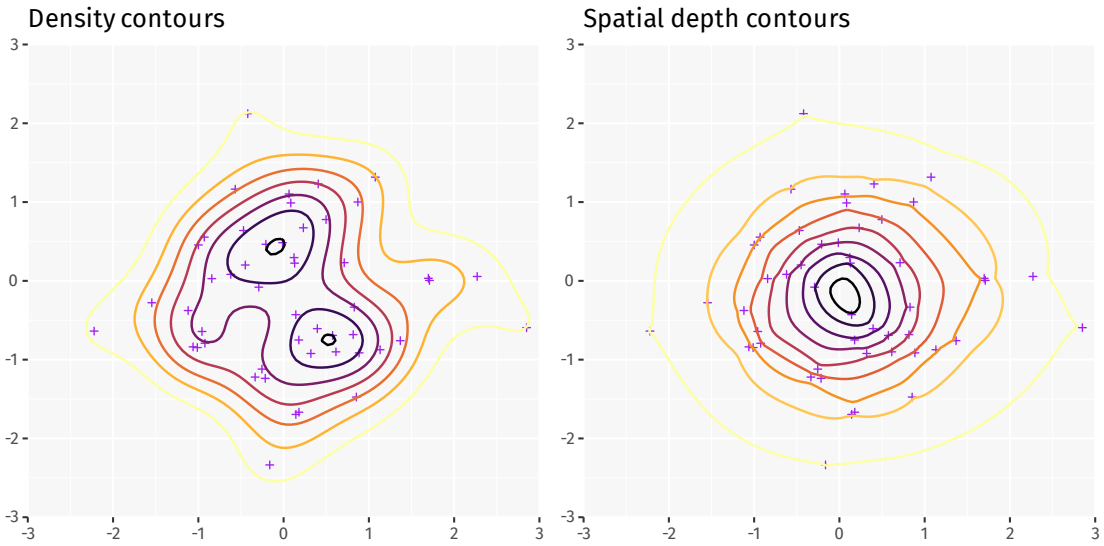


Figure 1.1: Density contours (via kernel density estimation) and spatial depth (Definition 2.1.4) contours for 50 points sampled from a standard bivariate normal distribution. The depth contours offer a clearer, more robust measure of centrality within the data cloud.

can be realized as the maximizer of the spatial depth (Definition 2.1.4). Similarly, Liu (1990) and Donoho and Gasko (1992) use simplicial and halfspace depths to define and study multivariate location estimators. Liu and Singh (1993) and Chenouri and Small (2012) use multivariate ranks induced by depth functions to devise homogeneity tests analogous to the univariate Wilcoxon rank-sum and Kruskal-Wallis tests. Liu et al. (1999) use depth functions for a variety of nonparametric multivariate procedures, including developing graphical tools such as the depth-depth plot (Definition 2.3.1) for exploratory data analysis. This idea later lead to the development of the DD classifier (Li et al., 2012) and the DD^G classifier (Cuesta-Albertos et al., 2017). A lot of these procedures exploit properties such as invariance, monotonicity, continuity, etc. of various depth functions as needed. Thus, depth functions provide a toolkit combining versatility, robustness and computational ease.

It may seem natural at first to rely upon the density function as an indication of centrality, the corresponding center being the mode. This approach presents a number of problems – not every distribution admits a density function (consider discrete distributions), and even when one exists, it may not be particularly informative (see Example 2.2.5). A more practical concern is that density functions become increasingly difficult to estimate for multivariate distributions on $\mathbb{R}^d$ as the dimension d increases – this is popularly known as the *curse of dimensionality*.

One method of estimating the density function f from a sample $\{\boldsymbol{x}_i\}_{i=1}^n$ is via a *kernel density estimator* (Wasserman, 2005), of the form

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i). \tag{1.0.4}$$

Here, $K_h \colon \mathbb{R}^d \to \mathbb{R}$ is a *kernel function* with *bandwidth(s)* h; a simple example is the Gaussian kernel

$$K_h(\boldsymbol{z}) = \frac{1}{\sqrt{2\pi}\prod_{i=1}^d h_i} \exp\left(-\sum_{i=1}^d \frac{z_i^2}{2h_i^2}\right) \tag{1.0.5}$$

The tuning parameters $\{h_i\}_{i=1}^d$ are typically determined via cross-validation; an optimal bandwidth is one that minimizes the integrated risk $\int_{\mathbb{R}^d} \mathbb{E}[(f(\boldsymbol{x}) - \hat{f}_h(\boldsymbol{x}))^2] d\boldsymbol{x}$. It can be shown that under certain circumstances, the integrated risk is of order $O(n^{-4/(4+d)})$, which grows swiftly with the dimension d. This in turn means that the number of sample points n required to achieve the same level of confidence explodes with increasing dimension d. Besides, it is often the case that the number of tuning parameters for the kernel K_h also increases. Of course, density estimation in the functional setting is much more complex. Altogether, density does not provide a tractable quantification of centrality. Depth functions will turn out to alleviate a majority of these theoretical and computational concerns.

In Chapter 2, we introduce depth functions on multivariate data and distributions on $\mathbb{R}^d$, with a brief discussion on some of their properties. We look at the depth-depth plot as a tool for exploratory data analysis, then move onto applications in testing, classification, and clustering tasks. In Chapter 3, we extend our understanding of depth functions to functional data and distributions on function spaces such as L^2 and $\mathcal{C}$. We see that although many procedures for multivariate data carry over naturally to this setting, functional data poses its own unique challenges. We explore applications of depth functions in classification and outlier detection tasks, and briefly examine the setting of partially observed data and the reconstruction problem. Finally, in Chapter 4, we discuss the concept of local depth, focusing on a general recipe for converting a global depth function into its local counterpart. Motivated by the construction of local depth regions, we propose a new kernel based regression procedure which works reasonably well for univariate, multivariate, and functional data.

Chapter 2

Multivariate Data

Let $\mathscr{F}$ be a class of distributions on $\mathbb{R}^d$. 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); these are sometimes referred to as the *Zuo-Serfling axioms*.

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 $\alpha \in [0, 1]$,

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

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

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

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

By demanding that D be non-negative and bounded, 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 $S > E > C > A > H$.

Mosler and Mozharovskyi (2022, Table 2) provides a detailed summary of the properties satisfied by the depth functions discussed in the following section.

2.1 Multivariate depth functions

The earliest formulation of a depth function may be attributed to Tukey (1975).

Definition 2.1.1 (Halfspace/Tukey depth)**.** Denote the collection of all closed halfspaces in $\mathbb{R}^d$ containing $\boldsymbol{x}$ by $\mathcal{H}_{\boldsymbol{x}}$. The halfspace depth, or Tukey depth, is defined as

$$D_H(\boldsymbol{x}, F) = \inf_{H \in \mathcal{H}_{\boldsymbol{x}}} P_F(H). \tag{2.1.1}$$

Remark. If $F \in \mathcal{F}$ is supported on a convex region $K \subset \mathbb{R}^d$, then $D(\,\cdot\,, F)$ vanishes outside K. More generally, for convex $K \subset \mathbb{R}^d$, we have $D_H(\boldsymbol{x}, F) \leq P_F(K^c)$ for all $\boldsymbol{x} \in K^c$. This is because one can choose a halfspace $H \in \mathcal{H}_{\boldsymbol{x}}$ entirely contained within K^c. Using this, we see that the halfspace depth obeys **P4**.

Proposition 2.1.2. *The halfspace depth can be formulated as*

$$D_H(\boldsymbol{x}, F) = \inf_{\boldsymbol{v} \in S^{d-1}} P_{\boldsymbol{X} \sim F}(\langle \boldsymbol{v}, \boldsymbol{X} \rangle \leq \langle \boldsymbol{v}, \boldsymbol{x} \rangle). \tag{2.1.2}$$

Remark. When $d = 1$, the halfspace depth reduces to

$$D_H(x, F) = \min\{P_F(-\infty, x],\ P_F[x, \infty)\}. \tag{2.1.3}$$

Definition 2.1.3 (Mahalanobis depth)**.** Let $\boldsymbol{X} \sim F$ have mean $\boldsymbol{\mu}$ and covariance matrix Σ. The Mahalanobis depth is defined as

$$D_M(\boldsymbol{x}, F) = \left(1 + (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)^{-1}. \tag{2.1.4}$$

Remark. The mean and covariance in the above definition may be replaced with more robust estimates $\boldsymbol{\mu}^*$ and Σ^*, for instance using the minimum covariance determinant (MCD) method. The corresponding depth function is called the robust Mahalanobis depth.

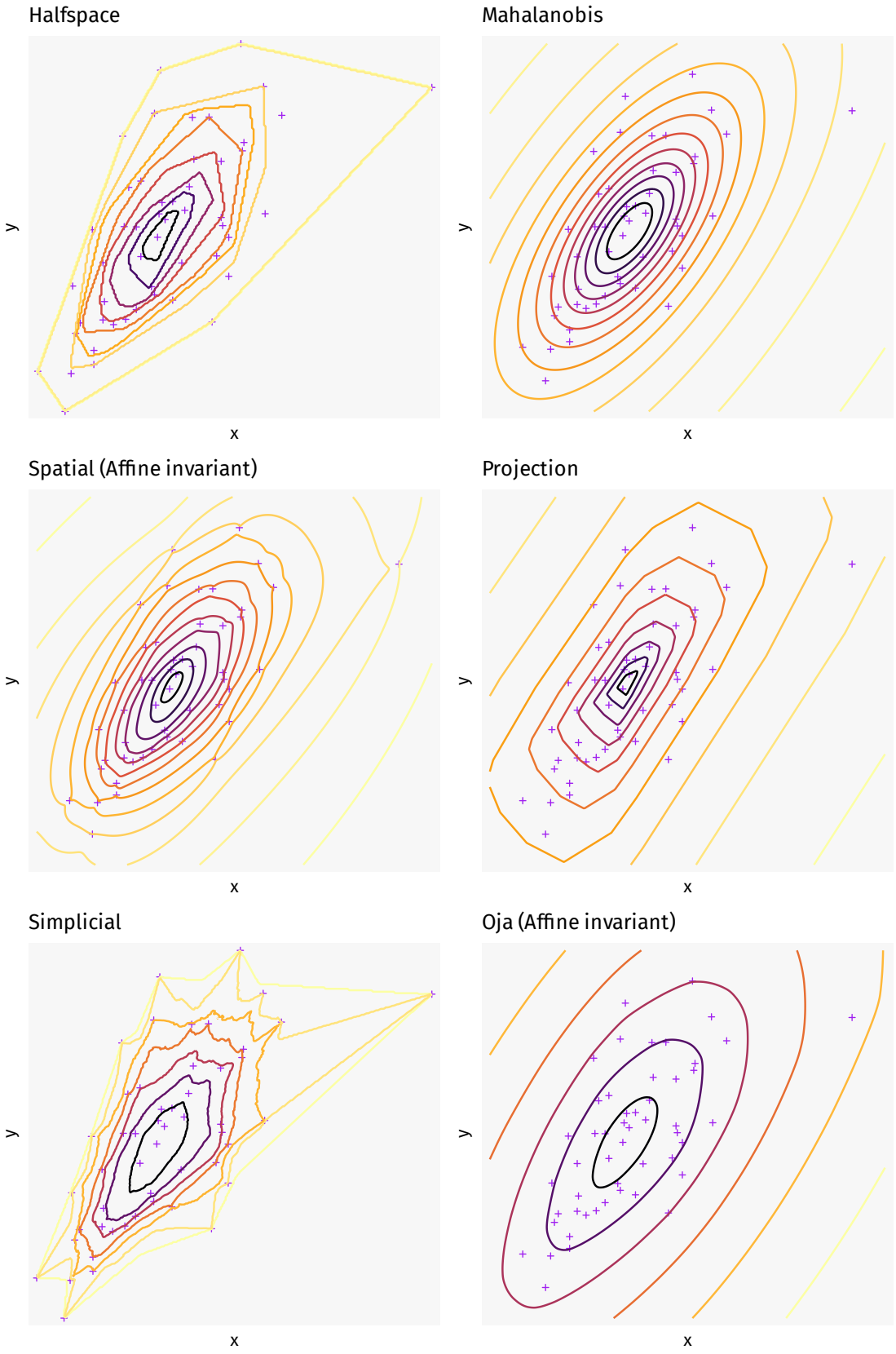


Figure 2.1: Depth contours with respect to purple points. Darker contours have higher depth.

The spatial depth was introduced in Serfling (2002),

Definition 2.1.4 (Spatial depth)**.** The spatial depth is defined as

$$D_{Sp}(\boldsymbol{x}, F) = 1 - \left\| \mathbb{E}_{\boldsymbol{X} \sim F} \left[\frac{\boldsymbol{x} - \boldsymbol{X}}{\|\boldsymbol{x} - \boldsymbol{X}\|} \right] \right\|. \tag{2.1.5}$$

We use the convention $\mathbf{0}/0 = \mathbf{0}$.

Remark. Spatial depth defined as above does not obey **P1**. Indeed, spatial depth is only invariant under spherical transformations of the form $U\boldsymbol{X} + \boldsymbol{b}$ for orthonormal U. We may define an affine invariant version of spatial depth as

$$D_{AISp}(\boldsymbol{x}, F) = 1 - \left\| \mathbb{E}_{\boldsymbol{X} \sim F} \left[\frac{\Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{X})}{\sqrt{(\boldsymbol{x} - \boldsymbol{X})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{X})}} \right] \right\|. \tag{2.1.6}$$

Remark. Nagy (2017) showed that spatial depth does not obey **P3**.

Definition 2.1.5 (Projection depth)**.** The projection depth is defined as

$$D_P(\boldsymbol{x}, F) = \left(1 + \sup_{\boldsymbol{v} \in S^{d-1}} \frac{|\langle \boldsymbol{v}, \boldsymbol{x} \rangle - \operatorname{med}(\langle \boldsymbol{v}, \boldsymbol{X} \rangle)|}{\operatorname{MAD}(\langle \boldsymbol{v}, \boldsymbol{X} \rangle)} \right)^{-1}, \quad \boldsymbol{X} \sim F. \tag{2.1.7}$$

Liu (1990) introduced the following depth function based on random simplices.

Definition 2.1.6 (Simplicial depth)**.** The simplicial depth is defined as

$$D_{Sim}(\boldsymbol{x}, F) = P_{\boldsymbol{X}_i \overset{\text{iid}}{\sim} F}(\boldsymbol{x} \in \operatorname{conv}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1})), \tag{2.1.8}$$

where $\operatorname{conv}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1})$ denotes the convex hull of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}\}$.

Definition 2.1.7 (Oja depth)**.** The simplicial volume depth, or Oja depth, is defined as

$$D_{Oja}(\boldsymbol{x}, F) = \left(1 + \mathbb{E}_{\boldsymbol{X}_i \overset{\text{iid}}{\sim} F} [\operatorname{vol}(\operatorname{conv}(\boldsymbol{x}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_d))] \right)^{-1}. \tag{2.1.9}$$

Remark. Oja depth does not obey **P1**, since

$$\operatorname{vol}(\operatorname{conv}(A\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, A\boldsymbol{x}_{d+1} + \boldsymbol{b})) = |\det(A)| \operatorname{vol}(\operatorname{conv}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1})). \tag{2.1.10}$$

Instead, we may define an affine invariant version of Oja depth as

$$D_{AIOja}(\boldsymbol{x}, F) = \left(1 + \mathbb{E}_{\boldsymbol{X}_i \overset{\text{iid}}{\sim} F} \left[\frac{\operatorname{vol}(\operatorname{conv}(\boldsymbol{x}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_d))}{\sqrt{\det(\Sigma)}} \right] \right)^{-1}, \tag{2.1.11}$$

where Σ is the covariance matrix of F.

Remark. The Mahalanobis, projection, and Oja depths all follow the pattern of $(1 + O(\boldsymbol{x}, F))^{-1}$, where $O(\boldsymbol{x}, F)$ measures some kind of outlyingess of $\boldsymbol{x}$ in F. We will often see this performed in reverse, extracting a measure of outlyingess $1/D(\boldsymbol{x}, F) - 1$ from a depth function D.

2.1.1 The projection property

Definition 2.1.8 (Projection property)**.** We say that a depth function D has the projection property if

$$D(\boldsymbol{x}, F_{\boldsymbol{X}}) = \inf_{\boldsymbol{v} \in S^{d-1}} D(\langle \boldsymbol{v}, \boldsymbol{x} \rangle, F_{\langle \boldsymbol{v}, \boldsymbol{X} \rangle}). \tag{2.1.12}$$

Depths which have this property can be approximated by calculating the univariate depths of the projected data along many directions $\boldsymbol{v}$.

Lemma 2.1.9 (Mosler and Mozharovskyi, 2022)**.** *The halfspace depth, Mahalanobis depth, and projection depth have the projection property.*

The halfspace depth in particular is often computationally challenging. Thus, the property motivates the definition of the random Tukey depth (Cuesta-Albertos & Nieto-Reyes, 2008).

Definition 2.1.10 (Random Tukey depth)**.** Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be a realization of an iid sample from $\mathcal{U}(S^{d-1})$. The random Tukey depth is defined as

$$D_{RT}(\boldsymbol{x}, F_{\boldsymbol{X}}) = \min_{1 \leq i \leq n} D_H(\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle, F_{\langle \boldsymbol{v}_i, \boldsymbol{X} \rangle}). \tag{2.1.13}$$

2.1.2 Continuity properties

It is also desirable for a depth function to obey some notions of continuity.

C1. *Continuity in* $\boldsymbol{x}$.

$$D(\boldsymbol{x}_n, F) \to D(\boldsymbol{x}, F) \text{ when } \boldsymbol{x}_n \to \boldsymbol{x}. \tag{2.1.14}$$

C2. *Continuity in* F.

$$D(\boldsymbol{x}, F_n) \to D(\boldsymbol{x}, F) \text{ when } F_n \xrightarrow{d} F. \tag{2.1.15}$$

C3. *Uniform continuity.*

$$\sup_{\boldsymbol{x} \in G} |D(\boldsymbol{x}, F_n) - D(\boldsymbol{x}, F)| \to 0 \text{ when } F_n \xrightarrow{d} F. \tag{2.1.16}$$

Property **C1** is rarely satisfied without imposing some regularity conditions on F, such as absolute continuity. Property **C2** helps bridge the gap between the population and empirical versions of depth. Property **C3** becomes relevant when dealing with the convergence of depth contours.

The Mahalanobis depth is trivially continuous in $\boldsymbol{x}$, i.e. obeys **C1**. Furthermore, it also satisfies **C2** as long as F has a regular covariance matrix (Mosler & Mozharovskyi, 2022).

The halfspace depth also enjoys all three notions of continuity, under mild restrictions on F.

Theorem 2.1.11 (Mizera and Volauf, 2002)**.** *Let $F \in \mathscr{F}$ be such that the probability of every hyperplane in $\mathbb{R}^d$ is zero, i.e. for all $\alpha \in \mathbb{R}$ and $\boldsymbol{v} \in S^{d-1}$,*

$$P_{\boldsymbol{X}\sim F}(\langle \boldsymbol{v}, \boldsymbol{X}\rangle = \alpha) \;=\; 0. \tag{2.1.17}$$

Then for $\boldsymbol{x}_n \to \boldsymbol{x}$ and $F_n \xrightarrow{d} F$, we have $D_H(\boldsymbol{x}_n, F_n) \to D_H(\boldsymbol{x}, F)$.

Remark. Equation 2.1.17 is satisfied whenever F is absolutely continuous.

Remark. It follows that if $F \in \mathscr{F}$ satisfies Equation 2.1.17, then the map $D_H(\,\cdot\,, F)$ is continuous.

Corollary 2.1.12. *Let $F \in \mathscr{F}$ satisfy Equation 2.1.17. Then, for $F_n \xrightarrow{d} F$, and compact $K \subset \mathbb{R}^d$,*

$$\sup_{\boldsymbol{x}\in K} |D_H(\boldsymbol{x}, F_n) - D_H(\boldsymbol{x}, F)| \to 0. \tag{2.1.18}$$

Proof. Denoting $g = D(\,\cdot\,, F)$, $g_n = D_H(\,\cdot\,, F_n)$, we have the continuity of g along with $g_n(\boldsymbol{x}_n) \to g(\boldsymbol{x})$ whenever $\boldsymbol{x}_n \to \boldsymbol{x}$ in K. If the given conclusion is false, we may pass to a subsequence of g_n and find $\epsilon > 0$ such that each $\sup_{\boldsymbol{x}\in K} |g_n(\boldsymbol{x}) - g(\boldsymbol{x})| \geq \epsilon$. Using the compactness of K, we pass to a further subsequence and find $\boldsymbol{x} \in K$ such that $\boldsymbol{x}_n \to \boldsymbol{x}$. This contradicts $|g_n(\boldsymbol{x}_n) - g(\boldsymbol{x}_n)| \geq \epsilon$. □

Theorem 2.1.13. *Let $F \in \mathscr{F}$ satisfy Equation 2.1.17. Then, for $F_n \xrightarrow{d} F$,*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} |D_H(\boldsymbol{x}, F_n) - D_H(\boldsymbol{x}, F)| \to 0. \tag{2.1.19}$$

Proof. Let $K_r = \{\boldsymbol{x} \in \mathbb{R}^d \colon \|\boldsymbol{x}\| \leq r\}$ be a continuity set of F. Observe that $D_H(\boldsymbol{y}, F) \leq P_F(K_r^c)$ for $\boldsymbol{y} \in K_r^c$, hence

$$\sup_{\boldsymbol{y}\in K_r^c} |D_H(\boldsymbol{y}, F_n) - D_H(\boldsymbol{y}, F)| \leq P_{F_n}(K_r^c) + P_F(K_r^c). \tag{2.1.20}$$

As $n \to \infty$, we have $P_{F_n}(K_r^c) \to P_F(K_r^c) = p_r$ (say). Thus, denoting $\delta_n(X) = \sup_{\boldsymbol{x}\in X} |D_H(\boldsymbol{x}, F_n) - D_H(\boldsymbol{x}, F)|$, we have

$$\limsup_{n\to\infty} \delta_n(\mathbb{R}^d) \leq \lim_{n\to\infty} \delta_n(K_r) + \limsup_{n\to\infty} \delta_n(K_r^c) \tag{2.1.21}$$

$$\leq 0 + 2p_r. \tag{2.1.22}$$

Using $p_r \to 0$ as $r \to \infty$ completes the proof. □

The spatial depth is similarly well behaved.

Theorem 2.1.14. *Spatial depth obeys **C1** when F is non-atomic, as well as **C2**.*

Proof. Consider the spatial map

$$S_F \colon \mathbb{R}^d \to \mathbb{R}^d, \qquad \boldsymbol{x} \mapsto \mathbb{E}_{\boldsymbol{X}\sim F}\left[\frac{\boldsymbol{x} - \boldsymbol{X}}{\|\boldsymbol{x} - \boldsymbol{X}\|}\right]. \tag{2.1.23}$$

The Dominated Convergence Theorem guarantees the continuity of S_F, hence of $D_{Sp}(\,\cdot\,, F) = 1 - \|S_F(\,\cdot\,)\|$. Furthermore, if $F_n \xrightarrow{d} F$, we have $S_{F_n}(\boldsymbol{x}) \to S_F(\boldsymbol{x})$ by the Portmanteau Lemma for all $\boldsymbol{x} \in \mathbb{R}^d$. □

Theorem 2.1.15 (Serfling, 2002)**.** *For $F \in \mathscr{F}$ and compact $K \subset \mathbb{R}^d$,*

$$\sup_{\boldsymbol{x} \in K} |D_{Sp}(\boldsymbol{x}, \hat{F}_n) - D_{Sp}(\boldsymbol{x}, F)| \xrightarrow{a.s.} 0. \tag{2.1.24}$$

Remark. This result can be generalized from compact subsets K to the whole of $\mathbb{R}^d$, using the following Lemma (2.1.16) and arguments similar to the proof of Theorem 2.1.13.

Lemma 2.1.16. *Spatial depth obeys **P4**, i.e. $D_{Sp}(\boldsymbol{x}, F) \to 0$ as $\|\boldsymbol{x}\| \to \infty$.*

Proof. Let $\epsilon > 0$, and let $M > 0$ such that $P_{\boldsymbol{X} \sim F}(\|\boldsymbol{X}\| > M) = \epsilon$. Denote $\boldsymbol{Y} = (\boldsymbol{x} - \boldsymbol{X})/\|\boldsymbol{x} - \boldsymbol{X}\|$, and observe that

$$\mathbb{E}_{\boldsymbol{X} \sim F}[\boldsymbol{Y}] = (1 - \epsilon)\,\mathbb{E}[\boldsymbol{Y} \,|\, \|\boldsymbol{X}\| \le M] + \epsilon\,\mathbb{E}[\boldsymbol{Y} \,|\, \|\boldsymbol{X}\| > M] \tag{2.1.25}$$

Thus, using $\|\boldsymbol{Y}\| = 1$ and the reverse triangle inequality,

$$\|\,\mathbb{E}[\boldsymbol{Y}]\,\| \ge (1 - \epsilon)\|\,\mathbb{E}[\boldsymbol{Y} \,|\, \|\boldsymbol{X}\| \le M]\,\| - \epsilon. \tag{2.1.26}$$

Let $\alpha = \arccos((1 - 2\epsilon)/(1 - \epsilon))$, and let $r_\alpha = M/\sin\alpha$. It follows that the ball $\{\boldsymbol{x} \in \mathbb{R}^d \colon \|\boldsymbol{x}\| \le M\}$ subtends an angle of at most 2α from any point $\boldsymbol{x}$ such that $\|\boldsymbol{x}\| > r_\alpha$. This gives $\|\,\mathbb{E}[\boldsymbol{Y} \mid \|\boldsymbol{X}\| \le M]\| \ge \cos\alpha$. Thus, for $\|\boldsymbol{x}\| > r_\alpha$,

$$\|\,\mathbb{E}[\boldsymbol{Y}]\| \ge (1 - \epsilon)\cos\alpha - \epsilon = 1 - 3\epsilon, \tag{2.1.27}$$

whence $D_{Sp}(\boldsymbol{x}, F) \le 3\epsilon$. □

2.1.3 Characterization properties

It seems natural to ask the question – do depth functions completely characterize a distribution, the way density functions do? In other words, can we recover $F \in \mathscr{F}$ from $D(\,\cdot\,, F)$? For most depth functions, the answer is 'no', unless we greatly restrict the class of functions $\mathscr{F}$ under consideration. For instance, the Mahalanobis depth $D_M(\,\cdot\,, F)$ only depends on the first two moments of F, and thus has no hope of distinguishing between distributions which differ in higher moments. On the other hand, if we only consider a family of elliptical distributions $\mathrm{Ell}(h; \cdot, \cdot)$ for strictly monotonically decreasing h, it can be shown that any depth D satisfying **P1** and **C1** uniquely determines F (Mosler & Mozharovskyi, 2022).

Definition 2.1.17 (Elliptical distributions)**.** We say that a distribution is elliptical if it has a density of the form

$$f(\boldsymbol{x}) = c\,|\Sigma|^{-1/2}\,h\left((\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right) \tag{2.1.28}$$

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

A positive result for the halfspace depth is that it fully characterizes empirical distributions.

Theorem 2.1.18 (Struyf and Rousseeuw, 1999)**.** *The empirical distribution of any dataset* $\{\boldsymbol{X}_i\}_{i=1}^n \subset \mathbb{R}^d$ *is uniquely determined by its halfspace depth function* $D(\,\cdot\,, \hat{F}_n)$.

Nagy (2020) offers a comprehensive overview of the halfspace depth characterization problem. Indeed, Nagy (2021) supplies examples of distinct probability distributions F_1, F_2 such that $D_H(\,\cdot\,, F_1) = D_H(\,\cdot\,, F_2)$. It can be shown that for an α-symmetric distribution F with $0 < \alpha \leq 1$ that $D_H(\boldsymbol{x}, F) = G(-\|\boldsymbol{x}\|_\infty)$, where G is the marginal distribution of the first component of $\boldsymbol{X} \sim F$. Using this idea, Nagy (2021) produces two such distributions with the same marginal G.

2.2 Depth contours

Given a depth function D and some fixed distribution $F \in \mathscr{F}$, we may examine contours produced by $D(\,\cdot\,, F)$. The following definitions are adapted from Liu et al., 1999.

Definition 2.2.1. The contour of depth t is the set $\{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}, F) = t\}$.

Definition 2.2.2. The region enclosed by the contour of depth t is the set

$$R_F(t) \;=\; \{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}, F) > t\}. \tag{2.2.1}$$

It is often more convenient to deal with depth contours and regions based on their probability content rather than a depth cutoff.

Definition 2.2.3. The p-th central region is the set

$$C_F(p) \;=\; \bigcap_t \{R_F(t) : P_F(R_F(t)) \geq p\}. \tag{2.2.2}$$

Definition 2.2.4. The p-th level contour, or center-outward contour surface, is the set $Q_F(p) = \partial C_F(p)$.

Example 2.2.5. 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.

Depth based central regions and contours may be approximated empirically as follows.

Definition 2.2.6. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$. We introduce depth based order statistics $\boldsymbol{X}_{[1]}, \ldots, \boldsymbol{X}_{[n]}$, which are a reordering of the sample in decreasing order of depth, i.e. $D(\boldsymbol{X}_{[1]}, F) \geq \cdots \geq D(\boldsymbol{X}_{[n]}, F)$.

With this, given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$, the sample p-th central region is given by

$$C_{\hat{F}_n}(p) = \text{conv}(\boldsymbol{X}_{[1]}, \ldots, \boldsymbol{X}_{[\lceil np \rceil]}). \tag{2.2.3}$$

The consistency of these sample central regions typically requires some continuity of type **C3** (Donoho & Gasko, 1992; He & Wang, 1997; Liu, 1990).

Depths such as the halfspace depth, the Mahalanobis depth, and the Oja depth produce convex central regions. Any depth satisfying **P3** produces star-shaped central regions. Notably, the spatial depth does not necessarily produce convex nor star-shaped central regions (Nagy, 2017).

2.2.1 The Monge-Kantorovich depth

Depths which produce convex, nested central regions are appropriate for a large class of unimodal distributions with some degree of symmetry. Indeed, the fact that depths satisfying **P1** and **C1** characterize elliptical distributions follows from the fact that the depth contours coincide with density contours. However, there are instances where this is unsuitable, such as in the distributions illustrated in Figures 4.1 and 4.2. In Chapter 4, we will examine the idea of *local depth*, which is capable of producing non-convex, non-star-shaped, un-nested central regions.

Here, we briefly look at the Monge-Kantorovich depth introduced by Chernozhukov et al. (2017), which is also capable of producing non-convex contours but retains their nestedness. This is based on the idea of 'transporting' contours from a reference distribution U_d, say $\mathcal{U}(B^d)$, to the target distribution F on $\mathbb{R}^d$ via a canonical *vector quantile map* $Q\colon B^d \to \mathbb{R}^d$. We say that Q 'pushes forward' U_d into F, denoted $Q\#F = U_d$; for $U \sim U_d$, we have $Q(U) \sim F$. The inverse map from F to the reference U_d is called the *vector rank map* R; we write $R\#F = U_d$. Now, Q is defined via the theory of optimal transport (Villani, 2003); it is the map which minimizes the quadratic cost $\mathbb{E}_{U\sim U_d}[(Q(U) - U)^2]$ subject to $Q(U) \sim F$. Under certain conditions, the maps Q, R exist and are unique; for instance, the Brenier-McCann theorem requires the absolute continuity of U, F supported within convex subsets of $\mathbb{R}^d$. With this, we supply a loose definition of the Monge-Kantorovich depth below.

Definition 2.2.7 (Monge-Kantorovich depth)**.** Let Q be the vector quantile map associated with F, and let R be its inverse, so that $R\#F = U_d$. The Monge-Kantorovich depth is defined as

$$D_{MK}(\boldsymbol{x}, F) = D_H(R(\boldsymbol{x}), U_d). \tag{2.2.4}$$

Similarly, the Monge-Kantorovich rank of $\boldsymbol{x} \in \mathbb{R}^d$ in F is given by $\|R(\boldsymbol{x})\|$. If $K_p = \partial C_{U_d}(p)$ is the p-th level contour of U_d, then the Monge-Kantorovich p-quantile of F is the image $Q(K_p)$. The reference distribution U_d and depth D_H may of course be replaced as necessary.

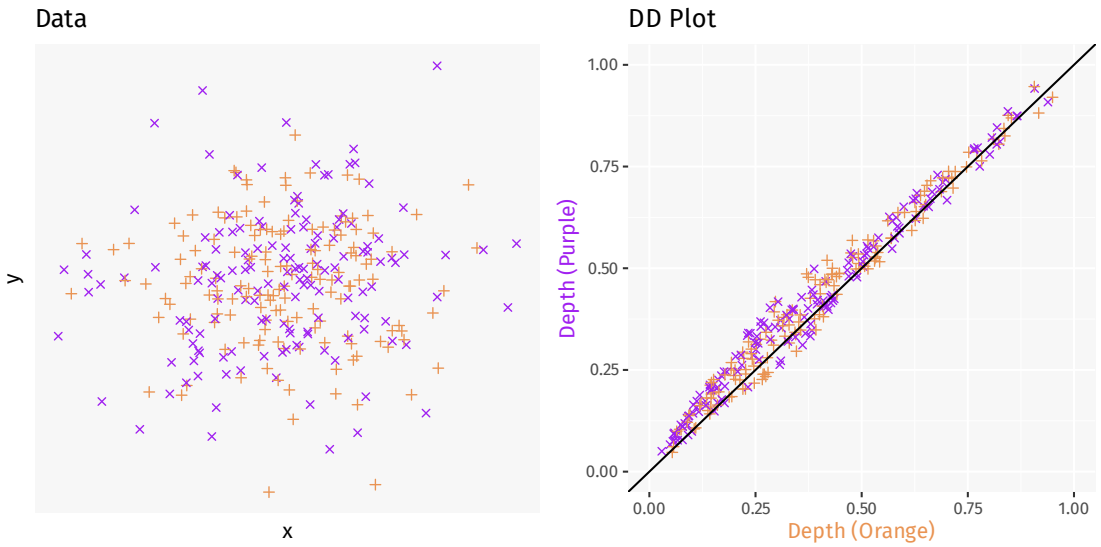


Figure 2.2: Empirical DD plot using spatial depth, where both underlying distributions (bivariate normal) are identical. Observe how the points in the DD plot stay close to the diagonal black line.

One major strength of this notion of depth is the distribution-free nature of the ranks produced. This has produced applications in areas such as distribution-free (nonparametric, multivariate) testing (Deb & Sen, 2023; Ghosal & Sen, 2022).

2.3 Depth-Depth plots

The *DD plot*, introduced by Liu et al. (1999), is a very useful tool for visualizing differences between distributions.

Definition 2.3.1 (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

$$\mathrm{DD}(F, G) \;=\; \{(D(\boldsymbol{z}, F), D(\boldsymbol{z}, G)) : \boldsymbol{z} \in \mathbb{R}^d\}. \tag{2.3.1}$$

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 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

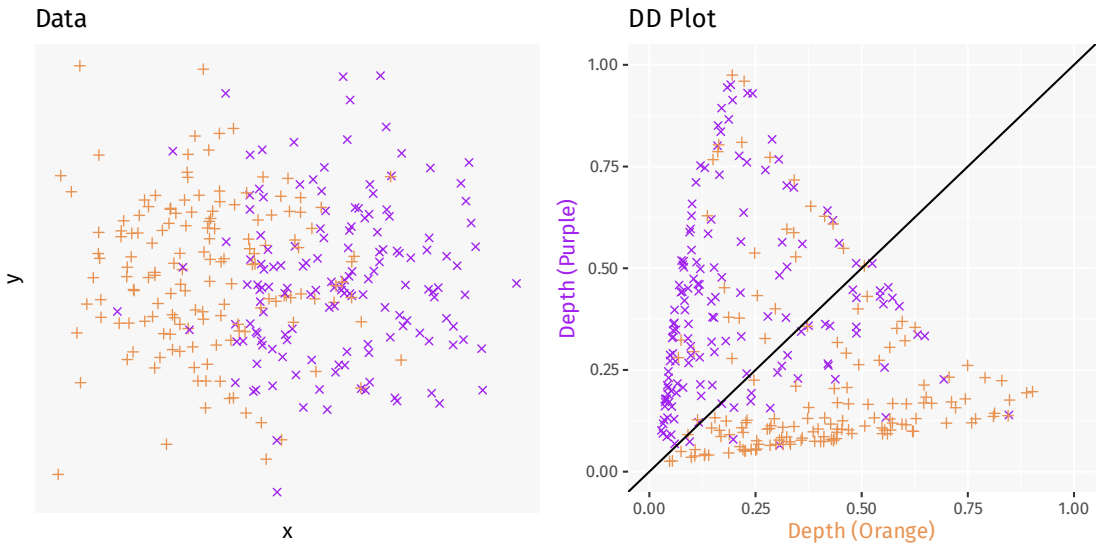


Figure 2.3: Empirical DD plot using spatial depth, where both underlying distributions (bivariate normal) differ only in location. Observe how most of the orange points fall in the lower triangle, while the purple ones fall in the upper triangle. The deepest point with respect to the orange distribution has fairly low depth with respect to the purple one, and vice versa.

distributions $\hat{F}_n, \hat{G}_m$. With this, we may examine the empirical DD plot

$$\mathrm{DD}(\hat{F}_n, \hat{G}_n) = \{(D(\boldsymbol{z}, \hat{F}_n), D(\boldsymbol{z}, \hat{G}_m)) : \boldsymbol{z} \in \mathscr{D}_F \cup \mathscr{D}_G\}. \tag{2.3.2}$$

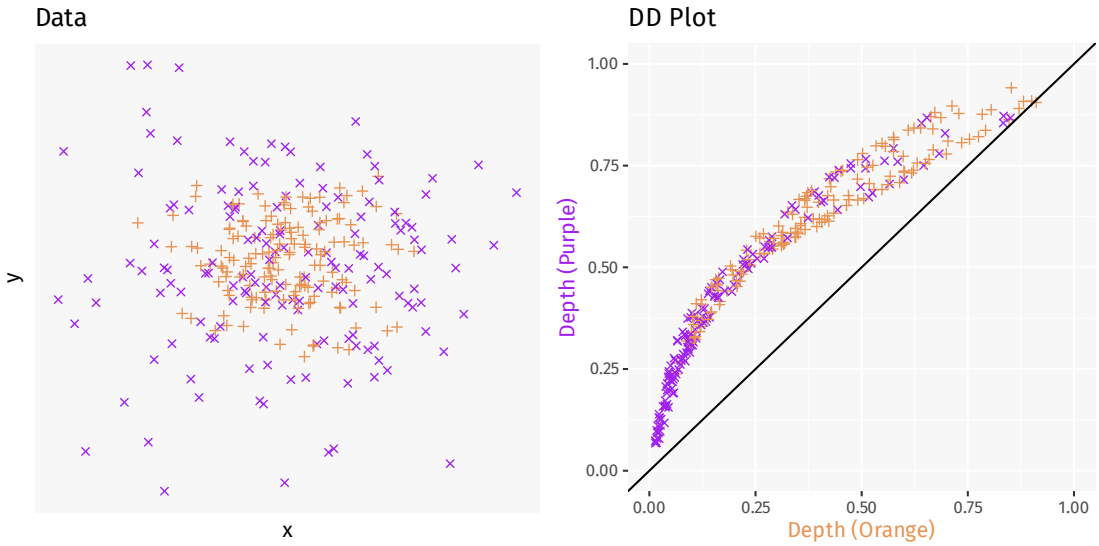
DD plots can be used as a diagnostic tool to detect differences in location and scale between two multivariate distributions.

1. If $F = G$, the points in $\mathrm{DD}(\hat{F}_n, \hat{G}_m)$ stay close to the diagonal. See Figure 2.2.
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 2.3.
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 2.4a.

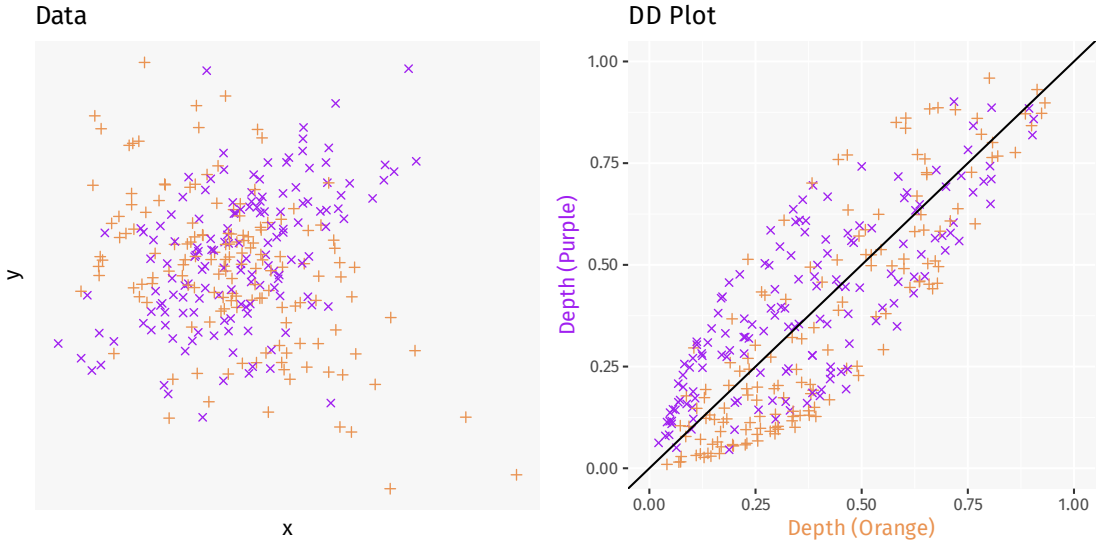
Liu et al. (1999) also demonstrate the use of DD plots to detect differences in skewness and kurtosis. This tool is especially convenient since the DD plot is always two dimensional regardless of the dimension d of the sample points.

2.4 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.



(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 2.4: 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.

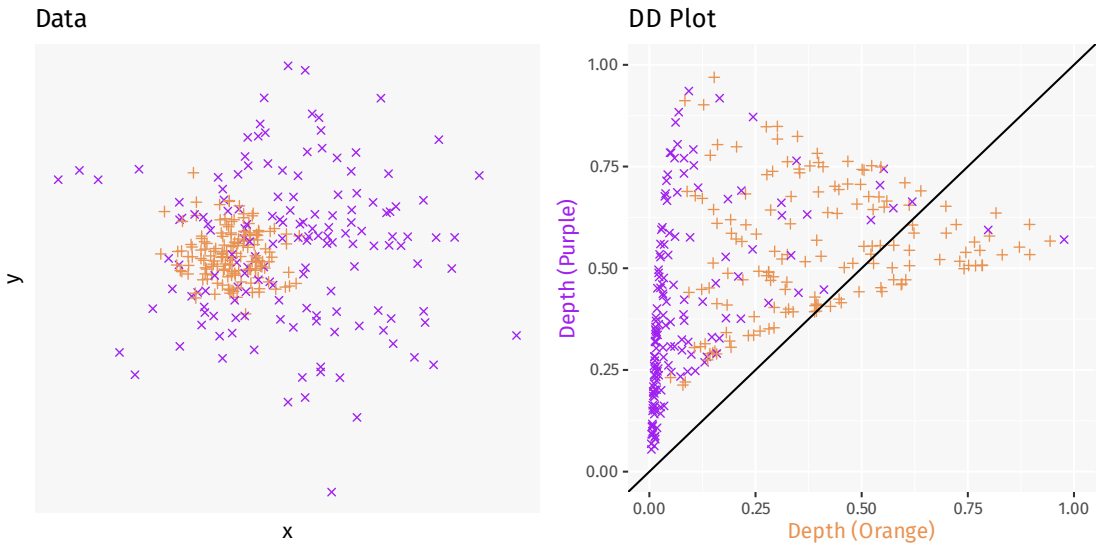


Figure 2.5: Empirical DD plot using spatial depth, where both underlying distributions (bivariate normal) differ in both location and scale. Observe that there is a clear separation between the orange and purple points in the DD plot, although not about the diagonal line.

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 2.4.1. *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 2.4.2. Denote

$$R(\boldsymbol{z}, F) = P_{\boldsymbol{X} \sim F}(D(\boldsymbol{X}, F) \leq D(\boldsymbol{z}, F)). \tag{2.4.1}$$

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 $\boldsymbol{z}$.

Proposition 2.4.3 (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 2.4.4. Denote the quality index

$$Q(F, G) = P(D(\boldsymbol{X}, F) \leq D(\boldsymbol{Y}, F) \mid \boldsymbol{X} \sim F, \boldsymbol{Y} \sim G). \tag{2.4.2}$$

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

$$Q(F, G) = \mathbb{E}_{\boldsymbol{Y} \sim G}[R(\boldsymbol{Y}, F)]. \tag{2.4.3}$$

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 $H_A : Q(F, G) < 1/2$. Specifically, we restrict our attention to elliptical distributions on $\mathbb{R}^d$.

Proposition 2.4.5 (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 2.4.6 (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*

$$G_\alpha = (1 - \alpha)F + \alpha G \tag{2.4.4}$$

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

$$R(\cdot, \hat{F}_{n_0}) = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{1}(D(\boldsymbol{Z}_i, \hat{F}_{n_0}) \leq D(\cdot, \hat{F}_{n_0})). \tag{2.4.5}$$

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 2.4.7 (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

$$Q(\hat{F}_n, \hat{G}_m) = \frac{1}{m} \sum_{j=1}^{m} R(\boldsymbol{Y}_j, \hat{F}_n) = \frac{1}{nm} \sum_{i,j} \mathbf{1}(D(\boldsymbol{X}_i, \hat{F}_n) \leq D(\boldsymbol{Y}_j, \hat{F}_n)). \tag{2.4.6}$$

This estimate is indeed consistent under mild assumptions.

Theorem 2.4.8 (Liu and Singh, 1993)**.** *Suppose that the distribution of* $D(\boldsymbol{Y}, F)$ *is continuous where* $\boldsymbol{Y} \sim G$*, and that*

$$\sup_{\boldsymbol{z} \in \mathbb{R}^d} |D(\boldsymbol{z}, \hat{F}_n) - D(\boldsymbol{z}, F)| \xrightarrow{a.s.} 0. \tag{2.4.7}$$

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 2.4.9 (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*

$$S(\hat{F}_n, \hat{G}_m) = \left[\frac{1}{12}\left(\frac{1}{n} + \frac{1}{m}\right)\right]^{-1/2} \left[Q(\hat{F}_n, \hat{G}_m) - \frac{1}{2}\right] \xrightarrow{d} \mathcal{N}(0, 1) \tag{2.4.8}$$

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

Later, Zuo and He (2006) show that under certain mild regularity conditions, the above asymptotic convergence can be extended to a broader class of depth functions, without the assumption $F = G$. They demonstrate that

$$\left[\frac{\sigma_{GF}^2}{n} + \frac{\sigma_{FG}^2}{m}\right]^{-1/2} \Big[Q(\hat{F}_n, \hat{G}_m) - Q(F, G)\Big] \xrightarrow{d} \mathcal{N}(0, 1), \tag{2.4.9}$$

where

$$\sigma_{FG}^2 = \int P_{\boldsymbol{X} \sim F}^2(D(\boldsymbol{X}, F) \leq D(\boldsymbol{y}, F))\, dG(\boldsymbol{y}) - Q^2(F, G), \tag{2.4.10}$$

$$\sigma_{GF}^2 = \int P_{\boldsymbol{Y} \sim G}^2(D(\boldsymbol{x}, F) \leq D(\boldsymbol{Y}, F))\, dF(\boldsymbol{x}) - Q^2(F, G). \tag{2.4.11}$$

Observe that given two samples, we have a choice between using $Q(\hat{F}_n, \hat{G}_m)$ or $Q(\hat{G}_m, \hat{F}_n)$. It may be advantageous to use the sample with a greater number of observations as the reference distribution. Shi et al. (2023) propose a weighted combination of the form

$$W_{n,m}^{\alpha} = \alpha S(\hat{F}_n, \hat{G}_m)^2 + (1 - \alpha) S(\hat{G}_m, \hat{F}_n)^2 \tag{2.4.12}$$

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

$$M_{n,m} = \max\{S(\hat{F}_n, \hat{G}_m)^2, S(\hat{G}_m, \hat{F}_n)^2\}. \tag{2.4.13}$$

Under similar assumptions, they show that both $W_{n,m}^{\alpha} \xrightarrow{d} \chi_1^2$ and $M_{n,m} \xrightarrow{d} \chi_1^2$ as $\min\{n, m\} \to \infty$ and n/m converges to a positive constant, under the null hypothesis $F = G$.

For multisample homogeneity testing, Chenouri and Small (2012) construct a Kruskal-Wallis-like test, with the role of univariate ranks replaced by the depth based ranks $R(\cdot, \hat{F})$ within the pooled data. The resulting test is shown to be powerful for both location and scale shifts.

2.5 Classification

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

Definition 2.5.1 (Classifier)**.** A classifier is a map $\hat{\iota}\colon \mathbb{R}^d \to \{1, \ldots, k\}$.

Example 2.5.2 (Bayes classifier)**.** Suppose that the population densities f_i for each $1 \leq i \leq k$ are known. The Bayes classifier assigns $\boldsymbol{x}$ to the $\hat{\iota}_B$-th population where

$$\hat{\iota}_B(\boldsymbol{x}) \;=\; \operatorname*{arg\,max}_{1 \leq i \leq k} \; \pi_i f_i(\boldsymbol{x}). \tag{2.5.1}$$

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

Definition 2.5.3 (Average misclassification rate)**.** The average misclassification rate of a classifier $\hat{\iota}$ is given by

$$\Delta(\hat{\iota}) \;=\; \sum_{i=1}^{k} \pi_i \, P_{\boldsymbol{X} \sim F_i}(\hat{\iota}(\boldsymbol{X}) \neq i). \tag{2.5.2}$$

Proposition 2.5.4. *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 2.5.5 (Maximum depth classifier)**.** Suppose that the prior probabilities π_i are equal. The maximum depth classifier $\hat{\iota}_D$ for a choice of depth function D is described by

$$\hat{\iota}_D(\boldsymbol{x}) \;=\; \operatorname*{arg\,max}_{1 \leq i \leq k} \; D(\boldsymbol{x}, F_i). \tag{2.5.3}$$

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

$$\mathscr{D} \;=\; \{(\boldsymbol{x}_{ij}, i)\} \subset \mathbb{R}^d \times \{1, \ldots, k\}, \tag{2.5.4}$$

where $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$ is an instance of an iid sample from F_i for each $1 \leq i \leq k$. The empirical maximum depth classifier simply replaces the population distributions F_i with their empirical counterparts $\hat{F}_i$ determined by $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$. Thus, it is given by

$$\hat{\iota}_D(\boldsymbol{x}) \;=\; \operatorname*{arg\,max}_{1 \leq i \leq k} \; D(\boldsymbol{x}, \hat{F}_i). \tag{2.5.5}$$

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

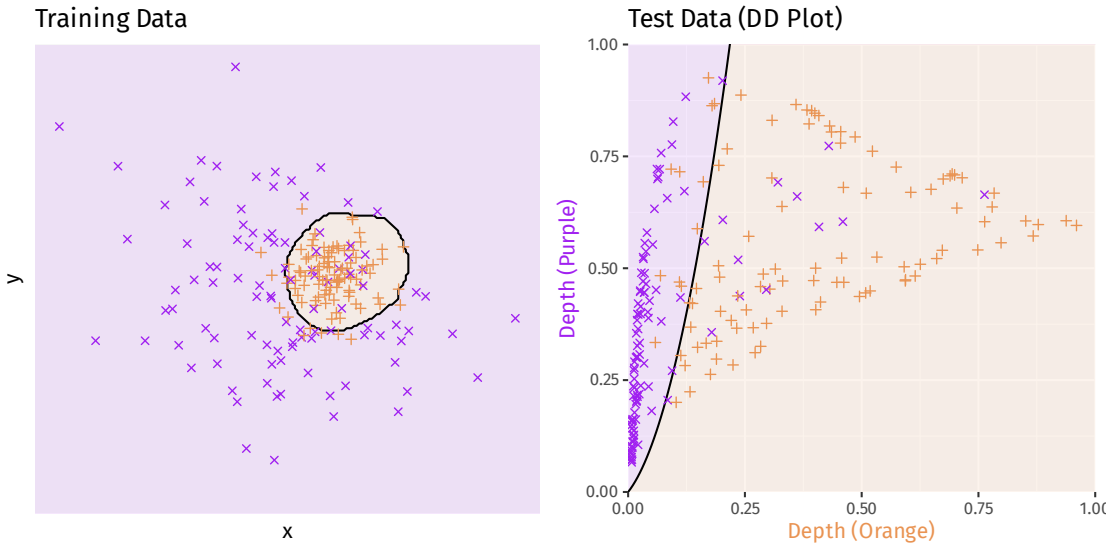


Figure 2.6: The DD classifier using spatial depth and a polynomial separating curve. The orange and purple shaded regions indicate the prediction rule learned from the training data; the black line marks the separating boundary. The classification accuracy here is 88%.

Theorem 2.5.6 (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 $g(k\boldsymbol{x}) \leq g(\boldsymbol{x})$ 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 $\min\{n_1, \ldots, n_k\} \to \infty$.*

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

$$f_i(\boldsymbol{x}) = c_i |\Sigma|^{-1/2} h_i\left((\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right) \tag{2.5.6}$$

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

$$\pi_i f_i(\boldsymbol{x}) > \pi_j f_j(\boldsymbol{x}) \iff D(\boldsymbol{x}, F_i) > r_{ij}(D(\boldsymbol{x}, F_j)) \tag{2.5.7}$$

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

$$\hat{\Delta}(\hat{\imath}) = \sum_{i=1}^{k} \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} \mathbf{1}(\hat{\imath}(\boldsymbol{x}_{ij}) \neq i). \tag{2.5.8}$$

Definition 2.5.8 (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

$$\hat{\imath}_{D,r}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } D(\boldsymbol{x}, F_2) \leq r(D(\boldsymbol{x}, F_1)), \\ 2, & \text{if } D(\boldsymbol{x}, F_2) > r(D(\boldsymbol{x}, F_1)). \end{cases} \tag{2.5.9}$$

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.

$$\hat{r} = \arg\min_{r \in \Gamma} \hat{\Delta}(\hat{\imath}_{D,r}). \tag{2.5.10}$$

Figure 2.6 shows the DD classifier applied on normal populations.

Remark. The maximum depth classifier $\hat{\imath}_D$ is simply the DD classifier $\hat{\imath}_{D,\mathrm{id}}$, where $\mathrm{id}(x) = x$. Figure 2.5 clearly illustrates how this choice of separating function may not always be appropriate, especially when the different populations have scale differences.

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 2.5.9. *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* $i \in \{1,2\}$*,*

$$\sup_{\boldsymbol{z} \in \mathbb{R}^d} |D(\boldsymbol{z}, \hat{F}_i) - D(\boldsymbol{z}, F_i)| \xrightarrow{a.s.} 0. \tag{2.5.11}$$

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

$$P_{\boldsymbol{Z} \sim F_i}(D(\boldsymbol{Z}, F_i) = \delta) = 0. \tag{2.5.12}$$

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 2.5.10. 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

$$\boldsymbol{x}^D = \left(D(\boldsymbol{x}, F_1), \ldots, D(\boldsymbol{x}, F_k)\right). \tag{2.5.13}$$

Remark. The graph

$$\mathrm{DD}(F_1, \ldots, F_k) = \{\boldsymbol{x}^D : \boldsymbol{x} \in \mathbb{R}^d\} \tag{2.5.14}$$

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

$$\hat{\imath}_D(\boldsymbol{x}) = i \iff \boldsymbol{x}^D \in R_i^D = \{\boldsymbol{y} \in [0,1]^k : y_i = \max_j y_j\}. \tag{2.5.15}$$

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

$$\mathscr{D}^D = \{(\boldsymbol{x}_{ij}^D, i)\} \subset [0,1]^k \times \{1, \ldots, k\} \tag{2.5.16}$$

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

2.6 Clustering

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

Definition 2.6.1 (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

$$\{I_1, \ldots, I_K\} \mapsto \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2, \tag{2.6.1}$$

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

$$\boldsymbol{\mu}_k = \frac{1}{|I_k|} \sum_{i \in I_k} \boldsymbol{x}_i = \arg\min_{\boldsymbol{\mu} \in \mathbb{R}^d} \sum_{i \in I_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2. \tag{2.6.2}$$

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 2.6.2. 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 2.6.3 (L_1-median)**.** The L_1-median of the k-th cluster is given by

$$\boldsymbol{\theta}_k = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \sum_{i\in I_k} \|\boldsymbol{x}_i - \boldsymbol{\theta}\|. \tag{2.6.3}$$

Definition 2.6.4. The between cluster depth of $\boldsymbol{x}_i$ is $D_i^b = D_\ell(\boldsymbol{x}_i)$, where

$$\ell = \underset{k\colon i\notin I_k}{\arg\min} \|\boldsymbol{x}_i - \boldsymbol{\theta}_k\|. \tag{2.6.4}$$

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

Definition 2.6.5 (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

$$\{I_1, \ldots, I_K\} \mapsto \frac{1}{N} \sum_{k=1}^{K} \sum_{i\in I_k} \mathrm{ReD}_i, \tag{2.6.5}$$

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

$$\bar{d}_k(\boldsymbol{z}) = \frac{1}{|\{i \in I_k\colon \boldsymbol{x}_i \neq \boldsymbol{z}\}|} \sum_{\substack{i\in I_k \\ \boldsymbol{x}_i\neq\boldsymbol{z}}} \|\boldsymbol{x}_i - \boldsymbol{z}\|. \tag{2.6.6}$$

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

$$\mathrm{Sil}_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \qquad a_i = \bar{d}_k(\boldsymbol{x}_i), \quad b_i = \min_{\ell\neq k} \bar{d}_\ell(\boldsymbol{x}_i). \tag{2.6.7}$$

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

$$\{I_1, \ldots, I_K\} \mapsto \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} (1 - \lambda) \operatorname{Sil}_i + \lambda \operatorname{ReD}_i \tag{2.6.8}$$

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.

Chapter 3

Functional Data

Consider a class of functions $\mathscr{X}$ of the form $\boldsymbol{x}\colon [0,1] \to \mathbb{R}^d$, equipped with a norm $\|\cdot\|$, and let $\mathscr{F}$ be a suitable class of distributions on $\mathscr{X}$. Typically, we choose $\mathscr{X}$ to be either $L_2[0,1]$ or $\mathcal{C}[0,1]$. It is desirable for a depth function $D\colon \mathscr{X} \times \mathscr{F} \to \mathbb{R}$ to satisfy the following properties (Gijbels & Nagy, 2017).

P0. *Non-degeneracy.* For $F \in \mathscr{F}$,

$$\inf_{\boldsymbol{x}\in\mathscr{X}} D(\boldsymbol{x}, F) < \sup_{\boldsymbol{x}\in\mathscr{X}} D(\boldsymbol{x}, F). \tag{3.0.1}$$

The property **P0** has been introduced to emphasize that generalizing multivariate depth to the functional setting requires some care; the natural functional analogue to halfspace/Tukey depth when $\mathscr{X}$ is a Banach space,

$$D_H(\boldsymbol{x}, F) = \inf_{\boldsymbol{v}\in\mathscr{X}^*} P_{\boldsymbol{X}\sim F}(\boldsymbol{v}(\boldsymbol{X}) \leq \boldsymbol{v}(\boldsymbol{x})), \tag{3.0.2}$$

turns out to be degenerate for a wide class of distributions $\mathscr{F}$ (Chakraborty & Chaudhuri, 2014). For example, when $\mathscr{X} = \mathcal{C}[0,1]$ with the supremum norm and $\boldsymbol{X}$ is a Gaussian process with a positive definite covariance kernel, we have $D_H(\,\cdot\,, F_{\boldsymbol{X}}) = 0$ almost surely. A similar result holds for the analogue to the projection depth. However, neither the functional random Tukey depth nor the functional spatial depth

$$D_S(\boldsymbol{x}, F_{\boldsymbol{X}}) = 1 - \left\| \mathbb{E}_{\boldsymbol{X}\sim F} \left[\frac{\boldsymbol{x} - \boldsymbol{X}}{\|\boldsymbol{x} - \boldsymbol{X}\|_2} \right] \right\|_2 \tag{3.0.3}$$

suffer this deficiency when $\mathscr{X} = L_2$ (Cuesta-Albertos & Nieto-Reyes, 2008; Gijbels & Nagy, 2017).

The remaining properties are analogues of the Zuo-Serfling properties for multivariate depth functions. First, the notion of affine invariance in **P1** can be generalized in many ways; Gijbels and Nagy (2017) recommend the following.

P1S. *Scalar-affine invariance.* For $a, b \in \mathbb{R}$ with a non-zero and $\boldsymbol{x} \in \mathscr{X}$,

$$D(a\boldsymbol{x} + b, F_{a\boldsymbol{X}+b}) = D(\boldsymbol{x}, F_{\boldsymbol{X}}). \tag{3.0.4}$$

P1F. *Function-affine invariance.* For $a, b, x \in \mathscr{X}$ with $ax \in \mathscr{X}$,

$$D(ax + b, F_{aX+b}) = D(x, F_X). \tag{3.0.5}$$

When generalizing **P2**, we must first define a notion of symmetry of $F \in \mathscr{F}$. To this end, we say that $F_{\boldsymbol{X}}$ is symmetric about $\boldsymbol{\theta} \in \mathscr{X}$ if for all $\varphi \in \mathscr{X}^*$, we have $\varphi(\boldsymbol{X})$ is symmetric about $\varphi(\boldsymbol{\theta})$. Again, we are free to choose our notion of univariate symmetry for $\varphi(\boldsymbol{X})$. Gijbels and Nagy (2017) consider central and halfspace symmetry.

P2C. *Maximality at center of central symmetry.* Any centrally symmetric $F \in \mathscr{F}$ is symmetric about $\boldsymbol{\theta} \in \mathscr{X}$ if and only if $D(\boldsymbol{\theta}, F) = \sup_{\boldsymbol{x} \in \mathscr{X}} D(\boldsymbol{x}, F)$.

P2H. *Maximality at center of halfspace symmetry.* Any halfspace symmetric $F \in \mathscr{F}$ is symmetric about $\boldsymbol{\theta} \in \mathscr{X}$ if and only if $D(\boldsymbol{\theta}, F) = \sup_{\boldsymbol{x} \in \mathscr{X}} D(\boldsymbol{x}, F)$.

Earlier, Nieto-Reyes and Battey (2016) proposed the following variant of **P2**.

P2G. *Maximality at Gaussian process mean.* For a zero-mean, stationary, almost surely continuous Gaussian process $F \in \mathscr{F}$, we have $D(\boldsymbol{\theta}, F) = \sup_{\boldsymbol{x} \in \mathscr{X}} D(\boldsymbol{x}, F)$ where θ is the zero mean function.

The above notions of maximality at the center are **P2H** > **P2C** > **P2G** in order of strength.

The properties **P3** and **P4** have straightforward generalizations.

P3D. *Monotonicity relative to deepest point.* For $F \in \mathscr{F}$ such that $D(\boldsymbol{\theta}, F) = \sup_{\boldsymbol{x} \in \mathscr{X}} D(\boldsymbol{x}, F)$, we have for $\alpha \in [0, 1]$,

$$D(\boldsymbol{x}, F) \leq D(\boldsymbol{\theta} + \alpha(\boldsymbol{x} - \boldsymbol{\theta}), F). \tag{3.0.6}$$

P4V. *Vanishing at infinity.* For any $F \in \mathscr{F}$,

$$D(\boldsymbol{x}, F) \to 0 \text{ as } \|\boldsymbol{x}\| \to \infty. \tag{3.0.7}$$

Nieto-Reyes and Battey (2016) and Gijbels and Nagy (2017) also deal with the notions of continuity in F. Let $d_{\mathscr{F}}$ metrize the topology of weak convergence in $\mathscr{F}$.

C2W. *Weak continuity in F.* For all $\epsilon > 0$ and $F \in \mathscr{F}$, there exists $\delta > 0$ such that for all $G \in \mathscr{F}$ such that $d_{\mathscr{F}}(F, G) < \delta$, we have $|D(\boldsymbol{x}, F) - D(\boldsymbol{x}, G)| < \epsilon$, F-almost surely.

C2U. *Uniform continuity in F.* For all $\epsilon > 0$ and $F \in \mathscr{F}$, there exists $\delta > 0$ such that for all $G \in \mathscr{F}$ such that $d_{\mathscr{F}}(F, G) < \delta$, we have $\sup_{\boldsymbol{x} \in \mathscr{X}} |D(\boldsymbol{x}, F) - D(\boldsymbol{x}, G)| < \epsilon$.

Gijbels and Nagy (2017, Table 1) provides a detailed summary of which of these properties are satisfied by the depth functions discussed in the following section.

3.1 Functional depth functions

3.1.1 Summary depths

Let D be a univariate or multivariate depth function. We can use this to define the depth of a curve $\boldsymbol{x}$ by first computing the multivariate D-depth of each time slice $\boldsymbol{x}(t)$, then 'summarizing' these depths over all $t \in [0, 1]$. One possibility is to take a simple or weighted time average, as in the integrated depth (Fraiman & Muniz, 2001).

Definition 3.1.1 (Fraiman-Muniz depth)**.** The integrated depth, or Fraiman-Muniz depth, is defined as

$$D_F(\boldsymbol{x}, F_{\boldsymbol{X}}) = \int_{[0,1]} D(\boldsymbol{x}(t), F_{\boldsymbol{X}(t)})\, w(t)\, dt. \tag{3.1.1}$$

Here, w is a weight function.

Alternatively, we may choose the lowest or 'worst' depth over time (Mosler, 2013). This way, low depth values over small portions of time, which indicate a deviation from centrality, are better reflected in the summary.

Definition 3.1.2 (Infimal depth)**.** The infimal depth is defined as

$$D_{Inf}(\boldsymbol{x}, F_{\boldsymbol{X}}) = \inf_{t \in [0,1]} D(\boldsymbol{x}(t), F_{\boldsymbol{X}(t)}). \tag{3.1.2}$$

Nagy et al. (2017), motivated by the problem of detecting *shape outliers*, extend the definitions of Fraiman-Muniz depth and infimal depth as follows. We will examine their significance briefly in Section 3.3.

Definition 3.1.3. The J-th order integrated depth is defined as

$$D_F^J(\boldsymbol{x}, F_{\boldsymbol{X}}) = \int_{[0,1]^J} D((\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_J))^\top, F_{(\boldsymbol{X}(t_1), \ldots, \boldsymbol{X}(t_J))^\top})\, w(\boldsymbol{t})\, d\boldsymbol{t}. \tag{3.1.3}$$

Definition 3.1.4. The J-th order infimal depth is defined as

$$D_{Inf}^J(\boldsymbol{x}, F_{\boldsymbol{X}}) = \inf_{\boldsymbol{t} \in [0,1]^J} D((\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_J))^\top, F_{(\boldsymbol{X}(t_1), \ldots, \boldsymbol{X}(t_J))^\top}. \tag{3.1.4}$$

Remark. It is often convenient to use Monte-Carlo approximations of the J-th order Fraiman-Muniz and infimal depths.

3.1.2 Band depths

López-Pintado and Romo (2009) later introduced the notion of band depth for univariate functional data.

Definition 3.1.5 (Band depth)**.** The band depth, for some index $J \geq 2$, is defined as

$$D_B^J(\boldsymbol{x}, F_{\boldsymbol{X}}) = \sum_{j=2}^{J} P_{\boldsymbol{X}_i \overset{\text{iid}}{\sim} F_{\boldsymbol{X}}}(\boldsymbol{x} \in \operatorname{conv}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j)). \tag{3.1.5}$$

The empirical version of band depth is defined as

$$D_B^J(\boldsymbol{x}, \hat{F}_n) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \leq i_1 < \cdots < i_j \leq n} \mathbf{1}(\boldsymbol{x} \in \operatorname{conv}(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_j})). \tag{3.1.6}$$

This is simply the proportion of j-tuples of curves (for $2 \leq j \leq J$) which envelope $\boldsymbol{x}$. Note that if two curves intersect at a point, a third curve is enveloped by them only when it passes through the point of intersection. For most commonly used $F_{\boldsymbol{X}}$, this happens with probability zero, making the band depth for $J = 2$ degenerate. Thus, we generally use $J = 3$.

Remark. The band depth may fail to satisfy **P0** even for $J \geq 3$. It follows from Chakraborty and Chaudhuri (2014, Theorem 3.2) that when $\mathscr{X} = \mathcal{C}[0,1]$ and $\boldsymbol{X}$ is a Feller process (for instance, Brownian motion) such that $P(X_0 = 0) = 1$ and the distribution of each $\boldsymbol{X}_t$ for $t \in (0,1]$ is non-atomic and symmetric about 0, the band depth $D_B^J(\,\cdot\,, F_{\boldsymbol{X}}) = 0$ almost surely. The following modification of the band depth resolves this issue.

Definition 3.1.6 (Modified band depth)**.** Define the enveloping time

$$\operatorname{ET}(\boldsymbol{x};\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_j) = m_1(\{t \in [0,1] \colon \boldsymbol{x}(t) \in \operatorname{conv}(\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_j(t))\}), \tag{3.1.7}$$

where m_1 is the Lebesgue measure on $\mathbb{R}$. The modified band depth is defined as

$$D_{MB}^J(\boldsymbol{x}, F_{\boldsymbol{X}}) = \sum_{j=2}^{J} \mathbb{E}_{\boldsymbol{X}_i \overset{\text{iid}}{\sim} F_{\boldsymbol{X}}} \left[\operatorname{ET}(\boldsymbol{x};\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_j)\right]. \tag{3.1.8}$$

The empirical version of modified band depth is defined as

$$D_{MB}^J(\boldsymbol{x}, \hat{F}_n) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \leq i_1 < \cdots < i_j \leq n} \operatorname{ET}(\boldsymbol{x};\, \boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_j}). \tag{3.1.9}$$

We generally use $J = 2$ for ease of computation, and denote the corresponding modified band depth simply as $D_{MB}(\,\cdot\,, \cdot\,)$, dropping the superscript.

$$D_{MB}(\boldsymbol{x}, \hat{F}_n) = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \operatorname{ET}(\boldsymbol{x}; \boldsymbol{x}_i, \boldsymbol{x}_j). \tag{3.1.10}$$

3.1.3 Half-region depths

Later, López-Pintado and Romo (2011) introduced the half-region depth.

Definition 3.1.7. We say that $\boldsymbol{y}$ is in the hypograph of $\boldsymbol{x}$, denoted, $\boldsymbol{y} \in H_{\boldsymbol{x}}$, if $\boldsymbol{y}(t) \leq \boldsymbol{x}(t)$ for all $t \in [0,1]$. Similarly, we say that $\boldsymbol{y}$ is in the epigraph of $\boldsymbol{x}$, denoted, $\boldsymbol{y} \in E_{\boldsymbol{x}}$, if $\boldsymbol{y}(t) \geq \boldsymbol{x}(t)$ for all $t \in [0,1]$.

Definition 3.1.8 (Half-region depth)**.** The half-region depth is defined as

$$D_{HR}(\boldsymbol{x}, F) = \min\{P_F(H_{\boldsymbol{x}}),\ P_F(E_{\boldsymbol{x}})\}. \tag{3.1.11}$$

The quantity $P_F(E_{\boldsymbol{x}})$ is called the epigraph index, which measures the proportion of curves that lie entirely above $\boldsymbol{x}$.

Remark. The half-region depth may also fail to satisfy **P0**, with the same counterexample used earlier for the degeneracy of the band depth (Chakraborty & Chaudhuri, 2014, Theorem 3.2).

Definition 3.1.9 (Modified half-region depth)**.** Denote the modified hypograph (MHI) and epigraph (MEI) indices

$$\mathrm{MHI}_F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{X}\sim F}[m_1(\{t \in [0,1] \colon \boldsymbol{x}(t) \geq \boldsymbol{X}(t)\})], \tag{3.1.12}$$
$$\mathrm{MEI}_F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{X}\sim F}[m_1(\{t \in [0,1] \colon \boldsymbol{x}(t) \leq \boldsymbol{X}(t)\})]. \tag{3.1.13}$$

The modified half-region depth is defined as

$$D_{MHR}(\boldsymbol{x}, F) = \min\{\mathrm{MHI}_F(\boldsymbol{x}),\ \mathrm{MEI}_F(\boldsymbol{x})\}. \tag{3.1.14}$$

3.2 Classification

Observe that the classification procedures for multivariate data described in Section 2.5 (the maximum depth classifier, the DD classifier, and the DD^G classifier) only depend on the data through the depth feature vectors

$$\boldsymbol{x}^D = (D(\boldsymbol{x}, F_1), \ldots, D(\boldsymbol{x}, F_k)) \in \mathbb{R}^k. \tag{3.2.1}$$

By simply choosing an appropriate functional data depth D, all of these classification procedures naturally generalize to the functional setting. The most flexible of these is the DD^G classifier (Cuesta-Albertos et al., 2017), which allows for any multivariate classification procedure on the transformed data $\mathscr{D}^D$.

[1] http://www-stat.stanford.edu/ElemStatLearn

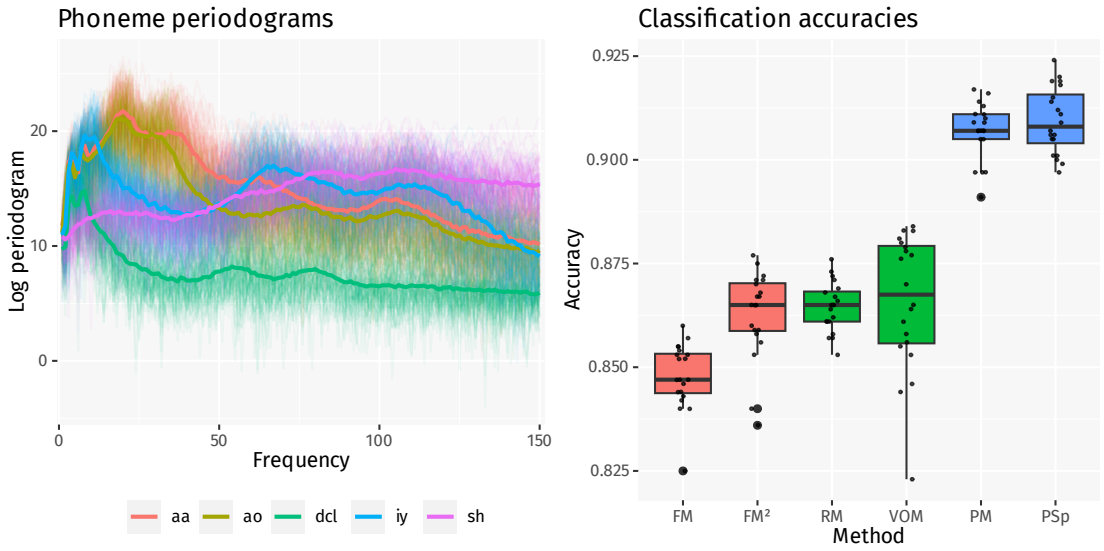


Figure 3.1: Classification of periodograms of digitized speech,[1] by phonemes ('aa', 'ao', 'dcl', 'iy', 'sh'). The thick colored lines in the plot on the left mark the median curve across 400 examples in each of the five groups. The boxplot shows classification accuracies for 20 runs of each of the following methods: maximum depth classification using the first order (FM) and second order (FM^2) Fraiman-Muniz depths in red, the RM and VOM classifiers in green, and the maximum depth Mahalanobis (PM) and spatial (PSp) depth classifiers on d-variate feature vectors obtained by $d = 10$ random projections in blue. In each run, 50% of the data was been aside for training.

3.2.1 Outlyingness matrices

Dai and Genton (2018) proposed a method which measures the outlyingness of $\boldsymbol{x}$ with respect to a population via depth as follows.

Definition 3.2.1. Let $\boldsymbol{X}$ be a d-variate stochastic process of continuous functions. At each time point $t \in [0, 1]$, the directional outlyingness is defined as

$$\boldsymbol{O}(t) = \boldsymbol{O}(\boldsymbol{X}(t), F_{\boldsymbol{X}(t)}) = \left(\frac{1}{D(\boldsymbol{X}(t), F_{\boldsymbol{X}(t)})} - 1 \right) \boldsymbol{v}(t), \tag{3.2.2}$$

where $\boldsymbol{v}(t)$ is the unit vector pointing from the median of $F_{\boldsymbol{X}(t)}$ to $\boldsymbol{X}(t)$.

Definition 3.2.2. The functional directional outlyingness is defined as

$$\mathrm{FO}(\boldsymbol{X}, F_{\boldsymbol{X}}) = \int_{[0,1]} \|\boldsymbol{O}(t)\|^2 \, w(t) \, dt. \tag{3.2.3}$$

Definition 3.2.3. The mean directional outlyingness is defined as

$$\mathbf{MO}(\boldsymbol{X}, F_{\boldsymbol{X}}) = \int_{[0,1]} \boldsymbol{O}(t) \, w(t) \, dt. \tag{3.2.4}$$

Definition 3.2.4. The variation of directional outlyingness is defined as

$$\mathrm{VO}(\boldsymbol{X}, F_{\boldsymbol{X}}) = \int_{[0,1]} \|\boldsymbol{O}(t) - \mathbf{MO}(t)\|^2 \, w(t) \, dt. \tag{3.2.5}$$

Here, w is a weight function on $[0,1]$. In our discussion, we set $w = 1$.

It is easily verified that

$$\mathrm{FO}^2 = \| \mathbf{MO} \|^2 + \mathrm{VO}\,. \tag{3.2.6}$$

Dai and Genton (2018) propose using the $(d+1)$-variate feature vectors

$$\boldsymbol{Y}(\boldsymbol{X}, F_{\boldsymbol{X}}) = (\mathbf{MO}^\top,\, \mathrm{VO})^\top \tag{3.2.7}$$

corresponding to the curve $\boldsymbol{X}$ for the purposes of classification. The **MO** gives a sense of how outlying the curve $\boldsymbol{X}$ is within $F_{\boldsymbol{X}}$ as a whole, while the VO measures the amount of variation in the outlyingness over time. Loosely speaking, **MO** is affected by the position, while VO is affected by the shape of $\boldsymbol{X}$ within $F_{\boldsymbol{X}}$. For instance, one may define the classifier

$$\hat{\imath}(\boldsymbol{X}) = \underset{1\le i\le k}{\arg\max}\, D'(\boldsymbol{Y}(\boldsymbol{X}, F_i), F_{\boldsymbol{Y}(\boldsymbol{X},F_i)}), \tag{3.2.8}$$

where D' is a multivariate depth function. This is simply a maximum depth classifier applied on the feature vectors $\boldsymbol{Y}$. When D' is chosen to be the robust Mahalanobis depth, we have the RM classifier

$$\hat{\imath}_{RM}(\boldsymbol{X}) = \underset{1\le i\le k}{\arg\max}\, D_{RM}(\boldsymbol{Y}(\boldsymbol{X}, F_i), F_{\boldsymbol{Y}(\boldsymbol{X},F_i)}). \tag{3.2.9}$$

Definition 3.2.5. The functional directional outlyingness matrix is defined as

$$\mathrm{FOM}(\boldsymbol{X}, F_{\boldsymbol{X}}) = \int_{[0,1]} \boldsymbol{O}(t)\, \boldsymbol{O}(t)^\top\, w(t)\, dt. \tag{3.2.10}$$

Definition 3.2.6. The functional directional outlyingness matrix is defined as

$$\mathrm{VOM}(\boldsymbol{X}, F_{\boldsymbol{X}}) = \int_{[0,1]} (\boldsymbol{O}(t) - \mathbf{MO}(t))\, (\boldsymbol{O}(t) - \mathbf{MO}(t))^\top\, w(t)\, dt. \tag{3.2.11}$$

Again, it is easily verified that

$$\mathrm{FOM} = \mathbf{MO}\ \mathbf{MO}^\top + \mathrm{VOM}, \tag{3.2.12}$$

and that

$$\mathrm{FO} = \mathrm{trace}(\mathrm{FOM}), \qquad \mathrm{VO} = \mathrm{trace}(\mathrm{VOM}). \tag{3.2.13}$$

We may also use the feature matrix VOM, or its matrix norm $\| \mathrm{VOM} \|_F$ corresponding to the curve $\boldsymbol{X}$ for the purposes of classification. Here, $\| \cdot \|_F$ denotes the Frobenius norm. For instance, a VOM based classifier may be defined as

$$\hat{\imath}_{\mathrm{VOM}}(\boldsymbol{X}) = \underset{1\le i\le k}{\arg\min} \| \mathrm{VOM}(\boldsymbol{X}, F_i)\|_F. \tag{3.2.14}$$

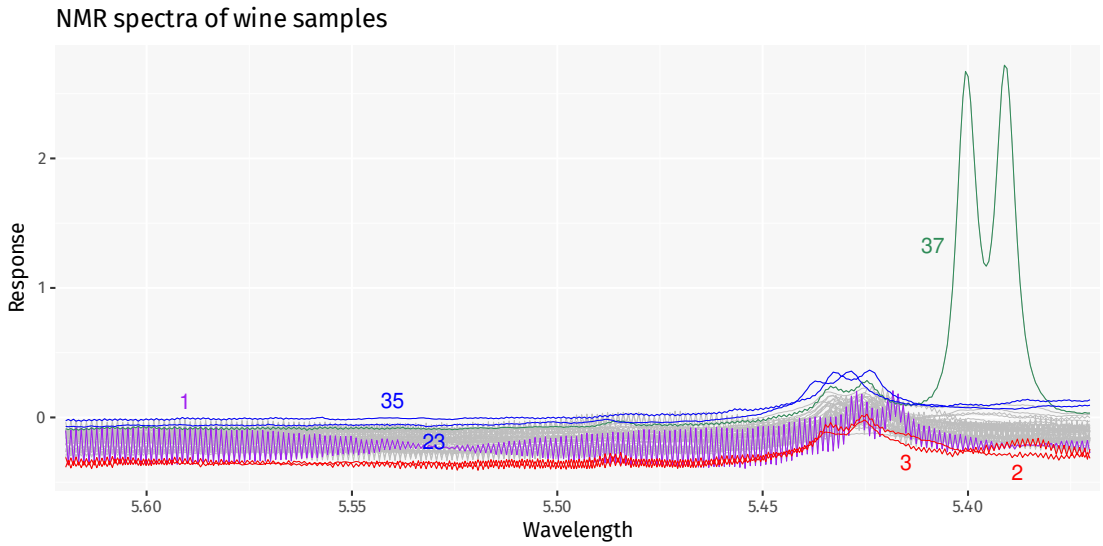


Figure 3.2: NMR spectra of 40 wine samples,[2] from the R package `speaq`, with some curves showing outlying behaviour highlighted. The green curve #37 is an isolated outlier, the blue and red curves are shift outliers, and the purple curve is a shape outlier.

3.2.2 Random projections

Another approach is to use a feature vector consisting of multiple projections of $\boldsymbol{X}$. Given functions $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ chosen at random, we examine the d-variate feature vectors

$$\boldsymbol{V}(\boldsymbol{X}, F_{\boldsymbol{X}}) = (\langle \boldsymbol{v}_1, \boldsymbol{X} \rangle, \ldots, \langle \boldsymbol{v}_d, \boldsymbol{X} \rangle) \tag{3.2.15}$$

and apply a depth based multivariate classifier. For instance, given a multivariate depth function D', we may define a classifier

$$\hat{\imath}^d_{D'}(\boldsymbol{X}) = \operatorname*{arg\,max}_{1 \le i \le k} D'(\boldsymbol{V}(\boldsymbol{X}, F_i), F_{\boldsymbol{V}(\boldsymbol{X}, F_i)}). \tag{3.2.16}$$

3.3 Outlier detection

A curve $\boldsymbol{x}\colon [0,1] \to \mathbb{R}$ may exhibit outlying behaviour with respect to a body of curves in many ways; we use the useful classification as detailed in Hubert et al. (2015). It may deviate significantly over a short interval, in which case we call it an *isolated outlier*. Alternatively, it may deviate over a large, or perhaps even the whole interval, in which case we call it a *persistent outlier*. If this deviation is in terms of shape – for instance, the curve may be rougher or smoother – we call it a *shape outlier*. Otherwise, if the curve has the same shape as the rest but appears above or below them, we call it a *shift outlier*. Another possibility is that the curve differs in scale, in which case we call it an *amplitude outlier*. Some of these behaviours have been illustrated in the dataset in Figure 3.2.

[2] https://ucphchemometrics.com/datasets/

A simple way of detecting shift outliers is using a *functional boxplot* (Sun & Genton, 2011), which is a natural analogue of the boxplot for univariate data. Here, curves are ranked according to their depths (say using the modified band depth), and the 50% central region is identified. A *fence* is created by inflating this central region envelope by a factor of 1.5; curves straying outside this fence are identified as outliers.

An important consideration when dealing with shape outliers is that each time slice $\boldsymbol{x}(t)$ may be fairly inconspicuous with respect to the marginal $F_{\boldsymbol{X}(t)}$. For instance, a shape outlier may be significantly more oscillatory than the rest, yet remain within the central region. This means that a tool like the functional boxplot may succeed in identifying shift or amplitude outliers, but fall short against shape outliers. In general, the basic algorithm of iteratively selecting curves with low functional depth as outliers is often insufficient.

A common approach towards examining the shapes of curves in a dataset is to bundle them with their derivatives. For instance, motion data often involves tracking both position and velocity over time. Thus, one may replace a curve x by $(x(\cdot), x'(\cdot))$ and examine its depth; the Fraiman-Muniz depth for such a curve would be of the form

$$D_F^{(2)}(x, F_X) = \int_{[0,1]} D((x(t), x'(t))^\top, F_{(X(t),X'(t))^\top})\, w(t)\, dt. \tag{3.3.1}$$

This naturally extends to $D_F^{(J)}$, taking derivatives of order $0, \ldots, J-1$. Nagy et al. (2017) point out several difficulties with this formulation, primarily the assumption of differentiability and the errors introduced when approximating derivatives. They make the notion of a J-th order shape outlier more precise as follows.

Definition 3.3.1. If there exists $\boldsymbol{t} \in [0,1]^J$ such that $(\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_J))^\top$ is outlying with respect to $F_{(\boldsymbol{X}(t_1),\ldots,\boldsymbol{X}(t_J))^\top}$, then we say that $\boldsymbol{x}$ is a J-th order outlier with respect to $F_{\boldsymbol{X}}$.

With this, the J-th order extension of Fraiman Muniz depth (Definition 3.1.3) is well equipped to detect J-th order outliers. Nagy et al. (2017) show that this depth D_F^J incorporates information about the J-th order derivatives of the curves (along with potentially additional information about its shape), which makes its performance comparable to or even better than $D_F^{(J)}$. They supply a simple algorithm based on D_F^J values for identifying J-th order outliers. A similar argument can be made for the J-th order extension of the infimal depth (Definition 3.1.4).

3.3.1 Outliergrams

Arribas-Gil and Romo (2014) combined the notions of the modified epigraph index (MEI) and the modified band depth (MBD), proposing the outliergram as a tool for detecting shape outliers. They show that for a sample $\{\boldsymbol{x}_i\}_{i=1}^n$, each

$$\mathrm{MBD}(\boldsymbol{x}_i) = D_{MB}(\boldsymbol{x}_i, \hat{F}_n) \leq a_0 + a_1\, \mathrm{MEI}(\boldsymbol{x}_i) + a_2 n^2\, \mathrm{MEI}(\boldsymbol{x}_i)^2 \tag{3.3.2}$$

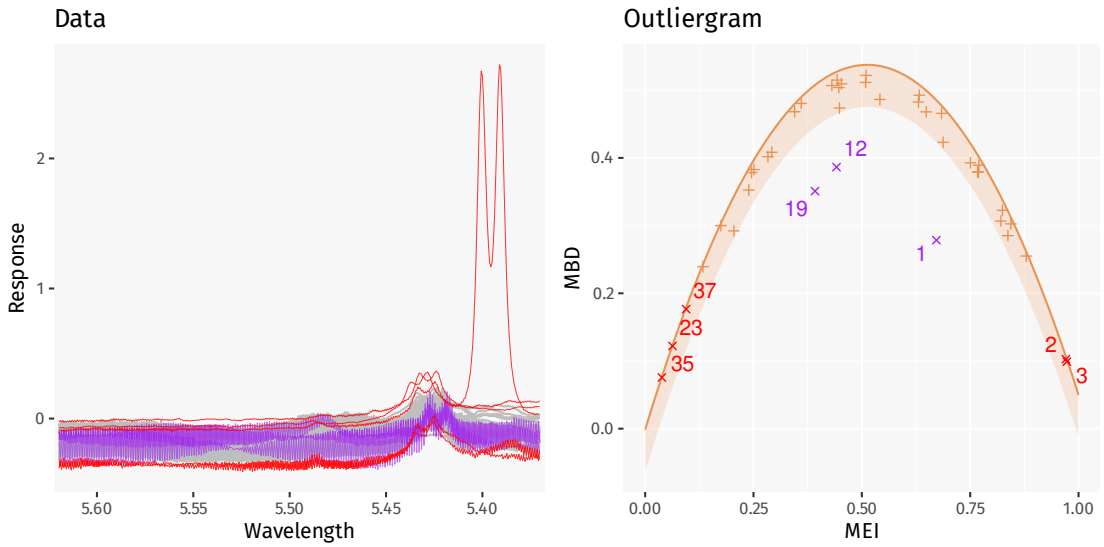


Figure 3.3: Outliergram for the NMR spectra of 40 wine samples. The three purple curves have been identified as shape outliers, as they fall outside the orange ribbon in the outliergram. Although the red curves lie on the orange parabola, they have low MBD and extreme MEI values, indicating that they lie above or below the main mass of curves.

where $a_0 = a_2 = -2/n(n-1)$ and $a_1 = 2(n+1)/(n-1)$. The distance

$$d_i = a_0 + a_1 \,\mathrm{MEI}(\boldsymbol{x}_i) + a_2 n^2 \,\mathrm{MEI}(\boldsymbol{x}_i) - \mathrm{MBD}(\boldsymbol{x}_i) \tag{3.3.3}$$

is indicative of the outlyingness of $\boldsymbol{x}_i$. Arribas-Gil and Romo (2014) consider shape outlying curves as those for which $d_i \geq d^* = Q_3 + 1.5\,\mathrm{IQR}$, where Q_3 and IQR are the third quartile and the interquartile range of $\{d_i\}_{i=1}^n$ respectively.

Definition 3.3.2 (Outliergram)**.** The outliergram for a dataset $\mathscr{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ is the graph

$$\{(\mathrm{MEI}(\boldsymbol{x}_i, \hat{F}_n),\, \mathrm{MBD}(\boldsymbol{x}_i, \hat{F}_n)) \colon 1 \leq i \leq n\}. \tag{3.3.4}$$

Shape outliers are curves $\boldsymbol{x}_i$ such that $(\mathrm{MEI}_i, \mathrm{MBD}_i)$ falls outside a ribbon of height d^* under the parabola $a_0 + a_1\,\mathrm{MEI} + a_2 n^2\,\mathrm{MEI}^2$.

Figure 3.3 illustrates the use of the outliergram. We have also highlighted curves with fairly low or high MEI values as shift outliers; a low MEI value indicates that the curve lies above the main mass of curves, and a high MEI indicates that it lies below.

3.3.2 Centrality-Stability diagrams

In their discussion of methods of functional outlier detection, Hubert et al. (2015) proposed the centrality-stability diagram, where both the 'centrality' of a curve (measured by depth) and its variability in cross-sectional outlyingness over time

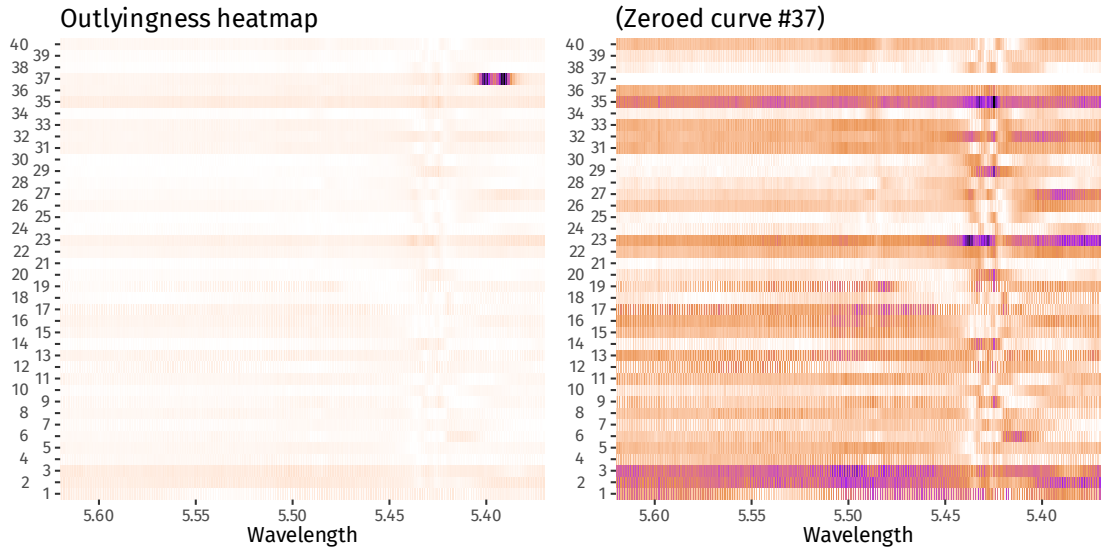


Figure 3.4: Outlyingness heatmap for the NMR spectra of 40 wine samples. The extreme curve #37 has been zeroed out in the second diagram to better illustrate the variation in outlyingness for the remaining curves.

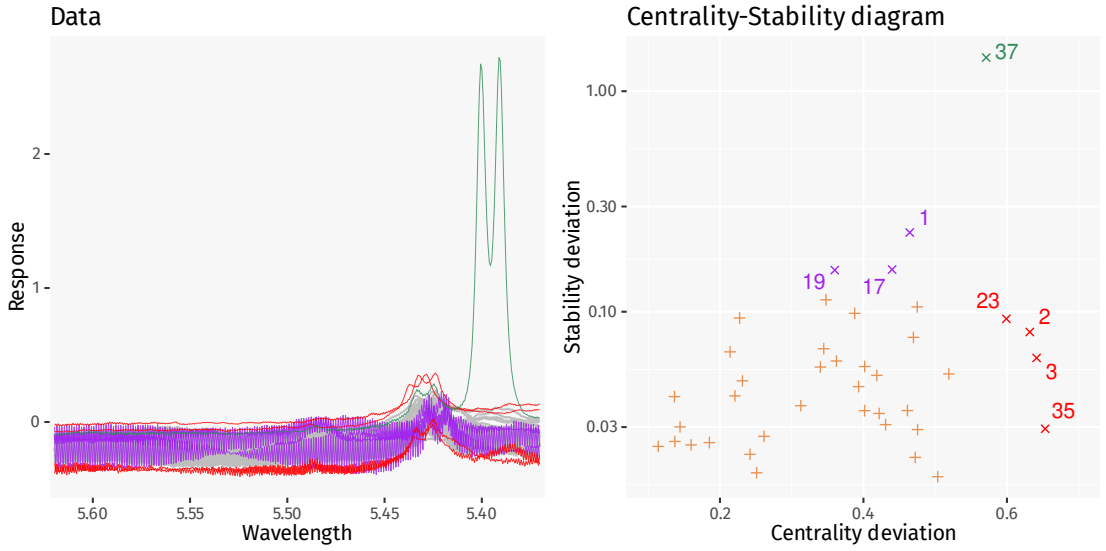


Figure 3.5: Centrality-Stability diagram for the NMR spectra of 40 wine samples. The red curves are seen to deviate in terms of centrality, indicated by the fact that the corresponding points in the centrality-stability diagram fall towards the right. The purple curves deviate in terms of stability, with the green curve showing extreme deviation.

are accounted for. A deviation in centrality may point towards a shift outlier, while a deviation in stability may point towards an isolated or shape outlier.

Hubert et al. (2015) begin by choosing a multivariate depth function of the form $D'(\boldsymbol{x}(t), F_{\boldsymbol{X}(t)}) = (1 + O(\boldsymbol{x}(t)))^{-1}$, where $O(\,\cdot\,)$ is an outlyingness function. The Mahalanobis, projection, and Oja depths clearly fit this description. Here, for the purposes of computation in the univariate case, we choose

$$O(x(t)) = \frac{|x(t) - \text{med}(X(t)|}{\text{MAD}(X(t))}, \tag{3.3.5}$$

instead of using the skew-adjusted version; the differences are minor enough for us to ignore. The variation in $O(\boldsymbol{x}(\,\cdot\,))$ over time for different curves, in the form of an *outlyingess heatmap*, is quite revealing; Figure 3.4 shows that curves may have large outlyingness for short or long intervals.

Corresponding to D', we have an integrated Fraiman-Muniz depth

$$D_F(\boldsymbol{x}, F) = \int_{[0,1]} (1 + O(\boldsymbol{x}(t)))^{-1}\, dt. \tag{3.3.6}$$

However, a spike in outlyingess over a short time interval, such as in curve #37 in Figure 3.4, may potentially be 'washed out' in this averaging. With this, we seek a method of detecting sharp bursts in outlyingness. Note that by setting

$$\widetilde{\text{MO}}(\boldsymbol{x}, F) = \int_{[0,1]} O(\boldsymbol{x}(t))\, dt, \tag{3.3.7}$$

Cauchy-Schwarz gives us the relation

$$D_F(\boldsymbol{x}, F) \cdot (1 + \widetilde{\text{MO}}(\boldsymbol{x}, F)) \geq 1. \tag{3.3.8}$$

Equality is achieved only when $O(\boldsymbol{x}(\,\cdot\,))$ remains constant over time. Any sudden variation in outlyingness over time will be detected by the *stability deviation*

$$\Delta S(\boldsymbol{x}, F) \;=\; (1 + \widetilde{\text{MO}}(\boldsymbol{x}, F)) - \frac{1}{D_F(\boldsymbol{x}, F)}, \tag{3.3.9}$$

the difference between the arithmetic and harmonic means of $1+O(\boldsymbol{x}(\,\cdot\,))$. Defining the *centrality deviation* simply as $\Delta C(\boldsymbol{x}, F) = 1 - D_F(\boldsymbol{x}, F)$, we have our *centrality-stability diagram*.

Definition 3.3.3 (Centrality-Stability diagram)**.** The centrality-stability diagram for a dataset $\mathscr{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ is the graph

$$\{(\Delta C(\boldsymbol{x}_i, \hat{F}_n),\, \Delta S(\boldsymbol{x}_i, \hat{F}_n)) \colon 1 \leq i \leq n\}. \tag{3.3.10}$$

Remark. We make a distinction between **MO** from Definition 3.2.3 and $\widetilde{\text{MO}}$; the outlyingness $O(\,\cdot\,)$ used in the latter is real and positive.

Figure 3.5 illustrates the use of the centrality-stability diagram as a summary of the outlyingness heatmap from Figure 3.4. This time, the isolated outlier curve #37 is well separated from the shift and shape outliers, unlike in the outliergram in Figure 3.3.

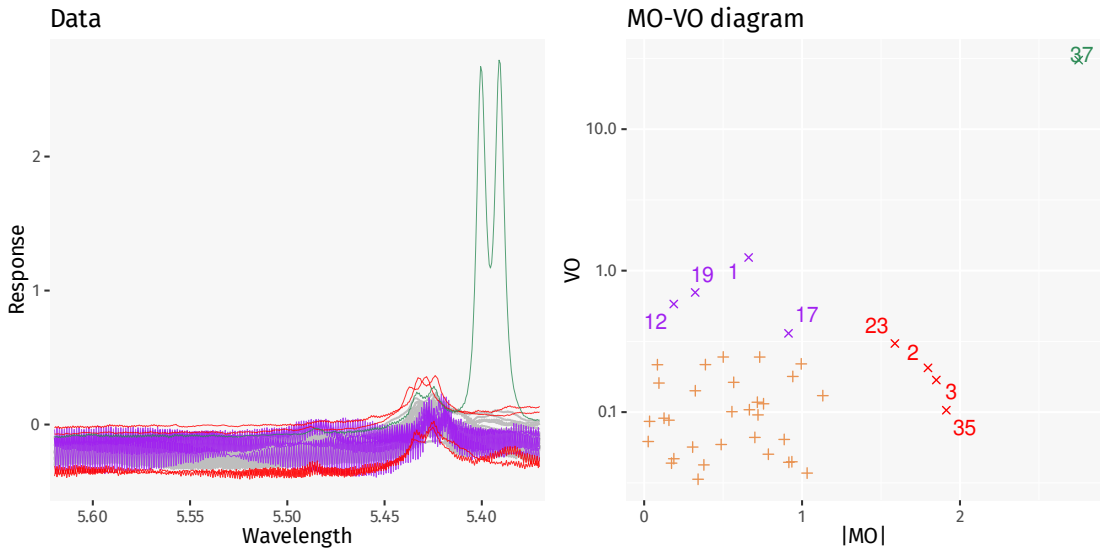


Figure 3.6: MO-VO diagram for the NMR spectra of 40 wine samples. We plot $|\mathrm{MO}|$ rather than MO here for better comparison with the centrality stability diagram in Figure 3.5; nevertheless, the signed MO values would reveal whether the shift outliers lie above or below the main mass of curves.

3.3.3 MO-VO diagrams

We observe that the MO-VO diagram from Dai and Genton (2018) neatly falls under a general category of centrality-stability diagrams. The quantity $\mathbf{MO}(\boldsymbol{x}, F)$ may indeed be treated as a measure of deviation from centrality of $\boldsymbol{x}$. Again, $\mathrm{VO}(\boldsymbol{x}, F)$ being the variance of $\boldsymbol{O}(t)$, may be treated as a measure of deviation from stability, since it captures the variability of outlyingness over time and is sensitive to changes over short intervals.

Definition 3.3.4 (MO-VO diagram)**.** The MO-VO diagram for a dataset $\mathscr{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ is the graph

$$\{(\mathbf{MO}(\boldsymbol{x}_i, \hat{F}_n),\ \mathrm{VO}(\boldsymbol{x}_i, \hat{F}_n)) \colon 1 \leq i \leq n\}. \tag{3.3.11}$$

For the purposes of computation in the univariate case, we use the directional outlyingess function

$$O(x(t)) = \frac{x(t) - \mathrm{med}(X(t))}{\mathrm{MAD}(X(t))}. \tag{3.3.12}$$

Figure 3.6 illustrates the use of the MO-VO diagram. Note the similarities with the centrality-stability diagram from Figure 3.5.

We demonstrate all of the above diagnostic tools on a different dataset in Figures 3.8 and 3.7.

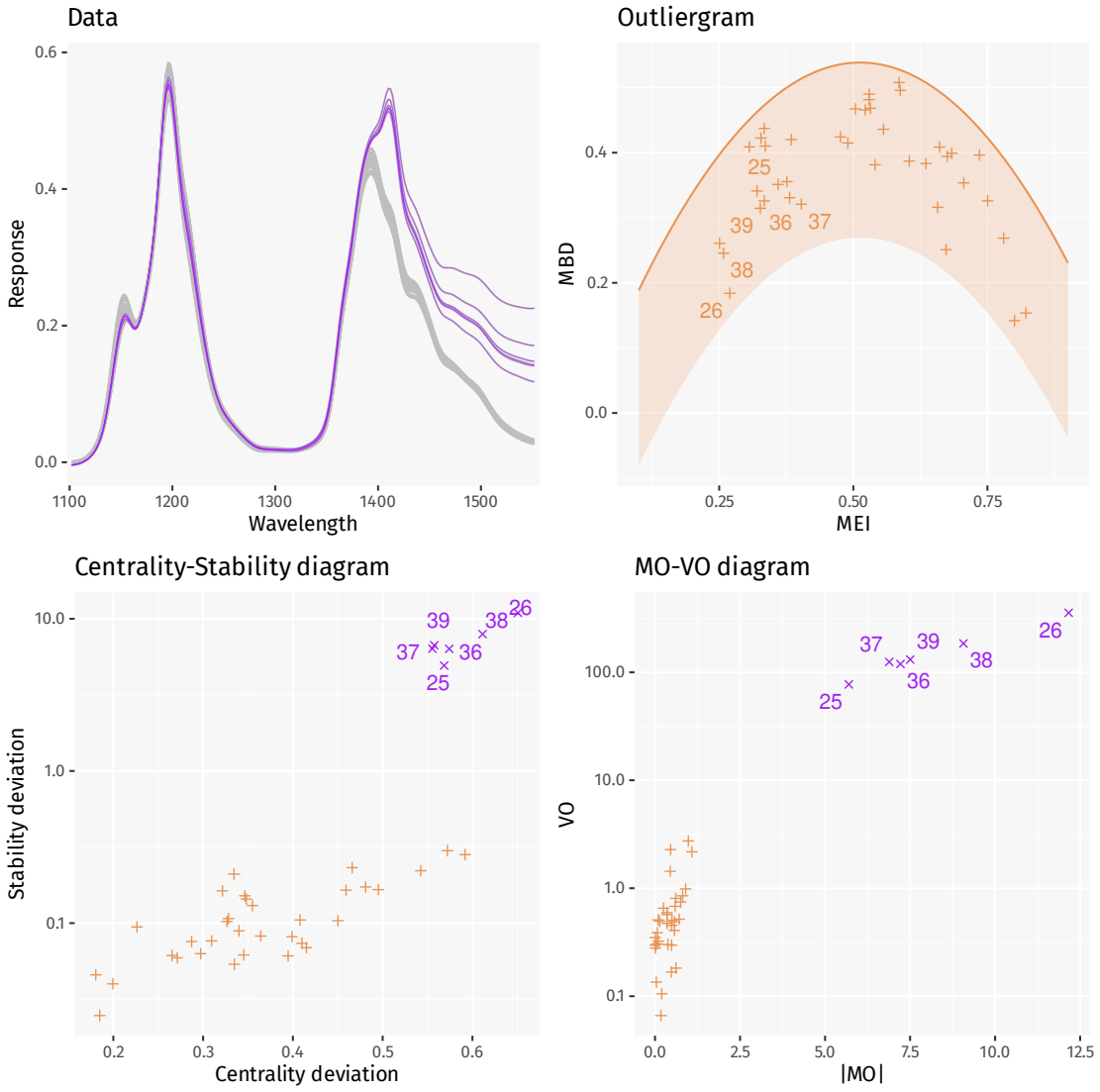


Figure 3.7: Outliergram, centrality-stability, and MO-VO diagrams for the NIR spectra of 39 gasoline samples, from the R package `rrcov`. The six purple curves #25, 26, 36-39 correspond to samples containing added alcohol. While the outliergram does not clearly identify these outliers, the centrality-stability and MO-VO diagrams show a marked separation from the main curves. Indeed, there is no cutoff d^* defining the lower boundary of the orange ribbon in the outliergram which properly excludes the six outliers.

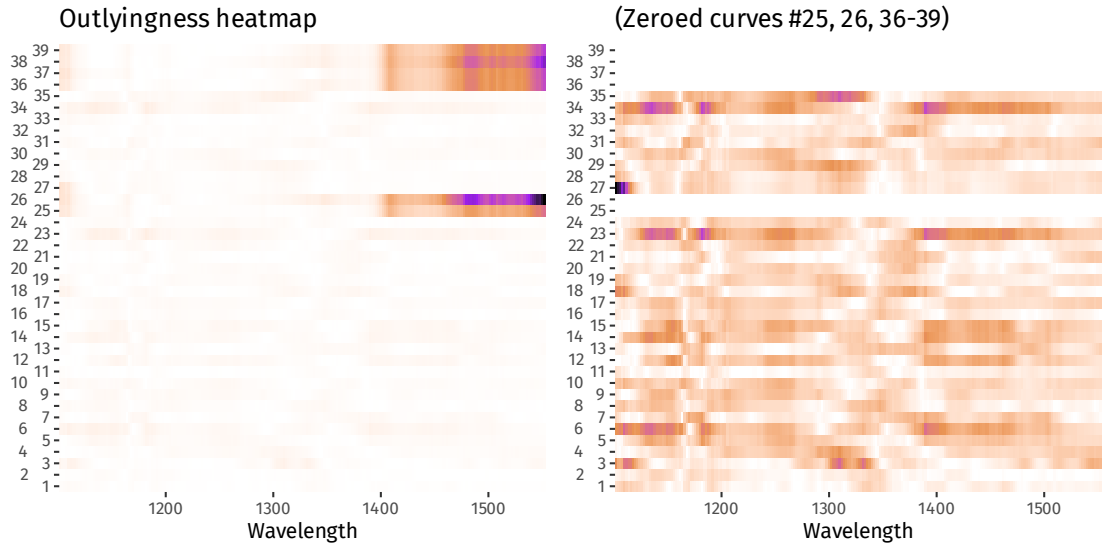


Figure 3.8: Outlyingness heatmap for the NIR spectra of 39 gasoline samples. The outlying curves have been zeroed out in the second diagram.

3.4 Partially observed functional data

Consider the setting where the stochastic process $\boldsymbol{X}$ of continuous functions is not observed on the entire interval $[0,1]$, but rather on a random subset $O \subseteq [0,1]$. Then, a dataset of partially observed curves is of the form $\mathscr{D} = \{(\boldsymbol{X}_i, O_i)\}_{i=1}^n$, where $\boldsymbol{X}_i \overset{\text{iid}}{\sim} F_{\boldsymbol{X}}$, $O_i \overset{\text{iid}}{\sim} Q$ where Q generates random compact subsets of $[0,1]$, independent of $\boldsymbol{X}_i$. In other words, $(\boldsymbol{X}_i, O_i) \overset{\text{iid}}{\sim} F_{\boldsymbol{X}} \times Q$. This setup is known as the 'missing completely at random' assumption (Kraus, 2015).

We set $\mathscr{J}(t) = \{j : t \in O_j\}$ to keep track of which curves $\boldsymbol{X}_i$ have been observed at time t. Furthermore, denote $q(t) = |\mathscr{J}(t)|$ as the number of curves $\boldsymbol{X}_i$ observed at time t.

Elías, Jiménez, Paganoni, and Sangalli (2023) propose the following modification of the Fraiman-Muniz depth for partially observed data.

Definition 3.4.1 (Partially observed integrated functional depth)**.** Let D be a d-variate depth function. The Partially Observed Integrated Functional Depth (POIFD) is defined as

$$D_{POIFD}((\boldsymbol{x}, o), F_{\boldsymbol{X}} \times Q) = \int_o D(\boldsymbol{x}(t), F_{\boldsymbol{X}(t)})\, w_o(t)\, dt, \tag{3.4.1}$$

where $w_o(t) = q(t) / \int_o q(t)\, dt$.

We can now proceed with tasks such as classification, outlier detection, etc. on our partially observed dataset, via depth based procedures using POIFD values. Another natural problem is one of curve reconstruction: given a partially observed

curve $(\boldsymbol{X}, O)$, can we estimate $\boldsymbol{X}$ on $M = [0,1] \setminus O$? For instance, we may search for a reconstruction operator $\mathcal{R}\colon L_2(O) \to L_2(M)$ that minimizes the mean integrated prediction squared error $\mathbb{E}\left[\|\boldsymbol{X}_M - \mathcal{R}(\boldsymbol{X}_O)\|^2\right]$. Here, $\boldsymbol{X}_O$ denotes the curve $\boldsymbol{X}$ restricted to O, and similarly for $\boldsymbol{X}_M$. The best predictor in this sense is the conditional expectation $\mathbb{E}[\boldsymbol{X}_M|\boldsymbol{X}_O]$, which is in general a non-linear operator. Thus, Kraus (2015) and Kneip and Liebl (2020) search for continuous linear operators $\mathcal{A}$, using methods based on estimating terms of the Karhunen-Loéve expansion of $\boldsymbol{X}$.

Elías, Jiménez, and Shang (2023) offer a depth based solution to the reconstruction problem, adapted from a similar algorithm for time-series forecasting (Elías et al., 2022). The main idea involves selecting a collection of curves, with indices $\mathscr{I}$, which best envelope $(\boldsymbol{X}, O)$, then taking a weighted linear combination. In particular, they suggest

$$\hat{\boldsymbol{X}}(t) = \frac{\sum_{i\in\mathscr{I}(t)} w_i \boldsymbol{X}_i(t)}{\sum_{i\in\mathscr{I}(t)} w_i}, \qquad w_i = \exp\left(-\theta\, \frac{\|(\boldsymbol{X}, O) - (\boldsymbol{X}_i, O_i)\|}{\delta}\right), \tag{3.4.2}$$

where $\mathscr{I}(t) = \mathscr{I} \cap \mathscr{J}(t) = \{i \in \mathscr{I} : t \in O_i\}$ and $\delta = \min_{i\in\mathscr{I}} \|(\boldsymbol{X}, O) - (\boldsymbol{X}_i, O_i)\|$. Here, θ is a tuning parameter, perhaps chosen by minimizing the mean squared error on $(\boldsymbol{X}, O)$. Furthermore, we have denoted

$$\|(\boldsymbol{X}, O) - (\boldsymbol{X}', O')\| = \frac{1}{m(O \cap O')} \left(\int_{O\cap O'} \|\boldsymbol{X}(t) - \boldsymbol{X}'(t)\|^2\, dt\right)^{1/2}. \tag{3.4.3}$$

Choosing the best envelope $\mathscr{I}$ involves both depth and distance. Elías, Jiménez, and Shang (2023) use the following three criteria to devise an algorithm that iteratively selects $\mathscr{I}$.

1. $(\boldsymbol{X}, O)$ should be as deep as possible in the collection of curves $\{(\boldsymbol{X}, O)\} \cup \{(\boldsymbol{X}_i, O_i)\}_{i\in\mathscr{I}}$, in the sense of POIFD.
2. $(\boldsymbol{X}, O)$ should be enveloped by $\{(\boldsymbol{X}_i, O_i)\}_{i\in\mathscr{I}}$ as much as possible, i.e. we want to maximize the enveloping time $\mathrm{ET}((\boldsymbol{X}, O); \{\boldsymbol{X}_i, O_i\}_{i\in\mathscr{I}})$.
3. $\{(\boldsymbol{X}_i, O_i)\}_{i\in\mathscr{I}}$ should contain as many near curves to $(\boldsymbol{X}, O)$ as possible, in the sense of the distance 3.4.3.

Chapter 4

Local Depth

The depth functions which we've encountered so far are well suited for unimodal, symmetric distributions. For instance, depth functions like the halfspace and Mahalanobis depths behave especially well with elliptical distributions: the depth and density contours coincide. These depth functions always produce convex, nested contours; the Zuo-Serfling property **P3** also forces star-shaped central regions. As a result, they may fail to capture the structure of distributions which are not convexly supported (Figure 4.1), or those with multiple modes (Figure 4.2).

Remedying this requires moving away from the *global* structure to *local* structures of distributions. One formulation of such a *local depth* is due to Agostinelli and Romanazzi (2011), where the halfspace and simplicial depths are modified so as to capture only local information around a given point $\boldsymbol{x} \in \mathbb{R}^d$. The local halfspace depth is obtained not by considering halfspaces containing $\boldsymbol{x}$ as in 2.1.1, but rather cuboidal slabs; similarly, the local simplicial depth is defined by restricting the volumes of the random simplices under consideration in 2.1.8.

In this chapter, we focus on the definition of *local depth* and *local depth neighbourhoods* proposed by Paindaveine and Van Bever (2013). This allows the construction of a local counterpart of any existing global depth function (multivariate, functional, or otherwise), giving it a distinct advantage over the previous treatment.

4.1 Local depth regions

Given a distribution $F_{\boldsymbol{X}}$, we may define a symmetrized distribution about a point $\boldsymbol{x} \in \mathscr{X}$ as

$$F_{\boldsymbol{X}}^{\boldsymbol{x}} = \frac{1}{2}F_{\boldsymbol{X}} + \frac{1}{2}F_{2\boldsymbol{x}-\boldsymbol{X}}. \tag{4.1.1}$$

With this, $\boldsymbol{x}$ becomes the point of central symmetry, hence the deepest point in $F_{\boldsymbol{X}}^{\boldsymbol{x}}$ with respect to a depth function D that obeys **P2**. Thus, the β-th central regions of $F_{\boldsymbol{X}}^{\boldsymbol{x}}$ behave like neighbourhoods of $\boldsymbol{x}$.

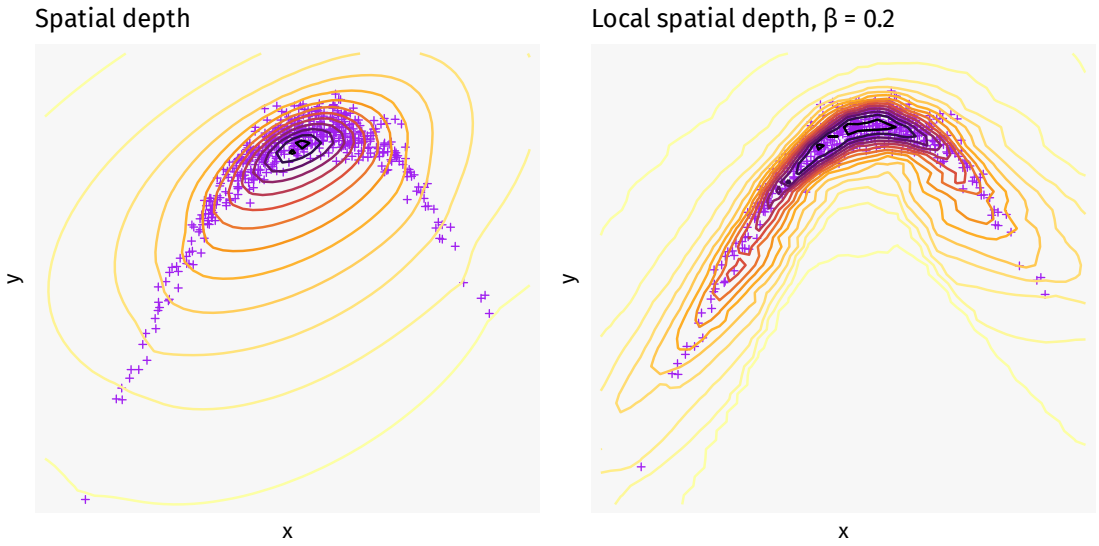


Figure 4.1: Depth contours with respect to a 'banana-shaped' distribution. Observe that the spatial depth contours fail to adequately capture the curved shape of the data cloud, in contrast with the local spatial depth (with $\beta = 0.2$) contours.

Definition 4.1.1 (Paindaveine and Van Bever, 2013)**.** The probability-β depth-based neighbourhood of $\boldsymbol{x}$ with respect to the distribution F is defined as

$$N_\beta^{\boldsymbol{x}}(F) = C_{F^{\boldsymbol{x}}}(\beta), \tag{4.1.2}$$

i.e. the β-th central region of F symmetrized about $\boldsymbol{x}$.

When working with a sample $\mathscr{D} = \{\boldsymbol{X}_i\}_{i=1}^n$ from F, we may obtain the β depth-based neighbourhood of $\boldsymbol{x}$ by first computing the reflected sample $\mathscr{D}' = \{2\boldsymbol{x} - \boldsymbol{X}_i\}_{i=1}^n$, then arranging the elements of the symmetrized sample $\mathscr{D}^{\boldsymbol{x}} = \mathscr{D} \cup \mathscr{D}'$ in descending order by their empirical depth values and choosing the first β proportion of elements. The neighbourhood $N_\beta^{\boldsymbol{x}}(\hat{F}_n)$ is the convex hull of these elements.

Definition 4.1.2 (Paindaveine and Van Bever, 2013)**.** Let D be a depth function, and let $F_\beta^{\boldsymbol{x}}$ denote the distribution F conditional on the neighbourhood $N_\beta^{\boldsymbol{x}}(F)$. The corresponding local depth function at locality level $\beta \in (0, 1]$ is defined as

$$LD_\beta(\boldsymbol{x}, F) = D(\boldsymbol{x}, F_\beta^{\boldsymbol{x}}) \tag{4.1.3}$$

Again, when working with a sample $\mathscr{D} = \{\boldsymbol{X}_i\}_{i=1}^n$, we obtain $LD_\beta(\boldsymbol{x}, \hat{F}_n)$ by arranging the elements of $\mathscr{D}$ in descending order by their empirical depth values in the symmetrized sample $\mathscr{D}^{\boldsymbol{x}}$, choosing the first β proportion of elements, and computing the depth of $\boldsymbol{x}$ with respect to these elements.

Remark. When $\beta = 1$, the local depth LD_1 reduces to the original global depth D.

Remark. The notions of depth based neighbourhoods and local depth make sense for any distribution F on a space $\mathscr{X}$ as long as the process of symmetrization around $\boldsymbol{x} \in \mathscr{X}$ can be achieved.

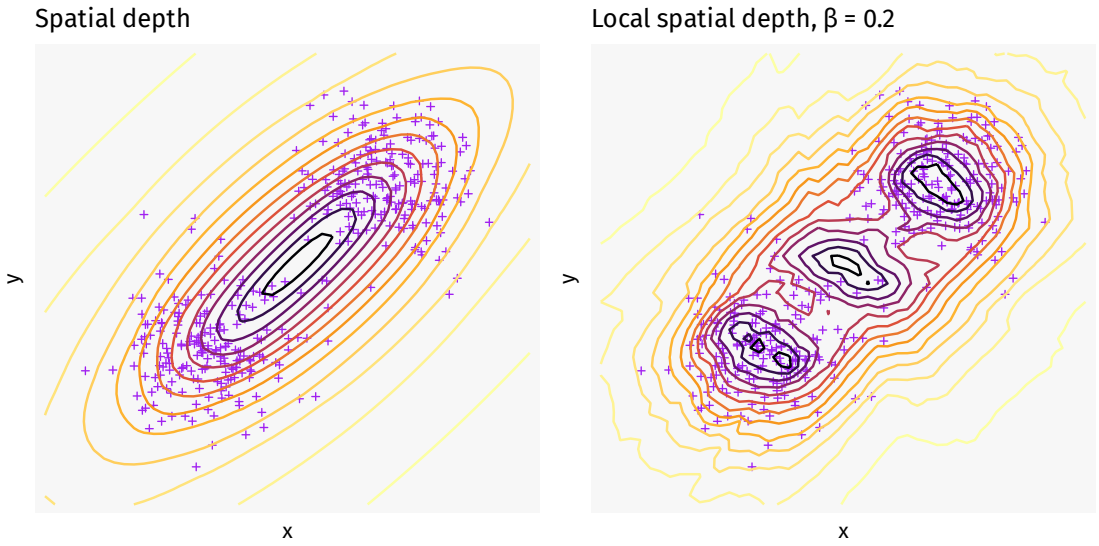


Figure 4.2: Depth contours with respect to a bimodal distribution. Although the local spatial depth contours capture the two modes correctly, it erroneously ascribes high depth values to a region in between them.

4.2 Regression based on local depth

Definition 4.2.1. Let D be a depth function, and let $\widetilde{F}_\beta^{\boldsymbol{x}}$ denote the symmetrized distribution $F^{\boldsymbol{x}}$ conditional on the neighbourhood $N_\beta^{\boldsymbol{x}}(F)$. Given $\boldsymbol{x} \in \mathscr{X}$, we may define a local depth kernel at locality level β centered at $\boldsymbol{x}$ as

$$K_\beta^{\boldsymbol{x}} \colon N_\beta^{\boldsymbol{x}}(F) \to \mathbb{R}, \qquad \boldsymbol{z} \mapsto D(\boldsymbol{z}, \widetilde{F}_\beta^{\boldsymbol{x}}). \tag{4.2.1}$$

This naturally extends to a map $\mathscr{X} \to \mathbb{R}$ as $K_\beta^{\boldsymbol{x}}(\boldsymbol{z}) = 0$ for $\boldsymbol{z} \notin N_\beta^{\boldsymbol{x}}(F)$.

Note that $\widetilde{F}_\beta^{\boldsymbol{x}}$ is angularly symmetric about $\boldsymbol{x}$. As a result, $K_\beta^{\boldsymbol{x}}$ is maximized at and decreases away from $\boldsymbol{x}$, for reasonably well behaved depth functions (**P2** and **P3** for multivariate depth functions).

With this, we propose the (linear) estimator

$$\hat{\boldsymbol{y}}_\beta(\boldsymbol{x}) = \sum_i w_i(\boldsymbol{x})\, \boldsymbol{y}_i, \qquad w_i(\boldsymbol{x}) = \frac{K_\beta^{\boldsymbol{x}}(\boldsymbol{x}_i)}{\sum_j K_\beta^{\boldsymbol{x}}(\boldsymbol{x}_j)}. \tag{4.2.2}$$

The locality level $\beta \in (0, 1]$ is a tuning parameter which may be chosen via methods such as cross-validation.

The kernel function $K_\beta^{\boldsymbol{x}}$ is supported on the neighbourhood $N_\beta^{\boldsymbol{x}}(F)$, whose shape may vary with changing $\boldsymbol{x} \in \mathscr{X}$. Indeed, since $N_\beta^{\boldsymbol{x}}(\hat{F}_n)$ contains the β proportion of points from $\{\boldsymbol{x}_i\}$ 'closest' to $\boldsymbol{x}$ (in the sense of being more central in the symmetrized dataset $\mathscr{D}^{\boldsymbol{x}}$), this neighbourhood ought to be smaller when $\boldsymbol{x}$ is more central, and larger when $\boldsymbol{x}$ has fewer points nearby. Thus, $K_\beta^{\boldsymbol{x}}$ behaves somewhat

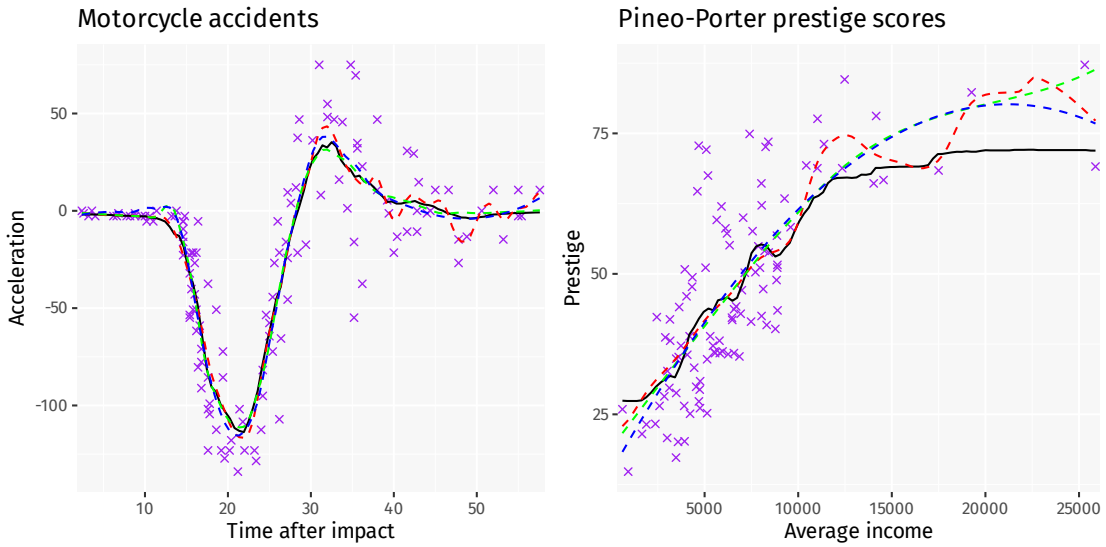


Figure 4.3: Regression curves for the `cars` and `carData::Prestige` datasets available in R. The black curve indicates the local depth based estimate, the dashed red curve indicates the Nadaraya-Watson kernel based estimate, and the dashed green and blue curves indicate the local linear and quadratic estimates respectively. The locality levels and relevant bandwidths have been obtained by leave-one-out cross validation. The local depth based estimate is the best and second-best in terms of MSE respectively.

like a variable bandwidth or adaptive kernel, whose shape adjusts to the surrounding data as $\boldsymbol{x}$ varies. Furthermore, the 'bandwidth' of $K_\beta^{\boldsymbol{x}}$ is controlled solely by the parameter β regardless of the dimensionality or nature of $\mathscr{X}$. This stands in contrast with more traditional kernels which often require a selection of multiple bandwidths. For instance, a Gaussian kernel of the form

$$\boldsymbol{z} \mapsto \frac{1}{\prod_{i=1}^d h_i} \exp\left(-\sum_i \frac{(x_i - z_i)^2}{2h_i^2}\right) \tag{4.2.3}$$

needs d parameters $\{h_i\}_{i=1}^d$ to be determined.

Equation 4.2.2 may also be thought of as a weighted KNN estimator, since $N_\beta^{\boldsymbol{x}}(\hat{F}_n)$ always captures the same number of points.

When the depth function D is chosen to be affine invariant, the estimator 4.2.2 is also affine invariant, in the sense that it is unchanged by an affine transformation of $\mathscr{X}$. This is because $N_\beta^{A\boldsymbol{x}+\boldsymbol{b}}(F_{A\boldsymbol{X}+\boldsymbol{b}})$ will simply be the affine image of $N_\beta^{\boldsymbol{x}}(F_{\boldsymbol{X}})$.

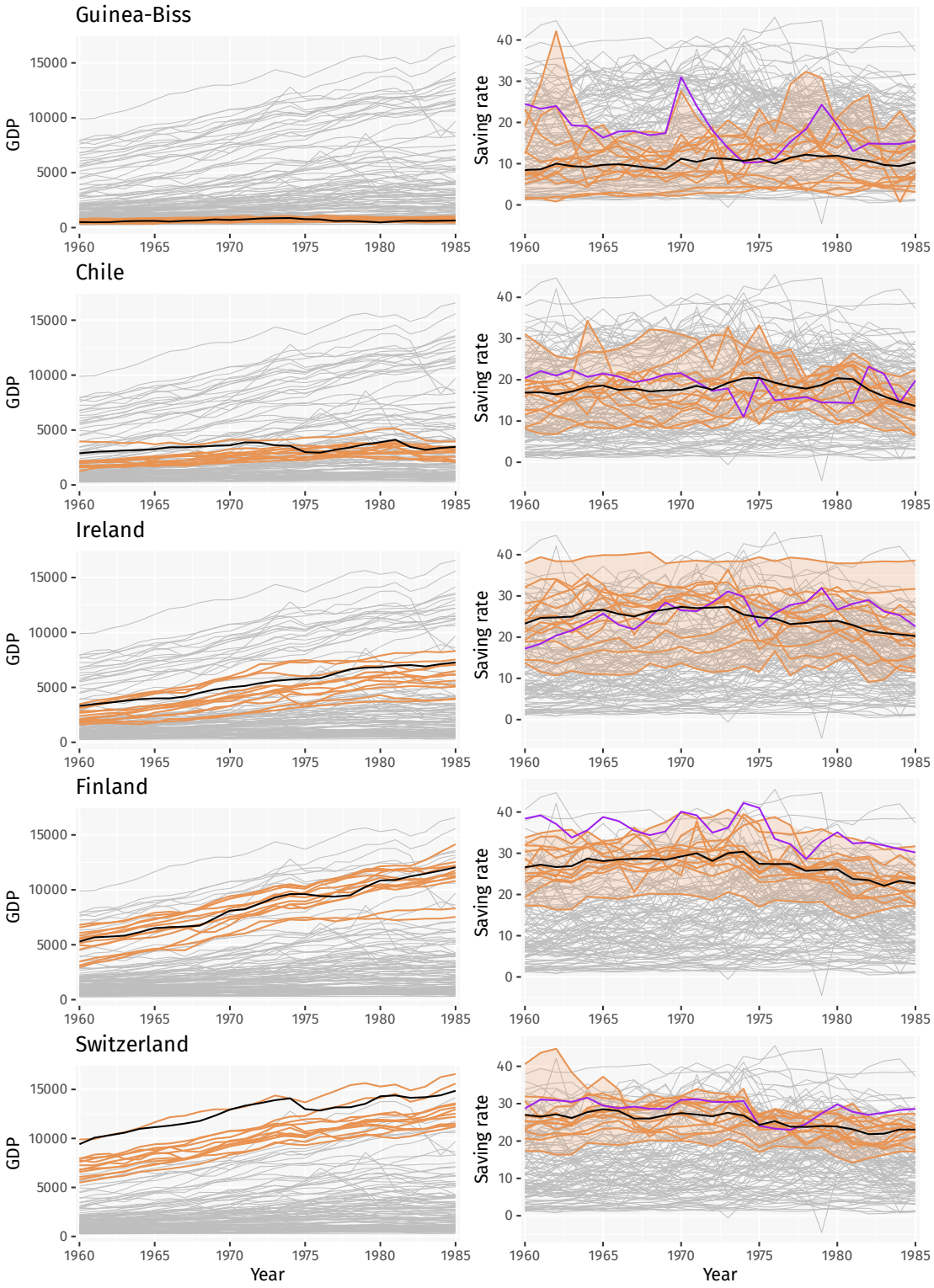


Figure 4.4: Regression curves for the Penn table dataset, avaiable as `Ecdat::SumHes` in *R*. The covariates are GDP curves of different countries in the left column. The new GDP curve of the indicated country is marked in black, and its $\beta = 0.1$ neighbours are marked in orange. On the left, the estimated savings rate curve is marked in black, with the true curve in purple. The orange response curves correspond to the orange covariates, and their envelope is shaded in.

Chapter 5

Conclusion