Functional classification and the random Tukey depth. Practical issues

Juan A. Cuesta-Albertos and Alicia Nieto-Reyes

Abstract Depths are used to attempt to order the points of a multidimensional or infinite dimensional set from the "center of the set" to the "outer of it". There are few definitions of depth which are valid in the functional case. One of them is the so-called random Tukey depth, which is based on some randomly chosen one-dimensional projections and thus varies (randomly) from computation to computation. Some theoretical properties of this depth are well-known, but it has not yet been studied from a practical point of view. The aim of this paper is to analyze its behavior in classification problems, the interest of this study being increased by the random character of the depth. To do this, we compare the performance of the random Tukey depth in a real data set with the results obtained with the López-Pintado and Romo depths.

1 Introduction

Given a probability P defined in a multidimensional or infinite-dimensional space $\mathcal{X}$, a depth attempts 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.

In the multidimensional setting, the first definition of depth was established by Mahalanobis (in [12]). This definition is based on the well known Mahalanobis distance. If μ and Σ are, respectively, the mean and covariance matrix of P, then, the Mahalanobis depth of x with respect to P is

$$D_H(x,P) := \frac{1}{1+(x-\mu)^t \Sigma^{-1}(x-\mu)},\ x \in \mathbb{R}^p.$$

Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Spain juan.cuesta@unican.es · alicia.nieto@unican.es

From this starting point, subsequent definitions of depth (see [9]) clarified that depths as well as having some robustness properties, are a highly flexible tool for handling nonparametrically statistical problems involving testing, classification, descriptive statistics,... This, in turn, has led to the study of the possibility of introducing depths in the functional setting. However, most of the known multidimensional depths cannot be generalized to the functional case because the dimension of the space under consideration plays a key role in them, or alternatively, because of the associated computational difficulties. For instance, the computation of the Tukey depth (a precise definition appears in (1)) is unfeasible for dimensions as low as eight if the sample size is only 100.

As far as we know, some definitions of depth valid for functional spaces have been proposed in [5], [6], [7] and [11]. In this paper, we are particularly interested in the so-called random Tukey depth which was studied in [3, 4] because these papers leave some practical issues open. Our goal here is to make a first attempt to show how these gaps can be filled when handling classification problems.

Let us begin with some definitions. Apart from its lack of robustness, the Mahalanobis depth has some flaws: it is not defined if the mean or the covariance matrix does not exist and it treats P as symmetric (because points at the same Mahalanobis distance from the mean have the same depth). A reasonable way to overcome these problems in the one dimensional case could be to define the depth of the point x with respect to P by

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

which is a monotone transformation of the Mahalanobis depth if μ and Σ exist and P is symmetric, thus providing the same order of the points.

The *Tukey depth* was introduced in [16] and can be defined as follows. Let P be a probability on $\mathbb{R}^p$ and $v \in \mathbb{R}^p$. If Π_v denotes the projection on the one-dimensional subspace generated by v and P_v the one-dimensional marginal of P on the same subspace, then, the Tukey depth of x with respect to P is

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

The computational problems we mentioned above, led the authors of [3] to introduce the random Tukey depth, which is a random approximation of the Tukey depth. In [3], the following generalization to Hilbert spaces was proposed:

Definition 1. Let $\mathcal{X}$ be a separable Hilbert space, P be a probability distribution on $\mathcal{X}$, ν be a Gaussian distribution with non-degenerated marginals on $\mathcal{X}$ and $v_1, ..., v_k$ be i.i.d. random vectors with distribution ν. The random Tukey depth of $x \in \mathcal{X}$ with respect to P based on k random vectors chosen with ν is

$$D_{T,k,\nu}(x,P) := \min\{D_1(\mathit{\Pi}_{v_i}(x), P_{v_i}) : i = 1, ..., k\}.$$

The random Tukey depth was used in [3], in the finite dimensional case, to handle several classification and testing problems and, in addition, it was shown there that this depth has some useful properties in the infinite and finite cases. In particular, it was shown that in the infinite dimensional case, it satisfies most of the requirements of the definition stated in [8] and formalized in [17] for a statistical depth.

However, in [3] nothing is said about the influence that the selection of ν and k might have in practice. The aim of this paper is to make a preliminary analysis of these issues from the point of view of a classification problem, and, at the same time, to compare the results obtained with the random Tukey depth with those provided in [10] with the depths proposed in [11].

The situation we have chosen to carry out this comparison is the supervised classification problem which was carried out in [10]. In this paper, the authors analyze 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. We represent the data in Figure 1.

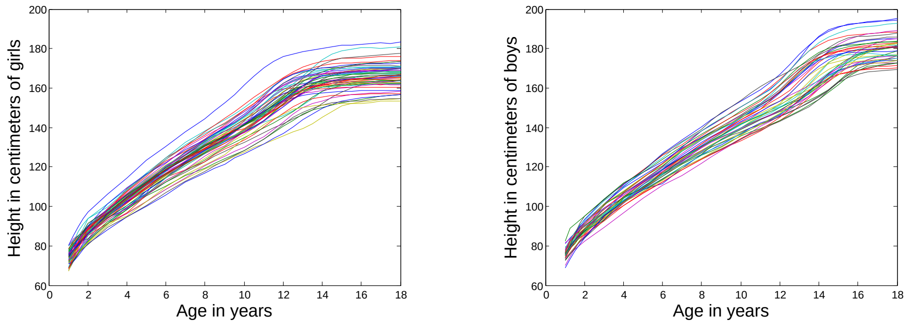

Fig. 1 Growth curves of 54 girls (left-hand side) and 39 boys (right-hand side) measured 31 times each between 1 and 18 years of age.

Heights were measured in centimeters 31 times in the period from one to eighteen years. In the period from one to two years, the measures were taken every three months, in the period from three to seven years one time a year and, finally, in the period from eight to eighteen years two times a year. The data are in the file growth.zip, downloaded from ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/Matlab. On this web-page, some notes that make use of the data can also be found. These notes were designed to accompany the books [13, 14]. In addition, these data are used in the recent book [15].

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, [13]). However, we only consider here the growth curves, as did [10], because our interest lies in comparing our results with those obtained by them.

It should be noted that the distribution ν which appears in Definition 1 does not need to be Gaussian. In fact, as shown in [1], any dissipative distribution works here. Thus, in the finite dimensional case, the uniform distribution on the unit sphere may be enough. Regrettably, in the functional setting, there is no distribution like this which can be taken as a reference. Although some papers have already appeared using random projections (in the finite and in the infinite dimensional cases), as far as we know, except for a small comment in [2] in the finite dimensional case, none of them has paid attention to the problem of the precise selection of ν.

A preliminary step in addressing this question is given in Section 2, where we also comment on the selection of the number of vectors used in the definition of the random Tukey depth. Then, in Section 3 we compare the results obtained with the random Tukey depth with those obtained in [10].

2 Distribution and number of vectors for the random Tukey depth in practice

In order to analyze the effect of the selection of ν in the random Tukey depth in classification problems, the idea is to analyze the same data using two strategies: firstly, one that does not admit variations in ν, i.e. ν is a fixed distribution. Secondly, one that selects ν from a parametric family of distributions, thus making it possible to chose the parameters which determine ν in a data-dependent way.

The parametric family we handle has two real parameters $a \geq 0$ and $c \geq 0$, and is defined forthwith. Let us assume that we are in a two-class classification problem and that we have two training samples $\mathbb{X} = \{X_1(t), ..., X_n(t)\}$ and $\mathbb{Y} = \{Y_1(t), ..., Y_m(t)\}$, where, $t \in [0, T]$. First, compute the point-wise median in both samples: $m_{\mathbb{X}}(t)$, and $m_{\mathbb{Y}}(t),\ t \in [0, T]$. Then, given a, c let $\nu = S_{a,c}$ be the solution of the 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.

The fixed distribution that we compare with is the standard Brownian motion, which is the member of the family corresponding to the case $a = c = 0$.

In the following section, we choose $a \in \{0, 1\}$. Note that when $a = 0$ 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 processes.

In practice, we will assume that the trajectories have been measured in the same finite set of values $t_1 < \ldots < t_h$. Then, given a and c, to simulate the random trajectories we have taken

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

where Z_i, $i = 2, ..., h$, are independent random variables with distribution $N(0, t_i - t_{i-1})$.

Concerning k, in [3] the authors carry out some simulations to select k in the finite dimensional case for dimensions ranging from $p = 2, 8, 50$ and several sample sizes. Those results suggest that high values for k are not required. The results that follow in Section 3 have been obtained by selecting $k \in \{1, ..., 100\}$. Although the upper bound for k might be considered too low, we have repeated the process replacing 100 by 1,000 and the results obtained have been similar.

We propose the use of leave-one-out cross validation to choose the right value of k, as well as those of a and c when required.

3 The procedure in practice

As stated, in this section, we compare the results of classifying the heights data set when employing the random Tukey depth with those obtained with the depths proposed in [11]. To do this, we have repeated the study made in [10] with three differences:

1. Most importantly, we have replaced the functional depths handled there with the random Tukey depth.
2. In [10], the authors consider the curves as elements in $L^1[0, 1]$, which is not possible here, because we need a separable Hilbert space. Thus, we have taken $\mathbb{H} = L^2[0, 1]$.
3. In [10], the authors smoothed the original data using a spline basis. We have omitted this step because it is not necessary for our method.

Regarding item 2., remember that the heights were measured 31 times on times $t_i \in [1, 18]$, $i = 1, ..., 31$ where

$$\begin{aligned} t_i &= 3/4 + i/4 \text{ for } i = 1, ..., 5, \\ t_i &= i - 3 \text{ for } i = 6, ..., 10 \\ t_i &= 2.5 + i/2 \text{ for } i = 11, ..., 31. \end{aligned}$$

If $i = 2, \ldots, 30$, then the observation $X(t_i)$ represents the height of the individual in the interval $((t_i+t_{i-1})/2, (t_{i+1}+t_i)/2)$. Taking into account that, in the last part of the study, the measurements were taken every half a year, we can assume that $X(t_{31})$ is valid for the period $(17.5, 18.5)$. Finally, it seems safe to assume that the $X(t_1)$ is not valid for representing previous heights. Therefore, we can assume that the interval in which the measurements have been taken is $[1, 18.5]$. In consequence, first, we need to modify the time in order to transform the interval $[1, 18.5]$ into $[0, 1]$ and, then, we can employ properties of the Rieman integral to make the approximation

$$< X, s_{a,c} > = \int_0^1 X\left(17.5u + 1\right) s_{a,c}(17.5u + 1) du \approx \sum_{i=1}^{31} X(t_i) s_{a,c}(t_i) \mathit{\Delta}_i,$$

where $s_{a,c}$ is drawn with distribution $S_{a,c}$ and $\mathit{\Delta}_i$ denotes the length of the interval associated to the point t_i. Then, if we define $t_0 = 1$ and $t_{32} = 18.5$, we have

$$\mathit{\Delta}_i = (t_{i+1} - t_{i-1})/35, \ i = 1, ..., 31.$$

In [10], the authors consider three possibilities for splitting the sample into training and validation sets. For the sake of brevity, we split the sample using only leave-one-out cross-validation.

Let us briefly explain how the whole process works. Note that we have a sample of size 93. Therefore, we have repeated 100 times the following: for each observation in the sample, we consider the training sample composed of 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.

Firstly, we have focused our attention on the $S_{0,0}$ distribution. Here we only have to select the value of k. As stated previously, this value is chosen by leave-one-out cross-validation applied to the remaining sample with 92 observations. Henceforth, this procedure is called $S_{0,0}$.

Moreover, we have applied the procedure allowing variations in a and c. Here, also using leave-one-out cross-validation, we have chosen the best combination of k, a and c. Henceforth, this procedure is denoted by $S_{a,c}$. Note that in this case, it may occur that the chosen a and c satisfy $a = c = 0$; thus, the $S_{a,c}$ procedure should give better results than $S_{0,0}$.

The results of the comparison appear in Table 1, which includes the obtained failure rates using the three methods proposed in [10] when applied to the random Tukey depth and to the depths proposed in [11]. These methods are: distance to the trimmed mean ($\mathrm{M}_{\alpha,\beta}$), weighted average distance (AM) and trimmed weighted average distance (TAM). We have chosen $\alpha = \beta = 0.2$ as in [10]. The depths handled in [10] are the band depth determined by three different curves (DS3), by four different curves (DS4) and the generalized band depth (DGS). Their error rates are contained in the last three columns of Table 1 and have been taken from Tables 1-3 in [10]. The previous two columns of Table 1 concern the random Tukey depth. The first includes the failure rates when using the procedure $S_{0,0}$ and the second when using $S_{a,c}$.

Classification	Random Tukey		Depths proposed in [10]		
method	$S_{0,0}$	$S_{a,c}$	DS3	DS4	DGS
$\mathrm{M}_{\alpha,\beta}$	.1858	.1825	.1828	.1828	.1613
AM	.1403	.1368	.2473	.2473	.1935
TAM	.1542	.1430	.2436	.2436	.1690

Table 1 Rates of mistakes when classifying the growth curves by sex for the shown methods and depths.

According to Table 1, if we employ the $M_{\alpha,\beta}$ method, the random Tukey depth with the procedure $S_{0,0}$ works worse than the other depths and, when coupled with $S_{a,c}$ performs similarly to the $DS3$ and $DS4$ depths but worst than the DGS. However, for the AM and TAM methods, the random Tukey depth provides better results than the depths used in [10] when we take the standard Brownian motion 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 $S_{0,0}$. In the case of $S_{a,c}$, the median of the number of random vectors has been 2 for the $\mathrm{M}_{\alpha,\beta}$ method and 1 for both of the other two methods.

4 Conclusions

The Tukey depth is one of the best-behaved multidimensional depths but it cannot be used in the functional setting. However, the random Tukey depth, which approximates it in multidimensional spaces, does work in functional settings.

The definition of the random Tukey depth involves choosing a distribution. We have seen how in practice the behavior of this functional depth varies depending on the chosen distribution. Specifically, its performance increases when the distribution is data driven.

Furthermore, to compute the random Tukey depth, a finite number of vectors have to be drawn with this chosen distribution. This number is of great importance since the computational time needed to compute the random Tukey depth depends on it. In [3] it was seen that this number is low in multidimensional spaces and, in view of our experience, it seems that it is also surprisingly low when dealing with functional data.

References

1. Cuesta-Albertos, J.A., del Barrio, T., Fraiman, R. and Matrán, C. (2007). The random projection method in goodness of fit for functional data. Computat. Statist. Data Anal. **51**(10), 4814–4831.
2. Cuesta-Albertos, J.A. and Febrero-Bande, M. (2010). Multiway ANOVA for Functional Data. To appear in Test.
3. Cuesta-Albertos J.A., and Nieto-Reyes, A. (2008). The random Tukey depth. Comput. Statist. Data Anal. 52(11):4979–4988
4. Cuesta-Albertos J.A., and Nieto-Reyes, A. (2008). A random functional depth. In: Dabo-Niang S, Ferraty F (eds) Functional and Operational Statistics. Springer, 121–126
5. Cuevas A., Febrero M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. Comput. Statist. 22(3):481–496
6. Cuevas A., and Fraiman, R. (2009). On depth measures and dual statistics. A methodology for dealing with general data. J. Multivariate Anal. 100(4):753–766
7. Fraiman R., and Muniz G. (2001). Trimmed means for functional data. Test. 10(2):419–440
8. Liu, R.Y. (1990). On a notion of data depth based on random simplices. Ann. Statist. 18:405–414
9. Liu, R.Y., Parelius, J.M., and Singh, K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann. Statist. 27(3):783–858
10. López-Pintado, S., and Romo, J. (2006). Depth-based classification for functional data. In: R. Liu, R. Serfling and D.L. Souvaine, Editors, Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, DIMACS Series vol. 72, American Mathematical Society:103–119
11. López-Pintado, S., and Romo, J. (2009). On the concept of depth for functional data. J. Amer. Statist. Assoc. 104(486):718–734
12. Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proc. Natl. Inst. Science 12:49–55
13. Ramsay, J.O. and Silverman, B.W. (1997). Functional Data Analysis. Springer.
14. Ramsay, J.O. and Silverman, B.W. (2002). Applied Functional Data Analysis: Methods and Case Studies. Springer.
15. Ramsay, J.O., Hooker, G. and Graves, S. (2009). Functional Data Analysis with R and MATLAB. Springer.
16. Tukey, J.W. (1975). Mathematics and picturing of data. Proc. ICM Vancouver 2:523–531
17. Zuo, Y., and Serfling, R. (2000). General notions of statistical depth function. Ann. Statist. 28(2):461–482