University of Tunis
Institut Supérieur de Gestion

University of Cassino and Southern Lazio

A dual degree PhD program

Depth-based classification approaches for directional data

Houyem Demni

Doctor of Philosophy in Statistics

PhD program in

'Imprese, Istituzioni e Comportamenti'
University of Cassino and Southern Lazio
'Sciences de Gestion: Management'
University of Tunis

Supervisors:
Prof.Giovanni Camillo Porzio
Prof.Amor Messaoud

April 20, 2021

University of Tunis
Institut Supérieur de Gestion

University of Cassino and Southern Lazio

Doctoral Thesis

Depth-based classification approaches for directional data

Houyem Demni

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics

PhD program in
'Imprese, Istituzioni e Comportamenti', University of Cassino and Southern Lazio
'Sciences de Gestion: Management', University of Tunis

Supervisors:
Prof.Giovanni Camillo Porzio
Prof.Amor Messaoud

April 20, 2021

Acknowledgements

At the end of a journey, it often happens to look back and see how far you have gone, how much you enjoyed the path, and to see who accompanied you alongs the adventure. All of you shall be thanked, it would be not possible to finish this course without you.

First, I show heartfelt gratitude to my Tunisian supervisor Dr.Amor Messaoud for his support, motivation and expertise. You have guided me since the very beginning of my PhD and you helped me to face many issues of my research. I will never be able to thank you enough.

I would like to express my sincere gratitude to my Italian supervisor Prof.Giovanni Camillo Porzio who introduced me the field of Directional Statistics during my first visit to the University of Cassino and Southern Lazio. I am extremely grateful to you not only for never letting me lose the track, and being always proactive in having me to explore new routes. But, also for your continuous support, for being always helpful and comprehensive and for doing your best to provide better conditions during my stay in Italy. I have learnt a lot from you and you inspired me a lot. This thesis would not have been possible without you.

I appreciate the University of Cassino and Southern Lazio for providing funds to my research and I am very grateful to the staff of the International Office, especially Laura Morone for her support during all the period of my stay in Italy. I would like also to deeply acknowledge Giuseppe Pandolfo, Alfonso Iodice D'Enza, Davide Buttarazzi and Mario Guarracino for both the stimulating discussions and the advice. I wish you all the best for the future.

I thank Dr.Ondrej Vencalek from the Palacky University in Olomouc, Czech Republic, for his feedback, for hosting me during my stay in Czech Republic as well as for the fruitful cooperation, which led to a joint paper.

I am extremely grateful to all my friends. Particularly, I thank all my friends in Tunisia for encouraging me as well as my friends in Italy for their substantial support from my very first days there. Thank you for making me feel like at home. I also thank all the friends that I met during this journey. It was a pleasure to meet you.

Last but not least, to my family who loves me unconditionally regardless of all my flaws. I would like to thank my parents, my sister and my brother, for having believed in me and for their unfailing support and continuous encouragement throughout all these years. No words can express how much I am grateful to you. I owe you everything. Thank you from the bottom of my heart.

Abstract

Supervised learning tasks aim to define a data-based rule by which new objects are assigned to one of the given classes. To this end, a training set containing objects with known memberships is exploited. Directional data are points lying on the surface of circles, spheres or hyper-spheres. Given that they lie on a non-linear manifold, directional observations require specific methods to be analyzed. In this thesis, the main interest is to present novel methodologies and to perform reliable inferences for directional data, within the framework of supervised classification. First, a supervised classification procedure for directional data is introduced. The procedure is based on the cumulative distribution of the cosine depth, that is a directional distance-based depth function. The proposed method is compared with the max-depth classifier, a well-known depth-based classifier within the literature, through simulations and a real data example. Second, we study the optimality of the depth distribution and the max-depth classifiers from a theoretical perspective. More specifically, we investigate the necessary conditions under which the classifiers are optimal in the sense of the optimal Bayes rule. Then, we study the robustness of some directional depth-based classifiers in the presence of contaminated data. The performance of the depth distribution classifier, the max-depth classifier and the DD-classifier is evaluated by means of simulations in the presence of both class and attribute noise. Finally, the last part of the thesis is devoted to evaluate the performance of depth-based classifiers on a real directional data set.

Preface

In the last decades, directional data analysis has emerged as an interesting field in statistics. In many contexts, the considered natural supports of data are Riemannian manifolds: the unit circle, the sphere and their extensions in higher dimensions. Such kind of data can be described by one or more directions.

Analyses of directional data through statistical methods focus on their graphical representation and on tasks related to regression, time series, correlation, image analysis, text mining and machine learning. Analyzing directional data implies specific methods given their specific nature. A wide survey on the theory and the methodology of directional statistics can be found in Mardia & Jupp (2009).

Machine learning methods are an integral part of many practical problems. For instance, classification is an innate task done by human beings every day for hundreds of times. Unconsciously, the human brain tends to classify all the objects that surround us based on past experiences and according to their properties.

The learning process that each one undertook creates a gained knowledge from which one can distinguish between classes, indicate their differences and determine the class membership of a new observed object. Over the last decades, this process was automatized due to the increasing of computing power. Thus, a new field known as machine learning has emerged with the main goal of developing computer algorithms for supervised learning.

Supervised learning procedures aim to define data-based rules by which new objects can be assigned to one of the existing classes. Generally speaking, objects are regarded as data points in the multivariate space where each is described by a set of features (explanatory variables) and a class label (dependent variable). Then, through analyzing the training set in which those data points have a known class membership,

one can infer a separating function to be used in order to assign new points to the class to which they most probably belong.

Supervised classification methods should figure out how close an observation is situated with respect to a class. This could be done through studying location, shape and scale of the underlying distribution. The closeness of a data point to a class can be defined as a measure of correspondence to the entire class or as a measure of distance to a predefined center.

Several existing classification procedures rely either on a certain parametric distributions for the data or a certain forms of separating curves. Parametric classifiers are not fully useful in real applications especially when little information about the underlying distributions is available. On the other hand, non parametric classifiers are more flexible and more desirable given their ability to accommodate different data structures.

The statistical function known as data depth is a measure of centrality and a representative way of ordering data points regarding to their centrality within the data group. In particular, depth functions have been employed in classification given their parameter-free nature which allow classifiers to have attractive theoretical properties. On the other hand, data depth is promising for directional observations since no standard ordering is available for such kind of data.

Within the one-dimensional space, it is always possible to order observations given their magnitude or their median (known as robust location measure). Moving to the higher dimensional space, there is no natural way to rank the data. This motivated the introduction of depth functions, which provide a center outward ordering of all data points from a center (deepest point) of a given multivariate distribution.

Based upon these considerations, the aim of this work is to propose depth-based approaches for dealing with directional data in the context of classification. This thesis contains five chapters. The first focuses on introducing the main concepts and notions that will be adopted and developed in the remaining part of the monograph. The second chapter, named *A Directional Depth Distribution Classifier based on the Cosine Depth*, introduces a supervised classification procedure for directional data, that is based on the cosine depth function. The proposed depth distribution classifier is based on the

distribution function of the cosine depth and it aims at assigning directional data observations to classes. The new method is compared to the max-depth classifier which is based on the depth value. Our results show the effectiveness of our proposal. The cosine depth distribution improves over the max-depth distributions in many different settings.

In Chapter 3, which is named *On the optimality of the max-depth and depth distribution classifiers for spherical data*, we investigate conditions under which the max-depth classifier and the depth distribution classifier (known also as max-rank classifier) are equivalent to the optimal Bayes rule. They are optimal if distributions are unimodal, rotational symmetric, differ only in location parameters and have equal prior probabilities.

Chapter 4, named *Distance-based directional depth classifiers: a robustness study*, tackles the problem of robustness of distance depth-based classifiers for directional data. We mainly consider in this study three classifiers which are the max-depth, the depth distribution and the DD-classifier. We compare their performance in presence of class noise and label noise with respect to the Bayes classifier. We introduce some directional specific contamination schemes: antipodality and orthogonality of the contaminated distribution mean, and the directional mean shift outlier model.

In Chapter 5, which is named *Directional supervised learning through depth functions: an application to ECG waves analysis*, we apply the depth-based classifiers to a real data set. Some final remarks are offered to conclude the manuscript and summarize the main contributions. Lastly, The Appendix A supplements Chapter 4 with additional material related to some computational aspects while Appendix B provides the main R function used in this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction to Directional data

Directions can be depicted in two, three or more dimensions (on a q-dimensional hypersphere). Directional data presented on these different manifolds come up in many areas such as Biology, Genetics, Astronomy, Physics, Earth Sciences, Meteorology, Medicine, and Social Sciences.

Working with directional data requires the use of special methods that take into account the angular nature of the information. For instance, 0 and 2π are the same angle and their average is 0 and not π (Fernandes & Cardoso, 2016). The periodical behavior makes linear statistics methods inappropriate for this kind of data.

Directional statistics offers a wide range of techniques as well as theoretical background to successfully explore and work with directional information. A recent survey on advances of directional statistics can be found in Pewsey & García-Portugués (2020) where developments in different fields including classification are considered.

The aim of this chapter is to provide an overview of notions and statistical tools useful for the analysis of directional data. It has been conceived in order to better understand the topics covered within the next chapters of this work. We provide background information related to circular data, spherical data and we also review supervised classification methods for directional data.

1.1 Circular data

A direction like wind direction can be recorded and graphically represented as an angle θ (in radians or degrees) on a circle. Otherwise, this kind of data refereed as circular data can be depicted as unit vectors connecting the origin of the circle to these points or as points on the circumference of the unit circle centered at the origin. This implies a reference direction (the zero direction) and a sense of rotation (clockwise or counter-clockwise) that can be chosen given the nature of the data and/or the goal of the analysis.

Once these two conditions are fulfilled, each directional point x can be defined by an angle θ or by a complex number z. Thus, the relation can be represented as

$$x = (\cos\theta, \sin\theta)^T,$$

or

$$z = e^{i\theta} = \cos\theta + i\sin\theta,$$

where θ is a measured angle.

In the directional domain, the use of the arithmetic mean as measure of center for circular data is meaninglessness because of its dependence on the reference direction and on the sense of rotation. A useful measure known as the mean direction or as the circular mean can be obtained by considering data points as unit vectors. Thus, given unit vectors $x_1, .., x_n$ with corresponding angles $\theta_1, .., \theta_n$, the mean direction $\bar{\theta}$ is obtained by solving the following equations

$$\bar{C} = \bar{R}\cos\bar{\theta},\ \bar{S} = \bar{R}\sin\bar{\theta},$$

and the mean resultant vector $\bar{R}$ is given by:

$$\bar{R} = ||R|| = (\bar{C}^2 + \bar{S}^2)^{1/2}.$$

The mean resultant length is the length of the mean vector while the mean direction is the projection of such vector on circle. Figure 1.1 shows the circular mean, depicted by the black arrow where some random circular data points are generated.

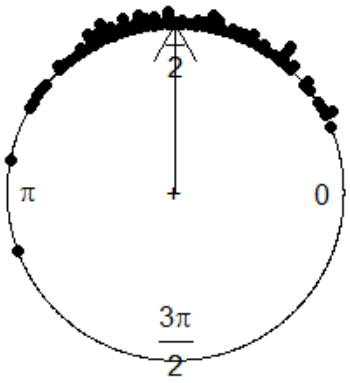


Figure 1.1: Circular representation of simulated circular random observations, the black arrow represents the sample mean direction.

The mean resultant vector $\bar{R}$ measures the dispersion of circular data. It lies in the range $[0, 1]$ in the sense that the smaller is the mean resultant length, the more the data are dispersed and the higher is, the higher circular observations are clustered. Hence, the circular variance can be defined as

$$V = 1 - \bar{R}.$$

The range of the circular variance is also $[0, 1]$. If $V = 0$, the distribution is highly concentrated.

The probability of a circular distribution is concentrated on the circumference of a unit circle, and the range of circular random variables (measured in radians) is $[0, 2\pi)$ or $[\pi, -\pi)$. The most common used distribution in circular statistics is the von Mises distribution (Von Mises, 1918) which is unimodal and symmetric.

The probability density function of the von Mises distribution can be defined as

$$f(\theta; \mu, c) = \frac{1}{2\pi I_0(c)} e^{c\cos(\theta-\mu)},$$

where I_0 is the modified Bessel function of the first kind and order 0, μ is the mean direction and c is the concentration parameter.

Figure 1.2 shows the circular density plot for a random sample generated from a von Mises distributions with $\mu = \pi$ and $c = 3$.

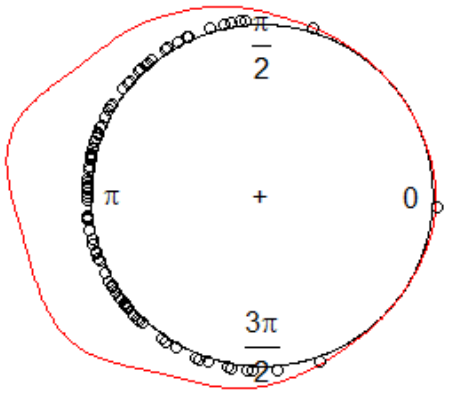


Figure 1.2: Density plot of random circular data generated from the von Mises distribution with circular mean equal to π and concentration equal to 3.

1.2 Spherical data

Directions observed in $\mathbb{R}^3$ such as the positions of stars on the celestial sphere can be expressed by a pair of angles (θ,φ) or by a 3×1 unit vector on the unit sphere S^2 and they are distinguished by the term spherical data. Directions in $q-$dimensions can be described as unit vectors x, as points on $S^{q-1} = \{x\colon x'x = 1\}$ on the $(q-1)$-dimensional sphere with unit radius and center at the origin.

Directional data can be stored as spherical coordinates or as polar coordinates. The spherical coordinates of any point x on the sphere (depicted in Figure 1.3) can be obtained as

$$x = (\cos\theta, \sin\theta\cos\varphi, \sin\theta\sin\varphi)^T,$$

where $\theta \in [0, 2\pi)$ and $\varphi \in [0, \pi]$.

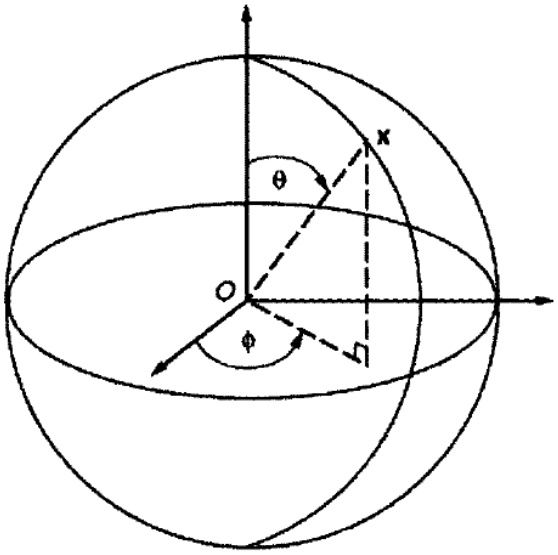


Figure 1.3: Spherical coordinates: θ= colatitude, φ= longitude (Mardia & Jupp, 2009).

On the other hand, the polar coordinates which represent the radius r, the inclination θ and the azimuth φ of any point x are obtained from its Cartesian coordinates (x, y, z) by

$$\begin{cases} r = \sqrt{x^2 + y^2 + z^2} \\ \theta = \arccos(\frac{z}{r}) \\ \varphi = \arctan(\frac{y}{x}) \end{cases}$$

Descriptive Measures

The mean direction and the mean resultant length: Assuming that $x_1, .., x_n$ are points on $S^{(q-1)}$, the location of these points can be given by their sample mean direction

$$\bar{x_0} = ||\bar{x}||^{-1}\bar{x},$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is the center mass of $x_1, .., x_n$.

As in the circular case, the vector $\bar{x}$ in polar coordinates can be expressed by

$$\bar{x} = \bar{R}\bar{x_0}.$$

where $\bar{R}$ is the mean resultant length, a common measure of concentration for spherical data

$$\bar{R} = ||\bar{x}||, \bar{R} \geq 0.$$

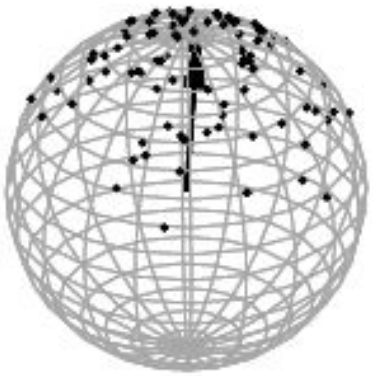


Figure 1.4: A random sample from a von Mises-Fisher distribution vMF ((0,0,1)',5). The black arrow shows the sample mean direction.

Figure 1.4 shows a group of 100 points from the von Mises-Fisher distribution vMF((0,0,1)',5) defined on the sphere S^2.
At the population level, we define analogously, the population mean resultant length ρ for a random unit vector as

$$\rho = \sum_{i=1}^{q} E[x_i]^2)^{1/2} = \{E[x]^T E[x]\}^{1/2}.$$

The population mean direction mean direction, when $\rho \leq 0$, can be defined as

$$\mu = \rho^{-1} E[x].$$

The mean direction of spherical data has the following equivariance properties. Assuming that U is an orthogonal transformation (i.e. a rotation) and $x_1, .., x_n$ a sample of spherical observations, the mean direction of $U_{x_1}, .., U_{x_n}$ is $U_{\bar{x_0}}$. The mean resultant length of $U_{x_1}, .., U_{x_n}$ is $\bar{R}$ which is invariant under rotation.

The mean resultant length has a minimization property, that is, $S(a)$ the arithmetic mean of the squared euclidean distance between x_i and a which attains its minimum at $a = \bar{x_0}$ and it is giving by

$$S(a) = \frac{1}{n}\sum_{i=1}^{n} \|x_i - a\|^2 = 2(1 - \bar{x}^T a) = 2(1 - \bar{R}\bar{x_0}^T a) \tag{1.1}$$

Sample spherical variance:

From Eqn. 5.1, when $S(a)$ is minimized at $a = \bar{x_0}$ (subject to the constraint $a^T a = 1$), we get the sample spherical variance

$$\min_a S(a) = 2(1 - \bar{R})$$

This quantity is a measure of clustering of data points around the mean direction in the sense that when $\bar{R} \simeq 0$, the data points $x_1, ..., x_n$ are widely dispersed. On the other hand, when $\bar{R} \simeq 1$, the observations $x_1, ..., x_n$ are highly concentrated.

Variance decomposition: The total variation can be decomposed as

$$2n(1 - \bar{C}) = \textstyle\sum_{i=1}^{n} \|x_i - \mu\|^2 = 2n(1 - \bar{R}) = 2n(\bar{R} - \bar{C}),$$

where μ is a unit vector and $\bar{C}$ is the sample mean of the components $x_1, ..., x_n$ along μ, such that $\bar{C}$ is given by

$$\bar{C} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} {x_i}^T \mu.$$

The moment of inertia: The spherical dispersion can be measured by the scatter matrix $\bar{T}$ about the origin and it is defined as

$$\bar{T} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} {x_i}^T x_i.$$

Let S be the sample variance matrix, that is giving by

$$S = \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n} (x_i - \bar{x_0})^T (x_i - \bar{x_0}).$$

Then, the spherical dispersion can be written as

$$\bar{T} = \tfrac{n-1}{n} S + \bar{x_0}^T \bar{x_0}.$$

If we put $\bar{x_0}^T \bar{x_0} = 1$, we get

$$tr(\bar{T}) = 1,$$

and thus

$$S = \frac{n-1}{n} tr(S) + \bar{R}^2. \tag{1.2}$$

Analogously, a connection between the population mean resultant vector ρ and the variance matrix Σ of x is defined as

$$tr(\Sigma) + \rho^2 = 1.$$

Additionally, since there is a connection between the mean $E[x]$ and the variance matrix Σ of the random vector x and since S is unbiased estimator of Σ, taking the expectation from Eqn. 1.2 yields

$$E[\bar{R}^2] = \rho^2 + \tfrac{1}{n}(1 - \rho^2).$$

1.3 Spherical models

The von Mises-Fisher distribution:

The von Mises-Fisher distribution (vMF) is the most common used distribution for spherical data in the field of directional statistics. The probability density function of the von Mises-Fisher distribution is given by

$$h(x; \mu, c) = \left(\frac{c}{n}\right)^{q/2-1} \frac{1}{\Gamma^{(q/2)} I_{q/2-1}(c)} \exp\{c\mu^T x\},$$

where $c \geq 0$, $||\mu|| = 1$, and I_v denotes the modified Bessel function (Mardia & Jupp, 2009) of the first kind and order v. The parameters μ and c are the mean direction and the concentration parameter, respectively.

The modified Bessel function I_q is defined as

$$I_q(c) = \frac{1}{2\pi} \int_0^{2\pi} \cos q\theta \exp^{c \cos\theta} d\theta.$$

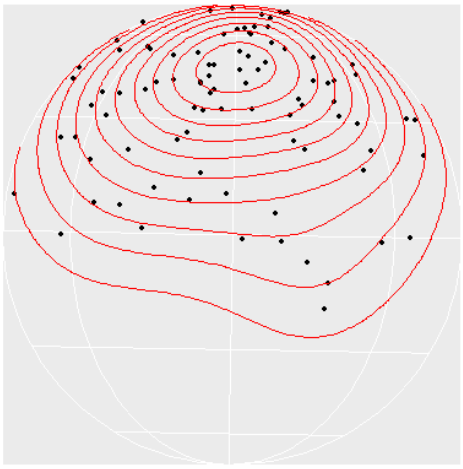


Figure 1.5: Von Mises-Fisher observations with longitude $\theta = 45°$, latitude $\varphi = 0°$ and concentration $c = 5$. The spherical density is drawn in red.

A random sample of spherical data from a von Mises-Fisher defined on S^2 is drawn in Figure 1.5 and the corresponding density is plotted in red.

The Uniform distribution:

The uniform distribution on the sphere S^{q-1} is the basic distribution. The von Mises-Fisher distribution reduces to the uniform distribution when the concentration $c = 0$.

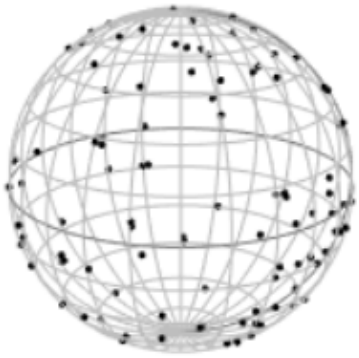


Figure 1.6: Spherical representation of simulated random data from a vMF((0,0,1)',0).

As much as the concentration parameter increases, as much as the data are concentrated on the sphere. In figure 1.6, data points are very sparse given that the concentration is equal to 0.

The Fisher-Bingham Distribution:

The Fisher-Bingham distribution is an important distribution on $S^{(q-1)}$ and it

can serve to generate a wide class of directional distributions (Kent et al., 2018). The probability density function of the Fisher-Bingham distribution can be expressed as

$$h_{FB} := \exp(c\mu_0^T x - x^T A x),$$

where μ_0, $c > 0$ are the mean direction and concentration parameters, respectively, and $A(q \times q)$ is a symmetric matrix. The smallest eigenvalue of A can be set equal to 0.

From a statistical point of view, the full family of Fisher-Bingham distributions is wide and too general. In practice, a special case of the aligned Fisher-Bingham distributions with unique mode at $x = \mu_0$ and for which μ_0 is an eigen vector of A can be considered.

In theory, the aligned models are easy to describe if the coordinate system is rotated so that $\mu_0 = (1, 0, .., 0)^T$ lies on the first coordinate axis and $A = diag(\lambda_1, ..., \lambda_d)$ is the diagonal. Some important special cases of unimodal aligned models include the the balanced five-parameter Fisher-Bingham (FB5b) and the extreme five-parameter Fisher-Bingham (FB5e) distributions.

The balanced five-parameter Fisher-Bingham (FB5b) distribution, that is also known as the kent distribution on S^2, is characterized by its matrix A where the eigenvalues $\lambda_1, ..., \lambda_j$ of A should be as follow: $\lambda_1 = 0$ and $\sum_{j=2}^{q} \lambda_j = 0$. On the other hand, the extreme five-parameter Fisher-Bingham (FB5e) is the case where $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 = \delta \geq 0$. An example of spherical observations generated from FB5b and FB5e distributions is depicted in Figure 1.7.

1.4 Supervised classification of directional data

The development of discriminant analysis methods for directional data has been a major research theme lately, particularly amongst the machine learning community. As for circular data, SenGupta & Roy (2005) introduced a chord-length based classification rule in order to classify circular observations. They compared their proposed methods with respect to the Fisher's rule based on the exact error probabilities and apparent error rates.

(a) (b)

Figure 1.7: Simulated data points from Fisher-Bingham on S^2. (a) Balanced FB5 distribution (kent), (b) extreme FB5 distribution. We reproduce the same plot designed in Kent et al. (2018).

Recently, Di Marzio et al. (2018) proposed a non-parametric classification method based on the kernel estimation of the population density and local logistic regression. They showed that the non parametric estimators are preferred against the classical parametric ones, especially in case of high complexity of the data scenario. Leguey et al. (2019) introduced a Bayesian classification rule for wrapped Cauchy circular predictors. They proved that circular classifiers improve over linear classifiers for the case of wrapped Cauchy distributions.

For spherical data, classification procedures for the Watson and the von Mises-Fisher distributions were proposed (Figueiredo & Gomes, 2006; Figueiredo, 2009). López-Cruz et al. (2015) extended the naive Bayes classifier for directional spherical von Mises-Fisher distributed predictors. They proposed various versions of naive Bayes and discussed conditions under which each method should be used.

Within the field of image textures classification, Kim & So (2018) considered a classification rule based on multi-resolution directional filters. Non parametric methods for supervised classification of spherical data, based on kernel density estimation, have been studied by Di Marzio et al. (2019).

More recently, Tsagris & Alenazi (2019) studied the use of maximum likelihood discriminant analysis function on the sphere considering different distributions. They

mainly considered rotational symmetric distributions as well as non-rotationally symmetric ones. They compared the performance of the discriminant rule based on the maximum likelihood with respect to the k-nearest neighbors classification rule and discussed under which conditions one should be preferred over the other.

Given the interest in this topic, this research work is then dedicated to some non-parametric supervised classification rules based on directional depth functions. In brief, non-parametric tools are preferred here because of their flexibiity, while depth functions are preferred because they do not need any additional parameter (like the bandwidth parameter) to be exploited.

Chapter 2

A Directional Depth Distribution Classifier based on the Cosine Depth

Abstract

Directions, rotations, axes, clock or calendar measurements can be represented as angles or equivalently as unit vectors. As points lying on the boundary of circles, spheres or hyper-spheres, they are also referred as directional data, and they require dedicated methods to be analyzed. In the framework of supervised classification, this work introduces a directional data classifier based on a data depth function. Depth functions provide an inner-outer ordering of the data in a reference space according to some centrality measure, and have appeared as a powerful tool in many fields of multivariate statistics. The recently introduced distance based depth functions for directional data are considered here. More specifically, this work introduces a cosine depth-based distribution method which aims at assigning directional data to classes, given that a training set with class labels is already available. A simulation study evaluating the performance of the proposed method and a real data example are provided.

Keywords: Max-depth classifier; Supervised classification; von Mises-Fisher.

This Chapter is based on the published paper: Demni, H., Messaoud, A. and Porzio, G.C. (2019) "The Cosine Depth Distribution Classifier for Directional Data". In: Ickstadt K., Trautmann H., Szepannek G., Lübke K., and Bauer N., (eds), *Applications in Statistical Computing. Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Cham, pp. 49-60.
Acknowledgments: Thanks are due to the Editors and the two anonymous referees of the book *Applications in Statistical Computing* for their comments and suggestions. Thanks are also due to Giuseppe Pandolfo and Alfonso Iodice D'Enza for the valuable discussions, the support and the suggestions. A part of this work has been presented at the Meeting "Statistics and Data Science: new Developments for Business and Industrial Applications (SDS 2018)" in Turin, Italy.

2.1 Introduction

Directional information arises whenever observations are recorded as directions in two or three dimensions, or as points lying on the surface of q-dimensional hyper-spheres. They can be recorded using either angles or unit vectors in the q-dimensional Euclidean space. A one-to-one mapping between these two representations exists, so that the Cartesian coordinates of the ending point of a unit vector can be transformed in spherical coordinates, and vice versa.

A two-dimensional observation (circular data) can be represented by an angle or as a point on the circumference of the unit circle centered at the origin, or as a unit vector connecting the origin to this point. Three-dimensional directional data (spherical data) can be depicted by two angles or as a unit vector in three dimensions, or as a point on the unit sphere, and so on.

Directional data can be found in many scientific areas in the context of measuring directions or cycles. For instance, they are used to estimate the relative rotations of tectonic plates in Earth sciences (Chang, 1993), to measure the electrical cardiological activity during a heartbeat in Medicine (Downs & Liebman, 1969), to study the animal navigation patterns in Biology (Batschelet, 1981), and to analyze wind and ocean directions in Metrology (Bowers et al., 2000). For a deeper introduction to directional data, the reader is referred to the book (Mardia & Jupp, 2009).

In directional data analysis, a special role is played by data depth functions. They characterize the centrality of a point with respect to a distribution or a sample so that a center-outward ordering of the points can be obtained. Depth functions are available for linear, functional, and directional data. For a review of directional depth functions, see the recent work by Pandolfo, Paindaveine, & Porzio (2018).

Depth functions have found interesting applications in supervised classification, where the aim is to assign new observations to labeled classes. Among the many proposals that exploit data depth functions, the most popular are probably the max-depth classifier (widely investigated in Ghosh & Chaudhuri (2005)), and the depth vs. depth (denoted by DD) classifier (Li et al., 2012). Unlike the many parametric and semi parametric classification methods, they neither assume any particular type of probability distribution, nor consider any specified parametric form for the separating surface.

In the directional data domain, both these popular depth-based classifiers have been developed, at least to a certain extent. The performance of directional max-depth classifiers

when different depth functions are adopted has been studied in Pandolfo, Paindaveine, & Porzio (2018), while the DD-classifier for circular data has been introduced in Pandolfo, D'Ambrosio, & Porzio (2018).

An additional depth-based classifier for linear data has been recently introduced by Makinde & Fasoranbaku (2018): the depth distribution classifier. The proposal seems promising given that it is optimal for a wider class of distributions if compared with the max-depth classifier. For this reason, the aim of this work is two-fold. First, it provides a review of the max-depth, the DD-classifier, and the depth distribution techniques. Then, it introduces a new supervised classification tool for directional data: the cosine depth distribution classifier.

Unlike the DD-classifier, the depth distribution classifier is naturally able to deal with multiclass classification issues. As a consequence, we compare its performance with others who can deal with multi classes. A simulation study is then offered to the reader where we compare the performance of the depth distribution classifier with respect to the max-depth classifier.

The chapter is organized as follows. Section 2.2 provides background on statistical depth functions and on depth-based classifiers. In Section 2.3, the proposed directional depth distribution classifier is introduced, while its performance is assessed through a simulation study in Section 2.4. Finally, some final remarks are offered in Section 2.5.

2.2 Depth-based classifiers for linear data

Generally speaking, the depth of a point measures to what extent that point is inner with respect to a given distribution or to a multivariate sample. The most internal point is called the deepest, and it is considered a measure of centrality. More precisely, by definition of a depth function (Liu et al., 1999), if a given distribution has a symmetry point, this point should be the deepest above all of that distribution, and a depth will decrease whenever the distance from the symmetry point increases.

Formally, we have that a depth function $D(.) : \mathbb{R}^q \to \mathbb{R}$ with respect to a distribution F is a mapping of a vector $x \in \mathbb{R}^q$ to a real-valued number $D(x; F) \in \mathbb{R}$. Within the literature, several depth functions have been introduced. Amongst the many, Tukey's half space depth (Tukey, 1975), the simplicial depth (Liu, 1990) and the zonoid depth (Koshevoy & Mosler, 1997) have reached some popularity.

The data depth concept provides center-outward ordering of points in any dimension

and it allows some non-parametric multivariate statistical analysis to be performed, in which no distributional assumptions are needed. That is, the distribution F in the expression $D(x;F)$ is typically substituted by its empirical counterpart $\hat{F}$, with no needs to assume a parametric form for it. Applications of data depth arise in statistical inference (location and scatter estimation Romanazzi, 2009), (statistical quality control Messaoud et al., 2008), (outlier detection and data visualization Rousseeuw et al., 1999; Buttarazzi et al., 2018).

The same concept of data depth could be applied to directional data. Within this framework, Figure 2.1 shows a random sample of directional observations on the sphere ($q = 3$) where the blue point is more central within the data cloud than the red point. Thus, the depth value of the blue point should be greater than the one of the red point.

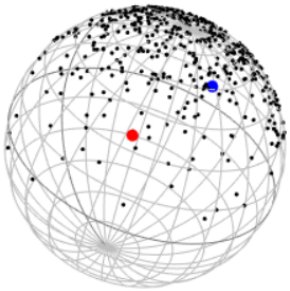


Figure 2.1: Graphical representation of a random sample of spherical data. The blue point is more central than the red point within the group of black points.

Depth functions have been also employed in supervised learning, where a classification rule is constructed from labeled training data to assign an arbitrary new data point to one of the labels. The main underlying idea is that the centrality of a new point with respect to the labeled classes (i.e., its depth) is a measure of the degree of closeness to each label.

Dissimilar to parametric and semi parametric classification methods, the depth-based classifiers neither assume any particular type of probability distribution nor consider any specified parametric form for the separating surface. Generally speaking, any depth function can be adopted to define a depth-based classifier.

In this section, the two main depth-based classifiers are reviewed. Namely, the max-depth classifier and the DD-classifier. Other depth-based classifiers have been proposed for linear data, such as the $DD-\alpha$ classification approach (Lange et al., 2014), and the class of depth-based functions associated with the $k$nn classification rule introduced by Paindaveine & Van Bever (2015). A recent comprehensive overview of depth-based classifiers can be found in Vencálek (2017). More recently, a depth distribution classifier was introduced by Makinde

& Fasoranbaku (2018). Given the focus of this work, it will be also briefly described in this section.

2.2.1 The max-depth classifier

The max depth classifier assigns the new data point to the class with respect to which it attains the highest depth value. This is because higher depth values correspond to more central areas with respect to the class. After Liu (1990), the concept of max-depth in classification has been developed by Ghosh & Chaudhuri (2005).

Let y be the new point to be assigned, and let $D(y, \hat{F})$ be the depth of y with respect to the empirical distribution $\hat{F}$. For the sake of simplicity, let us consider the case of two groups (i.e., we have available $\hat{F}_1$ and $\hat{F}_2$ from a training data set). Then, the max-depth classification rule is given by

$$\begin{cases} D(y, \hat{F}_1) > D(y, \hat{F}_2) \Longrightarrow \text{assign } \ y \ \text{ to population 1} \\ D(y, \hat{F}_1) < D(y, \hat{F}_2) \Longrightarrow \text{assign } \ y \ \text{ to population 2} \end{cases}$$

If $D(y, \hat{F}_1) = D(y, \hat{F}_2)$, the classification rule will randomly assign the observation to one of the two groups with equal probability.

The max-depth classifier is equivalent to the optimal Bayes classifier with equal prior probabilities in case all the populations are elliptically distributed with density function strictly decreasing when moving away from the ellipsoid center, and with populations differing only in location parameters (Ghosh & Chaudhuri, 2005). For this condition to hold, the adopted depth functions must be continuous, positive over the entire q-dimensional space and decreasing too.

When the populations differ not only in location, a modified version of the classification approach based on a function of the half-space depth has been proposed (Ghosh & Chaudhuri, 2005), where the empirical half space depth of a point y with respect to a multivariate sample is given by the minimum number of points that lie in any closed half-space containing y Tukey (1975). A modified version of the max-depth classifier based on the projection depth function has been introduced as well (Cui et al., 2008). It outperforms the method by Ghosh & Chaudhuri (2005) only in normal settings.

Finally, the use of the max-depth classifier for directional data has also been studied (Pandolfo, Paindaveine, & Porzio, 2018). By means of a simulation study, it has been shown that the distance based depth classifiers outperform classifiers based on the angular Tukey's

(Liu & Singh, 1992) and on the angular simplicial depth (Liu & Singh, 1992) if data are drawn from a von Mises-Fisher distribution (Mardia & Jupp, 2009), either with equal or different concentration levels.

2.2.2 The DD-classifier

The depth vs. depth classifier (or DD-classifier) is a non parametric two-class classification method introduced by Li et al. (2012). It is based on the depth vs. depth (or DD)-plot, which is a graphical tool allowing the comparison of two multivariate distributions or samples through their corresponding depth values.

Briefly, the DD-plot is a scatterplot where each plotted point has coordinates given by the depths of the corresponding point in the original multivariate space with respect to the two examined groups. In this way, it is possible to transform two multivariate samples to a simple two-dimensional scatter plot regardless of the dimensions of the original sample space.

The main idea behind the DD-classifier is to find the best polynomial separating function in a DD-plot. Consequently, the generic form of the DD-classifier is given as follows. Let $r(.)$ be some real increasing function, and $\hat{F}_1$ and $\hat{F}_2$ be the empirical cdf's of two multivariate samples (the two samples are the training set, where each of the two sample has its own class label). Then, the classification rule is defined by

$$\begin{cases} D(y, \hat{F}_1) > r(D(y, \hat{F}_2)) \Longrightarrow \text{assign } \ y \ \text{ to } \ \text{population 1} \\ D(y, \hat{F}_1) < r(D(y, \hat{F}_2)) \Longrightarrow \text{assign } \ y \ \text{ to } \ \text{population 2} \end{cases}$$

In case of equality, y will be randomly classified to group 1 or 2 with equal probability.

If $r(.)$ is set equal to the 45 degree line, and apart from the case of equality, the DD-classification rule would assign y to group 1 if $D(y, \hat{F}_1) > D(y, \hat{F}_2)$ and assign y to group 2 otherwise. This will reduce the DD classifier to the max-depth classifier described above. If F_1 and F_2 differ and they both admit a density from the elliptically contoured family, then the DD-classifier will be optimal in the Bayes sense whenever the used depths are strictly increasing functions of the densities themselves.

The performance of the DD-classifier associated with different depth functions (Mahalanobis depth Mahalanobis, 1936), (projection depth Y. Zuo, 2003), (half-space depth Tukey, 1975), and (simplicial depth Liu, 1990) has been compared by Li et al. (2012) with the performance of other classifiers such as the Linear Discriminant Analysis (James et al., 2013), the Quadratic Discriminant Analysis (James et al., 2013), the Support Vector Machine (Vencálek,

2017), and the max-depth (Ghosh & Chaudhuri, 2005) classifier. It seems the DD-classifier shows a better performance in many of the cases, or a similar performance otherwise.

For this reason, a recent interest eventually arised on the use of the DD-classifier for directional data. It has been investigated for the case of circular data by Pandolfo, D'Ambrosio, & Porzio (2018).

2.2.3 The depth distribution classifier

Based on the cumulative distribution function (cfd) of depths, a new depth-based classifier has been very recently introduced Makinde & Fasoranbaku (2018). That is, the depth distribution classifier.

Let $F_D^G(x)$ be the cumulative distribution function of a depth function $D(X, G)$ evaluated in x:

$$F_D^G(x) := P(D(X, G) \leq D(x, G))$$

where X is a random variable, and G is a generic distribution with respect to which the depth is evaluated. Both X and G are defined on the original sample space.

The value of $F_D^G(x)$ provides information on how central is x with respect to G. If x is a central observation, then $D(x, G)$ will be large, and hence $F_D^G(x)$ will be large too. At the extreme, if x_0 is a deepest point of G, we will have $F_D^G(x_0) = 1$. On the other hand, if x is far from the center of G (from its deepest point), then $F_D^G(x)$ will be small.

Accordingly, a depth distribution classifier can be defined Makinde & Fasoranbaku (2018). Let y be the new point to be assigned. And, for the sake of simplicity, let us consider the case of two groups G_1 and G_2. Then, the depth distribution classification rule is given by

$$\begin{cases} F_D^{\hat{G}_1}(y) > F_D^{\hat{G}_2}(y) \Longrightarrow \text{assign } \ y \ \text{ to } \ \text{population 1} \\ F_D^{\hat{G}_1}(y) < F_D^{\hat{G}_2}(y) \Longrightarrow \text{assign } \ y \ \text{ to } \ \text{population 2} \end{cases}$$

If $F_D^{\hat{G}_1}(y) = F_D^{\hat{G}_2}(y)$, the classification rule will randomly assign the observation to one of the two groups with equal probability.

2.3 The cosine depth distribution classifier for directional data

The cosine depth distribution classifier is proposed here as a tool to classify points lying on the surface of hyper-spheres, in analogy with the work by Makinde & Fasoranbaku (2018).

Directions in q-dimensions can be represented as unit vectors z on the sphere $S^{(q-1)} = \{z : z^T z = 1\}$ with unit radius and center at the origin. Let $H_1, ..., H_J$ be a set of directional distributions, and let $F_D^H(z)$ be the cumulative distribution function of a depth function defined on hyper-spheres $D(Z, H)$ evaluated in z:

$$F_D^H(z) := P(D(Z, H) \leq D(z, H))$$

Let w be the new point to be assigned, and, again for the sake of simplicity, let us consider the case of two groups (i.e., G_1 and G_2). Then, the directional depth distribution classification rule will be given by

$$\begin{cases} F_D^{\hat{G}_1}(w) > F_D^{\hat{G}_2}(w) \Longrightarrow \text{assign } w \text{ to population 1} \\ F_D^{\hat{G}_1}(w) < F_D^{\hat{G}_2}(w) \Longrightarrow \text{assign } w \text{ to population 2} \end{cases}$$

If $F_D^{\hat{G}_1}(w) = F_D^{\hat{G}_2}(w)$, the classification rule will randomly assign the observation to one of the two groups with equal probability.

The performance of the just introduced directional depth distribution classifier depends on the choice of the depth function. Many depths for directional data were introduced and are available in the literature Pandolfo, Paindaveine, & Porzio (2018). Here, distance based directional depths will be considered. They are briefly reviewed below.

Let $d()$ be a bounded and nonnegative directional distance function with $d^{\sup}$ its supremum over $S^{(q-1)}$. By definition, a directional distance based depth of a point $z \in S^{(q-1)}$ with respect to a distribution H on $S^{(q-1)}$ is given by

$$D(z, H) := d^{\sup} - E_H[d(z, W)],$$

where $E[.]$ is the expected value, and W is a random variable from a distribution H.

To obtain a directional depth function enjoying nice properties, suitable distances should be adopted. Particularly, they must be rotation invariant. As a consequence, they will be of

the form $\delta(z's)$ for some function $\delta = [-1, 1] \to \mathbb{R}^+$ where s is also a point on $S^{(q-1)}$. For instance, three of these distances that will yield rotational invariant directional depths are discussed in Pandolfo, Paindaveine, & Porzio (2018), and briefly reported below.

- **Cosine depth:** Adopting the cosine distance, i.e. $\delta(t) = 1 - t$, the cosine depth is obtained as $D_{cos} := 2 - E_H[(1 - z'W)]$.
- **Arc distance depth:** Adopting the arc distance, i.e. $\delta(t) = \arccos(t)$, the arc distance depth is obtained as $D_{arc} := \pi - E_H[\arccos(z'W)]$
- **Chord depth:** Adopting the chord distance, i.e. $d_{chord} = ||z - t|| = \sqrt{2(1 - z't)}$, the chord depth is obtained as $D_{chord} := 2 - E_H[\sqrt{2(1 - z'W)}]$

Finally, amongst these three directional depth functions, the cosine depth is preferred here. This is because of two reasons. First, it can be easily computed. Then, it provided good performances when associated to the max-depth classifier on hyper-spheres Pandolfo, Paindaveine, & Porzio (2018). This yields the cosine depth distribution classifier.

2.4 Simulation Study

The performance of the cosine depth distribution classifier is evaluated by means of a simulation study. A comparison with the max-depth classifier for directional data based on the same depth function is also offered to the reader.

2.4.1 Study design

This study is based on the assumption of equal prior probabilities and considering a two class classification problem.

Let G_1 and G_2 be two von Mises-Fisher distributions (vMF). That is, G_1 and G_2 have their probability density function given by

$$h(z; \mu, c) = \left(\frac{c}{n}\right)^{q/2-1} \frac{1}{\Gamma^{(q/2)} I_{q/2-1}(c)} \exp\{c\mu^T z\},$$

where $c \geq 0$, $||\mu|| = 1$, and I_v denotes the modified Bessel function Mardia & Jupp (2009) of the first kind and order v. The parameters μ and c are the mean direction and the concentration parameter, respectively.

For the sake of comparison, the simulation scheme was designed in analogy with the simulations in Pandolfo, Paindaveine, & Porzio (2018), where the performance of the directional max-depth classifier was investigated with respect to the choice of the depth function.

Within the first two setups, the two groups are both unimodal with different locations parameters and same/different concentration levels (setup 1 and 2, respectively). On the contrary, the third setup will investigate the case of bimodality for one of the two groups. The other will be unimodal with its mode lying between the modes of the first group.

- In Setup 1, we study the case of difference in location, same concentration parameter for both G_1 and G_2, data on the sphere ($q = 3$) and on a hyper-sphere in dimension $q = 10$. The location parameters are set $\mu_1 = (0, 0, 1)$ and $\mu_2 = (1, 0, 0)$ for dimension $q = 3$, and $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$ and $\mu_2 = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ for dimension $q = 10$, respectively. The performance is then investigated for different concentration levels. We set $c \in \{2, 5\}$.

- In Setup 2, the two distributions differ also in concentration. With the same location parameters of Setup 1, the concentration c is set equal to 2 for group 1, and 5 for group 2, considering again dimensions $q \in \{3, 10\}$.

- In Setup 3 we consider discrimination between a vMF distribution with $\mu = \mu_1$, and a bimodal density obtained as an equal weight mixture of two von Mises-Fisher, with means μ_{21} and μ_{22}. For dimension $d = 3$, we set $\mu_1 = (0, 0, 1)$, $\mu_{21} = (1, 0, 0)$, $\mu_{22} = (1, 0, 1)$. For dimension $d = 10$, we set $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, \cos 7\pi/4, \sin 7\pi/4)$, $\mu_{21} = (0, 0, 0, 0, 0, 0, 0, 0, 1, 0)$ and $\mu_{22} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$. All the vMF's have the same concentration levels $c = 4$.

As in Pandolfo, Paindaveine, & Porzio (2018), we set the training set size equal to 200 with 100 observations generated from G_1 and 100 observations generated from G_2 and the testing set size equal to 100 with 50 observations generated from G_1 and 50 observations generated from G_2. For each simulation condition, the experiment is replicated 250 times.

Setup			AMR (*standard deviation*)	
			DistD	*MaxD*
Setup 1	$q = 3$	$c = 2$	**0.236** *(0.037)*	0.257 *(0.045)*
		$c = 5$	**0.066** *(0.028)*	0.074 *(0.022)*
	$q = 10$	$c = 2$	**0.375** *(0.052)*	0.380 *(0.052)*
		$c = 5$	**0.150** *(0.030)*	0.167 *(0.037)*
Setup 2	$q = 3$		**0.104** *(0.022)*	0.170 *(0.030)*
	$q = 10$		**0.238** *(0.045)*	0.274 *(0.037)*
Setup 3	$q = 3$		0.496 *(0.022)*	**0.440** *(0.030)*
	$q = 10$		**0.170** *(0.037)*	0.185 *(0.037)*

Table 2.1: Average misclassification rate (AMR) and standard deviations of the depth distribution (DistD) and the max-depth (MaxD) classifiers in different simulation setups. Best achieved results are highlighted in bold.

2.4.2 Results

The detailed result of our simulation studies are reported in this Section. The performance of the classifiers is evaluated by means of the misclassification rate. That is, the number of observations misclassified over the sample size in each replicated sample.

For each simulation setup, the distribution of the misclassification rates obtained by the cosine depth distribution classifier (DistD) and by the max-depth classifier (MaxD) are summarized through boxplots (Figures 2.2-2.5). The corresponding average misclassification rates (AMR) and standard deviations are instead reported in Table 2.1, where the best achieved results are highlighted in bold.

Considering Setup 1 in dimension 3 (results in Figure 2.2), the cosine depth distribution classifier achieved a slightly better performance than the max-depth classifier for both scenarios of concentration parameters, i.e., $c = 2$ and $c = 5$. On the other hand, the misclassification rate is lower when the concentration parameter is higher, and this is because data are more separated and less sparse on the sphere, and hence they can be better discriminated.

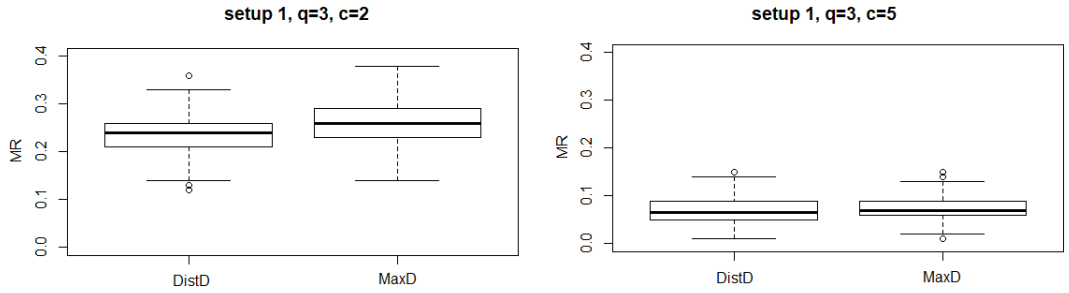


Figure 2.2: Boxplots of the misclassification rates (MR) of the depth distribution (DistD) and the max-depth (MaxD) classifiers obtained from 250 independent replications in Setup 1 with concentration parameters $c = 2$ and $c = 5$ in dimension $q = 3$.

For $q = 10$, results from Setup 1 indicate that the cosine depth distribution classifier performs better than the max-depth classifier, in case of equal concentration, also in higher dimensions (Figure 2.3). However, the overall performance deteriorates, especially for quite sparse data ($c = 2$).

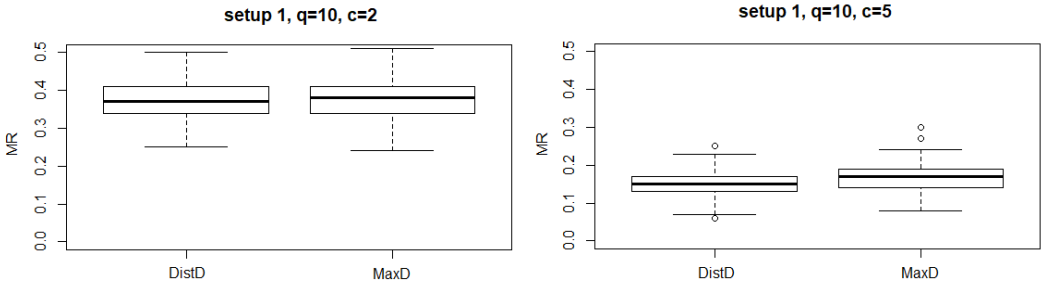


Figure 2.3: Boxplots of the misclassification rates (MR) of the depth distribution (DistD) and the max-depth (MaxD) classifiers obtained from 250 independent replications in Setup 1 with concentration parameters 2 or 5 for dimension q=10.

In the case of different concentration levels across the two groups (Setup 2), the cosine depth distribution classifier shows highly satisfactory performance, with a much lower misclassification rate compared to the max-depth classifier, especially in lower dimension (Figure 2.4).

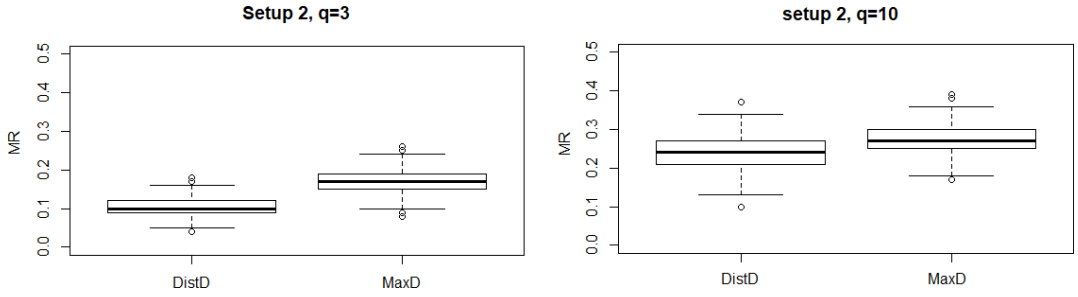


Figure 2.4: Boxplots of the misclassification rates (MR) of the depth distribution (DistD) and the max-depth (MaxD) classifiers obtained from 250 independent replications in Setup 2 for dimensions $q \in \{3, 10\}$.

In the third setup, the case of the bimodal group, if both the dimension and the concentration level are low ($q = 3$, $c = 4$), the two classifiers essentially fail (Figure 2.5, left). Although the max-depth classifier slightly outperform the new introduced method, both their average misclassification rates approximate the 50% rate, which can be attained by just assigning randomly each new observations to one of the two groups. This happens because data from the two groups are largely mixed up on the sphere. When the dimension increases ($q = 10$), data are not largely mixed up any more, the two classifiers perform better, with the cosine distribution depth slightly beating the max-depth classifier (Figure 2.5, right).

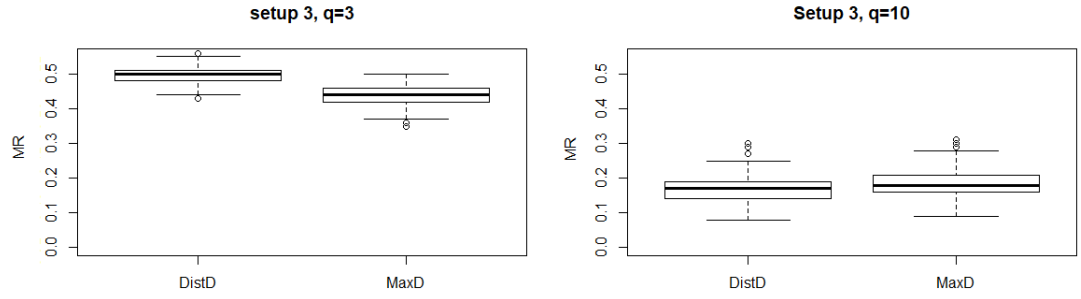


Figure 2.5: Boxplots of the misclassification rates (MR) of the depth distribution (DistD) and the max-depth (MaxD) classifiers obtained from 250 independent replications in Setup 3 for dimension $q \in \{3, 10\}$.

2.5 Concluding Remarks

This chapter first reviews supervised classification methods based on data depth, and then it introduces a procedure to classify directional data. Directional data are a special class of

quantitative measures which requires dedicated methods to be analyzed properly. They refer to point lying on the surface of hyper-spheres.

The proposed classifier is based on the distribution function of the cosine depth, and for this reason it is called cosine depth distribution classifier. The performance of the proposed classification method is investigated in lower and higher dimension settings with a comparison to the max-depth classifier through simulations. The simulation results suggest that the cosine depth distribution classifier might improve over the max-depth classifier in many scenarios.

The use of the cosine depth classifier for directional data seems thus promising. As a further study, it would be of interest to investigate to what extent, and in which cases, such a classifier would provide better performances if different depth functions are adopted in place of the cosine depth. Some real data applications would also be of interest.

Chapter 3

On the optimality of the max-depth and max-rank classifiers for spherical data

Abstract

The main goal of supervised learning is to construct a function from labeled training data which assigns arbitrary new data points to one of the labels. Classification tasks may be solved by using some measures of data point centrality with respect to the labeled groups considered. Such a measure of centrality is called data depth. In this paper, we investigate conditions under which depth-based classifiers for directional data are optimal. We show that such classifiers are equivalent to the Bayes (optimal) classifier when the considered distributions are rotationally symmetric, unimodal, differ only in location and have equal priors. The necessity of such assumptions is also discussed.

Keywords: Depth-based classifier; von Mises-Fisher distribution; Directional data; Cosine depth

---

This Chapter has been published as: Vencalek, O., Demni, H., Messaoud, A. and Porzio, G.C. (2020) "On the optimality of the max-depth and max-rank classifiers for spherical data", *Applications of Mathematics*, 65(3), 331-342.

Acknowledgments: This work was partially developed while Houyem Demni was visiting the Department of Mathematical Analysis and Applications of Mathematics, Palacký University in Olomouc. Thanks are due to the University of Cassino and Southern Lazio for supporting the Mobility. Grateful acknowledgment is made to Ondrej Vencalek for his help, his suggestions and the very constructive discussions. Thanks are also due to the Editor and the two anonymous referees of the Journal *Applications of Mathematics* for their comments and suggestions.

3.1 Introduction

Supervised classification techniques enjoy a wide range of applications in many fields. Given a training set of observations and their membership of certain groups, new observations with unknown membership should be accordingly assigned. A fairly large number of classification rules are available in the literature (e.g James et al., 2013).

Within this setting, depth-based classification procedures have been introduced. Depths provide center-outward ordering of points in multidimensional spaces with respect to a given distribution, and their applications often lead to effective robust statistical procedures. As a consequence, depth-based supervised classification techniques are typically able to deal with the presence of outliers or mislabeled observations in the training set (Hubert et al., 2017). Many depth based classifiers are available, and for a review we refer to the work of Vencálek (2017). On the other hand, depth-based supervised classification procedures have only recently been introduced in the directional data framework (Demni et al., 2019; Pandolfo, D'Ambrosio, & Porzio, 2018; Pandolfo, Paindaveine, & Porzio, 2018).

Spherical (or directional) data are data lying on the unit hyper-sphere. They occur naturally when a direction or an angle in space is of interest (e.g. wind direction), but also when data consist of time points and the interest is in cycles (time points on a watch can be treated as angles). In higher dimensions, locations on the Earth and/or any kind of information recordable as unit vectors can be analyzed from a directional data perspective.

Such data can be encountered in many fields of science and technology such as Earth sciences (Chang, 1993; N. Fisher, 1989), meteorology (Bowers et al., 2000), neurosciences (Leong & Carlile, 1998) or biology (Batschelet, 1981) to capture the direction of some phenomena of interest. Other interesting applications of directional data include shape analysis and its use in economics (Klecha et al., 2018; Kosiorowski, 2007).

Spherical data have their own specific features and therefore classical statistical methods need to be adjusted to these kinds of data. In this context, depths have been successfully applied (Agostinelli & Romanazzi, 2013; Ley et al., 2014; Liu & Singh, 1992; Pandolfo, Paindaveine, & Porzio, 2018) and some robustness aspects have also been investigated. For instance, Pandolfo, Paindaveine, & Porzio (2018) showed that the cosine depth deepest point achieves the highest directional breakdown point in terms of lower bound when compared to the chord and arc distance depth deepest points in the case of von Mises-Fisher distributions. A discussion on the finite-sample maximum bias of the cosine depth deepest point (the spher-

ical mean) and the arc distance deepest point (the spherical median) is instead available in Kirschstein et al. (2019).

However, although the recently introduced depth-based classifiers for directional data performed well in simulation studies (Demni et al., 2019; Pandolfo, D'Ambrosio, & Porzio, 2018; Pandolfo, Paindaveine, & Porzio, 2018), corresponding theoretical results are still lacking. For all the above reasons, this work investigates properties of depth-based classifiers for directional data. It introduces the conditions under which these classifiers are optimal. That is, they are equivalent to the Bayes classifier – the classifier with the lowest achievable probability of misclassification. Special attention is paid to the case of von Mises-Fisher distributions, since they play a central role among models for directional data.

The chapter is organized as follows. Section 3.2 introduces some basic concepts of directional data and the directional distance-based depth functions. Furthermore, it describes the max-depth and the depth distribution classifiers for spherical data. Section 3.3 includes the main results. It provides a discussion about the assumptions under which the depth-based classifiers are optimal as well as the necessity of such assumptions. Final comments are provided in Section 3.4.

3.2 Background material

This section reviews basic concepts of directional data and their corresponding depth measures. Furthermore, it introduces directional data depth based classifiers.

3.2.1 Directional data

In q-dimensional space, directions can be depicted as points on the sphere $S^{q-1} = \{x \in \mathbb{R}^q \colon x'x = 1\}$ or as vectors with unit radius and center at the origin. Note that in the two-dimensional case, any direction can be also described by an angle (circular data). In the three-dimensional case, data points can also be described by two angles corresponding to longitude and latitude.

As highlighted in Chapter 1, the basic location parameter of spherical data is the mean direction $\mu = \frac{\mathrm{E}X}{\|\mathrm{E}X\|}$ (defined iff the value in the denominator is positive). A possible measure of variability, denoted traditionally as ρ, is the mean resultant length and is defined as $\rho = \|\mathrm{E}X\| = (\mathrm{E}X'\mathrm{E}X)^{1/2}$.

In this chapter, the class of rotationally symmetric distributions is considered. The

distribution H of a random variable X is said to be rotationally symmetric about some vector $\mu \in S^{q-1}$ iff the distribution of OX is again H for all $q \times q$ orthogonal matrices O satisfying $O\mu = \mu$. This class of distributions was first studied by Saw (1978). Any distribution which is rotationally symmetric about μ and absolutely continuous w.r.t. a surface area measure on S^{q-1} has a density of the form $h(x) = g(x'\mu)$ for some (univariate) function $g\colon [-1,1] \to \mathbb{R}_0^+$, e.g. Paindaveine & Verdebout (2015).

The most widely used distribution on the sphere is the von Mises-Fisher distribution, which is also rotationally symmetric (e.g Ley et al., 2014). Here, we recall the probability density function of the von Mises-Fisher distribution which is defined as

$$h(x;\mu,\kappa) = c_{\kappa,q} \exp\{\kappa\mu' x\},$$

where μ is the mean direction, $\kappa \geq 0$ is a concentration parameter, and $c_{\kappa,q} > 0$ is a normalizing constant (depending on parameters κ and q). Its value is $c_{\kappa,q} = \left(\frac{\kappa}{2}\right)^{q/2-1} \frac{1}{\Gamma(q/2) I_{q/2-1}(\kappa)}$, where I_v is the modified Bessel function of the first kind and order v (e.g Mardia & Jupp, 2009).

3.2.2 Data depth for directional data

The concept of data depth for directional data was first introduced by Small (1987). Later, it was extended by Liu & Singh (1992). They introduced the arc distance depth and at the same time extended the simplicial depth (originally introduced in Liu (1990)) to the directional angular simplicial depth and the halfspace depth (originally introduced in Tukey (1975)) to the directional angular Tukey depth.

In this work, the class of depth based on rotational invariant distance is considered. It was introduced by Pandolfo, Paindaveine, & Porzio (2018).

- The directional distance-based depth is recalled here:

  Let $d\colon S^{q-1} \times S^{q-1} \to \mathbb{R}_0^+$ be a bounded distance on the sphere S^{q-1}. Let H be a probability distribution on S^{q-1}. The directional d-depth of a point $x \in S^{q-1}$ with respect to the distribution H is defined as

  $$D(x,H) = d^{sup} - \mathrm{E}_H\left(d(x,X)\right), \tag{3.1}$$

  where d^{sup} is the upper bound of the distance between any two points on S^{q-1}, E_H is

the expected value and X is a random directional variable from H.

- Rotational invariance is an important property of distance and subsequently depth.

  A distance d is rotationally-invariant if $d(Ox, Oy) = d(x, y)$ for all $x, y \in S^{q-1}$ and all $q \times q$ orthogonal matrices O.

  Any rotationally-invariant (bivariate) distance $d(x, y)$ can be expressed as a univariate function δ of the scalar product $x'y$, i.e.

  $$d(x, y) = \delta(x'y), \tag{3.2}$$

  as shown in Pandolfo, Paindaveine, & Porzio (2018) (Proposition 1). It is easy to see that any directional depth based on rotationally-invariant distance is also rotationally-invariant, i.e. $D(x, H) = D(Ox, H_O)$ for all $x \in S^{q-1}$ for all $q \times q$ orthogonal matrices O, where H_O denotes distribution of OX when X has distribution H. See also Theorem 1 in Pandolfo, Paindaveine, & Porzio (2018).

- Let us now recall the three most widely used rotationally-invariant distance-based depth functions: the cosine depth, the arc distance depth, and the chord depth.

  - The cosine depth of a point $x \in S^{q-1}$ w.r.t. the distribution H of a random directional variable X is defined as $D_{cos}(x, H) := 2 - \mathrm{E}_H\left[(1 - x'X)\right] = 1 + \mathrm{E}_H(x'X)$ using the cosine distance $\delta(t) = 1 - t$.
  - The arc distance depth of a point $x \in S^{q-1}$ w.r.t. the distribution H of a random directional variable X is defined as $D_{arc}(x, H) := \pi - \mathrm{E}_H\left[\arccos\left(x'X\right)\right]$ using the arc distance $\delta(t) = \arccos(t)$.
  - The chord depth of a point $x \in S^{q-1}$ w.r.t. the distribution H of a random directional variable X is defined as $D_{chord}(x, H) := 2 - \mathrm{E}_H\left[\sqrt{2(1 - x'X)}\right]$ using the chord distance $\delta(t) = \sqrt{2(1-t)}$.

3.2.3 Max-depth and depth distribution classifiers for directional data

This section introduces the max-depth and the max rank classifiers. The above classifiers can be associated with all the available depth functions for directional data within the literature. In this study, the cosine depth is preferred for the following reasons. First, the cosine

depth does not require a large computational effort, unlike the other depths. Secondly, both classifiers provide good performance when associated with the cosine depth on hyper-spheres (Demni et al., 2019; Pandolfo, Paindaveine, & Porzio, 2018). Finally, the cosine depth (deepest point) can be considered a robust location estimator (Pandolfo, Paindaveine, & Porzio, 2018).

Consider now K different distributions $H_1, \ldots, H_K$ on hyper-sphere S^{q-1}. A classification rule in the directional framework is a function

$$c\colon S^{q-1} \to \{1, \ldots, K\},$$

which assigns points on the hyper-sphere to distributions from which they are likely to come. Here we restrict our attention to the two-class problem ($K = 2$).

Directional max-depth classifier

The concept of max-depth classifier for multivariate data was developed by Ghosh & Chaudhuri (2005). More recently, Pandolfo, D'Ambrosio, & Porzio (2018) extended the max depth classifier to the directional framework.

Let x be the new observation to be classified, and let $D(x, H_i)$, $i = 1, \ldots, K$ be the depth of x with respect to the distributions $H_1, \ldots, H_K$, respectively. The max depth classification rule is then given by

$$c_m(x) = \underset{i}{\operatorname{argmax}}\, D(x; H_i) \tag{3.3}$$

In practice, theoretical distributions are unknown and need to be estimated. Therefore, one uses empirical distribution functions $\widehat{H}_i$ based on data points in the training set instead of theoretical distributions H_i.

Directional depth distribution classifier

The depth distribution classifier known also as the max rank classifier was introduced by Makinde & Fasoranbaku (2018) for multivariate data and then extended to directional data by Demni et al. (2019).

The cumulative distribution function of the depth function $D(\cdot, H)$, denoted as $F_D(\cdot, H)$

is defined as

$$F_D(x, H) = \mathrm{P}\left(D(X, H) \leq D(x, H)\right), \tag{3.4}$$

where X is a random directional variable from the distribution H.

The directional depth distribution classification rule Demni et al. (2019) is then defined as

$$c_{edd}(x) = \underset{i}{\operatorname{argmax}}\, F_D(x, H_i).$$

In practice, the unknown distributions H_i are again replaced by their corresponding empirical distributions based on training set observations.

3.3 Properties of the max-depth and the depth distribution classifiers

The properties of the max-depth and max-rank classifiers are studied in this section. To the best of our knowledge, the optimality property of the depth-based classifiers has not been investigated elsewhere in the context of directional data.

The optimality of the considered depth-based classifiers was studied by Ghosh & Chaudhuri (2005) (the max-depth classifier) and by Makinde & Fasoranbaku (2018) (the max-rank classifier) in the context of multivariate (unconstrained) data. Both classifiers were shown to be equivalent to the optimal Bayes classifier (the classifier with the lowest total probability of misclassification) in some situations. More precisely, optimality is achieved if the considered distributions are elliptically symmetric with density strictly decreasing from the center (which implies unimodality of the distributions), differing only in location and having equal prior probabilities. While the assumptions on symmetry and unimodality are not too restrictive in practice, the assumptions on equal dispersions (implied by difference only in location) and equal priors reduce the applicability of the classifiers in practice quite substantially. We show that similar assumptions are needed for optimality also in the case of directional data.

Theorem 3.3.1. *Let H_1 and H_2 be rotational symmetric unimodal continuous distributions on the sphere S^{q-1} differing only in their mean directions (denoted by μ_1 and μ_2, respectively), i.e. their densities $h_i()$, $i = 1, 2$ can be expressed as $h_i(x) = h(\mu_i' x)$, $i = 1, 2$ for all $x \in$*

S^{q-1}, where $h()$ is some strictly increasing function. Let the distributions have equal prior probabilities $p_1 = p_2$. Then for any rotation-invariant distance-based depth, both the max-depth classifier and the max-rank classifier are equivalent to the (optimal) Bayes classifier.

Proof. First, we simplify the form of the Bayes classifier in the considered settings. The Bayes classification rule assigns x to group 1 iff

$$p_1 h_1(x) > p_2 h_2(x).$$

In the case of equal priors and rotational symmetric distributions the inequality simplifies to $h(\mu_1' x) > h(\mu_2' x)$. Since $h()$ is a strictly increasing function, the inequality can be rewritten as

$$\mu_1' x > \mu_2' x.$$

Now we show that the max-depth classifier can be expressed in the same way. This results directly from Theorem 3 of Pandolfo, Paindaveine, & Porzio (2018) who showed that in the considered situation depth can be expressed as a strictly increasing function of the cosine distance from the mean direction, i.e. $D(x, H_i) = \phi(\mu_i' x), i = 1, 2$, for some strictly increasing function $\phi\colon [-1, 1] \to \mathbb{R}_0^+$. Since the function ϕ is the same for both distributions, the inequality $D(x, H_1) > D(x, H_2)$ holds iff

$$\mu_1' x > \mu_2' x.$$

Finally, we deal with the max rank classifier. For the distribution H_i, $i = 1, 2$, the cumulative distribution function of depth (3.4) can be expressed as

$$F_D(x, H_i) = \mathrm{P}\left(D(X, H_i) \leq D(x, H_i)\right) = \int_{S(x)} h_i(y)\, dy = \int_{S(x)} h(\mu_i' y)\, dy,$$

where $S(x) = \{y \in S^{q-1}\colon \mu_i' y < \mu_i' x\}$. Since we are integrating a non-negative function, the value of the integral increases with expanding the set $S(x)$. Therefore, the higher is the product $\mu_i' x$, the higher is the value of the integral, and hence $F_D(x, H_1) > F_D(x, H_2)$ iff

$$\mu_1' x > \mu_2' x.$$

□

In the following, we discuss the conditions which guarantee Bayes optimality. The depth-based classifiers employ rotation-invariant distance-based depth functions and therefore the depth is a function of the cosine distance from the mean direction. To achieve correspondence between depth and density (used in the Bayes classifier), we have to assume that the density is also a function of the cosine distance from the mean direction, i.e. the rotational symmetry of the distribution. We further need assumption of monotonicity of a function $h()$ to avoid situations in which the density is low in points close to the mean direction. As already mentioned at the beginning of this section, the other assumptions – on equal variability and equal priors – reduce the applicability of the classifiers in practice quite substantially. Therefore, we investigated the performance of the classifiers in the case of unequal concentrations.

3.3.1 The max-depth classifier in a more general case

The following theorem clarifies the form of the max depth classifier for the cosine depth in the situation in which the considered distributions may differ not only in location but also in dispersion.

Theorem 3.3.2. *Let H_1 and H_2 be two distributions on the sphere S^{q-1}, having mean directions μ_1 and μ_2, respectively, and mean resultant lengths ρ_1 and ρ_2, respectively. If the cosine depth is employed, the max-depth classifier (3.3) has the following form:*

$$c(x) = \underset{i}{\operatorname{argmax}}\, \rho_i \mu_i' x. \tag{3.5}$$

and therefore, the distributions are "separated" by the hyperplane

$$(\rho_1 \mu_1 - \rho_2 \mu_2)' x = 0. \tag{3.6}$$

Proof. The theorem directly follows from the form of the cosine depth in this case:

$$D(x, H_i) = 1 + \mathrm{E}_{H_i} x' X = 1 + \rho_i x' \mu_i.$$

□

The separating hyperplane is determined by the parameters of location (μ_1 and μ_2) and parameters reflecting variability of the distributions (ρ_1 and ρ_2). Clearly, the max-depth classifier does not include (and hence does not account for) information on priors. Also, the

whole information on distribution is reduced only to its mean direction and mean resultant length.

In the case of equal mean resultant lengths, the max depth classifier simplifies to the form $c(x) = \operatorname{argmax} \mu_i' x$. The separating hyperplane is then determined by the equation $(\mu_1 - \mu_2)'x = 0$. It is a hyperplane orthogonal to the hyperplane determined by vectors μ_1 and μ_2 which halves the angle between them.

Note that the formula of the max-depth classifier cannot be simplified in this way when using nonlinear transformations of the scalar product $\mu_i' x$ in the depth function even if the transformation is monotone, i.e. for the arc distance depth and the chord depth.

Geometrically, we can imagine the above-described situation as follows. Denote the angle between μ_1 and μ_2 as θ ($\cos\theta = \mu_1'\mu_2$). There exists an orthogonal matrix R such that $R\mu_1 = (\cos\frac{\theta}{2}, \sin\frac{\theta}{2}, 0, \ldots, 0)' =: \mu_1^0$ and $R\mu_2 = (\cos\frac{\theta}{2}, -\sin\frac{\theta}{2}, 0, \ldots, 0)' =: \mu_2^0$. Hence, we can assume that $\mu_1 = \mu_1^0$ and $\mu_2 = \mu_2^0$.

In this situation, the cosine depth of a point $x = (x_1, x_2, \ldots, x_q)'$ can be expressed in the following form:

$$\begin{aligned} D(x, H_1) &= 1 + \rho_1(x_1 \cos\frac{\theta}{2} + x_2 \sin\frac{\theta}{2}) \\ D(x, H_2) &= 1 + \rho_2(x_1 \cos\frac{\theta}{2} - x_2 \sin\frac{\theta}{2}). \end{aligned}$$

The separating hyperplane is then determined by the equation

$$(\rho_1 - \rho_2)\left(\cos\frac{\theta}{2}\right)x_1 + (\rho_1 + \rho_2)\left(\sin\frac{\theta}{2}\right)x_2 = 0,$$

which simplifies to the form $x_2 = 0$ in the case of equal mean resultant lengths.

3.3.2 Studied class of spherical distributions

We studied a broad subclass of unimodal rotational symmetric distributions on the sphere S^{q-1} for which the Bayes classifier can be derived and subsequently compared to the max depth classifier discussed above.

Let us consider a density function $h(x)$ proportional to a sum $v + g(\mu' x)$, where $v > 0$ is a positive real constant, $\mu \in S^{q-1}$ mean direction and $g\colon [-1, 1] \to \mathbb{R}$ is an odd strictly increasing function.

Note that the higher the value of the constant v, the closer is the distribution to the uniform distribution. Therefore, higher values of v imply higher variability. Parameter v can

be thus understood as a measure of variability.

After plugging in the normalizing constant, the density can be expressed as

$$h(x) = \frac{1}{A_q} + \frac{g(\mu' x)}{v A_q},$$

where A_q denotes the surface area of the sphere S^{q-1}. Assuming that $\mu = (1, 0, \ldots, 0)'$, which can be achieved by a rotation of the distribution, one can derive a relation between the variability parameter v and the mean resultant length ρ of the considered distribution:

$$\rho = \mathrm{E}X_1 = \int_{S^{q-1}} x_1 h(x)\, dx = \int_{S^{q-1}} x_1 \left(\frac{1}{A_q} + \frac{g(x_1)}{v A_q} \right) dx = \frac{1}{v A_q} G_q,$$

where $G_q = \int_{S^{q-1}} x_1 g(x_1)\, dx$ is a constant. The density can thus be expressed as a function of its mean direction $\boldsymbol{\mu} \in S^{q-1}$ and mean resultant length ρ (using above-defined constants A_q and G_q) in the following way:

$$h(x) = \frac{1}{A_q} + \frac{1}{G_q} \rho g(\mu' x). \tag{3.7}$$

Let us now consider a classification problem for two distributions with densities of the above-mentioned form (3.7) with possibly different mean directions and mean resultant lengths, but with the same function $g()$, i.e. we assume $h_i(x) = \frac{1}{A_q} + \frac{1}{G_q} \rho_i g(\mu_i' x),\ i = 1, 2$. Assuming equal prior probabilities, the Bayes classifier can be expressed as

$$c(x) = \underset{i}{\operatorname{argmax}}\, \rho_i g(\mu_i' x).$$

If g is identity, i.e. $g(y) = y$, the Bayes classifier is equivalent to the max-depth classifier if the cosine depth is employed (see Theorem 3.3.2 above).

We have shown that equal variability (expressed by the mean resultant length) is not a necessary condition for optimality. We found a class of distributions, in which optimality is achieved even if distributions differ in variability (identity or some multiple of identity are the only cases of $g()$ in which the Bayes classifier coincides with the max depth classifier).

3.3.3 Bayes classifier in the case of von Mises-Fisher distributions

The class of distributions studied in the previous section does not include the most well-known distribution on the sphere, namely the von Mises-Fisher (vMF) distribution. In this section, we briefly discuss this important case. Let us consider two different vMF distributions, i.e. distributions with densities

$$h_i(x;\mu_i,\kappa_i) = c_{\kappa_i,q}\exp\{\kappa_i\mu_i'x\},\ i=1,2.$$

The equation defining the separating subspace for the Bayes classifier given by equality $\pi_1 h_1(x) = \pi_2 h_2(x)$ can be rewritten as

$$(\kappa_2\mu_2 - \kappa_1\mu_1)'x = \ln\left(\frac{\pi_1 c_{\kappa_1,q}}{\pi_2 c_{\kappa_2,q}}\right). \tag{3.8}$$

As with the max-depth classifier (3.6), the separation subspace is a hyperplane. However, mean directions are multiplied by concentration parameters κ here, not by mean resultant lengths ρ. The relationship between these parameters of variability is not straightforward. The following holds:

$$\rho = \frac{I_{q/2}(\kappa)}{I_{q/2-1}(\kappa)}, \tag{3.9}$$

where I_v is the modified Bessel function of the first kind and order v, see section 9.3.2 of Mardia & Jupp (2009). Note that the ratio (3.9) is strictly increasing in κ. Moreover, the constant on the right-hand side of (3.8) is non-zero in the case of differing priors and concentration parameters (if the considered ratio is not equal to one by chance).

3.4 Final Remarks

This chapter reviewed two depth-based classifiers for directional data, namely the max-depth and max-rank classifiers, and discussed conditions under which they are equivalent to the Bayes (optimal) classifier. Conditions under which optimality is guaranteed include (rotational) symmetry and unimodality of the underlying distributions, with a difference only in location and equal prior probabilities.

These conditions are not necessary and we found a class of distributions for which

the max-depth classifier based on the cosine depth can be optimal even if distributions also differ in variability. On the other hand, such a class does not include the von Mises-Fisher distributions, and the max depth classifier is generally not optimal when groups present different variability levels. Moreover, it was shown that the above classifiers ignore information on prior probabilities.

Chapter 4

Distance-based directional depth classifiers: a robustness study

Abstract

Contaminated training sets can highly affect the performance of classification rules. For this reason, robust supervised classifiers have been introduced. Amongst the many, this work focuses on depth-based classifiers, a class of methods which have been proven to enjoy some robustness properties. However, no robustness studies are available for them within a directional data framework. Here, their performance under some directional contamination schemes is evaluated. A comparison with the directional Bayes rule is also provided. Different directional specific contamination schemes are introduced and discussed: antipodality and orthogonality of the contaminated distribution mean, and the directional mean shift outlier model.

Keywords: DD-plot; Distribution depth; von Mises-Fisher; Max-depth; Spherical data

---

This Chapter has been conditionally accepted for publication in the journal *Communication in Statistics: Simulation and computation* as: Demni, H., Messaoud, A. and Porzio, G.C. "Distance-based directional depth classifiers: a robustness study".

Acknowledgments: The authors wish to thank the two anonymous referees for their valuable comments which led to a considerable improvement of a first version of the present work. Thanks are also due to Mario Guarracino for his precious suggestions. A part of this work has been presented at the International Conference on Classification and Data Analysis (CLADAG2019) in Cassino, Italy and at the workshop Theory and Practice of Statistical Data Processing in Nova Seninka, Czech Republic. Thanks are due to the participants for their precious feedback.

4.1 Introduction

Given a set of labeled data, with the labels expressing their membership to some groups (training set), supervised classification methods aim at predicting the class of new unlabeled observations. Supervised classifiers, which are widely exploited in many scientific fields, have thus the main goal of discriminating between two or more classes. The performance of a classifier is problem specific and it is generally evaluated by considering its ability to correctly assign observations from a test set to the class they belong to.

However, such a performance can be dramatically influenced by the presence of contaminated data points in the training set. As a consequence, one of the issues to be considered when a classifier is adopted is its robustness. That is, its ability to provide reasonable results even in the presence of some kind of contamination. Note that the need for robust classifiers is well recognized in applications (e.g. for gene expression data, Chandra & Gupta, 2011).

Amongst the many available robust techniques, depth-based classifiers have been widely considered. Hubert & Van der Veeken (2010) studied the robustness of projection depth classifiers for skewed data, while Li et al. (2012) examined the robustness properties of the DD-classifier and showed that it is robust against outliers and extreme observations. Dutta & Ghosh (2012) investigated the robustness of projection depth classifiers and showed their superiority against classifiers based on the half-space depth. Exploiting the properties of depth functions, Hubert et al. (2017) proposed a robustified classifier that relies on depth functions and distances. Vencálek & Pokotylo (2018) introduced depth weighted classifiers for multivariate data and investigated their robustness with respect to the Bayes classification rule.

Depth-based classifiers have been introduced in the directional domain as well. More specifically, the directional max-depth classifier (Pandolfo, Paindaveine, & Porzio, 2018), the DD-classifier (Pandolfo, D'Ambrosio, & Porzio, 2018) and the depth distribution classifier (Demni et al., 2019) are available. They have been also successfully employed in different fields of applications (see Demni, 2021; Pandolfo & D'Ambrosio, 2021). Here, directional domain refers to the analysis of distributions whose support is either the boundary of a circle, or a sphere, or a hyper-sphere.

On the other hand, how these directional methods perform in case of contaminated training sets is substantially unknown. We only found some preliminary notes on the robustness of the circular DD-classifier in Pandolfo (2017). Furthermore, robustness in the directional

domain must be evaluated apart: the location parameter space is bounded, and hence the bias is bounded too. As a consequence, as already discussed in Ko & Guttorp (1988), standard robustness measures like influence function or breakdown point are not directly applicable. See also the discussion in Kirschstein et al. (2019), and references therein.

Henceforth, given the lack of studies and the specificity of the robustness issues in the directional domain, this work investigates to what extent the available directional depth-based classifiers are able to deal with contaminated training sets. Two different contamination models are considered: the general outlier case (anomalous data lies somewhere within the admittable space, far from all the labeled groups) and the more specific mislabeling model (some training data have been erroneously labeled).

For the general outlier case, directional specific contamination schemes are introduced and discussed here. Particularly, conditions are investigated under which the mean of the contaminated distribution is antipodal (maximum impact) or orthogonal (half-a-way impact) to the mean of the uncontaminated distribution. The directional mean shift outlier model is also examined.

Finally, for each depth-based classifier, one has to choose the depth function to employ. In the directional domain, several of them are available: the angular simplicial and half-space depths (Liu & Singh, 1992), the angular Mahalanobis depth (Ley et al., 2014), and the distance-based depths (Pandolfo, Paindaveine, & Porzio, 2018). Because of their much lower computational cost, the distance-based depths are considered in this work.

A common practice in supervised classification is to evaluate the robustness of a method through more or less extensive numerical simulations. Accordingly, the comparison of the performance of the discussed methods will be done by means of a thorough simulation study under the von Mises-Fisher distribution, the main directional distribution on spheres. Performance of the depth-based classifiers will be also compared against the directional Bayes classifier (whose optimality properties are well known in the literature).

The chapter is structured as follows. Section 4.2 provides a review on robust supervised classification methods and on how their robustness has been evaluated. Section 4.3 illustrates why robustness studies for directional data need a specific attention, and it introduces directional specific contamination schemes. Section 4.4 reviews directional depth functions and discusses their robustness properties, and presents depth-based classifiers. The simulation study is provided in Section 4.5. Section 4.6 offers some final remarks.

4.2 Robustness of supervised classifiers

The maximum achievable accuracy in classification problems depends on the chosen classification rule as well as on the quality of the training set. Information contained in such labeled data sets is used to classify the remaining unlabeled data. Therefore, contaminated data (also known as unreliable data) can highly affect the performance of classifiers.

In a supervised classification context, two main sources of contamination can be distinguished: outliers and mislabeled data. The former refers to some data points in the training set that are misplaced far away from each cluster, whereas the latter is related to inaccurate group memberships, that is caused by labeling errors. Those kinds of anomalous data are also known as attribute noise and class noise, respectively (see Zhu & Wu, 2004; Frénay & Verleysen, 2013).

Formally, the two sources of contamination can be described using two different kinds of mixture models. Let thus $\mathcal{F} := \{F_i, i = 1, .., K\}$ be a set of K unknown distributions where each F_i characterizes the i-th class. Then, the contamination model for the case of outliers in the i-th group is given by:

$$F_i^{out} := (1 - \epsilon_i)F_i + \epsilon_i G_i \tag{4.1}$$

where F_i^{out} is the distribution associated to the i-th group contaminated by a proportion ϵ_i of outliers, and $G_i \notin \mathcal{F}$. This contamination model derives from the standard contamination model in robustness studies (Huber & Ronchetti, 2009, page 12), and it has been largely adopted to evaluate the performance of classifiers as well.

A sub-model of Eqn.(4.1) is the mean shift outlier model, where the original distribution F_i and the contaminating distribution G_i differ only in location. The mean shift outlier contamination model has been widely considered to evaluate robustness of supervised classifiers.

More recently, a contamination model more specific to supervised classification methods arised: the mislabeled data model. We also define such a model as a mixture of two distributions. We have:

$$F_i^{mis} := \frac{1}{1 + \epsilon_{ji}} F_i + \frac{\epsilon_{ji}}{1 + \epsilon_{ji}} F_j \tag{4.2}$$

where F_i^{mis} is the distribution associated to the i-th group contaminated by a proportion ϵ_{ji} of

observations from F_j. The contamination rate ϵ_{ji} refers thus to the proportion of observations belonging to the j-th group which have been erroneously labeled as if from group i: the mistake happened to the observations from F_j, the impact is on F_i (there will be also an impact on the size of each group in the training set). Equation (4.2) assumes the distribution F_i is contaminated only through mislabeling of observations from F_j. Such a definition can be clearly extended to the case of mislabeling in more classes.

Several techniques have been proposed to deal with both issues in the literature. A first option is to analyze the training set by applying data pre-processing tools in order to detect unreliable observations before doing the classification task. Within this context, many tools are available. Hodge & Austin (2004) provided a survey of techniques for outlier detection and discussed their advantages and disadvantages for clustering, classification and/or recognition. Debruyne (2009) introduced an outlier map for support vector machines. He showed that this tool is efficient to detect outliers and/or mislabeled data when using support vector classifiers. The limit of this first approach is that the analyst needs to decide upon each candidate outlier if it is so or not, and the analysis must be run in two steps (outlier detection, and then classification).

Alternatively, several robust classification methods are available. Hawkins & McLachlan (1997) introduced high-breakdown linear discriminant analysis and showed their method is robust against contamination by outliers through real data sets. Thereafter, robust discriminant rules based on high-breakdown estimators, on robust estimates of location and scatter, and on the minimum covariance determinant estimates of the mean and covariance have been proposed under the hypothesis that each group is normally distributed (see Croux & Dehon, 2001; Hubert & Van Driessen, 2004; Abebe & Nudurupati, 2009).

Within the same framework, Joossens & Croux (2004) studied the effect of outliers on the behavior of robust linear and quadratic discriminant analysis through an extensive simulation study. They compared the total probabilities of misclassification under different contamination schemes and discussed which method should be preferred. Recently, a semi-supervised classification method has been suggested in order to improve the classifier performance and to identify outlying observations (Cappozzo et al., 2020). Nakayama (2019) also proposed a robust version of a support vector machine classifier, that can handle high-dimensional imbalanced data.

Hawkins & McLachlan (1997) discussed the breakdown point (i.e. the maximum number of outliers that can be accommodated) of their method under the mean-shift outlier model. All

the others evaluated the performance of their classification methods by means of simulations. Mainly, simulation studies considered normal distributions. Abebe & Nudurupati (2009) included the Cauchy distribution.

4.3 Directional data, robustness, and directional contamination scenarios

Directional data can be viewed as unit vectors on the unit hyper-sphere $S^{(q-1)} := \{x \in \mathbb{R}^q, ||x|| := \sqrt{x'x} = 1\}$. Such data can be encountered in many fields such as neurosciences, astrophysics, oceanography and biology, to cite a few. For a recent review on directional statistics, the reader is referred to Pewsey & García-Portugués (2020). For discriminant analysis on the sphere, see in particular (Tsagris & Alenazi, 2019).

Being a hyper-sphere, the support of a directional variable is bounded by definition. As a consequence, directional location parameters have bounded parameter space, and their robustness features need to be analyzed apart. For instance, given that their influence function is also bounded, Ko & Guttorp (1988) introduced the standardized influence function. Later, He & Simpson (1992) discussed breakdown points for compact parameter space, while Kirschstein et al. (2019) analyzed the finite sample maximum bias of some directional location estimators.

To illustrate the issue, consider the robustness properties of the expected value of a linear variable (we call a linear variable a variable whose support is a subset of the real line). Clearly, by Eqn.(4.1), if the expected value of the contaminating distribution G_i goes to infinity, the expected value of the contaminated distribution F_i^{out} goes to infinity as well, irrespective of the value of the contamination level $\epsilon > 0$. This is a well known story who suggests to use the median as a robust location parameter (see e.g. Wilcox, 2011, Sect. 2.1.3).

Unlikely, directional location parameters cannot go to infinity. Under which conditions they go the furthest possible becomes thus a case of interest.

To exemplify, let us focus on the directional mean. By definition, the directional mean of a directional random variable $X \sim H$ is defined as $\mu_X := E(X)/||E(X)||$, provided that $||E(X)|| \neq 0$, where E() is the expected value operator, and $||\cdot||$ is the L_2 Euclidean norm. By definition of directional variables, we also have $||E(X)|| \leq 1$. Note that $||E(X)|| =: \rho_X$ is a measure of dispersion and it is called the mean resultant length of the directional variable

X.

Consider then the contamination model introduced above (Eqn.4.1) for a directional distribution. Provided that it exists, it is possible to derive its directional mean, as stated in the following Theorem.

Theorem 4.3.1 (Directional mean of a contaminated distribution)**.** *Let $X_1 \sim H_1$ and $X_{c.ing} \sim H_{c.ing}$ be two directional random variables whose directional mean exists. Let also denote with μ_1 and $\mu_{c.ing}$ their directional means, respectively. Consider then the contaminated directional distribution H_C with a level ϵ of contamination:*

$$H_C := (1-\epsilon)H_1 + \epsilon H_{c.ing} \qquad 0 < \epsilon < \frac{1}{2}. \tag{4.3}$$

The directional mean μ_C of the contaminated distribution H_C is given by:

$$\mu_C = \frac{(1-\epsilon)\mu_1||E(X_1)|| + \epsilon\mu_{c.ing}||E(X_{c.ing})||}{||(1-\epsilon)\mu_1||E(X_1)|| + \epsilon\mu_{c.ing}||E(X_{c.ing})||||}, \tag{4.4}$$

provided that $||E(X_C)|| \neq 0$, with $X_C \sim H_C$.

The proof follows from standard properties of expected values. The statement is similar to Eqn.(4.4) in Kirschstein et al. (2019). However, in that work the statement referred to an estimator of the directional mean. Here, the interest is on the parameter itself. There, the contaminating distribution was a Dirac distribution (worst case). Here, the contaminating distribution is a generic directional distribution with $||E(X_{c.ing})|| \neq 0$.

After Theorem 4.3.1, we are able to introduce and investigate three different directional contamination scenarios. That is, the cases of antipodality and orthogonality of the directional mean of the contaminated distribution, and the mean shift outlier model. These cases will be exploited to evaluate the performance of the classifiers later in Section 4.5.

First of all, Theorem 4.3.1 allows investigating conditions under which a contamination achieves its maximum impact on the directional mean. That is, under which conditions the directional mean of the contaminated distribution H_C is antipodal to the mean of the uncontaminated distribution (i.e., $\mu_C = -\mu_1$). The result is stated as a Corollary.

Corollary 4.3.2 (Maximum impact on the directional mean)**.** *Let the conditions of Theorem 4.3.1 hold. If $\mu_{c.ing} = -\mu_1$ and*

$$(1-\epsilon)||E(X_1)|| - \epsilon||E(X_{c.ing})|| < 0, \tag{4.5}$$

then

$$\mu_C = -\mu_1. \tag{4.6}$$

In other words, if the contaminating distribution is antipodally located ($\mu_{c.ing} = -\mu_1$) and its mean resultant length $||E(X_{c.ing})||$ is larger than $(1-\epsilon)/\epsilon||E(X_1)||$, then the directional mean moves by an angle of width equal to π radians (it moves the furthest possible from its original location). Note that, when $\mu_{c.ing} = -\mu_1$ and instead $(1-\epsilon)||E(X_1)||-\epsilon||E(X_{c.ing})|| > 0$, we have that $\mu_C = \mu_1$. Hence, in this latter case, an antipodally located contaminating distribution has no impact at all on the directional mean of the uncontaminated distribution.

From Theorem 4.3.1 it is also possible to derive conditions under which the mean of the contaminated distribution is $\pi/2$ radians far away from its original value (or orthogonal to the original mean, $\mu_C^T\mu_1 = 0$). That is, under which conditions a contamination moves the directional mean half a way between the maximum impact and no impact on the contaminated distribution. This result is stated as a Corollary as well.

Corollary 4.3.3 (Impact on the directional mean: orthogonality)**.** *Let the conditions of Theorem 4.3.1 hold. Then* $\mu_C^T\mu_1 = 0$ *iff*

$$\mu_{c.ing}^T\mu_1 = -\frac{(1-\epsilon)}{\epsilon}\frac{||E(X_1)||}{||E(X_{c.ing})||}, \tag{4.7}$$

with $\frac{(1-\epsilon)}{\epsilon}\frac{||E(X_1)||}{||E(X_{c.ing})||} < 1$.

The condition in Eqn.(4.7) has many solutions. For instance,

$$\begin{cases} \mu_1 & = (0, 0, \cdots, 1)^T, \\ \mu_{c.ing} & = \big(-\sqrt{1-\big(\frac{(1-\epsilon)}{\epsilon}\frac{||E(X_1)||}{||E(X_{c.ing})||}\big)^2}, 0, \cdots, 0, -\frac{(1-\epsilon)}{\epsilon}\frac{||E(X_1)||}{||E(X_{c.ing})||}\big)^T. \end{cases} \tag{4.8}$$

Finally, the mean shift outlier model within the directional setting can be also investigated thanks to Theorem 4.3.1. This is the case where the original and the contaminating distributions have equal mean resultant length ($||E(X_1)|| = ||E(X_{c.ing})||$). Under such a condition, Eqn.(4.4) simplifies and the following Corollary can be stated.

Corollary 4.3.4 (Directional Mean Shift Outlier Model: Impact on the directional mean)**.**

Let the conditions of Theorem 4.3.1 hold. Let also $||E(X_1)|| = ||E(X_{c.ing})|| \neq 0$. *Then,*

$$\mu_C^T \mu_1 = \frac{(1-\epsilon) + \epsilon \mu_{c.ing}^T \mu_1}{||(1-\epsilon)\mu_1 + \epsilon \mu_{c.ing}||}. \tag{4.9}$$

To illustrate Eqn.(4.9), the functions which relate the angles in degrees between the means of the contaminated and the uncontaminated distributions ($\arccos(\mu_C^T \mu_1)$) with the angle in degrees between the means of the contaminating and the uncontaminated distributions ($\arccos(\mu_{c.ing}^T \mu_1)$) have been depicted in Figure 4.1 for different levels of contamination ϵ ($\epsilon = \{0.05, 0.10, 0.20, 0.30, 0.40, 0.49\}$). They are all non-monotone functions, with a maximum which depends on the level of contamination and which is roughly located within a range from 80° for $\epsilon = 0.05$ to 165° for $\epsilon = 0.49$.

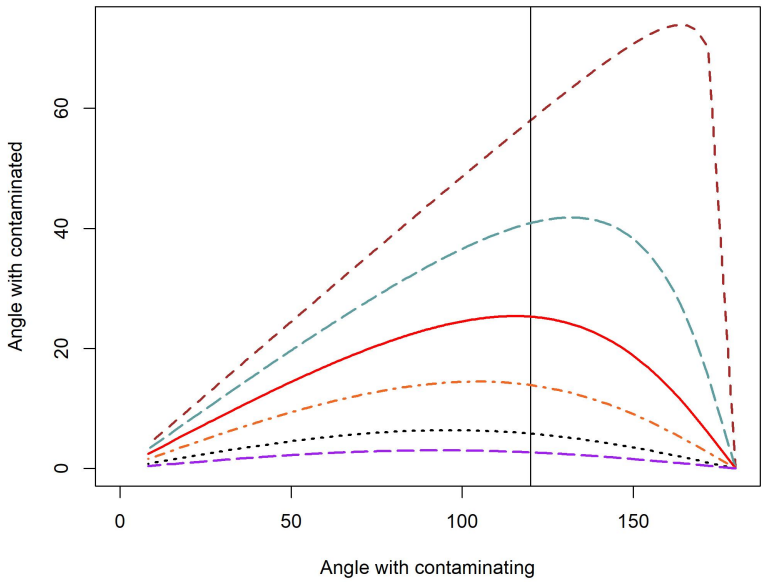


Figure 4.1: Directional mean shift outlier model. Impact on the directional mean of a certain level of contamination ϵ for a value of the angle between the means of the uncontaminated (H_1) and the contaminating distribution ($H_{c.ing}$). Each function refers to a different level of contamination ($\epsilon = \{0.05, 0.10, 0.20, 0.30, 0.40, 0.49\}$; long-dashed, dotted, dot-dashed, solid, dashed, and two-dashed lines, respectively). The vertical axis represents the angle in degrees between the directional mean of the original distribution μ_1 and the directional mean of the contaminated distribution μ_C. The maximum of a function tells the value at which the directional mean of the contaminating distribution $\mu_{c.ing}$ must lie apart from μ in order to have the largest possible impact on the mean of the contaminated distribution (H_C) for a given level of contamination. To exemplify, a vertical line is depicted at 115.5°: a value for which the function corresponding to $\epsilon = 0.30$ (solid red line) is approximately maximized.

Results from Corollaries (4.3.2), (4.3.3), and (4.3.4) allow illustrating some specificity of robustness issues within the directional domain. First, unlike the mean of linear variables, the directional mean is robust up to a certain extent (it cannot reach its antipodal value unless $(1-\epsilon)||E(X_1)|| - \epsilon||E(X_{c.ing})|| < 0$, by Corollary (4.3.2)). Second, unlike the linear case, its robustness properties depend on the level of dispersion of both the uncontaminated and contaminating distributions. Third, the mean shift outlier model - largely adopted to discuss robustness of linear variables - does not yield the worst contamination case (even if $\mu_{c.ing} = -\mu_1$, no $\epsilon < \frac{1}{2}$ yields $\mu_C = -\mu_1$).

Corollaries (4.3.2), (4.3.3), and (4.3.4) will be exploited to design the simulation study in Section 4.5. The robustness of the mentioned directional depth-based classifiers will be investigated under the case of antipodality, orthogonality, and mean shift outlier model.

For the mean shift outlier model, the performance will be evaluated at $\mu_{c.ing}^T\mu_1 = -0.43$, i.e. at $\arccos(\mu_{c.ing}^T\mu_1) \approx 115.5°$). This is the value at which the impact on the directional mean of the uncontaminated distribution is approximately maximized under $\epsilon = 0.30$ (to illustrate, a vertical line at 115.5° has been superimposed on the plot in Figure 1).

4.4 Directional data depths and depth-based classifiers

This section reviews distance-based depth functions and their robustness properties. It then introduces depth-based classifiers for directional data.

4.4.1 Directional depth functions

Statistical depth functions extend univariate ordering to higher dimensions by ordering multivariate data with respect to a center. Particularly, they offer a center-outward ordering by providing a measure of how central a point is with respect to a certain distribution. To cite, the first and most adopted depth function is the well known Tukey's half-space depth (Tukey, 1975). Many other depth notions are available (see e.g. Liu et al., 1999; Y. Zuo & Serfling, 2000).

In analogy with data depth for data points in $\mathbb{R}^q$, the depth idea has been extended to the directional domain as well (see Small, 1987; Liu & Singh, 1992). Accordingly, directional depth functions measure the degree of centrality of a point in the sample space with respect

to a directional distribution, and they provide a center-outward ordering on circles or on hyper-spheres.

In the current study, for computational reasons, the focus is on three rotational invariant distance-based depth functions: the arc distance, the chord, and the cosine depth (Pandolfo, Paindaveine, & Porzio, 2018). A directional distance-based depth of a point $x \in S^{(q-1)}$ with respect to a directional distribution H is given by:

$$D(x, H) := d^{sup} - E_H(d(x, X)),$$

where d is a bounded distance on $S^{(d-1)}$, d^{sup} is the upper bound of such a distance between any two points on $S^{(q-1)}$, $E(.)$ is the expected value, and X is a random variable from H.

Accordingly, the following rotational invariant distance-based depth can be defined:

- The cosine depth: $D_{cos} := 2 - E_H(1 - x'X)$;
- The arc distance depth: $D_{arc} := \pi - E_H(\arccos(x'X))$;
- The chord depth: $D_{chord} := 2 - E_H(\sqrt{2(1 - x'X)})$.

The empirical version of each of the depth functions is obtained by replacing H by $\hat{H}$. General properties of the distance-based depth functions have been widely discussed in Pandolfo, Paindaveine, & Porzio (2018).

4.4.2 Robustness of distance-based directional depth functions

Robustness properties of the distance-based depth functions have been essentially investigated in terms of the robustness of the deepest point $z(H) := \sup_{x \in S^{(d-1)}} D(x, H)$. Mainly, a general result on the directional breakdown point of the deepest points for the depth functions discussed above is available.

Let consider the contamination model defined by Eqn.(4.3), and let us define the breakdown point of a directional deepest point as the infimum of the contamination level ϵ such that the deepest point of the contaminated distribution is antipodal to the deepest point of the uncontaminated distribution. It has been proved that the breakdown point of the deepest point $z(H)$ is greater or equal than $[D(z(H), H) - D(-z(H), H)]/(2d^{sup})$. For von Mises-Fisher distributions, such a bound is larger for the cosine depth deepest point when compared with the bounds for the arc and chord distance deepest points (Pandolfo, Paindaveine, & Porzio, 2018).

Further robustness properties of the deepest points are depth specific, and are inherited by the properties of the corresponding parameter. For instance, given that the cosine depth deepest point is the spherical mean, its maximum bias is given by Theorem 3 in Ko & Guttorp (1988).

On the other hand, it is possible to derive SB-robustness properties of the cosine and arc distance deepest points by Theorem 6.2 in He & Simpson (1992). We have that the arc distance deepest point is SB-robust at the von Mises-Fisher distribution, while the cosine depth deepest point does not enjoy such a property.

4.4.3 Directional depth-based classifiers

A directional classifier is a function $class : S^{(q-1)} \to \{1, ..., i, ..., K\}$ where the integer i refers to one of the K different directional distributions $H_1, ..., H_i, ..., H_K$. Three main directional depth-based classifiers are available: the max-depth classifier, the depth distribution classifier (also called max-rank classifier) and the depth versus depth classifier (DD-classifier). The first assigns observations to the distribution H_i where they attains the highest depth value. The second exploits the cumulative distribution function of the depth within each labeled group. The third defines a discriminating function within the DD-plot, a plot where each observation has coordinates based on its depth value with respect to two directional distributions.

In particular, given a training set composed by K empirical distributions $\hat{H}_i$, $i = 1, .., K$, the max-depth classification rule is given by

$$class_{max}(x) := \text{argmax}_i \, D(x; \hat{H}_i) \qquad i = (1, ..., K),$$

where x is a new observation to be classified, and $D(x, \hat{H}_i)$, $i = 1, .., K$, is the empirical depth of x with respect to the distribution $\hat{H}_i$.

The depth distribution classifier is based on the value

$$F_D(x, \hat{H}_i) := P(D(X, \hat{H}_i) \leq D(x, \hat{H}_i)),$$

where $X \sim H_i$. Then, to classify a new observation, the directional depth distribution classification rule is given by:

$$class_{dd}(x) := \text{argmax}_i \, F_D(x, \hat{H}_i) \qquad i = (1, ..., K).$$

The DD-classification rule has been thought for the case of two or few more classes and it is given by

$$class_{DD}(x) := \text{argmax}_i \, r(D(x; \hat{H}_i)) \qquad i = (1, ..., K),$$

where $r(.)$ is a real increasing function which has the aim of discriminating points in the DD-space. Several separating functions can be considered for discrimination in the DD-space. Li et al. (2012) suggested a polynomial function whose degree needs to be chosen by minimizing the average error rate on the training set. In case r is a straight line, Li et al. (2012) advice it should pass through the origin. If such a line is the 45 degree line, we obtain the max-depth classifier as a special case. Other choices can be also made such as the linear and quadratic discriminant separating rules and the k-nearest neighbors classification rule (see e.g. Pandolfo & D'Ambrosio, 2021).

Optimality properties of the max-depth classifier and of the depth distribution classifier for directional data were investigated by Vencálek et al. (2020). They showed that, under rotational invariance of the adopted depth functions, both classifiers are equivalent to the optimal Bayes classifier under the following assumptions on the underlying distributions: unimodality, rotational symmetry, difference only in location and equal priors. They also proved that, under some circumstances, the cosine max-depth can be optimal even when distributions are enjoying different concentration parameters.

4.5 Robustness of directional depth-based classifiers: a simulation study

The robustness of the max-depth, the depth distribution and the DD classifiers is investigated in this section through a simulation study. To evaluate their performance, the two contamination models mentioned in Section 4.2 are considered. For the first of them, the three contamination scenarios introduced in Section 4.3 are adopted.

To simplify, a binary classification problem is discussed, and only one of the two groups will be contaminated within the training set. To avoid confusion, we will use the following notation throughout this Section. Two directional distributions H_1 and H_2 will be considered: the first will be contaminated by a third distribution $H_{c.ing}$, while the second will be not. Accordingly, for the case of outliers (Eqn.(4.1)), we will have:

$$H_1^{out} := (1 - \epsilon_1)H_1 + \epsilon_1 H_{c.ing} \qquad 0 \leq \epsilon_1 < 1/2. \tag{4.10}$$

For the case of the mislabeled data, by Eqn.(4.2) we have:

$$H_1^{mis} := \frac{1}{1+\epsilon_{21}} H_1 + \frac{\epsilon_{21}}{1+\epsilon_{21}} H_2 \qquad 0 \leq \epsilon_{21} < 1/2. \tag{4.11}$$

Data are generated from von Mises–Fisher distributions, the most widely used distribution to model data on spheres and in particular in the framework of supervised classification (see Figueiredo, 2009; López-Cruz et al., 2015). Under the hypothesis that each class H_i, $i = 1, 2$ follows a von Mises-Fisher distribution on $S^{(q-1)}$, their probability density function is given by

$$h_i(x; \mu_i, c_i) := \left(\frac{c_i}{n}\right)^{q/2-1} \frac{1}{\Gamma^{(q/2)} I_{q/2-1}(c_i)} \exp\{c_i \mu_i^T x\}, \tag{4.12}$$

where the parameters μ_i and c_i denote the mean direction and the concentration parameter, respectively, with $||\mu_i|| = 1$, $c_i \geq 0$, and I_v denotes the modified Bessel function of the first kind and order v. The higher is the value of the concentration parameter c_i, the more the data are concentrated on the sphere. The case of $c_i = 0$ yields the uniform distribution on the sphere. The concentration level c_i is a strictly increasing function of the mean resultant length of the distribution. Tabled values of such a function for the case $q = 3$ are available in (Mardia & Jupp, 2009, Appendix 3.2).

4.5.1 The directional Bayes classifier as a benchmark

As a benchmark, because of its optimality, the directional Bayes classifier is considered. The Bayes rule is optimal as it minimizes the total probability of misclassification. In the directional domain, it is defined as

$$class_{Bayes}(x) := \text{argmax}_i \, h_i(x) p_i$$

where h_i, is the directional density function which corresponds to the distribution H_i, and p_i is its corresponding prior probability, $i = 1, .., K$.

The empirical version of the Bayes classifier is adopted in practice and the case of equal prior is considered here. That is, the density parameters are estimated on the training set, and - within our setting - we obtain

$$class_B(x) := \text{argmax} \, h_i(x, \hat{\mu}_i, \hat{c}_i),$$

where $h_i(.)$ is now the density in Eqn.(4.12), and $\hat{\mu}_i$ and $\hat{c}_i$ are estimates of the corresponding parameters.

Given the aim of this work, three versions of the empirical Bayes classifiers are considered within the simulation study. Each version differs from the others for the way the location and concentration parameters μ_i and c_i are estimated on the training set. We adopt the standard maximum likelihood estimators, and both the robust M-type estimators introduced by Kato & Eguchi (2016). Note that the optimality is reached in case the parameters are not estimated but known in advance (this case being known as theoretical Bayes classifier).

4.5.2 Simulation design

All the data were generated according to a von Mises-Fisher (VmF) distribution. The location parameters of the distributions H_1 and H_2 were set to be orthogonal. Because of the rotational invariance of the procedure, without loss of generality, we set $\mu_1 = (0, 0, 1)$ and $\mu_2 = (1, 0, 0)$.

The following simulation scheme is adopted.

(a) Three directional depth-based classifiers: max-depth, depth distribution, DD-classifier.

As mentioned in Sect. 4.3, the classifiers considered for the comparison are the max-depth, the depth distribution classifiers and the DD-classifier.

(b) Three depth functions: cosine depth, chord depth, arc distance depth.

As discussed, these rotational invariant distance depth functions are computationally feasible even in high dimensions (unlike the angular simplicial and half-space depths). In addition, optimality of the max-depth and depth distribution classifiers when associated with the cosine depth is ensured under von Mises-Fisher and equal prior if distributions differ only in location.

(c) Three discriminating functions in the DD-space (for the DD-classifier): linear discriminant analysis rule (lda), quadratic discriminant analysis rule (qda) and k-nearest neighbors classification rule ($k$nn).

We adopt the linear discriminant analysis, the quadratic discriminant analysis and the k-nearest neighbor rules as separating functions in the DD-classifier. The tuning parameter k for the k-nearest neighbor ($k$nn) is chosen by cross validation.

(d) Three outlier contamination settings (Eqn.4.1), one contamination setting for mislabeled data (Eqn.4.2).

Within the training set, the first group is contaminated, the second group it is not. The three outlier contamination settings are based on the discussion in Section 4.3. For the first two of them, we consider the cases for which under 30% of contamination we have either antipodality or orthogonality of the directional mean of the contaminated distribution H_1^{out} with respect to the mean of the uncontaminated distribution H_1, respectively. The third is the mean outlier shift model with maximum impact under $\epsilon = 30\%$.

Accordingly, for the antipodal case, the training observations from H_1 are contaminated with observations generated from a VmF with location parameter equal to $\mu_{c.ing} = (0, 0, -1)$ and mean resultant length $||E(X_{c.ing})|| = \rho_{c.ing} = 0.9$ (which corresponds to $c_{c.ing} = 10$).

To have the mean of the contaminated distribution orthogonal to the original mean of H_1, the contamination set is from a VmF with location parameter equal to $\mu_{c.ing} = (-0.855, 0, -0.519)$ (calculated from Eqn.4.7), and mean resultant length $\rho_{c.ing} = 0.9$ ($c_{c.ing} = 10$). The mean resultant length for H_1 is equal to $\rho_1 = 0.2$ ($c_1 = 0.615$) in both cases of antipodality and orthogonality.

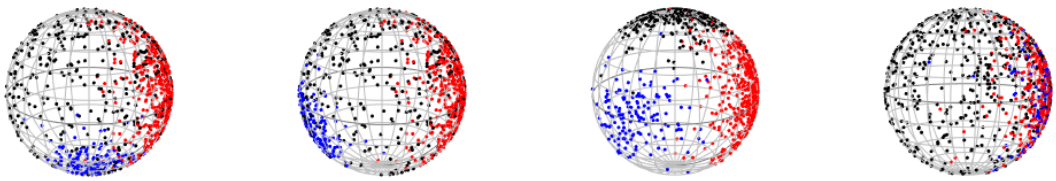
The location parameter of the contamination set in H_1 within the mean shift outlier model is $\mu_{c.ing} = (-0.43, -0.903, 0)$ according to the discussion in Section 4.3 (see also Figure 4.1). The mean resultant length of the contaminating distribution is equal to the mean resultant length for H_1 ($\rho_1 = \rho_{c.ing} = 0.9$, $c_{1_{c.ing}} = c_{c.ing} = 10$).

In case of contamination with mislabeled data, a percentage ϵ_{21} of observations from H_2 is labeled as from group 1. That is, according to Eqn.4.2, the training observations from H_1 are contaminated with observations coming from H_2.

To provide insights on this simulation design, training sets-one for each of the adopted contamination settings (for $\epsilon = 0.3$), are plotted in Figure 4.2 (panel a, b, c and d). Non-contaminated observations from H_1 are depicted in black, those from H_2 in red, and the contaminating set is depicted in blue. Panel (a) represents the case of antipodal mean to H_1 while panel (b) illustrates the case where the mean of the contaminated distribution H_{cont} is orthogonal to the original mean of H_1. The mean shift outlier model is depicted in panel (c), while the contamination case with mislabeled data is in panel (d).

(e) Three levels of contamination on the training set from H_1: $\epsilon = 0\%, 10\%$, and 30%.

For $\epsilon = 0\%$ we have no contamination at all, and hence optimality of the theoretical Bayes classifier. Then, the case of a certain degree of contamination ($\epsilon = 10\%$) and of a more unfavorable case ($\epsilon = 30\%$) are investigated. See the discussion above on the effect of 30% of



(a) (b) (c) (d)

Figure 4.2: Illustration of the simulation design. Contamination $\epsilon = 0.30$. Mean resultant length $\rho_1 = 0.2$, $\rho_2 = 0.8$ and $\rho_{c.ing} = 0.9$. Contamination models: mean of the contaminated distribution H_1^{out} is antipodal to the original uncontaminated mean of group 1 (panel a), orthogonal mean of H_1^{out} to the original uncontaminated mean of group 1 (panel b), mean shift outlier model (panel c), mislabeled data (panel d). The black and red points refer to the two main groups (contaminated group in black), while the blue points represent the contaminating data.

contamination in our setting.

(f) Two cases of mean resultant length for the group which is not contaminated (H_2): $\rho_2 \in \{0.8, 0.9\}$ ($c_2 \in \{5, 10\}$).

Conditions (d), (e) and (f) are considered for the three Bayes classifiers (maximum likelihood and robust) as well. Hence, the complexity of the design is $(2 \times 3 \times 4 \times 3 \times 2 + 3 \times 3 \times 4 \times 3 \times 2 + 3 \times 4 \times 3 \times 2) = 432$ simulation conditions.

In order to preserve the numerical stability of the methods, we use a multiple holdouts simulation generating scheme. Under each simulation condition, 150 training sets of size 1000 are generated (500 from each group). For each of them, testing sets made up of 500 observations (250 from each H_i) are generated and the obtained misclassification rate is recorded. Furthermore, under contamination, each training set from H_1 is contaminated by 10 different contamination sets generated according to $H_{c.ing}$ (each set of size $500 \cdot \epsilon$), and the average misclassification rate over the 10 contamination sets is recorded.

Under the mislabeled data model, we will have $500 \cdot (1 + \epsilon_{21})$ observations in group 1, and $500 \cdot (1 - \epsilon_{21})$ observations in group 2. As a matter of fact, this implies that the two groups have the same size under no contamination, they will differ in size in case of contaminated data.

4.5.3 Simulation results

The performance of each classifier is evaluated by looking at the empirical distribution of the misclassification rate (the proportion of misclassified observations in each replicated sample), for 150 replications. In order to illustrate and discuss the obtained results, the misclassification rate distribution under each setting is summarized through boxplots (Figure 4.3-4.6).

Within each contamination scheme, we keep the same scale on the vertical axis. For the sake of comparison, a line corresponding to the average misclassification rate of the theoretical Bayes under no contamination in each simulation scheme has been superimposed in each plot. Results are discussed for all the contamination cases in detail.

Comparison of depth-based classifiers. Under no contamination, the DD-classifier provides the overall best performance with respect to the max-depth and depth distribution classifiers (Figure 4.3-4.6, first column of plots). If outliers are present, the overall better performance is still achieved by the DD-classifier expect for the case of orthogonality. In this latter case, we observe that the DD-classifier is not the best if associated with a linear separating function. (Figure 4.4, middle and right panels).

Worth of note is the case of the mean shift outlier model, under no contamination, when H_1 and H_2 differ only in location (Figure 4.5, bottom left corner plot, Appendix A, Table A.6): the performance of all the depth-based classifiers is equivalently good and comparable to the Bayes rule. This result seems to suggest that the optimality property of the cosine max-depth and the cosine depth distribution classifiers (Vencálek et al., 2020) holds for the DD-classifier and for the arc and chord depth functions as well.

Under the same scheme, when H_1 and H_2 differ in both location and in concentration (Figure 4.5, upper left corner plot), the behavior of all the classifiers is still comparable. This is probably related to the fact that data are separated and not too sparse on the sphere within this scheme, and hence they can be fairly discriminated.

Within the case of mislabeled data (Figure 4.6, Appendix A, Table A.7-8), the less robust clearly appears to be the depth distribution classifier (Figure 4.6, right panels).

Effect of contamination. The effect of contamination depends on the setting and on the classifier. In case of antipodality and orthogonality, the contamination level has substantially no effect on the classifier performances.

In case of mean shift model, we observe an impact of the contamination level on the empirical Bayes based on the maximum likelihood estimates, while the Bayes based on the

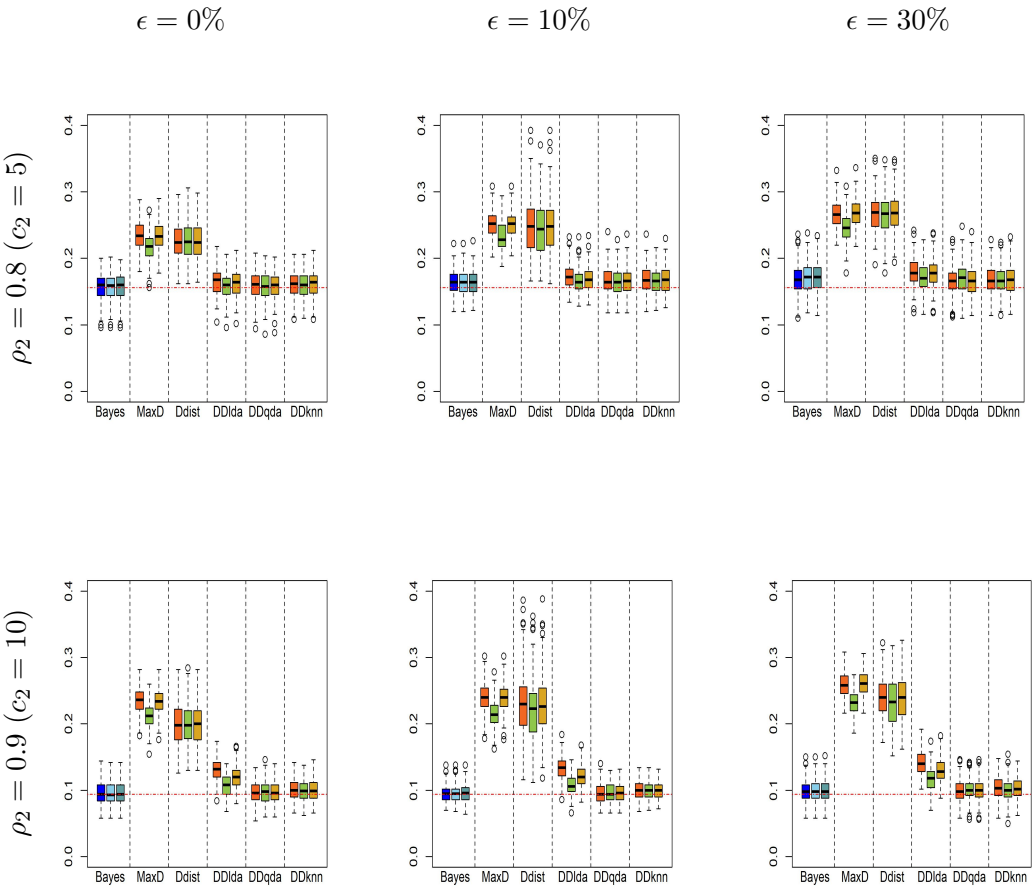


Figure 4.3: Antipodality. Under $\epsilon = 0.30$, the mean of the contaminated distribution H_1^{out} is antipodal to the mean of the original uncontaminated distribution H_1. Boxplots of misclassification rates of the empirical Bayes, max-depth (MaxD), depth distribution (Ddist) and DD-classifiers (DD). Plots by column: no contamination ($\epsilon = 0$), contamination ($\epsilon = 10\%, 30\%$). Plots by row: concentration of the uncontaminated group H_2 $c_2 = 5$ ($\rho_2 = 0.8$), $c_2 = 10$ ($\rho_2 = 0.9$). Within each plot each graph-box refers to the empirical Bayes, MaxD, Ddist and DD classifiers (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn), respectively. The blue box-plot refers to the empirical Bayes, the sky blue box-plot refers to the robust Bayes type 0 estimators and the cadet blue box-plot refers to the robust Bayes type 1 estimators. The orange box-plots refer to the cosine depth, the green ones to the chord depth and the yellow ones to the arc distance depth. The horizontal dashed red line gives the average misclassification rate of the theoretical Bayes under no contamination.

Kato and Eguchi type 0 estimator seems to be the most robust solution. However, we should note that differences in the average mislcassification rates are negligible in absolute value between the different empirical Bayes and the DD-plot classifiers for any contamination level.

In case of mislabeled data, a slight effect of contamination is observed for all the clas-

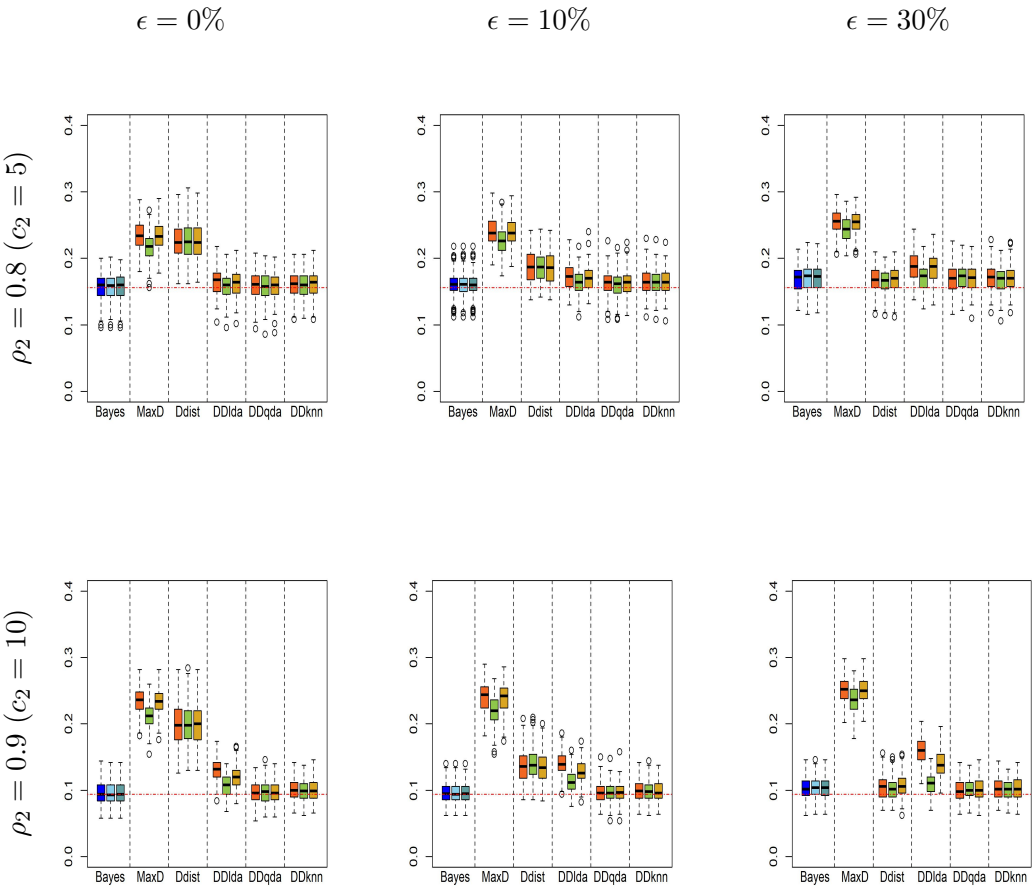


Figure 4.4: Orthogonality. Under $\epsilon = 0.30$, the mean of the contaminated distribution H_1^{out} is orthogonal to the mean of the original uncontaminated distribution H_1. Box-plots of misclassification rates of the empirical Bayes, max-depth (MaxD), depth distribution (Ddist) and DD-classifiers (DD). Plots by column: no contamination ($\epsilon = 0$), contamination ($\epsilon = 10\%, 30\%$). Plots by row: concentration of the uncontaminated group H_2 $c_2 = 5$ ($\rho_2 = 0.8$), $c_2 = 10$ ($\rho_2 = 0.9$). Within each plot each graph-box refers to the empirical Bayes, MaxD, Ddist and DD classifiers (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn), respectively. The blue box-plot refers to the empirical Bayes, the sky blue box-plot refers to the robust Bayes type 0 estimators and the cadet blue box-plot refers to the robust Bayes type 1 estimators. The orange box-plots refer to the cosine depth, the green ones to the chord depth and the yellow ones to the arc distance depth. The horizontal dashed red line gives the average misclassification rate of the theoretical Bayes under no contamination.

sifiers.

Effect of depth functions. The misclassifcation rates of the cosine depth, chord depth and arc distance depth classifiers are more or less comparable in each of the examined schemes. The performance of the depth-based classifiers does not seem to be mainly related to the choice

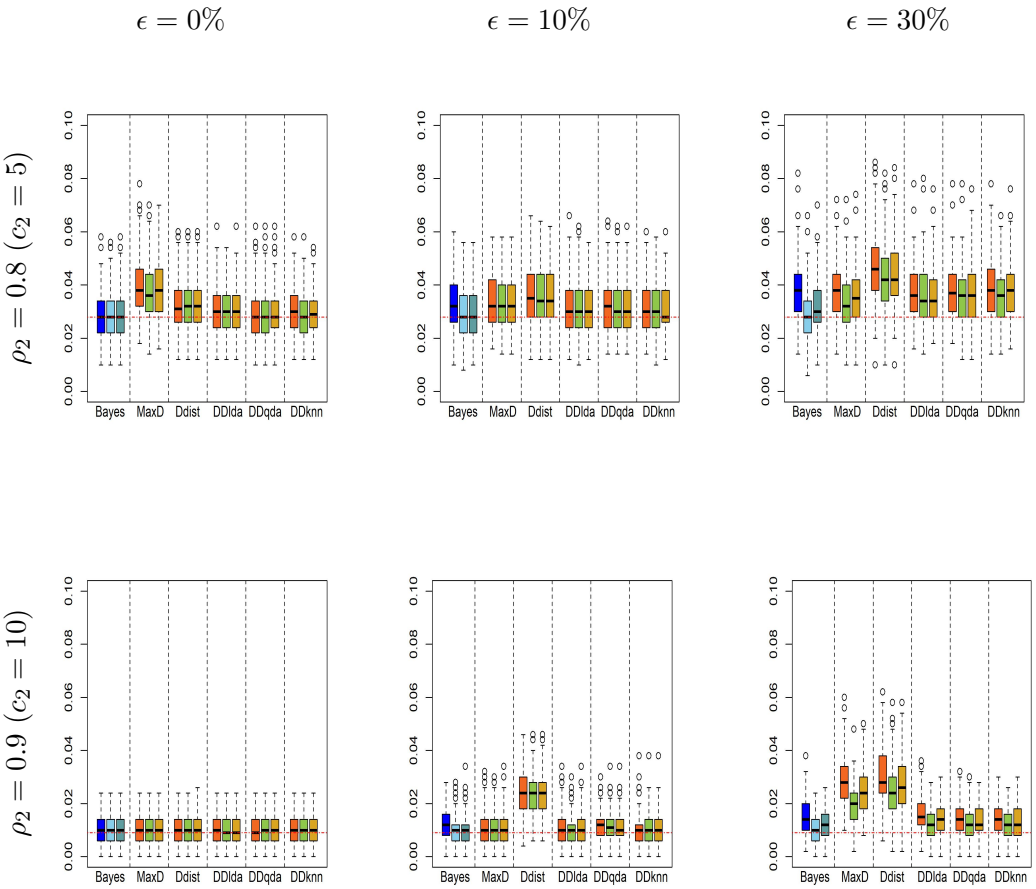


Figure 4.5: Mean shift outlier model. The original uncontaminated distribution H_1 and the contaminating distribution $H_{c.ing}$ differ only for their directional mean. Means are 115° far from each other: this yields the maximum achievable impact on the mean of the original uncontaminated distribution H_1 under $\epsilon = 0.30$. Boxplots of misclassification rates of the empirical Bayes, max-depth (MaxD), depth distribution (Ddist) and DD-classifiers (DD). Plots by column: no contamination ($\epsilon = 0$), contamination ($\epsilon = 10\%, 30\%$). Plots by row: concentration of the uncontaminated group H_2 $c_2 = 5$ ($\rho_2 = 0.8$), $c_2 = 10$ ($\rho_2 = 0.9$). Within each plot each graph-box refers to the empirical Bayes, MaxD, Ddist and DD classifiers (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn), respectively. The blue box-plot refers to the empirical Bayes, the sky blue box-plot refers to the robust Bayes type 0 estimators and the cadet blue box-plot refers to the robust Bayes type 1 estimators. The orange box-plots refer to the cosine depth, the green ones to the chord depth and the yellow ones to the arc distance depth. The horizontal dashed red line gives the average misclassification rate of the theoretical Bayes under no contamination.

of the depth function in these settings.

Effect of the separating function within the DD-plot. The DD-classifier associated with the quadratic discriminant function (qda) and the k-nearest neighbors classification rule ($k$nn)

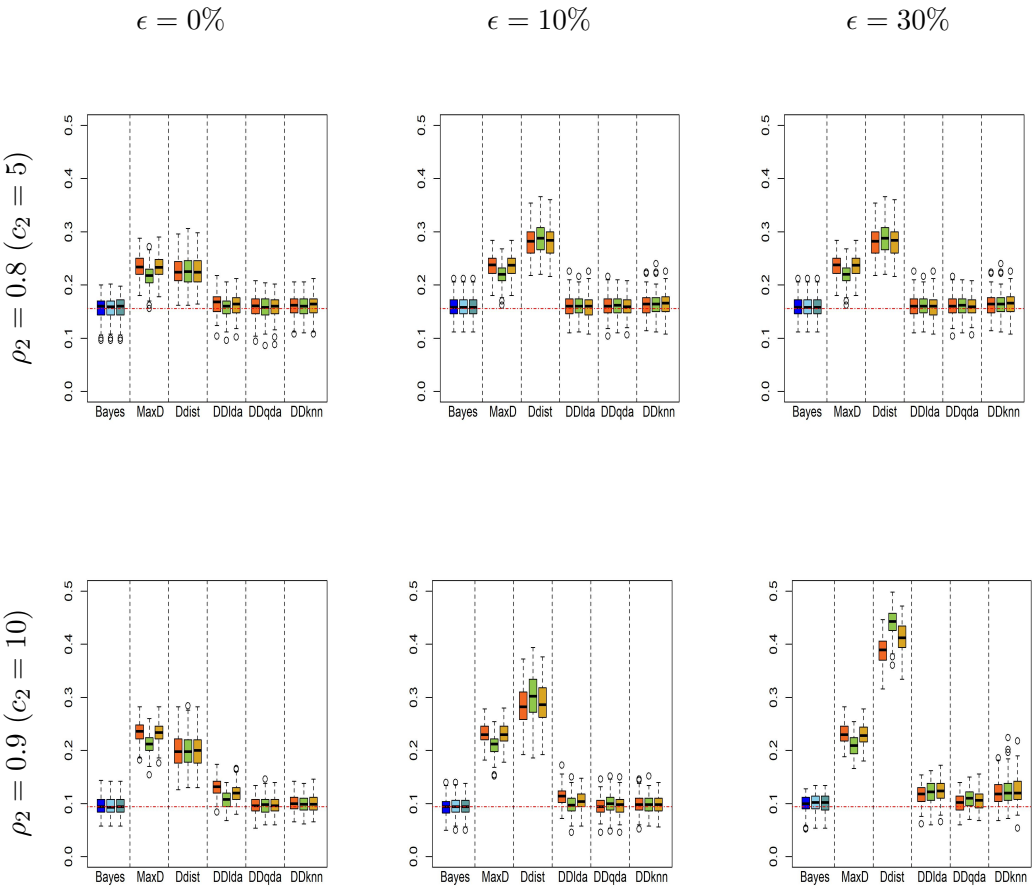


Figure 4.6: Mislabeled data. A percentage ϵ_{21} of observations from the uncontaminated distribution H_2 is added up to the original uncontaminated distribution H_1. Boxplots of misclassification rates of the empirical Bayes, max-depth (MaxD), depth distribution (Ddist) and DD-classifiers (DD). Plots by column: no contamination ($\epsilon = 0$), contamination ($\epsilon = 10\%, 30\%$). Plots by row: concentration of the uncontaminated group H_2 $c_2 = 5$ ($\rho_2 = 0.8$), $c_2 = 10$ ($\rho_2 = 0.9$). Within each plot each graph-box refers to the empirical Bayes, MaxD, Ddist and DD classifiers (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn), respectively. The blue box-plot refers to the empirical Bayes, the sky blue box-plot refers to the robust Bayes type 0 estimators and the cadet blue box-plot refers to the robust Bayes type 1 estimators. The orange box-plots refer to the cosine depth, the green ones to the chord depth and the yellow ones to the arc distance depth. The horizontal dashed red line gives the average misclassification rate of the theoretical Bayes under no contamination.

generally outperform or perform equivalently to the case of the linear discriminant analysis (lda) separating function (see Appendix A).

Effect of mean resultant length of the uncontaminated distribution H_2. When the mean resultant length of the group from H_2 is lower, the misclassification rate is higher. This is because

the higher is the mean resultant length, the more concentrated are the data within the group, and thus they can be better discriminated with respect to the first group, regardless of the effect of the contamination on this latter (Figure 4.3-4.5).

Comparison with the empirical Bayes classifiers. Under no contamination, the performance of the DD-classifier is equivalent to the performance of the empirical maximum likelihood Bayes and robust Bayes classifiers. In case of contamination, the DD-classifier is comparable to the empirical Bayes classifiers. In particular, under antipodality, orthogonality and mean shift outlier model, the best overall performance is achieved by the empirical maximum likelihood Bayes, robust Bayes and the DD classifiers. Considering the case of mislabeled data, the DD-classifier and the empirical Bayes classifiers also provide the overall best performance in terms of average misclassification rates. The robust empirical Bayes classifiers show a comparable behavior to the empirical (maximum likelihood) Bayes and to the DD-classifier in almost all the settings and they are slightly better within the case of mean shift outlier model (although not substantially better).

4.6 Final remarks

The unique feature of directional data asks for dedicated studies. In particular, robustness of directional methods needs to be evaluated apart.

To exemplify, in the standard linear data domain, the further are the outliers from the data center, the larger is their impact on the estimate of the mean. In the directional domain, if a bunch of anomalous data is located at the furthest point from the mean of the uncontaminated data set (i.e. they are located at the point antipodal to the mean), their effect in terms of bias on the directional mean estimate can be negligible, if not null (unless outliers are in great proportion and much more concentrated than the uncontaminated data; see the discussion in Section 4.3 in the present work).

For this reason, this work investigated the robustness of some directional depth-based classifiers under both class and attribute noise, a contribution not yet available in the literature. A comparison of the performance of the max-depth classifier, the depth distribution classifier and the DD-classifier (associated with different separating functions) with respect to the empirical Bayes rules was thus provided. For a further comparison, the empirical Bayes was evaluated both under maximum likelihood and robust estimators of the parameters.

With that aim, this work also introduces and investigates three contamination schemes

which are specific for the case of directional data under class noise. Under the first scheme, the mean of the contaminated distribution is antipodal to the mean of the uncontaminated distribution; under the second, the two directional means are orthogonal; under the third, we have a directional mean shift outlier model.

The main findings of our study follow. When there is no contamination in the training set, the DD-classifier perform equivalently well with respect to the empirical Bayes classifiers, regardless of the choice of the depth function, and if the discriminant rule adopted within the DD-plot is the quadratic or the $k$nn separator. On the other hand, the DD-classifier definitely shows better performances with respect to the max-depth and the empirical depth classifiers (unless the distributions differ only in location, where they perform equally well). This first finding may suggest as further work to investigate if the DD-classifier enjoys any optimality property under rotationally symmetric distributions.

In case of contamination, empirical results suggest the DD-classifier has again merits with respect to the max-depth and the depth distribution classifiers. Furthermore, the DD-classifier is fairly comparable with respect to the empirical Bayes rule both in terms of center and variability of the misclassification rate distributions (the center being the average misclassification rate). That is, DD and Bayes classifiers enjoy a robustness level which is generally equivalent both in terms of average than in terms of reliability of the results.

More specifically, in the presence of outliers in the training set, the DD-classifier and the empirical Bayes yield the best in terms of average misclassification rate within all the examined class attribute contamination settings (antipodality, orthogonality, mean shift outlier model). These results are consistent with those obtained in the linear domain. Li et al. (2012) found indeed that the empirical Bayes and the DD-classifier are more robust if compared with a broad range of classifiers for Euclidean data. The same level of robustness is not achieved by the max-depth and the depth distribution classifiers.

In case of mislabeled data, the best classification accuracy is still obtained by the DD and the Bayes classifier, while the performance of the depth distribution deteriorates in the presence of a high level of class noise. The max-depth seems to be less affected by mislabeling under some scenarios, although it is not in general a real competitor to the other depth-based classifiers.

As for the depth function to be adopted, the choice is a minor issue also in case of contamination. The quadratic discriminant rule and the k-nn should be preferred under contamination as well.

Finally, from a data analysis viewpoint, the following recommendations arise. If the training data are not contaminated and they clearly follow some well known parametric directional distribution, then use the Bayes rule. If some attribute or class noise is suspected, and the group directional distributions are known, use the Bayes rule as well. If no information about the distributions is available, use the DD-classifier: it is a non-parametric technique which works equally well under both contaminated and non-contaminated von Mises-Fisher, probably not so badly under other distributional set up too.

On the other hand, the current study focused on the von Mises-Fisher distribution, the main rotational symmetric distribution on the sphere. It would be definitely of interest to investigate the robustness of depth-based classifiers under non-rotational invariant contaminated and non-contaminated distributions. An issue which is left as further work.

Chapter 5

Directional supervised learning through depth functions: an application to ECG waves analysis

Abstract

Detecting cardiac arrhythmia is important to prevent sudden and untimely deaths. Therefore, the present work investigates arrhythmias from Electrocardiography (ECG) waves. Directional depth-based classifiers are employed to predict the presence of cardiac arrhythmia. A comparison of their performance with respect to the directional Bayes rule is provided.

Keywords: Distance-based depth, directional variables, supervised classification

---

This Chapter has been accepted for publication as: Demni, H. "Directional supervised learning through depth functions: an application to ECG waves analysis". In: Balzano S., Porzio G.C., Salvatore R., Vistocco D., and Vichi M. (eds), *Statistical Learning and Modeling in Data Analysis. Studies in Classification, Data Analysis and Knowledge Organization* Springer, to appear, (2021)
Acknowledgments: Thanks are due to the Editors and the referees of the book *Statistical Learning and Modeling in Data Analysis* for their valuable comments and suggestions.

5.1 Introduction and motivations

Over many decades, linearization were used to explore spherical data by trying to circumvent with their non-linear nature. Then, R. A. Fisher (1953) showed that linear approximations hamper studying some specific phenomena such as the remanent magnetism in sedimentary rocks. Thereafter, several studies have been dedicated to analyze directional data in an appropriate way due to their distinctive properties (e.g. Mardia (1975); Jupp & Mardia (1989)).

The use of directional statistical methods have been motivated by interesting applications in many fields such as astronomy, bioinformatics, neurology, genetics, aeronautics, medicine and machine learning. Here, we focus on the application of directional supervised learning techniques to Electrocardiography (ECG) waves analysis. The aim is to find a function that assigns new patients to either the class of healthy or ill people, based on values obtained from their ECG waves. To this end, the predictive variables in our problem are not treated as linear continuous variables any more, but as directional variables measured in angles.

Within the context of directional supervised classification, new depth-based classifiers have been quite recently introduced: the max-depth classifier (Pandolfo, Paindaveine, & Porzio, 2018), the DD-classifier (Pandolfo, D'Ambrosio, & Porzio, 2018) and the depth distribution classifier (Demni et al., 2019). For these classifiers, both optimality properties and simulation results are available.

Vencálek et al. (2020) derived the conditions under which some of these classifiers are optimal in the Bayes sense. For instance, they found that the max-depth and the depth distribution classifiers are optimal if the underlying distributions are rotationally symmetric, unimodal, differ only in location, and gave equal prior probabilities.

Robustness properties of the max-depth, depth versus depth and depth distribution classifiers were investigated under different contamination schemes in Demni et al. (2020). It came out that the DD-classifier performs better or equivalently to the empirical Bayes while it outperforms the max-depth and the depth distribution classifiers in the presence of noise.

What is still lacking is to evaluate how these depth-based directional classifiers perform on real data. This short note has thus the goal of starting fulfilling this gap. With that aim, this work analyzes the performance of the max-depth, the depth versus depth, and the depth distribution classifiers on a real data set which is well known in the supervised learning

literature. It refers to some arrhythmia data used to discriminate between healthy and ill people. In our study, we focus on the directional predictors which comes from ECG waves. The performance of such classifiers is also compared with the performance of the directional Bayes classifier under the hypothesis of a von-Mises Fisher distribution.

This chapter is organized as follows. Section 5.2 presents the arrhythmia data set, a description of the directional variables and the overall aim of the analysis. In section 5.3, we briefly present the mentioned depth-based classifiers for directional data. Section 5.4 reports results on the performance of the depth-based classifiers when applied to the ECG waves problem. In Section 5.5, some final remarks are offered to the reader.

5.2 The arrhythmia data set

Arrhythmia refers to irregular heartbeats, and it can be evaluated by looking at the electrical activity of the heart, recorded through Electrocardiogram (ECG) waves. Analyzing ECG waves can provide insights on heart health issues. Waves which can be in turn evaluated as angular variables.

The arrhythmia data is one of the data sets available within the UCI Machine Learning Repository (Frank & Asuncion, 2010). It reports the presence of different types of cardiac-arrhythmia from ECG as well as its absence. The original data set contains 452 patients records described by 279 predictive variables (measurements, patient data and ECG recording) and 16 classes: the first refers to normal ECG (healthy patients) while classes 2 to 15 correspond to different types of arrhythmia and class 16 refers to the unclassified patients.

5.2.1 Standard classification methods for Cardiac Arrhythmia

Various Machine learning and data mining methods have been applied for the detection of arrythmia through electrocardiograms (ECG). In this section, we review machine learning methods for the diagnosis of cardiac arrythmia within the context of classification.

For instance, Guvenir et al. (1997) proposed an inductive supervised classification algorithm which is based on the majority of votes among the class predictions made by each feature. They showed that their proposed method provides better performance if compared to other standard methods. Gao et al. (2005) developed a detection system for arrhythmia based on a Bayesian Artificial Neural Network Classifier which is able to deal with missing feature values and unclassified classes. A comparison with a wide range of classifiers such

as naive Bayes, decision trees, logistic regression and radial basis function networks has been also provided.

W. Zuo et al. (2008) introduced a kernel weighted k-nearest neighbor classifier for the diagnosis of cardiac arrhythmia. They considered the multi-class classification problem and they showed their method outperforms both the naive Bayes classifier and the one introduced in Guvenir et al. (1997). An effective automated Artificial Neural Network based system has been also suggested by Jadhav et al. (2010). They showed that the multilayer perceptron (MLP) feedforward neural network model is able to ensure the true estimation of the complex decision boundaries.

5.2.2 Directional classification methods for Cardiac Arrhythmia

Amongst the many, the studies which exploited the arrhythmia data while adopting directional data techniques are considered here. All of them dealt with the waves as angular variables.

First, López-Cruz et al. (2015) proposed an extension to directional data of the naive Bayes classifier and the selective naive Bayes for von-Mises and von-Mises Fisher distributions. They showed their superiority with respect to other versions of the naive Bayes.

Then, Fernandes & Cardoso (2016) introduced a discriminative binary classifier for mixed data (linear and angular) and they showed that their method is competitive to traditional classifiers. More recently, Pernes et al. (2019) proposed several versions of directional support vector machines that supports both angular and linear predictors and compared them to several directional classifiers.

The best average misclassification rate for the arrhythmia data is 0.209 and it was obtained in Pernes et al. (2019) by a directional logistic regression model. However, we note that the performances reported within these three papers are not directly comparable, given that different simulation settings have been adopted within each of them. Furthermore, such performances are not comparable with our case study results as well, where the focus is on the discriminant power of the directional variables on their own.

5.2.3 Scope of the analysis and variables description

In line with the previous mentioned studies (López-Cruz et al., 2015; Fernandes & Cardoso, 2016; Pernes et al., 2019), unclassified samples were removed and the study goal was trans-

Number of dir. variables	Classes	Number of obs.	Number of obs. per class
4	2	430	class 1, normal: 245 class 2, arrhythmia: 185

Table 5.1: Summary of the main characteristics of the data used in this work, including the number of directional (dir.) features and the number of observations (obs.) per class (class 1: normal vs class 2: arrhythmia).

formed into a binary classification problem (normal vs.arrhythmia).

As predictors, the four angular variables characterizing ECG waves are considered. That is, the aim of our study is to discriminate between healthy and non-healthy patients with arrhythmia based on the values obtained from their ECG waves. Table 5.1 summarizes the number of directional variables, the number of classes and the number of observations per class of the evaluated dataset.

The angular variables characterize the vector angles from the front plane of four ECG waves and they are measured in degrees in the original data set. The P-wave reflects the atrial depolarization, the QRS-wave represents the depolarization of the ventricles, the T-wave describes the rapid re-polarization of contractile cells while the QRST-wave corresponds to the global ventricular re-polarization.

By looking at the rose diagram of each observed distribution separately for the groups of healthy and ill people, we saw they are unimodal. Hence, their distribution can be properly investigated by means of circular boxplots (Buttarazzi et al., 2018), which are here represented in Figure 5.1.

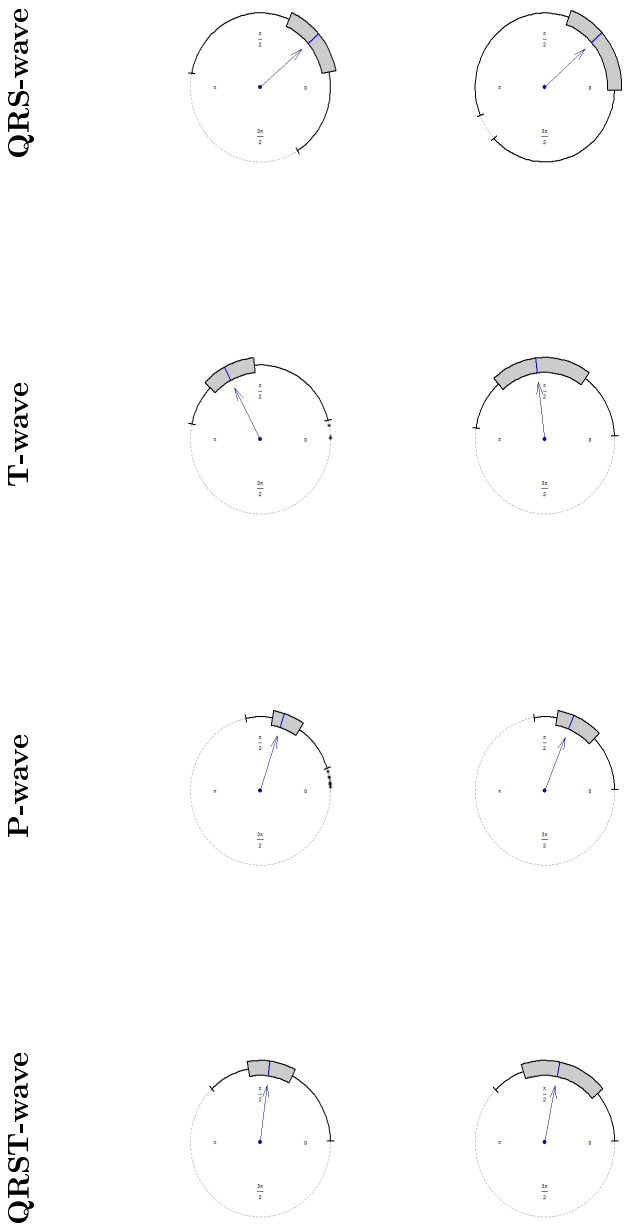


Figure 5.1: Circular box-plots of the angular variables exploited in this study. By column: healthy patients (left) and patients with arrhythmia (right). By row: QRS-wave, T-wave, P-wave and QRST-wave.

We note that the distribution of the QRS-wave angles span over more than half a circle, while all the others have angles in $(0, \pi)$. As a consequence, they can be mapped

into a 4-dimensional hyper-sphere embedded in a $5D$ space. Directional supervised learning procedures act directly on such a hyper-sphere.

Some of the observed marginal distributions are substantially symmetric, others are clearly asymmetrically distributed (e.g. the P-wave angles and, to a certain extent, the QRST waves). Looking at differences between the two groups, the T-waves seems to have the higher marginal ability to discriminate.

5.3 Directional depth-based supervised learning techniques

In this section, the three main directional depth-based supervised classification methods are briefly reviewed: the max-depth classifier, the depth versus depth classifier (DD-classifier) and the depth distribution classifier.

Considering K empirical distributions $\hat{H}_i$, $i = 1, .., K$, the directional max-depth classifier is given by

$$class_{max}(x) := \text{argmax}_i \, D(x; \hat{H}_i),$$

where $x \in S^{(q-1)}$ is a new observation to be classified and, $D(x, \hat{H}_i), i = 1, .., K$ is the empirical depth of x with respect to the directional empirical distributions $\hat{H}_1, .., \hat{H}_K$, respectively.

The directional DD-classifier is a generalization of the max-depth classifier and it is given by

$$class_{DD}(x) := \underset{i}{\text{argmax}} \, r(D(x; \hat{H}_i)), \tag{5.1}$$

where $r(.)$ is a real increasing function which has the aim of well separate points in the depth versus depth space (DD-plot). Different choices have been considered for $r(.)$. Li et al. (2012) suggested to consider a polynomial discriminating function, whose degree have to be estimated, while Mosler & Mozharovskyi (2017) adopted a $k$nn decision rule.

The directional depth distribution classifier is given by

$$class_{dd}(x) := \text{argmax}_i \, F_D(x, \hat{H}_i),$$

with

$$F_D(x, \hat{H}_i) := P(D(X, \hat{H}_i) \leq D(x, \hat{H}_i)),$$

where $D(x, \hat{H}_i)$, $i = 1, .., K$ is the empirical depth of x with respect to the empirical distributions $\hat{H}_1, .., \hat{H}_K$, respectively, and hence $F_D(., \hat{H}_i)$ is the cdf of the depth function under $\hat{H}_i$.

For each classifier, a depth function must be adopted. Here, distance-based depth functions are considered. They are defined as follows (Pandolfo, Paindaveine, & Porzio, 2018):

- The cosine depth: $D_{cos}(x, H) = 2 - E_H[(1 - x'X)]$;
- The arc distance depth: $D_{arc}(x, H) = \pi - E_H[\arccos(x'X)]$;
- The chord depth: $D_{chord}(x, H) = 2 - E_H[\sqrt{2(1 - x'X)}]$.

Here, $x \in S^{(q-1)}$ is a point whose depth is evaluated with respect to the directional distribution H, $E[.]$ is the expected value, and X is a random variable from H. The empirical depth is obtained by replacing H by $\hat{H}$ for each depth function.

Finally, for the sake of completeness, we recall how the empirical Bayes classifier is defined. We have:

$$class_{Bayes}(x) := \text{argmax}_i \, \hat{h}_i(x) p_i$$

where p_i is the prior probability corresponding to the distribution H_i, $i = 1, .., K$, and $\hat{h}_i()$ is the estimated assumed density for the i-th group. In directional supervised learning, the Bayes classifiers has been used with the $h_i()$'s being von Mises-Fisher densities with different location and concentration parameters (López-Cruz et al., 2015).

5.4 Performance of depth-based classifiers on ECG-waves

As discussed, the aim of this study is to evaluate the performance of depth-based classifiers on a set of real data arising from an ECG analysis. With that goal, the angular variables were transformed to their Euclidean coordinates (units vectors) and a simulation study was performed. In line with the existing literature (López-Cruz et al., 2015; Fernandes & Cardoso, 2016; Pernes et al., 2019), a 3-fold stratified cross-validation method where the percentage of samples for each class is preserved was considered. The experiment was repeated 100 times.

Ten different possible solutions were evaluated and compared. Each of the three mentioned classifiers was combined with three different directional depth functions (cosine, chord,

Classifier		AMR	Average macro F_1-score
Empirical Bayes		0.36	0.63
Max-depth	cosine	0.40	0.60
	chord	0.39	0.60
	arc-distance	0.42	0.52
Depth distribution	cosine	0.36	0.63
	chord	0.35	0.64
	arc-distance	0.35	0.63
DD-plot with $k$nn	cosine	0.35	0.64
	chord	**0.33**	**0.67**
	arc-distance	0.34	0.66

Table 5.2: Average misclassification rate (AMR) and average macro F_1-score of the Bayes, max-depth, depth distribution and DD-classifiers when associated to the cosine, chord, and arc distance depth functions. Best achieved results are highlighted in bold.

arc-distance), and all of them were compared against the empirical Bayes classifier under the von-Mises Fisher assumption.

For the $r(.)$ function in Eq.5.1, the k-Nearest Neighborhood ($k$nn) discriminant rule has been adopted in line with (Pandolfo, Paindaveine, & Porzio, 2018; Demni et al., 2020), with the tuning parameter k chosen by cross validation. The performance of the classifiers was evaluated by means of the misclassification rate which is the number of misclassified observations over the sample size in each replicated sample, and by the macro F_1-score which is the unweighted mean value of the individual F_1-scores of each class.

The distribution of the misclassification rates obtained by the max-depth, depth distribution, DD and Bayes classifiers when associated with different distance based-depth functions are here provided through box-plots (Figure 5.2) and summarized through the average misclassification rates (Table 5.2). The macro F_1-scores of the directional classifiers are also given in Table 5.2.

Although the two classes are imbalanced, the average macro F_1-score and the average accuracy of the classifiers are consistent with each other: the best classifier in terms of average accuracy is the best classifier in terms of average macro F_1-score too (Table 5.2).

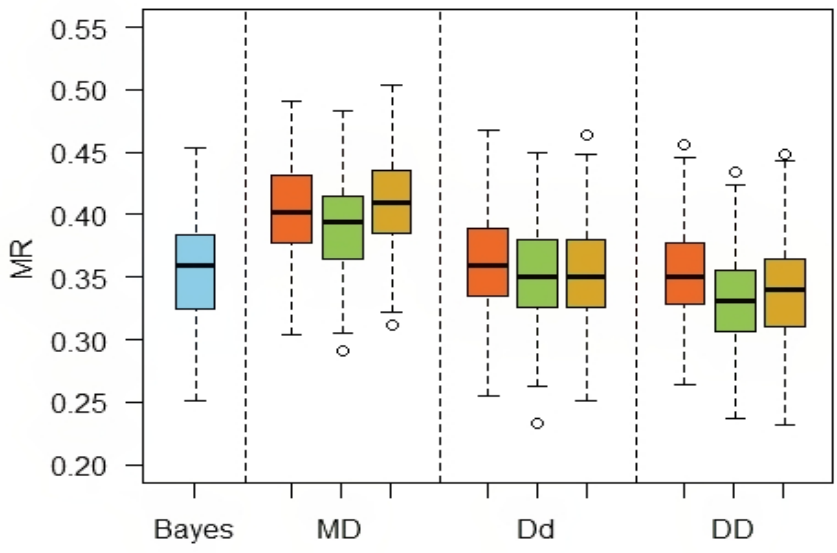


Figure 5.2: Box-plots of misclassification rates (MR) of the Bayes, max-depth (MD), depth distribution (Dd) and DD-classifiers (DD). In each graph-box (excluding the Bayes), the most left box-plot refers to the cosine depth, the middle one to the chord depth and the most right to the arc distance depth. The best performance is achieved by the DD-classifier associated with the chord depth.

The DD-classifier achieves the best overall performance in terms of average misclassification rate (Figure 5.2, most right graph-box). Furthermore, it performs better than the Bayes rule independently from the choice of the depth function. The depth distribution and the DD classifiers performs equivalently to the empirical Bayes classifier (if not slightly better). The worst performance is given by the max-depth classifier.

In general, the choice of the depth function seems not to be particularly influential on the performance of the three classifiers, although some small differences arise. In addition, by looking at the confusion matrix of the classifiers, it appears that it is in general more difficult to classify patients with arrhythmia. The higher proportion of misclassified observations arises indeed from class 2 (observations are wrongly assigned to class 1 while they are coming from class 2).

5.5 Final remarks

In this chapter, directional distance-based depth classifiers were applied to some arrhythmia data in order to distinguish between the presence or absence of cardiac diseases. We investigated the performance of the max-depth, depth distribution, depth versus depth classifiers and the Bayes rule. Angular variables arising from ECG recordings were considered.

In directional supervised learning, the standard Bayes rule assumes data in each group come from a von-Mises Fisher distribution. If so, the Bayes rule yields the best available discriminant procedure. On the other hand, real data not necessarily fulfill such or any other parametric assumption. This is why it is always of interest to compare the performance of new methods against the Bayes rule on specific fields of application.

On the considered data, we had that the DD-classifier largely outperforms the max-depth and it performs better than the depth distribution and the empirical Bayes classifier. On the other hand, the performance of the depth distribution classifier is equivalent to the Bayes rule, and the max-depth classifier definitely provides the worst behavior over all the considered methods.

As further research, we see the necessity of developing new depth-based methods which combine both linear and directional variables to fully exploit the information available within the data set. It would be also of interest to test the discussed directional supervised learning methods on other real data applications within this field.

Conclusions

A supervised classification method exploits a labeled training data set to classify a new data point by assigning it to one of the labeled groups. With that aim, given two or more labeled data groups, data depth functions can be adopted for solving classification tasks. These functions measure the centrality of a data point with respect to each group.

This work analyzed the use of depth functions to obtain classifiers within a directional domain. Directional observations are data points lying on the boundary of circles, spheres and their extensions. For this kind of data, standard statistical techniques are not suitable. Within the following, we summarize our main contributions and we report the main perspectives and future works.

Main contributions

This work contributes to the existing literature in many ways. First, it introduces a new directional depth distribution classifier based on the cumulative distribution of the cosine depth function. Under different scenarios, experimental results suggest that the cosine depth distribution classifier is an improvement over the existing max-depth classifier. Both classifiers are well suited for multiple classes problems.

The second contribution concerns studying the optimality of the introduced cosine depth distribution and the cosine max-depth classifiers. Conditions under which the aforementioned classifiers are optimal in the sense of the Bayes rule are discussed. It is formally proved that both classifiers achieve Bayes optimality when group distributions are rotationally symmetric, unimodal, have equal priors and differ only in location. It is also proved that the cosine max-depth classifier is optimal when distributions differ in both location and dispersion and group distributions belong to a specific distribution family.

Furthermore, robustness of directional depth-based supervised classifiers is deeply in-

vestigated. According to the recent literature, two contamination sources are considered: attribute and label noise. For the first case, we examine conditions under which the mean of the contaminated distribution is antipodal or orthogonal to the mean of the uncontaminated distribution. The directional mean shift outlier model is also considered. Through an extensive simulation study, performances in terms of robustness of the max-depth classifier, the DD-classifier and the depth distribution classifier were compared against the standard and some robust Bayes classifiers. We adopted three distance depth functions to be associated with the classifiers. Results show that the directional DD-classifier performs well when dealing with noise in many different settings. Some recommendations from a data analysis point of view were also provided.

Finally, a new application of the aforementioned methods was considered. Directional depth based-classifiers were exploited to identify groups of patients based on their Electrocardiogram waves. Results show that the DD-classifier yields the best in term of average misclassification rate.

Future works

This work opens many and diverse perspectives for future research. From a theoretical point of view, it would be of interest to study properties and optimality conditions of the directional DD-classifier. Our empirical results suggest that it can achieve Bayes optimality under a large class of distributions.

It would be also of interest to assess the performance of depth-based classifiers not only when they are combined with distance-based depth functions. Other directional depth functions are indeed available. However, they have not been considered here because of their prohibitive computational cost. In case new algorithms should be made available, their adoption can even improve the achieved performances. Another line of research is to investigate the robustness of depth-based classifiers under non-rotational symmetric distributions, such as the Fisher-Bingham distributions.

Finally, while an extremely recent attempt is available in Pandolfo & D'Ambrosio (2021), investigating the performance of directional depth-based classifiers on some other real data applications is certainly worthy. On the other hand, the application examined in this thesis work suggests the need to develop new depth-based classifiers that are able to deal with mixed data (linear and directional variables): emerging field with many possible potential

applications.

Appendix A

In this Appendix, we report the average misclassification rates of the empirical Bayes classifiers (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the depth distribution and the DD classifiers (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn), respectively, associated with the cosine depth, the chord and the arc distance depths for the robustness study conducted in Chapter 4. We consider three outlier contamination scenarios: antipodality and orthogonality of the contaminated distribution mean, and the directional mean shift outlier model and one mislabeled data case.

Table A.1: Antipodality. Average misclassification rates AMR of the empirical Bayes (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the DD-classifier (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn) and the depth distribution associated with the cosine depth, the chord and the arc distance depths when the mean resultant length of the uncontaminated group H_2 $\rho_2 = 0.8$ ($c_2 = 5$), best achieved results for each contamination level are in bold.

Setup		contamination level		
$\rho_2 = 0.8, c_2 = 5$		0%	10%	30%
Empirical Bayes	ML	**0.157**	**0.163**	0.168
	Robust type 0	**0.157**	**0.163**	0.170
	Robust type 1	**0.157**	**0.163**	0.170
Max-depth	Cosine	0.234	0.250	0.267
	Chord	0.216	0.231	0.246
	Arc distance	0.232	0.250	0.268
Depth distribution	Cosine	0.227	0.249	0.267
	Chord	0.227	0.244	0.265
	Arc distance	0.227	0.249	0.267
DD-plot with lda	Cosine	0.165	0.173	0.179
	Chord	0.159	0.164	0.171
	Arc distance	0.163	0.169	0.177
DD-plot with qda	Cosine	0.159	0.166	**0.165**
	Chord	0.158	0.165	0.169
	Arc distance	0.159	0.166	0.166
DD-plot with $k$nn	Cosine	0.161	0.167	0.168
	Chord	0.160	0.166	0.167
	Arc distance	0.162	0.166	0.167

Table A.2: Antipodality. Average misclassification rates AMR of the empirical Bayes (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the DD-classifier (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn) and the depth distribution associated with the cosine depth, the chord and the arc distance depths when the mean resultant length of the uncontaminated group H_2 $\rho_2 = 0.9$ ($c_2 = 10$), best achieved results for each contamination level are in bold.

Setup		contamination level		
$\rho_2 = 0.9, c_2 = 10$		0%	10%	30%
Empirical Bayes	ML	**0.095**	**0.095**	**0.098**
	Robust type 0	**0.095**	**0.095**	0.099
	Robust type 1	**0.095**	**0.095**	0.099
Max-depth	Cosine	0.235	0.241	0.259
	Chord	0.212	0.216	0.232
	Arc distance	0.233	0.240	0.261
Depth distribution	Cosine	0.199	0.230	0.239
	Chord	0.198	0.220	0.231
	Arc distance	0.199	0.229	0.238
DD-plot with lda	Cosine	0.131	0.133	0.140
	Chord	0.107	0.107	0.116
	Arc distance	0.120	0.120	0.129
DD-plot with qda	Cosine	0.096	0.096	0.099
	Chord	0.097	0.097	0.101
	Arc distance	0.096	0.096	0.099
DD-plot with $k$nn	Cosine	0.100	0.100	0.104
	Chord	0.099	0.099	0.101
	Arc distance	0.100	0.100	0.102

Table A.3: Orthogonality. Average misclassification rates AMR of the empirical Bayes (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the DD-classifier (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn) and the depth distribution associated with the cosine depth, the chord and the arc distance depths when the mean resultant length of the uncontaminated group H_2 $\rho_2 = 0.8$ ($c_2 = 5$), best achieved results for each contamination level are in bold.

Setup		contamination level		
$\rho_2 = 0.8, c_2 = 5$		0%	10%	30%
Empirical Bayes	ML	**0.157**	**0.160**	0.169
	Robust type 0	**0.157**	**0.160**	0.171
	Robust type 1	**0.157**	**0.160**	0.171
Max-depth	Cosine	0.234	0.240	0.254
	Chord	0.216	0.225	0.243
	Arc distance	0.232	0.239	0.253
Depth distribution	Cosine	0.227	0.187	0.167
	Chord	0.227	0.187	**0.164**
	Arc distance	0.227	0.186	0.168
DD-plot with lda	Cosine	0.165	0.172	0.187
	Chord	0.159	0.163	0.170
	Arc distance	0.163	0.170	0.185
DD-plot with qda	Cosine	0.159	0.163	0.169
	Chord	0.158	**0.160**	0.172
	Arc distance	0.159	0.162	0.170
DD-plot with $k$nn	Cosine	0.161	0.164	0.170
	Chord	0.160	0.163	0.168
	Arc distance	0.162	0.164	0.169

Table A.4: Orthogonality. Average misclassification rates AMR of the empirical Bayes (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the DD-classifier (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn) and the depth distribution associated with the cosine depth, the chord and the arc distance depths when the mean resultant length of the uncontaminated group H_2 $\rho_2 = 0.9$ ($c_2 = 10$), best achieved results for each contamination level are in bold.

Setup		contamination level		
$\rho_2 = 0.9, c_2 = 10$		0%	10%	30%
Empirical Bayes	ML	**0.095**	**0.095**	0.101
	Robust type 0	**0.095**	**0.095**	0.103
	Robust type 1	**0.095**	**0.095**	0.102
Max-depth	Cosine	0.235	0.241	0.251
	Chord	0.212	0.216	0.235
	Arc distance	0.233	0.240	0.250
Depth distribution	Cosine	0.199	0.230	0.105
	Chord	0.198	0.220	0.102
	Arc distance	0.199	0.229	0.107
DD-plot with lda	Cosine	0.131	0.133	0.159
	Chord	0.107	0.107	0.109
	Arc distance	0.120	0.120	0.140
DD-plot with qda	Cosine	0.096	0.096	**0.100**
	Chord	0.097	0.097	0.101
	Arc distance	0.096	0.096	0.101
DD-plot with $k$nn	Cosine	0.100	0.100	0.102
	Chord	0.099	0.098	0.101
	Arc distance	0.100	0.100	0.102

Table A.5: Mean shift outlier model. Average misclassification rates AMR of the empirical Bayes (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the DD-classifier (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn) and the depth distribution associated with the cosine depth, the chord and the arc distance depths when the mean resultant length of the uncontaminated group H_2 $\rho_2 = 0.8$ ($c_2 = 5$), best achieved results for each contamination level are in bold.

Setup $\rho_2 = 0.8, c_2 = 5$		contamination level 0%	10%	30%
Empirical Bayes	ML	**0.028**	0.032	0.038
	Robust type 0	**0.028**	**0.028**	**0.029**
	Robust type 1	**0.028**	**0.028**	0.032
Max-depth	Cosine	0.039	0.034	0.037
	Chord	0.037	0.033	0.033
	Arc distance	0.038	0.033	0.035
Depth distribution	Cosine	0.031	0.036	0.046
	Chord	0.031	0.035	0.042
	Arc distance	0.031	0.035	0.043
DD-plot with lda	Cosine	0.031	0.031	0.036
	Chord	0.030	0.031	0.035
	Arc distance	0.030	0.031	0.035
DD-plot with qda	Cosine	**0.028**	0.032	0.037
	Chord	0.028	0.031	0.036
	Arc distance	0.029	0.031	0.036
DD-plot with $k$nn	Cosine	0.029	0.031	0.038
	Chord	0.029	0.031	0.035
	Arc distance	0.029	0.030	0.037

Table A.6: Mean shift outlier model. Average misclassification rates AMR of the empirical Bayes (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the DD-classifier (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn) and the depth distribution associated with the cosine depth, the chord and the arc distance depths when the mean resultant length of the uncontaminated group H_2 $\rho_2 = 0.9$ ($c_2 = 10$), best achieved results for each contamination level are in bold.

Setup $\rho_2 = 0.9, c_2 = 10$		contamination level 0%	10%	30%
Empirical Bayes	ML	**0.009**	0.011	0.015
	Robust type 0	**0.009**	**0.009**	**0.010**
	Robust type 1	**0.009**	0.010	0.012
Max-depth	Cosine	**0.009**	0.010	0.028
	Chord	**0.009**	0.010	0.019
	Arc distance	**0.009**	0.010	0.024
Depth distribution	Cosine	**0.009**	0.023	0.030
	Chord	0.010	0.023	0.024
	Arc distance	0.010	0.023	0.026
DD-plot with lda	Cosine	**0.009**	0.010	0.015
	Chord	**0.009**	0.010	0.012
	Arc distance	**0.009**	0.010	0.013
DD-plot with qda	Cosine	**0.009**	0.011	0.013
	Chord	**0.009**	0.011	0.012
	Arc distance	**0.009**	0.011	0.013
DD-plot with $k$nn	Cosine	0.010	0.010	0.013
	Chord	0.010	0.010	0.012
	Arc distance	0.010	0.010	0.012

Table A.7: Mislabeled case. Average misclassification rates AMR of the empirical Bayes (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the DD-classifier (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn) and the depth distribution associated with the cosine depth, the chord and the arc distance depths when the mean resultant length of the uncontaminated group H_2 $\rho_2 = 0.8$ ($c_2 = 5$), best achieved results for each contamination level are in bold.

Setup		contamination level		
$\rho_2 = 0.8, c_2 = 5$		0%	10%	30%
Empirical Bayes	ML	**0.157**	**0.159**	**0.162**
	Robust type 0	**0.157**	**0.159**	**0.162**
	Robust type 1	**0.157**	**0.159**	**0.162**
Max-depth	Cosine	0.234	0.236	0.235
	Chord	0.216	0.219	0.219
	Arc distance	0.232	0.235	0.234
Depth distribution	Cosine	0.227	0.280	0.374
	Chord	0.227	0.287	0.406
	Arc distance	0.227	0.281	0.381
DD-plot with lda	Cosine	0.165	0.173	0.179
	Chord	0.159	**0.159**	0.204
	Arc distance	0.163	0.160	0.205
DD-plot with qda	Cosine	0.159	0.160	0.163
	Chord	0.158	0.161	0.170
	Arc distance	0.159	0.160	0.165
DD-plot with $k$nn	Cosine	0.161	0.163	0.213
	Chord	0.160	0.164	0.211
	Arc distance	0.162	0.164	0.214

Table A.8: Mislabeled case. Average misclassification rates AMR of the empirical Bayes (maximum likelihood (ML), robust type 0, robust type 1 estimators), the max-depth, the DD-classifier (according to the considered classification rules in the DD-space, i.e. lda, qda and $k$nn) and the depth distribution associated with the cosine depth, the chord and the arc distance depths when the mean resultant length of the uncontaminated group H_2 $\rho_2 = 0.9$ ($c_2 = 10$), best achieved results for each contamination level are in bold.

Setup		contamination level		
$\rho_2 = 0.9, c_2 = 10$		0%	10%	30%
Empirical Bayes	ML	**0.095**	**0.095**	**0.099**
	Robust type 0	**0.095**	**0.095**	0.100
	Robust type 1	**0.095**	**0.095**	0.100
Max-depth	Cosine	0.235	0.232	0.231
	Chord	0.212	0.210	0.210
	Arc distance	0.233	0.231	0.229
Depth distribution	Cosine	0.199	0.284	0.388
	Chord	0.198	0.304	0.440
	Arc distance	0.199	0.290	0.412
DD-plot with lda	Cosine	0.131	0.113	0.117
	Chord	0.107	0.097	0.121
	Arc distance	0.120	0.104	0.123
DD-plot with qda	Cosine	0.096	0.096	0.101
	Chord	0.097	0.099	0.108
	Arc distance	0.096	0.097	0.105
DD-plot with $k$nn	Cosine	0.100	0.100	0.121
	Chord	0.099	0.099	0.124
	Arc distance	0.100	0.100	0.124

Appendix B

In this Appendix, the main R functions which were used within the thesis are reported. For completeness, some of the codes were reproduced. Thanks are due to Giuseppe Pandolfo for providing the cosine depth, the chord depth and the arc distance depth functions as well as to Davide Buttarazzi for sharing his function to draw the spherical density on the sphere.

Code details in R language

R code to draw data on the circle and their corresponding mean direction

```
1 library (circular) # install.packages("circular", dep=T)
2 #Generate 100 observations from a von Mises distribution
3 # with mean direction pi/2  and concentration 5
4 x <- rvonmises(n=100, mu=circular(pi/2), kappa=5)
5 # Plot the data
6 plot(x, stack=TRUE, bins=150)
7 # Plot the circular mean
8 arrows.circular(mean.circular(x))
```

R code to draw the Density plot of circular data

```
1 library (circular) # install.packages("circular", dep=T)
2 #Generate 100 observations from a von Mises distribution
3 # with mean direction pi  and concentration 5
4 y <- rvonmises(n=100, mu=circular(pi), kappa=5)
5 Density <- density((y), bw=25)
6 plot(Density, points.plot=TRUE, xlim=c(-1.5,1), col=2)
```

R code to draw Fisher-Bingham data on the sphere

```
require (simd) # install.packages("simd", dep=T)
#Generate 1000 observations from a Kent distribution
# with mean direction (1,0,0) and concentration 80
X_bFB5=rFisherBingham(1000,c(80,0,0),c(0,30,-30))

#Generate 1000 observations from a the extreme FB5 distribution
# with mean direction (1,0,0) and concentration 80
X_eFB5=rFisherBingham(1000,c(80,0,0),c(0,40,-40))

require(sphereplot) # install.packages("sphereplot", dep=T)
# converting into spherical coordinates
Xsph_bFB5<-car2sph(X_bFB5[,1],X_bFB5[,2],X_bFB5[,3],deg=TRUE)

require(rgl) # install.packages("rgl", dep=T)

# 3D sphere plot
rgl.sphgrid(radius = 1, add =TRUE,col.long=1, col.lat=1,longtype =D
    )
#grid : adding points
rgl.sphpoints(Xsph_bFB5,deg=TRUE,col=1,cex=3.5)
```

R code to draw the spherical Density plot

```
library(ggplot2)      # For plotting
library(cowplot)      # grid arrangement of plots
library(Directional) # For spherical density functions
library(maps)         # vector maps of the world
library(hrbrthemes)  # hrbrmstr themes
library(magick)       # For animation
library(mapproj)      # Needed for projection

# Generate data from von Mises starting from latitude and longitude
```

```
10 random_points <- function(n_points, lat, lon, concentration) {
11   # Directional defines lat + long as 0-180 and 0-360 respectively
12   # we have to shift back and forth
13   mu <- euclid(c(lat + 90, lon + 180))[1,]
14   pts <- euclid.inv(rvmf(n_points, mu, concentration))
15   pts[,1] <- pts[,1] - 90
16   pts[,2] <- pts[,2] - 180
17   data.frame(pts)
18 }
19
20 # Make a density grid to be plotted
21 # The arguments of the function vmf.kerncontour can be changed
22 vmf_density_grid <- function(u, ngrid = 100) {
23   # Translate to (0,180) and (0,360)
24   u[,1] <- u[,1] + 90
25   u[,2] <- u[,2] + 180
26   res <- vmf.kerncontour(u, thumb = "none", den.ret = T, full =
       TRUE,
27                          ngrid = ngrid)
28
29   # Translate back to (-90, 90) and (-180, 180) and create a grid
       of coordinates
30   ret <- expand.grid(Lat = res$lat - 90, Long = res$long - 180)
31   ret$Density <- c(res$den)
32   ret
33 }
34
35 # Generate spherical data (sample size, latitude of mean, longitude
      of mean, concentration)
36 # Here we generate a bimodal distribution
37 x0 <- random_points(100, 0, 0, 15)
38 x1 <- random_points(100, 45, 0, 15)
39 x <- rbind.data.frame(x0,x1)
```

```
40
41 # Graphical parameter to be used later
42 no.axis <- theme(axis.ticks.y = element_blank(), axis.text.y =
       element_blank(),
43                  axis.ticks.x = element_blank(), axis.text.x =
                        element_blank(),
44                  axis.title.x = element_blank(), axis.title.y =
                        element_blank())
45
46 # Define plot in bivariate space
47 myplot <- ggplot(x, aes(x = Long, y = Lat)) +
48           scale_y_continuous(breaks = (-2:2) * 30, limits = c
                (-90,90)) + scale_x_continuous(breaks = (-4:4) * 45,
                limits = c(-180, 180))
49           +geom_point(size = 1)
50           +geom_contour(data = vmf_density_grid(x, ngrid = 300),
51                         aes(x=Long, y=Lat, z=Density), color = "red"
                              )
52
53 # Plot bivariate myplot
54 # Plot on a hemisphere starting from the bivariate setting (you
       should choose the orientation)
55 # Orientation = An optional vector c(latitude,longitude,rotation)
56 # which describes where the "North Pole" should be when computing
       the projection.
57 # Orientation could be set as the mean direction
58 ortho.projections <- plot_grid(
59   myplot + coord_map("ortho", orientation = c(0,0,0)) + no.axis,
60       labels = NULL, align = 'h')
61
62 ortho.projections
```

Directional depth functions

```
### Cosine depth function ###
# arguments must be in cartesian coordinates
CosDepth <- function(X, Y){
  # computing the cosine dissimilarity
  cos_dis <- function(ma, mb){
    mat=tcrossprod(ma, mb)
    t1=sqrt(apply(ma, 1, crossprod))
    t2=sqrt(apply(mb, 1, crossprod))
    sim <- 1-(mat / outer(t1,t2))
    return(sim)}
  M<-cos_dis(X,Y)
  # computing the cosine distance depth
  mat <- apply(M, 1, mean)
  res <- 2 - mat
  return(res)
}
```

```
### Arc distance depth function ###
sArcDeptht <- function(x,X){
  # computing the arc distance length
  spherdist <- acos(x%*%t(X))
  options(warn=-1)
  spherdist[is.nan(spherdist)] = 0
  # computing the arc distance depth
  mat <- apply(spherdist,1,mean)
  res <- pi - mat
  return(res)
}
```

```
### Chord depth function in dimension 3 ###
sChordDepth3D <- function(X, Y){
  # computing chord distance
```

```
4    M <- matrix(NA, nrow = nrow(X), ncol = nrow(Y))
5    for(i in 1:nrow(X)){
6      for(j in 1:nrow(Y)){
7        M[i,j] <- sqrt(((X[i,1]-Y[j,1])^2)+((X[i,2]-Y[j,2])^2)
8         +((X[i,3]-Y[j,3])^2))
9      }
10   }
11   # computing the chord distance depth
12   mat <- apply(M, 1, mean)
13   res <- 2 - mat
14   return(res)
15 }
```

```
1  ### Chord depth function in 10 dimensions ###
2  sChordDepth10D <- function(X, Y){
3    # computing the chord distance
4    M <- matrix(NA, nrow = nrow(X), ncol = nrow(Y))
5    for(i in 1:nrow(X)){
6      for(j in 1:nrow(Y)){
7        M[i,j] <- sqrt(((X[i,1]-Y[j,1])^2)+((X[i,2]-Y[j,2])^2)
8         +((X[i,3]-Y[j,3])^2)+((X[i,4]-Y[j,4])^2)+((X[i,5]-Y[j,5])^2)
9         +((X[i,6]-Y[j,6])^2)+((X[i,7]-Y[j,7])^2)+((X[i,8]-Y[j,8])^2)
10        +((X[i,9]-Y[j,9])^2)+((X[i,10]-Y[j,10])^2))
11     }
12   }
13   # computing the chord distance depth
14   mat <- matrix(NA,ncol=1,nrow=nrow(M))
15   SpherChordDepth <- matrix(NA,ncol=1,nrow=nrow(M))
16   for (k in 1:nrow(M)){
17     mat[k,] <- mean(M[k,])
18     SpherChordDepth [k,] <- 2 - mat[k,]
19   }
20   return(as.matrix(SpherChordDepth))
```

```
21 }
```

Cosine depth distribution classifier function

```
1  # Distribution depth classifier - Use cosine distance depth (
       requires the CosDepth function)
2  # Input: testset under H1, testset under H2, training set under H1,
        training set under H2
3  # Output: misclassification rate
4  # (# of times obs. from testsetH1 are assigned to group 2 + # of
       times obs. from testsetH2 are assigned to group 1)/total
5
6  ##################################
7  ddc.cos <- function(H1test, H2test, H1train, H2train){
8
9  ############################ define internal function "edf()", called
       later
10 # "edf()" will takes as input a set of values rather than a single
       directional observation
11 # the function "edf" should take the same input of the CosDepth
       function: a test and a training set
12 # the function "edf" should give as output a set of values (a
       vector):
13 # each element of the vector is the empirical cdf of the
       corresponding test set value wrt to the whole training set
14
15 edf <- function(Htest,Htrain){
16
17 # Compute Cosine Distance Depth for the sample wrt itself
18   DepthHtrain <-CosDepth(Htrain,Htrain)
19
20 # Compute Cosine Distance Depth for Htest wrt Htrain
21   DepthHtest <-CosDepth(Htest,Htrain)
22
```

```
23 # I need a vector, where each element is the empirical cdf of the
       test set wrt to the training depth set
24 # how many of the depths in depthtest are lower than the value of
       depthtrain?
25
26 result <- rep(NA,length(DepthHtest))
27
28 for(i in 1:length(DepthHtest)){
29  result[i] <- length(DepthHtrain[DepthHtrain<=DepthHtest[i]])/
        length(DepthHtrain)
30  } ### end for
31
32 return(result)
33
34 } ### end function
35 ############################
36   cdd1 <- edf(H1test, H1train) ## depth of testH1 wrt trainH1
37   cdd2 <- edf(H1test, H2train) ## depth of testH1 wrt trainH2
38   cdd3 <- edf(H2test, H2train) ## depth of testH2 wrt trainH2
39   cdd4 <- edf(H2test, H1train) ## depth of testH2 wrt trainH1
40
41   ab <- cbind(cdd1, cdd2) ## 2 columns, depth of testH1 wrt trainH1
         on the first c., depth of testH1 wrt trainH2 on the second
42
43   cd <- cbind(cdd3, cdd4) ## 2 columns, depth of testH2 wrt trainH2
         on the first c., depth of testH2 wrt trainH1 on the second
44
45  ## the matrix "ad" has 2 columns:
46  ## depth of testH1 wrt trainH1, and of testH2 wrt trainH2 on the
       first column
47 ########## if these depths are greater than those on the second
      column, then correct classification happened
48 ########## i.e., if the value on the first column is greater than
```

```
    the value on the second, then OK
49  ## depth of testH1 wrt trainH2, and of testH2 wrt trainH1 on the
      second column
50   ad <- rbind(ab, cd)
51
52   res <- c()
53   for(i in 1:nrow(ad)){
54     if(ad[i,1] == ad[i,2]){  ## to adjust for equal values of the
         depths
55       if(rbinom(1,1,0.50) > 0)
56         (res[i] <- 2) else (res[i] <- 1)
57     }
58     else if (ad[i,1] > ad[i,2]) {res[i] <- 1} else res[i] <- 2
59   }
60
61 ### res <- 1 stay for correct misclassification
62 ### res <- 2 stay for misclassification (misclassified observation)
63
64 #print(res)
65   misrates <- 1-(length(res[res==1]))/(length(res))
66   return(misrates)
67 }
```

Cosine max-depth classifier function

```
1 # Max depth classifier - Use cosine distance depth (requires the
    CosDepth function)
2 # Input: testset under H1, testset under H2, training set under H1,
     training set under H2
3 # Output: misclassification rate
4 # (# of times obs. from testsetH1 are assigned to group 2 + # of
    times obs. from testsetH2 are assigned to group 1)/total
5
6 ################################
```

```
cddDD <- function(H1test, H2test, H1train, H2train){
  cdd1 <- CosDepth(H1test, H1train)
  cdd2 <- CosDepth(H1test, H2train)
  cdd3 <- CosDepth(H2test, H2train)
  cdd4 <- CosDepth(H2test, H1train)
  ab <- cbind(cdd1, cdd2)
  cd <- cbind(cdd3, cdd4)
  ad <- rbind(ab, cd)
  res <- c()
  for(i in 1:nrow(ad)){
    if(ad[i,1] == ad[i,2]){
      if(rbinom(1,1,0.50) > 0)
        (res[i] <- 2) else (res[i] <- 1)
    }
    else if (ad[i,1] > ad[i,2]) {res[i] <- 1} else res[i] <- 2
  }
  misrates <- 1-(length(res[res==1]))/(length(res))
  return(misrates)
}
```

Directional Bayes classifier function

```
require(Directional) # install.packages("Directional", dep=T)
require(movMF) # install.packages("movMF", dep=T)
# Classify an observation using a random uniform variate
rand_classify<-function(){
  x<-runif(1)
  if(x>0.5)
    class_testpt<-1
  else
    class_testpt<-2

  misrates <- 1-(length(class_testpt[class_testpt==1]))/(length(
      class_testpt))
```

```
12    return(misrates)
13  }
14
15  MLestimatesvm<-function(train1,train2,fast=FALSE,tol=1e-07){
16    # Estimate the parameters of the first distribution
17    par1<-vmf(train1,fast,tol=tol)
18    # Estimate the parameters of the second distribution
19    par2<-vmf(train2,fast,tol=tol)
20    return(c(par1,par2))
21  }
22
23  classify_testpt<-function(testpt1,testpt2,par_1,par_2){
24    d1<-dmovMF(testpt1,par_1$kappa*par_1$mu)
25    d2<-dmovMF(testpt1,par_2$kappa*par_2$mu)
26    d3<-dmovMF(testpt2,par_1$kappa*par_1$mu)
27    d4<-dmovMF(testpt2,par_2$kappa*par_2$mu)
28    ab <- cbind(d1, d2)
29    cd <- cbind(d3, d4)
30    ad <- rbind(ab, cd)
31    class_testpt <- c()
32
33    for(i in 1:nrow(ad)){
34      if(ad[i,1]>ad[i,2])
35        class_testpt[i]<-1
36      else{
37        if(ad[i,1]<ad[i,2])
38          class_testpt[i]<-2
39        else
40          class_testpt[i]<-rand_classify()
41      }
42
43    }
44    misrates <- (sum(class_testpt[1:length(d1)]==2)+sum(class_testpt[
```

```
        length(d1)+1:length(d3)]==1))/(length(class_testpt))
45    return(misrates)
46  }
```

References

Abebe, A., & Nudurupati, S. V. (2009). Rank-based classification using robust discriminant functions. *Communications in Statistics—Simulation and Computation*, *38*(2), 199–214.

Agostinelli, C., & Romanazzi, M. (2013). Nonparametric analysis of directional data based on data depth. *Environmental and ecological statistics*, *20*(2), 253–270.

Batschelet, E. (1981). Circular statistics in biology. *Academic Press, 111 Fifth Ave., New York, NY 10003, 1981, 388*.

Bowers, J., Morton, I., & Mould, G. (2000). Directional statistics of the wind and waves. *Applied Ocean Research*, *22*(1), 13–30.

Buttarazzi, D., Pandolfo, G., & Porzio, G. C. (2018). A boxplot for circular data. *Biometrics*, *74*(4), 1492–1501.

Cappozzo, A., Greselin, F., & Murphy, T. B. (2020). A robust approach to model-based classification based on trimming and constraints. *Advances in Data Analysis and Classification*, *14*, 327–354.

Chandra, B., & Gupta, M. (2011). Robust approach for estimating probabilities in naïve–bayes classifier for gene expression data. *Expert Systems with Applications*, *38*(3), 1293–1298.

Chang, T. (1993). Spherical regression and the statistics of tectonic plate reconstructions. *International Statistical Review/Revue Internationale de Statistique*, 299–316.

Croux, C., & Dehon, C. (2001). Robust linear discriminant analysis using s-estimators. *Canadian Journal of Statistics*, *29*(3), 473–493.

Cui, X., Lin, L., & Yang, G. (2008). An extended projection data depth and its applications to discrimination. *Communications in Statistics—Theory and Methods*, *37*(14), 2276–2290.

Debruyne, M. (2009). An outlier map for support vector machine classification. *The Annals of Applied Statistics*, 1566–1580.

Demni, H. (2021). Directional supervised learning through depth functions: an application to ecg waves analysis. Balzano S., Porzio G.C., Salvatore R., Vistocco D., and Vichi M. (eds) Statistical Learning and Modeling in Data Analysis. Studies in Classification, Data Analysis and Knowledge Organization, to appear.

Demni, H., Messaoud, A., & Porzio, G. C. (2019). The cosine depth distribution classifier for directional data. In: Bauer N., Ickstadt K., Lübke K., Szepannek G., Trautmann H., Vichi M. (eds), *Applications in Statistical Computing. Studies in Classification, Data Analysis, and Knowledge Organization*, Chapter 4, Springer, Cham, 49–60, doi: 10.1007/978-3-030-25147-5-4.

Demni, H., Messaoud, A., & Porzio, G. C. (2020). Distance-based directional depth classifiers: a robustness study. *submitted*.

Di Marzio, M., Fensore, S., Panzera, A., & Taylor, C. C. (2018). Nonparametric classification for circular data. *Applied Directional Statistics: Modern Methods and Case Studies*, 200.

Di Marzio, M., Fensore, S., Panzera, A., & Taylor, C. C. (2019). Kernel density classification for spherical data. *Statistics & Probability Letters*, *144*, 23–29.

Downs, T. D., & Liebman, J. (1969). Statistical methods for vectorcardiographic directions. *IEEE Transactions on Biomedical Engineering*, *16*(1), 87–94.

Dutta, S., & Ghosh, A. K. (2012). On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics*, *64*(3), 657–676.

Fernandes, K., & Cardoso, J. S. (2016). Discriminative directional classifiers. *Neurocomputing*, *207*, 141–149.

Figueiredo, A. (2009). Discriminant analysis for the von Mises-Fisher distribution. *Communications in Statistics—Simulation and Computation*, *38*(9), 1991–2003.

Figueiredo, A., & Gomes, P. (2006). Discriminant analysis based on the watson distribution defined on the hypersphere. *Statistics*, *40*(5), 435–445.

Fisher, N. (1989). Smoothing a sample of circular data. *Journal of Structural Geology*, *11*(6), 775–778.

Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *217*(1130), 295–305.

Frank, A., & Asuncion, A. (2010). UC Irvine Machine Learning Repository. *https://archive.ics.uci.edu/ml/datasets/Arrhythmia. Accessed on 3 June 2020*.

Frénay, B., & Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869.

Gao, D., Madden, M., Chambers, D., & Lyons, G. (2005). Bayesian ann classifier for ecg arrhythmia diagnostic system: A comparison study. In *Proceedings. 2005 ieee international joint conference on neural networks, 2005.* (Vol. 4, pp. 2383–2388).

Ghosh, A. K., & Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, *32*(2), 327–350.

Guvenir, H. A., Acar, B., Demiroz, G., & Cekin, A. (1997). A supervised machine learning algorithm for arrhythmia analysis. In *Computers in cardiology 1997* (pp. 433–436).

Hawkins, D. M., & McLachlan, G. J. (1997). High-breakdown linear discriminant analysis. *Journal of the American Statistical Association*, *92*(437), 136–143.

He, X., & Simpson, D. G. (1992). Robust direction estimation. *The Annals of Statistics*, 351–369.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Huber, P. J., & Ronchetti, E. M. (2009). Robust statistics.

Hubert, M., Rousseeuw, P., & Segaert, P. (2017). Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*, *11*(3), 445–466.

Hubert, M., & Van der Veeken, S. (2010). Robust classification for skewed data. *Advances in Data Analysis and Classification*, *4*(4), 239–254.

Hubert, M., & Van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, *45*(2), 301–320.

Jadhav, S. M., Nalbalwar, S., & Ghatol, A. (2010). Artificial neural network based cardiac arrhythmia classification using ecg signal data. In *2010 international conference on electronics and information engineering* (Vol. 1, pp. V1–228).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer, New York.

Joossens, K., & Croux, C. (2004). Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis. In: Hubert M., Pison G., Struyf A., Van Aelst S. (eds) *Theory and Applications of Recent Robust Methods. Statistics for Industry and Technology.* Birkhäuser, Basel, 131–140, https://doi.org/10.1007/978-3-0348-7958-3-12.

Jupp, P., & Mardia, K. (1989). A unified view of the theory of directional statistics, 1975-1988. *International Statistical Review/Revue Internationale de Statistique*, 261–294.

Kato, S., & Eguchi, S. (2016). Robust estimation of location and concentration parameters for the von mises–fisher distribution. *Statistical Papers*, *57*(1), 205–234.

Kent, J. T., Ganeiber, A. M., & Mardia, K. V. (2018). A new unified approach for the simulation of a wide class of directional distributions. *Journal of Computational and Graphical Statistics*, *27*(2), 291–301.

Kim, N. C., & So, H. J. (2018). Directional statistical gabor features for texture classification. *Pattern Recognition Letters*, *112*, 18–26.

Kirschstein, T., Liebscher, S., Pandolfo, G., Porzio, G. C., & Ragozini, G. (2019). On finite-sample robustness of directional location estimators. *Computational Statistics & Data Analysis*, *133*, 53–75.

Klecha, T., Kosiorowski, D., Mielczarek, D., & Rydlewski, J. P. (2018). New proposals of a stress measure in a capital and its robust estimator. *arXiv preprint arXiv:1802.03756*.

Ko, D., & Guttorp, P. (1988). Robustness of estimators for directional data. *The Annals of Statistics*, 609–618.

Koshevoy, G., & Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, *25*(5), 1998–2017.

Kosiorowski, D. (2007). About phase transitions in kendall's shape space. *Acta Universitatis Lodziensis Folia Oeconomica, Łódź, Poland*, *206*, 137-155.

Lange, T., Mosler, K., & Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers*, *55*(1), 49–69.

Leguey, I., Bielza, C., & Larrañaga, P. (2019). Circular bayesian classifiers using wrapped cauchy distributions. *Data & Knowledge Engineering*, *122*, 101–115.

Leong, P., & Carlile, S. (1998). Methods for spherical data analysis and visualization. *Journal of Neuroscience Methods*, *80*(2), 191–200.

Ley, C., Sabbah, C., & Verdebout, T. (2014). A new concept of quantiles for directional data and the angular Mahalanobis depth. *Electronic Journal of Statistics*, *8*(1), 795–816.

Li, J., Cuesta-Albertos, J. A., & Liu, R. Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association*, *107*(498), 737–753.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 405–414.

Liu, R. Y., Parelius, J. M., & Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics*, *27*(3), 783–858.

Liu, R. Y., & Singh, K. (1992). Ordering directional data: concepts of data depth on circles and spheres. *The Annals of Statistics*, 1468–1484.

López-Cruz, P. L., Bielza, C., & Larrañaga, P. (2015). Directional naive bayes classifiers. *Pattern Analysis and Applications*, *18*(2), 225–246.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. In Proceedings of the National Institute of Science, Calcutta, 2, 49-55.

Makinde, O. S., & Fasoranbaku, O. A. (2018). On maximum depth classifiers: depth distribution approach. *Journal of Applied Statistics*, *45*(6), 1106–1117.

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *37*(3), 349–371.

Mardia, K. V., & Jupp, P. E. (2009). *Directional statistics* (Vol. 494). John Wiley & Sons, Chichester.

Messaoud, A., Weihs, C., & Hering, F. (2008). Detection of chatter vibration in a drilling process using multivariate control charts. *Computational Statistics & Data Analysis*, *52*(6), 3208–3219.

Mosler, K., & Mozharovskyi, P. (2017). Fast DD-classification of functional data. *Statistical Papers*, *58*(4), 1055–1089.

Nakayama, Y. (2019). Robust support vector machine for high-dimensional imbalanced data. *Communications in Statistics—Simulation and Computation*, 1–17.

Paindaveine, D., & Van Bever, G. (2015). Nonparametrically consistent depth-based classifiers. *Bernoulli*, *21*(1), 62–82.

Paindaveine, D., & Verdebout, T. (2015). Optimal rank-based tests for the location parameter of a rotationally symmetric distribution on the hypersphere. In: Hallin M., Mason D., Pfeifer D., Steinebach J. (eds) Mathematical Statistics and Limit Theorems. Springer, Cham, 249–269, https://doi.org/10.1007/978-3-319-12442-1-14.

Pandolfo, G. (2017). Robustness aspects of DD-classifiers for directional data. In Greselin, F., Mola F., and Zenga, M. (eds), CLADAG 2017 Book of Short Papers. Universitas Studiorum S.r.l. Casa Editrice, Mantova.

Pandolfo, G., & D'Ambrosio, A. (2021). Depth-based classification of directional data. *Expert Systems with Applications*, *169*, 114433.

Pandolfo, G., D'Ambrosio, A., & Porzio, G. C. (2018). A note on depth-based classification of circular data. *Electronic Journal of Applied Statistical Analysis*, *11*(2), 447–462.

Pandolfo, G., Paindaveine, D., & Porzio, G. C. (2018). Distance-based depths for directional data. *Canadian Journal of Statistics*, *46*(4), 593–609.

Pernes, D., Fernandes, K., & Cardoso, J. S. (2019). Directional support vector machines. *Applied Sciences*, *9*(4), 725.

Pewsey, A., & García-Portugués, E. (2020). Recent advances in directional statistics. *arXiv preprint arXiv:2005.06889*.

Romanazzi, M. (2009). Data depth, random simplices and multivariate dispersion. *Statistics & Probability Letters*, *79*(12), 1473–1479.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, *53*(4), 382–387.

Saw, J. G. (1978). A family of distributions on the m-sphere and some hypothesis tests. *Biometrika*, *65*(1), 69–73.

SenGupta, A., & Roy, S. (2005). A simple classification rule for directional data. In *Advances in ranking and selection, multiple comparisons, and reliability* (pp. 81–90). Springer.

Small, C. G. (1987). Measures of centrality for multivariate and directional distributions. *Canadian Journal of Statistics*, *15*(1), 31–39.

Tsagris, M., & Alenazi, A. (2019). Comparison of discriminant analysis methods on the sphere. *Communications in Statistics: Case Studies, Data Analysis and Applications*, *5*(4), 467–491.

Tukey, J. W. (1975). Mathematics and the picturing of data. In (Vol. 2). In Proceedings of the International Congress of Mathematicians, Vancouver, 2, 523–531.

Vencálek, O. (2017). Depth-based classification for multivariate data. *Austrian Journal of Statistics*, *46*(3-4), 117–128.

Vencálek, O., Demni, H., Messaoud, A., & Porzio, G. C. (2020). On the optimality of the max-depth and max-rank classifiers for spherical data. *Applications of Mathematics*, *65*(3), 331–342.

Vencálek, O., & Pokotylo, O. (2018). Depth-weighted Bayes classification. *Computational Statistics & Data Analysis*, *123*, 1–12.

Von Mises, R. (1918). Uberdie Ganzzahligkeit der Atom gewicht und verwandte Frogren. *Physikal*, *19*, 490–500.

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing.* third Edition. Elsevier. ISBN 978-0-12-386983-8.

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, *22*(3), 177–210.

Zuo, W., Lu, W., Wang, K., & Zhang, H. (2008). Diagnosis of cardiac arrhythmia using kernel difference weighted knn classifier. In *2008 computers in cardiology* (pp. 253–256).

Zuo, Y. (2003). Projection-based depth functions and associated medians. *The Annals of Statistics*, *31*(5), 1460–1490.

Zuo, Y., & Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, 461–482.