The Random Tukey Depth*

J.A. Cuesta-Albertos and A. Nieto-Reyes
Departamento de Matemáticas, Estadística y Computación,
Universidad de Cantabria, Spain

January 28, 2008

Abstract

The computation of the Tukey depth, also called halfspace depth, is very demanding, even in low dimensional spaces, because it requires that all possible one-dimensional projections be considered. In this paper we propose a random depth which approximates the Tukey depth. It only takes into account a finite number of one-dimensional projections which are chosen at random. Thus, this random depth requires a reasonable computation time even in high dimensional spaces. Moreover, it is easily extended to cover the functional framework.

We present some simulations indicating how many projections should be considered depending among others on the kind of problem, sample size and dimension of the sample space. We also compare this depth with some others proposed in the literature. It is noteworthy that the random depth, based on a very low number of projections, obtains results very similar to those obtained with other depths.

Key words and phrases: Random Tukey depth, one-dimensional projections, multidimensional data, functional data, homogeneity test, supervised classification.
A.M.S. 1980 subject classification: Primary 62H05; Secondary: 62G07, 62G35.

1 Introduction

This paper is written in the same spirit as [8]. In the abstract of this paper, D.J. Hand states that *"...simple methods typically yield performance almost as good as more sophisticated methods to the extent that the difference in performance may be swamped by other sources of uncertainty..."*. Hand's work is related to classification techniques. Here we analyze a conceptually simple and easy to compute multidimensional depth that can be applied to functional problems and that provides results comparable to those obtained with more involved depths.

*Research partially supported by the Spanish Ministerio de Ciencia y Tecnología, grant MTM2005-08519-C02-02 and the Consejería de Educación y Cultura de la Junta de Castilla y León, grant PAPIJCL VA102/06.

Depths are intended to order a given set in the sense that if a datum is moved toward the center of the data cloud, then its depth increases and if the datum is moved toward the outside, then its depth decreases.

More generally, given a probability distribution P defined in a multidimensional (or even infinite-dimensional) space $\mathcal{X}$, a depth tries to order the points in $\mathcal{X}$ from the "center (of P)" to the "outer (of P)". Obviously, this problem includes data sets if we consider P as the empirical distribution associated to the data set at hand. Thus, in what follows, we will always refer to the depth associated to a probability distribution P.

In the one-dimensional case, it is reasonable to order the points using the order induced by the function

$$x \to D_1(x, P) := \min\{P(-\infty, x], P[x, \infty)\}. \tag{1}$$

Thus, the points are ordered following the decreasing order of the absolute values of the differences between their percentiles and 50, and the deepest points are the medians of P.

Several multidimensional depths have been proposed (see, for instance, the recent book [10]) but here we are mainly interested in the *Tukey (or halfspace) depth* (see [17]). If $x \in \mathbb{R}^p$, then, the Tukey depth of x with respect to P, $D_T(x, P)$, is the minimal probability which can be attained in the closed halfspaces containing x. According to [18], this depth behaves very well in comparison with various competitors.

An equivalent definition of $D_T(x, P)$ is the following. Given $v \in \mathbb{R}^p$, let Π_v be the projection of $\mathbb{R}^p$ on the one dimensional subspace generated by v. Thus, $P \circ \Pi_v^{-1}$ is the marginal of P on this subspace, and it is obvious that

$$D_T(x, P) = \inf\{D_1(\Pi_v(x), P \circ \Pi_v^{-1}) : v \in \mathbb{R}^p\},\ x \in \mathbb{R}^p. \tag{2}$$

I.e., $D_T(x, P)$ is the infimum of all possible one-dimensional depths of the one-dimensional projections of x, where those depths are computed with respect to the corresponding (one-dimensional) marginals of P.

Some other depths based on the consideration of all possible one-dimensional projections have been proposed (see, for instance, [19]). We consider that what follows could be applied to all of them, but we have chosen the Tukey depth to test it specifically.

Perhaps the most important drawback of the Tukey depth is the required computational time. This time is more or less reasonable if $p = 2$, but it becomes prohibitive even for $p = 8$ [15, pag. 54]. To reduce the time, in [20, pag. 2234] what we propose is to approximate their values using randomly selected projections.

On the other hand, in [5], a random depth is defined. In this paper, given a point x, the authors propose to choose at random a finite number of vectors $v_1, ..., v_k$, and then, take as depth of x the mean of the values $D_1(\Pi_{v_i}(x), P \circ \Pi_{v_i}^{-1})$, $i = 1, ..., k$.

Our approach follows more closely the suggestion in [20]: we simply replace the infimum in (2) by a minimum over a finite number of randomly chosen projections, obtaining a random approximation to the Tukey depth. Moreover, in Section 2 (Theorem 2.3) we show that this approximation satisfies the definition of depth given in [18] (but with the convergence being in probability) and, then, it can also be considered as a depth, which we call the *random Tukey depth*. Section 2 closes with Theorem 2.4, where the consistency

of the random Tukey depth is proved. The proofs of these theorems are included in the Appendix.

The main problem of the random Tukey depth as an approximation to the Tukey depth is to find the number of random projections required to obtain a good approximation. This question is addressed in Section 3. This number could depend on the kind of application of the depth in which we are interested as well as on the dimension of the underlying space and on the size of the random sample we are employing. However, the simulations carried out in Section 3 suggest that a maximum of 100 randomly chosen projections are enough for a wide range of conditions. Section 3 ends with a comparison of the time required to compute the random Tukey depth and the time to compute the Mahalanobis depth.

One of the main advantages of the random Tukey depth is that it can be extended to infinite-dimensional functional spaces despite the definition of depth not being fully satisfied in this case. This is studied in Section 4. There we also apply the random Tukey depth to a well known supervised classification problem where we are required to classify an individual, as female or male, based on its growth curve. In addition, we compare the random Tukey depth with the depths proposed in [12] as well as with classification methods based on the k-nearest neighbors (k-NN) and kernels. Moreover, taking into account that, in fact, the data are only 31-dimensional, we also compare it with the random forest method.

The computations of the random forests have been done with a software downloaded from http://www.cs.waikato.ac.nz/ml/weka. The other computations have been carried out with MatLab. Computational codes are available from the authors upon request.

2 The random Tukey depth

In this section we define the random Tukey depth and show that it satisfies the definition of statistical depth given in [18] and its consistency.

Concerning the notation, in this section $\mathcal{P}$ denotes the class of distributions on the Borel sets of $\mathbb{R}^p$ and P_X the distribution of a general random vector X. In addition, the symbols $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ respectively denote the usual norm and scalar product in $\mathbb{R}^p$.

Now, let us formally define the random Tukey depth.

Definition 2.1 *Let $P \in \mathcal{P}$. Let $\nu \in \mathcal{P}$ absolutely continuous, and let $v_1, ..., v_k$ be independent and identically distributed random vectors with distribution ν. The random Tukey depth of $x \in \mathbb{R}^p$ with respect to P based on k random vectors chosen with ν is*

$$D_{T,k,\nu}(x, P) = \min\{D_1(\Pi_{v_i}(x), P \circ \Pi_{v_i}^{-1}) : i = 1, ..., k\},\ x \in \mathbb{R}^p.$$

In order to simplify the notation, unless some risk of confusion appears, we will delete the subscript ν in the notation and simply write $D_{T,k}$.

Obviously, $D_{T,k}(x, P)$ is a random variable. It may seem somewhat strange to take a random quantity to measure the depth of a point, which is inherently not-random. We have two reasons for taking this point of view.

Firstly, Theorem 4.1 in [4] shows that if P and Q are probability distributions on $\mathbb{R}^p$, such that the absolute moments $m_k := \int \|x\|^k dP(x)$ are finite and satisfy $\sum_{k\geq 1} m_k^{-1/k} = \infty$, ν is an absolutely continuous distribution on $\mathbb{R}^p$ and

$$\nu\{v \in \mathbb{R}^p : P \circ \Pi_v^{-1} = Q \circ \Pi_v^{-1}\} > 0,$$

then $P = Q$. In other words, if we have two different distributions, and we randomly choose a marginal of them, those marginals are almost surely different. According to this result, one randomly chosen projection is enough to distinguish between two p-dimensional distributions. Since the depths determine one-dimensional distributions, a depth computed on just one random projection allows us to distinguish between two distributions.

Secondly, if the support of ν is $\mathbb{R}^p$, and, for every k, $\{v_1, ..., v_k\} \subset \{v_1, ..., v_{k+1}\}$, then

$$D_{T,k}(x, P) \geq D_{T,k+1}(x, P) \to D_T(x, P), \quad \text{a.s.} \tag{3}$$

Therefore, if we choose a large enough k, the effect of the randomness in $D_{T,k}$ will be negligible. Of course, the question of interest is to learn how large k must be, because too large values of k would make this definition useless. We will analyze that point in Sections 3 and 4.

Here, we will show that, for every k, $D_{T,k}$ a.s. satisfies the definition of statistical depth function given by Zuo and Serfling in [18]. This definition consists of four key properties desirable for depths. They are affine invariance, maximality at center, monotonicity relative to deepest point and vanishing at infinity. In [18], it is shown that the Tukey depth satisfies this definition.

Concerning the maximality at center, this states that, having a distribution with a unique center of symmetry (with respect to some notion of symmetry), the depth function should attain the maximum at this center.

Definition 2.2 *The bounded and nonnegative mapping $D(\cdot, \cdot) : \mathbb{R}^p \times \mathcal{P} \longrightarrow \mathbb{R}$ is called a* statistical depth function *if it satisfies the following properties:*

1. *$D(Ax + b, P_{AX+b}) = D(x, P_X)$ holds for any $\mathbb{R}^p$-valued random vector X, any $p \times p$ nonsingular matrix A and any $b \in \mathbb{R}^p$.*
2. *$D(\theta, P) = \sup_{x \in \mathbb{R}^p} D(x, P)$ holds for any $P \in \mathcal{P}$ having center θ.*
3. *For any $P \in \mathcal{P}$ having deepest point θ, $D(x, P) \leq D(\theta + \alpha(x - \theta), P)$ holds for $\alpha \in [0, 1]$.*
4. *$D(x, P) \to 0$ as $\|x\| \to \infty$, for each $P \in \mathcal{P}$.*

Concerning point 2, various notions of symmetry are possible, among them, central, angular and halfspace symmetry. As central symmetry implies angular, which implies halfspace, we will identify center with the point of halfspace symmetry.

Theorem 2.3 *The random Tukey depth is a bounded and non-negative mapping which satisfies items 1, 2 and 3 in Definition 2.2.*

Moreover, let $P \in \mathcal{P}$ and $k > 0$. If $\|x\| \to \infty$ with $x \in \mathbb{R}^p$, then $D_{T,k}(x, P)$ converges to zero in probability.

Remark 2.3.1 As a consequence of Theorem 2.3 we can say that the random Tukey depth is a depth in the sense of Definition 2.2 if we replace the convergence in item 4 by convergence in probability.

The randomness only affects item 4. The problem is that it would be possible for all the k vectors to be included in the same hyperplane, and, in this case, it is obvious that property 4 is not satisfied, for instance, for some sequence of points orthogonal to this hyperplane, if $D_{T,k}(0, P) > 0$.

Another desirable property for depths is that its sample version converges to the population counterpart. Slightly more general than this is that, almost surely, $\sup_x |D(x, P_n) - D(x, P)| \to 0$ where P_n denotes the empirical distribution (i.e. if $x_1, ..., x_n$ is a random sample, $P_n[A] = \#(A \cap \{x_1, ..., x_n\})/n$). This property is satisfied by the Tukey depth (see [18]) and Theorem 2.4 shows that the random Tukey depth also enjoys this property.

Theorem 2.4 *Let $\nu \in \mathcal{P}$ and $v_1, ..., v_k$ be independent and identically distributed random vectors with distribution ν. Let $P \in \mathcal{P}$ and let $\{P_n\}$ be a sequence of empirical distributions computed on a random sample taken from P which is independent of the vectors $v_1, ..., v_k$.*

Then, conditionally on $v_1, ..., v_k$, we have that

$$\sup_{x \in \mathbb{R}^p} |D_{T,k}(x, P_n) - D_{T,k}(x, P)| \to 0, \text{ almost surely [P].}$$

Remark 2.4.1 In Theorem 2.4, the almost surely convergence is with respect to the empirical samples taken from P and the random vectors employed in the computation of the depths are chosen independently of these samples.

In fact, this result holds for every fixed vector in $\mathbb{R}^p$, randomly chosen with the distribution ν or not, with the only condition that it be independent of the random sample taken from P.

3 How many random projections? Testing homogeneity

Obviously, Theorem 4.1 in [4] also holds if ν is a probability distribution absolutely continuous with respect to the surface measure on the unit sphere in $\mathbb{R}^p$. Then, in this section, we fix ν to be the uniform distribution on the unit sphere to analyze the question of the selection of k. Our proposal is to make this selection depending on the problem we have at hand; for instance with bootstrap (as in Section 3.1) or with cross-validation (as in Section 4). However, it is good to first have some idea about the range in which to look for this value.

The obvious way to do this is to make some comparisons between D_T and $D_{T,k}$ for several dimensions, sample sizes and distributions; however, the long computational times required to obtain D_T makes those comparisons impractical. Then, instead of doing these comparisons, we have placed ourselves in some situations in which the deepness of the points are clearly defined and can easily be computed with a different depth. However, a comparison from the point of view of the results is carried out in Section 3.1.

If P is an elliptical distribution with centralization parameter μ and dispersion matrix Σ, then P is centrally symmetric around μ and it seems that every reasonable depth should consider those points at the same Mahalanobis distance of μ to have the same depth, and that differences in depth should correspond with differences in Mahalanobis distance of μ. Then, in this situation, every depth should be a monotone function of the Mahalanobis depth [13], where, given $x \in \mathbb{R}^p$, this depth is

$$D_M(x, P) := \frac{1}{1 + (x - \mu)^t \Sigma^{-1} (x - \mu)}. \tag{4}$$

Therefore, we can have an idea about the right k in $D_{T,k}$ as follows: if P is elliptical, $D_T(\cdot, P)$, is a monotone function of $D_M(\cdot, P)$. Thus, from (3), the larger the k, the larger the resemblance between $D_{T,k}(\cdot, P)$ and a monotone function of $D_M(\cdot, P)$. However, there should exist a value k_0 from which this resemblance starts to stabilize. This is the value for k we are looking for.

Taking into account that depths only try to rank points according to their closeness to the center of P, it is reasonable to measure the resemblance between $D_{T,k}(\cdot, P)$ and $D_M(\cdot, P)$ looking only at the ranks of the points. This is equivalent to employing the Spearman correlation coefficient, ρ. Thus, the resemblance that we handle here is

$$r_{k,P} := \rho \left(D_{T,k}(X, P), D_M(X, P) \right), \tag{5}$$

where X is a random variable with distribution P.

If P is an elliptical distribution, then the function $k \to r_{k,P}$ is strictly increasing. We try to identify the point k_0 from which the increments become negligible.

However, in practice, we will not have a distribution P, but a random sample $x_1, ..., x_n$ taken from P. This leads us to replace P in (5) by the empirical distribution P_n.

To illustrate the behavior of the function $k \to r_{k,P_n}$, we have represented it in Figure 1 for different distributions, sample sizes and dimensions. In this figure, the first column corresponds to centered Gaussian distributions with covariance matrices with ones on the diagonal and 0.9 in all positions off-diagonal. Remaining columns in Figure 1 represent, from left to right, standard Gaussian distributions, distributions with independent double exponential marginals and with independent Cauchy marginals.

Dimensions and sample sizes vary in rows. We consider, from top to bottom, sample sizes $n = 25, 100$ for $\mathbb{R}^2$, $n = 50, 100$ for $\mathbb{R}^8$ and $n = 100, 500$ for $\mathbb{R}^{50}$. Thus, the case $n = 100$ (second, fourth and fifth rows) can be used to see how the dimension affects the function for a fixed sample size.

The last row is different. In this row we take advantage of the fact that we know the exact covariance matrix of the theoretical distribution, so, in row number seven Mahalanobis depth is computed with the exact value of Σ. In this case, we have taken $n = 500$ in $\mathbb{R}^{50}$.

A last comment is related to the computation of the location center and the dispersion matrix (except in the last row) of P_n, to be employed in D_M. Those parameters should depend on the distribution which generated the sample. What we mean is the following:

the covariance matrix is an appropriate parameter in the Gaussian and exponential case. But it is not adequate for the Cauchy distribution, where we have identified Σ with the robust covariance matrix proposed in [14, page 206]. Furthermore, we have replaced μ by the sample mean in the Gaussian case and by the coordinate-wise median in the exponential and Cauchy settings.

In the graphs k varies in the set $\{1, ..., 25\}$ in the first and second rows, in $\{1, ..., 100\}$ in third, fourth and fifth rows, and in $\{1, ..., 500\}$ in last two rows. Moreover, there are no obvious differences between using the theoretical covariance matrices or their estimation or between the case of independent marginals and dependent ones. We have checked more cases (not shown here) with similar results, among which we have analyzed some intermediate dimensions, other sample sizes, and dispersion matrices with 0.5 in all off-diagonal elements for the Gaussian, exponential and Cauchy distributions.

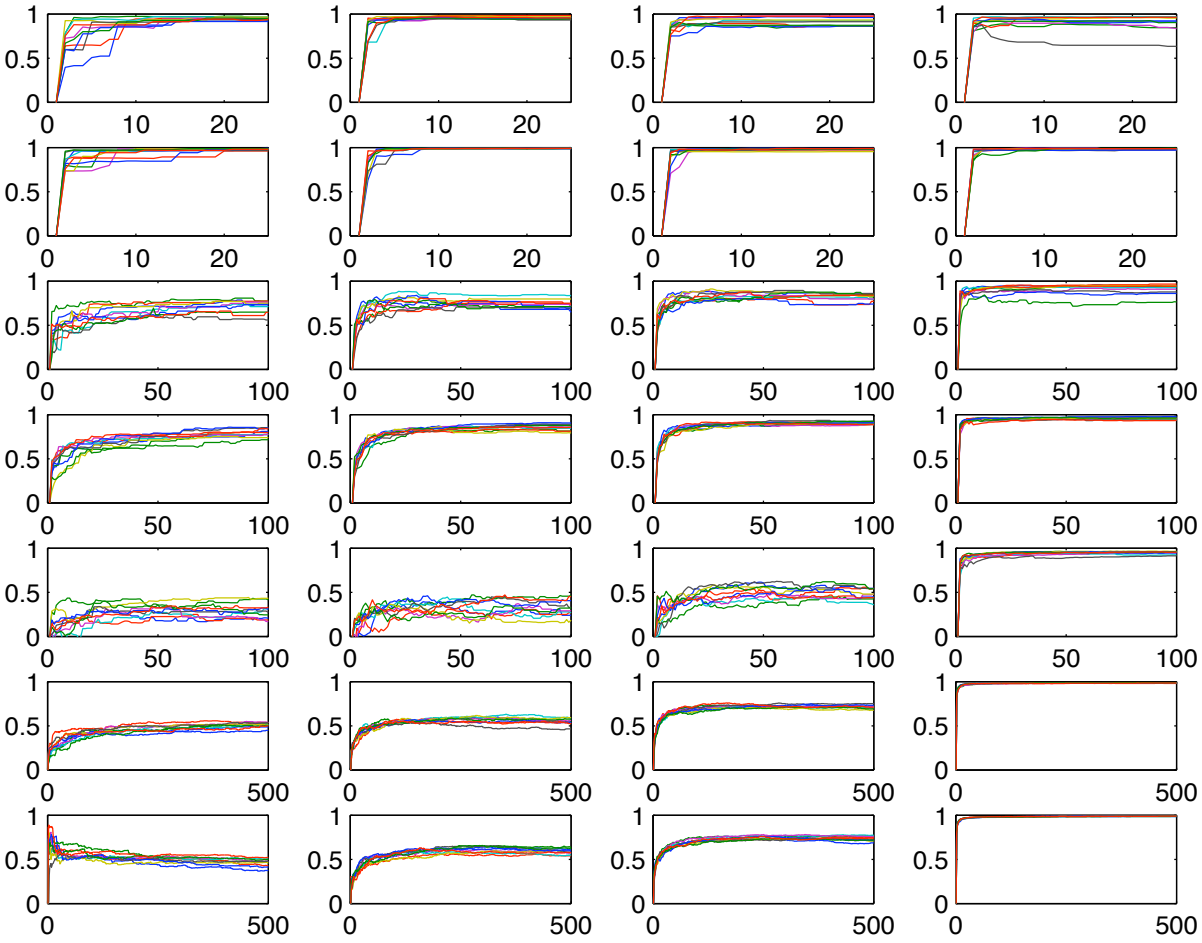

Figure 1: Representation of the function $k \to r_{k,P_n}$ defined in (5) for several dimensions, sample sizes and distributions. The underlying models are described in the text.

It seems that the graphs stabilize for $k \leq 10$ if $p = 2$, $k \leq 60$ if $p = 8$ and $k \leq 250$ if

$p = 50$ which are suitable values for computations and, of course, well below those usually employed to compute the Tukey depth.

An important fact which appears only in the graph in the lower-left corner of Figure 1 is that since P_n does not follow the model exactly, then the function r_{k,P_n} is not necessarily increasing and, in fact, sometimes, after an initial increasing part, it starts to decrease. In our opinion, the reason for this is that $\lim_k D_{T,k}(x, P_n) = D_T(x, P_n) \neq D_M(x, P_n)$.

3.1 Testing homogeneity

Our goal in this subsection is to show how the random Tukey depth, with values for k of the order suggested by Figure 1, provides results which are similar to those obtained in practice with the Tukey depth. In this section we will select k, with bootstrap, between the values $\{1, ..., 50\}$.

To this end, we are going to reproduce the simulation study carried out in [11], where the authors apply depth measures to test differences in homogeneity between two distributions. Let us begin by giving a brief description of the problem and the procedure. Additional details can be found in [11].

Assume that we have two random samples $\{X_1, ..., X_{n_1}\}$ and $\{Y_1, ..., Y_{n_2}\}$ taken from the centered distributions P and Q respectively. Let us assume that those distributions coincide except for a scale factor, i.e., we are assuming that there exists $r > 0$ such that the r.v.'s $\{rX_1, ..., rX_{n_1}\}$ and $\{Y_1, ..., Y_{n_2}\}$ are identically distributed. The problem consists in testing the hypotheses:

$$\begin{aligned} H_0: &\quad r = 1 \text{ (both scales are the same)} \\ H_a: &\quad r > 1 \text{ (Q has a larger scale).} \end{aligned}$$

The idea is that, under the alternative, the observations in the second sample should appear in the outside part of the joint sample $\{X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2}\}$, and, consequently, should have lower depths than the points in the first sample. Thus, it is possible to test H_0 against H_a by computing the depths of the points $\{Y_1, ..., Y_{n_2}\}$ in the joint sample, replacing them by their ranks and rejecting H_0 if those ranks are small.

The Wilcoxon rank-sum test can be used to test when the ranks of the points $\{Y_1, ..., Y_{n_2}\}$ are small. In [11] several possibilities to break the ties are proposed. We have tried all of them, with no relevant differences. Thus, we have chosen random tie-breaking as the only method to be shown here.

Concerning the selection of the number of random projections, we have employed bootstrap as follows. Let us assume that we want to carry out the homogeneity test at the level $\alpha = .05$ and that we have the samples $X := \{X_1, ..., X_{n_1}\}$ and $Y := \{Y_1, ..., Y_{n_2}\}$. We begin by selecting 50 vectors at random and fixing $r_b = 1.2$ as the first value of $r > 1$ that we consider interesting to be identified. Next, we apply bootstrap, 100 times, to the joint sample $Z := \{X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2}\}$ as follows:

We take, from Z, two bootstrap samples, I and J, respectively with sizes n_1 and and n_2. We center both in mean or median depending on the distribution we employed to generate the samples X and Y, and, then, multiply the vectors in J by r_b.

Now we compute the random Tukey depth of this new joint sample using the first of the 50 vectors selected and check if the null hypothesis is rejected. If not, we continue adding the following vector to the calculation of the random Tukey depth until the null hypothesis is rejected, and keep record of the number of vectors needed.

If the null hypothesis is not rejected despite using 50 vectors, we forget those samples and start with the following bootstrap couple. Remember that, the random Tukey depths are approximations of the Tukey depth. Since the Tukey depth is only able to detect the differences in homogeneity around 18% of times when $r = 1.2$ and $n_1 = n_2 = 20$ (see Table 3.1), we should expect that the random depths fails to reject 82% of times, and those cases are useless to determine k_0.

At the end of the bootstrap step, we have some values of k, which correspond to the cases in which we have rejected the homogeneity hypothesis. We have taken its 80% percentile as k_0 to be used in D_{T,k_0} to test the homogeneity in the sample at hand, Z. Notice that the value r_b is fixed at 1.2 in every case, independently of when the null hypothesis or the alternative holds.

In Table 3.1 we show the rate of rejections under the exposed conditions when we carry out the test for dimension $p = 2$ at the significance level $\alpha = .05$. The reason for doing the simulations only in dimension $p = 2$ is that our goal is to compare the results of the random Tukey depth with those of the Tukey depth, which can only be computed quite quickly when $p = 2$. Thus, the table also includes, between parenthesis, the rejection rates when the random depth is replaced by the Tukey depth computed using 1,000 directions uniformly scattered on the upper halfspace.

The distributions used in the simulations are the 2-dimensional standard Gaussian, and the double exponential and Cauchy with independent marginals. We have centered the samples from the Gaussian distribution in mean, and in component-wise median the samples from the double exponential and the Cauchy distributions. We have considered the values $r = 1, 1.2, 2$, and $n_1 = n_2 = n$ with $n \in \{20, 30, 100\}$, and have done 5,000 simulations for each combination of distribution, sample size and r.

As explained, the values of k depend on the samples we have each time. However, the medians of the obtained values along the repetitions have been $k = 2$, for all three distributions for sample size $n = 100$ and respectively $k = 3$ for $n = 30$ and $k = 4$ for $n = 20$, thus reinforcing the impression provided by Figure 1 that k should be well below 10 for dimension $p = 2$.

In [11] previous ideas are also applied to check the homogeneity between K samples, $K > 2$. The problem is the following. Let $\{X_{1,1}, ..., X_{1,n_1}\}, ..., \{X_{K,1}, ..., X_{K,n_K}\}$ be random samples obtained, respectively, from the distributions $P_1, ..., P_K$ and let us assume that there exist $r_1, ..., r_{K-1} > 0$ such that the random vectors $r_1X_{1,1}, ..., r_1X_{1,n_1}$,..., $r_{K-1}X_{K-1,1}, ..., r_{K-1}X_{K-1,n_{K-1}}$, $X_{K,1}, ..., X_{K,n_K}$ are identically distributed.

We are interested in testing the following hypotheses:

$$H_0: \quad r_i = 1, i = 1, ..., K-1 \text{ (all scales are the same)}$$
$$H_a: \quad \text{there exists } r_i \neq 1 \text{ (scales are different).}$$

If we center each sample separately, join all the observations in a single sample, compute

the depths of all the points and transform those depths in ranks, then, we can apply the Kruskal-Wallis test [9] to check if there are lacks of homogeneity between the ranks in each sub-sample.

Table 3.1 *Rate of rejections in 5,000 simulations using the random Tukey depth (between parenthesis, the rate with D_T) for the considered distributions, sample sizes and values of r. The dimension is $p = 2$. The significance level is .05.*

Sample size	Scale factor	Distribution					
		Cauchy		Gaussian		D. exponential	
$n = 20$	$r = 1$	.051	(.055)	.061	(.055)	.049	(.057)
	$r = 1.2$	.124	(.125)	.254	(.240)	.177	(.174)
	$r = 2$	.547	(.539)	.959	(.960)	.819	(.824)
$n = 30$	$r = 1$	.056	(.049)	.051	(.060)	.055	(.050)
	$r = 1.2$	.143	(.146)	.314	(.322)	.217	(.223)
	$r = 2$	.706	(.704)	.995	(.995)	.938	(.943)
$n = 100$	$r = 1$	.047	(.048)	.054	(.049)	.053	(.048)
	$r = 1.2$	.290	(.291)	.664	(.725)	.473	(.495)
	$r = 2$	.992	(.991)	1	(1)	1	(1)

We have carried out a simulation study applying previous procedure to the Tukey depth and to the random Tukey depth in the 2-dimensional case with Gaussian distributions, $K = 3$ and sample sizes $n_1 = n_2 = n_3 = n$, where $n \in \{20, 30\}$. We have carried out 5,000 replications in each case at the significance level $\alpha = .05$.

To select k, we have applied bootstrap in a similar way to that employed previously. The only difference is that now we have taken three bootstrap samples and that, after centering, we have multiplied just one of them by $r_b = 1.2$. On the other hand, the Tukey depth has been computed similarly to the previous case.

Results are shown in Table 3.2, where the results obtained applying the same procedure with the Tukey depth appear between parenthesis.

In this case, the median of k have been 4 for all cases with $n = 20$ and for $r_1 = r_2 = 1, 1.2$ and sample size $n = 30$. It was 3 when $n = 30$ and $r_1 = r_2 = 1.2$, $r_1 = 2$, $r_2 = 1.2$.

Table 3.2 *Rate of rejections in 5,000 simulations using $D_{T,k}$ (between parenthesis the rate with D_T) to test the homogeneity in three samples of Gaussian distributions with independent, identically distributed marginals and the exposed values of r. The dimension is $p = 2$. The significance level is .05.*

Covariance matrices	Sample sizes			
	$n = 20$		$n = 30$	
$r_1 = r_2 = 1$	.05	(.05)	.05	(.06)
$r_1 = r_2 = 1.2$	.15	(.15)	.20	(.21)
$r_1 = 2$, $r_2 = 1.2$	.89	(.89)	.98	(.98)
$r_1 = r_2 = 2$	.96	(.97)	1	(1)

The results of both studies in this subsection are quite encouraging, because there are no important differences between the rejection rates with both depths in spite of the comparatively low number of directions employed to compute the random Tukey depth.

3.2 Computational time

We end this section by paying some attention to the required computational time to compute the random Tukey depth. As a comparison we have selected the time to compute the Mahalanobis depth, which is one of the quickest depths according to Table 1 in [15].

In Table 3.3 we present the mean time, obtained from 200 simulations, employed to compute the random Tukey and Mahalanobis depths for all points in a sample with the shown sizes and dimensions. The numbers of employed random directions correspond with those obtained in Figure 1.

To make a reliable comparison between the computational times it is necessary to compare the time needed to compute the random Tukey depth on a sample with the one to compute the Mahalanobis depth on the same sample. The problem is, the first depth to be computed may have an advantage as the RAM memory may be cleaner than when the second depth is computed. In order to avoid this, we have computed the random Tukey depth first 100 times and the Mahalanobis depth first 100 times.

The computations have been carried out on a computer Xserve G5, PowerPC G5 Dual 2.3 GHz and 2Gb of RAM memory.

Table 3.3 *Time, in seconds, to compute the random Tukey and the Mahalanobis depths of all points in a sample with size n taken from a standard Gaussian distribution.*

Dimension	Random vectors	Sample size	Random Tukey	Mahalanobis
$p = 2$	$k = 10$	$n = 25$	$4.349 \cdot 10^{-4}$	.0014
		$n = 100$	$6.322 \cdot 10^{-4}$	.0024
$p = 8$	$k = 60$	$n = 50$	.0047	.0017
		$n = 100$	.0105	.0028
$p = 50$	$k = 250$	$n = 100$	.1153	.0047
		$n = 500$	.5596	.0158

It can be observed that the time needed to compute the random Tukey depth is acceptable in every case. Moreoever, it is better than the one needed to compute the Mahalanobis depth for low dimensions like $p = 2$, of the same order for $p = 8$ and worse for dimensions around 50.

4 Functional random Tukey depth. Functional classification

An interesting possibility of the random Tukey depth is that it can be straightforwardly extended to functional spaces. The only requirement of the main result in [4] is that the

sample space be a separable Hilbert space. Thus, in this section we will assume that we are considering a distribution P defined on this kind of space.

To fix ideas, we will handle the space, $I\!H$, of square-integrable functions in a given interval which, after re-scaling, we can assume to be $[0,1]$. Thus, $I\!H = L^2[0,1]$ and given $f, g \in I\!H$ we have that $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$ and $\|f\| = \langle f, f\rangle^{1/2}$.

The random Tukey depth is not a statistical depth in the functional case. The proofs for items 1-3 in Definition 2.2 (with obvious modifications such as replacing matrices with linear operators) are the same as in Theorem 2.3. However, the following example shows that item 4 fails in this case even for statistical convergences.

Example 4.1 Let $\{\delta_n\}_n \subset \mathbb{R}^+$ with $\lim_n \delta_n = 0$. Let $x_n \in I\!H$ such that $x_n(t) = 1/\delta_n$ if $t \in [0, \delta_n)$ and zero otherwise. Obviously, $\|x_n\| = \delta_n^{-1/2}$ and, then $\lim_n \|x_n\| = \infty$.

Let us take ν equal to the distribution of the standard Brownian motion and let $P = \nu$. Obviously Theorem 4.1 in [4] works with this distribution.

If X is a random element with distribution ν, then $\langle x_n, X\rangle$ converges to zero in probability because

$$\begin{aligned} E|\langle x_n, X\rangle| &= E\left|\int_0^{\delta_n} X(t)\delta_n^{-1}dt\right| \\ &\leq \int_0^{\delta_n} E|X(t)|\delta_n^{-1}dt = \int_0^{\delta_n} (2t/\pi)^{1/2}\delta_n^{-1}dt \leq (2\delta_n/\pi)^{1/2}, \end{aligned} \tag{6}$$

where the last equality holds because the distribution of $X(t)$ is $N(0,t)$. Thus, if $v_1, ..., v_k \in I\!H$ are randomly chosen with distribution ν, we have that

$$\lim_n D_1(< v_i, x_n >, P \circ \Pi_{v_i}^{-1}) = D_1(0, P \circ \Pi_{v_i}^{-1}) = \max_x D_1(x, P \circ \Pi_{v_i}^{-1}) = 2^{-1},$$

because $P \circ \Pi_{v_i}^{-1}$ is a centered Gaussian distribution.

Thus, in this setting the following results hold. Their proofs appear in the Appendix.

Theorem 4.2 *The random Tukey depth is a bounded and non-negative mapping which satisfies items 1, 2 and 3 in Definition 2.2.*

Theorem 4.3 *Let $v_1, ..., v_k \in I\!H$. Let P be a probability distribution on $I\!H$, and let $\{P_n\}$ be a sequence of empirical distributions computed on a random sample taken from P which is independent of the vectors $v_1, ..., v_k$.*

Then, conditionally on $v_1, ..., v_k$, we have that

$$\sup_{x \in \mathbb{R}^p} |D_{T,k}(x, P_n) - D_{T,k}(x, P)| \to 0, \text{ almost surely [P].}$$

Concerning the number of random directions to take, we follow the same procedure as in the finite dimensional case, i.e., the choice of method depends on the kind of problem at hand, for instance with bootstrap or cross-validation.

However, in this setting we have an additional problem. In the finite dimensional case, it seems reasonable to choose the random directions employing the uniform distribution on the sphere because of its invariance properties. Regrettably, in infinite dimensional spaces, there is no distribution with those nice properties, which makes the selection of the random directions more arduous.

An interesting possibility is to choose the distribution depending on the problem. This would lead designing a procedure which, given the problem, selects a distribution with some optimality properties.

In the subsection which follows we only try to give some hints as to what results could be obtained with a complete development of the theory, since the work required to do this exceeds the limits of this paper. There, we have taken first ν equal to the distribution of the standard Brownian motion. Then, we have tried some modifications of this distribution, which have improved the behavior of the procedure. Moreover, since those modifications are, in fact, parameter-dependent, we have chosen the values of the parameters with cross-validation.

4.1 Application to classification

Here we study a real example. To compare our depth with some other functional depth, we have repeated the classification problem carried out in [12], where the authors handle a data set consisting of the growth curves of a sample of 39 boys and 54 girls, the aim being to classify them, by sex, using just this information. Heights were measured in meters at 31 times in the period from one to eighteen years. The data were taken from the file growth.zip, downloaded from ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/Matlab. We represent the data in Figure 4.1.

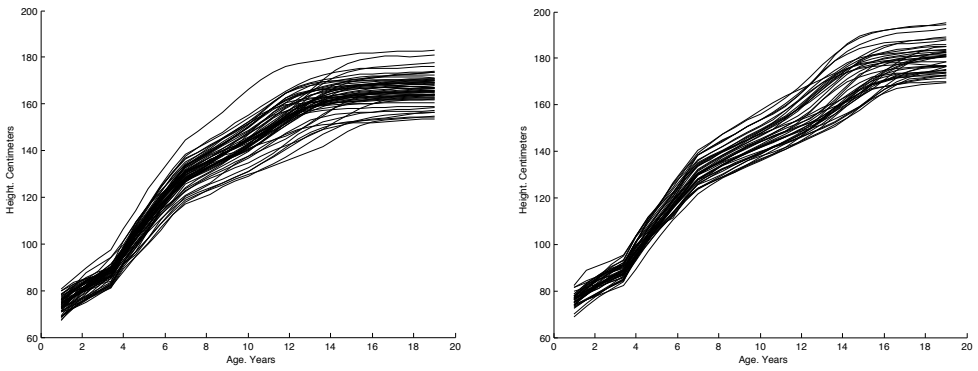

Figure 2: Growth curves of 54 girls (left-hand side) and 39 boys (right-hand side) measured 31 times each between 1 and 18 years of age.

Additionally, we have compared our results with those obtained with some other procedures. We have chosen two functional ones: the k-NN, and a kernel procedure to estimate

the probability that an observation belongs to a group conditional to the height curve. In the last procedure, we have employed two kernels, the indicator of the interval $[0, 1]$ and the quadratic kernel.

Since we only have 31 observations for each individual, we can also consider the data as multidimensional ones and so, it is of some interest to make a comparison with a multidimensional classification procedure. To this end, we have used the Random Forest procedure.

It is well known that when handling this kind of data, it is useful to consider not only the growth curve but also accelerations of height (see, for instance, [16]). However, since we are mainly interested in comparing our results with those obtained in [12], where only growth curves were considered, here we will do the same.

The classification procedure can be extended to an arbitrary number of groups, but, merely to keep the notation as simple as possible, we will assume that we have just two groups. Thus, let us assume that we have two samples $X_1, ..., X_n$ and $Y_1, ..., Y_m$ in $I\!H$ selected from two populations and that we are interested in classifying another curve $Z \in I\!H$ in one of those groups using a depth D to be chosen later. Three classification methods are proposed in [12]:

1.- Distance to the trimmed mean ($\mathbf{M}_{\alpha,\beta}$)

Compute the depths of the points in the sample $X_1, ..., X_n$. Choose $\alpha \in [0, 1)$. The α-trimmed mean of this sample, $\mu_\alpha(X)$, is the mean of the $n \times (1 - \alpha)$ deepest points.

Given $\beta \in [0, 1)$, compute similarly $\mu_\beta(Y)$, the β-trimmed mean of the sample $Y_1, ..., Y_m$.

Now, we classify Z in the first group if

$$\|Z - \mu_\alpha(X)\| < \|Z - \mu_\beta(Y)\|.$$

Otherwise, we classify Z in the second group.

2.- Weighted average distance (AM)

In some sense, in method M, each group is represented by its trimmed mean. Here, we compute the distance between Z and the group as a weighted mean of the distances between Z and the members of the group where the weights are the depths of the points.

Thus, we would classify the function Z in the first group only if

$$\frac{\sum_{i=1}^n \|Z - X_i\| D_X(X_i)}{\sum_{i=1}^n D_X(X_i)} < \frac{\sum_{j=1}^m \|Z - Y_j\| D_Y(Y_j)}{\sum_{j=1}^n D_Y(Y_j)}, \tag{7}$$

where the subscripts in D_X and D_Y mean that the depths are computed with respect to the empirical distribution associated to the corresponding sample.

3.- Trimmed weighted average distance (TAM)

In the AM method, the result of the classification could be affected by the number of elements in each sample if $n \neq m$. The solution for this consists in taking a third value

$l \leq \min(n, m)$ and replacing (7) by

$$\frac{\sum_{i=1}^{l} \|Z - X_{(i)}\| D_X(X_{(i)})}{\sum_{i=1}^{l} D_X(X_{(i)})} < \frac{\sum_{i=1}^{l} \|Z - Y_{(i)}\| D_Y(Y_{(i)})}{\sum_{i=1}^{l} D_Y(Y_{(i)})},$$

where $X_{(1)}$ is the deepest point in the X-sample, $X_{(2)}$ is the second deepest point in the X-sample,... and similarly for the Y-sample. When handling this procedure, we take $l = \min(n, m)$.

We have also included two additional procedures not considered in [12]:

4.- Maximum Depth (MD)

This has been used, for instance, in [5]. It consists, simply, of the following: If we try to classify the observation Z, then add it to the two training samples, compute its depth in each one and classify this point in the sample in which its depth is greater.

5.- Mixture (MiA and MiT)

The idea is to try to compensate the misclassifications of one method with the remaining ones. To do this, we apply to a given observation all previous methods and classify it in the group where it has been classified most often.

There is a technical problem because we have four methods and even cases may occur. Taking into account that the procedures AM and TAM are very close, we have decided to employ only one of them. And, to avoid biases in the selection we have, in fact, divided this method in two: the method MiA (resp. MiT) classifies the new observation in the group which obtain two votes between the procedures $\mathrm{M}_{\alpha,\beta}$, AM (resp. TAM) and MD.

In [12] the authors consider three possibilities to split the sample into training and validation sets. In order to shorten the exposition we only present the results corresponding to the leave-one-out cross-validation setting.

Regarding the selection of the distribution ν used to select the directions to project, we have tried several procedures. The first one is to choose ν as the distribution of the standard Brownian motion. The remaining possibilities are intended to try to take advantage of the differences which appear between the training samples.

To do this, we first compute the functions containing the point-wise medians of the trajectories in both training samples. I.e., for every $t \in [0, 1]$ we compute

$$m_X(t) = \text{median}\{X_1(t), ..., X_n(t)\}, \text{ and } m_Y(t) = \text{median}\{Y_1(t), ..., Y_m(t)\}.$$

Now, we take for ν the distribution of the solution of the following stochastic differential equation

$$S_{a,c}(0) = c \text{ and } dS_{a,c}(t) = |m_X(t) - m_Y(t)|^a dB(t),$$

where B is a standard Brownian motion. Here we choose $a \in \{0, 1\}$. In the first case, the difference between the functions m_X and m_Y has no influence on ν. The constant c specifies the initial value for the solution. We have tried the values $c = 0, 1, 5$. The reason

for introducing c is that the Brownian motion always starts at 0 and is continuous, thus erasing the differences in the early states of the process. In particular, the distribution of $S_{0,0}$ is the standard Brownian motion.

Given a, c, to simulate the random trajectories, first we have taken the values $t_i = i/31, i = 1, ..., 31$ in $[0, 1]$ because we have only 31 measurements for each growth curve. Then we have defined

$$\begin{aligned} S_{a,c}(t_1) &= c \\ S_{a,c}(t_i) &= S_{a,c}(t_{i-1}) + |m_X(t_i) - m_Y(t_i)|^a Z_i,\ i = 2, ..., 31, \end{aligned}$$

where Z_i are independent, identically distributed $N(0, 1/31)$ random variables.

Concerning k, the simulations in Section 3 suggest that not too high values for k are required. In particular, the results which follow have been obtained by selecting k between 1 and 100. In spite of this value perhaps looking too low, we have repeated the process replacing 100 by 1000 and the results have been similar. The right value of k has been obtained by leave-one-out cross validation.

Let us explain, briefly, how the whole process continues. Note first that we have a sample with size 93. In order to check our procedure, we have repeated 1,000 times the following: for each observation in the sample, we consider the training sample composed by the remaining 92 observations. Then, we have generated at random 100 vectors with each of the distributions of the random variables $S_{a,c}$ for $a = 0, 1$ and $c = 0, 1, 5$, which gives 6 different samples of random directions with size 100 each.

First, we have focused our attention on the $S_{0,0}$ distribution. Here we only have to select the value of k. This value is chosen by leave-one-out cross-validation. This procedure is called $S_{0,0}$ in what follows.

Then, we have applied the procedure, allowing variations in a and c. Here we have chosen, also using leave-one-out cross-validation, the best combination of k, a and c. This is the procedure denoted as $S_{a,c}$ in what follows.

4.1.1 Comparison with the results in López-Pintado, S. and Romo, J., 2006.

In this section we compare our depth with those proposed in [12]. To do this, we have repeated the study [12] with only three differences:

1. Most importantly, we have replaced the functional depths handled there by the random Tukey depth.

2. In [12] the authors consider the curves as elements in $L^1[0, 1]$, which is not possible here, because we need a separable Hilbert space. As stated, we take $I\!H = L^2[0, 1]$.

3. In [12], the authors smoothed the original data using a spline basis. We have skipped this step because it did not seem necessary to us.

The results of the comparison appear in Table 4.4, which includes the obtained failure rates using the methods proposed in [12] when applied to the random Tukey depth and to

the depths proposed in [12]. Thus, only the classification methods $M_{\alpha,\beta}$, AM and TAM are employed here. Moreover, we have chosen $\alpha = \beta = 0.2$ as done in [12] .

The last three columns contain the error rates obtained with the depths handled in [12]. They are the band depth determined by three different curves (DS3), by four different curves (DS4) and the generalized band depth (DGS). Their values have been taken from Tables 1-3 in [12]. The first column includes the failure rates when using the distribution of $S_{0,0}$ and the second one when using the distribution of $S_{a,c}, a = 0, 1,\ c = 0, 1, 5$ where a and c are chosen with cross-validation.

Table 4.4 *Rate of mistakes when classifying the growth curves by sex for the shown methods and depths.*

Classification	Random Tukey		Depths proposed in [12]		
method	$S_{0,0}$	$S_{a,c}$	DS3	DS4	DGS
$M_{\alpha,\beta}$	.1368	.1339	.1828	.1828	.1613
AM	.1398	.1248	.2473	.2473	.1935
TAM	.1538	.1359	.2436	.2436	.1690

According to Table 4.4, the random Tukey depth provides better results than those depths proposed in [12] and even better when parameters a, c in $S_{a,c}$ are chosen with cross-validation. The medians of the number of random vectors used have been 1 for each of the three methods with the standard Brownian motion and 2 for each of the methods with $S_{a,c}$.

4.1.2 Comparison with other classification procedures

Here we compare our results with other classification methods. Concerning the random Tukey depth, we use all methods: $M_{\alpha,\beta}$, AM, TAM, MD, MiA and MiT. We combine them with the two methods to select the random directions which we have presented: methods $S_{0,0}$ and $S_{a,c}$, where the values of a and c are chosen by cross-validation. The results for methods $M_{\alpha,\beta}$, AM and TAM appear in Table 4.4 and in Table 4.5 for MD, MiA and MiT. The median of the number of employed random directions is 20 for the MD method when combined with $S_{0,0}$ and 10 when combined with $S_{a,c}$.

The conclusion is that the MD method is slightly better than $M_{\alpha,\beta}$, AM and TAM and that the combined methods provide better results than each method alone.

Table 4.5 *Rate of mistakes when classifying the growth curves by sex for the shown methods and the random Tukey depth.*

Classification method	$S_{0,0}$	$S_{a,c}$
MD	.1255	.1131
MiA	.1024	.0976
MiT	.1010	.0972

One possibility which we have not pursued is to modify α and β in $M_{\alpha,\beta}$. The reason for this is that, in this section, we are mostly interested in comparing our depth with those proposed in [12]. However, we have tried the values $\alpha = 1/38$ and $\beta = 1/53$ (I.e., we look for the deepest point in each training sample and classify the point under consideration according to its closeness to them) lowering the rate of mistakes to around .105 for the $M_{\alpha,\beta}$ method, which has given rised to rates even slightly below .09 when joined in the MiA and MiT methods.

The next step is to classify the same data using a multidimensional procedure (the random forest) and three functional ones (the k-NN and two kernel methods).

As for the random forest, we only wish to state here that this is a procedure which is a combination of tree predictors in which each tree is determined by the values of a random vector which has the same distribution for all the trees in the forest. See [3] for more details. We have employed a random forest with 100 trees.

Concerning the k-NN method (see, for instance, [2]), we have used three possibilities for selecting the number of nearest neighbors. The first two have consisted in fixing $k = 1$ and 3 respectively. In the last one we have chosen k between $1, 3, ..., 91$ with leave-one-out cross-validation. It so happened that in the last possibility, the chosen value was 3 every time. This explains that the rates of failures were the same in both cases (see Table 4.6).

Finally, we have applied two kernel methods (see [1] and [7]). The idea is that, if we consider a new random variable $Z \in \{0, 1\}$ which contains the group to which the observation belongs, and x_0 is the observed curve, it is possible to apply a kernel method to estimate the conditional probability $\mathbb{P}[Z = i/X = x_0]$ for $i = 0, 1$ and, then, to classify the observation in the group in which this probability is highest.

The first kernel tried is $K(u) = 1_{[0,1]}(u)$ and the second one is the quadratic one, $K(u) = (1 - u^2)1_{[0,1]}(u)$. The selection of the window was accomplished as follows: given the training sample $X_1, ..., X_{92}$, we have considered the values

$$\begin{aligned} h_m &= \min\{\|X_i - X_j\|,\ i \neq j, i, j = 1, ..., 92\}, \\ h_M &= \max\{\|X_i - X_j\|,\ i, j = 1, ..., 92\}. \end{aligned}$$

We have chosen the window applying leave-one-out cross-validation to the grid of values $h_m + i(h_M - h_m)/50, i = 0, 1, ..., 50$. The rates of mistakes appear in Table 4.6.

Table 4.6 *Rate of mistakes when classifying the growth curves by sex using cross validation for the shown methods.*

Random Forest	k-NN			Kernel	
	1-NN	3-NN	cross-val.	Indicator	Quadratic
.0968	.0753	.0323	.0323	.0645	.0430

Therefore, the rate of mistakes of both random Tukey procedures is well above those obtained with the k-NN and the kernel method. However, it is similar to the one obtained with Random Forests.

5 Discussion

We introduce the random Tukey depth, which can be considered as a random approximation of the Tukey depth. The new depth is interesting because of the little effort required in its computation and because it can be extended to cover Hilbert valued data.

This depth, in the finite dimensional case is, in fact, a depth according to the definition in [18]. In the infinite dimensional case, only the first three properties of a depth are fulfilled. Moreover, both in the finite as well as in the infinite dimensional settings this depth can be consistently estimated from a random sample.

The interest of the random Tukey depth lies in the fact that by taking only a few one-dimensional projections, it is possible to obtain similar, or even better, results than those obtained with more involved depths, even in the infinite dimensional setting. The number of required projections is surprisingly low, indeed.

This is shown in the comparisons with some other depths that we have carried out. Those studies do not show relevant differences between the results obtained with the considered depths and with the random Tukey depth. Thus, we conclude that, at least under the considered conditions, the random Tukey depth is an alternative which is worth considering because of the little time required to compute it.

However, when this depth is applied to classification problems, the results are similar to those obtained with random forests but worse than those obtained with k-NN or kernel methods. But, the improvement which appears between the first and the second row in Table 4.5, added to the one obtained when changing α and β in the $\mathrm{M}_{\alpha,\beta}$ procedure, makes us relatively optimistic about the results which could be obtained if an optimal procedure to select the distribution ν were applied.

There exists the possibility to extend the results of this paper to more general spaces due to the generalization of the main results in [4], which appears in [6].

Appendix. The proofs

The proofs are identical for finite or infinite dimensional spaces (excepting, of course, the proof of item 4 in Theorem 2.3, which only works for the finite dimensional case). Then, in this appendix the symbol $\mathcal{X}$ will refer here, indistinctly, to $\mathbb{R}^p$ or $I\!H$.

Given a set $B \subset \mathcal{X}$, B^c and ∂B will respectively be their topological complement and boundary. If $x, v \in \mathcal{X}$ and $P \in \mathcal{P}$, we will denote

$$S_{x,v}^P := \begin{cases} \{y : \langle y - x, v\rangle \leq 0\} & \text{if } P\{y : \langle y - x, v\rangle \leq 0\}] \leq P[y : \langle y - x, v\rangle \geq 0] \\ \{y \, \langle y - x, v\rangle \geq 0\} & \text{otherwise} \end{cases}.$$

To simplify, if there is no risk of confusion, the super-index P will be omitted. With this notation, we have that $D_1(\Pi_v(x), P \circ \Pi_v^{-1}) = P(S_{x,v})$.

Proof of Theorems 2.3 and 4.2. Clearly, the random Tukey depth is bounded and nonnegative because it is a minimum of probabilities. To check the remaining properties, let ν be an absolutely continuous distribution on $\mathcal{X}$, $k > 0$ and $P \in \mathcal{P}$.

1. Affine invariance. It is straightforward due to the linearity of the projections.

2. Maximality at center. First remember that a distribution P is halfspace symmetric about θ if $P[H] \geq 1/2$ for every closed halfspace H containing θ. Assume that $\theta \in \mathcal{X}$ is the center of P, and that there exists $x \in \mathcal{X}$ satisfying that

$$D_{T,k}(x, P) > D_{T,k}(\theta, P). \tag{8}$$

By definition, there exists $v \in \{v_1, \ldots, v_k\}$ such that $D_{T,k}(\theta, P) = D_1(\Pi_v(\theta), P \circ \Pi_v^{-1}) = P(S_{\theta,v}) \geq 1/2$, due to the halfspace symmetry. Thus, from (8) we get

$$P(S_{x,v}) > 1/2. \tag{9}$$

We have three possibilities for the sets $S_{x,v}$ and $S_{\theta,v}$. The first one is that $S_{x,v} \subseteq S_{\theta,v}$. But, then $D_{T,k}(x, P) \leq P(S_{x,v}) \leq P(S_{\theta,v}) = D_{T,k}(\theta, P)$ which contradicts (8).

The second one is $S_{\theta,v} \subset S_{x,v}$. From here and $P(S_{\theta,v}) \geq 1/2$, we obtain that $P(S_{x,v}^c \cup \partial S_{x,v}) \leq 1/2$. By (9), $\min(P(S_{x,v}^c \cup \partial S_{x,v}), P(S_{x,v})) = P(S_{x,v}^c \cup \partial S_{x,v})$, which contradicts the definition of $S_{x,v}$.

The final possibility is that $S_{\theta,v} \subset S_{x,v}^c$. Thus, $P(S_{x,v}^c) \geq 1/2$. Therefore, by (9) $1 = P(S_{x,v}) + P(S_{x,v}^c) > 1$.

3. Monotonicity relative to deepest point. Let assume that P has a deepest point θ and there exist $x \in \mathcal{X}$ and $\alpha \in [0, 1]$ with

$$D_{T,k}(x, P) > D_{T,k}(\theta + \alpha(x - \theta), P). \tag{10}$$

Obviously, cases $\alpha = 0$ and $\alpha = 1$ are not possible. Then, $\alpha \in (0, 1)$.

Since θ is the deepest point, we have that

$$D_{T,k}(\theta, P) \geq D_{T,k}(y, P), \text{ for all } y \in \mathcal{X}. \tag{11}$$

Let $v \in \{v_1, ..., v_k\}$ such that $D_{T,k}(\theta + \alpha(x - \theta), P) = P(S_{\theta+\alpha(x-\theta),v})$. From (10) and (11) it is inferred that

$$P(S_{\theta,v}) > P(S_{\theta+\alpha(x-\theta),v}). \tag{12}$$

Since $\alpha \in (0, 1)$, we have that the point $\theta + \alpha(x - \theta)$ lies in the open segment joining the points x and θ. Then, from (11) and (12), following a reasoning similar to the final part of the proof of Statement 2, we have that $S_{x,v} \subset S_{\theta+\alpha(x-\theta),v}$. Thus,

$$D_{T,k}(\theta + \alpha(x - \theta), P) = P(S_{\theta+\alpha(x-\theta),v}) \geq P(S_x, v) \geq D_{T,k}(x, P),$$

which contradicts (10).

4. Vanishing at infinity (only for Theorem 2.3). We will show this property for the case in which ν is not a probability, but the Lebesgue measure. The proof for probabilities follows the same steps till statement (15) below. From this point on, only some additional technicalities are required.

Let $\epsilon > 0$. Since $\lim_{H\to\infty} P\{y : \|y\| \leq H\} = 1$, we have that there exists $H_\epsilon > 0$ such that $P\{y : \|y\| \leq H_\epsilon\} > 1 - \epsilon$.

Furthermore, if $v \in \mathbb{R}^p$ then $\Pi_v^{-1}[-H_\epsilon\|v\|, H_\epsilon\|v\|] \supset \{y : \|y\| \leq H_\epsilon\}$. Thus, $P \circ \Pi_v^{-1}[-H_\epsilon\|v\|, H_\epsilon\|v\|] > 1 - \epsilon$, for all $v \in \mathbb{R}^p$. In consequence, if we have

$$\sup\left(D_1(-H_\epsilon\|v\|, P \circ \Pi_v^{-1}), D_1(H_\epsilon\|v\|, P \circ \Pi_v^{-1})\right) < \epsilon, \text{ for all } v \in \mathbb{R}^p. \tag{13}$$

Let $M > 0$ and let $x \in \mathbb{R}^p$ with $\|x\| \geq M$. We have that

$$\begin{aligned} \nu^k\left\{(v_1, ..., v_k) \in (\mathbb{R}^p)^k : D_{T,k}(x, P) < \epsilon\right\} &\geq \nu^k\left\{(v_1, ..., v_k) \in (\mathbb{R}^p)^k : D_{T,1}(x, P) < \epsilon\right\} \\ &= \nu\left\{v \in \mathbb{R}^p : D_{T,1}(x, P) < \epsilon\right\}, \end{aligned} \tag{14}$$

where we assume that $D_{T,1}$ is computed using v_1.

If $v \in \mathbb{R}^p$ satisfies that $|\langle x, v\rangle| \geq H_\epsilon\|v\|$, then by (13), it happens that $D_1(\Pi_v(x), P \circ \Pi_v^{-1})) < \epsilon$. Therefore, from (14),

$$\begin{aligned} \nu^k\left\{(v_1, ..., v_k) \in (\mathbb{R}^p)^k : D_{T,k}(x, P) < \epsilon\right\} &\geq \nu\left\{v \in \mathbb{R}^p : |\langle x, v\rangle| \geq H_\epsilon\|v\|\right\} \\ &\geq \nu\left\{v \in \mathbb{R}^p : \frac{|< x, v >|}{\|x\|\|v\|} \geq \frac{H_\epsilon}{M}\right\} \quad (15) \\ &= \nu\left\{v \in \mathbb{R}^p : \frac{|< e_1, v >|}{\|v\|} \geq \frac{H_\epsilon}{M}\right\}, \end{aligned}$$

where (15) comes from $\|x\| > M$. In the last equality, e_1 denotes the first element in a fixed orthonormal base of $\mathbb{R}^d$ and the equality holds because ν is rotationally invariant. Therefore, from this chain we have that

$$\inf_{x:\|x\|\geq M} \nu^k\left\{(v_1, ..., v_k) \in (\mathbb{R}^p)^k : D_{T,k}(x, P) < \epsilon\right\} \geq \nu\left\{v \in \mathbb{R}^p : \frac{|< e_1, v >|}{\|v\|} \geq \frac{H_\epsilon}{M}\right\},$$

and the proof ends because, trivially,

$$\lim_{M\to\infty} \nu\left\{v \in \mathbb{R}^p : \frac{|< e_1, v >|}{\|v\|} \geq \frac{H_\epsilon}{M}\right\} = 1.$$

•

Proof of Theorems 2.4 and 4.3. Let $k > 0$ and $v_1, ..., v_k \in \mathcal{X}$, which will remain fixed during the proof. Let P be a probability distribution on $\mathcal{X}$, let $x_1, ..., x_n \in \mathcal{X}$ be a random sample taken from P and let P_n be the associated empirical distribution.

Let $v \in \{v_1, ..., v_k\}$. It is obvious that $\Pi_v(x_1), ..., \Pi_v(x_n)$ is a random sample taken from the distribution $P \circ \Pi_v^{-1}$ and that the empirical distribution associated to those projections coincide with $P_n \circ \Pi_v^{-1}$. Moreover, $P \circ \Pi_v^{-1}$ is a distribution on the real line and, then, the Glivenko-Cantelli theorem gives that

$$\sup_{y\in\mathbb{R}} \sup\left(\left|P_n \circ \Pi_v^{-1}(-\infty, y] - P \circ \Pi_v^{-1}(-\infty, y]\right|, \left|P_n \circ \Pi_v^{-1}[y, \infty) - P \circ \Pi_v^{-1}[y, \infty)\right|\right) \to 0, \text{a.s.}$$

From here, we obtain that

$$\sup_{y\in\mathbb{R}}\left(|D_1(y,P_n\circ\Pi_v^{-1})-D_1(y,P\circ\Pi_v^{-1})|\right)\to 0, \text{a.s.} \tag{16}$$

Therefore,

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}|D_{T,k}(x,P_n)-D_{T,k}(x,P)| \\
&\qquad = \sup_{x\in\mathcal{X}}|\min_{i=1,\ldots,k} D_1(\Pi_{v_i}(x),P_n\circ\Pi_{v_i}^{-1})-\min_{i=1,\ldots,k} D_{T,k}(\Pi_{v_i}(x),P\circ\Pi_{v_i}^{-1})| \\
&\qquad \leq \sup_{x\in\mathcal{X},i=1,\ldots k}|D_1(\Pi_{v_i}(x),P_n\circ\Pi_{v_i}^{-1})-D_1(\Pi_{v_i}(x),P\circ\Pi_{v_i}^{-1})| \\
&\qquad = \sup_{y\in\mathbb{R},i=1,\ldots k}|D_1(y,P_n\circ\Pi_{v_i}^{-1})-D_1(y,P\circ\Pi_{v_i}^{-1})|,
\end{aligned}$$

which converges a.s. to zero because of (16). •

References

[1] Abraham, C., Biau, G. and Cadre, B., 2006. On the kernel rule for function classification. Ann. Inst. Statist. Math., 58, 619-633.

[2] Biau, G., Bunea, F. and Wegcamp, M.H., 2005. Functional Classification in Hilbert Spaces. IEEE Transact. Informat. Theo. 51, 2163-2172.

[3] Breiman, L., 2001. Random Forests. Machine Learning, 45, 5-32.

[4] Cuesta-Albertos, J., Fraiman, R. and Ransford, T., 2007. A sharp form of the Cramér-Wold theorem. J. Theoret. Probab., 20 201-209.

[5] Cuevas, A., Febrero, M. and Fraiman, R., 2007. Robust estimation and classification for functional data via projection-based depth notions. To appear in Computation. Statist.

[6] Cuevas, A. and Fraiman, R., 2007 On depth measures and dual statistics: A methodology for dealing with general data. Preprint.

[7] Ferraty, F. and Vieu, P., 2003. Curves discrimination: a nonparametric functional approach. Computat. Statist. Data Anal. 44, 161-173.

[8] Hand, D.J., 2006. Classifier technology and the illusion of progress. Statist. Sci., 21(1) 1-14.

[9] Hettmansperger, T.P., 1984. Statistical inference based on ranks. John Wiley & Sons, New York.

[10] Liu, R., Serfling, R. and Souvaine, D.L., editors, 2006. Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. American Mathematical Society, DIMACS Series, Vol. 72.

[11] Liu, R.Y. and Singh, K., 2006. Rank tests for nonparametric description of dispersion. In: R. Liu, R. Serfling and D.L. Souvaine (Ed.), Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, American Mathematical Society, DIMACS Series, Vol. 72, 17-35.

[12] López-Pintado, S. and Romo, J., 2006. Depth-based classification for functional data. In: R. Liu, R. Serfling and D.L. Souvaine (Ed.), Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, American Mathematical Society, DIMACS Series, Vol. 72, 17-35.

[13] Mahalanobis, P. C., 1936. On the Generalized Distance in Statistics. Proceed. Nat. Academy of India, 12 49-55.

[14] Maronna, R.A., Martin, R. D. and Yohai, V.J., 2006. Robust Statistics. Theory and Methods. John Wiley & Sons, Chichester.

[15] Mosler, K. and Hoberg, R., 2006. Data analysis and classification with the zonoid depth. In: R. Liu, R. Serfling and D.L. Souvaine (Ed.), Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, American Mathematical Society, DIMACS Series, Vol. 72, 17-35.

[16] Ramsay, J.O. and Silverman, B.W., 1997. Functional Data Analysis. Springer Verlag, New York.

[17] Tukey, J.W., 1975. Mathematics and picturing of data. Proceedings of ICM, Vancouver, 2 523-531.

[18] Zuo, Y. and Serfling, R., 2000. General notions of statistical depth function. Ann. Statist., 28(2) 461-482.

[19] Zuo, Y., 2003. Projection-based depth functions and associated medians. Ann. Statist., 31(5) 1460-1490.

[20] Zuo, Y., 2006. Multidimensional trimming based on projection depth. Ann. Statist., 34(5) 2211-2251.