Computational Statistics (2007) 22:481–496
DOI 10.1007/s00180-007-0053-0

ORIGINAL PAPER

# Robust estimation and classification for functional data via projection-based depth notions

**Antonio Cuevas · Manuel Febrero · Ricardo Fraiman**

Published online: 21 March 2007

**Abstract** Five notions of data depth are considered. They are mostly designed for functional data but they can be also adapted to the standard multivariate case. The performance of these depth notions, when used as auxiliary tools in estimation and classification, is checked through a Monte Carlo study.

**Keywords** Depth measures · Functional data · Projections method · Supervised classification

**Mathematics Subject Classification (2000)** Primary 62G07 · Secondary 62G20

## 1 Introduction

In the statistical methodology the problems with one-dimensional data are exceptional at least in an important aspect: in this simple situation the existence of the standard real-valued order statistics provides a natural way to define a

Research partially supported by Spanish grants MTM2004-00098 (A. Cuevas and R. Fraiman) and MTM2005-00820 (M. Febrero).

A. Cuevas (✉)
Departamento de Matemáticas, Univ. Autónoma de Madrid, Madrid, Spain
e-mail: antonio.cuevas@uam.es

M. Febrero
Departamento de Estatística e Inv. Operativa, Univ. de Santiago de Compostela,
Santiago de Compostela, Spain

R. Fraiman
Departamento de Matemática, Univ. de San Andrés, Buenos Aires, Argentina

data depth notion leading to well-known applications in robust statistics and data analysis.

There is no obvious simple way of defining depth for multivariate data. The papers by Liu et al. (1999) and Zuo and Serfling (2000) provide a broad perspective, in the finite dimensional case, of different notions of data depth and their applications (see below for more details). In spite of the recent progress in this field there is still considerable room for further research. Thus, some proposed notions of depth show a reasonable performance in dimensions 2 or 3 but entail a prohibitive computational burden when used in higher dimensions. However, in some practical applications (for example in those involving image analysis or microarrays data sets in genomic studies), the so-called high-dimensional data arise in a natural way. In such applications the dimensión $d$ can be very large, often larger than the sample size. On the other hand, there is a need of depth notions valid for infinite-dimensional data in connection with the increasing demand of statistical tools for functional data analysis (FDA) where the available data are functions $x = x(t)$ defined on some real interval (say $[0, 1]$) see e.g., Ramsay and Silverman (2005a,b) or Ferraty and Vieu (2006) for general accounts on FDA and Fraiman and Muniz (2001) for a proposal of functional depth (to be considered below in some detail). The high-dimensional situation and the FDA analysis are closely related in practice, as the functional data usually come in a discretized fashion so that a function $X_i$ in the sample is in fact given by $(X_i(t_1), \ldots, X_i(t_d))$. There are, however, at least two distinctive features with respect to the typical high-dimensional problems: first, the dimension $d$ depends on the discretization order which is not given in advance and, in principle, can be arbitrarily increased. Second, the data coming from the discretized functions tend to be highly correlated, which entails serious problems in the accurate estimation of the covariance matrices. Many functional approaches use either a “regularization procedure” (Hastie et al. 1995) (that tends to exclude from consideration the functions “too wiggly”) or a “filtering method”, which leads to replace every function by its coefficients with respect to the basis of a suitable finite-dimensional subspace.

As we have mentioned above, a suitable definition of data depth has an obvious usefulness in order to define robust estimators for a “location parameter” (even in high-dimensional or functional frameworks): for example we can use as an estimator the deepest datum (that is the median). Other robust estimators, as the well-known trimmed means, can also be defined in the obvious way when a definition of depth is available. Likewise, the different concepts of data depth can be used in classification: The idea is to assign a coming datum according to its relative depth in the different training samples. The purpose of this paper is to explore these ideas by checking the usefulness of some notions of data depth as auxiliary tools in robust estimation and supervised classification, with a special (but not exclusive) emphasis in problems involving functional data. It is clear that the different concepts of data depth, especially those applicable to functional data, are the core of our approach. From the classical paper by Tukey (1975), where the notion of “halfspace depth” is introduced, the research on data depth has received a considerable attention. Other well-known approaches to this notion include the “simplicial depth” proposed by Liu (1990) and the “regression depth” (see Rousseeuw and Hubert 1999). A projection-based method is proposed in Zuo (2003); see also Zuo and Serfling (2000). Further depth notions based on projections are considered below. In our case the directions are randomly chosen along the lines proposed in Cuesta-Albertos et al. (2006a).

The paper is organized as follows. In Sect. 2 below we define some notions of data depth. They are suitable either for high-dimensional or functional data, but we are particularly concerned with the latter. We check the performance of these depth notions, when used as auxiliary tools in both supervised classification (Sect. 3) and estimation (Sect. 4).

## 2 Some notions of data depth for high-dimensional or functional data

We will consider here five proposals, namely: the Fraiman and Muniz (2001) method, the $h$-mode depth, and three variants of a procedure (inspired in the work by Cuesta-Albertos et al. 2006a) relying on the use of random projections. All of them are suitable for dealing with functional data but the first three proposals can also be adapted for finite-dimensional data.

To our knowledge, only the first of these procedures has been so far explicitly proposed as a notion of depth for functional data, although the $h$-mode has been also considered, from a somewhat different point of view, in FDA: see e.g., Cuevas et al. (2006) and Ferraty and Vieu (2006).

### The Fraiman and Muniz (FM) method

It is based upon the following notion of (functional) integrated depth: for every $t \in [0, 1]$ let $F_{n,t}$ be the empirical distribution of the sample $x_1(t), \ldots, x_n(t)$ and let $Z_i(t)$ denote the univariate depth of the data $x_i(t)$ in this sample, given by $D_i(t) = 1 - |\frac{1}{2} - F_{n,t}(x_i(t))|$. Then, define for $i = 1, \ldots, n$,

$$I_i = \int_0^1 D_i(t)dt \tag{1}$$

and rank the observations $x_i(t)$ according to the values of $I_i$.

This procedure was first proposed by Fraiman and Muniz (2001) who used it mainly as a estimation tool, in order to define functional versions of the well-known L-estimators. The adaptation of the FM-method to the case of finite dimensional (multivariate) data is straightforward by replacing the integral in (1) by an appropriate finite sum.

### The $h$-mode (hM) method

According to this notion, the population $h$-depth of a datum $z$ is given by the function

$$f_h(z) = E(K_h(\|z - X\|)) \tag{2}$$

where $X$ is the random element describing the population, $\|.\|$ is a suitable norm (for example the $L_2$-norm in the functional case) and $K_h(t)$ is a re-scaled kernel of type $K_h(t) = \frac{1}{h}K(\frac{t}{h})$, $K$ being a given kernel function (we will use the Gaussian kernel $K(t) = (1/\sqrt{2\pi})\exp(-t^2/2)$) and $h$ is a fixed tuning parameter. We define the $h$-mode of $X$ as the deepest value of $z$ obtained by maximizing (in $z$) the function (2).

Given a random sample $X_1, \ldots, X_n$ of $X$, the sample version of the $h$-depth is defined in a natural way by replacing (2) by the empirical version

$$\hat{f}_h(z) = \frac{1}{n}\sum_{i=1}^{n} K_h(\|z - X_i\|) \tag{3}$$

When used with functional data, this notion is clearly reminiscent of the multivariate depth concepts based on likelihood (see e.g., Fraiman and Meloche 1999 and references therein) but the analogy is severely limited by the fact that, unlike the finite-dimensional case, there is no real underlying density to be estimated by (3) making $h$ tend to 0, slowly enough, as $n \to \infty$. This is associated with the lack of a standard translation-invariant reference measure playing the role that Lebesgue measure does in the Euclidean space $\mathbb{R}^d$. However, in our functional setup we can just preserve some partial aspects of the idea, by keeping $h$ constant and considering (2) as an auxiliary function indicating “how surrounded” is the function $z$ in the sample space. In our simulations the $h$-modal depth (3) is replaced by a “normalized” version $(f_h(z) - \min f_h(z))/(\max f_h(z) - \min f_h(z))$.

### The random projection (RP) method

Given a sample $X_1, \ldots, X_n$ let us take a random direction $a$ (independent from the $X_i$) and project the data along this direction. Then, the sample depth of a datum $X_i$ is defined as the univariate depth of the corresponding one-dimensional projection (expressed in terms of order statistics so that the median is the deepest point). When the sample is made of functional data, we will assume throughout that the $X_i$ belong to the Hilbert space $L^2[0,1]$ so that the projection of a datum $X$ is given by the standard inner product $\langle a, X\rangle = \int_0^1 a(t)X(t)dt$. In the finite-dimensional case the projection of $X = (\xi_1, \ldots, \xi_d)$ along the direction $a$ is evaluated through the usual Euclidean inner product $a_1\xi_1 + \cdots + a_d\xi_d$, denoted also by $\langle a, X\rangle$.

It is clear that this definition leads to a random measure of depth, as it is is based on the rank of the projections along a random direction. Thus, the population depth of a datum $z$ with respect to a random element $X$ is defined as the (random) quantile corresponding to the value $\langle a, z\rangle$ in the distribution of $\langle a, X\rangle$. Of course, a single representative value can be obtained by averaging on $a$.

As for the distribution of $a$, different possibilities make sense. In the simulations below the direction $a$ is chosen according to a Gaussian distribution, in the appropriate space, standardized to norm 1; see Fishman (1996, p. 234).

In order to reduce variability, the depth of every datum can be assessed by averaging the depths obtained with a large number of different random directions. This has been done in our simulation study, where 50 random directions have been used.

As indicated above, our use of random projections relies on some ideas of Cuesta-Albertos et al. (2006a,b). These authors have obtained an extension, applicable to Hilbert spaces, of the classical Cramer–Wold theorem which characterizes a probability distribution in terms of one-dimensional projections. They also propose a statistical application to goodness of fit techniques. The present work could provides a further exploration of these ideas in the classification and estimation frameworks.

### The double random projection (RP2 and RPD) methods

These methods are valid only for functional data. We will present them here in sample versions, but the population counterparts can be readily defined with some obvious changes. Let $X_1, \ldots, X_n$ be a sample of differentiable functions defined on $[0,1]$. The basic idea is to use the method of random projections simultaneously for the functions and their derivatives thus incorporating the information on the function smoothness provided which is relevant in some practical applications. To be more precise, the sample of functions $X_1, \ldots, X_n$ is reduced to a sample in $\mathbb{R}^2$ defined by $(\langle a, X_1\rangle, \langle a, X_1'\rangle), \ldots, (\langle a, X_n\rangle, \langle a, X_n'\rangle)$, where $a$ is a randomly chosen direction. Now, depending on the treatment of this bi-dimensional sample there are several alternative possibilities. We will consider here two of them: the random projection method could be used again for the bi-dimensional projections $(\langle a, X_1\rangle, \langle a, X_1'\rangle), \ldots, (\langle a, X_n\rangle, \langle a, X_n'\rangle)$. We will denote this method by RP2. A further possibility is to use a different procedure in order to evaluate the depth of the bidimensional (projected) sample data. For example, we could use the $h$-modal depth in this second step: the resulting method will be denoted by RPD.

## 3 Applications to supervised classification with functional data

The supervised classification problem for two populations can be stated as follows: assume we have two independent "training samples" $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ taken, respectively, from the populations $P_0$ and $P_1$ described by the random variables $X$ and $Y$. The aim of classification procedures is to assign a new coming observation $Z$ to either $P_0$ or $P_1$ using the information provided by the training samples. Of course, this problem could also be stated in a more general version, for $k$ populations, but we will restrict ourselves to the simplest more usual case $k = 2$. Supervised classification is a relevant problem, sometimes referred to as "discrimination" in the statistical community. It is also a topic of leading current interest in the fields of "machine learning" (in computing science) and "pattern recognition" (in engineering). The systematic study of this subject goes back to the classical work by Fisher in the mid thirties, which led to well-known linear classifier.

The above mentioned sampling model, based on independent samples, is sometimes replaced by a "regression model" where the data consists of a sample $(X_i, \delta_i)$, $i = 1, \ldots, n$, $X_i$ being the "input" variables (on which the classification is based) and $\delta_i$ an indicator taking values 0 or 1, according to the membership of the corresponding observation to $P_0$ or $P_1$. Here the target is to predict the value of the variable $\delta$ from the observation of the corresponding $X$ in a new individual with $Z = (X, \delta)$. Most classification procedures can be used in both models, with obvious modifications and slightly different interpretations.

Besides the linear classifier (and its variants), the currently available methods include non-parametric alternatives (based on nearest neighbors or kernel classifiers), as well as some procedures relying on logistic regression and other more sophisticated approaches based on neural networks and support vector machines. A recent proposal by Ghosh and Chaudhuri (2005) combines linear discrimination ideas with some notions of data depth to get some interesting distribution-free properties. The books by Devroye et al. (1996) and Hastie et al. (2001) provide deep (in some sense complementary) accounts of classification topics.

The standard framework for supervised classification has been mainly developed for the case that the predictor variables $X_i$, $Y_j$ are $\mathbb{R}^d$-valued, with small or moderate values of the dimension $d$. Instead, the methods we propose here (as a by product of the ideas presented in Sect. 2) are suitable for functional data analysis. This means that the data $X_i$, $Y_j$ can be viewed as real functions defined on the unit interval $[0, 1]$.

While the theory of classification in FDA is still far from complete, some interesting developments have been recently made. An extension of the classical linear discriminant analysis to the case of functional data has been considered in James and Hastie (2001). Ferraty and Vieu (2003) have analyzed in the functional context the use of nonparametric kernel estimators of the function $P(\delta = 1|X = x)$ as auxiliary tools for supervised classification. A result of consistency for nearest-neighbors (k-NN) classifiers for infinite-dimensional data can be found in Cérou and Guyader (2005). It turns out that, unlike the finite-dimensional case (see Stone 1977; Devroye et al. 1996), the k-NN classifier is not universally consistent when it is applied to functional data. A classifier providing universal consistency when the data take values in Hilbert spaces is proposed in Biau et al. (2005). Further insights and references can be found in the books by Ramsay and Silverman (2005a,b) and Ferraty and Vieu (2006).

A general paradigm in classification theory can be expressed as follows: in order to classify the new coming datum $Z$ let's simply evaluate how deep it is when incorporated successively to samples 0 and 1. We will assign $Z$ to $P_0$ when its depth inside sample 0 is larger than that in sample 1. Such a principle is more or less implicitly used in different classification methods. For example, it is a well-known fact that the classical Fisher discriminant method amounts to classify the observation $Z$ by minimizing (on $i = 0, 1$) the Mahalanobis distance $(Z - \mu_i)'\Sigma^{-1}(Z - \mu_i)$, where $\mu_1, \mu_2$ are the means of $P_0$ and $P_1$, respectively, and $\Sigma$ is the common covariance matrix which of course must be estimated from the data. In this classical example the data depth in the population $P_i$ is measured by the Mahalanobis distance. Fisher's linear discriminant method is still, by far, the most popular procedure for supervised classification with finite-dimensional data buy our interest in this section is focused on the infinite-dimensional (functional) case and the extension of Fisher's ideas to that framework is not straightforward. Thus, we will instead analyze the applicability of the depth notions proposed in Sect. 2.

In the simulations below the idea is always the same: given the training data with observations correctly classified in $P_0$ or $P_1$, the observations of a new test sample are classified into $P_0$ or $P_1$ according to their depths in the respective training samples. We will use the nearest neighbor classifier (k-NN) as a benchmark in our comparisons. This method is a simple nonparametric alternative easy to apply in all situations, even with functional data. The definition is well-known: in order to classify a datum $Z$, let us evaluate the $k$ nearest data in the training sample (where $k$ is a tuning parameter fixed in advance) and decide the classification of $Z$ by "majority vote", according to the membership, to $P_0$ or $P_1$, of these nearest neighbors; see Devroye et al. (1996) and Cérou and Guyader (2005) for details on the k-NN classification method. We will use throughout the $L_2$ distance as a criterion of proximity between functions.

### 3.1 Simulation comparisons of several functional classifiers

Two models have been simulated in order to generate the functional samples:

*Model 1*: The population $P_0$ consists of trajectories of the process $X(t) = m_0(t) + e(t)$, where $m_0(t) = 30(1 - t)t^{1.2}$ and $e(t)$ is a Gaussian process with mean 0 and $Cov(X(s), X(t)) = 0.2 \exp(-|s - t|/0.3)$. The process corresponding to $P_1$ differs from $X(t)$ only in the mean function and is given by $Y(t) = m_1(t) + e(t)$, with $m_1(t) = 30(1 - t)^{1.2}t$.

*Model 2*: The population $P_0$ consists of trajectories of the process $X(t) = m_0(t) + e(t)$, where $m_0(t) = 30(1 - t)t^2 + 0.5|\sin(20\pi t)|$ and $e(t)$ a Gaussian process with mean 0 and $Cov(X(s), X(t)) = 0.2 \exp(-|s - t|/0.3)$. Population $P_1$ is made of spline approximations (with 8 knots) of trajectories from the previous process.

It can be seen that Model 1 corresponds to a more or less regular situation with smooth mean functions; they are shown in Fig. 1 (left). By contrast, Model 2 provides a more irregular situation where $m_0$ is a wiggly version of the smooth

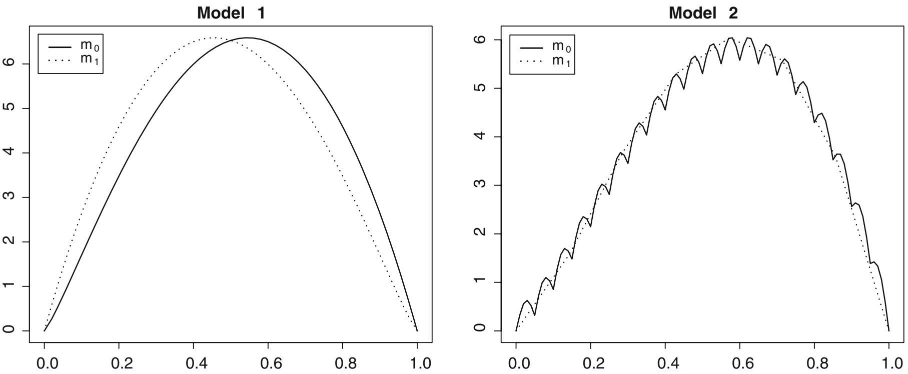

**Fig. 1** Mean functions for the processes $P_0$ and $P_1$ in Model 1 (*left*) and 2 (*right*)

**Table 1** Descriptive statistics for the correct classification proportions under Model 1

| | *h*-modal | FM | k-NN | RP | RP2 | RPD |
|---|---|---|---|---|---|---|
| Minimum | 0.9200 | 0.9200 | 0.9800 | 0.9700 | 0.9300 | 0.9400 |
| First quartile | 0.9800 | 0.9700 | 1.0000 | 0.9900 | 0.9800 | 0.9700 |
| Median | 0.9900 | 0.9800 | 1.0000 | 1.0000 | 0.9900 | 0.9800 |
| Mean | 0.9847 | 0.9773 | 0.9977 | 0.9961 | 0.9849 | 0.9810 |
| Third quartile | 1.0000 | 0.9900 | 1.0000 | 1.0000 | 0.9925 | 0.9900 |
| Maximum | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

function $m(t) = 30(1-t)t^2$ and the population $P_1$ is made of rough, less wiggly, approximations of the functions in $P_0$. The mean functions of $P_0$ and $P_1$ are shown in Fig. 1 (right). Of course, the purpose of such a model is to check the performance of the different methods to detect the small waves in the trajectories of $P_0$. The results of our Monte Carlo study are summarized in Tables 1 and 2. These are the relevant technical details:

*Number of runs*: Our simulation results are based on 100 independent runs.
*Training and test samples*: In every run, 200 "training" observations are generated for each model: 100 come from $P_0$ and 100 from $P_1$. Likewise, a test sample of 100 observations (50 from $P_0$ and 50 from $P_1$) is generated and the proportion of correct classifications is recorded for these observations.
*Interpretation of the tables*: The outputs in the tables cells provide the main descriptive statistics (mean, median, quantiles,...) for the distribution of the correct classification proportions over the 100 runs.
*Tuning parameters*: The parameter $h$ in the $h$-mode is chosen as the 20 percentile in the $L_2$ distances between the functions in the training sample. The k-NN procedure is used with $k = 5$.
*Discretization*: The functions are handled in a discretized version based on 51 equispaced grid points on $[0, 1]$.
*Choice of the projection directions*: The RP methods are implemented by projecting the discretized trajectories along 50 independent normalized

**Table 2** Descriptive statistics for the correct classification proportions under Model 2

| | *h*-modal | FM | k-NN | RP | RP2 | RPD |
|---|---|---|---|---|---|---|
| Minimum | 0.7200 | 0.5900 | 0.8200 | 0.6300 | 0.9500 | 0.9000 |
| First quartile | 0.7900 | 0.7000 | 0.8700 | 0.7800 | 0.9975 | 0.9800 |
| Median | 0.8200 | 0.7300 | 0.8900 | 0.8000 | 1.0000 | 0.9900 |
| Mean | 0.8197 | 0.7332 | 0.8912 | 0.8042 | 0.9962 | 0.9884 |
| Third quartile | 0.8525 | 0.7700 | 0.9100 | 0.8400 | 1.0000 | 1.0000 |
| Maximum | 0.9200 | 0.8400 | 0.9700 | 0.9300 | 1.0000 | 1.0000 |

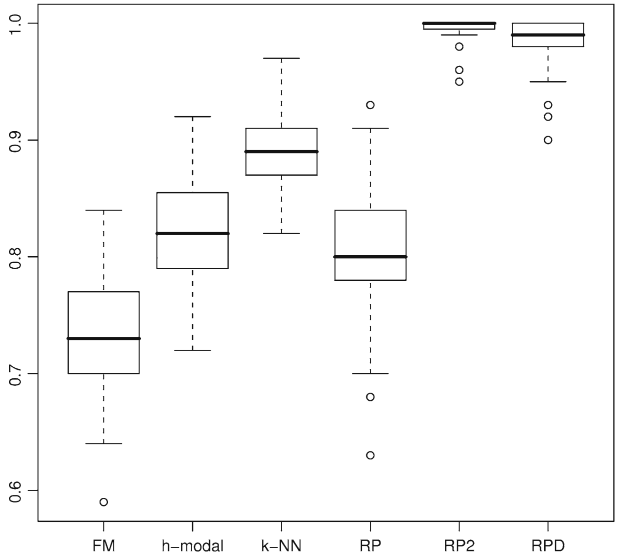

**Fig. 2** *Box* plots for the proportions of correct classification corresponding to Table 2

(‖.‖ = 1) standard Gaussian random directions with dimension 51 (the same as the discretization size).

Thus, in summary, the k-NN method is the winner, in efficiency and variability, under Model 1 but the performance of the RP method is quite similar. Under Model 2 there is a much higher variability but the methods based on derivatives, RP2 and RPD (especially the first one) clearly outperform the remaining classifiers. This is visualized in the box plots of Fig. 2. So, when differences in smoothness between the functions of $P_0$ and $P_1$ are expected, there is a good case to use the double projection methods.

## 3.2 A comparison based on a real data example

We will consider here a data set already used for other authors (e.g., Ramsay and Silverman 2002). The data correspond to the heights of 54 girls and 39 boys

**Table 3** Proportions of correct classification obtained for the growth data

| | $h$-modal | FM | k-NN | RP | RP2 | RPD |
|---|---|---|---|---|---|---|
| First quartile | 0.8696 | 0.6522 | 0.9130 | 0.7717 | 0.7826 | 0.8696 |
| Median | 0.9130 | 0.7391 | 0.9565 | 0.8261 | 0.8696 | 0.9130 |
| Mean | 0.8822 | 0.7243 | 0.9596 | 0.8135 | 0.8513 | 0.9096 |
| Third quartile | 0.9130 | 0.7826 | 1.0000 | 0.8696 | 0.9130 | 0.9565 |

recorded between the ages of 1 and 18 years. We have selected 70 subsamples with 40 girls and 30 boys each and used them to predict the group membership for the remaining 23 curves. For each subsample, the proportion of correct classification has been obtained. The corresponding relevant values (quartiles, median, mean) (over the 70 subsamples) are given in Table 3.

The results are summarized in Fig. 3 below. In this case the k-NN method is the winner and RPD ranks second.

To give an idea of the relative computational burden of the different methods, let us mention that if we take the smaller CPU time (0.12 s for the k-NN method) as a unit, the computational performances, defined as the corresponding quotients (CPU($h$-modal)/CPU(k-NN), etc.) of CPU times, are as follows: $h$-modal = 2.0833, k-NN = 1, FM = 17.75, RP = 28.1667, RP2 = 1398.9, RPD = 79.9167.

## 4 Applications to estimation

### 4.1 Monte Carlo comparisons in finite-dimensional examples

In all models considered in this subsection the samples consist of 80 observations drawn from a Gaussian "central" distribution $N(\mu, \Sigma)$ plus 20 additional observations (outliers) coming from a "contaminating" distribution far away from the central one.

Model 3 provides an example of asymmetric contamination where the central parameters are

$$\mu = (0, 0)', \quad \Sigma = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 0.5 \end{pmatrix},$$

and the outliers come from another Gaussian distribution with the same covariance matrix and means vector $\mu_o = (3, 3)'$.

In Model 4 we consider a situation of "symmetric" contamination: the central distribution is the same as in Model 3 but every outlier comes, with equal probability from either $N(\mu_o, \Sigma)$ or $N(-\mu_o, \Sigma)$, where $\mu_o = (3, 3)'$.

In Model 5 the central distribution is a ten-dimensional Gaussian with $\mu = (0, 0, \ldots, 0)$ and $\Sigma$ a diagonal matrix with equispaced diagonal elements $\sigma_{1,1} = 0.5, \ldots, \sigma_{10,10} = 1$. In this case the 20 outliers come from $N(\mu_o, \Sigma_o)$, with $\mu_o = (3, 3, \ldots, 3)$ and $\Sigma_o = 0.5\Sigma$.

The numerical outputs in the Tables 4, 5 and 6 below provide the average distances, over 500 runs, from the different estimators to the "central value"

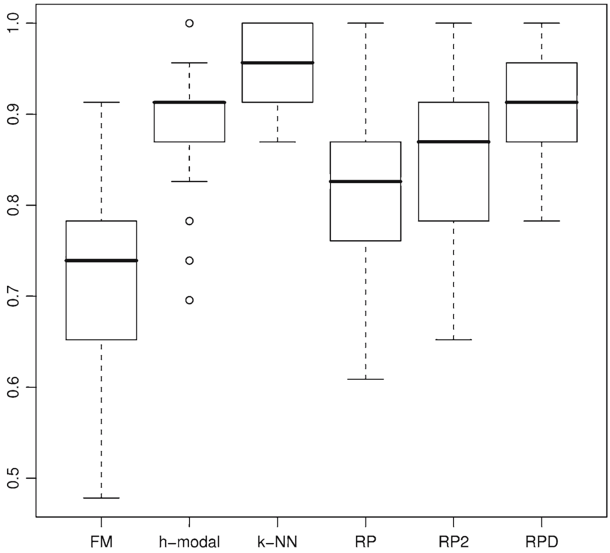

**Fig. 3** *Box* plots for the proportions of correct classification obtained by using the different methods with the growth data

represented by the means vector $\mu$. So the aim is to check the robustness of the estimators against the outliers. Three distances have been used: the usual Euclidean metric ($L_2$), the sum distance ($\sum_i |x_i - y_i|$), denoted by $L_1$, and the maximum distance ($\max_i |x_i - y_i|$), denoted by $L_\infty$. Of course, the main conclusions are usually the same under all the three distances but sometimes arise differences worth considering. In all cases we have used three of the depth measures: FM, $h$-modal and RP (based on 50 random directions, as in Sect. 3.1). For each of them, two location estimators have been considered: the sample median (i.e. the deepest sample observation) and the 0.25-trimmed mean (i.e. the average of the 75% deepest sample observations). As a reference for comparisons we have also included the ordinary sample mean and the partial average of all the sample observations excluding the 20 outliers. We call this artificial statistic the "tricky mean" as it is obvious that it only can be used in the context of a simulation study when the outlying observations are identified with no error.

In the two bi-dimensional examples (Models 3 and 4), the simplicial depth (SD) has been included in our simulations as an additional procedure to define medians and trimmed means. The SD depth of a vector $z$ is defined in terms of the number of simplices (triangles in the bi-dimensional case), with vertices at the sample points, where $z$ is included. This is a well-known notion of depth which has received considerable attention in the literature; see, e.g., Liu et al. (1999). It is more suitable for the bi-dimensional case, as its use in high dimensions entails a prohibitive computational burden.

Some conclusions:

1. The best performance among the depth measures corresponds, uniformly, to the $h$-modal depth.

**Table 4** Comparison of different location estimators under Model 3

| | Medians | | | Trimmed means | | |
|---|---|---|---|---|---|---|
| | $L_2$ | $L_1$ | $L_\infty$ | $L_2$ | $L_1$ | $L_\infty$ |
| SD | 0.7444442 | 0.9301753 | 0.6905463 | 0.549245 | 0.7659218 | 0.4344239 |
| FM | 0.4022886 | 0.5306568 | 0.3497064 | 0.4092533 | 0.5658106 | 0.3303732 |
| *h*-modal | 0.2086686 | 0.2711216 | 0.1841219 | 0.1502731 | 0.1932119 | 0.1336811 |
| RP | 0.3866646 | 0.5232521 | 0.3239475 | 0.4779928 | 0.6636608 | 0.3834305 |
| | Mean | | | Tricky mean | | |
| | 0.8519566 | 1.200490 | 0.6410113 | 0.1178718 | 0.1501102 | 0.1062546 |

**Table 5** Comparison of different location estimators under Model 4

| | Medians | | | Trimmed means | | |
|---|---|---|---|---|---|---|
| | $L_2$ | $L_1$ | $L_\infty$ | $L_2$ | $L_1$ | $L_\infty$ |
| SD | 0.734442 | 0.9230511 | 0.67915 | 0.547715 | 0.7644198 | 0.4330717 |
| FM | 0.4155896 | 0.5512968 | 0.3598973 | 0.4069715 | 0.563881 | 0.3269491 |
| *h*-modal | 0.2065237 | 0.2695839 | 0.1809284 | 0.1535719 | 0.1982366 | 0.1367727 |
| RP | 0.3830504 | 0.518503 | 0.3214765 | 0.4855656 | 0.6756114 | 0.3868042 |
| | Mean | | | Tricky mean | | |
| | 0.84858 | 1.19602 | 0.6365693 | 0.1227913 | 0.1561855 | 0.1105188 |

**Table 6** Comparison of different location estimators under Model 5

| | Medians | | | Trimmed means | | |
|---|---|---|---|---|---|---|
| | $L_2$ | $L_1$ | $L_\infty$ | $L_2$ | $L_1$ | $L_\infty$ |
| FM | 1.746196 | 4.350537 | 1.139259 | 0.3220637 | 0.8297678 | 0.1992795 |
| *h*-modal | 1.354274 | 3.508538 | 0.8196287 | 0.3125208 | 0.8050044 | 0.1912802 |
| RP | 1.594079 | 4.107386 | 0.9760013 | 0.3624315 | 0.9407486 | 0.221241 |
| | Mean | | | Tricky mean | | |
| | 1.918720 | 6.0181 | 0.7297597 | 0.2964808 | 0.7648882 | 0.1813836 |

2. The best estimator is the trimmed mean (when combined with the *h*-modal depth) in all considered cases of models and distance measures. However, the differences between the different depth measures are smaller in the high dimensional case (Model 5).
3. While the "loss of performance" of the trimmed mean when compared with the ideal unfeasible "tricky mean" is quite acceptable, the gains with respect to the ordinary sample mean are impressive.

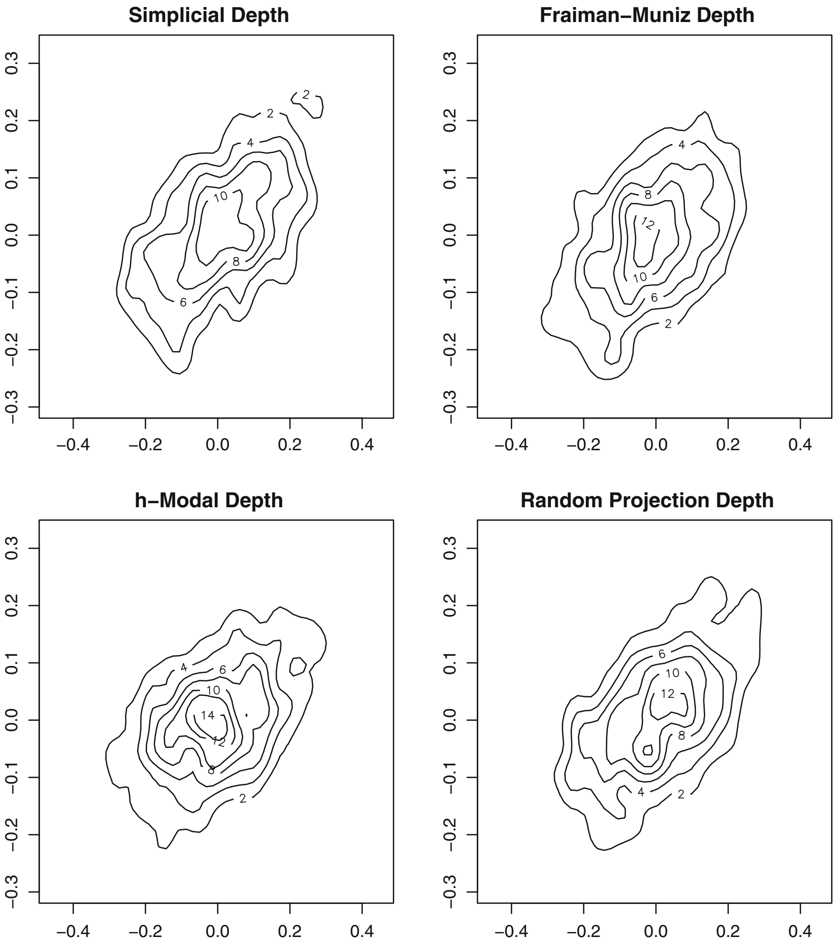

**Fig. 4** Estimated density of the distance from the trimmed mean to the target point $(0, 0)$

4. The general conclusions are quite consistent along the different distances but the relative differences between the estimators are smaller in the $L_\infty$ metric.

The behavior of the four depth measures is visualized in Fig. 4 that shows the level curves of the estimated density of the $L_2$ distance from the values of trimmed mean to the target $(0, 0)$.

## 4.2 Monte Carlo comparisons in examples with functional data

The approach in this subsection is similar to that followed in the previous one: Every sample consist of 100 functions (trajectories) which are generated as follows. There is a central process (whose mean function is denoted by $m = m(t)$, $t \in [0, 1]$) which provides 80 trajectories out of 100 in every sample.

**Table 7** Comparison of different location estimators under Model 6

| | Medians | | | Trimmed means | | |
|---|---|---|---|---|---|---|
| | $L_2$ | $L_1$ | $L_\infty$ | $L_2$ | $L_1$ | $L_\infty$ |
| FM | 2.102944 | 12.07222 | 0.7029948 | 0.5200344 | 3.114435 | 0.1479091 |
| *h*-modal | 1.704636 | 9.765081 | 0.5807322 | 0.4467659 | 2.668422 | 0.1302641 |
| RP | 2.125785 | 12.32344 | 0.6907975 | 0.6641425 | 4.021498 | 0.1782564 |
| RP2 | 2.257558 | 13.16562 | 0.7197539 | 0.5523836 | 3.320757 | 0.1541845 |
| RPD | 1.930039 | 11.07713 | 0.6413013 | 0.4412892 | 2.63681 | 0.1288688 |
| | Mean | | | Tricky mean | | |
| | 2.918682 | 18.48780 | 0.6240016 | 0.4088806 | 2.429827 | 0.1213473 |

**Table 8** Comparison of different location estimators under Model 7

| | Medians | | | Trimmed means | | |
|---|---|---|---|---|---|---|
| | $L_2$ | $L_1$ | $L_\infty$ | $L_2$ | $L_1$ | $L_\infty$ |
| FM | 2.086320 | 11.90465 | 0.7121903 | 0.5823091 | 3.504831 | 0.1610501 |
| *h*-modal | 1.727930 | 9.890137 | 0.5933656 | 0.450189 | 2.708423 | 0.1297920 |
| RP | 2.128670 | 12.37386 | 0.6887023 | 0.7452996 | 4.539249 | 0.1967323 |
| RP2 | 2.211855 | 12.87983 | 0.7115233 | 0.6460436 | 3.903646 | 0.1760337 |
| RPD | 1.962318 | 11.33803 | 0.6475603 | 0.4461833 | 2.677843 | 0.1289595 |
| | Mean | | | Tricky mean | | |
| | 2.19362 | 13.89988 | 0.4785969 | 0.4095047 | 2.453991 | 0.1203139 |

The remaining 20 sample data are outliers coming from a process with the same structure as the central one but with a different mean function, $m_o = m_o(t)$.

Two models (numbered 6 and 7) will be considered. In both of them the central and the outlying processes are identical to those in Sect. 3.1 except for the mean functions which are, respectively, $m(t) = 30(1-t)t^3$, $m_o(t) = 30(1-t)^3 t$ (Model 6) and $m(t) = 30(1-t)t^{1.5}$, $m_o(t) = 30(1-t)^{1.5} t$ (Model 7). It can be seen that $m$ and $m_o$ are closer together in Model 7 than in Model 6. As a consequence the outliers are easier to identify, but potentially more harmful, in Model 6.

The interpretation of the Tables 7 and 8 below is completely similar to that of Tables 4, 5 and 6 in the previous subsection. In particular, every output in the tables is the average over 500 runs of the distances (evaluated with the respect to the functional distances $L_2$, $L_1$ and $L_\infty$ from the respective estimators to the "central function" $m(t)$.

The best results are obtained for the trimmed mean based on the RPD depth though the performance of this estimator with the *h*-modal depth is only slightly lower.

### 4.3 A remark on breakdown points

The notion of *replacement breakdown point* (or *finite sample breakdown point*), introduced by Donoho (1982), can also be used for estimators taking values in general metric spaces in order to assess their global robustness. Suppose we have a sample of random elements taking values on an arbitrary space $\mathbb{E}$. Let $\mathbb{E}^n$ denote the $n$-product space, $(\Gamma, d)$ a metric space, and let $T_n : \mathbb{E}^n \to \Gamma$ be a sequence of $\Gamma$-valued estimators.

Given a sample $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{E}^n$ the replacement breakdown point (RBP) of $T_n$ is defined as

$$\text{RPB}(T_n(\mathbf{X})) = min\left\{\frac{m}{n} : sup_{\mathbf{X^m}} d\left(T_n(\mathbf{X}), T_n(\mathbf{X^m})\right) = \infty\right\},$$

where $\mathbf{X^m}$ denotes the data set obtained by replacing $m$ coordinates of $\mathbf{X}$ with arbitrary values on $\mathbb{E}$, and $\infty$ stands for $sup_{v\in\mathbb{E}} d(T_n(\mathbf{X}), v)$ (just to cover also the case of bounded distances).

Roughly speaking, the RPB is the proportion of data which must be moved to $\infty$ in order to get that the estimator will do the same. Typically, the breakdown point is achieved when we replace the $m$ coordinates all together with the same value $\infty$. In this sense, the definition is quite pessimistic with respect to the behaviour of nature.

It is not difficult to see that the FM-median and the RP-median have RPB $\geq \frac{1}{2}$ for all $\mathbf{X}$, and for any dimension (even for infinite dimension). This is not the case of the $h$-modal median, where the RPB will depend on the parameter $h$. However, in order to get the breakdown of this estimator we will need to have enough replacement points close enough to each other, i.e. within a ball of radius $h$. In the case of $d$-dimensional data, for large enough sample sizes, the RPB of the $h$-mode will be of order $h^d$. The basic idea is that, under some regularity conditions, the breakdown ratio $m/n$ should fulfil $\frac{m}{n} \sim max_x f(x) h^d \lambda_d$, where $f$ stands for the underlying density of the data and $\lambda_d$ for the Lebesgue measure of the unit ball in $\mathbb{R}^d$. A formal closer look at the behavior of the RPB for this estimator goes beyond the scope of this paper.

**Acknowledgments** The constructive comments from an anonymous referee are gratefully acknowledged.

## References

Biau G, Bunea F, Wegkamp M (2005) Functional classification in Hilbert spaces. IEEE Trans Inf Theory 51:2163–2172

Cérou F, Guyader A (2005) Nearest neighbor classification in infinite dimension. Preprint

Cuesta-Albertos JA, Fraiman R, Ransford T (2006a) A sharp form of the Cramer–Wold theorem. J Theor Probab (in press)

Cuesta-Albertos JA, Fraiman R, Ransford T (2006b) Random projections and goodness-of-fit tests in infinite-dimensional spaces. Boletim da Sociedade Brasileira de Matematica 37:1–25

Cuevas A, Febrero M, Fraiman R (2006) On the use of the bootstrap for estimating functions with functional data. Comput Stat Data Anal 51:1063–1074

Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer, Heidelberg

Donoho DL (1982) Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Dept. Statist., Harvard University

Ferraty F, Vieu P (2003) Curves discrimination: a nonparametric functional approach. Comput Stat Data Anal 44:161–173

Ferraty F, Vieu P (2006) Nonparametric modelling for functional data. Springer, Heidelberg

Fishman GS (1996) Monte Carlo: concepts, algorithms and applications. Springer, Heidelberg

Fraiman R, Meloche J (1999) Multivariate L-estimation. Test 8:255–317

Fraiman R, Muniz G (2001) Trimmed means for functional data. Test 10:419–440

Ghosh AK, Chaudhuri P (2005) On data depth and distribution-free discriminant analysis using separating surfaces. Bernoulli 11:1–27

Hastie T, Buja A, Tibshirani R (1995) Penalized discriminant analysis. Ann Stat 23:73–102

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Data mining, inference, and prediction. Springer, New York

James GM, Hastie TJ (2001) Functional linear discriminant analysis for irregularly sampled curves. J R Stat Soc B 63:533–550

Liu RY (1990) On a notion of data depth based on random simplices. Ann Stat 18:405–414

Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference. With discussion and a reply by the authors. Ann Stat 27:783–858

Ramsay JO, Silverman BW (2002) Applied functional data analysis. Methods and case studies. Springer, New York

Ramsay JO, Silverman BW (2005a) Applied functional data analysis. Springer, Heidelberg

Ramsay JO, Silverman BW (2005b) Functional data analysis, 2nd edn. Springer, Heidelberg

Rousseeuw PJ, Hubert M (1999) Regression depth (with discussion). J Am Stat Assoc 94:388–433

Stone CJ (1977) Consistent nonparametric regression. With discussion and a reply by the author. Ann Stat 5:595–645

Tukey JW (1975) Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, pp 523–531. Canad. Math. Congress, Montreal

Zuo Y (2003) Projection based depth functions and associated medians. Ann Stat 31:1460–1490

Zuo Y, Serfling R (2000) General notions of statistical depth function. Ann Stat 28:461–482