Depth Functions as Measures of Representativeness

Ye Dong Stephen M. S. Lee[1]

e-mail: dongyecarol@gmail.com *e-mail: smslee@hku.hk*

Department of Statistics and Actuarial Science, The University of Hong Kong

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

Abstract

Data depth provides a natural means to rank multivariate vectors with respect to an underlying multivariate distribution. Most existing depth functions emphasize a centre-outward ordering of data points, which may not provide a useful geometric representation of certain distributional features, such as multimodality, of concern to some statistical applications. Such inadequacy motivates us to develop a device for ranking data points according to their "representativeness" rather than "centrality" with respect to an underlying distribution of interest. Derived essentially from a choice of goodness-of-fit test statistic, our device calls for a new interpretation of "depth" more akin to the concept of density than location. It copes particularly well with multivariate data exhibiting multimodality. In addition to providing depth values for individual data points, depth functions derived from goodness-of-fit tests also extend naturally to provide depth values for subsets of data points, a concept new to the data-depth literature.

Keywords and phrases: centre-outward ordering; data depth; goodness-of-fit tests; multimodality; representativeness.

2000 MSC code: 62G99

1 Introduction

In the past decades, a variety of definitions of data depth have been proposed to provide a natural means to rank multivariate data. Such depth functions are formulated primarily to measure the "centrality" of a single point relative to a specified distribution function F or to a sample of observations $X_1, \ldots, X_n$ drawn from F. The deepest point found by a depth function is often referred to as the "centre" of the distribution F. Much emphasis has been put on monotonicity of the depth function relative to this deepest point, so much so that it has become one of the four characterising properties of centre-outward ordering depth functions as introduced by Liu (1990) and Zuo and Serfling (2000). Examples of data depths possessing such properties include Mahalanobis's depth (Mahalanobis, 1936), Tukey's depth (Tukey, 1975), simplicial depth (Liu, 1990) and majority depth (Singh, 1991). Ironically, the requirement that a depth function provide a centre-outward ordering has restricted the scope of depth-based inferences. Indeed, in many applications such as cluster analysis, classification or tests for equality of populations, usefulness of a depth function relies on the tacit assumption that a "deep" point of a distribution, or sample, should also be a point which is "representative" of that distribution, or sample. While this often holds true for unimodal distributions, such assumption is less plausible when the underlying distribution or sample exhibits some degree of

multimodality. More generally, if our objective is to rank data points in order of their "representativeness" of the reference distribution, we are in need of an alternative notion of depth which can genuinely measure representativeness and, in particular, endow data points with an ordering sufficiently responsive to multimodal features of the reference distribution. The problem has not received much attention so far. Exceptions include Baggerly and Scott (1999), who argue for an interpretation of multivariate median as the highest density "contour" encompassing a 50% probability mass under F. Zuo and Serfling (2000) believe it important to choose between "sensitivity to multimodality" and "centre-outward ordering" in the derivation of a proper notion of data depth. The first constructive attempt at a shift of emphasis from "centrality" to "representativeness" has been signalled by Fraiman and Meloche's (1999) likelihood depth. More recently, Chen, Dang, Peng and Bart (2009) propose a kernelized spatial depth function for detecting outliers in non-unimodal data patterns. Hlubinka, Kotík and Vencálek (2010) modify the halfspace depth by reweighting the probability contents of halfspaces. By controlling the volumes of simplices or halfspaces, Agostinelli and Romanazzi (2011) generalise the classical simplicial and halfspace depths to a local depth which can reveal local distributional features. Paindaveine and van Bever (2012) introduce a different notion of local depth which can be viewed as a localised measure of centrality.

We propose in this paper a general scheme of formulating data depths that can measure representativeness of data points, or subsets of data points, with respect to multivariate distributions. Our formulation hinges on a choice of goodness-of-fit test applicable to data of any dimension, and provides a very general method for constructing depth functions. In general, different choices of goodness-of-fit tests give rise to different formulations, leading to a rich class of depth functions of which many are new to the literature. In particular, goodness-of-fit tests based on inter-point distances are shown to be especially effective in formulating depth functions which provide satisfactory rankings of data points in order of representativeness.

The paper is organised as follows. Section 2 reviews three main classes of goodness-of-fit tests, based on which new classes of depth functions are formulated. Particularly promising as a tool for measuring "representativeness" is the class of depth functions derived from interpoint distances, which are investigated in more detail in Section 3. Section 4 illustrates an application of our new depth functions to supervised classification problems based on simulated data. Section 5 provides a real-life example which contrasts the ranking of a macroeconomic bivariate data set made by simplicial depth with that made by one of our new depth functions derived from within-triplet distances. Section 6 concludes our findings. Technical proofs are given in the Appendix.

2 Depth function based on goodness-of-fit test

Consider a random sample $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ drawn from a distribution F on the sample space $\mathcal{S}$. A goodness-of-fit test typically refers to a test of the null hypothesis that $F = F_0$, for some specified distribution function F_0.

Literature on goodness-of-fit tests is abundant. To fix ideas we consider for our formulation of depth functions three main classes of goodness-of-fit tests: the Kolmogorov-Smirnov-Cramér type, the Cressie-Read type and the interpoint-distance type. The above choices provide a sufficiently broad selection of goodness-of-fit tests for the generation of depth functions, although the list is by no means exhaustive.

Denote generically by $T(\mathcal{X}_n, F)$ the goodness-of-fit test statistic, large values of which indicate a lack of fit of the distribution F to the observed data $\mathcal{X}_n$, or in other words, a lack of "representativeness" of $\mathcal{X}_n$ with respect to the distribution F. This motivates our new formulation of a depth function applicable to a pattern of data points, under which "depth" acquires a new meaning as a measure of "representativeness". Specifically, for any collection of points $\{x_1, \ldots, x_n\} \subset \mathcal{S}$ and any distribution function F on $\mathcal{S}$, the depth of the pattern $\{x_1, \ldots, x_n\}$ with respect to F is defined to be

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

for some decreasing function $\eta(\cdot|F)$ on $\mathbb{R}$, which can be chosen arbitrarily. For exam-

ple, we may set $\eta(t|F) = (1+t)^{-1}$ for any non-negative test statistic $T(\{x_1, \ldots, x_n\}, F)$. As a canonical choice, we can set $\eta(t|F) = \mathbb{P}_F\left(T(\mathcal{X}_n, F) > t\right)$, in which case the depth function admits at once an additional interpretation as a goodness-of-fit p-value associated with the "sample" $\{x_1, \ldots, x_n\}$. In many examples the distribution of $T(\mathcal{X}_n, F)$ under F either does not depend on F or can be estimated by Monte Carlo simulation of random samples from F.

Under our new formulation, a "deep" point pattern $\{x_1, \ldots, x_n\}$ relative to a distribution F can be viewed as a "sample" in no essential conflict with F, or one which is reasonably "representative" of F. This formulation can easily be specialised, by considering the case $n = 1$, to provide a depth measure for a single point, in which sense a depth function has traditionally been understood. Without confusion we write $D(F, x) = D(F, \{x\})$ for the depth function of the point x with respect to the distribution F.

2.1 Kolmogorov-Smirnov-Cramér type

Goodness-of-fit tests of the Kolmogorov-Smirnov-Cramér type are further divided into two subclasses, one of the Kolmogorov-Smirnov type and the other of the Cramér-von Mises type.

Test statistics of the Kolmogorov-Smirnov type have the form

$$T(\mathcal{X}_n, F) = \sup \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \int f \, dF \right| : f \in \mathcal{F}_n \right\}, \tag{1}$$

for some pre-specified collection $\mathcal{F}_n$ of measurable functions. A two-sample version of (1) has been considered by Præstgaard (1995), who establishes conditions for consistency of its permutation and bootstrap distributions. The formulation (1) encompasses a variety of Kolmogorov-Smirnov tests found in the literature. In the case $\mathcal{S} = \mathbb{R}^d$, Wolfowitz (1954) defines the Kolmogorov-Smirnov distance by setting $\mathcal{F}_n = \{\mathbf{1}_H : H \in \mathcal{H}\}$, the class of indicator functions for the collection $\mathcal{H}$ of all closed halfspaces in $\mathbb{R}^d$. Cabaña and Cabaña (1997) construct classes of goodness-of-fit tests by setting $\mathcal{F}_n = \{\mathcal{T}(a\mathbf{1}_A) : A \in \mathcal{A}\}$, where $\mathcal{T}$ is an isometry on $L_2(\mathbb{R}^d, \mathbb{P}_F)$ with range equal to the orthogonal complement of the constant function 1, $a \in L_2(\mathbb{R}^d, \mathbb{P}_F)$ depends on the sequence of alternatives of interest and $\mathcal{A}$ is a class of subsets in $\mathbb{R}^d$ sufficiently rich to generate Borel sets. For a separable Hilbert space $\mathcal{S}$ endowed with scalar product $\langle \cdot, \cdot \rangle$, Cuesta-Albertos, Fraiman and Ransford (2006) suggest taking $\mathcal{F}_n = \{\mathbf{1}\{\langle \cdot, h \rangle \leq t\} : t \in \mathbb{R}\}$, where h is a random direction generated according to a non-degenerate Gaussian law on $\mathcal{S}$.

Test statistics of the Cramér-von Mises type have the form

$$T(\mathcal{X}_n, F) = n^{\alpha/2} \int_{\Theta} \left| \frac{1}{n} \sum_{i=1}^{n} f_\theta(X_i) - \int f_\theta \, dF \right|^{\alpha} d\lambda_F(\theta), \tag{2}$$

for some fixed $\alpha > 0$, some collection of measurable functions $\{f_\theta : \theta \in \Theta\}$ and some positive measure λ_F on the index space Θ. For $\mathcal{S} = \mathbb{R}$, it is common to set $\alpha = 2$, $\Theta = \mathbb{R}$, $f_\theta = \mathbf{1}_{(-\infty,\theta]}$ and $d\lambda_F(\cdot) = h(F(\cdot))\, dF(\cdot)$, for some positive function h on $\mathbb{R}$. A recent study of the power function of the test based on the choice $h(\theta) = \theta^{2\beta}$, for $\beta > -1$, can be found in Makhoukhi (2008). For $\mathcal{S} = \mathbb{R}^d$, we may take $\alpha = 2$, $\Theta = \mathbb{S}^{d-1} \times \mathbb{R}$ and $f_{(\theta_1,\theta_2)}(x) = \mathbf{1}\{\theta_1^{\mathrm{T}} x \leq \theta_2\}$, where $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathcal{R}^d$ centred at the origin. Zhu, Fang and Bhatti (1997) consider this setup with $d\lambda_F(\theta_1, \theta_2) = d\mu(\theta_1)\, d\mathbb{P}_F(\theta_1^{\mathrm{T}} X \leq \theta_2)$, where μ denotes the uniform probability measure on $\mathbb{S}^{d-1}$, whereas Baringhaus and Franz (2004) take λ_F to be the product of μ and the Lebesgue measure on $\mathbb{R}$. Alba-Fernández, Jiménez-Gamero and Muñoz-García (2008) consider a two-sample version of (2) with $\alpha = 2$, $\Theta = \mathbb{R}^d$, $f_\theta = e^{i\theta^{\mathrm{T}}(\cdot)}$ and λ_F an arbitrary probability measure on $\mathbb{R}^d$.

It follows easily from (1) and (2) that, for a singleton $\{x\}$,

$$D(F, x) = \eta\left(\sup\left\{\left| f(x) - \int f\, dF \right| : f \in \mathcal{F}_n\right\} \,\middle|\, F\right) \tag{3}$$

based on the Kolmogorov-Smirnov test, and

$$D(F, x) = \eta\left(\int_\Theta \left| f_\theta(x) - \int f_\theta\, dF \right|^\alpha d\lambda_F(\theta) \,\middle|\, F\right) \tag{4}$$

based on the Cramér-von Mises test.

On taking $\mathcal{F}_n$ to be the collection of indicators of closed halfspaces in $\mathbb{R}^d$ and $\eta(t|F) = 1 - t$, (3) reduces to Tukey's depth on $\mathbb{R}^d$. Furthermore, in the special

case where $d = 1$, setting $\eta(t|F) = 3/2 - t$ reduces (3) to the majority depth. More generally, the random functional depth introduced by Cuesta-Albertos and Nieto-Reyes (2008) can be derived from (3) using the random projection approach of Cuesta-Albertos, Fraiman and Ransford (2006).

Following Baringhaus and Franz's (2004) formulation, (4) reduces to

$$\begin{aligned} D(F,x) &= \eta\left(\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}^d} |\theta_1^{\mathrm{T}}(x-y)|\, dF(y)\, d\mu(\theta_1)\right. \\ &\qquad \left. -\frac{1}{2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}^d\times\mathbb{R}^d} |\theta_1^{\mathrm{T}}(y-z)|\, d(F\otimes F)(y,z)\, d\mu(\theta_1) \,\middle|\, F\right) \\ &= \eta\left(\int_{\mathbb{R}^d} \|x-y\|\, dF(y) - \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d} \|y-z\|\, d(F\otimes F)(y,z) \,\middle|\, F\right), \quad (5) \end{aligned}$$

where $\|\cdot\|$ denotes the Euclidean norm. On the other hand, if $d = 1$ and we set $\alpha = 2$, $\Theta = \mathbb{R}$, $f_\theta = \mathbf{1}_{(-\infty,\theta]}$ and $d\lambda_F(\cdot) = h(F(\cdot))\, dF(\cdot)$, then (4) becomes

$$D(F,x) = \eta\left(F(x)^3 + (1-F(x))^3 \,\middle|\, F\right)$$

if h is a constant function, and

$$D(F,x) = \eta\left(F(x)^{-1}(1-F(x))^{-1} \,\middle|\, F\right)$$

if $h(u) = u^{-1}(1-u)^{-1}$. The latter leads to the simplicial depth as a special case.

2.2 Cressie-Read type

Let $\{C_j : j = 1, \ldots, k\}$ be a partition of $\mathcal{S}$. Cressie and Read (1984) introduce a family of power-divergence statistics of the form

$$T(\mathcal{X}_n, F) = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} \mathbf{1}\{X_i \in C_j\} \right) \left\{ \left(\frac{\sum_{i=1}^{n} \mathbf{1}\{X_i \in C_j\}}{n \int_{C_j} dF} \right)^{\lambda} - 1 \right\}, \tag{6}$$

which is suitable for testing the fit of the null distribution F, for any real constant λ. Special cases include the Pearson's chi-squared test statistic ($\lambda = 1$), the log-likelihood ratio statistic ($\lambda \to 0$), the Freeman-Tukey statistic ($\lambda = -1/2$) and the Neyman modified chi-squared test statistic ($\lambda = -2$). It can be shown that the Cressie-Read statistics are asymptotically chi-square on $k - 1$ degrees of freedom under the null distribution.

Setting $n = 1$ in (6), we obtain, for a singleton $\{x\}$, the depth function

$$D(F, x) = \eta \left(\frac{2}{\lambda(\lambda+1)} \left\{ \left[\int_{C_{j(x)}} dF(y) \right]^{-\lambda} - 1 \right\} \middle| F \right), \tag{7}$$

where $j(x) \in \{1, \ldots, k\}$ identifies the subset $C_{j(x)}$ that contains x.

2.3 Interpoint-distance type

Let $\delta(\cdot, \cdot)$ be an arbitrary distance measure on $\mathcal{S}$. Tests of the interpoint-distance type require calculations of δ-distances between points in the sample. They are often

introduced in the form of a multivariate two-sample test. Examples include those based on minimum spanning trees (Friedman and Rafsky, 1979), nearest neighbours (Schilling, 1986; Henze, 1988), interpoint distances within a triplet (Bartoszyński, Pearl and Lawrence, 1997), optimal cross-matches (Rosenbaum, 2005) and a notion of minimum energy (Aslan and Zech, 2005). The above tests are designed primarily to test for equality of two populations, from which two independent random samples, say $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and $\mathcal{Y}_m = \{Y_1, \ldots, Y_m\}$, are available. They can nevertheless be converted into one-sample goodness-of-fit tests by considering, under the assumption $Y_i \sim F$, either the limiting case $m \to \infty$ or the expected value of the test statistic with respect to the drawing of finite samples $\{Y_1, \ldots, Y_m\}$ from F for a fixed m.

As no single unifying formulation exists of the test statistics of the interpoint-distance type, we describe below three important examples which are distinct enough to reflect the diversity of this class of goodness-of-fit tests. In each example a closed-form expression can be obtained of the test statistic $T(\mathcal{X}_n, F)$.

(i) Tests based on nearest neighbours —

Tests of this type, as discussed by Schilling (1986) and Henze (1988), are designed to handle general multivariate two-sample problems. The test statistic is derived from the proportion of all k nearest neighbour comparisons in which observations

and their neighbours belong to the same sample.

Let $N_r(Z)$ be the rth nearest δ-neighbour in the combined sample $\mathcal{X}_n \cup \mathcal{Y}_m$ of the sample point $Z \in \mathcal{X}_n \cup \mathcal{Y}_m$. Then an unweighted version of the two-sample test statistic can be written as

$$k^{-1}(m+n)^{-1}\left(\sum_{i=1}^{n}\sum_{r=1}^{k}\mathbf{1}\left\{N_r(X_i)\in\mathcal{X}_n\right\}+\sum_{j=1}^{m}\sum_{r=1}^{k}\mathbf{1}\left\{N_r(Y_j)\in\mathcal{Y}_m\right\}\right). \qquad (8)$$

We derive below an adaptation of (8) to the one-sample problem by considering the limiting case where $m \to \infty$, $k/m \to \gamma \in (0,1)$ and $Y_1, Y_2, \ldots$ are independently distributed under F.

Define, for any fixed $x \in \mathcal{S}$ and any distribution F on $\mathcal{S}$,

$$F_x(t) = \mathbb{P}_F(\delta(x,X) \le t \mid X \sim F), \quad t \in \mathbb{R}.$$

For $m, k \gg n$, the combined sample $\mathcal{X}_n \cup \mathcal{Y}_m$ is so dominated by $\mathcal{Y}_m$ that $N_k(X_i)$ is essentially the observation in $\mathcal{Y}_m$ which is the kth nearest to X_i. Thus $N_k(X_i)$ lies at a distance D from X_i that satisfies approximately $F_{X_i}(D) = k/m \approx \gamma$, suggesting that $D \approx F_{X_i}^{-1}(\gamma)$. In the limiting case, we have, for each $i = 1, \ldots, n$,

$$\begin{aligned}\sum_{r=1}^{k}\mathbf{1}\left\{N_r(X_i)\in\mathcal{X}_n\right\} &= \sum_{j=1}^{n}\sum_{r=1}^{k}\mathbf{1}\left\{N_r(X_i)=X_j\right\}\\ &\approx \sum_{j=1}^{n}\mathbf{1}\left\{\delta(X_i,X_j)\le F_{X_i}^{-1}(\gamma)\right\}.\end{aligned}$$

The second term in (8) is essentially non-informative in the limiting case, since $k^{-1}(m+n)^{-1}\sum_{j=1}^{m}\sum_{r=1}^{k}\mathbf{1}\left\{N_r(Y_j)\in\mathcal{Y}_m\right\} \approx 1$ for m, k large. It thus follows that a

one-sample version of the statistic (8) can be taken as

$$T(\mathcal{X}_n, F) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}\left\{\delta(X_i, X_j) \leq F_{X_i}^{-1}(\gamma)\right\}. \tag{9}$$

Consider first the case $n = 2$ in which x_1, x_2 differ by an infinitesimally small distance $\epsilon > 0$. Then it follows from (9) that

$$D(F, \{x_1, x_2\}) = \eta\left(2 + \mathbf{1}\{F_{x_1}^{-1}(\gamma) \geq \epsilon\} + \mathbf{1}\{F_{x_2}^{-1}(\gamma) \geq \epsilon\} \,\middle|\, F\right),$$

which, for fixed ϵ, decreases as $F_{x_i}^{-1}(\gamma)$ increases, for $i = 1, 2$. The result suggests that, on setting $\epsilon \to 0$, we may define the depth of a single point x to be

$$D(F, x) = \eta\left(F_x^{-1}(\gamma) \,\middle|\, F\right). \tag{10}$$

(ii) Energy tests —

Aslan and Zech (2005) propose a two-sample energy test statistic of the form

$$\begin{aligned} \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} R(\delta(X_i, X_j)) + \frac{1}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} R(\delta(Y_i, Y_j)) \\ - \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} R(\delta(X_i, Y_j)), \end{aligned} \tag{11}$$

where R, known as the energy function, is monotonically decreasing on $[0, \infty)$. In the case $\mathcal{S} = \mathbb{R}^d$, we can take, for example, δ to be the Euclidean distance and define $R(r) = (r^{-\kappa} - 1)/\kappa$ for some fixed $\kappa \in [0, d/2)$. Note that the case $\kappa = 0$ corresponds, by considering the limiting case $\kappa \to 0$, to the choice $R(r) = -\ln(r)$.

Taking $R(r) = -r$, on the other hand, reduces the energy test to Baringhaus and Franz's (2004) multivariate two-sample test.

By taking expectation with respect to sampling of $\mathcal{Y}_m$ under F or considering the stochastic limit as $m \to \infty$, it is easily seen that (11) can be converted into a one-sample energy test statistic of the form

$$T(\mathcal{X}_n, F) = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} R(\delta(X_i, X_j)) - \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{S}} R(\delta(X_i, y))\, dF(y) + \frac{1}{2} \int_{\mathcal{S}^2} R(\delta(y_1, y_2))\, d(F \otimes F)(y_1, y_2). \quad (12)$$

Thus, for a single $x \in \mathcal{S}$, (12) leads to the depth function

$$D(F, x) = \eta \left(\frac{1}{2} \int_{\mathcal{S}^2} R(\delta(y, z))\, d(F \otimes F)(y, z) - \int_{\mathcal{S}} R(\delta(x, y))\, dF(y) \,\middle|\, F \right). \quad (13)$$

Note in the case $\mathcal{S} = \mathbb{R}^d$ that if we set δ to be the Euclidean distance and $R(r) = -r$, (13) reduces to the depth function given by (5). Fraiman and Meloche (1999) propose an affine invariant version of likelihood depth, which can be regarded as a special case of (13) if we set $\eta(t|F) = \frac{1}{2} \int_{\mathcal{S}^2} R(\delta(y, z))\, d(F \otimes F)(y, z) - t$, $R(t) = K(t/h)/h^d$ for some kernel function K and bandwidth h, and, with slight abuse of notation, $\delta(y, z) = (y - z)^{\mathrm{T}} \Sigma_F^{-1} (y - z)$, where Σ_F denotes the dispersion matrix of F.

(iii) Tests based on within-triplet distances —

Bartoszyński, Pearl and Lawrence (1997) introduce a multidimensional goodness-

of-fit test based on within-triplet distances by appealing to the fact that

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

and

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

have zero means under the null distribution, a result which forms the basis of their proposed goodness-of-fit test statistic

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

for some positive semidefinite matrix A designed to give good power properties against specific alternatives. Setting $n = 1$ in the above, we may define the depth of a point x to be

$$D(F, x) = \eta\left(\nu(F, x)^{\mathrm{T}} A\, \nu(F, x) \,\middle|\, F\right),$$

where $\nu(F, x) = [\nu_1(F, x), \nu_2(F, x), \nu_3(F, x)]^{\mathrm{T}}$,

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

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

and $\nu_2(F, x) = -\nu_1(F, x) - \nu_3(F, x)$. We shall henceforth focus, for simplicity, on the special case where A is diagonal such that

$$D(F, x) = \eta\left(\sum_{j=1}^{3} w_j \nu_j(F, x)^2 \,\middle|\, F\right), \tag{14}$$

for some weights $w_1, w_2, w_3 \geq 0$.

2.4 Numerical illustration with multidimensional data

To investigate their effectiveness in "representing" a reference distribution, we study empirically a number of depth functions derived from the three classes of goodness-of-fit tests. Special attention is paid to the question of whether the shapes of the depth functions preserve the multimodal feature of the underlying distribution F. The underlying distributions chosen for analysis are bimodal mixtures of multivariate normal distributions, namely $0.5\,N(-2\mathbf{1}_d, I_d) + 0.5\,N(2\mathbf{1}_d, I_d)$, for $d = 2$ and 10 respectively, where $\mathbf{1}_d$ denotes the d-vector of one's and I_d the $d \times d$ identity matrix. Only depth values of singletons $x \in \mathbb{R}^d$ are calculated for comparison. In the ten-dimensional case, the direction crossing the two modes of F is considered. Throughout the study δ is taken to be the Euclidean distance and $\eta(t|F) = -t$. Depth functions under investigation consist of the following examples drawn from the three classes of goodness-of-fit tests:

1. *Kolmogorov-Smirnov-Cramér type* —

   Tukey's depth, obtained by taking $\mathcal{F}_n$ in (3) to be the collection of indicators of closed halfspaces in $\mathbb{R}^d$; and the depth function (5);

2. *Cressie-Read type* —

   depth function (7) based on the Pearson's chi-squared test ($\lambda = 1$);

3. *interpoint-distance type* based on

   (i) *nearest neighbours* — depth function (10) with $\gamma = 0.05$;

   (ii) *energy tests* — depth function (13) with $R(r) = -\ln(r)$;

   (ii) *within-triplet distances* — depth function (14) with $w_1 = w_2 = 0.5$ and $w_3 = 0$.

Depth values based on within-triplet distances for cases $d = 2$ and 10 are approximated by Monte Carlo simulation of 1,000 and 3,000 observations from F, respectively, whereas depth values based on nearest neighbours for the case $d = 10$ are obtained by Monte Carlo simulation of 5,000 observations. In the remaining cases we compute the depth values with respect to F directly from closed-form expressions.

Depth functions of the Kolmogorov-Smirnov-Cramér type are found to be centre-outward ordering and fail to capture the bimodal shape of the underlying distribution. Results for the other two types are displayed in Figure 1, which shows that depth functions constructed by Pearson's chi-squared tests or by tests based on interpoint distances are effective in capturing bimodality in their depth graphs, except for the ten-dimensional case where depth values based on energy tests are unimodal along the direction crossing the two modes of F. Depth plots based on within-triplet distances reveal a multimodal shape with two conspicuous modes standing at the two modes of the underlying F. Our empirical evidence suggests that depth func-

tions derived from tests of the interpoint-distance type are preferable as a measure of representativeness to those based on the chi-squared test, as the latter suffers considerably from a lack of smoothness over the sample space. In the remaining sections we turn our attention to the properties and applications of depth functions based on interpoint distances.

3 Depth functions based on interpoint distances

3.1 Theoretical properties

Many practical applications require that a depth function be evaluated with respect to a random sample $\mathcal{Y}_m = (Y_1, \ldots, Y_m)$ drawn from F rather than to F directly, for the latter is often unavailable. We thus define the sample depth function of the point pattern $\{x_1, \ldots, x_n\}$ to be $D(F_{\mathcal{Y}_m}, \{x_1, \ldots, x_n\})$, where $F_{\mathcal{Y}_m}$ denotes the empirical distribution of $\mathcal{Y}_m$. We comment below briefly on the conditions sufficient for consistency of sample depth functions of the interpoint-distance type, in the sense that $D(F_{\mathcal{Y}_m}, \{x_1, \ldots, x_n\})$ converges in probability to $D(F, \{x_1, \ldots, x_n\})$ as $m \to \infty$. In each case we assume η to be a continuous function.

Consistency of sample depth functions based on nearest neighbours follows from strong consistency of the sample γth quantile of the m distances $\delta(x, Y_1), \ldots, \delta(x, Y_m)$,

which converges in probability to $F_x^{-1}(\gamma)$ for any $x \in \mathcal{S}$: see, for example, Serfling (1980, Section 2.3.1).

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

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

and

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

provided that the limits exist. Similarly we can show that

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

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

For more insight into the relationship between $D(F, \cdot)$ and F, we consider a univariate setting where F has a bounded, positive and continuously differentiable density $f = F'$, and where the point pattern consists of a singleton $x \in \mathbb{R}$. Given the resemblance between Fraiman and Meloche's (1999) likelihood depth and the depth function based on energy tests, we refer to the aforesaid paper for general properties of the latter depth function. We hereby restrict attention to depth functions derived from nearest neighbours and within-triplet distances.

The following proposition, which we prove in the Appendix, provide conditions that characterise local maxima or minima of depth functions based on nearest neighbours.

Proposition 1 *The depth function (10) based on nearest neighbours has a local maximum or local minimum at* x_0 *which satisfies, for some* $r_0 > 0$, $F(x_0 + r_0) - F(x_0 - r_0) = \gamma$ *and* $f(x_0 + r_0) = f(x_0 - r_0)$, *according as* $f'(x_0 + r_0) - f'(x_0 - r_0) < 0$ *or* > 0 *respectively.*

Proposition 1 asserts that (10) has a local maximum or minimum at the midpoint x_0 of an interval which has probability mass γ and the same densities at its two endpoints. The gradients of f at the two endpoints characterise the depth nature of x_0. Typically, the depth function is locally maximised at x_0 if the density over the interval is relatively higher than the density outside it, and vice versa. We note, however, that a sharp peak of f at x_0 may be interpreted by the depth function as an outlier and assigned a small depth value. Similarly, a local minimum x_0 of f may be deemed representative of f, and assigned a large depth, if it lies between two close peaks of f. Decreasing γ increases the sensitivity of the depth function to local features of f.

It is clear that the depth function based on nearest neighbours can be made affine invariant and vanishing at $\pm\infty$ if we choose, for example, δ to be the Euclidean

distance and $\eta(t|F) = \left\{1 + t/\Lambda_F^{-1}(q)\right\}^{-1}$ for any fixed $q \in (0,1)$, where $\Lambda_F(t) = \mathbb{P}_F(\delta(Y_1, Y_2) \leq t)$. The next proposition shows that the depth function based on nearest neighbours possesses the same properties as those typically required of a conventional depth function under appropriate unimodality conditions on f. The proof is given in the Appendix.

Proposition 2 *Suppose that f has a unique mode at x_0, strictly decreases on (x_0, ∞) and strictly increases on $(-\infty, x_0)$. Then the depth function (10) based on nearest neighbours has a unique deep centre and decreases strictly as x moves away from the centre in either direction.*

We consider next the depth function (14) based on within-triplet distances. Note that if we set δ to be the Euclidean distance, then the depth function is invariant under rotations, as well as under location and scale changes. The properties of (14) depend primarily on the constituent functions $\nu_j(F, x)$. It is easy to show in the univariate setting that

$$\nu_3(F, x) = 2F(x)(1 - F(x)) - 1/3$$

and

$$\nu_2(F, x) = 2\int_0^\infty \{1 - F(2w + x)\} f(w + x)\, dw + 2\int_{-\infty}^0 F(2w + x) f(w + x)\, dw - 1/3.$$

Elementary calculus shows that the function $\nu_3(F,x)^2$ has a W-shape with one local maximum at $x = F^{-1}(1/2)$ and two local minima at $x = F^{-1}(1/2 \pm 1/\sqrt{12})$. That its shape is determined essentially by only three quantiles of F does not render the function $\nu_3(F,x)^2$ very effective in representing F. We see, on the other hand, that

$$\frac{\partial \nu_2(F,x)}{\partial x} = 2f(x)(2F(x)-1) + 2\int_0^\infty \mathrm{sgn}(w) f(2w+x) f(w+x)\, dw$$

and

$$\frac{\partial^2 \nu_2(F,x)}{\partial x^2} = 2f'(x)(2F(x)-1) - 2\int_0^\infty \mathrm{sgn}(w) f'(2w+x) f(w+x)\, dw,$$

where $\mathrm{sgn}(w) = \mathbf{1}_{(0,\infty)}(w) - \mathbf{1}_{(-\infty,0)}(w)$. Dependence of the function $\nu_2(F,x)^2$ on F is too intricate to be described in interpretive terms under a general F. If we specialise to the case of a symmetric unimodal density f centred at x_0, then we see that $\nu_2(F,x)^2$ has a local minimum or maximum at $x = x_0$ according as

$$\int_0^\infty \{1 - F(2w+x_0)\} f(w+x_0)\, dw - 1/12 \tag{15}$$

is positive or negative. For example, if $f(x) \propto (1+x^2)^{-\kappa}$, then (15) is positive whenever $\kappa < 3/2$, in which case $\nu_2(F,x)^2$ has a local minimum at $x = 0$. Thus, if we set $w_1 = w_3 = 0$ in the definition of (14), the depth function will return a local maximum at 0 if $\kappa < 3/2$ and a local minimum there if $\kappa > 3/2$.

In general the depth function (14) is a decreasing function of a weighted sum of the $\nu_j(F,x)^2$, and has its maxima and minima governed by corresponding weighted

sums of the conditions applicable to each of the constituents $\nu_j(F,x)^2$. Given its rather insensitivity to the shape of F, it seems preferable to give $\nu_3(F,x)^2$ a small weight w_3.

3.2 Numerical illustration

For more concrete illustration, we calculate explicit expressions for various depth functions under a bimodal distribution F which consists of a mixture of two univariate normal distributions, $0.85\,N(3,1) + 0.15\,N(-3,1)$. We set $\eta(t|F) = -t$ in all cases. For the depth function (13) based on energy tests, we take $R(r)$ to be the $N(0,h^2)$ density function so that it resembles Fraiman and Meloche's (1999) likelihood depth based on a normal kernel with bandwidth h. For comparison we include also the local simplicial depth proposed by Agostinelli and Romanazzi (2011), which is defined to be the probability that the point x is contained in a random simplex, generated under F, of volume less than some threshold τ. They show that by choosing a small τ, the local simplicial depth reflects to some extent the shape of the underlying density function, thus sharing similar properties as the likelihood depth. We note that setting $\tau = \infty$ reduces the local simplicial depth to the classical simplicial depth, which is necessarily unimodal. Unlike the local simplicial or likelihood depths, the depth functions based on nearest neighbours or within-triplet

distances do not purport to recover the shape of the underlying density function, as has been discussed in Section 3.1.

Figure 2 plots the depth functions under different parameter settings, with reference to the underlying bimodal density function. We see that the local simplicial depth, the likelihood depth and the depth based on nearest neighbours all exhibit a bimodal shape when their respective control parameters are set at relatively small values, and gradually become unimodal as the parameter values increase. Depth functions based on within-triplet distances are in general responsive to the changing shape of the underlying density, although no clear trend can be deciphered across different combinations of the weights (w_1, w_2, w_3). It appears from the figures that the choice $(w_1, w_2, w_3) = (0.5, 0.5, 0)$ yields the most satisfactory results.

3.3 Choice of control parameters

We have seen that properties of depth functions based on interpoint distances depend sensitively on control parameters, that is the threshold γ for the nearest neighbour depth, the bandwidth h for the likelihood depth, and the weights (w_1, w_2, w_3) for the depth based on within-triplet distances. It is therefore desirable to have some practical guidance on the choice of such parameters. If our main object is to define a depth function $D(F, x)$ to best "represent" the underlying distribution F, it is

reasonable then to select parameters which provide the strongest "correlation" between the shapes of the depth function and the distribution. One possible measure of "correlation" can be evaluated, by analogy with the Pearson correlation coefficient, using the formula

$$\frac{\int_{\mathcal{S}} D(F,x)\varpi(x)\,dF(x) - \int_{\mathcal{S}} \varpi(x)\,dF(x) \int_{\mathcal{S}} D(F,x)\varpi(x)\,dx}{\sqrt{\int_{\mathcal{S}} D(F,x)^2\varpi(x)\,dx - \left(\int_{\mathcal{S}} D(F,x)\varpi(x)\,dx\right)^2}}, \tag{16}$$

where $\varpi(\cdot)$ denotes a weight function which can be taken conveniently to be a proper density function on $\mathcal{S}$ such that the integrals in (16) are finite. In practical situations where a random sample $\mathcal{Y}_m = (Y_1, \ldots, Y_m)$ from F, rather than F itself, is available, (16) can be approximated by its sample version

$$\frac{m^{-1}\sum_{i=1}^{m} D(F_{\mathcal{Y}_m}, Y_i)\varpi(Y_i) - m^{-1}\sum_{i=1}^{m} \varpi(Y_i) \int_{\mathcal{S}} D(F_{\mathcal{Y}_m}, x)\varpi(x)\,dx}{\sqrt{\int_{\mathcal{S}} D(F_{\mathcal{Y}_m}, x)^2\varpi(x)\,dx - \left(\int_{\mathcal{S}} D(F_{\mathcal{Y}_m}, x)\varpi(x)\,dx\right)^2}}, \tag{17}$$

maximisation of which leads to an empirical choice of the control parameters necessary for fixing $D(F_{\mathcal{Y}_m}, \cdot)$.

For illustration we maximise (16) for the four depth functions considered in the example of Section 3.2, with $\varpi(\cdot)$ set to be the uniform density function over the interval $[-10, 10]$. The maximising parameters are found to be $\gamma = 0.14$, $h = 0.055$ and $(w_1, w_2, w_3) = (0, 0.6, 0.4)$ for depth functions based on nearest neighbours, energy tests and within-triplet distances, respectively, and $\tau = 3.5$ for the local simplicial depth. We may refer to Figure 2 for a rough estimate of the shapes of the

depth functions given by the above maximising parameter values.

4 Application to supervised classification

4.1 Maximum depth classifier

Our new notion of data depth that emphasizes "representativeness" is most relevant to statistical problems which demand a high level of sensitivity towards features of F not restricted to location and scale. Such applications include, for example, cluster analysis, support estimation, classification and nonparametric multi-sample tests.

We consider in this section a statistical application of our depth functions to the problem of supervised classification. Suppose that we have available two labelled training samples, $\mathcal{Y}^{[1]} = (Y_1^{[1]}, \ldots, Y_{n_1}^{[1]})$ and $\mathcal{Y}^{[2]} = (Y_1^{[2]}, \ldots, Y_{n_2}^{[2]})$, drawn respectively from two distinct distributions F_1 and F_2. We are interested in classifying a new set of data points $\{x_1, \ldots, x_n\}$, which are known to come from the same distribution, to one of the two distributions. Our extended notion of depth function $D(F, \{x_1, \ldots, x_n\})$ provides a natural procedure for classification, which we describe below.

Denote by $F_{\mathcal{Y}^{[j]}}$ the empirical distribution of $\mathcal{Y}^{[j]}$, $j = 1, 2$. For each $j = 1, 2$ and

$i = (i_1, \ldots, i_n) \in \{1, \ldots, n_j\}^n$, calculate the depth values

$$d_i^{[j]} = D(F_{\mathcal{Y}^{[j]}}, \{Y_{i_1}^{[j]}, \ldots, Y_{i_n}^{[j]}\}).$$

Then, under the assumption of a uniform prior, we classify $\{x_1, \ldots, x_n\}$ as coming from F_1 if

$$\begin{aligned} & n_1^{-n}\text{card}\left(\left\{i : d_i^{[1]} \leq D(F_{\mathcal{Y}^{[1]}}, \{x_1, \ldots, x_n\})\right\}\right) \\ > \;\; & n_2^{-n}\text{card}\left(\left\{i : d_i^{[2]} \leq D(F_{\mathcal{Y}^{[2]}}, \{x_1, \ldots, x_n\})\right\}\right), \end{aligned} \tag{18}$$

and from F_2 if the above inequality is reversed. The classification is inconclusive if the two sides of (18) are equal. We note that in the special case where $n = 1$, the above classifier has the form of a maximum depth classifier as discussed by Ghosh and Chaudhuri (2005), with the depth function $D(F, x)$ rescaled by its cumulative distribution function $\mathbb{P}_F\left(D(F, X) \leq \cdot\right)$ under $X \sim F$. Making use of the DD-plot, Li, Cuesta-Albertos and Liu (2012) propose a more general depth-based approach to classification; see also Lange, Mosler and Mozharovskyi (2012).

Instead of maximising (17), we consider it more natural in the present context to set the control parameters for the depth functions by minimising some estimate of the misclassification rate. Define a classifier $\mathcal{C}\left(\mathcal{Y}^{[1]}, \mathcal{Y}^{[2]}, \{x_1, \ldots, x_n\}\right) = 1$ if (18) holds, 2 if the reverse of (18) holds, and 0 otherwise. For $j = 1, 2$, define $\mathcal{N}_j$ to be the set of all $(i_1, \ldots, i_n) \in \{1, \ldots, n_j\}^n$ with $i_1 < \cdots < i_n$, $\mathcal{Y}_i^{[j]} = \{Y_{i_1}^{[j]}, \ldots, Y_{i_n}^{[j]}\}$

and $\mathcal{Y}_{-i}^{[j]} = \mathcal{Y}^{[j]} \setminus \mathcal{Y}_i^{[j]}$, where $i = (i_1, \ldots, i_n) \in \mathcal{N}_j$. The misclassification rate of the classifier $\mathcal{C}$ can be estimated by leave-n-out cross validation to be

$$\begin{aligned}
&\left[\binom{n_1}{n} + \binom{n_2}{n}\right]^{-1} \\
&\times \left\{ \sum_{i \in \mathcal{N}_1} \left[\mathbf{1}\left\{ \mathcal{C}\left(\mathcal{Y}_{-i}^{[1]}, \mathcal{Y}^{[2]}, \mathcal{Y}_i^{[1]}\right) = 2 \right\} + 0.5\, \mathbf{1}\left\{ \mathcal{C}\left(\mathcal{Y}_{-i}^{[1]}, \mathcal{Y}^{[2]}, \mathcal{Y}_i^{[1]}\right) = 0 \right\} \right] \right. \\
&\qquad \left. + \sum_{i \in \mathcal{N}_2} \left[\mathbf{1}\left\{ \mathcal{C}\left(\mathcal{Y}^{[1]}, \mathcal{Y}_{-i}^{[2]}, \mathcal{Y}_i^{[2]}\right) = 1 \right\} + 0.5\, \mathbf{1}\left\{ \mathcal{C}\left(\mathcal{Y}^{[1]}, \mathcal{Y}_{-i}^{[2]}, \mathcal{Y}_i^{[2]}\right) = 0 \right\} \right] \right\},
\end{aligned} \tag{19}$$

in which any inconclusive case is given a count of 0.5. Minimisation of (19) then leads to an empirical choice of the control parameters.

4.2 Numerical examples

In the following numerical examples we set $n_1 = n_2 = 50$ and consider the two cases $n = 1$ and $n = 2$. For $n = 1$, we take F_1 to be a mixture of bivariate normal distributions, $0.2\, N([-4, 0]^{\mathrm{T}}, 4I_2) + 0.8\, N([4, 0]^{\mathrm{T}}, I_2)$, and F_2 to be the bivariate normal distribution with mean zero and dispersion matrix $\begin{pmatrix} 2 & -2.2 \\ -2.2 & 9.64 \end{pmatrix}$. For $n = 2$, we take F_1 to be the univariate normal mixture $0.2\, N(-2.5, 4) + 0.8\, N(2.5, 1)$ and F_2 to be $N(0, 3.24)$. Under the above settings the Bayes misclassification rates based on the uniform prior are found to be 0.0560 for the case $n = 1$ and 0.1814 for the case $n = 2$, which can be obtained by evaluating the integrals $2^{-1} \int_{\mathbb{R}^2} \min\{f_1(x), f_2(x)\}\, dx$ and

$2^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\min\{f_1(x)f_1(y), f_2(x)f_2(y)\}\,dx\,dy$, respectively, where f_j denotes the density of F_j, $j = 1, 2$.

Example (i) $n = 1$ —

Figure 3 shows the classification of $x \in \mathbb{R}^2$ using four different depth functions for the case $n = 1$. Points lying in the light grey region are classified as F_1 and those in the dark grey region as F_2. The white area indicates those points which cannot be conclusively classified. We see that the local simplicial depth and the likelihood depth based on a small bandwidth suffer from serious overfitting, leaving behind a large inconclusive area. Varying the threshold τ for the local simplicial depth does not make much difference, as points outside the convex hull of each training sample are given a depth of zero with respect to that training sample, rendering them indistinguishable between the two samples. Depth functions based on nearest neighbours with $\gamma = 0.05$ and on within-triplet distances appear to provide more satisfactory classification. We note also that larger choices of γ (for nearest neighbours) and bandwidth h (for likelihood depth) tend to suppress bimodality of the depth function, leading to a bigger chance of misclassification.

For a more detailed study, we consider for each depth function seven different control parameter settings, under each of which the misclassification rate is estimated by leave-one-out cross validation (19) and marked by the letter "**V**" in

Figure 4. Minimising over the seven parameter settings for each depth function, the cross-validated choices of the parameters are found to be $\gamma = 0.05$, $h = 1$, $(w_1, w_2, w_3) = (0, 1, 0)$ and $\tau = 10$ for classifiers based on nearest neighbour depth, likelihood depth, within-triplet distances and local simplicial depth, respectively.

Next we generate from each distribution F_j a test sample of 50 observations, which are to be classified based on the training data shown in Figure 3. The misclassification rates are summarised in Figure 4. We see that with the exception of local simplicial depth, all the other three depth functions succeed in returning misclassification rates very close to the Bayes rate 0.0560 under at least one of the control parameter settings being considered. The local simplicial depth has a misclassification rate considerably bigger than 0.1 for all choices of the threshold value τ including the case $\tau = 10000$ which corresponds to the classical centre-outward ordering simplicial depth. We also see that cross validation based on (19) is very effective in identifying the optimal, or nearly optimal, choice of the control parameter for each depth function.

Example (ii) $n = 2$ —

For the case $n = 2$, only the energy test and the within-triplet distance test provide useful expressions for calculating the depth $D(F_{\mathcal{Y}[j]}, \{x_1, x_2\})$ of a point-pair (x_1, x_2), which is displayed in Figure 5 for $j = 1$ (left panel) and $j = 2$ (right

panel). The corresponding classification is shown in Figure 6. We see again that the likelihood depth based on a large bandwidth $h = 10$ does not capture bimodality of the first sample, leading to a classification quite different from the other cases.

For a study of misclassification rates, we consider for each depth function the same seven control parameter settings as those given in example (i). As shown in Figure 7, leave-two-out cross validation based on (19) suggests setting $h = 0.5$ for the likelihood depth and $(w_1, w_2, w_3) = (0.5, 0.5, 0)$ for the depth based on within-triplet distances. As in example (i), we generate a test sample of 50 random point-pairs from each of the two distributions in order to evaluate the performance of the classifiers. Figure 7 reports the rates of misclassifying the test samples using classifiers trained on the supervised data shown in Figure 5. The results are similar to the $n = 1$ case. For both depth functions, at least one of the control parameter settings yields misclassification rates close to, or even lower than, the Bayes rate 0.1814. Cross validation is again very effective in identifying the best settings of the control parameters.

5 Application to economic data

To illustrate the practical relevance of a shift in emphasis from "centrality" to "representativeness", we compare the depth function based on within-triplet distances

with the classical simplicial depth, calculated with respect to a set of bivariate economic data available from the World Bank. The dataset consists of observations on two World Development Indicators, namely the life expectancy at birth and the gross national income (GNI) per capita, covering 162 countries for the year 2008. The GNI has been converted to international dollars using purchasing power parity rates. Data depths are useful for ranking the 162 countries in order of the extent to which each country's development typifies a general world trend. For the case of within-triplet distances, we set $(w_1, w_2, w_3) = (0.09, 0.66, 0.25)$, which maximises (17) over the simplex $\{(w_1, w_2, 1 - w_1 - w_2) : w_1, w_2 \geq 0, w_1 + w_2 \leq 1\}$, with ϖ taken to be the uniform density function over the rectangle $[0, 65] \times [40, 90]$. Figures 8 and 9 display, using both 2D contours and 3D plots, the two depth functions calculated with respect to the data. The data points observed for the 162 countries, shown also on the contour plots, cluster in a crescent and do not exhibit clear unimodality.

As shown in Figure 8, the centre-outward ordering simplicial depth identifies a unique deep centre, near which can be located the three most "central" countries, namely Thailand, Ukraine and Belarus. Yet a closer look at their positions, which are somewhat peripheral relative to the main data crescent, casts doubt on the representativeness of these three countries, despite their apparent centrality as revealed by the simplicial depth. Indeed, Thailand and Ukraine are rather atypical of the

world trend in view of their relatively short life expectancies compared to countries having similar levels of GNI per capita.

On the other hand, we see in Figure 9 that the central region identified above by simplicial depth turns now into a conspicuous dip in the depth surface calculated using within-triplet distances, and is surrounded by deep areas more representative of the entire dataset. The three deepest, or for that matter most representative, countries are Estonia, Croatia and Hungary, all of which lie on one side of the central dip. The three most "central" countries previously found by simplicial depth are ranked 131 (Thailand), 124 (Ukraine) and 111 (Belarus) respectively by within-triplet distances, and can hardly be deemed representative of the world trend, a result in agreement with our actual observations.

We highlight on both Figures 8 and 9 the two extreme cases, namely Thailand and Liberia, where the most positive and negative differences between the two ranks are found. Thailand is ranked the deepest (Figure 8) by simplicial depth but only 131st (Figure 9) by within-triplet distances. Liberia, on the contrary, is ranked 29th (Figure 9) by within-triplet distances. The simplicial depth, however, finds Liberia among the four least deep, or most outlying, countries in the dataset, a somewhat counter-intuitive result (Figure 8).

This example reiterates again the importance and practical relevance of develop-

ing a new notion of data depth which can more satisfactorily "represent" observed data patterns, especially in the absence of clear unimodality.

6 Conclusion

We have proposed an alternative interpretation of data depth as a measure of "representativeness", which is arguably more relevant to many important applications of depth in statistical problems. In this new perspective, goodness-of-fit tests come naturally to the fore with their provision of test statistics which can at once be identified with an appropriate measure of representativeness. Such connection gives rise to a new method of defining data depth, which now applies not only to a single point of interest but also to any pattern of points. Our procedure thus provides an alternative motivation for some existing depth functions such as Tukey's depth and the likelihood depth, and introduces new classes of depth functions which broaden the scope of practical applications of data depth in general.

Although it is not our objective in this paper to recommend a definitive choice of data depth, which is without doubt specific to the problem in hand, our numerical examples suggest that depth functions derived from a consideration of interpoint distances, especially those based on within-triplet distances, possess nice properties so far as representativeness is concerned under multimodal situations. Based as they

are on calculation of interpoint distances alone, these depth functions also enjoy the advantage of being computationally more efficient than the local simplicial depth which involves the handling of simplices, especially in high-dimensional settings. For the setting of control parameters of depth functions based on interpoint distances, we have proposed to maximise a correlation measure between the depth function and the underlying distribution or, in the special context of depth-based classification, to minimise a cross-validated estimate of the misclassification rate. Both approaches have found satisfactory empirical support in our numerical examples.

Appendix

Proof of Proposition 1

For any $x \in \mathbb{R}$, let $r(x)$ satisfy $F(x + r(x)) - F(x - r(x)) = \gamma$. Differentiation of the latter condition twice with respect to x gives $\{f(x-r(x))+f(x+r(x))\}r'(x) = f(x-r(x))-f(x+r(x))$ and $\{f(x-r(x))+f(x+r(x))\}r''(x) = (1-r'(x))^2 f'(x - r(x)) - (1+r'(x))^2 f'(x+r(x))$. Thus $r(x)$ has a local minimum or local maximum at $x = x_0$ satisfying $r'(x_0) = 0$, that is $f(x_0 - r(x_0)) = f(x_0 + r(x_0))$, according as $r''(x_0) > 0$ or < 0 respectively. The proposition then follows by noting that $r''(x_0) = \{2f(x_0 - r(x_0))\}^{-1}\{f'(x_0 - r(x_0)) - f'(x_0 + r(x_0))\}$ and that $D(F, x)$ is a decreasing function of $r(x)$.

Proof of Proposition 2

The unimodality condition ensures that there exists some $w \in \mathbb{R}$ and $R > 0$ such that $F(w+R) - F(w-R) = \gamma$ and $f(w+R) = f(w-R) = k$, say. Then necessarily f increases at $w - R$, decreases at $w + R$ and $f(x) > k$ for all $x \in (w - R, w + R)$. Clearly the depth function (10) is locally maximised at w by Proposition 1. Fix any $x_1 > x_2 > w$ and let $R_1, R_2 > 0$ satisfy $F(x_i + R_i) - F(x_i - R_i) = \gamma$, $i = 1, 2$. Note that $x_2 - R_2 > w - R$ and $x_2 + R_2 > w + R$, or otherwise $[x_2 - R_2, x_2 + R_2]$ either contains or is contained in $[w-R, w+R]$ strictly, contrary to the definition of R_2. It follows that $f(x) > f(x_2 + R_2)$ for all $x \in [w - R, x_2 + R_2)$. Consider two cases: (i) $x_1 - R_2 > x_2 + R_2$, (ii) $x_1 - R_2 \leq x_2 + R_2$. Under (i), we have $f(x_1 - R_2) < f(x)$ for all $x \in [x_2 - R_2, x_2 + R_2]$, so that $\int_{x_1-R_2}^{x_1+R_2} f(u)\, du < \int_{x_2-R_2}^{x_2+R_2} f(u)\, du = \gamma$, which implies that $R_1 > R_2$. Under (ii), we have $f(x_2 + R_2) < f(x)$ for all $x \in [x_2 - R_2, x_1 - R_2)$, so that

$$\int_{x_1-R_2}^{x_1+R_2} f(u)\, du - \int_{x_2-R_2}^{x_2+R_2} f(u)\, du = \int_{x_2+R_2}^{x_1+R_2} f(u)\, du - \int_{x_2-R_2}^{x_1-R_2} f(u)\, du < 0.$$

It follows that $F(x_1 + R_2) - F(x_1 - R_2) < \gamma$ and hence $R_1 > R_2$. Thus we have, under either (i) or (ii), that the depth function at x_1 is strictly smaller than that at x_2, so that it decreases strictly on (w, ∞). Similar arguments show that it increases strictly on $(-\infty, w)$.

References

[1] Agostinelli, C. and Romanazzi, M. (2011). Local depth. *J. Statist. Plann. Inference*, **141**, 817–830.

[2] Alba-Fernández, V., Jiménez-Gamero, M. D. and Muñoz-García, J. (2008). A test for the two-sample problem based on empirical characteristic functions. *Comput. Statist. Data Anal.*, **52**, 3730–3748.

[3] Aslan, B. and Zech, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *J. Stat. Comput. Simul.*, **75**, 109–119.

[4] Baggerly, K. A. and Scott, D. W. (1999). Comment on "Multivariate analysis by data depth: description statistics, graphics and inference" by R. Y. Liu, J. M. Parelius and K. Singh. *Ann. Statist.*, **27**, 843–844.

[5] Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.*, **88**, 190–206.

[6] Bartoszyński, R., Pearl, D. K. and Lawrence J. (1997). A multidimensional goodness-of-fit test based on interpoint distances. *J. Amer. Statist. Assoc.*, **92**, 577–586.

[7] Cabaña, A. and Cabaña, E. M. (1997). Transformed empirical processes and modified Kolmogorov-Smirnov tests for multivariate distributions. *Ann. Statist.*, **25**, 2388–2409.

[8] Chen, Y., Dang, X., Peng, H. and Bart, H. L. J. (2009). Outlier detection with the kernelized spatial depth function. *IEEE Trans. Patt. Anal. Mach. Int.*, **31**, 288–305.

[9] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B*, **46**, 440–464.

[10] Cuesta-Albertos, J. A., Fraiman, R. and Ransford, T. (2006). Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bull. Braz. Math. Soc. New Ser.*, **37**, 477–501.

[11] Cuesta-Albertos, J. and Nieto-Reyes, A. (2008). A random functional depth. *Functional and Operatorial Statistics* (S. Dabo-Niang and F. Ferraty, eds.), Springer, pp. 121–126.

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

[13] Friedman, J. and Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Sminov two-sample tests. *Ann. Statist.*, **7**, 697–717.

[14] Ghosh, A. K. and Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scand. J. Statist.*, **32**, 327–350.

[15] Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.*, **16**, 772–783.

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

[17] Lange, T., Mosler, K. and Mozharovskyi, P. (2012). Fast nonparametric classification based on data depth. *Statist. Papers*. doi: 10.1007/s00362-012-0488-4.

[18] Li, J., Cuesta-Albertos, J. A. and Liu, R. Y. (2012). *DD*-classifier: nonparametric classification procedure based on *DD*-plot. *J. Amer. Statist. Assoc.*, **107**, 737–753.

[19] Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.*, **18**, 405–414.

[20] Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Acad. Sci. India*, **12**, 49–55.

[21] Makhoukhi, M. B. (2008). An approximation for the power function of a nonparametric test of fit. *Statist. Probab. Lett.*, **78**, 1034–1042.

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

[23] Præstgaard, J. T. (1995). Permutation and bootstrap Kolmogorov-Smirnov tests for the equality of two distributions. *Scand. J. Statist.*, **22**, 305–322.

[24] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. Roy. Statist. Soc. Ser. B*, **67**, 515–530.

[25] Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.*, **81**, 799–806.

[26] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

[27] Singh, K. (1991). A notion of majority depth. Technical Report, Rutgers University, Department of Statistics.

[28] Tukey, J. W. (1975). Mathematics and picturing data. Proceedings of the International Congress on Mathematics, vol. 2, 523–531.

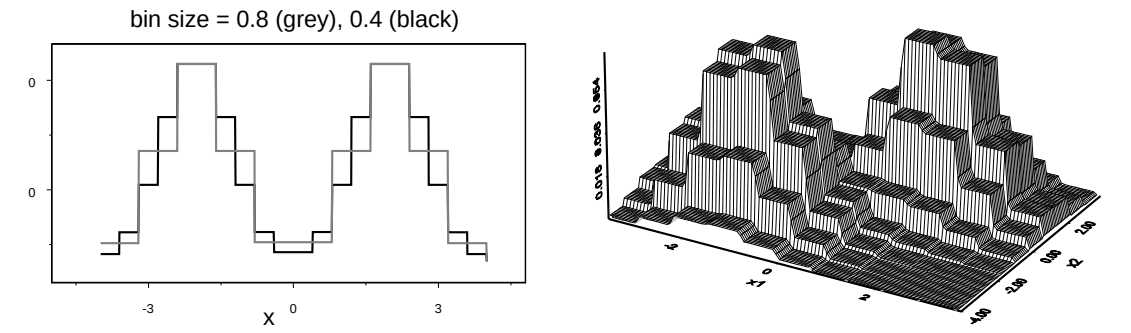
[29] Wolfowitz, J. (1954). Generalization of the theorem of Glivenko-Cantelli. *Ann. Math. Statist.*, **25**, 131–138.

[30] Zhu, L.-X., Fang, K.-T. and Bhatti, M. I. (1997). On estimated projection pursuit-type Cramér-von Mises statistics. *J. Multivariate Anal.*, **63**, 1–14.

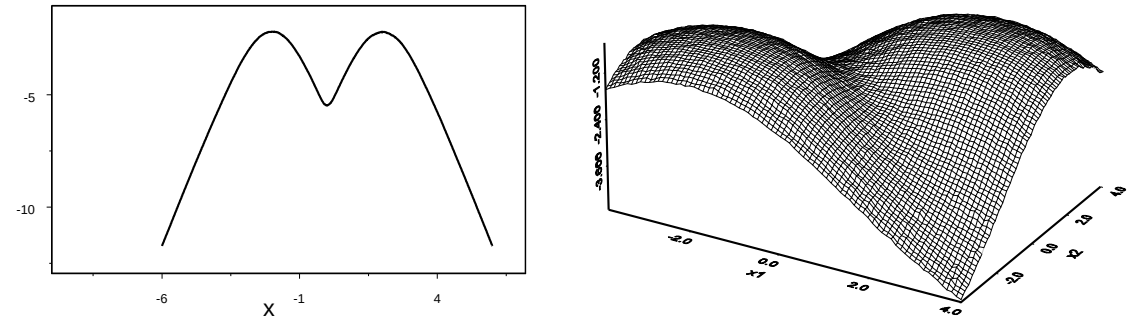
[31] Zuo, Y. and Serfling, R. (2000). Structural properties and convergence results for contours of sample statistical depth functions. *Ann. Statist.*, **28**, 483–499.

$d = 10$ $d = 2$

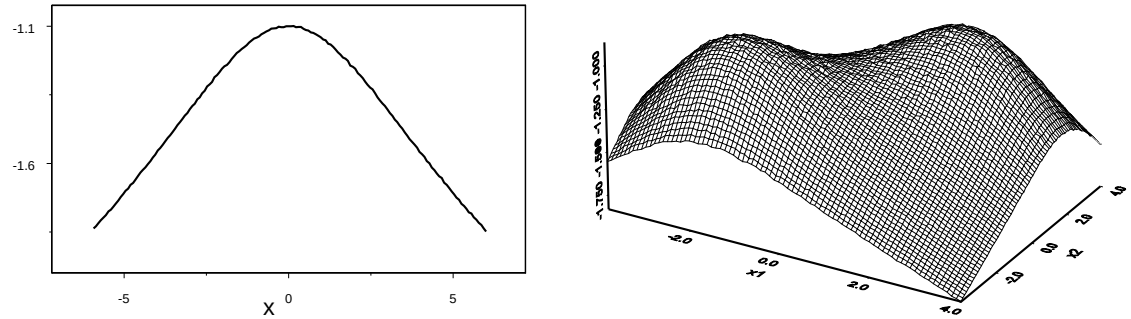
Pearson's chi-squared tests



Nearest neighbours: $\gamma = 0.05$



Energy tests: $R(r) = -\ln(r)$



Within-triplet distances: $(w_1, w_2, w_3) = (0.5, 0.5, 0)$

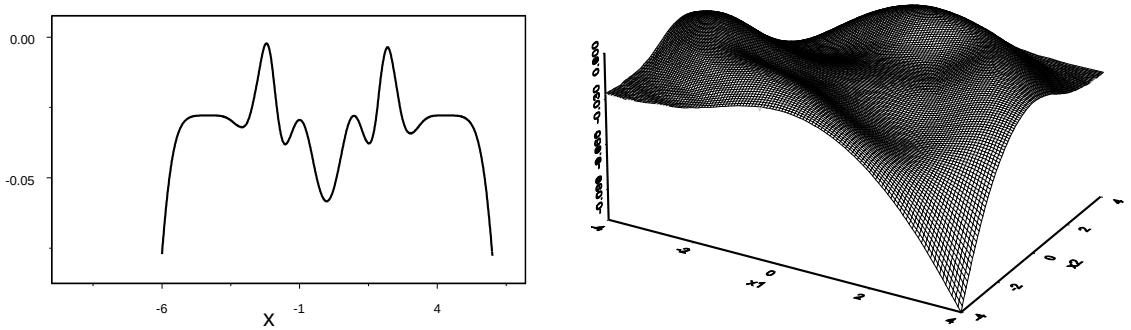


Figure 1: Depth of a single point $[x, \ldots, x]^{\mathrm{T}} \in \mathbb{R}^{10}$ (left panel) and $x \in \mathbb{R}^2$ (right panel), under a normal mixture with two equal modes.

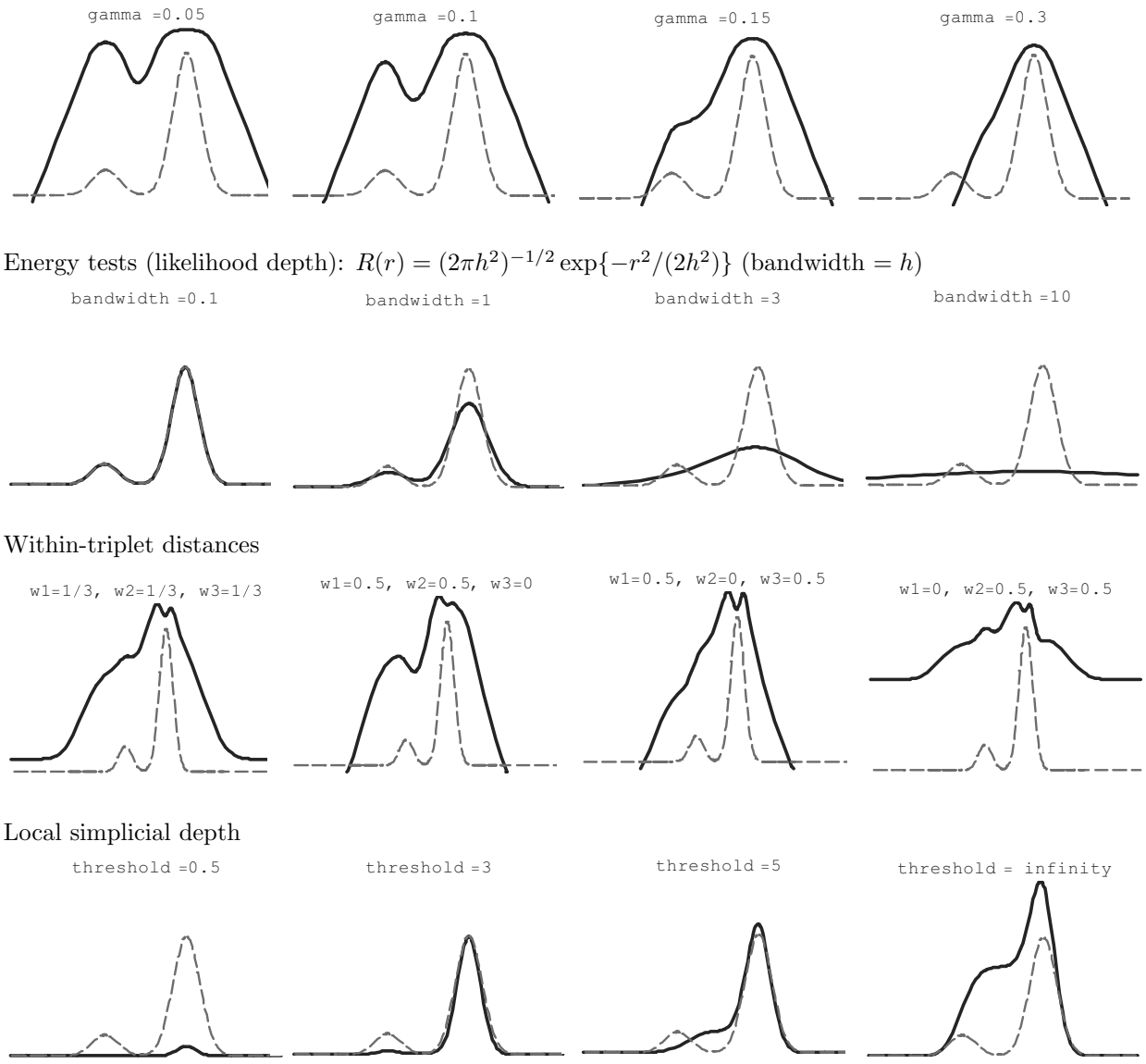
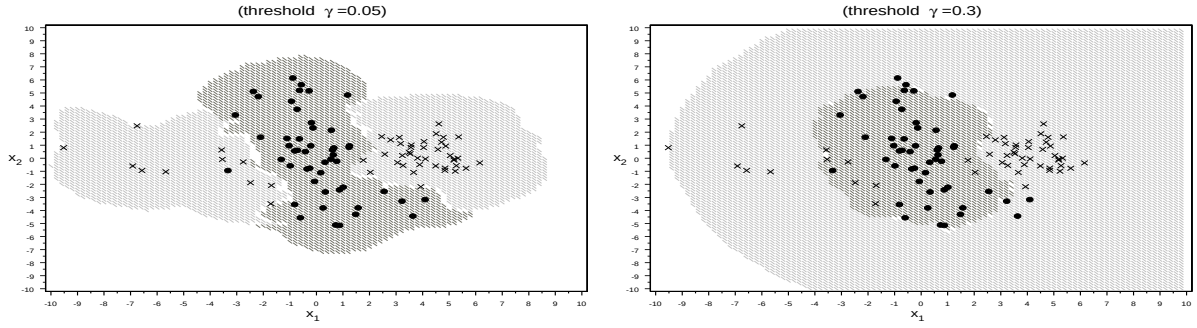
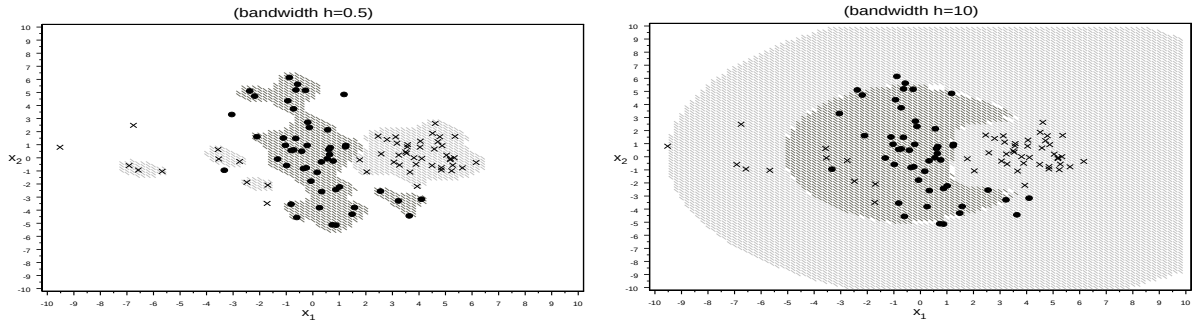


Figure 2: Depth of a single point $x \in \mathbb{R}$ under a normal mixture with two unequal modes. The grey dashed curves indicate the normal mixture density plotted on the same x scale.

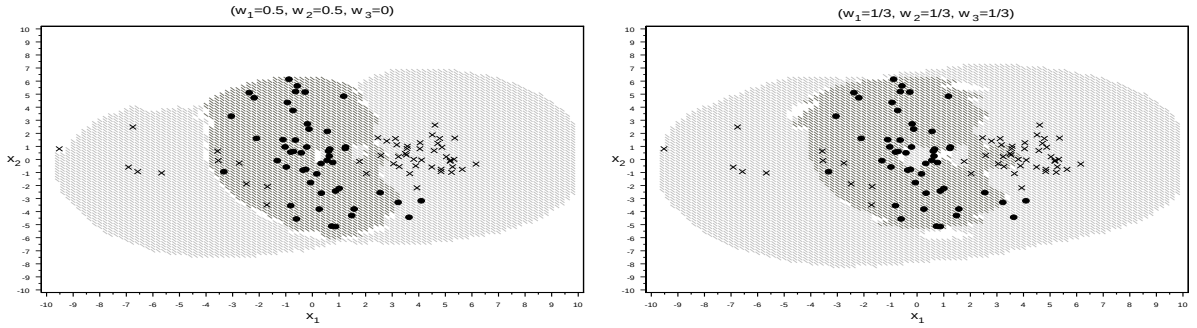
Nearest neighbours



Energy tests (likelihood depth): $R(r) = (2\pi h^2)^{-1/2} \exp\{-r^2/(2h^2)\}$ (bandwidth $= h$)



Within-triplet distances



Local simplicial depth

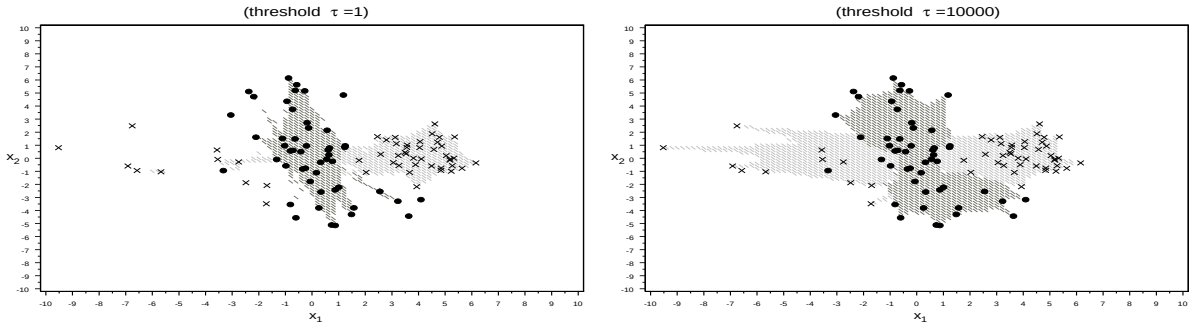


Figure 3: Classification of $x = [x_1, x_2]^{\mathrm{T}} \in \mathbb{R}^2$ to a bivariate normal mixture with two modes (light grey region) and a bivariate zero-mean normal distribution (dark grey region). Training samples, each of size 50, are indicated by "$\times$" and "$\bullet$" for the two distributions respectively.

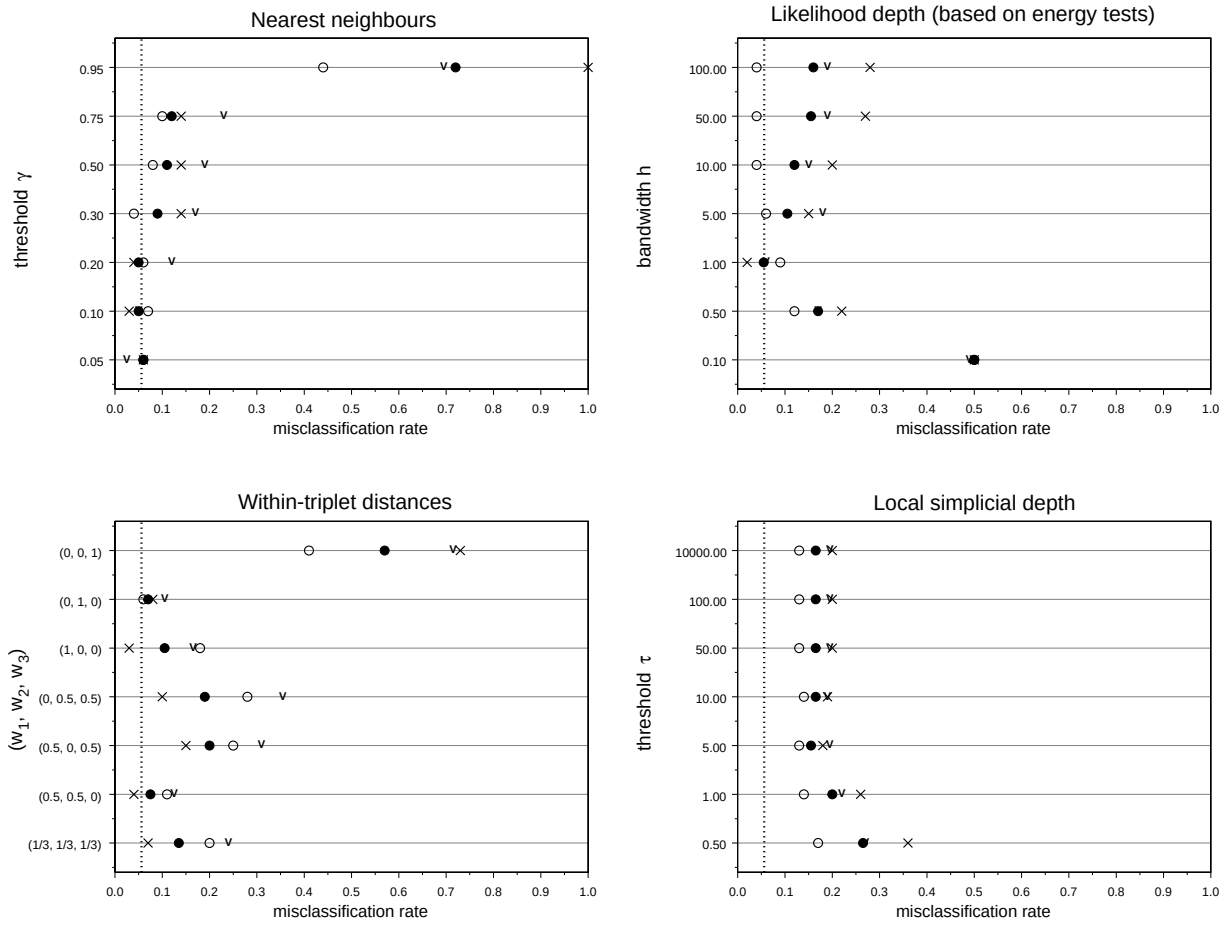


Figure 4: Rates of misclassification: (i) Bayes rate (dotted vertical line); (ii) leave-one-out cross-validated estimates based on training data ("V"); (iii) F_1 misclassified as F_2 ("○") based on test sample of 50 observations from F_1; (iv) F_2 misclassified as F_1 ("×") based on test sample of 50 observations from F_2; (v) average of (iii) and (iv) ("●").

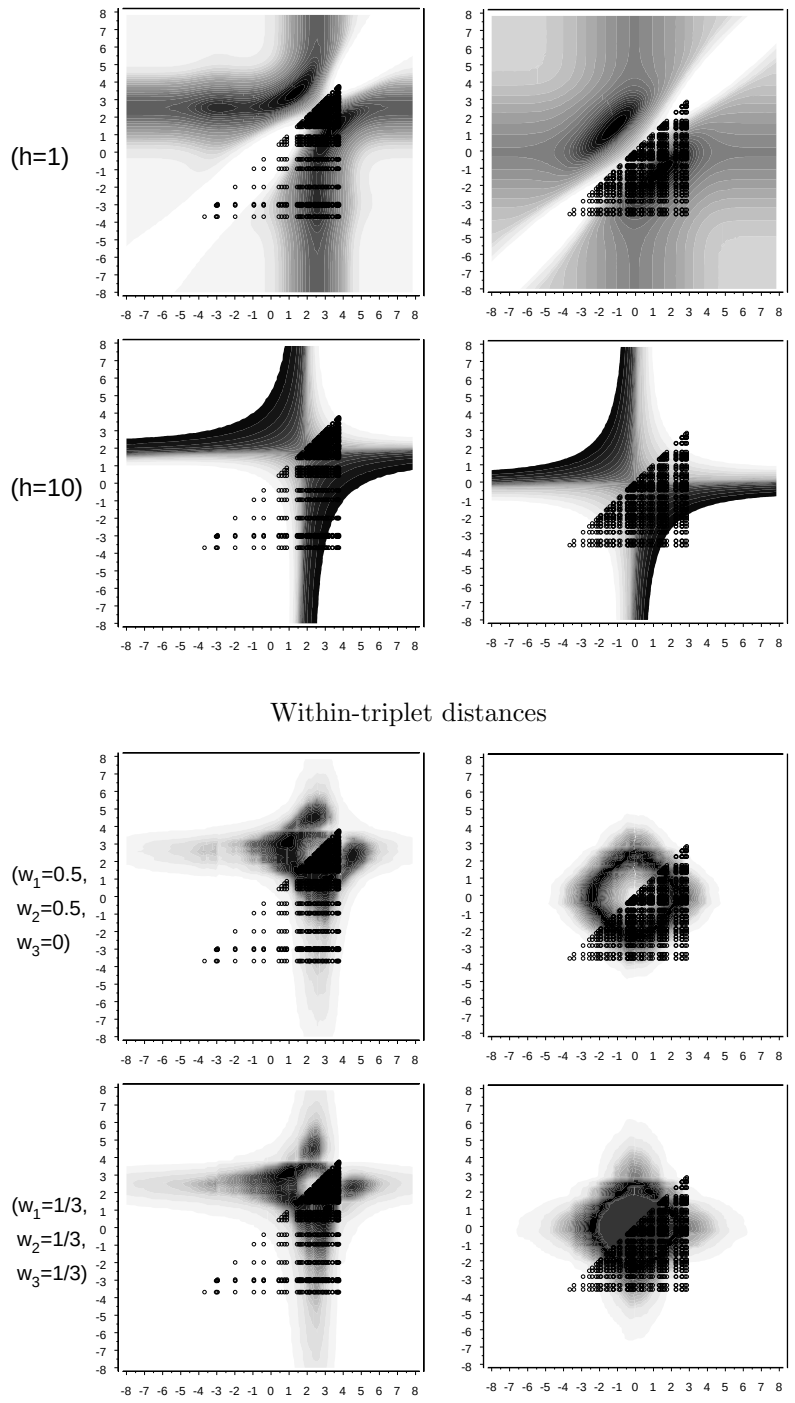
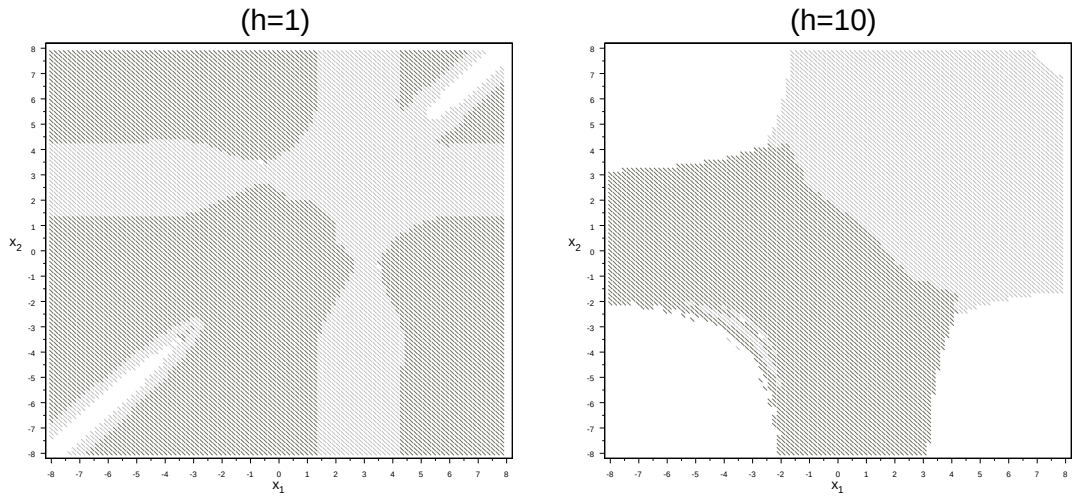


Figure 5: Depth of point-pair (x_1, x_2) with reference to random sample $X_1, \ldots, X_{50} \in \mathbb{R}$. Circles in triangular patterns indicate positions of $\{(X_i, X_j) : X_i \geq X_j, \;\; i, j = 1, \ldots, 50\}$. Samples 1 and 2 are drawn from a normal mixture with 2 modes and a zero-mean normal distribution, respectively.

Energy tests (likelihood depth): $R(r) = (2\pi h^2)^{-1/2} \exp\{-r^2/(2h^2)\}$ (bandwidth $= h$)



Within-triplet distances

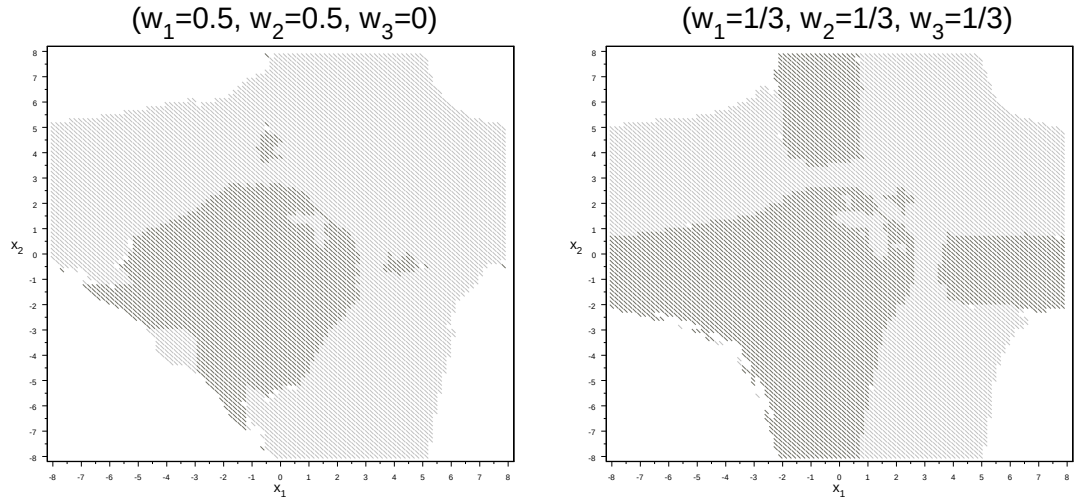


Figure 6: Classification of point-pair (x_1, x_2) to a normal mixture with two modes (light grey region) and a zero-mean normal distribution (dark grey region).

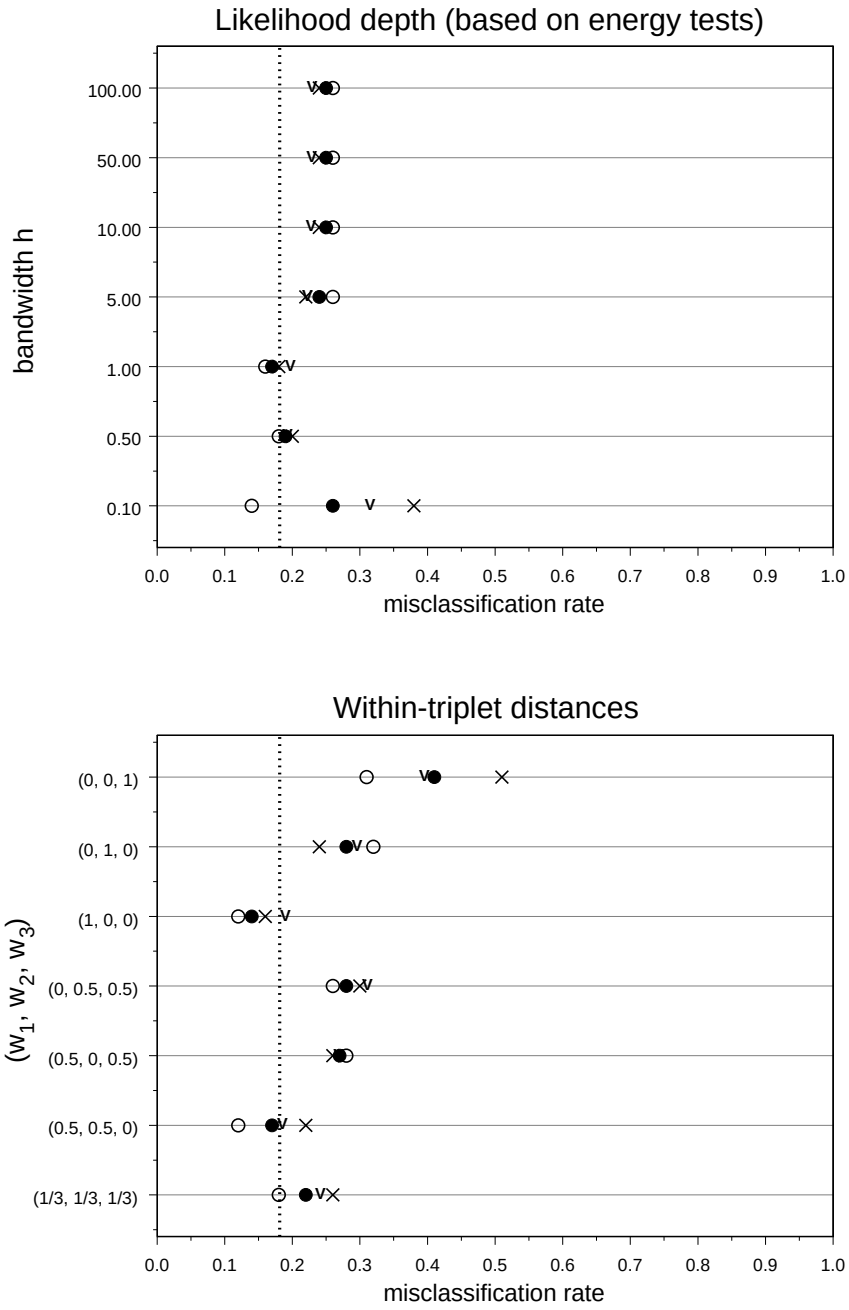


Figure 7: Rates of misclassification: (i) Bayes rate (dotted vertical line); (ii) leave-two-out cross-validated estimates based on training data ("V"); (iii) F_1 misclassified as F_2 ("$\circ$") based on test sample of 50 point-pairs from F_1; (iv) F_2 misclassified as F_1 ("$\times$") based on test sample of 50 point-pairs from F_2; (v) average of (iii) and (iv) ("$\bullet$").

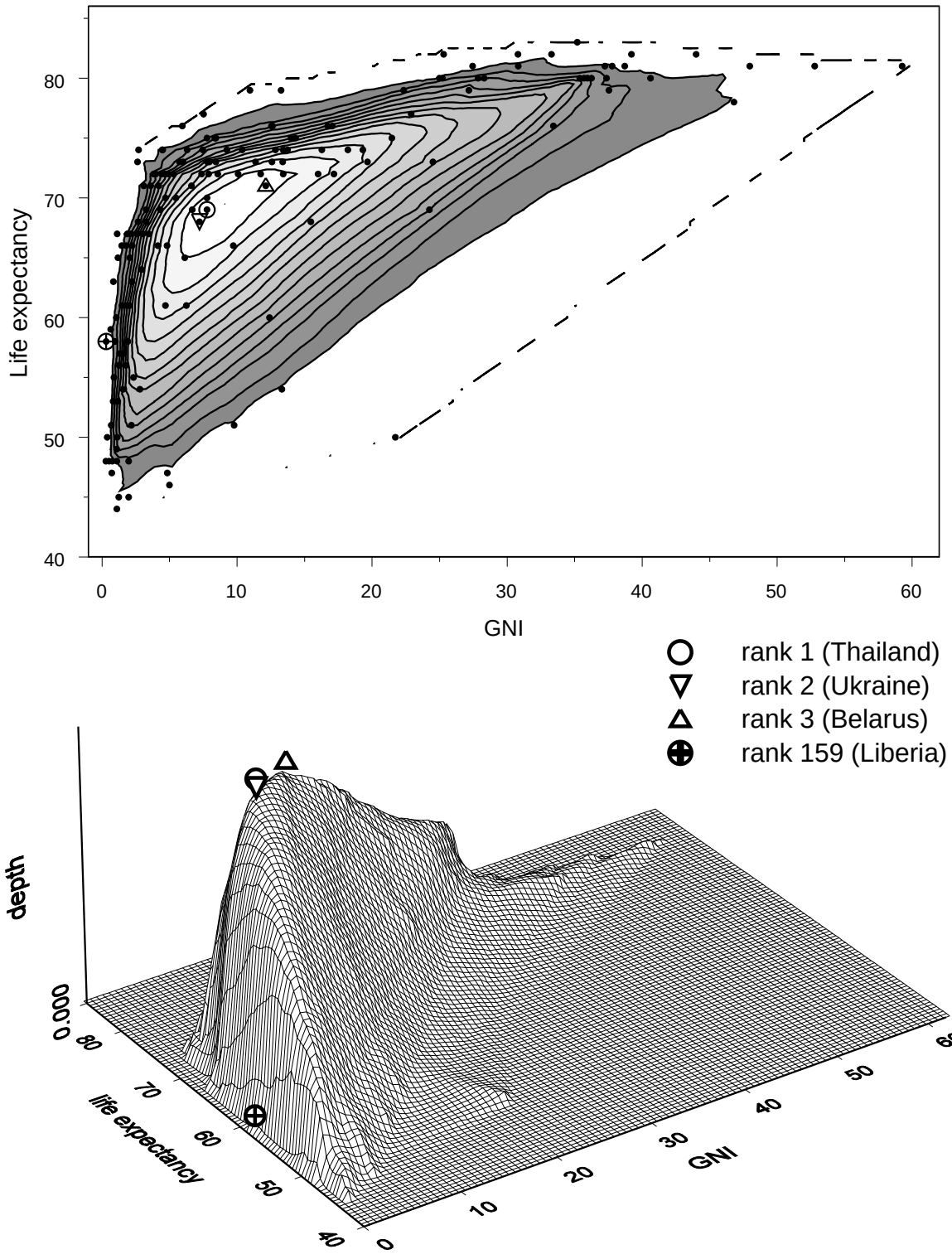


Figure 8: World Bank data — simplicial depth plots with respect to life expectancy and GNI indicators of 162 countries in 2008.

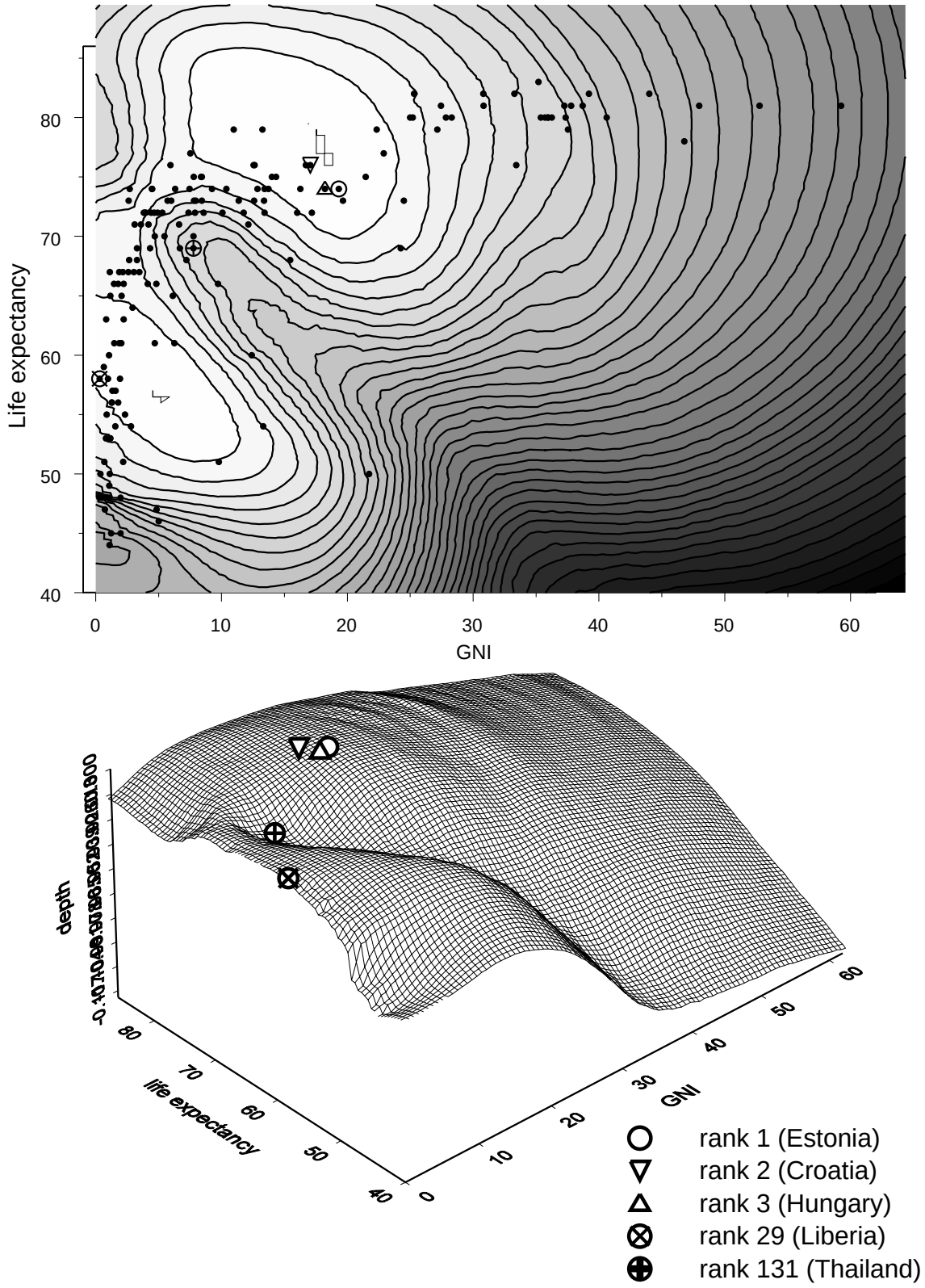


Figure 9: World Bank data — depth function plots, based on within-triplet distances with $(w_1, w_2, w_3) = (0.09, 0.66, 0.25)$, with respect to life expectancy and GNI indicators of 162 countries in 2008.