Ph.D. THESIS

Statistical Distances and Probability Metrics for Multivariate Data, Ensembles and Probability Distributions

GABRIEL MARTOS VENTURINI
ADVISOR: ALBERTO MUÑOZ GARCÍA

Department of Statistics
UNIVERSIDAD CARLOS III DE MADRID
Leganés (Madrid, Spain)
June, 2015

Universidad Carlos III de Madrid

PH.D. THESIS

Statistical Distances and Probability Metrics for Multivariate Data, Ensembles and Probability Distributions

Author:

Gabriel Martos Venturini

Advisor:

Alberto Muñoz García

DEPARTMENT OF STATISTICS

Leganés (Madrid, Spain), June, 2015

Acknowledgements

I would like to thank to my family and friends. Without the encouragement and support of all of them it would be impossible to reach the goal.

I also thank to Alberto Muñoz, my PhD adviser, for giving me the opportunity to learn and discover many things to his side. Throughout these years we shared many things and he has been a great teacher and friend. To Javier González for his support and advice and to Laura Sangalli, my adviser in Milán, for the time and effort dedicated to my work.

En primer lugar quiero agradecer profundamente a mi familia y mis amigos por acompañarme en este proyecto. Sin el apoyo de estas personas de seguro hubiera sido imposible llegar hasta el final.

En segundo lugar, quiero agradecer a Alberto Muñoz, mi director de tesis, por darme la oportunidad de aprender y descubrir muchas cosas a su lado. A lo largo de estos años de trabajo conjunto hemos compartido muchas cosas y ha sido un gran maestro y amigo. A Javier González por el apoyo, los consejos y la disposición. A Laura Sangalli, mi tutora en Milán, por recibirme y dedicarme tiempo y esfuerzo.

Abstract

The use of distance measures in Statistics is of fundamental importance in solving practical problems, such us hypothesis testing, independence contrast, goodness of fit tests, classification tasks, outlier detection and density estimation methods, to name just a few.

The Mahalanobis distance was originally developed to compute the distance from a point to the center of a distribution taking into account the distribution of the data, in this case the normal distribution. This is the only distance measure in the statistical literature that takes into account the probabilistic information of the data. In this thesis we address the study of different distance measures that share a fundamental characteristic: all the proposed distances incorporate probabilistic information.

The thesis is organized as follows: In Chapter 1 we motivate the problems addressed in this thesis. In Chapter 2 we present the usual definitions and properties of the different distance measures for multivariate data and for probability distributions treated in the statistical literature.

In Chapter 3 we propose a distance that generalizes the Mahalanobis distance to the case where the distribution of the data is not Gaussian. To this aim, we introduce a Mercer Kernel based on the distribution of the data at hand. The Mercer Kernel induces distances from a point to the center of a distribution. In this chapter we also present a plug-in estimator of the distance that allows us to solve classification and outlier detection problems in an efficient way.

In Chapter 4 of this thesis, we present two new distance measures for multivariate data that incorporate the probabilistic information contained in the sample. In this chapter we also introduce two estimation methods for the proposed distances and we study empirically their convergence. In the experimental section of Chapter 4 we solve classification problems and obtain better results than several standard classification methods in the literature of discrimi-

nant analysis.

In Chapter 5 we propose a new family of probability metrics and we study its theoretical properties. We introduce an estimation method to compute the proposed distances that is based on the estimation of the level sets, avoiding in this way the difficult task of density estimation. In this chapter we show that the proposed distance is able to solve hypothesis tests and classification problems in general contexts, obtaining better results than other standard methods in statistics.

In Chapter 6 we introduce a new distance for sets of points. To this end, we define a dissimilarity measure for points by using a Mercer Kernel, that is extended later to a Mercer Kernel for sets of points. In this way, we are able to induce a dissimilarity index for sets of points that it is used as an input for an adaptive k-mean clustering algorithm. The proposed clustering algorithm considers an alignment of the sets of points by taking into account a wide range of possible wrapping functions. This chapter presents an application to clustering neuronal spike trains, a relevant problem in neural coding.

Finally, in Chapter 7, we present the general conclusions of this thesis and the future research lines.

Resumen

En Estadística el uso de medidas de distancia resulta de vital importancia a la hora de resolver problemas de índole práctica. Algunos métodos que hacen uso de distancias en estadística son: Contrastes de hipótesis, de independencia, de bondad de ajuste, métodos de clasificación, detección de atípicos y estimación de densidad, entre otros.

La distancia de Mahalanobis, que fue diseñada originalmente para hallar la distancia de un punto al centro de una distribución usando información de la distribución ambiente, en este caso la normal. Constituye el único ejemplo existente en estadística de distancia que considera información probabilística. En esta tesis abordamos el estudio de diferentes medidas de distancia que comparten una característica en común: todas ellas incorporan información probabilística.

El trabajo se encuentra organizado de la siguiente manera: En el Capítulo 1 motivamos los problemas abordados en esta tesis. En el Capítulo 2 de este trabajo presentamos las definiciones y propiedades de las diferentes medidas de distancias para datos multivariantes y para medidas de probabilidad existentes en la literatura.

En el Capítulo 3 se propone una distancia que generaliza la distancia de Mahalanobis al caso en que la distribución de los datos no es Gaussiana. Para ello se propone un Núcleo (kernel) de Mercer basado en la densidad (muestral) de los datos que nos confiere la posibilidad de inducir distancias de un punto a una distribución. En este capítulo presentamos además un estimador plug-in de la distancia que nos permite resolver, de manera práctica y eficiente, problemas de detección de atípicos y problemas de clasificación mejorando los resultados obtenidos al utilizar otros métodos de la literatura.

Continuando con el estudio de medidas de distancia, en el Capítulo 4 de esta tesis se proponen dos nuevas medidas de distancia para datos multivariantes incorporando información

probabilística contenida en la muestra. En este capítulo proponemos también dos métodos de estimación eficientes para las distancias propuestas y estudiamos de manera empírica su convergencia. En la sección experimental del Capítulo 4 se resulven problemas de clasificación con las medidas de distancia propuestas, mejorando los resultados obtenidos con procedimientos habitualmente utilizados en la literatura de análisis discriminante.

En el Capítulo 5 proponemos una familia de distancias entre medidas de probabilidad. Se estudian también las propiedades teóricas de la familia de métricas propuesta y se establece un método de estimación de las distancias basado en la estimación de los conjuntos de nivel (definidos en este capítulo), evitando así la estimación directa de la densidad. En este capítulo se resuelven diferentes problemas de índole práctica con las métricas propuestas: Contraste de hipótesis y problemas de clasificación en diferentes contextos. Los resultados empíricos de este capítulo demuestran que la distancia propuesta es superior a otros métodos habituales de la literatura.

Para finalizar con el estudio de distancias, en el Capítulo 6 se propone una medida de distancia entre conjuntos de puntos. Para ello, se define una medida de similaridad entre puntos a través de un kernel de Mercer. A continuación se extiende el kernel para puntos a un kernel de Mercer para conjuntos de puntos. De esta forma, el Núcleo de Mercer para conjuntos de puntos es utilizado para inducir una métrica (un índice de disimilaridad) entre conjuntos de puntos. En este capítulo se propone un método de clasificación por k-medias que utliza la métrica propuesta y que contempla, además, la posibilidad de alinear los conjuntos de puntos en cada etapa de la construcción de los clusters. En este capítulo presentamos una aplicación relativa al estudio de la decodificación neuronal, donde utilizamos el método propuesto para encontrar clusters de neuronas con patrones de funcionamiento similares.

Finalmente en el Capítulo 7 se presentan las conclusiones generales de este trabajo y las futuras líneas de investigación.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of distance and similarity measures are of fundamental importance in Statistics. Distances (and similarities) are essentially measures that describe how close two statistical objects are. The Euclidean distance and the linear correlation coefficient are usual examples of distance and similarity measures respectively. The distance measures are often used as input for several data analysis algorithms. For instance in Multidimensional Scaling (MDS), where a matrix of distances or a similarity matrix D is used in order to represent each object as a point in an affine coordinate space so that the distances between the points reflect the observed proximities between the objects at hand.

Another usual example of the use of a distance measure is the problem of outlier (extreme value) detection. The Mahalanobis (1936) distance, originally defined to compute the distance from the point to the center of a distribution is then ideal to solve the outlier detection problem. In Figure 1.1, we represent the density functions of two normally distributed populations: $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively. The parameters that characterize the distribution of the populations are: $\mu_{\mathbb{P}_1} = \mu_{\mathbb{P}_2}$ and $\sigma^2_{\mathbb{P}_1} > \sigma^2_{\mathbb{P}_2}$, where $\mu_{\mathbb{P}_i}$ and $\sigma^2_{\mathbb{P}_i}$ for $i = 1, 2$, are the mean and the variance of each distribution. A standard procedure to determine if a point in the support of a distribution it is an outlier (an extreme value) is to consider a distance measure $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ to the center (in this case the mean) of the distribution.

It is clear in this context that a suitable distance measure to solve the extreme value detection problem should depend on the distribution of the data at hand. From a probabilistic point of view, if we observe the value $X = x$ as is shown in Figure 1.1, then it is more likely that x was generated from the population distributed according to $\mathbb{P}_1$. Therefore the distance criterion should agree with the probabilistic criterion, that is:

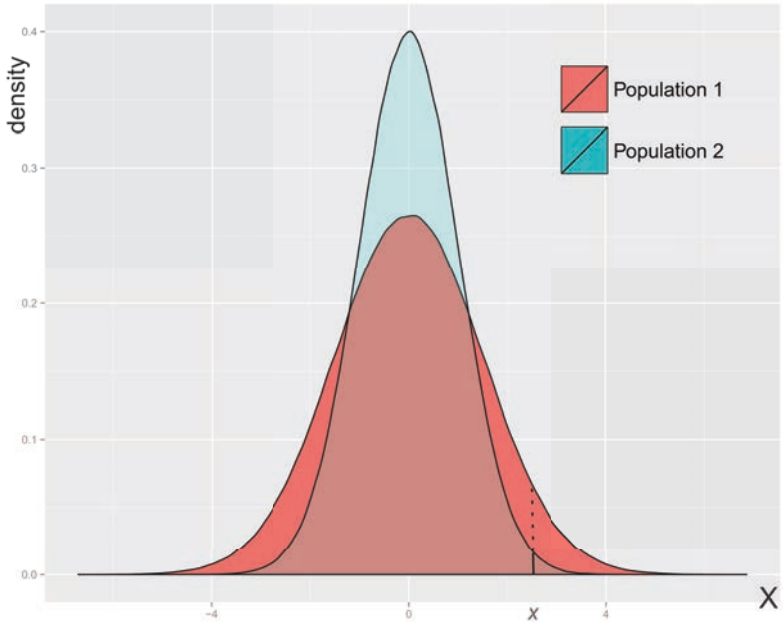


Figure 1.1: Distributions (density functions) of populations 1 and 2.

$$P(X = x|X \sim \mathbb{P}_1) > P(X = x|X \sim \mathbb{P}_2) \Rightarrow d_{\mathbb{P}_1}(x, \mu) < d_{\mathbb{P}_2}(x, \mu),$$

in order to conclude that if we observe a value $X = x$, as in Figure 1.1, then it is more likely that this value it is an outlier if the data is distributed according to $\mathbb{P}_2$ than if the data is distributed according to $\mathbb{P}_1$.

The Mahalanobis distance is the only distance measure of the statistical literature that takes into account the probabilistic information in the data and fulfills the probabilistic criterion: $d_{\mathbb{P}_1}(x, \mu) < d_{\mathbb{P}_2}(x, \mu)$ if $P(X = x|X \sim \mathbb{P}_1) > P(X = x|X \sim \mathbb{P}_2)$ but only in the cases of normally distributed data (as in Figure 1.1). In Chapter 3 of this thesis, we propose and study a distance measure from a point to the center of a distribution that is able to fulfill the probabilistic criterion even in the case of non-Gaussian distributions. The proposed metric generalizes in this way the Mahalanbois distance and is useful to solve, for instance, the outlier detection problem in cases when the Mahalanobis distance is not able.

Another restriction of the Mahalanobis distance is that was originally developed to compute distances from a point to a distribution center. In this way, in Chapter 4 of this thesis we propose new distance measures for general multivariate data, that is $d(x, y)$ where neither x

nor y are the center of a distribution, that take into account the distributional properties of the data at hand. We show in Chapter 4 that the proposed distance measures are helpful to solve classification problems in statistic and data analysis.

Probability metrics are distance measures between random quantities as random variables, data samples or probability measures. There are several probability metrics treated in the statistical literature, for example: The Hellinger distance, the Bhattacharyya distance, the Energy distance or the Wasserstein metric, to name just a few. These distance measures are important in order to solve several statistical problems, as for example to perform Hypothesis, Independence or Goodness of Fit tests. In Chapter 5 of this thesis we concentrate our efforts in the study of a new family of distance measures between probability measures. The proposed family of distance measures allows us to solve typical statistical problems: Homogeneity tests and classification problems among other interesting applications developed in Chapter 5.

A distance function it is also a fundamental input for several methods and algorithms in data analysis: k-means clustering, k-nearest neighbor classifier or support vector regression constitutes fundamental examples of methods that are based in the use of a *metric* to solve clustering, classification and regression problems respectively. Real world applications in the fields of Machine Learning and Data Mining that rely on the computation of distance and similarity measures are document categorization, image retrieval, recommender system, or target marketing, to name a few. In this context we usually deal with high-dimensional or even functional data, and the use of standard distance measures, e.g. the Euclidean distance or a standard similarity measure as the linear correlation coefficient, do not always adequately represent the similarity (or dissimilarity) between the objects at hand (documents, images, etc).

For instance, in Neuroscience in order to classify neurons according to its firing activity, we need to establish a suitable distance measure between neurons. In this context, data are usually represented as continuous time series $x(t)$, where $x(t) = 1$ if we observe a spike at the time t in the neuron x and $x(t) = 0$ otherwise. In Figure 1.2-a) we show a raster plot, this plot represents the spike train pattern of several neurons (vertical axes) against time (horizontal axes). In order to adequately classify the brain spike train paths, we need to define and compute a convenient measure d, such that for two neurons: x_i and x_j, then $d(x_i(t), x_j(t))$ is small when the firing activity of neuron x_i is similar to the firing activity of neuron x_j. It is clear that the use of standard distance measures between these time series, for example the Euclidean distance (in

the L_2 sense):

$$d(x_i(t), x_j(t)) = \int (x_i(t) - x_j(t))^2 dt,$$

will not provide a suitable metric to classify neurons.

In this thesis we also develop and study distance and similarity measures for high-dimensional and functional data. The proposed measures take into account the distributional properties of the data at hand in order to adequately solve several statistical learning problems. For instance, in Chapter 6 we propose and study the fundamental input to solve the brain spike train classification problem: a suitable matrix D of distances, where $[D]_{i,j} = d(x_i(t), x_j(t))$ is a suitable measure of distance between two neurons, that combined with the standard k-means algorithm, will allow us to adequately classify a set of neurons according to its firing activity.

Another example of the importance of the study of distance measures for functional data objects is in Bioinformatics. In this field, the genomic information of a biological system is represented by time series of cDNA micro-arrays expressions gathered over a set of equally spaced time points. In this case we can represent the DNA time series by using some functional base (splines, RKHS, polynomial bases, among others), as it is exemplified in Figure 1.2-b). A suitable distance measure between the DNA sequences is of fundamental importance when we need to classify time series of cDNA micro-arrays expressions. In Chapter 5 we propose a distance for probability metrics that can be adapted to solve classification problems of cDNA time series of micro-arrays data.

Another interesting example of the use of distance measures in the context of high-dimensional and complex data is in Social Network Analysis. The use of graphs helps to describes information about patterns of ties among social actors. In Figure 1.2-c), we exemplify the structure of a social network with the aid of a graph. In order to make predictions about the future evolution of a social network, we need a suitable measure of distance between networks, that is a suitable metric between graphs. A similar example is related to Image Recognition. Image and shapes, 2D and 3D objects respectively, can be represented as sets of points. In Figure 1.2-d) we represent the image of a heart as a uniformly distributed set of points. In this context a suitable measure of distance between sets of points, as the one presented in Chapters 5 and 6, will be helpful in order to classify and organize sets of images or surfaces.

Next section summarize the main contributions made in this thesis regarding the study of

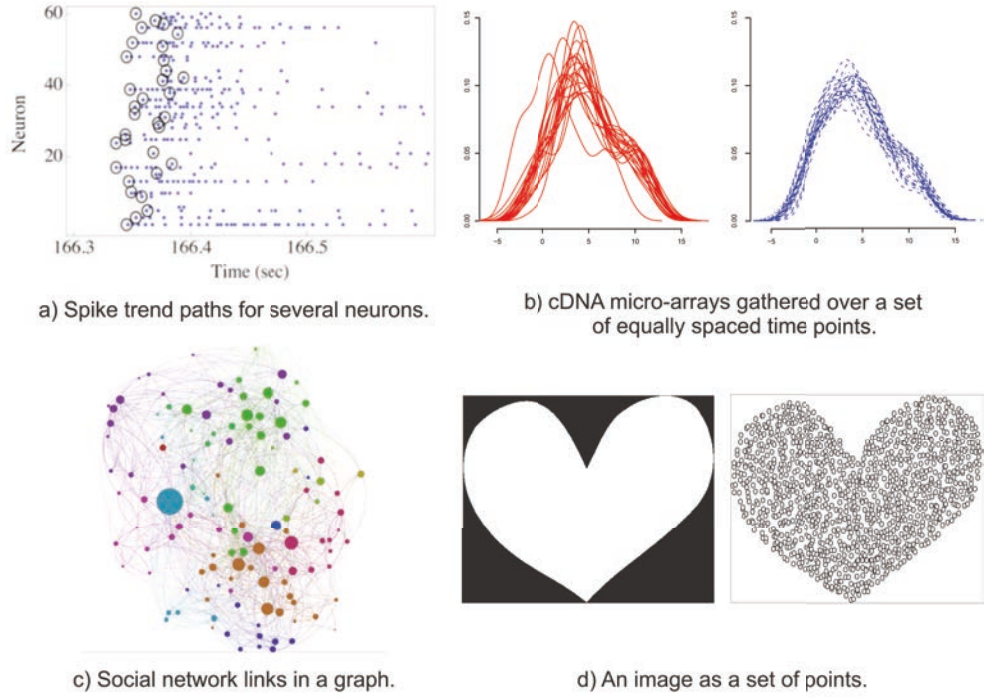


Figure 1.2: Four examples of high-dimensional data objects.

distance, similarity and related measures in different contexts: from a point to the center of a distribution, between multivariate data, between between probability measures, between sets of points (high-dimensional and complex data).

1.1 Overview of the Thesis and Contributions

Overview of Chapter 2

In Chapter 2 we give review the concept of distance in the context of Statistics and data analysis. The aim of this chapter is to give a wide context for the study of distance measures and a reference for the following chapters.

Problem to solve in Chapter 3

The Mahalanobis distance is a widely used distance in Statistics and related areas to solve classification and outlier detection problems. One of the fundamental inconvenience in the use of this metric is that only works well when the data at hand follows a Normal distribution.

Contributions of Chapter 3

In Chapter 3 we introduce a family of distances from a point to a center of a distribution. We propose a generalization of the Mahalanobis distance by introducing a particular class of Mercer kernel, the density kernel, based on the underlying data distribution. Density kernels induce distances that generalize the Mahalanobis distance and that are useful when data do not fit the assumption of Gaussian distribution. We provide a definition of the distance in terms of the data density function and also a computable version for real data samples of the proposed distance that is based on the estimation of the level sets. The proposed Generalized Mahalanobis distance is test on a variety of artificial and real data analysis problems with outstanding results.

Problem to solve in Chapter 4

There exist diverse distance measures in data analysis, all of them devised to be used in the context of Euclidean spaces where the idea of density is missing. When we want to consider distances that takes into account the distribution of the data at hand, the only distance left is the Mahalanobis distance. Nevertheless, the Mahalanobis distance only considers the distance from a point to the center of a distribution and assumes a Gaussian distribution in the data.

Contributions of Chapter 4

In Chapter 4 we propose two new distance measures for multivariate data that take into account the distribution of the data at hand. In first place, we make use of the distribution function to propose the Cumulative Distribution Function distance and study later its properties. Also in this chapter we propose a distance, the Minimum Work Statistical distance, that is based on the minimization of a functional defined in terms of to the density function. We study its properties and propose an estimation procedure that avoid the need of the explicit density estimation. In the experimental section of Chapter 4 we show the performance of the proposed distance measures to solve classification problems when they are combined with standard classification methods in statistics and data analysis.

Problem to solve in Chapter 5

Probability metrics are distances that quantifies how close (or similar) two random objects are, in particular two probability measures. The use of probability metrics is of fundamental

importance to solve homogeneity, independence and goodness of fit tests, to solve density estimation problems or to study the stochastic convergence among many other applications.

There exist several theoretical probability metrics treatise in the statistical literature. Nevertheless, in practice, we only have available a finite (and usually not huge) data sample. The main drawback to compute a distance between two probability measures is therefore that we do not know explicitly the density (or distribution) functions corresponding to the samples under consideration. Thus we need first to estimate the density and then compute the distance between the samples using the estimated densities. As it is well known, the estimation of general density functions it is an ill-posed problem: the solution is not necessarily unique and the solution it is not necessarily stable. And the difficulty in density estimation rises when the dimension of the data increases. This motivates the need of seeking for probability metrics that do not explicitly rely on the estimation of the corresponding probability distribution functions.

Contributions of Chapter 5

In Chapter 5 we study distances between probability measures. We show that probability measures can be considered as continuous and linear functionals (generalized functions) that belong to some functional space endowed with an inner product. We derive different distance measures for probability distributions from the metric structure inherited from the ambient inner product. We propose particular instances of such metrics for probability measures based on the estimation of density level sets regions, avoiding in this way the difficult task of density estimation.

We test the performance of the proposed metrics on a battery of simulated and real data examples. Regarding the practical applications, new probability metrics have been proven to be competitive in homogeneity tests (for univariate and multivariate data distributions), shape recognition problems, and also to classify DNA micro-arrays time series of genes between groups of patients.

Problem to solve in Chapter 6

In several cases in statistics and data analysis, in particular when we work with high-dimensional or functional data, there are problems in the registration of the data. In the fields of Neuro and Bio-Informatics it is frequent to observe registration problems: amplitude and phase variability in the registered neuronal and proteomic raw data respectively. In this context, in order to

solve classification and regression problems, it is important to *align* the data before the analysis. For this purpose, we need to define a proper metric that help to produce a suitable alignment procedure. In the case of raw data that contains rich and useful probabilistic information, we should incorporate this information in the metric that produce the alignment.

In several problems related to Neural coding, the use of classification and clustering methods that includes an alignment for the raw data are necessary. There are few classification and clustering methods treated in the statistical literature that incorporate these alignment step and that are prepared to deal with high-dimensional and complex data.

Contributions of Chapter 6

In Chapter 6 we propose a novel and flexible k-means clustering method for sets of points that incorporates an alignment step between the sets of points. For this purpose, we consider a Mercer kernel function for data points with reference to a distribution function, the probability distribution of the data set at hand. The kernel for data points it is extended to a Mercer kernel for sets of points. The resulting kernel for sets of points induces an affine invariant measure of dissimilarity for sets of points that contains distributional information of the sets of points.

The metric proposed in Chapter 6 is used to optimize an alignment procedure between sets of points and it is also combined with a k-means algorithm. In this way, the k-means clustering algorithm proposed in this Chapter incorporates an *alignment* step that make uses of a broad family of wrapping functions. The clustering procedure is flexible enough to work in different real data contexts. We present an application of the proposed method in Neuronal coding: spike train classification. Nevertheless, the given k-means procedure is also suitable in more general contexts, for instance: In image segmentation or time series classification, among others possible uses.

Chapter 7

In Chapter 7 we summarize the work done in the thesis and its main contributions. We also point out the most important future research lines.

List of Symbols

In Table 1.1 we introduce the main list of symbols we use in this thesis.

X	A compact set.
$C(X)$	Banach space functions on X.
$C_c(X)$	Space of all compactly supported and piecewise continuous functions on X.
$\mathcal{H}$	Hilbert space of functions.
$\mathcal{D}$	Space of test functions.
$\mathcal{T}$	The family of affine transformations.
K	A kernel function.
$\mathcal{F}$	σ-algebra of measurable subsets of X.
μ	Lebesgue measure in X.
ν	Borel measure in X.
PM	Probability measure.
$\mathbb{P},\ \mathbb{Q}$	Two σ-additive finite PMs absolutely continuous w.r.t. μ.
$f_{\mathbb{P}},\ f_{\mathbb{Q}}$	Density functions.
$F_{\mathbb{P}},\ F_{\mathbb{Q}}$	Distribution functions.
$\mathbb{E}$	Expectation operator.
$\mathbb{V}$	Variance operator.
MD	Mahalanobis Distance.
BD	Bregman Divergence.
ζ	A strictly convex and differentiable function.
d	A distance function.
$s_n(\mathbb{P})$	Random sample of n observations drawn from the PM $\mathbb{P}$.
$\boldsymbol{\alpha}_{\mathbb{P}}^m$	A non-decreasing α-sequence for the PM $\mathbb{P}: 0 = \alpha_1 \leq \ldots \leq \alpha_m = \max_x f_{\mathbb{P}}(x)$.
$S_\alpha(f_{\mathbb{P}})$	Minimum volume sets $\{x \in X \mid f_{\mathbb{P}}(x) \geq \alpha\}$.
$A_i(\mathbb{P})$	Constant density set: $A_i(\mathbb{P}) = S_{\alpha_i}(f_{\mathbb{P}}) - S_{\alpha_{i+1}}(f_{\mathbb{P}})$ of a PM $\mathbb{P}$ with respect to $\alpha_i \in \boldsymbol{\alpha}_{\mathbb{P}}^m$.

Table 1.1: The list with the most important symbols and acronyms used in the thesis.

Chapter 2

Background: Distances and Geometry in Statistic

The study of the geometrical properties of a random variable, a probability distribution or a sample of data collected from a population is not in the core of traditional Statistic. However in recent years, with the rising of complex data structures in fields like Computational Biology, Computer Vision or Information Retrieval, the research on the geometry and the study of intrinsic metrics of the statistical objects at hands is taking more relevance.

Several authors in the recent statistical literature have emphasizes in the importance of the study of the geometry of a statistical system and its intrinsic metric, a not exhaustive list of examples are Chenĥsov (1982); Kass (1989); Amari et al. (1987); Belkin et al. (2006); Lebanon (2006); Marriott and Salmon (2000); Pennec (2006). In these articles the authors applies techniques of differential geometry. The usual approach is to consider the statistical objects, for example probability distributions, as points in a Riemannian manifold endowed with an inner product. With the metric derived from the inner product, it is possible to solve several Statistical problems.

From our point of view, the study of *distance and similarity measures* that take into account the *geometrical structure* and also its *probabilistic contents* is of crucial importance in order to adequately solve relevant data analysis problems. Following this idea, we propose in this thesis to address the problem of construct distance measures that takes into account the probabilistic information of the data at hand in order to solve the usual statistical tasks: regression, classification and density estimation problems.

Next in this chapter, we provide a general overview to the concept of distance and similiarity measures, and also a review of other related concepts, for instance Divergences. We also introduce several probability metrics, including the metrics induced by Mercer kernels, providing in this way a context for the work done in this thesis.

2.1 Distances and Similarities between Points

A distance measure is in essence a function that measures how similar are two data. The study of distance and similarity measures appears in the early history of modern Statistics. For instance P.S. Mahalanobis, one of the most influential statistician in XX century, is remembered for the introduction of a measure of distance with his name: The Mahalanobis (1936) distance, that appears in the definition of multivariate Normal distribution, in the study of discriminant analysis or in several outlier detection techniques among many other statistical procedures. In what follows we formally introduce the concept of distance and its properties.

For any set X, a *distance function* or a *metric* is an application $d : X \times X \to \mathbb{R}$, such that for all $x, y \in X$:

(i) $d(x, y) \geq 0$ (non-negativity),

(ii) $d(x, y) = 0$ if and only of $x = y$ (identity of indiscernible),

(iii) $d(x, y) = d(y, x)$ (symmetry),

(iv) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

In order to clarify the terminology, in this thesis we distinguish a *metric* from a *semimetric* when the function d do not necessarily fulfill the triangle inequality, from a *quasimetric* when the function d do not necessarily satisfy the symmetry condition and from a *pesudometric* when the function d do not necessarily achieve the identity of indiscernible property. Other generalized definitions of metric can be obtained by relaxing the axioms i to iv, refer to Deza and Deza (2009) for additional details.

A space X equipped with a metric d it is call a *metric space* (X, d). Given two metric spaces (X, d_X) and (Y, d_Y), a function $\psi : X \to Y$ is called an *isometric embedding* of X into Y if ψ is injective and $d_Y(\psi(x), \psi(y)) = d_X(x, y)$ holds for all $x, y \in X$. An *isometry* (or congruence

mapping) is a bijective isometric embedding. Two metric spaces are called **isometric** (or isometrically isomorphic) if there exist an isometry between them. Two metric spaces (X, d_X) and (Y, d_Y) are called *homeomorphic* (or topologically isomorphic) if there exists a homeomorphism from X to Y , that is a bijective function $\psi : X \to Y$ such that ψ and ψ^{-1} are continuous.

In several data analysis problems, some additional properties to $i) \ - \ iv)$ are desirable on the metric d. For example, in Pattern Recognition the use of invariant metrics under certain type of transformations are usual Hagedoorn and Veltkamp (1999); Simard et al. (2012). Let $\mathcal{T}$ be a class of transformations (e.g. the class of affine transformations), we say that the metric d is invariant under the class $\mathcal{T}$ if $d(x, y) = d\left(h(x), h(y)\right)$ for all $x, y \in X$ and $h \in \mathcal{T}$. As an example, consider the case of the Euclidean distance which is invariant under the class of orthogonal transformations.

A *similarity* on a set X is a function $s : X \times X \to \mathbb{R}$ such that s is non-negative, symmetric and $s(x, y) \leq s(x, x)$ for all $x, \ y \in X$, with equality if and only if $x = y$. A *similarity* increases in a monotone fashion as x and y becomes more similar. A *dissimilarity* measure, it is also a non-negative and symmetric measure, but opposite to a similarity, the *dissimilarity* decreases as x and y becomes more similar. Several algorithms in Machine Learning and Data Mining work with similarities and dissimilarities, but sometimes it may become necessary to convert similarities into dissimilarities or vice versa. There exist several transformation of similarities into dissimilarities and vice versa, the trick consist in apply a monotone function. Next we give some examples.

From similarities to dissimilarities: Let s be a normalized similarity, that is $0 \leq s(x, y) \leq 1$ for all $x, y \in X$, typical conversions of s into a dissimilarity d are:

- $d(x, y) = 1 - s(x, y)$,
- $d(x, y) = \sqrt{s(x, x) + s(y, y) - 2s(x, y)}$,
- $d(x, y) = -\log(s(x, y))$,

alternative transformations can be seen in Gower (2006); Gower and Legendre (1986).

From dissimilarities to similarities: In the other way around, let d be a dissimilarity measure, typical conversions of d into a similarity measure s are:

- $s(x, y) = \exp(-\frac{d(x,y)}{\sigma})$,

- $s(x,y) = \frac{1}{1+d(x,y)}$,
- $s(x,y) = \frac{1}{2}\left(d^2(x,c) + d^2(y,c) - d^2(x,y)\right)$,

where c is a point in X (for example the mean point or the origin). Next we explore other concepts related to distances.

2.1.1 Bregman divergences

A Divergence arises as a weaker concept than a distance because does not necessarily satisfy the symmetric property and does not necessarily satisfy the triangle inequality. A divergence is defined in terms of a strictly convex and differentiable function as follows.

Definition 2.1. (**Bregman Divergence**): Let X be a compact set and ζ a strictly convex and differentiable function $\zeta : X \to \mathbb{R}$. The *Bregman Divergence* (BD) for a pair of points $(x,y) \in X$ is defined as follows

$$BD_\zeta(x,y) = \zeta(x) - \zeta(y) - \langle x - y, \nabla\zeta(y)\rangle, \tag{2.1}$$

where $\nabla\zeta(y)$ is the gradient vector evaluated in the point y. The BD is a related concept to distance that satisfies the following properties (see Bregman (1967) for further details):

- **Non-negativity**: $BD_\zeta(x,y) \geq 0$ and $BD_\zeta(x,y) = 0$ if and only if $x = y$,
- **Taylor Expansion**: for small Δx we can approximate

$$BD_\zeta(x, x+dx) = \frac{1}{2}\sum_{i,j}\frac{\partial^2\zeta(x)}{\partial x_i \partial x_j}dx_i dx_j,$$

- **Convexity**: A Bregman divergence is a convex function respect the first argument, i.e. $x \longmapsto BD_\zeta(x,y)$ is a convex function for all $y \in X$.
- **Linearity**: $BD_{\alpha\zeta_1+\beta\zeta_2} = \alpha BD_{\zeta_1} + \beta BD_{\zeta_2}$ for all ζ_1 and ζ_2 strictly convex and differentiable functions and positive constants α and β.
- **Affine Invariance**: let g be an affine function (i.e. $g(x) = Ax + c$, for a constant matrix $A \in \mathbb{R}^{d\times d}$ and a fix vector $c \in \mathbb{R}^d$), then $BD_\zeta(g(x), g(y)) = BD_{\zeta\circ g}(x,y) = BD_\zeta(x,y)$.

The convexity of the Bregman Divergences is an important property for many Machine Learning algorithms that are base in the optimization of this measure, see for instance Banerjee et al. (2005); Davis et al. (2007); Si et al. (2010).

Several well known distances are particular cases of Bregman divergences. For example, the Mahalanobis (1936) distance (d_M), a widely used distance in statistics and data analysis for classification and outlier detection tasks. The Mahalanobis distance is a scale-invariant metric that provides a measure of distance between a point $\mathbf{x} \in \mathbb{R}^d$ generated from a given distribution $\mathbb{P}$ and the mean $\boldsymbol{\mu} = \mathbb{E}_{\mathbb{P}}(\mathbf{x})$ of the distribution. Assume $\mathbb{P}$ has finite second order moments and denote by $\boldsymbol{\Sigma} = \mathbb{E}_{\mathbb{P}}\left((\mathbf{x}-\boldsymbol{\mu})^2\right)$ the covariance matrix. Then the Mahalanobis distance is defined by:

$$d_M(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}.$$

It is easy to see that when $\zeta(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}\mathbf{x}$, then $d_M^2(\mathbf{x}, \boldsymbol{\mu}) = BD_\zeta(\mathbf{x}, \boldsymbol{\mu})$. The Mahalanobis distance, or equivalently the Mahalanobis Bregman Divergence, has a very interesting geometrical interpretation, it can be seen as the composition of a linear transformation $T_M : \mathbf{x} \xrightarrow{T_M} \mathbf{x}' = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{x}$, plus the computation of the ordinary Euclidean distance (d_E) between the transformed data. This is illustrated in Figure 2.1 for two data points from a bivariate Normal distribution.

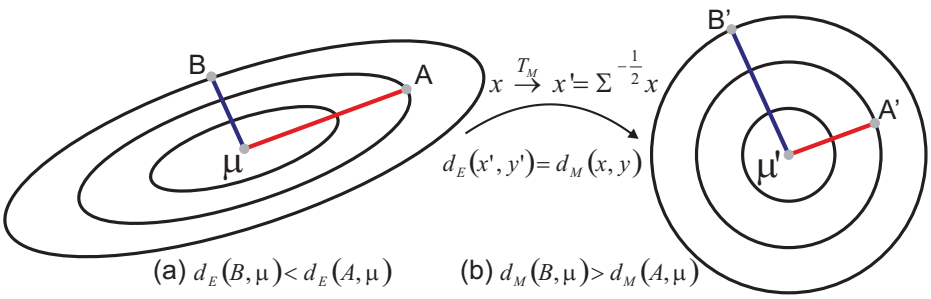


Figure 2.1: The effect of the Mahalanobis transformation.

When the data follows a Normal distribution then the Mahalanobis distance preserves a essential property: "all the points that belong to the same probability curve, that is $L_c(f_{\mathbb{P}}) = \{\mathbf{x} | f_{\mathbb{P}}(\mathbf{x}) = c\}$ where $f_{\mathbb{P}}$ is the density function of the normally distributed r.v. $\mathbf{x}$, are equally distant from the center (the densest point) of the distribution".

Related to this metric, in Chapter 3 we elaborate on the generalization of the Mahalanobis distance via the introduction of a Mercer density kernels. Density kernels induce metrics that generalize the Mahalanobis distance for the case of non Gaussian data distributions. We introduce the distance induced by kernel functions later in this chapter.

2.2 Probability Metrics

In Statistics, a *probability metric* (also known as a statistical distance) is a measure that quantifies the dissimilarity between two random quantities, in particular between two probability measures. Probability metrics are widely used in statistic, for example in the case of homogeneity test between populations. Given two random samples $S_{\mathbb{P}}^n = \{x_i\}_{i=1}^n$ and $S_{\mathbb{Q}}^m = \{y_i\}_{i=1}^m$, drawn from the probability measures (two different populations) $\mathbb{P}$ and $\mathbb{Q}$ respectively, we have to decide if there is enough empirical evidence to reject the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ using the information contained in the random samples $S_{\mathbb{P}}^n$ and $S_{\mathbb{Q}}^m$. This problem is solved by using a statistical distance. Is easy to see that this problem is equivalent to testing $H_0 : d(\mathbb{P}, \mathbb{Q}) = 0$ versus $H_1 : d(\mathbb{P}, \mathbb{Q}) > 0$, where d is a distance or a dissimilarity measure (a metric on the space of the involved probability measures). We will find evidence to not reject H_0 when $d(S_{\mathbb{P}}^n, S_{\mathbb{Q}}^m) \gg 0$. Other examples of the use of probability metrics in Statistics are independence and goodness of fit tests, to solve density estimation problems or to study convergence laws of random variables among many other applications.

In what follows $\mathbb{P}$ and $\mathbb{Q}$ denotes two probability measures and $f_{\mathbb{P}}$ and $f_{\mathbb{Q}}$ the respective density functions. The functions $F_{\mathbb{P}}$ and $F_{\mathbb{Q}}$ denotes the distributions functions respectively. Some examples of *probability metrics* are:

- The **Hellinger** distance between $\mathbb{P}$ and $\mathbb{Q}$ is computed as

$$d_H^2(\mathbb{P}, \mathbb{Q}) = \int_X \left(\sqrt{f_{\mathbb{P}}(x)} - \sqrt{f_{\mathbb{Q}}(x)} \right)^2 dx.$$

- The **Bhattacharyya** distance between $\mathbb{P}$ and $\mathbb{Q}$ is computed as

$$d_B(\mathbb{P}, \mathbb{Q}) = -\log \left(\int_X \left(\sqrt{f_{\mathbb{P}}(x) f_{\mathbb{Q}}(x)} \right) dx \right).$$

- The **Wasserstein-Kantorovich** distance between $\mathbb{P}$ and $\mathbb{Q}$ is computed as

$$d_W(\mathbb{P}, \mathbb{Q}) = \int_X |F_{\mathbb{P}}(x) - F_{\mathbb{Q}}(x)| \, dx.$$

- The **Total Variation** distance between $\mathbb{P}$ and $\mathbb{Q}$ is computed as

$$d_T(\mathbb{P},\mathbb{Q}) = \int_X |f_{\mathbb{P}}(x) - f_{\mathbb{Q}}(x)|\, d\mu,$$

  where μ is a positive measure such that both $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with respect to it.

- The **Komogorov-Smirnov** distance between $\mathbb{P}$ and $\mathbb{Q}$ is computed as

$$d_{K-S}(\mathbb{P},\mathbb{Q}) = \sup_x |F_{\mathbb{P}}(x) - F_{\mathbb{Q}}(x)|\,.$$

Probability metrics are mostly not proper metrics, usually they are semimetrics, quasimetrics or pseudometrics. More examples on distances between probability measures and its relationships can be seen in Deza and Deza (2009); Gibbs and Su (2002); Müller (1997); Zolotarev (1983). Several probability metrics in the statistical literature are referred as divergences, next we introduce this concept.

2.2.1 Statistical divergences

We already mention that a Divergence arises as a weaker concept than distance because does not necessarily fulfill the symmetric property and does not necessarily satisfy the triangle inequality. Divergences can also be used to measure the proximity between two probability measures.

Definition 2.2. (**Statistical Bregman Divergence**): Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures and denote by $f_{\mathbb{P}}$ and $f_{\mathbb{Q}}$ the respective density functions, let ζ be a strictly convex and differentiable function $\zeta : X \to \mathbb{R}$. The *Functional Bregman Divergence* (BD) for a pair $\mathbb{P}$ and $\mathbb{Q}$ is defined as follows

$$BD_\zeta(\mathbb{P},\mathbb{Q}) = \int_X \Big(\zeta(f_{\mathbb{P}}) - \zeta(f_{\mathbb{Q}}) - (f_{\mathbb{P}} - f_{\mathbb{Q}})\,\nabla\zeta(f_{\mathbb{Q}})\Big) d\mu(x),$$

where μ is a positive measure such that both $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with respect to it and $\nabla\zeta(f_{\mathbb{Q}})$ is the derivative of ζ evaluated at $f_{\mathbb{Q}}$ (see Jones and Byrne (1990); Csiszár (1995) for further details).

Some examples of Statistical Bregman divergences can be obtained making $\zeta(t) = t^2$, then

$BD_\zeta(\mathbb{P},\mathbb{Q})$ yields the Euclidean distance between $\mathbb{P}$ and $\mathbb{Q}$ (in the L_2 sense); when $\zeta(t) = t\log(t)$ then $BD_\zeta(\mathbb{P},\mathbb{Q})$ is the *Kullback Leibler* Divergence between $\mathbb{P}$ and $\mathbb{Q}$, that is:

$$BD_\zeta(\mathbb{P},\mathbb{Q}) = \int_X \left(f_{\mathbb{P}} \log\left(\frac{f_{\mathbb{P}}}{f_{\mathbb{Q}}}\right)\right) d\mu(x).$$

The Bregman divergences are intimately related to the Fisher-Rao metric. Fisher and Rao are the precursors of the idea of consider probability distributions as points that belongs to a manifold, and then take advantage of the manifold structure to derive appropriate metrics for distributions Burbea and Rao (1982); Amari et al. (1987); Atkinson and Mitchell (1981). This point of view is used, for instance, in Image and Vision Pennec (2006); Srivastava et al. (2007).

A divergence function induce a Riemannian metric in the space of probability measures. In the case of Bregman divergences, the metric tensor (denoted as g_{ij}) is defined in terms of the strictly convex and differentiable function ζ:

$$g_{ij}(\mathbf{z}) = \frac{\partial^2}{\partial z_i \partial z_j}\zeta(\mathbf{z}),$$

where the vector $\mathbf{z}$ represents the local coordinates on a (smooth) manifold $\mathcal{M}$. When the metric tensor g_{ij} is derived from a Bregman Divergence, we obtain a dually flat Riemannian structure. The flatness of the geometrical structure induced by a Bregman Divergence simplifies considerably the computation of geodesic paths between distributions, facilitating in this way the computation of the distance between two distributions.

In addition to the properties of Bregman Divergences mentioned in Section 2.1.1, further properties can be derived by the connections between Bregman divergence and the geometrical structures derived therefrom. In particular a canonical divergence, a generalized Pythagorean theorem and a projection theorem, refer to Amari (2009a,b); Amari and Cichocki (2010) for further details.

Another interesting way to measure the similarity between two probability measures is by using the structure of a Reproducing Kernel Hilbert Space, as is explained in next section.

2.2.2 Dissimilarity measures in the RKHS framework

It is possible to define distances between sets of points, curves, surfaces, distribution functions and even more general objects in a Reproducing Kernel Hilbert Space (RKHS). Next we give

the basic definitions and properties of a RKHS Aronszajn (1950), more details about RKHS in the context of Regularization Problems can be seen in Wahba (1990); Poggio and Smale (2003); Poggio and Shelton (2002).

Definition 2.3. (**RKHS**): A Reproducing Kernel Hilbert Space, denoted as $\mathcal{H}_K$, is a Hilbert Space $\mathcal{H}$ of functions defined on a compact domain X where every linear evaluation functional $F_x : \mathcal{H} \to \mathbb{R}$ is bounded: there exists $M > 0$ such that

$$\mid F_x(f(x)) \mid = \mid f(x) \mid \leq M\|f(x)\|, \tag{2.2}$$

where $\|.\|$ is the norm in the Hilbert space $\mathcal{H}$.

Definition 2.4. (**Mercer Kernel**): Let X be a compact domain and $K : X \times X \to \mathbb{R}$ a continuous and symmetric function. If we assume that the matrix $\boldsymbol{K}|_{\mathbf{x}}$ is positive definite, that is, for any arbitrary set $\mathbf{x} = \{x_1, ..., x_n\} \in X$ the matrix $\boldsymbol{K}|_{\mathbf{x}}$ with elements $(\boldsymbol{K}|_{\mathbf{x}})_{i,j} = K(x_i, x_j)$ is a positive definite matrix, then K is a Mercer Kernel.

The Moore-Aronszajn (1950) theorem establish a one to one correspondence between positive definite kernels and Reproducing Kernel Hilbert Spaces. For each RKHS space of functions on X there exists a **unique** reproducing kernel K which is positive definite. Conversely, any RKHS can be characterized by a positive definite Kernel.

Theorem 2.1. *(**Generation of RKHSs from Kernels**): Let X be a compact domain and let $K : X \times X \to \mathbb{R}$ be a continuous, symmetric and positive definite function. Define $K_x : X \to \mathbb{R}$ as the function $K_x(t) = K(x, t)$. Then for every K there exists a unique RKHS $(\mathcal{H}_K, \langle ., . \rangle_{\mathcal{H}_K})$ of functions on X satisfying that:*

- *For all $x \in X$, $K_x \in \mathcal{H}_K$*
- *The span of $\{K_x : x \in X\}$ is dense in $\mathcal{H}_K$*
- *For all f in $\mathcal{H}_K$ and $x \in X$ then $f(x) = \langle K_x, f \rangle_{\mathcal{H}_K}$*

We can construct $\mathcal{H}_K$ given a kernel function K. Let $\mathcal{H}$ be the space of functions spanned by finite linear combinations: $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ where $n \in \mathbb{N}$, $x_i \in X$ and $\alpha_i \in \mathbb{R}$ and define the inner product

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j K(x_i, x_j), \tag{2.3}$$

for $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ and $g(x) = \sum_{j=1}^{n} \beta_j K(x_j, x)$. Then $\mathcal{H}_K$ is the completion of $\mathcal{H}$. Now if $\mathcal{H}$ is a RKHS then by Riesz representation theorem there exists a unique function $K_x \in \mathcal{H}$, such

that $\langle K_x, f\rangle_{\mathcal{H}} = f(x)$ for all $x \in X$. The function K_x is called the point evaluation functional at the point x. Aronszajn (1950) proved that this function exists, is unique, symmetric and positive definite (is a Mercer Kernel).

Next we show a connection between the theory of reproducing kernels and integral operators via the Mercer theorem, for further details see for example Aronszajn (1950); Berlinet and Thomas-Agnan (2004). Let X be a compact metric space, equipped with a finite, strictly positive Borel measure ν and let K be a positive definite kernel as in Definition 2.4 satisfying

$$\|K\|_\infty = \sup_{x\in X} \sqrt{K(x,x)} < \infty. \tag{2.4}$$

Let $L^2_\nu(x)$ be the space of square integrable functions in X where ν is a Borel measure, the linear map $L_K : L^2_\nu(x) \to L^2_\nu(x)$ defined by the integral operator:

$$(L_K(f))(x) = \int_X K(x,t)f(t)d\nu(t),$$

is well defined and the function K is called the Kernel of L_K. If K is a Mercer Kernel then L_K is a self adjoint, positive, compact operator with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq 0$. By the Spectral theorem, the corresponding set of eigenfunctions $\{\phi_i(x)\}_{i=1}^\infty$ form an orthonormal basis for $L^2_\nu(X)$, where

$$\phi_i(x) = \frac{1}{\lambda_i}\int_X K(x,t)\phi_i(t)d\nu(t). \tag{2.5}$$

Therefore for any pair of eigenfunctions $\{\phi_i(x), \phi_j(x)\}$ we have the following results:

- $\|\phi_i\|_{L^2_\nu(X)} = 1$,
- $\langle \phi_i(x), \phi_j(x)\rangle = \delta_{ij}$ where $\delta_{ij} = 1$ when $i = j$ and 0 otherwise,
- For any $f \in L^2_\nu(X)$ then $f(x) = \sum_{j=1}^\infty \langle f, \phi_j\rangle \phi_j$.

Theorem 2.2. *(**Mercer's Theorem**): Let X be a compact domain, ν a non degenerate Borel measure in X, and $K : X \times X \to \mathbb{R}$ a Mercer kernel. Let $\{\lambda_i\}_{i\geq 1}$ the eigenvalues of L_K and $\{\phi_i\}_{i\geq 1}$ the corresponding eigenfunctions. Then, for all $x, y \in X$;*

$$K(x;y) = \sum_{i=1}^\infty \lambda_i \phi_i(x)\phi_i(y), \tag{2.6}$$

where the series converges absolutely (for each $x, y \in X$) and uniformly (in $x, y \in X$).

The Theorem 2.2 allows us to interpret K as a dot product in the feature space via the map $\mathbf{\Phi}$ as follows:

$$\begin{aligned} &\mathbf{\Phi} : X \to l^2, \\ &x \mapsto \phi(x) = (\sqrt{\lambda_i}\phi_i(x))_{i \in \mathbb{N}}, \end{aligned}$$

where l^2 is the linear space of square summable sequences. Now according to Theorem 2.2 for all $x, y \in X \times X$ then $K(x, y) = \langle \mathbf{\Phi}(x), \mathbf{\Phi}(y) \rangle$. Thus K acts as a dot product in the embedding (the image of the often nonlinear mapping $\mathbf{\Phi}$).

Some examples of Mercer Kernels are:

- *Linear*: $K(x, y) = x^T y$,
- *Polynomial*: $K(x, y) = \langle x, y \rangle^d$,
- *Gaussian*: $K(x, y) = \exp(-\frac{\| x - y \|^2}{2\sigma^2})$ where $\sigma > 0$.

Next we explore the connection between RKHS and dissimilarity measures for sets of points and statistical distributions. These measures are the core of a large list of techniques based on distances in Statistics. In first place we demonstrate how to use RKHS to induce a distance for sets of points.

Definition 2.5. (**Kernel Dissimilarity for Sets of Points**): Let $S_{\mathbb{P}}^n = \{x_i\}_{i=1}^n$ and $S_{\mathbb{Q}}^m = \{y_i\}_{i=1}^m$ be two sets of points independently drawn from the probability measures $\mathbb{P}$ and $\mathbb{Q}$ respectively. The *kernel dissimilarity* induced by a positive definite kernel K between the sets $S_{\mathbb{P}}^n$ and $S_{\mathbb{Q}}^m$ is computed as

$$D_K^2(S_{\mathbb{P}}^n, S_{\mathbb{Q}}^m) = \underbrace{\sum_{x \in S_{\mathbb{P}}^n} \sum_{x' \in S_{\mathbb{P}}^n} K(x, x') + \sum_{y \in S_{\mathbb{Q}}^m} \sum_{y' \in S_{\mathbb{Q}}^m} K(y, y')}_{\text{self similarity}} \underbrace{- 2 \sum_{x \in S_{\mathbb{P}}^n} \sum_{y \in S_{\mathbb{Q}}^m} K(x, y)}_{\text{cross similarity}}, \tag{2.7}$$

where the kernel K is a symmetric, positive definite similarity measure that satisfy the condition $K(x, y) = 1$ if and only if $x = y$ and $K(x, y) \to 0$ when the distance between the points x and y increases.

Metrics induced by kernels can be rewritten in terms of the associated feature map $\mathbf{\Phi}$. Using the linearity of the inner product and Mercer's theorem,we can express the kernel similarity in terms of the lifting map $\mathbf{\Phi}$. As an example consider the kernel metric for points in Definition 2.5, thus we have:

$$\begin{aligned}
D_K^2(S_{\mathbb{P}}^n, S_{\mathbb{Q}}^m) &= \sum_{x\in S_{\mathbb{P}}}\sum_{x'\in S_{\mathbb{P}}} K(x,x') + \sum_{y\in S_{\mathbb{Q}}}\sum_{y'\in S_{\mathbb{Q}}} K(y,y') - 2\sum_{x\in S_{\mathbb{P}}}\sum_{y\in S_{\mathbb{Q}}} K(x,y), \\
&= \sum_{x\in S_{\mathbb{P}}}\sum_{x'\in S_{\mathbb{P}}} \langle \mathbf{\Phi}(x), \mathbf{\Phi}(x')\rangle + \sum_{y\in S_{\mathbb{Q}}}\sum_{y'\in S_{\mathbb{Q}}} \langle \mathbf{\Phi}(y), \mathbf{\Phi}(y')\rangle - 2\sum_{x\in S_{\mathbb{P}}}\sum_{y\in S_{\mathbb{Q}}} \langle \mathbf{\Phi}(x), \mathbf{\Phi}(y)\rangle, \\
&= \left\langle \sum_{x\in S_{\mathbb{P}}} \mathbf{\Phi}(x), \sum_{x\in S_{\mathbb{P}}} \mathbf{\Phi}(x) \right\rangle + \left\langle \sum_{y\in S_{\mathbb{Q}}} \mathbf{\Phi}(y), \sum_{y\in S_{\mathbb{Q}}} \mathbf{\Phi}(y) \right\rangle - 2\left\langle \sum_{x\in S_{\mathbb{P}}} \mathbf{\Phi}(x), \sum_{y\in S_{\mathbb{Q}}} \mathbf{\Phi}(y) \right\rangle, \\
&= \| \sum_{x\in S_{\mathbb{P}}} \mathbf{\Phi}(x) - \sum_{y\in S_{\mathbb{Q}}} \mathbf{\Phi}(y) \|^2 . \qquad (2.8)
\end{aligned}$$

Therefore we can adopt the following definition:

Definition 2.6. (**Hilbertian Metric Kernel Dissimilarity**): Let $S_{\mathbb{P}}^n = \{x_i\}_{i=1}^n$ and $S_{\mathbb{Q}}^m = \{y_i\}_{i=1}^m$ be two sets of points independently drawn from the probability measures $\mathbb{P}$ and $\mathbb{Q}$ respectively and a positive definite kernel K. The *kernel distance* induced by a positive definite kernel K between the sets $S_{\mathbb{P}}^n$ and $S_{\mathbb{Q}}^m$ is defined as

$$D_K^2(S_{\mathbb{P}}^n, S_{\mathbb{Q}}^m) = \| \sum_{x\in S_{\mathbb{P}}} \mathbf{\Phi}(x) - \sum_{y\in S_{\mathbb{Q}}} \mathbf{\Phi}(y) \| . \qquad (2.9)$$

There are important consequences of recomputing the distance in this way:

- The kernel distance embeds isometrically in an Euclidean space. While in general $\mathcal{H}$ might be infinite dimensional, the Hilbert space structure implies that for any finite collection of inputs, the effective dimensionality of the space can be reduced via projection to a much smaller finite number.
- Most analysis problems are "easier" in Euclidean spaces. This includes problems like clustering and regressions. The embedding of the kernel in such spaces allows us to represent complex functional objects in a single vector: $\mathbf{\Phi}(S_{\mathbb{P}}) = \sum_{x\in S_{\mathbb{P}}} \mathbf{\Phi}(x)$ in the RKHS.
- The embedding "linearises" the metric by mapping the input space to a vector space. Then many problems in functional data analysis can be solved easily by exploiting the linear structure of the lifted space.

- Computational cost in large scale problems is reduced. If $\mathcal{H}$ is approximated (assuming a fixed error) with a $\rho \ll n$ dimensional space, then in this space the operational cost for computing the kernel distance between two point sets of total size n is reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n\rho)$.

The definition of the kernel dissimilarity for sets of points can be extended to a kernel dissimilarity for density functions in the following way.

Definition 2.7. (**Kernel Dissimilarity for Density Functions**): Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures and denote by $f_{\mathbb{P}}$ and $f_{\mathbb{Q}}$ the respective density functions. The *kernel similarity* induced by a positive definite kernel K between $\mathbb{P}$ and $\mathbb{Q}$ is computed as

$$D_K^2(\mathbb{P},\mathbb{Q}) = \underbrace{\int_X\int_X f_{\mathbb{P}}(x)K(x,y)f_{\mathbb{P}}(y)dxdy + \int_X\int_X f_{\mathbb{Q}}(x)K(x,y)f_{\mathbb{Q}}(y)dxdy}_{\text{self similarity}} - \underbrace{2\int_X\int_X f_{\mathbb{P}}(x)K(x,y)f_{\mathbb{Q}}(y)dxdy}_{\text{cross similarity}}.$$

We can relate the dissimilarity given in Definition 2.7 to standard metrics in functional analysis. For example, when $K(x,y) = \delta(x-y)$, where δ is the Dirac delta generalized function, then $D_K^2(\mathbb{P},\mathbb{Q})$ is the Euclidean distance (in the L_2 sense) between $\mathbb{P}$ and $\mathbb{Q}$. There is a sufficient condition to ensure that the dissimilarity measure induced by a kernel function is a distance: as was pointed out by Gretton et al. (2006); Sriperumbudur et al. (2010b), K must be a strictly integrable positive definite kernel. Otherwise the dissimilarity measures induced by kernel functions are pseudometrics. More details about distances and (dis)similarities induced by kernel functions can be seen in Phillips and Venkatasubramanian (2011); Zhou and Chellappa (2006); Scholkopf (2001).

In the most general case, for example when one deals with sets of curves or surfaces, one has to consider an alternative free coordinate system as is described in Bachman (2012). Denote by $t_P(p)$ to the *unit tangent vector* at the point p over the variety $\mathcal{P}$ and $t_Q(q)$ to the unit tangent vector at the point q over the variety $\mathcal{Q}$ (in this context $\mathcal{P}$ and $\mathcal{Q}$ represents general curves or surfaces). Then the *pointwise* kernel similarity between $p \in \mathcal{P}$ and $q \in \mathcal{Q}$ is given by $K(p,q)\langle t_P(p), t_Q(q)\rangle$. Therefore in order to obtain a distance induced by a kernel function between $\mathcal{P}$ and $\mathcal{Q}$, also known as a *current distance*, we simply integrate over the differential

form on $\mathcal{P} \times \mathcal{Q}$ and thus:

$$D_K^2(\mathcal{P}, \mathcal{Q}) = \iint_{\mathcal{P}\times\mathcal{P}} K(p, p')\langle t_P(p), t_P(p')\rangle + \iint_{\mathcal{Q}\times\mathcal{Q}} K(q, q')\langle t_Q(q), t_Q(q')\rangle - 2\iint_{\mathcal{P}\times\mathcal{Q}} K(p, q)\langle t_P(p), t_Q(q)\rangle.$$

We will use the theory of RKHS several times along the thesis in order to introduce different distance measures. For example, in next chapter we make use of density kernels to introduce a new distance measure from a point to the center of a distribution that generalize the Mahalanobis distance to cases when the distribution of the data is not Gaussian.

Chapter 3

On the Generalization of the Mahalanobis Distance[1]

Chapter abstract

The Mahalanobis distance (MD) is a distance widely used in Statistics, Machine Learning and Pattern Recognition in general. When the data come from a Gaussian distribution, the MD uses the covariance matrix to evaluate the distance between a data point and the distribution mean. In this chapter we generalize the MD for general unimodal distributions introducing a particular class of Mercer kernel, the density kernel, based on the underlying data density. Density kernels induce distances that generalize the MD and that are useful when data do not fit to the Gaussian distribution. We study the theoretical properties of the proposed distance and show its performance on a variety of artificial and real data analysis problems.

Chapter keywords: Mahalanobis distance, Bergman divergences, exponential family, density kernel, level sets, outlier detection, classification.

3.1 Introduction

The Mahalanobis distance (MD) Mahalanobis (1936); De Maesschalck et al. (2000), is a scale-invariant metric that provides a measure of distance between a point $\mathbf{x} \in \mathbb{R}^d$ generated from a given (probability) distribution $\mathbb{P}$ and the mean $\boldsymbol{\mu} = \mathbb{E}_{\mathbb{P}}(\mathbf{x})$ of the distribution. Assume $\mathbb{P}$ has

[1] The contents of this chapter are published in the Journal of Intelligent Data Analysis (Martos et al., 2014) and in the Proceedings of the Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications conference (Martos et al., 2013)

finite second order moments and let us denote by $\mathbf{\Sigma} = \mathbb{E}_{\mathbb{P}}\left((\mathbf{x}-\boldsymbol{\mu})^2\right)$ the covariance matrix. Then the MD is defined by:

$$d_M(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

The Mahalanobis distance arises naturally in the problem of comparing two populations (distributions) with the same covariance matrix. Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be the two distributions, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ the respective mean vectors and $\mathbf{\Sigma}$ the common covariance matrix. Then the Mahalanobis distance between the two populations (distributions) $\mathbb{P}_1$ and $\mathbb{P}_2$ is computed as:

$$d_M(\mathbb{P}_1, \mathbb{P}_2) = \sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}.$$

In practice we are not given neither the theoretical mean vectors nor the covariance matrix, and we have to work with samples. Given two samples $\{\mathbf{x}_{1,i}\}_{i=1}^n$ (from $\mathbb{P}_1$) and $\{\mathbf{x}_{2,i}\}_{i=1}^m$ (from $\mathbb{P}_2$) in $\mathbb{R}^d$, denote by $\bar{\mathbf{x}}_i$ the sample estimator of $\boldsymbol{\mu}_i$, $i = 1, 2$, and by $\mathbf{S}$ the sample estimator of $\mathbf{\Sigma}$; then the sample estimation of the distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ is:

$$\hat{d}_M(\mathbb{P}_1, \mathbb{P}_2) = \sqrt{(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)}.$$

Now consider a classical discrimination problem: Given two normally distributed populations/distributions denoted as $\mathbb{P}_1$ and $\mathbb{P}_2$, with the same covariance matrix, and a point $\mathbf{x} \in \mathbb{R}^d$, we want to classify $\mathbf{x}$ as belonging to $\mathbb{P}_1$ or to $\mathbb{P}_2$. Then the discrimination functions can be expressed by $y_i(\mathbf{x}) = \ln p(\mathbf{x}|\mathbb{P}_i) \propto d_M(\mathbf{x}, \boldsymbol{\mu}_i)^2$, $i = 1, 2$ Flury (1997). In this way, $\mathbf{x}$ will be assigned to the class with highest discriminant function value or, equivalently, to the class with smallest MD. Thus, for classification problems, the true knowledge of $p(\mathbf{x}|\mathbb{P})$ is essential to classify correctly new data points.

As we have just shown, the MD estimates adequately (the logarithm of) such probability in the case of normal distributions (the case of non-equal covariance matrices is slightly different but the same conclusions apply). Next we show that MD fails to estimate $p(\mathbf{x}|\mathbb{P})$ if the involved distributions are not normal anymore.

Figure 3.1-a) illustrates a graphical interpretation of the MD behaviour: points $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to the same level set of the distribution, that is: $p(\mathbf{x}_1|\mathbb{P}) = p(\mathbf{x}_2|\mathbb{P})$ and $p(\mathbf{x}_3|\mathbb{P}) < p(\mathbf{x}_1|\mathbb{P})$. The MD can be interpreted as the composition of the linear transformation $T_M : \mathbf{x} \xrightarrow{T_M} \mathbf{x}' = S^{-\frac{1}{2}}\mathbf{x}$, plus the computation of the ordinary Euclidean distance (ED) between the transformed

data point and the center μ of the distribution (this can be seen with the aid of Figure 3.1-b)). As expected, the MD preserves the order between level set curves. In particular, before transforming, $d_E(\mathbf{x}_1, \boldsymbol{\mu}) > d_E(\mathbf{x}_3, \boldsymbol{\mu}) > d_E(\mathbf{x}_2, \boldsymbol{\mu})$ but $P(\mathbf{x}_1|\mathbb{P}) = P(\mathbf{x}_2|\mathbb{P}) > P(\mathbf{x}_3|\mathbb{P})$, that is, the ED fails to correctly rank the 3 data points. After transforming (equivalently, using the MD), the order given by the level sets and distances are the same.

In Figure 3.1-c) we consider again three points generated from a (now) non normal distribution satisfying that $P(\mathbf{x}_1|\mathbb{P}) = P(\mathbf{x}_2|\mathbb{P}) > P(\mathbf{x}_3|\mathbb{P})$. However, in this case, $\boldsymbol{\mu}$ does not coincide with the mode (the densest point), and the MD fails to reflect the order given by the level sets, which gives the correct rule to classify the data points: $d_M(\mathbf{x}_1, \boldsymbol{\mu}) > d_M(\mathbf{x}_3, \boldsymbol{\mu}) > d_M(\mathbf{x}_2, \boldsymbol{\mu})$.

We propose in this chapter introduce of a family of kernels based on the underlying density function of the sample at hand. The proposed density kernels induces new distances that generalize the Mahalanobis distance. The family of distances proposed in this chapter preserve the essential property of the Mahalanobis distance: "all the points that belong to the same probability curve, that is $L_c(f_{\mathbb{P}}) = \{\mathbf{x}|f_{\mathbb{P}}(\mathbf{x}) = c\}$ where $f_{\mathbb{P}}$ is the density function of the r.v. $\mathbf{x}$, are equally distant from the center (the densest point) of the distribution".

The chapter is organized as follows: In Section 3.2 we introduce density kernels and the Generalized Mahalanobis distance, a distributional distance, induced by the density kernels. We provide a computable version on of the proposed distance based on the estimation of level sets. In Section 3.3 we show the performance of the generalized MD for a battery of outlier detection and classification problems.

3.2 Generalizing the Mahalanobis Distance via Density Kernels

In this section we introduce a family of distances induced by a specific family of kernels defined below.

3.2.1 Distances induced by density kernels

Consider a measure space $(X, \mathcal{F}, \mu)$, where X is a sample space (here a compact set of $\mathbb{R}^d$), $\mathcal{F}$ a σ-algebra of measurable subsets of X and $\mu : \mathcal{F} \to \mathbb{R}^+$ the ambient σ-additive measure, the Lebesgue measure. A probability measure $\mathbb{P}$ is a σ-additive finite measure absolutely continuous w.r.t. μ that satisfies the three Kolmogorov axioms. By Radon-Nikodym theorem, there exists a measurable function $f_{\mathbb{P}} : X \to \mathbb{R}^+$ (the density function) such that $\mathbb{P}(A) = \int_A f_{\mathbb{P}} d\mu$,

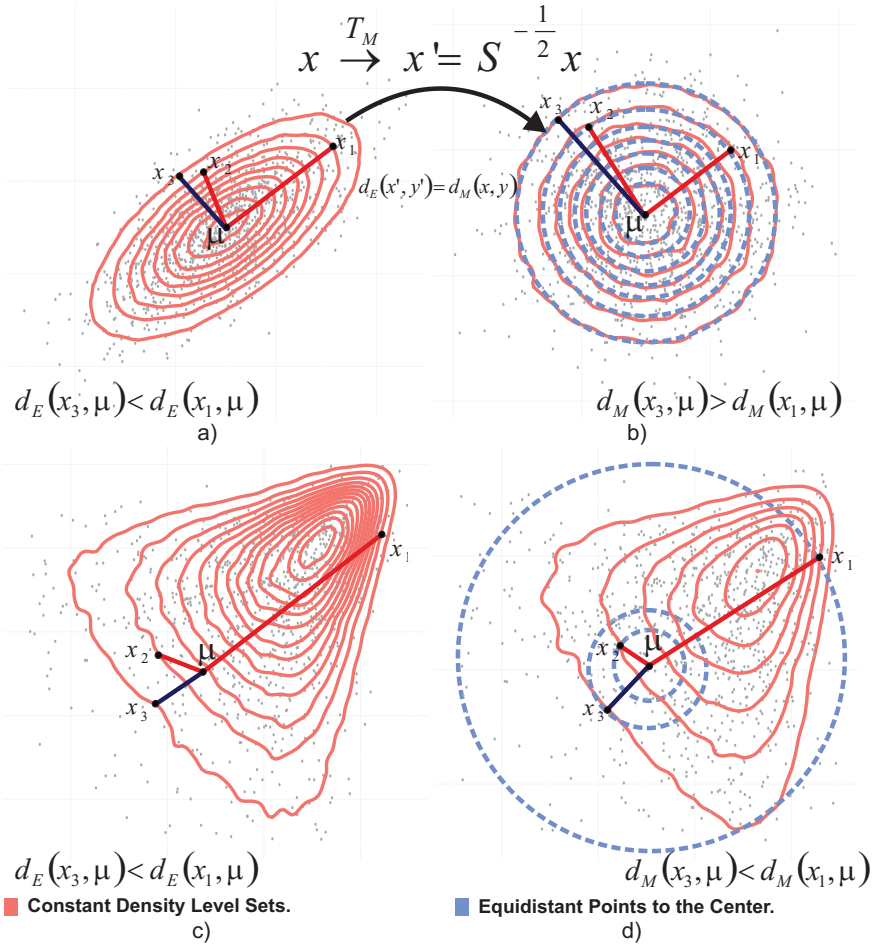


Figure 3.1: $a)$ Sample points from a normal distribution and level sets. $b)$ Sample points after Mahalanobis transformation. $c)$ Sample points from a non normal distribution and level sets. $b)$ Sample points after Mahalanobis transformation.

and $f_{\mathbb{P}} = \frac{d\mathbb{P}}{d\mu}$ the Radon-Nikodym derivative.

In this chapter we assume that the density function is unimodal. Before introducing the concept of density kernel, we elaborate on the definitions of f-monotone and asymptotic f-monotone functions.

Definition 3.1 (**f-monotonicity**). Let $\mathbf{X}$ be a random vector in $\mathbb{R}^d$ that follows the distribution induced by the probability measure $\mathbb{P}$ and denote by $f_{\mathbb{P}} : X \to \mathbb{R}^+$ the corresponding density function. A function $g : \mathcal{X} \to \mathbb{R}$ is f-monotone if:

$$f_{\mathbb{P}}(\mathbf{x}) \geq f_{\mathbb{P}}(\mathbf{y}) \Rightarrow g(\mathbf{x}, \mathbb{P}) \geq g(\mathbf{y}, \mathbb{P}).$$

The notation $g(\mathbf{x}, \mathbb{P})$ indicates that the function g depends on some explicit parameters of the $\mathbb{P}$ distribution. Next we give examples of f-monotone functions.

Example 3.1. Assume that the random vector $\mathbf{X}$ follows a d-dimensional multivariate normal distribution. In this case $f_{\mathbb{P}}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$, where $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix. The following functions g_1, g_2, g_3 and g_4 are f-monotone:

i) $g_1(\mathbf{x}, \mathbb{P}) = g_1(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \propto f_{\mathbb{P}}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$,

which is proportional to the density function. Also considers:

ii) $g_2(\mathbf{x}, \mathbb{P}) = g_2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -d_M^2(\mathbf{x}, \boldsymbol{\mu})$,

which is the (negative square) Mahalanobis distance from the point $\mathbf{x}$ to the mean of the distribution $\mathbb{P}$. Let ζ be a continuously-differentiable real-valued and strictly convex function, then:

iii) $g_3(\mathbf{x}, \mathbb{P}, \zeta) = e^{-\zeta(f_{\mathbb{P}}(\mathbf{x})) + \zeta(f_{\mathbb{P}}(\boldsymbol{\mu})) + \zeta'(f_{\mathbb{P}}(\boldsymbol{\mu}))(f_{\mathbb{P}}(\mathbf{x}) - f_{\mathbb{P}}(\boldsymbol{\mu}))} = e^{-BD_\zeta(f_{\mathbb{P}}(\mathbf{x}), f_{\mathbb{P}}(\boldsymbol{\mu}))}$,

where ζ' denotes the first derivative of the function ζ. The function $g_3(\mathbf{x}, \mathbb{P}, \zeta)$ is the exponential Bregman divergence and obeys the f-monotonicity property. For the last example of a f-monotone function consider the rational quadratic kernel $K_f(\mathbf{x}, \boldsymbol{\mu}) = 1 - \frac{||f_{\mathbb{P}}(\mathbf{x}) - f_{\mathbb{P}}(\boldsymbol{\mu})||^2}{||f_{\mathbb{P}}(\mathbf{x}) - f_{\mathbb{P}}(\boldsymbol{\mu})||^2 + c}$ where c is a positive constant, then:

iv) $g_4(\mathbf{x}, \mathbb{P}) = g_4(\mathbf{x}, \boldsymbol{\mu}) = K_f(\mathbf{x}, \boldsymbol{\mu})$,

which also has the f-monotone property.

• • •

In practice $\mathbb{P}$ is unknown and only a random sample $S_n = \{\mathbf{x}_i\}_{i=1}^n$ is available. Therefore we need a working definition $g(\mathbf{x}, \mathbb{P})$ when $\mathbb{P}$ is not explicitly known. Next, we provide the sample counterpart of Definition 3.1.

Definition 3.2 (asymptotic f-monotonicity). Consider a random sample $S_n = \{\mathbf{x}_i\}_{i=1}^n$ drawn from $\mathbb{P}$. A function $g(\mathbf{x}, S_n)$ is asymptotically f-monotone if:

$$f_{\mathbb{P}}(\mathbf{x}) \geq f_{\mathbb{P}}(\mathbf{y}) \Rightarrow \lim_{n \to \infty} P\left(g(\mathbf{x}, S_n) \geq g(\mathbf{y}, S_n)\right) = 1.$$

We obtain asymptotically f-monotone functions substituting the parameters in $g(\mathbf{x}, \mathbb{P})$ by its sample estimations as in Example 3.2.

Example 3.2. Consider a random vector $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $S_n = \{\mathbf{x}_i\}_{i=1}^n$ be a independent random sample drawn from $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then:

$$\begin{aligned} g_1(\mathbf{x}, S_n) &= e^{-\frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}})}, \\ g_2(\mathbf{x}, S_n) &= -(\mathbf{x}-\hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}), \\ g_3(\mathbf{x}, S_n, \zeta) &= e^{-BD_\zeta(f_{\mathbb{P}}(\mathbf{x}), f_{\mathbb{P}}(\hat{\boldsymbol{\mu}}))}, \\ g_4(\mathbf{x}, S_n) &= K_f(\mathbf{x}, \hat{\boldsymbol{\mu}}), \end{aligned}$$

are asymptotic f-monotone functions, where $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are consistent sample estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively.

• • •

Using the convergence of the parametric estimations: $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \xrightarrow{P} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and assuming also that the function g is continuous with respect to $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have:

$$\lim_{n\to\infty} P\left(|g_i(\mathbf{x}, S_n) - g_i(\mathbf{x}, \mathbb{P})| \leq \varepsilon\right) = 1,$$

therefore $g_i(\mathbf{x}, S_n) \xrightarrow{P} g_i(\mathbf{x}, \mathbb{P})$ for $i = 1, 2, 3, 4$ (converges in probability). Now, for any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $f_{\mathbb{P}}(\mathbf{x}) > f_{\mathbb{P}}(\mathbf{y})$, then $\lim_{n\to\infty} P\left(g(\mathbf{x}, S_n) > g(\mathbf{y}, S_n)\right) = 1$ and thus, the functions g_1, g_2, g_3 and g_4, given in Example 3.2, are asymptotic f-monotone functions.

Other non parametric asymptotic f-monotone functions can be considered, for example $g_5(\mathbf{x}, S_n) \propto \hat{f}_{S_n}(\mathbf{x})$, where $\hat{f}_{S_n}$ can be any consistent and nonparametric estimator of the density $f_{\mathbb{P}}$. Another example is $g_6(\mathbf{x}, S_n, k) = e^{-d_{S_n,k}(\mathbf{x})}$, where $d_{S_n,k}(\mathbf{x})$, is the distance from the point $\mathbf{x}$ to its k-nearest neighbours.

The function g is a *positive* (asymptotic) f-monotone functions if $g(\mathbf{x}, \mathbb{P}) \geq 0$ ($g(\mathbf{x}, S_n) \geq 0$) for all $\mathbf{x}$ in the support of $\mathbb{P}$. The function g is a *negative* (asymptotic) f-monotone functions if $-g$ is a positive (asymptotic) f-monotone function. Now we are ready to introduce the concept of density kernel.

Definition 3.3 (Density kernel). Let $\mathbf{X}$ be a random vector in $\mathbb{R}^d$ that follows a distribution according to the probability measure $\mathbb{P}$ and let $g(\mathbf{x}, \mathbb{P})$ be a positive f-monotone function. Consider the mapping $\phi : X \to \mathbb{R}^+$ given by $\phi_{\mathbb{P}}(\mathbf{x}) = g(\mathbf{x}, \mathbb{P})$. The density kernel is defined as:

$$K_{\mathbb{P}}(\mathbf{x}, \mathbf{y}) = \phi_{\mathbb{P}}(\mathbf{x})\phi_{\mathbb{P}}(\mathbf{y}).$$

$K_{\mathbb{P}}$ is a positive definite kernel, that is, a Mercer kernel (see the appendix A for a proof). The associated density kernels to the functions g_1, g_2, g_3 and g_4 defined in Example 3.1 are as follows. For $g_1(\mathbf{x}, \mathbb{P})$ we get:

$$K_{\mathbb{P}}^{[1]}(\mathbf{x}, \mathbf{y}) = e^{-\frac{1}{2}(d_M^2(\mathbf{x},\boldsymbol{\mu})+d_M^2(\mathbf{y},\boldsymbol{\mu}))},$$

which is the exponential mean of the Mahalanobis distance. For $g_2(\mathbf{x}, \mathbb{P})$:

$$K_{\mathbb{P}}^{[2]}(\mathbf{x}, \mathbf{y}) = 1 + d_M^2(\mathbf{x}, \boldsymbol{\mu}) d_M^2(\mathbf{y}, \boldsymbol{\mu}),$$

where we add a constant in order to normalize the kernel. For $g_3(\mathbf{x}, \mathbb{P})$:

$$K_{\mathbb{P}}^{[3]}(\mathbf{x}, \mathbf{y}) = e^{-\frac{1}{2}(BD_\zeta(f(\mathbf{x}),f(\boldsymbol{\mu}))+BD_\zeta(f(\mathbf{y}),f(\boldsymbol{\mu})))},$$

is the exponential mean of the Bregman divergences. For $g_4(\mathbf{x}, \mathbb{P})$:

$$K_{\mathbb{P}}^{[4]}(\mathbf{x}, \mathbf{y}) = K_f(\mathbf{x}, \boldsymbol{\mu}) K_f(\mathbf{y}, \boldsymbol{\mu}),$$

is the product of two kernel functions, that is, another kernel function.

3.2.2 Dissimilarity measures induced by density kernels

Next we study dissimilarities induced by the density kernels defined above.

Definition 3.4 (Density kernel dissimilarity). Let $\mathbf{X}$ be a random vector in $\mathbb{R}^d$ that follows a distribution according to the probability measure $\mathbb{P}$ and let $K_{\mathbb{P}}(\mathbf{x}, \mathbf{y})$ be a density kernel, then define the density kernel (squared) dissimilarity function as:

$$d^2_{K_{\mathbb{P}}}(\mathbf{x}, \mathbf{y}) = -\log K_{\mathbb{P}}(\mathbf{x}, \mathbf{y}).$$

For the proposed f-monotone functions of Example 3.1 we get:

$$\begin{aligned}
d^2_{K_{\mathbb{P}}^{[1]}}(\mathbf{x}, \mathbf{y}) &= \frac{1}{2}(d_M^2(\mathbf{x}, \boldsymbol{\mu}) + d_M^2(\mathbf{y}, \boldsymbol{\mu})), \\
d^2_{K_{\mathbb{P}}^{[2]}}(\mathbf{x}, \mathbf{y}) &= -\log(1 + d_M^2(\mathbf{x}, \boldsymbol{\mu}) d_M^2(\mathbf{y}, \boldsymbol{\mu})), \\
d^2_{K_{\mathbb{P}}^{[3]}}(\mathbf{x}, \mathbf{y}) &= \frac{1}{2}(BD_\zeta(f_{\mathbb{P}}(\mathbf{x}), f_{\mathbb{P}}(\boldsymbol{\mu})) + BD_\zeta(f_{\mathbb{P}}(\mathbf{y}), f_{\mathbb{P}}(\boldsymbol{\mu}))), \\
d^2_{K_{\mathbb{P}}^{[4]}}(\mathbf{x}, \mathbf{y}) &= \log\left(\frac{1}{K_f(\mathbf{x}, \boldsymbol{\mu}) K_f(\mathbf{y}, \boldsymbol{\mu})}\right).
\end{aligned}$$

The measures $d^2_{K^{[i]}_{\mathbb{P}}}(\mathbf{x},\mathbf{y})$ for $i = 1,2,3,4$ are non-negative: $d^2_{K^{[i]}_{\mathbb{P}}}(\mathbf{x},\mathbf{y}) \geq 0$, and symmetric: $d^2_{K^{[i]}_{\mathbb{P}}}(\mathbf{x},\mathbf{y}) = d^2_{K^{[i]}_{\mathbb{P}}}(\mathbf{y},\mathbf{x})$.

We are particularly interested in the definition and computation of a distance from a point to the center of a distribution in order to solve classification and outlier detection problems. Therefore, following the spirit of the Mahalanobis distance, we propose the following definition of a generalized Mahalanobis (GM) distance associated to a density kernel $K_{\mathbb{P}}$.

Definition 3.5 (GM distance associated to $K_{\mathbb{P}}$). Let $\mathbf{X}$ be a random vector in $\mathbb{R}^d$ that follows a distribution according to the probability measure $\mathbb{P}$ and let $K_{\mathbb{P}}(\mathbf{x},\mathbf{y})$ be a density kernel. Then define the GM distance associated to the density kernel $K_{\mathbb{P}}$, from a point $\mathbf{x}$ to the center of the distribution $\mathbb{P}$, by:

$$d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x},\mathbf{m_o}) = -\log K_{\mathbb{P}}(\mathbf{x},\mathbf{m_o}),$$

where $\mathbf{m_o}$ is the mode of the distribution: $\mathbf{m_o} = \max_{\mathbf{x}} f_{\mathbb{P}}(\mathbf{x})$.

Proposition 3.1 (Generalization of Mahalanobis distance). *The Mahalanobis distance is a particular case of the $d_{GM_{K_{\mathbb{P}}}}$ distance.*

Proof. Let $\phi_{\mathbb{P}}(\mathbf{x}) = e^{-(\mathbf{x}-\mathbf{m_o})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_o})}$, by Definition 3.3:

$$\begin{aligned} d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x},\mathbf{m_o}) &= -\log K_{\mathbb{P}}(\mathbf{x},\mathbf{m_o}), \\ &= -\log\left(\phi_{\mathbb{P}}(\mathbf{x})\phi_{\mathbb{P}}(\mathbf{m_o})\right), \\ &= -\log e^{-(\mathbf{x}-\mathbf{m_o})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_o})}, \\ &= (\mathbf{x}-\mathbf{m_o})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_o}), \\ &= d^2_M(\mathbf{x},\mathbf{m_o}). \end{aligned}$$

□

The distance $d_{GM_{K_{\mathbb{P}}}}$ has the following property, characteristic of the Mahalanobis Distance:

Proposition 3.2 (Density coherence). *If $f_{\mathbb{P}}(\mathbf{x}) = f_{\mathbb{P}}(\mathbf{y})$ then $d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x},\mathbf{m_o}) = d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{y},\mathbf{m_o})$.*

Proof. This property is obtained by using Definition 3.1: Let $f_{\mathbb{P}}(\mathbf{x}) = f_{\mathbb{P}}(\mathbf{y})$ then $\phi_{\mathbb{P}}(\mathbf{x}) = \phi_{\mathbb{P}}(\mathbf{y})$

which implies that

$$\begin{aligned} d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m_o}) &= -\log K_{\mathbb{P}}(\mathbf{x}, \mathbf{m_o}), \\ &= -\left(\log \phi_{\mathbb{P}}(\mathbf{x}) + \log \phi_{\mathbb{P}}(\mathbf{m_o})\right), \\ &= -\left(\log \phi_{\mathbb{P}}(\mathbf{y}) + \log \phi_{\mathbb{P}}(\mathbf{m_o})\right), \\ &= d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{y}, \mathbf{m_o}). \end{aligned}$$

□

In real life applications, usually we do not know the underlying distribution $\mathbb{P}$; we will use sample density kernels, that arise from density kernels by replacing f-monotone functions by asymptotic f-monotone functions.

For instance, for $\phi_{S_n}(\mathbf{x}) = g_1(\mathbf{x}, S_n) = e^{-\frac{1}{2}(\mathbf{x}-\widehat{\mathbf{m}}_\mathbf{o})^T \hat{\mathbf{\Sigma}}^{-1}(\mathbf{x}-\widehat{\mathbf{m}}_\mathbf{o})}$, we obtain:

$$\begin{aligned} \hat{d}^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m_o}) &= -\log K_{S_n}(\mathbf{x}, \mathbf{m_o}), \\ &= -\log\left(\phi_{S_n}(\mathbf{x})\phi_{S_n}(\mathbf{m_o})\right), \\ &= -\log\left(e^{-\frac{1}{2}(\mathbf{x}-\widehat{\mathbf{m}}_\mathbf{o})^T \hat{\mathbf{\Sigma}}^{-1}(\mathbf{x}-\widehat{\mathbf{m}}_\mathbf{o})} e^{-\frac{1}{2}(\mathbf{m_o}-\widehat{\mathbf{m}}_\mathbf{o})^T \hat{\mathbf{\Sigma}}^{-1}(\mathbf{m_o}-\widehat{\mathbf{m}}_\mathbf{o})}\right), \\ &= \hat{d}^2_M(\mathbf{x}, \widehat{\mathbf{m}}_\mathbf{o}) + \hat{d}^2_M(\widehat{\mathbf{m}}_\mathbf{o}, \mathbf{m_o}). \end{aligned}$$

We can interpret $\hat{d}^2_M(\widehat{\mathbf{m}}_\mathbf{o}, \mathbf{m_o})$ as a bias term: when the sample size increases the estimation of the mode is more precise, in the limit $\lim_{n\to\infty} \hat{d}^2_M(\widehat{\mathbf{m}}_\mathbf{o}, \mathbf{m_o}) = 0$. Therefore $\hat{d}^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m_o})$ takes into account the bias term and gives a more precise estimation of the Mahalanobis distance when the sample size is not large enough. When the sample size increases, and provided consistent estimator of the mode and the covariance matrix, the estimation $\hat{d}^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m_o})$ converges to the Mahalanobis distance.

To study the performance in practice of the proposed family of distances we concentrate our attention on the choice $\phi_{\mathbb{P}}(\mathbf{x}) = \frac{f_{\mathbb{P}}(\mathbf{x})}{f_{\mathbb{P}}(\mathbf{m}_o)}$, where $f(\mathbf{m}_o) = \max_{\mathbf{x}} f(\mathbf{x})$. Then by the definition of the density kernel:

$$d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m}_o) = \log\left(\frac{f_{\mathbb{P}}(\mathbf{m}_o)}{f_{\mathbb{P}}(\mathbf{x})}\right).$$

When $\mathbf{x} = \mathbf{m}_o$, $d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m}_o) = \log(1) = 0$, and $d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m}_o)$ increases when $\mathbf{x}$ moves off from the mode $\mathbf{m}_o$.

The theoretical advantages of using the GM distance over using the Mahalanobis distance are twofold: First, the GM distance always fulfill the fundamental property that if two points belongs to the same probability level curve, that is $f_{\mathbb{P}}(\mathbf{x}) = f_{\mathbb{P}}(\mathbf{y})$, then they are located at the same distance to the center of the distribution:

$$d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m_o}) = d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{y}, \mathbf{m_o}).$$

In the case of the Mahalanobis distance this property is achieved only for few distributions, in particular when the underlying data distribution is Gaussian. Second, it is possible to derive a sample version of the GM distance by just providing an estimator of $f_{\mathbb{P}}(\mathbf{x})$, while the sample MD needs the estimation of the covariance matrix, that is quite problematic when dimensionality increases or there are outliers Zhang et al. (2012); Hsieh et al. (2011).

To have a working definition of the GM distance (given a sample), we need to estimate the density function $f_{\mathbb{P}}$. For most practical problems, including classification and outlier detection problems, we do not need an exact knowledge of $f_{\mathbb{P}}$: it will be enough to know the relative order among the distances from data points to the mode of the distribution (due to the use of the Bayes rule). That is, given $\mathbf{x}$ and $\mathbf{y}$, it is enough to know if $f_{\mathbb{P}}(\mathbf{x}) < f_{\mathbb{P}}(\mathbf{y})$ or $f_{\mathbb{P}}(\mathbf{x}) > f_{\mathbb{P}}(\mathbf{y})$. Therefore, we just need to estimate the α-level sets of $f_{\mathbb{P}}$: Given a probability measure $\mathbb{P}$ with density function $f_{\mathbb{P}}$, the α-level sets (or minimum volume sets) are defined by $S_\alpha(f_{\mathbb{P}}) = \{\mathbf{x} \in X | \, f_{\mathbb{P}}(\mathbf{x}) \geq \alpha\}$, such that $P(S_\alpha(f_{\mathbb{P}})) = 1 - \nu$, where $0 < \nu < 1$. If we consider an ordered sequence $\alpha_1 < \ldots < \alpha_m$, then $S_{\alpha_{i+1}}(f_{\mathbb{P}}) \subseteq S_{\alpha_i}(f_{\mathbb{P}})$. Let us define $A_i(\mathbb{P}) = S_{\alpha_i}(f_{\mathbb{P}}) - S_{\alpha_{i+1}}(f_{\mathbb{P}})$, $i \in \{1, \ldots, m-1\}$. We can choose $\alpha_1 \simeq 0$ and $\alpha_m \geq \max_{\mathbf{x}} f_{\mathbb{P}}(\mathbf{x})$ (which exists, given that X is compact and $f_{\mathbb{P}}$ is continuous). If the $\{\alpha_i\}_{i=1}^m$ sequence is long enough, we can assume constant density for the points contained in $A_i(\mathbb{P})$, that is, they have the same value $f_{\mathbb{P}}(\mathbf{x})$.

If $\mathbf{x} \in A_i(\mathbb{P})$, and because of the definition of $A_i(\mathbb{P})$, then $f_{\mathbb{P}}(\mathbf{x})$ is in correspondence with α_i, that is $f_{\mathbb{P}}(\mathbf{x}) \propto \alpha_i$, and thus:

$$d^2_{GM_{K_{\mathbb{P}}}}(\mathbf{x}, \mathbf{m}_o) = \log\left(\frac{f_{\mathbb{P}}(\mathbf{m}_o)}{f_{\mathbb{P}}(\mathbf{x})}\right) \propto \log\left(\frac{\alpha_m}{\alpha_i}\right). \tag{3.1}$$

Next we introduce the algorithm to estimate the $A_i(\mathbb{P})$ sets.

3.2.3 Level set estimation

Usually the available data are given as a finite sample. We consider an iid sample $s_n(\mathbb{P}) = \{\mathbf{x}_i\}_{i=1}^n$ drawn from the density function $f_{\mathbb{P}}$. To estimate level sets from a data sample (useful to obtain $\hat{S}_\alpha(f_{\mathbb{P}})$) we present the following definitions and theorems, concerning the One-Class Neighbor Machine Muñoz and Moguerza (2006, 2005).

Definition 3.6 (Neighbourhood Measures). Consider a random variable X with density function $f(\mathbf{x})$ defined on $\mathbb{R}^d$. Let S_n denote the set of random independent identically distributed (i.i.d.) samples of size n (drawn from f). The elements of S_n take the form $s_n = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^d$. Let $M : \mathbb{R}^d \times S_n \longrightarrow \mathbb{R}$ be a real-valued function defined for all $n \in \mathbb{N}$. (a) If $f(\mathbf{x}) < f(\mathbf{y})$ implies $\lim_{n\to\infty} P(M(\mathbf{x}, s_n) > M(\mathbf{y}, s_n)) = 1$, then M is a **sparsity measure**. (b) If $f(\mathbf{x}) < f(\mathbf{y})$ implies $\lim_{n\to\infty} P(M(\mathbf{x}, s_n) < M(\mathbf{y}, s_n)) = 1$, then M is a **concentration measure**.

The Support Neighbour Machine Muñoz and Moguerza (2006, 2005) solves the following optimization problem:

$$\begin{array}{ll} \max\limits_{\rho,\zeta} & \nu n\rho - \sum\limits_{i=1}^{n} \zeta_i \\ \text{s.t.} & g(\mathbf{x}_i) \geq \rho - \zeta_i \,, \\ & \zeta_i \geq 0, \qquad\qquad i = 1, \ldots, n \,, \end{array} \tag{3.2}$$

where $g(\mathbf{x}) = M(\mathbf{x}, s_n)$ is a sparsity measure, $\nu \in [0, 1]$, ζ_i with $i = 1, \ldots, n$ are slack variables and ρ is a threshold induced by the sparsity measure.

Theorem 3.1. *The set* $R_n = \{\mathbf{x} : h_n(\mathbf{x}) = \text{sign}(\rho_n^* - g_n(\mathbf{x})) \geq 0\}$ *converges to a region of the form* $S_\alpha(f) = \{\mathbf{x}|f(\mathbf{x}) \geq \alpha\}$, *such that* $P(S_\alpha(f)) = 1 - \nu$.

Therefore, the Support Neighbour Machine estimates a density contour cluster $S_\alpha(f)$ (around the mode). Theorem 3.1 (see Muñoz and Moguerza (2006, 2005) for a formal proof) can be expressed in algorithmic form as in Table 3.1:

Hence, we take $\hat{A}_i(\mathbb{P}) = \hat{S}_{\alpha_i}(f_{\mathbb{P}}) - \hat{S}_{\alpha_{i+1}}(f_{\mathbb{P}})$, where $\hat{S}_{\alpha_i}(f_{\mathbb{P}})$ is estimated by R_n defined in Table 3.1. For further details on the estimation and its rate of convergence refers to Muñoz and Moguerza (2004); Moguerza and Muñoz (2004); Muñoz and Moguerza (2006, 2005).

With the estimation of level sets and the relation presented in Equation 3.1, we will test with some experiment the performance of the proposed distance.

Table 3.1: Algorithmic formulation of Theorem 3.1.

	Estimation of $\mathbf{R_n} = \hat{\mathbf{S}}_\alpha(\mathbf{f})$:
1	Choose a constant $\nu \in [0, 1]$.
2	Consider the order induced in the sample s_n by the sparsity measure $g_n(\mathbf{x})$, that is, $g_n(\mathbf{x}_{(1)}) \leq \cdots \leq g_n(\mathbf{x}_{(n)})$, where $\mathbf{x}_{(i)}$ denotes the i^{th} sample, ordered after g.
3	Consider the value $\rho_n^* = g(\mathbf{x}_{(\nu n)})$ if $\nu n \in \mathbb{N}$, $\rho_n^* = g_n(\mathbf{x}_{([\nu n]+1)})$ otherwise, where $[\mathbf{x}]$ stands for the largest integer not greater than $\mathbf{x}$.
4	Define $h_n(\mathbf{x}) = \text{sign}(\rho_n^* - g_n(\mathbf{x}))$.

3.3 Experimental Section

In this section we compare the performance of the generalized Mahalanobis distance and the classical Mahalanobis distance in a variety of artificial and real data classification and outlier detection problems.

3.3.1 Artificial experiments

In the first experiment we test the ability of the GM distance to detect outliers in a non-Gaussian scenario. To this aim we first generate a sample of 200 points $\mathbf{z_i} = (x_i, f(x_i))$ using a mixture $f(x) = 0.95f_1(x) + 0.05f_2(x)$, where

$$\begin{aligned} f_1(x) &= \sin(x) + \varepsilon, \\ f_2(x) &= \frac{1}{2}\sin(2x - \frac{\pi}{2}) + \varepsilon, \end{aligned}$$

$x \in [0, \pi]$, $\varepsilon \sim N(\mu_\varepsilon = 0, \sigma_\varepsilon^2 = 0.1)$. Thus, we are considering a proportion of 5% outlying points.

Next, for each of the GM distance, the Mahalanobis Distance and the Euclidean Distance we calculate the vectors $(d_{GM}(\mathbf{z_i}, \mathbf{m}_o))$, $(d_{MD}(\mathbf{z_i}, \boldsymbol{\mu}))$, $(d_E(\mathbf{z_i}, \boldsymbol{\mu}))$ and consider as outliers for each of the distances the 5% largest distances. In addition, we use five alternative outlier detection algorithms, pc-Outlier, sign-Outlier and locoutPercent Filzmoser et al. (2008); Zimek et al. (2012), fastPCS Vakili and Schmitt (2014) and DMwR Torgo (2010), for the sake of comparison. The results are summarized in Table 3.2.

The GM distance outperforms the Mahalanobis distance and also the other methods, by correctly identifying all the outliers without any false-positive result. In Figure 3.3 we show

Table 3.2: Comparison of different outliers detection methods.

Metric/Technique	% of:	Outliers captured	False-positives (Type I error)	False-negatives (Type II error)
pc-Outlier		5.0%	20.0%	95.0%
sign-Outlier		0%	7.0%	100%
locoutPercent		5.0%	9.5%	95%
fastPCS		5.0%	50%	95%
DMwR		40.0%	2.0%	60%
Percentile 5% ED		0%	5.5%	100%
Percentile 5% MD		0%	5.5%	100%
Percentile 5% GM		**100%**	**0.0%**	**0.0%**

the level curves of the main distribution, together with the outlying points. Among the more sophisticated methods, only DMwR is able to capture a significative proportion of the outliers, 40% of them, and pc-Outlier, locoutPercent and fastPCS are only able to detect 5% of the outliers. The rest of alternative methods fail by labelling as outliers points that are not distant from the center of the distribution.

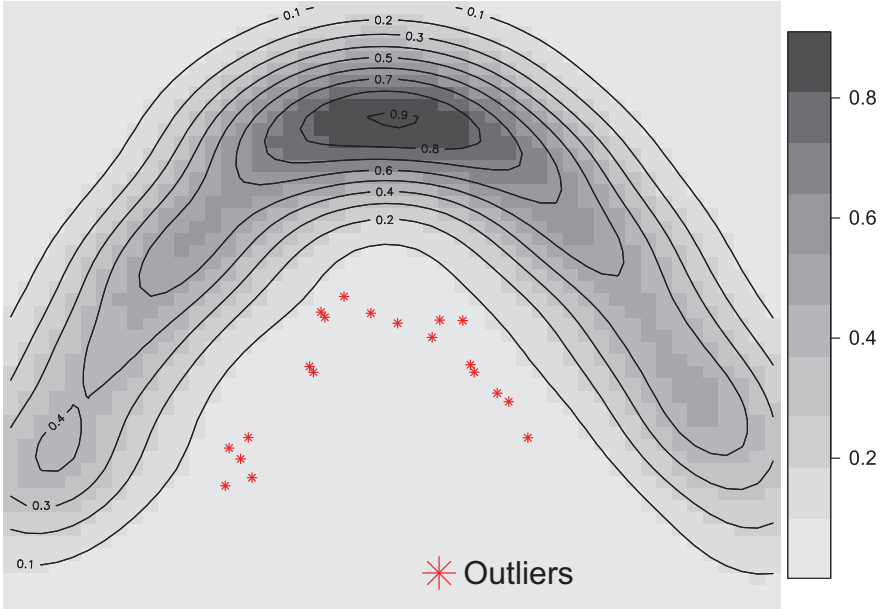


Figure 3.2: Level sets of the main distribution plus outlying data points.

In the second experiment, we consider again a mixture of distributions. In this case the 95% of the observations come from a bi-logistic distribution $BL(\alpha = 0.5, \beta = 0.9)$ Smith et al. (1990), and the rest from a normal distribution $\mathbf{N}(\boldsymbol{\mu} = (3,3), \boldsymbol{\Sigma} = 5I_{2\times 2})$. We take $n = 1000$

Table 3.3: Comparison of different outliers detection methods.

Metric/Technique	% of:	Outliers captured	False-positives (Type I error)	False-negatives (Type II error)
pc-Outlier		36.5%	23.2%	65.8%
sign-Outlier		23.1%	7.4%	76.9%
locoutPercent		13.4%	**7.3**%	86.4%
fastPCS		15.2%	8.5%	76.9%
DMwR		12.6%	13.2%	70.1%
Percentile 5% ED		3.8%	10.7%	96.1%
Percentile 5% MD		23.1%	10.4%	76.9%
Percentile 5% GM		**38.5**%	10.3%	**65.4**%

data points. In this case, the problem is more difficult to solve, given that the contaminating distribution is located in the center of the main data cloud. This means that many points of the second distribution will not be distinguishable from the main cloud.

We consider the same outlier detection procedures than in the preceding example, and the results are summarized in Table 3.3.

Again, the use of the Generalized Mahalanobis distance gives the best results, followed by the pc-Outlier method. Besides, the Mahalanobis distance performs better than the Euclidean distance, and the Generalized Mahalanobis distance outperforms the Mahalanobis distance, showing again the usefulness of the proposed generalization.

In Figure 3.3, we show the main distribution together with the detected outliers for the GM distance.

3.3.2 Real data experiments

In the first experiment we test, in a classification problem, the performance of the GM distance against the classical procedures that use the Mahalanobis distance: linear and quadratic discriminant analysis.

We consider a collection of 860 documents, organized in three main topics, extracted from three bibliographic data bases (LISA, INSPEC and Sociological Abstracts). Each document is represented as a vector in the Latent Semantic Space (in this example $\mathbb{R}^{364}$) using the Singular Value Decomposition Deerwester et al. (1990).

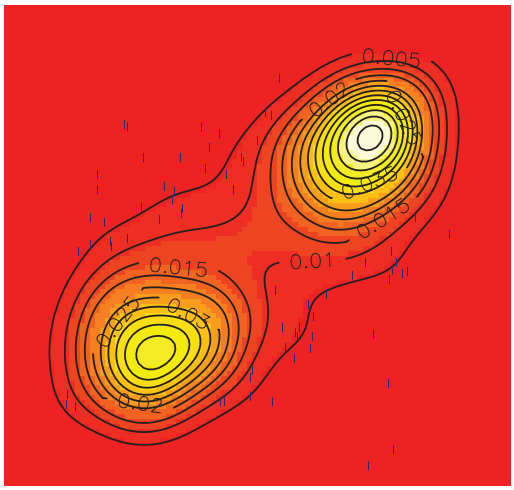


Figure 3.3: Contaminated points detected for the GM distance. The rest (belonging to a normal distribution) are masked with the main distribution cloud.

Table 3.4: Classification percentage errors for a three-class text database and three classification procedures. In parenthesis the St. Error on the test samples.

Method	% of:	**Train Error**	**Test Error**
Linear Discriminant		6.100%	7.035% (0.007)
Quadratic Discriminant		6.426%	6.960% (0.001)
Generalized Mahalanobis		**2.553**%	**2.761**% (0.002)

The topics (classes of documents) are: "dimensionality reduction" and "feature selection" (311 documents), "optical cables" and "power semiconductor devices" (384 documents) and "rural areas" and "retirement communities" (165 documents).

We randomly split the data set into training (60% of the documents, that is, 516) and testing (40% of the documents, that is, 344). Regarding the use of GM distance, we first estimate the level sets for each of the three classes using the training set, and then a document d from the test set is classified using the following procedure:

$$C_d = \arg\min_i d_{GM}(d, m_{0i}),$$

where m_{0i} is the estimated mode for class i ($i \in \{1, 2, 3\}$). We repeat 100 times the generation procedure (the split of the data set intro training and testing subsets) and average the resulting errors for each of the classification procedures. The results are shown in Table 3.4.

The use of the GM distance outperforms the Mahalanobis distance again. In this case we

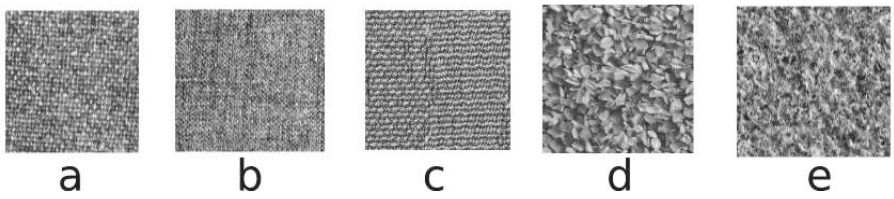


Figure 3.4: Textures images: a) blanket, b) canvas, c) seat, d) linseeds and e) stone.

Table 3.5: Comparison of different outliers detection methods.

Metric/Technique	% of: Outliers captured	False-positives (Type I error)	False-negatives (Type II error)
pc-Outlier	60%	13.23%	28.65%
sign-Outlier	40%	5.13%	37.75%
locoutPercent	35%	**2.80%**	39.39%
DMwR	20.0%	4.38%	80%
Percentile 5% ED	25%	4.00%	42.85%
Percentile 5% MD	35%	3.60%	39.39%
Percentile 5% GM	**100%**	5.10%	**0.00%**

have 860 points in $\mathbb{R}^{364}$, and the estimation of the covariance matrix, inaccurate in high dimensional scenarios, contributes to the lower performance of the MD in linear and quadratic discriminant analysis.

The second real data example considers the detection of outliers in a sample of texture images. We consider the Kylberg texture database Kylberg (2011). We use a similar experimental set up as those described in Section 3.1 but we do not report the results of fastPCS method because it is implemented to work with data until dimension 25. For this experiment we use 500 texture images (with a resolution of 576×576 pixels). The first 480 texture images are rather homogeneous (Figure 3.4 a) to c)). We also consider 20 "outlying" images with apparently different patterns (Figure 3.4 d) and e)). We represent each image using the 32 parameters of the wavelet coefficient histogram proposed in Mallat (1989).

We report the results in Table 3.5. Although the type I error of the GM (5.10%) is slightly larger than in cases like ED (4.00%) or MD (3.60%), the GM distance is the only method able to capture all the outliers in the sample, and it is the unique procedure without false-negative outliers.

Chapter Summary

In this chapter we have proposed a family of density kernels based on the underlying distribution of the data. This family of density kernels induces a Generalized Mahalanobis distance introduced in Definition 3.5. In the case of the exponential kernel, the proposed distance is exactly the Mahalanobis distance.

The proposed distance outperform the classical Mahalanobis distance and other more sophisticated methods in Statistics and Machine Learning to solve classification and outlier detection problems. Thus, the proposed distance really generalized the classical Mahalanobis distance, devised for the normal distribution case.

Chapter 4

New Distance Measures for Multivariate Data Based on Probabilistic Information

Chapter abstract

In this chapter we study distances for points in an affine space taking into account the probabilistic information in the data. We propose two new distance measures for points and study its properties. We also propose and study estimation methods for the proposed distances and show its performance in classification.

Chapter keywords: Distance functions. Density and Distribution. Classification.

4.1 Introduction

There are diverse distance measures in data analysis: Canberra, Euclidean, Cosine or Manhattan, to name just a few. These distance measures are devised to be used in the context of Euclidean spaces, where the idea of density is missing and only the coordinate positions of the points in an affine space are taken into consideration to compute the distances.

In the case when we want to consider also the distribution of the data at hand there is only one distance left: the Mahalanobis distance. Nevertheless, the Mahalanobis distance only consider the distance from a point to the center of a distribution. In this chapter we propose distances for points in an affine space X taking into account the distribution of the data at hand.

The rationale of the new distances is similar to the "earth mover intuition" in the definition of the Wasserstein metric Rüschendorf (1985); Rachev and Rüschendorf (1998). Our starting point is a random sample $S^n_{\mathbb{P}} = \{x_1, \ldots, x_n\}$ drawn from a PM $\mathbb{P}$. The sample points can be considered as particles of total mass 1 placed in X. For each $i = 1, \ldots, n$ the particle i is placed at the coordinate $x_i \in X$ with a mass $w_i \geq 0$. Since the total mass is 1, we require that $\sum_{i=1}^n w_i = 1$. Following the "earth mover intuition", we can assume that the cost associated to move the particle located in the position x_i to the position x_j is proportional to the total mass contained between these two particles. In other words, if we consider the ordered sample $\{x_{(1)}, \ldots, x_{(n)}\}$, then the new distance measures $d_{\mathbb{P}}$ between two points, say $x_{(i)}$ and $x_{(j)}$, are proportional to the number of points in the sample $S^n_{\mathbb{P}}$ included in the interval $[x_{(i)}, x_{(j)}]$.

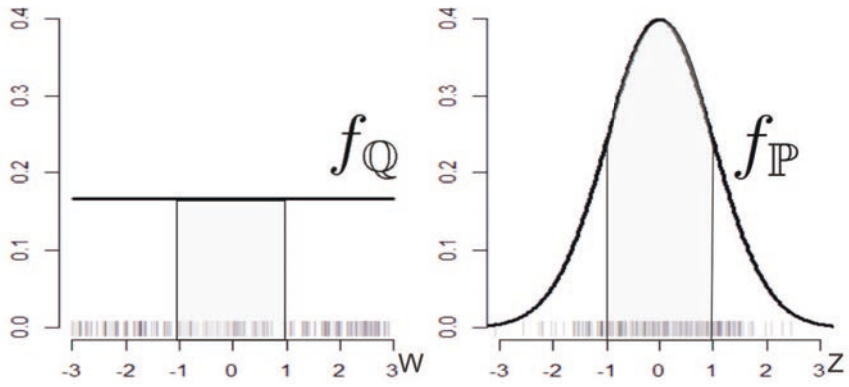


Figure 4.1: Two density functions: Uniform density $f_{\mathbb{Q}}$ (left) and Normal density $f_{\mathbb{P}}$ (right). Sample points from the two different distributions are represented with black bars in the horizontal x-axes.

Let $\mathbb{Q} = U[-3, 3]$ be the uniform PM in the interval $[-3, 3]$ and $\mathbb{P} = N(\mu = 0, \sigma = 1)$ the standard normal PM. Figure 4.1-left shows a uniformly distributed random sample $S^n_{\mathbb{Q}}$, and Figure 4.1-right a normally distributed random sample $S^n_{\mathbb{P}}$ where the sample size $n = 100$. The coordinates of the sample points are represented with black bars in the horizontal axes. Then

$$\frac{1}{n} \sum_{w \in S^n_{\mathbb{Q}}} \mathbb{1}_{[w \in [-1,1]]}(w) \approx P(-1 \leq W \leq 1) = \frac{1}{3},$$

and

$$\frac{1}{n}\sum_{z\in S_{\mathbb{P}}^n}\mathbb{1}_{[z\in[-1,1]]}(z)\approx P(-1\leq Z\leq 1)=0.682.$$

Thus we expect that

$$\frac{1}{n}\sum_{w\in S_{\mathbb{Q}}^n}\mathbb{1}_{[w\in[-1,1]]}(w)<\frac{1}{n}\sum_{z\in S_{\mathbb{P}}^n}\mathbb{1}_{[z\in[-1,1]]}(z),\quad \text{for } n\gg 0.$$

We want that the new distances fulfills this inequality, that is: $d_{\mathbb{Q}}(-1,1)<d_{\mathbb{P}}(-1,1)$. It is apparent then that the density function will play a central role in what follows.

This chapter is organized in the following way: In Section 4.2 we present the cumulative distribution function distance and study its properties. In Section 4.3 we present the Minimum Statistical Work distance and study its properties. We also present in this section an estimation procedure to compute the proposed distance and demonstrate its convergence to the proposed theoretical distance. In Section 4.4 we show the performance of the proposed distance and compare its performance to solve classification problems with other several standard metrics in Statistics and Machine Learning.

4.2 The Cumulative Distribution Function Distance

Through this chapter we consider a measure space $(X,\mathcal{F},\mu)$, where X is a sample space (here a compact set of $\mathbb{R}^d$), $\mathcal{F}$ a σ-algebra of the measurable subsets of X and $\mu:\mathcal{F}\to\mathbb{R}^+$ the Lebesgue measure. A probability measure $\mathbb{P}$ is a σ-additive finite measure absolutely continuous w.r.t. μ. By Radon-Nikodym theorem, there exists a measurable function $f:X\to\mathbb{R}^+$ (the density function) such that $\mathbb{P}(A)=\int_A f d\mu$, and $f=\frac{d\mathbb{P}}{d\mu}$ is the Radon-Nikodym derivative.

The first distance that we propose in this chapter makes use of the cumulative distribution function (CDF). The CDF associated to the PM $\mathbb{P}$ is defined as:

$$F_{\mathbb{P}}(\mathbf{x})=F_{\mathbb{P}}\left(x_1,\ldots,x_d\right)=\int_{-\infty}^{x_1}\ldots\int_{-\infty}^{x_d}f_{\mathbb{P}}(x_1,\ldots,x_d)dx_1\ldots dx_d$$

The CDF $F_{\mathbb{P}}:X\to[0,1]$ is a non-decreasing and right-continuous function. Furthermore $\lim_{x_1,\ldots,x_d\to-\infty}F_{\mathbb{P}}(\mathbf{x})=0$ and $\lim_{x_1,\ldots,x_d\to\infty}F_{\mathbb{P}}(\mathbf{x})=1$, and thus $F_{\mathbb{P}}$ is bounded. We propose in first place a univariate definition of the CDF distance and extend later the definition of the distance to the multivariate context.

There exists a well known relationship between the uniform distribution and the CDF: If U is a uniformly distributed random variable in the interval $[0,1]$, then $Q(U) = F_{\mathbb{P}}^{-1}(U)$ has a cumulative distribution function given by $F_{\mathbb{P}}$. In this way, the CDF allows us to embed the PM $\mathbb{P}$ into the $[0,1]$ interval and moreover, the distribution of the r.v. defined as $U = F_{\mathbb{P}}(x)$ is uniform in the interval $[0,1]$. This motivates the definition of the CDF distance (in the univariate case) as follows.

Definition 4.1 (**CDF distance: the univariate case**). Let $\mathbb{P}$ be a PM and denoted by $F_{\mathbb{P}}$ to the respective CDF. Define the CDF distance between two points x and y in $X \subset \mathbb{R}$ as follows:

$$d_{F_{\mathbb{P}}}(x,y) = \sqrt{\left(F_{\mathbb{P}}(x) - F_{\mathbb{P}}(y)\right)^2} = |F_{\mathbb{P}}(x) - F_{\mathbb{P}}(y)|.$$

We can interpret the distance in Definition 4.1 as a composition of a (non-linear) transformation: $F_{\mathbb{P}} : X \to [0,1]$, plus the computation of the ordinary Euclidean distance $||F_{\mathbb{P}}(x) - F_{\mathbb{P}}(y)||$. The proposed distance has the required property that: "the distance between two points is proportional to the probability contained between the two points".

In the univariate case, the proposed distance it is the distance between the quantiles of the PM $\mathbb{P}$. The definition of the CDF distance is easily extensible to the multivariate case. Denote by $F_{\mathbb{P}}^i(x)$ to the i^{th}-marginal cumulative distribution function, that is:

$$F_{\mathbb{P}}^i(x) = \lim_{x_1,\ldots,x_{i-1},x_{i+1},\ldots x_d \to \infty} F_{\mathbb{P}}(\mathbf{x}), \quad \text{for } i = 1,\ldots,d,$$

then the multivariate CDF distance is defined as follows.

Definition 4.2 (**CDF distance: the multivariate case**). Let $\mathbb{P}$ be a PM and denoted by $F_{\mathbb{P}}$ to the respective CDF and by $F_{\mathbb{P}}^i(x)$ to the i^{th}-marginal cumulative distribution function. Define the CDF distance between two points $\mathbf{x}$ and $\mathbf{y}$ in $X \subset \mathbb{R}^d$ by

$$d_{F_{\mathbb{P}}}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{d} \left(F_{\mathbb{P}}^i(x_i) - F_{\mathbb{P}}^i(y_i)\right)^2}.$$

The map $\phi_{\mathbb{P}} : X \to [0,1]^d$ assigns $\mathbf{x} = (x_1,\ldots,x_d) \longmapsto (F_{\mathbb{P}}^1(x_1),\ldots,F_{\mathbb{P}}^d(x_d))$. In this way, $\phi_{\mathbb{P}}$ embeds the random vector $\phi_{\mathbb{P}}(\mathbf{x})$ into a hyper-cube $[0,1]^d$. The proposed distance fulfills all the properties to be a proper metric. Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be three points in the support of the PM $\mathbb{P}$,

then $d_{F_{\mathbb{P}}}$ has the following properties:

- *Non-negativity*: $d_{F_{\mathbb{P}}}(\mathbf{x}, \mathbf{y}) \geq 0$, and *Indiscernibility of identicals*: $d_{F_{\mathbb{P}}}(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.

  Proof: The $d_{F_{\mathbb{P}}}$ distance is non-negative by definition. $d_{F_{\mathbb{P}}}(\mathbf{x}, \mathbf{y}) = 0$ if $F^i_{\mathbb{P}}(x_i) = F^i_{\mathbb{P}}(y_i)$ for $i = 1, \ldots, d$. This happens only if $\mu([x_i, y_i]) = 0$ for $i = 1, \ldots, d$, that is $x_i = y_i$ for $i = 1, \ldots, d$ or x_i, y_i or both do not belong to the support of the distribution and hence the distance between these points does not make sense.

- *Symmetry*: $d_{F_{\mathbb{P}}}(\mathbf{x}, \mathbf{y}) = d_{F_{\mathbb{P}}}(\mathbf{y}, \mathbf{x})$, and *Triangle inequality*: $d_{F_{\mathbb{P}}}(\mathbf{x}, \mathbf{y}) \leq d_{F_{\mathbb{P}}}(\mathbf{x}, \mathbf{z}) + d_{F_{\mathbb{P}}}(\mathbf{z}, \mathbf{y})$.

  Proof: The symmetry is a property inherited from the definition of the distance. To prove the the triangle inequality observe that:

$$
\begin{aligned}
d_{F_{\mathbb{P}}}(\mathbf{x}, \mathbf{y}) &= \sqrt{\sum_{i=1}^{d} \left(F^i_{\mathbb{P}}(x_i) - F^i_{\mathbb{P}}(y_i)\right)^2} \\
&= \sqrt{\sum_{i=1}^{d} \left(F^i_{\mathbb{P}}(x_i) - F^i_{\mathbb{P}}(z_i) + F^i_{\mathbb{P}}(z_i) - F^i_{\mathbb{P}}(y_i)\right)^2} \\
&\leq \sqrt{\sum_{i=1}^{d} \left(F^i_{\mathbb{P}}(x_i) - F^i_{\mathbb{P}}(z_i)\right)^2 + \sum_{i=1}^{d} \left(F^i_{\mathbb{P}}(z_i) - F^i_{\mathbb{P}}(y_i)\right)^2} \\
&\leq \sqrt{\sum_{i=1}^{d} \left(F^i_{\mathbb{P}}(x_i) - F^i_{\mathbb{P}}(z_i)\right)^2} + \sqrt{\sum_{i=1}^{d} \left(F^i_{\mathbb{P}}(z_i) - F^i_{\mathbb{P}}(y_i)\right)^2} \\
&\leq d_{F_{\mathbb{P}}}(\mathbf{x}, \mathbf{z}) + d_{F_{\mathbb{P}}}(\mathbf{z}, \mathbf{y}).
\end{aligned}
$$

The computation of $d_{F_{\mathbb{P}}}$ it is straightforward when we know explicitly the PM $\mathbb{P}$. In the cases when we do not know the PM $\mathbb{P}$ but there is available a data sample $S^n_{\mathbb{P}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ drawn from $\mathbb{P}$, we can use the empirical CDF as a plug-in estimator of the distance proposed in Definition 4.2. We denote by $\hat{F}^i_{S^n_{\mathbb{P}}}$ to the empirical i^{th} marginal CDF, that is:

$$
\hat{F}^i_{S^n_{\mathbb{P}}}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[x_i \leq t]} \quad \text{for } i = 1, \ldots, d.
$$

The strong law of large numbers ensures the convergence of the estimator: $\hat{F}^i_{S^n_{\mathbb{P}}} \xrightarrow{a.s.} F^i_{\mathbb{P}}$. By using this result we define the empirical CDF distance $\hat{d}_{F_{S^n_{\mathbb{P}}}}$ as follows.

Definition 4.3 (Empirical CDF distance). Let $S_{\mathbb{P}}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a random sample drawn from $\mathbb{P}$, denoted by $\hat{F}^i_{S_{\mathbb{P}}^n}$ the i^{th}-empirical marginal cumulative distribution function. We define the empirical CDF distance between two points $\mathbf{x}$ and $\mathbf{y}$ in X by

$$\hat{d}_{F_{S_{\mathbb{P}}^n}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} \left(\hat{F}^i_{S_{\mathbb{P}}^n}(x_i) - \hat{F}^i_{S_{\mathbb{P}}^n}(y_i) \right)^2}.$$

It is important to highlight that the empirical CDF distance also converges $\hat{d}_{F_{S_{\mathbb{P}}^n}} \xrightarrow{a.s.} d_{F_{\mathbb{P}}}$ by the strong law of large numbers.

An example of the use of $d_{F_{\mathbb{P}}}$ in a real data context

In this example we motivate the use of the metric $d_{F_{\mathbb{P}}}$ to compare households by taking into account the income distribution. For this purpose we take income data from the U.S. Census Bureau (Current Population Survey 2014: Annual Social and Economic Supplement). Sample is based on $272,814$ households and the variable Income represent the total yearly-income in the survey respondent households during the year 2013. As we can see in Figure 4.2, the distribution of the income is characterized by its lack of symmetry.

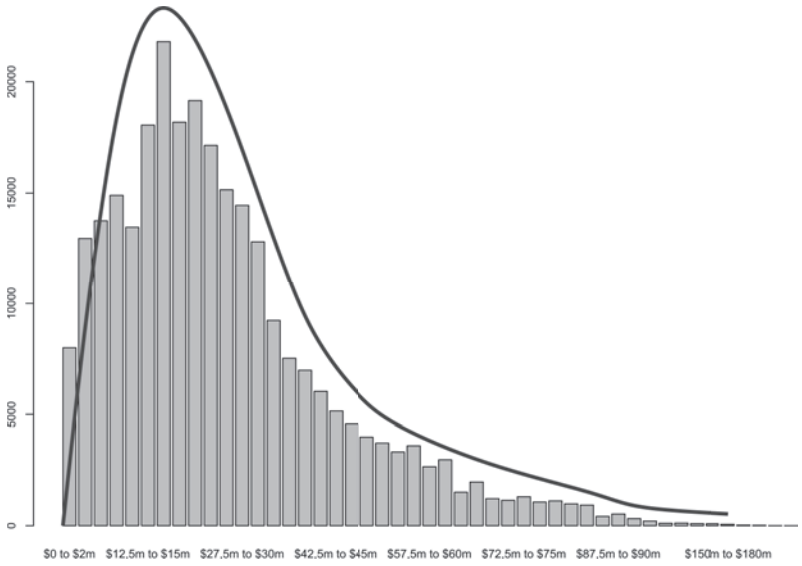


Figure 4.2: The distribution of the income data from the U.S. Census Bureau (2014 survey).

We can measure the similarity between households by using a distance function that quantify the difference in income. It will be desirable that the similarity measure takes into account

the distribution of the income. Following the economic theory of welfare Easterlin (2005); Hovenkamp (1990), the "Law of Diminishing Marginal Returns of the Income" ensures that an additional unit of income produce an increment in the "utility" (welfare) of the household which is inversely related to the amount of income before the increment. The distance used to measure the similarity between households should reflect this law.

We compare the similarity between the households with yearly incomes of: 10, 15, 100 and 105 (income is measured in a scale of $1000\times$USD) by using 3 different metrics. We give next the distance matrices for the 3 distances: the Euclidean distance (d_E), the Mahalanobis distance (d_M) and the proposed CDF distance ($d_{F_{S^n_{\mathbb{P}}}}$):

$$d_E = \begin{pmatrix} 0 & 5 & 90 & 95 \\ 5 & 0 & 85 & 90 \\ 90 & 85 & 0 & 5 \\ 95 & 90 & 5 & 0 \end{pmatrix} \quad d_M = \begin{pmatrix} 0 & 0.50 & 0.90 & 0.95 \\ 0.50 & 0 & 0.85 & 0.90 \\ 0.90 & 0.85 & 0 & 0.50 \\ 0.95 & 0.90 & 0.50 & 0 \end{pmatrix} \quad d_{F_{S^n_{\mathbb{P}}}} = \begin{pmatrix} 0 & 0.23 & 0.84 & 0.85 \\ 0.23 & 0 & 0.62 & 0.63 \\ 0.84 & 0.62 & 0 & 0.01 \\ 0.85 & 0.63 & 0.01 & 0 \end{pmatrix}$$

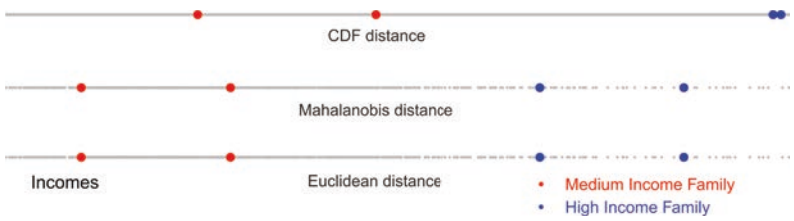


Figure 4.3: Income distribution represented via MDS for the metrics $d_{F_{S^n_{\mathbb{P}}}}$, d_M and d_E respectively.

In Figure 4.3 we represent via Multidimensional Scaling (MDS) the Income sample data and mark the 4 households with $10, 15$ (red points) and $100, 105$ (blue points) yearly thousand USD.

For the Income data presented in this experiment, the metric $d_{F_{S^n_{\mathbb{P}}}}$ is able to adequately represent the similarity between the households according to the law of diminishing marginal returns of the income: The distance between medium-income families is enlarged and the distance between the high-income families is reduced in comparison with the other two metrics.

4.3 The Minimum Work Statistical Distance

In this section we introduce the concept of distance between points by making an analogy with the physical concept of the *work* generated when an object is moved.

In Physics, the **work** is a measurable quantity that represents the magnitude of the **force** done in order to move an object a certain distance. Denote by W the work done by a constant force F that moves an object[1] a distance d in the direction of the force, then the relationship between these magnitudes is represented with the equation:

$$W = Fd$$

The object can be moved along a non-linear trajectory, for example by using a curved path γ on the space, that is a smooth curve parametrized by $t \in [0, 1]$ (we require that $\gamma \in C^2[0, 1]$). In this case, the instantaneous work (denoted as dW) that takes place over an instant of time (dt) when we move the object along the trajectory $\gamma(t)$ is given by:

$$dW = F \cdot v dt,$$

where v is the (constant) velocity of the object that moves along the path $\gamma(t)$ at the time t. The product vdt represents the infinitesimal (also known as "elemental") displacement of the object during dt. The quantity $F \cdot v$ measures the *power* applied during an infinitesimal period of time dt. Therefore the sum of all the infinitesimal amounts of work made along the trajectory γ yields the total work:

$$W_\gamma = \int_\gamma F \cdot v dt.$$

From this elementary definition, we are able to establish a connection between Physics and Statistics in order to derive a metric for points. Our aim is to relate the concept of work in Physics with the concept of distribution in Statistics. Let $f_{\mathbb{P}}$ be the density function relative to a probability measure $\mathbb{P}$. The density function $f_{\mathbb{P}}$ features a continuous random variable X. We can measure the instantaneous *statistical work* necessary to change the feature $X = x$ by:

$$dW(x) = \langle f_{\mathbb{P}}, \delta_x \rangle = \int_{-\infty}^{\infty} \delta(s - x) f(s)\, ds = f_{\mathbb{P}}(x)$$

where δ is the Dirac-Delta generalized function (refer to the appendix B for further details

[1] The object can be assumed to be a particle in a constant gravitational environment.

regarding generalized functions theory). This analogy results natural if we think for example that the r.v. X represents the height of a group of students in centimeters and $f_{\mathbb{P}}$ is the respective density function that describes its distribution. Then, the number of students with a height in the range of x_0 cm. and $x_0 + \varepsilon$ cm. (with $\varepsilon > 0$) in the population is proportional to $P(x_0 \leq x \leq x_0 + \varepsilon)$. Therefore the number of students that we need to inspect (to measure) when we want to find a student which is ε cm. taller than a student of x_0 cm. represents the statistical work and it will be then proportional to $P(x_0 \leq x \leq x_0 + \varepsilon)$.

Equivalently, the instantaneous *statistical work* can also be measured in terms of the distribution function. Let $F_{\mathbb{P}}$ be the distribution function relative to a probability measure $\mathbb{P}$, then:

$$dW(x) = -\langle F_{\mathbb{P}}, \delta_x' \rangle = \langle F_{\mathbb{P}}', \delta_x \rangle = \langle f_{\mathbb{P}}, \delta_x \rangle, \tag{4.1}$$

where we use the definition of the distributional derivative to arrive to the proposed equivalence.

By analogy with the concept of total work in Physics we can define then the Statistical Work as follows.

Definition 4.4 (Statistical Work Distance: the univariate case). Let $\mathbb{P}$ be a PM and denote by $f_{\mathbb{P}}$ to the density function associated to $\mathbb{P}$. Then define the distance d_{SW} between x_1 and x_2, two points that belongs to $X \subset \mathbb{R}$, as the sum (the integral) of the instantaneous statistical work done between the respective points, that is:

$$d_{SW}(x_1, x_2) = \int_{x_1}^{x_2} dW = \Big| \int_{x_1}^{x_2} f_{\mathbb{P}}(s)\, ds \Big| = \Big| F_{\mathbb{P}}(x_1) - F_{\mathbb{P}}(x_2) \Big| = P(x_1 \leq x \leq x_2).$$

In Figure 4.4 we give a geometrical interpretation of the proposed distance. The $d_{SW}(x_1, x_2)$ distance is the amount of density contained between x_1 and x_2 (in the univariate case). Thus $d_{SW} = d_{F_{\mathbb{P}}}$ in the univariate case.

Next we show that the d_{SW} distance between the points x_1 and x_2 it is related with the arc-length (a distance) of the $F_{\mathbb{P}}$ curve between the points $(x_1, F_{\mathbb{P}}(x_1))$ and $(x_2, F_{\mathbb{P}}(x_2))$.

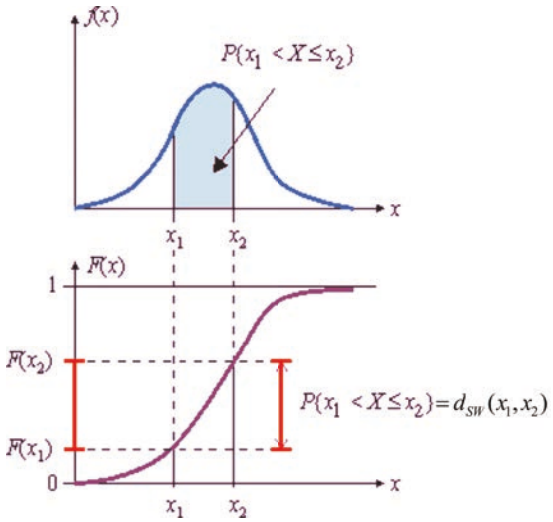


Figure 4.4: Schema of the d_{SW} distance and its geometrical interpretation.

The d_{SW} distance and its relationship with arc-length of the CDF function

We can compute the arc-length distance in the in the curve $F_{\mathbb{P}}$ (denoted as $L_{F_{\mathbb{P}}}$) between the points x_1 and x_2 (assuming that $x_1 \leq x_2$) by using the calculus of variations (see the appendix B for further details):

$$L_{F_{\mathbb{P}}}(x_1, x_2) \quad = \quad \int_{x_1}^{x_2} \sqrt{1 + \left(F'_{\mathbb{P}}(s)\right)^2} ds.$$

When x_1 and x_2 are neighbor points, that is when $x_2 = x_1 + \varepsilon$ for a small value of ε, then:

$$Ł^2_{F_{\mathbb{P}}}(x_1, x_1 + \varepsilon) \approx d^2_{SW}(x_1, x_1 + \varepsilon) + d^2_E(x_1, x_1 + \varepsilon) = \left|F_{\mathbb{P}}(x_1) - F_{\mathbb{P}}(x_1 + \varepsilon)\right|^2 + \varepsilon^2,$$

and the approximation is more accurate as ε approximates to 0. This last expression is just the approximation of the arc-length of the curve $F_{\mathbb{P}}$ between the points x_1 and $x_1 + \varepsilon$ by using the Pythagoras theorem. Figure 4.5 describes this approximation in the neighbor of a point x. The arc length of the line that joints the points $(x, F_{\mathbb{P}}(x))$ and $(x + \varepsilon, F_{\mathbb{P}}(x + \varepsilon))$ through the CDF $F_{\mathbb{P}}$ is approximately equal to sum of two distances: the CDF distance and the Euclidean distance between the points x and $x + \varepsilon$.

Next we extend the definition of d_{SW} to the multivariate context.

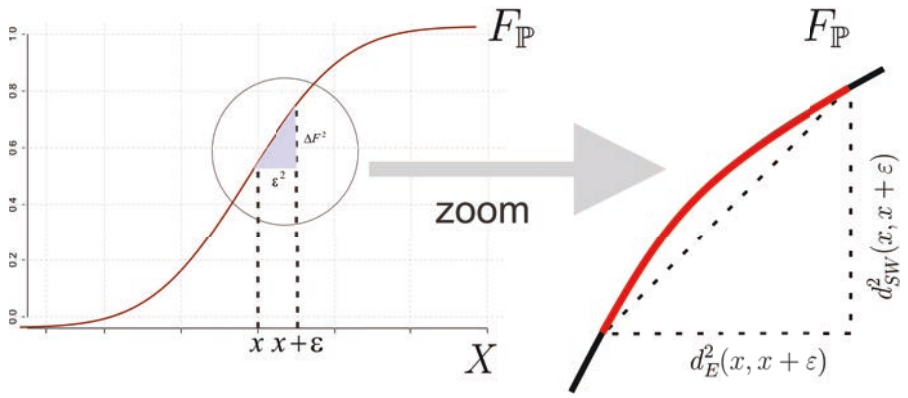


Figure 4.5: The relationship between the d_{SW} distance and the arc-length through $F_{\mathbb{P}}$.

Definition 4.5 (Statistical Work Distance: the multivariate case). Let $\mathbb{P}$ be a PM and $f_{\mathbb{P}}$ the density function associated to $\mathbb{P}$ and let $\gamma : [0,1] \to X$ be a smooth curve ($\gamma \in C^2$) parametrized by $t \in [0,1]$. Then define the minimum work statistical distance between the points $\mathbf{x}$ and $\mathbf{y}$ in X by:

$$d_{SW}(\mathbf{x}, \mathbf{y}) = \inf_{\gamma \in C^2} \int_0^1 f_{\mathbb{P}}(\gamma(t))\, dt, \tag{4.2}$$

such that $\gamma(0) = \mathbf{x}$ and $\gamma(1) = \mathbf{y}$.

The geometrical interpretation of the proposed distance is simple, $d_{SW}(\mathbf{x}, \mathbf{y})$ is the result of line integral of the density function $f_{\mathbb{P}}$ connecting the points $\mathbf{x}$ and $\mathbf{y}$ (that is: $\gamma(0) = \mathbf{x}$ and $\gamma(1) = \mathbf{y}$) with minimal area.

We illustrate the definition of the distance with the aid of an example. In Figure 4.6 we shown the density function of a normal distribution with parameters $\boldsymbol{\mu} = (0,0)$ and $\boldsymbol{\Sigma} = \frac{1}{2}\mathbf{I}_{2\times 2}$, the vector of means and the covariance matrix respectively. It can be clearly seen that the minimization of the area under the density function between the points $\mathbf{x} = (-1,0)$ and $\mathbf{y} = (1,0)$ is obtained when we consider the line integral through the arc of the level set curve.

Proposition 4.1 (On the existence, uniqueness and properties of the d_{SW} distance). *The solution to the problem stated in Equation 4.2 is given by the path $\gamma^* \in C^2$ that minimizes the line-integral between the points $\mathbf{x}$ and $\mathbf{y}$. The solution to the problem stated in Equation 4.2 exists provided that the density function $f_{\mathbb{P}}$ is an integrable function. The uniqueness of the path that solves this problem is not required: if γ_1 and γ_2 are two different solutions to the problem stated in 4.2, then the Statistical Work done with both solutions are equal and the distance proposed in Definition 4.5 remains the same.*

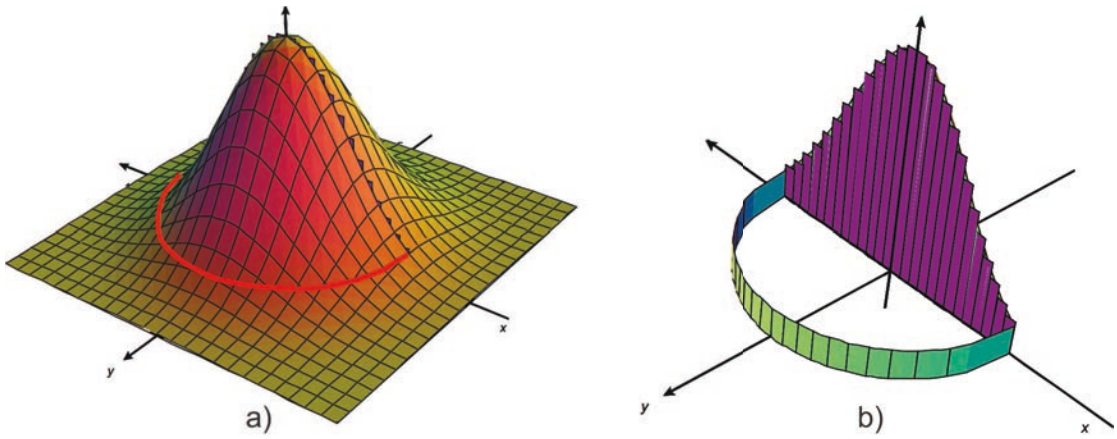


Figure 4.6: Two possible integration paths of the bi-variate normal density function in a) and the resulting integrals in b).

The proposed d_{SW} distance fulfills all the requisites to be a proper metric: it is non-negative (and $d_{SW}(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$), it is symmetric and satisfy the triangle inequality.

Proof for the properties of d_{SW}. The non-negativity follows directly from the definition of the distance. Being γ^* the solution path to the problem stated in Definition 4.5 a smooth curve in X, then $d_{SW}(\mathbf{x},\mathbf{y}) = 0$ if and only if $\int_{\gamma^*} f_{\mathbb{P}}\ dt = 0$. That is: $\mathbf{x} = \mathbf{y}$ or there is no density through the γ^* path. Therefore from the probabilistic point of view the points $\mathbf{x}$ and $\mathbf{y}$ are indistinguishable and $d_{SW}(\mathbf{x},\mathbf{y}) = 0$. The symmetry of the distance is derived from the definition: If $d_{SW}(\mathbf{x},\mathbf{y}) \neq d_{SW}(\mathbf{y},\mathbf{x})$ these contradicts the facts that we reach the infimum in Definition 4.5. That is the paths in one of the directions $\mathbf{x} \xrightarrow{to} \mathbf{y}$ or $\mathbf{x} \xleftarrow{to} \mathbf{y}$ are not optimal. Same argument can be applied to verify the triangle inequality: If there exists a point $\mathbf{z} \in X$ such that $d_{SW}(\mathbf{x},\mathbf{y}) > d_{SW}(\mathbf{x},\mathbf{z}) + d_{SW}(\mathbf{z},\mathbf{y})$, then this is inconsistent with $d_{SW}(\mathbf{x},\mathbf{y}) = \inf_\gamma \int_0^1 f_{\mathbb{P}}\ dt$ (provided that $\gamma(0) = \mathbf{x}$ and $\gamma(1) = \mathbf{y}$). □

Next section presents an estimation procedure to compute the d_{SW} distance when we do not have information about the PM $\mathbb{P}$.

4.3.1 The estimation of the Minimum Work Statistical distance

In this section we discuss how to compute the distance d_{SW} when the parameters that characterize the PM $\mathbb{P}$ are unknown. In the most general case, we only have available a finite sample of data points. Denote by $S_{\mathbb{P}}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ the iid sample drawn from the PM $\mathbb{P}$. Our intention is to compute an estimation of $d_{SW}(\mathbf{x},\mathbf{y})$ by using the information contained in the data

sample $S_{\mathbb{P}}^n$.

There are several possible approaches in order to compute the proposed distance. The first one consists of the estimation of the parameters of the underlying distribution $\mathbb{P}$, using $S_{\mathbb{P}}^n$ as the source of information. Then we could compute, with the aid of numerical methods, the d_{SW} distance between any given pairs of points in the support of the PM $\mathbb{P}$ by using the Definition 4.5. This alternative is usually not relevant, since in most of the cases we do not know to which parametric family of distributions the data belongs to.

Alternatively, we could follow a non-parametric approach. In this case, we can estimate the PM $\mathbb{P}$ (the density or the cumulative distribution function regarding the PM $\mathbb{P}$) by using a non-parametric estimator. There are several drawbacks associated to this approach. In first place the rate of convergence of the non-parametric estimators is usually slow. Additionally, it is also known that the non-parametric estimation of the density becomes intractable as dimension arises.

We propose a different and indirect approach that avoids the need of the explicit estimation of the density. Given a random sample $S_{\mathbb{P}}^n$, then the expected number of sample points that fall between x and y is proportional to the proposed distance, that is:

$$\mathbb{E}_{\mathbb{P}}\left(\sum_{z\in S_{\mathbb{P}}^n} \mathbb{1}_{[z\in[x,y]]}(z)\right) = n\int_x^y f_{\mathbb{P}}(s)\,ds \propto d_{SW}(x,y).$$

We illustrate this fact with the aid of Figures 4.7 where we have plotted the density function $f_{\mathbb{P}}$, the cumulative distribution function $F_{\mathbb{P}}$ and a sample of data points $S_{\mathbb{P}}^n$ on the horizontal axis. As can be seen in Figure 4.7-left, if the points x and y are located near to the center of the distribution (near to the mode), then the number of sample points lying in the interval $[x, y]$ is large, as it is also large the distance $d_{SW}(x, y)$ (and also the distance $d_F(x, y)$ as they are equivalent in dimension 1).

If we assume that in order to move from the point x to the point y we need to "jump" in between the sample points, then in order to reach the point y departing from x in Figure 4.7-left, we will need to make several "jumps". In the other way around, if the points x and y are located in a region of low density, then the number of "jumps" to reach y departing from x (or vice-verse) diminish according to number of sampled points in between x and y, as can seen in Figure 4.7-right. Following this basic principle we identifies the statistical work with the

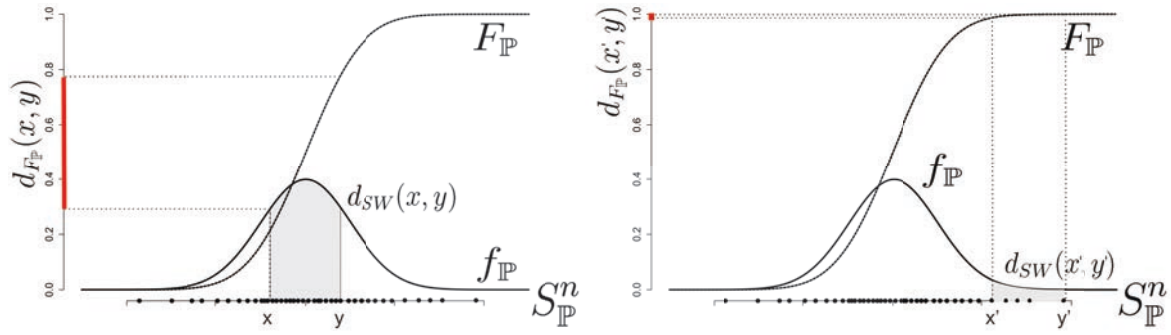


Figure 4.7: The relationship between the PM $\mathbb{P}$, the metric d_{SW} and the random sample $S^n_{\mathbb{P}}$.

number of jumps, and define the estimated minimum work statistical distance, for univariate data, as follows.

Definition 4.6 (Estimation of the Minimum Work Statistical Distance). Let $\mathbb{P}$ be a PM and denotes by $S^n_{\mathbb{P}} : \{x_1, ..., x_n\}$ the set of iid sample points drawn from the PM $\mathbb{P}$. The estimated minimum work statistical distance $\hat{d}_{SW}$ between the points x and y in X is defined by:

$$\hat{d}_{SW}(x, y) = \frac{1}{n} \sum_{z \in S^n_{\mathbb{P}}} \mathbb{1}_{[z \in [x,y]]}(z).$$

It is clear from the previous definition that in the univariate case then $\hat{d}_{SW} = \hat{d}_{F_{S^n_{\mathbb{P}}}}$, as it is expected since both distances are defined as equivalents in the theory. Next we extend the distance estimation procedure to the multivariate context. We estimates d_{SW} in the multivariate case by a minimization of a functional that depends on the number of "jumps" between (local) points. For this purpose we need first to introduce a graph with vertices defined in the coordinates of the sample points $S^n_{\mathbb{P}}$.

We consider a graph $\mathcal{G}(S^n_{\mathbb{P}}) = \{\mathbf{V}, \mathbf{E}\}$, made up of a set of n vertices, where the i^{th}-vertex is represented with the coordinates of the i^{th}-sampled point $\mathbf{x}_i \in S^n_{\mathbb{P}}$. The set of edges (or links) in the matrix $\mathbf{E}$, help us to join the vertices in $\mathbf{V}$ following a neighbor rule specified next. For all $\mathbf{x}_i \in S^n_{\mathbb{P}}$ we denote as $N_k(\mathbf{x}_i) \subset S^n_{\mathbb{P}}$ the set of k-nearest neighbours of $\mathbf{x}_i$ for $i = 1, ..., n$. For $\mathbf{x}_i$ and $\mathbf{x}_j \in S^n_{\mathbb{P}}$ define the function $I_k : S^n_{\mathbb{P}} \times S^n_{\mathbb{P}} \to \{0, 1\}$, where k is a parameter of the function I, in the following way:

$$I_k(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{1}_{[N_k(\mathbf{x}_i)]}(\mathbf{x}_j) + \mathbb{1}_{[N_k(\mathbf{x}_j)]}(\mathbf{x}_i) - \mathbb{1}_{[N_k(\mathbf{x}_i)]}(\mathbf{x}_j)\mathbb{1}_{[N_k(\mathbf{x}_j)]}(\mathbf{x}_i),$$

where $\mathbb{1}_{[N_k(\mathbf{x}_i)]}(\mathbf{x}_j) = 1$ if $\mathbf{x}_j$ belongs to $N_k(\mathbf{x}_i)$ and $\mathbb{1}_{[N_k(\mathbf{x}_i)]}(\mathbf{x}_j) = 0$ otherwise. The parameter k in the function I is defined in the context of the problem. We fill the adjacency matrix

$\mathbf{E}$ by doing: $[\mathbf{E}]_{i,j} \leftarrow I_k(\mathbf{x}_i, \mathbf{x}_j)$. Thus $\mathbf{E}$ is a symmetric matrix that represents the neighbor adjacency relationships between the sampled points.

We estimate the minimum work statistical distance $\hat{d}_{SW}(\mathbf{x}_s, \mathbf{x}_t)$ (for $\mathbf{x}_s$ and $\mathbf{x}_t$ in $S_{\mathbb{P}}^n$) as a proportion of the minimum number of jumps necessary to reach $\mathbf{x}_t$ departing from $\mathbf{x}_s$ on the graph $\mathcal{G}(S_{\mathbb{P}}^n)$. To this aim consider the auxiliary and binary variables $\chi_{i,j}$, such that $\chi_{i,j} = 1$ indicates that the path that connects the nodes i and j is select and $\chi_{i,j} = 0$ otherwise. Denotes by Ω to the set of indexes associated to feasible paths, that is: $\Omega = \{i, j \mid [\mathbf{E}]_{i,j} = 1\}$, then the shortest path problem can be written in the following way:

$$\begin{aligned} &\text{minimize} \sum_{i,j \in \Omega} \chi_{i,j}, \\ &\text{subject to} \begin{cases} \sum_j \chi_{i,j} - \sum_j \chi_{j,i} = 1 & \text{if } i = s, \\ \sum_j \chi_{i,j} - \sum_j \chi_{j,i} = -1 & \text{if } i = t, \\ \sum_j \chi_{i,j} - \sum_j \chi_{j,i} = 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4.3}$$

The restrictions ensures that we depart from the starting point $\mathbf{x}_s$ (the s^{th} node in $\mathbf{V}$) and arrive to the destination point $\mathbf{x}_t$ (the t^{th} node in $\mathbf{V}$). The linear problem stated above it is well studied in the literature, for further details refer to Cherkassky et al. (1996); Ahuja et al. (1988) and reference therein.

Let $\{\chi^*_{i,j}\}_{i,j=1}^n$ be a solution for the problem stated in Equation 4.3, then the estimated minimum work statistical distance $\hat{d}_{SW}(\mathbf{x}_s, \mathbf{x}_t)$ is defined as:

$$\hat{d}_{SW}(\mathbf{x}_s, \mathbf{x}_t) = \frac{1}{n} \sum_{i,j} \chi^*_{i,j}, \tag{4.4}$$

in this way the quantity $\hat{d}_{SW}(\mathbf{x}_s, \mathbf{x}_t)$ represent the approximation to the minimum density path according to Definition 4.5.

In order to exemplify the use of the proposed estimation method we compute the proposed distance for a well known parametrized distribution and compare the theoretical distance with the estimation provided in Equation 4.4. To this end, we generate data from a bi-variate normal distribution with mean vector $\boldsymbol{\mu} = (0, 0)$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}_{2\times 2}$. We considers 6 different scenarios with sample sizes $n = 100$, 200, 500, 1000, 5000 and 10000, respectively. The sample data is represented in Figure 4.8.

We choose 4 different points in the support of the distribution: $\mathbf{x} = (-2, -2)$, $\mathbf{y} = (2, 2)$, $\mathbf{z} = (\frac{1}{2}, 1)$ and $\boldsymbol{\mu} = (0, 0)$ in order to compute 3 distances: $d_{SW}(\mathbf{x}, \mathbf{y})$, $d_{SW}(\mathbf{x}, \boldsymbol{\mu})$ and $d_{SW}(\mathbf{x}, \mathbf{z})$. In the first case, by using the calculus of variations (see the appendix for further details), we derive that the optimal integration path in order to minimize the statistical work between the points $\mathbf{x}$ and $\mathbf{y}$ is given by the path that goes over the level set curve of the density function. With the aid of numerical integration methods we approximate $d_{SW}(\mathbf{x}, \mathbf{y}) \approx 0.00069$. In Figure 4.8 we demonstrates the convergence of the estimated optimal integral path (blue lines) with respect to the theoretical one.

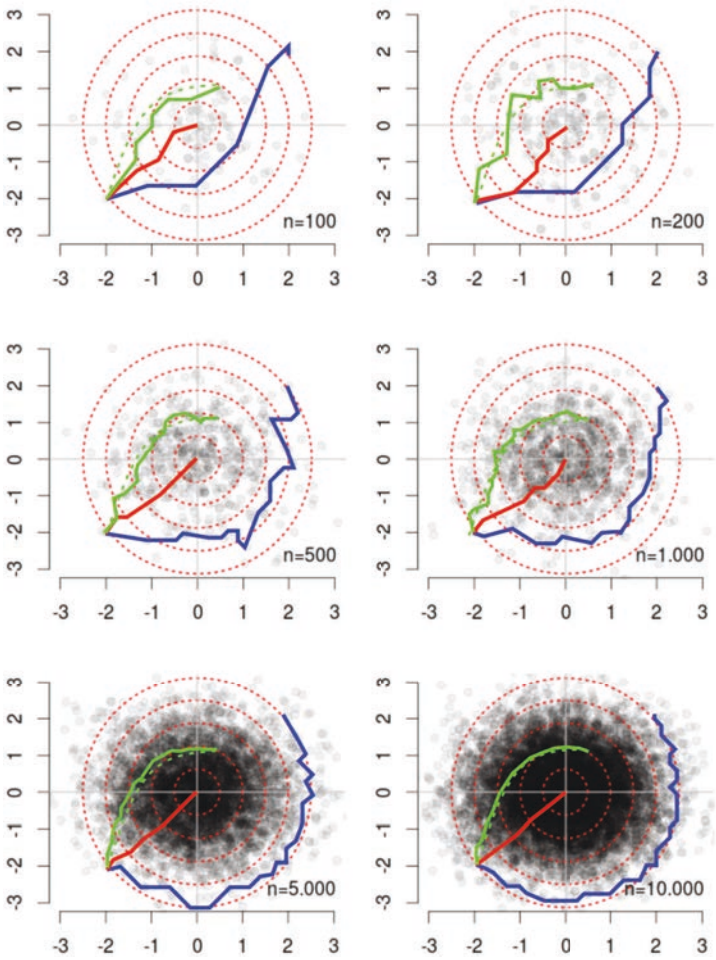


Figure 4.8: The convergence of the estimated integral path from the point $\mathbf{x} = (-2, -2)$ to $\mathbf{y} = (2, 2)$ (in blue), from $\mathbf{x} = (-2, -2)$ to $\boldsymbol{\mu} = (0, 0)$ (in red) and from the point $\mathbf{x} = (-2, -2)$ to $\mathbf{z} = (\frac{1}{2}, 1)$ (in green) for different sample sizes.

The integral path that minimize the area under the density function from the point $\mathbf{x}$ to the center of the distribution is given by a straight line from the point $\mathbf{x}$ to $\boldsymbol{\mu}$. By using numerical integration method we approximates the distance as $d_{SW}(\mathbf{x}, \boldsymbol{\mu}) \approx 0.0011$. In Figure 4.8 we demonstrates the convergence of the estimated optimal integral path (red line) with respect to the theoretical one.

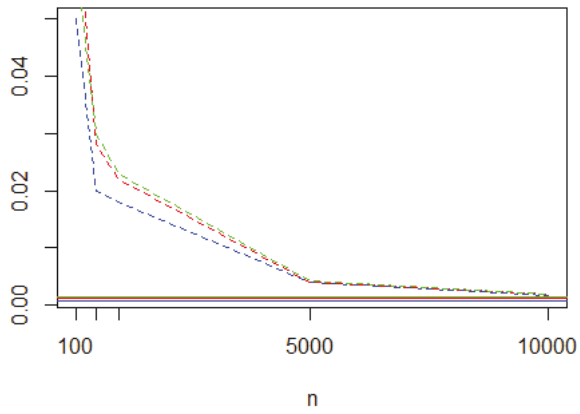


Figure 4.9: Estimated distances (doted lines) vs real distances (horizontal lines).

For the last distance we are not able to find a theoretical solution for the Euler-Lagrange first order differential condition (see the appendix). In this case we prove several paths and obtain numerically an approximated solution of $d_{SW}(\mathbf{x}, \mathbf{z}) \approx 0.0014$. In Figure 4.8 we demonstrates the convergence of the estimated optimal integral path (green line) with respect to the approximated theoretical solution (doted green line) for this case.

In Figure 4.9 we shown the convergence of the estimated distances (doted lines) with respect to the its theoretical values (horizontal lines) for the sample sizes considered in this illustration.

4.4 Experimental Section

In this section we present two experiments to demonstrate the potential of the proposed distances in the context of classification.

An artificial experiment

The aim of the first experiment is to shown the adequacy of the proposed distances in the context of classification. To this end, we generate a sample of size $n = 100$ points drawn from a bivariate normal distribution $\mathbb{P} = N(\boldsymbol{\mu} = (0, 0), 0.05\mathbf{I}_{2\times 2})$ and another 100 points drawn from a uniform distribution $\mathbb{U}$ in the annulus $A(R, r, c)$, that is the region bounded by two concentric circles of radius $R > r$ and centre in c. In this experiment we chose $R = 1$, $r = 0.05$ and the centre in the origin. The generated data is represented in Figure 4.10.

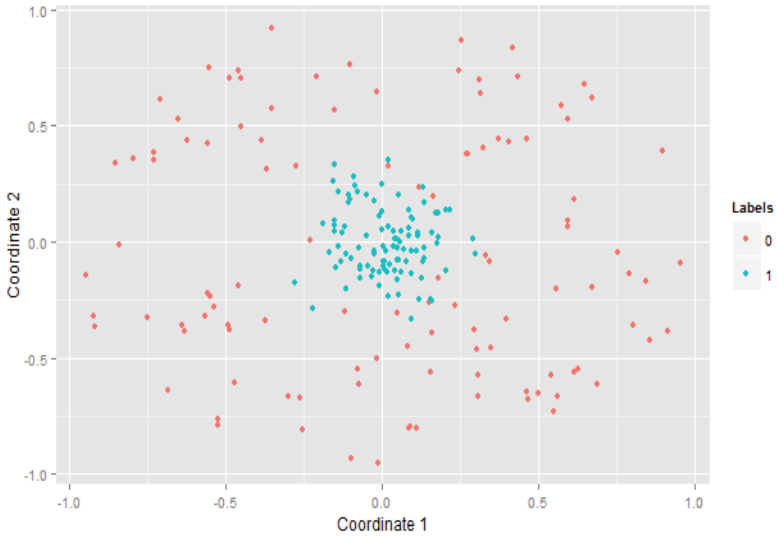


Figure 4.10: Sample points from the distributions $\mathbb{P}$ and $\mathbb{U}$.

We start by using two standard clustering algorithm: k-means clustering and hierarchical cluster combined within several distance measures. In Table 4.1 we demonstrate the performance of the proposed distances compared to another frequently used distances in Statistics and Machine Learning.

The number of points wrongly classified as points that belong to the distribution $\mathbb{P}$ appear in the the column False $\mathbb{P}$. In the column entitled as False $\mathbb{U}$ of Table 4.1 appears the points wrongly classified as points that belongs to the distribution $\mathbb{U}$. The last column is Table 4.1 describes the global percentage of misclassification rate. As can be seen, the small error rate is obtain when we combine the proposed distances with the hierarchical cluster algorithm. In particular, the use of the $\hat{d}_{SW}$ distance produce the best classification result.

Table 4.1: Misclassification performance and total error rate for several standard metrics implemented together with the hierarchical cluster algorithm.

Metric	False $\mathbb{P}$	False $\mathbb{U}$	% total error
k-means	51	41	44%
Euclidean	4	62	33%
Maximum	6	60	31%
Manhattan	6	46	26%
Canberra	53	43	46%
Cosine	50	28	39%
Mahalanobis	5	51	28%
$\hat{d}_F$	4	44	24%
$\hat{d}_{SW}$	**4**	**34**	**19%**

In order to compare the proposed metric with more sophisticated methods of classification we define two kernels in terms of the proposed metrics: $K_F(\mathbf{x}, \mathbf{y}, \sigma) = \exp(-\frac{d_F^2(\mathbf{x},\mathbf{y})}{2\sigma^2})$ and $K_{SW}(\mathbf{x}, \mathbf{y}, \sigma) = \exp(-\frac{d_{SW}^2(\mathbf{x},\mathbf{y})}{2\sigma^2})$. We compare the classification results obtained with a Support Vector Machine trained with the proposed distance based kernels proposed in this chapter and other standard kernels. The result is shown in Table 4.2.

Table 4.2: Misclassification performance and total error rate with Support Vector Machines.

Metric	False $\mathbb{P}$	False $\mathbb{U}$	% total error
svm K_{Lineal}	22	53	35.5%
svm $K_{Polynomial}$	12	50	31.0%
svm K_{RBF}	5	4	4.5%
svm K_F	2	3	2.5%
svm K_{SW}	**2**	**2**	**2.0%**

All the parameters of the classification methods presented in Table 4.2 where optimized by using a 10-folds cross-validation process. Here again the two proposed distances implemented as kernel functions obtains the best classification results. We show in this way that the proposed distances are suitable to be used in classification problems with outstanding results.

A real data experiment

The purpose of this experiment is to identify each of a large number of black-and-white rectangular images that display 3 vowels letters: "A", "I" and "U". We use for this aim the Letter Image Recognition Data base Slate (1991); Frey and Slate (1991). Each image is represented with 16 numerical attributes (statistical moments and edge counts) that represents the main features of the image. In order to make the analysis of the date more simple we decide to use only the first two principal components of the original data.

In Figure 4.11 we can see the 3 groups of letters. Every point represent an image of a vowel letter. We first split the data set into two groups: Training sample (1650 observations) and Test sample (710 observations).

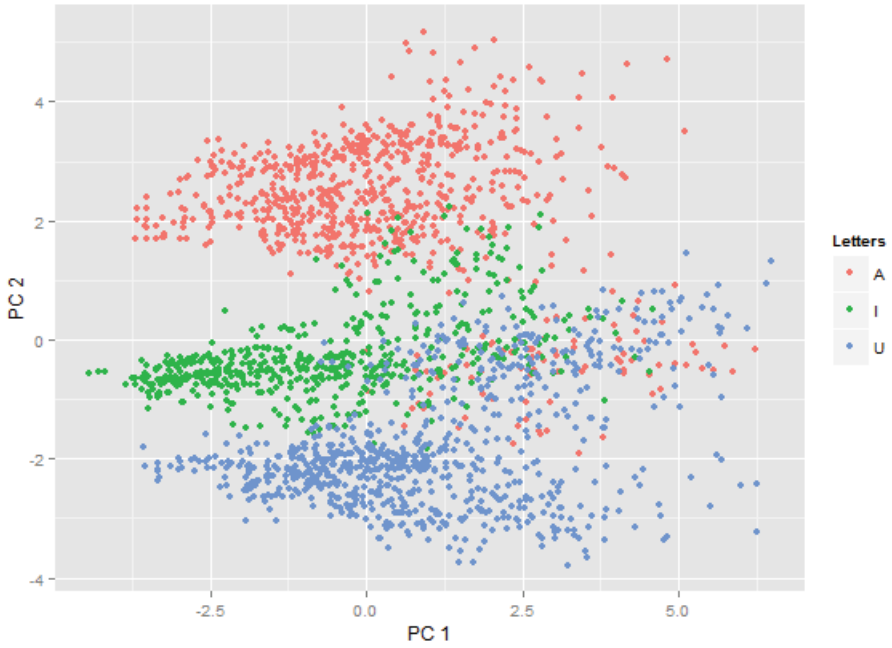


Figure 4.11: A 2D representation of the groups of vowels: "A", "I" and "U".

We train several standard classification methods in Machine Learning and Data Mining. We include also the the classification results obtained with a Support Vector Machine trained with kernels that are based in the proposed distances: $K_F(\mathbf{x},\mathbf{y},\sigma) = \exp(-\frac{d_F^2(\mathbf{x},\mathbf{y})}{2\sigma^2})$ and $K_{SW}(\mathbf{x},\mathbf{y},\sigma) = \exp(-\frac{d_{SW}^2(\mathbf{x},\mathbf{y})}{2\sigma^2})$. A new version of the k-Nearest Neighbors algorithm (k-NN for short)is also

proposed based in the definition of the CDF and minimum statistical work distances. The parameters of all the methods and algorithms are optimized via a 10-fold cross-validation process. The classification results for all the considered methods in the Train and Test data sets are shown in Table 4.3.

Table 4.3: Classification performance for several standard methods in Machine Learning.

Metric	% Error in Train	% Error in Test
svm K_{Lineal}	47.3%	48.8%
svm $K_{Polynomial}$	32.7%	34.4%
svm K_{RBF}	15.2%	15.8%
svm K_F	14.7%	15.1%
svm K_{SW}	14.5%	14.9%
Random Forest	10.5%	11.1%
K-NN	-	11.7%
K-NN based on d_F	-	10.9%
K-NN based on d_{SW}	-	**10.7**%

The best classification result is obtained with the k-NN algorithm based on the d_{SW} distance, that is coherent with the results in Muñoz and Moguerza (2006). The Support Vector Machines implemented with the kernels $K_F(\mathbf{x}, \mathbf{y})$ and K_{SW} outperforms to the RBF-SVM, the usual choice to solve classification problems with SVM's.

The experiments presented in this experimental section show that the proposed distances, combined with standard classification algorithms, outperform the standard methods in Machine Learning and Data Mining that make uses of distances that do not take into account the relevant probabilistic information in the data.

Chapter Summary

In this chapter we propose two distances for multivariate data that takes into account the probabilistic information of the data at hand. We present estimation methods to compute the proposed distances given a data sample. We shown, with the aid of two experiments, that the proposed metrics work well in classification problems.

Chapter 5

A New Family of Probability Metrics in the Context of Generalized Functions[1]

Chapter abstract

In this chapter we study Probability Measures (PM) from a functional point of view: we show that PMs can be considered as functionals (generalized functions) that belong to some functional space endowed with an inner product. This approach allows us to introduce a new family of distances for PMs, based on the action of the PM functionals on 'interesting' functions of the sample. We propose a specific (non parametric) metric for PMs belonging to this class, based on the estimation of density level sets. Some real and simulated data sets are used to measure the performance of the proposed distance against a battery of distances widely used in Statistics and related areas.

Chapter keywords: Probability measures, generalized functions, level sets, distances for data sets, homogeneity tests.

5.1 Introduction

The study of distances between probability measures (PM) is increasingly attracting attention in the fields of Statistics, Data Analysis and Pattern Recognition. For example, the use of probability metrics is of fundamental importance in homogeneity tests, independence tests and goodness of fit problems. These problems can be solved by choosing an appropriate distance

[1] The work presented in this chapter is partially included in the Proceeding of the Artificial Neural Networks and Machine Learning conference (Muñoz et al., 2012) and in (Muñoz et al., 2015).

between PMs. For instance there are some goodness of fit tests based on the use of the χ^2 distance and others that use the Kolmogorov-Smirnoff statistics, which corresponds to the choice of the supremum distance.

More examples of the use of distance measures between PMs can also be found in Clustering Banerjee et al. (2005), Image Analysis Dryden et al. (2009), Time Series Analysis Moon et al. (1995), Econometrics Marriott and Salmon (2000) and Text Mining Lebanon (2006), just to name a few.

For a review of interesting distances between probability distributions and theoretical results, see for instance, Deza and Deza (2009); Zolotarev (1983); Müller (1997) and references therein. Non parametric estimators often play a role in estimating such distances. In practical situations there is usually available a (not huge) data sample, and the use of purely non parametric estimators often results in poor performance Gretton et al. (2006).

An appealing point of view, initiated by Fisher and Rao Burbea and Rao (1982); Amari et al. (1987); Atkinson and Mitchell (1981) and continued with recent development of Functional Data Analysis and Information Geometry Methods Ramsay and Silverman (2002); Amari and Nagaoka (2007), is to consider probability distributions as points belonging to some manifold, and then take advantage of the manifold structure to derive appropriate metrics for distributions. This point of view is used, for instance, in Image and Vision Pennec (2006).

In this chapter we elaborate on the idea that consists of considering PMs as points in a functional space endowed with an inner product, and then derive different distances for PMs from the metric structure inherited from the ambient inner product. We propose particular instances of such metrics for PMs based on the estimation of density level sets regions.

This chapter is organized as follows: In Section 5.2 we review some distances for PMs and represent probability measures as generalized functions; next we define general distances acting on the Schwartz distribution space that contains the PMs. Section 5.3 presents a new distance built according to this point of view. Section 5.4 illustrates the theory with some simulated and real data sets.

5.2 Distances for Probability Distributions

Several well known statistical distances and divergence measures are special cases of f-divergences Csiszár and Shields (2004). Consider two PMs, say $\mathbb{P}$ and $\mathbb{Q}$, defined on a measurable space $(X, \mathcal{F}, \mu)$, where X is a sample space, $\mathcal{F}$ a σ-algebra of measurable subsets of X and $\mu : \mathcal{F} \to \mathbb{R}^+$ the Lebesgue measure. For a convex function f and assuming that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, then the f-divergence from $\mathbb{P}$ to $\mathbb{Q}$ is defined by:

$$d_f(\mathbb{P}, \mathbb{Q}) = \int_X f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q}. \tag{5.1}$$

Some well known particular cases: for $f(t) = \frac{|t-1|}{2}$ we obtain the *Total Variation* metric; $f(t) = (t-1)^2$ yields the χ^2-distance; $f(t) = (\sqrt{t}-1)^2$ yields the *Hellinger* distance.

The second important family of dissimilarities between probability distributions is made up of Bregman Divergences: Consider a continuously-differentiable real-valued and strictly convex function φ and define:

$$d_\varphi(\mathbb{P}, \mathbb{Q}) = \int_X \left(\varphi(p) - \varphi(q) - (p-q)\varphi'(q)\right) d\mu(x), \tag{5.2}$$

where p and q represent the density functions for $\mathbb{P}$ and $\mathbb{Q}$ respectively and $\varphi'(q)$ is the derivative of φ evaluated at q (see Frigyik et al. (2008); Cichocki and Amari (2010) for further details). Some examples of Bregman divergences: $\varphi(t) = t^2$ $d_\varphi(\mathbb{P}, \mathbb{Q})$ yields the Euclidean distance between p and q (in L_2); $\varphi(t) = t\log(t)$ yields the *Kullback Leibler* (KL) Divergence; and for $\varphi(t) = -\log(t)$ we obtain the *Itakura-Saito* distance. In general d_f and d_φ are not metrics because the lack of symmetry and because they do not necessarily satisfy the triangle inequality.

A third interesting family of PM distances are integral probability metrics (IPM) Zolotarev (1983); Müller (1997). Consider a class of real-valued bounded measurable functions on X, say $\mathcal{H}$, and define the IPM between $\mathbb{P}$ and $\mathbb{Q}$ as

$$d_\mathcal{H}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{H}} \left| \int f d\mathbb{P} - \int f d\mathbb{Q} \right|. \tag{5.3}$$

If we choose $\mathcal{H}$ as the space of bounded functions such that $h \in \mathcal{H}$ if $\|h\|_\infty \leq 1$, then $d_\mathcal{H}$ is the Total Variation metric; when $\mathcal{H} = \{\prod_{i=1}^d \mathbb{1}(-\infty, x_i) : x = (x_1, \ldots, x_d) \in \mathbb{R}^d\}$, $d_\mathcal{H}$ is the Kolmogorov distance; if $\mathcal{H} = \{e^{\sqrt{-1}\langle \omega, \cdot \rangle} : \omega \in \mathbb{R}^d\}$ the metric computes the maximum differ-

ence between characteristics functions. In Sriperumbudur et al. (2010b) the authors propose to choose $\mathcal{H}$ as a Reproducing Kernel Hilbert Space and study conditions on $\mathcal{H}$ to obtain proper metrics $d_{\mathcal{H}}$.

In practice, the obvious problem to implement the above described distance functions is that we do not know the density (or distribution) functions corresponding to the samples under consideration. For instance suppose we want to estimate the KL divergence (a particular case of Equation (5.1) taking $f(t) = -\log t$) between two continuous distributions $\mathbb{P}$ and $\mathbb{Q}$ from two given samples. In order to do this we must choose a number of regions, N, and then estimate the density functions for $\mathbb{P}$ and $\mathbb{Q}$ in the N regions to yield the following estimation:

$$\widehat{KL}(\mathbb{P},\mathbb{Q}) = \sum_{i=1}^{N} \hat{p}_i \log \frac{\hat{p}_i}{\hat{q}_i}, \tag{5.4}$$

see further details in Boltz et al. (2009).

As it is well known, the estimation of general distribution functions becomes intractable as dimension arises. This motivates the need of metrics for probability distributions that do not explicitly rely on the estimation of the corresponding probability/distribution functions. For further details on the sample versions of the above described distance functions and their computational subtleties see Scott (2009); Cha (2007); Wang et al. (2005); Nguyen et al. (2010); Sriperumbudur et al. (2010a); Goria et al. (2005); Székely and Rizzo (2004) and references therein.

To avoid the problem of explicit density function calculations we will adopt the perspective of the generalized function theory of Schwartz (see Zemanian (1982), for instance), where a function is not specified by its values but by its behavior as a functional on some space of testing functions.

5.2.1 Probability measures as Schwartz distributions

Consider a measure space $(X, \mathcal{F}, \mu)$, where X is a sample space, here a compact set of a real vector space: $X \subset \mathbb{R}^d$ (a not restrictive assumption in real scenarios, see for instance Moguerza and Muñoz (2006)), $\mathcal{F}$ a σ-algebra of measurable subsets of X and $\mu : \mathcal{F} \to \mathbb{R}^+$ the ambient σ-additive measure (here the Lebesgue measure). A probability measure $\mathbb{P}$ is a σ-additive finite measure absolutely continuous w.r.t. μ that satisfies the three Kolmogorov axioms. By Radon-Nikodym theorem, there exists a measurable function $f : X \to \mathbb{R}^+$ (the density function) such

that $\mathbb{P}(A) = \int_A f d\mu$, and $f_{\mathbb{P}} = \frac{d\mathbb{P}}{d\mu}$ is the Radon-Nikodym derivative.

A PM can be regarded as a Schwartz distribution (a generalized function, see Strichartz (2003) for an introduction to Distribution Theory): We consider a vector space $\mathcal{D}$ of test functions. The usual choice for $\mathcal{D}$ is the subset of $C^\infty(X)$ made up of functions with compact support. A distribution (also named generalized function) is a continuous linear functional on $\mathcal{D}$. A probability measure can be regarded as a Schwartz distribution $\mathbb{P} : \mathcal{D} \to \mathbb{R}$ by defining $\mathbb{P}(\phi) = \langle \mathbb{P}, \phi \rangle = \int \phi d\mathbb{P} = \int \phi(x) f(x) d\mu(x) = \langle \phi, f \rangle$. When the density function $f \in \mathcal{D}$, then f acts as the representer in the Riesz representation theorem: $\mathbb{P}(\cdot) = \langle \cdot, f \rangle$.

In particular, the familiar condition $\mathbb{P}(X) = 1$ is equivalent to $\left\langle \mathbb{P}, \mathbb{1}_{[X]} \right\rangle = 1$, where the function $\mathbb{1}_{[X]}$ belongs to $\mathcal{D}$, being X compact. Note that we do not need to impose that $f \in \mathcal{D}$; only the integral $\langle \phi, f \rangle$ should be properly defined for every $\phi \in \mathcal{D}$.

Hence a probability measure/distribution is a continuous linear functional acting on a given function space. Two given linear functionals $\mathbb{P}_1$ and $\mathbb{P}_2$ will be identical (similar) if they act identically (similarly) on every $\phi \in \mathcal{D}$. For instance, if we choose $\phi = Id$, $\mathbb{P}_1(\phi) = \langle f_{\mathbb{P}_1}, x \rangle = \int x d\mathbb{P} = \mu_{\mathbb{P}_1}$ and if $\mathbb{P}_1$ and $\mathbb{P}_2$ are 'similar' then $\mu_{\mathbb{P}_1} \simeq \mu_{\mathbb{P}_2}$ because $\mathbb{P}_1$ and $\mathbb{P}_2$ are continuous functionals. Similar arguments apply for variance (take $\phi(x) = (x - \mu)^2$) and in general for higher order moments. For $\phi_\xi(x) = e^{ix\xi}$, $\xi \in \mathbb{R}$, we obtain the Fourier transform of the probability measure (called characteristic functions in Statistics), given by $\hat{P}(\xi) = \left\langle \mathbb{P}, e^{ix\xi} \right\rangle = \int e^{ix\xi} d\mathbb{P}$.

Thus, two PMs can be identified with their action as functionals on the test functions if the set of test functions $\mathcal{D}$ is rich enough and hence, distances between two distributions can be defined from the differences between functional evaluations for appropriately chosen test functions.

Definition 5.1. Identification of PM's. Let $\mathcal{D}$ be a set of test functions and $\mathbb{P}$ and $\mathbb{Q}$ two PM's defined on the measure space $(X, \mathcal{F}, \mu)$, then we say that $\mathbb{P} = \mathbb{Q}$ on $\mathcal{D}$ if:

$$\langle \mathbb{P}, \phi \rangle = \langle \mathbb{Q}, \phi \rangle \quad \forall \phi \in \mathcal{D}.$$

The key point in our approach is that if we appropriately choose a finite subset of test functions $\{\phi_i\}$, we can compute the distance between the probability measures by calculating a finite number of functional evaluations. In the next section we demonstrate that when $\mathcal{D}$ is

composed by indicator functions that indicates the regions where the density remains constant, then the set $\mathcal{D}$ is rich enough to identify PM. In the next section we define a distance based on the use of this set of indicator functions.

5.3 A Metric Based on the Estimation of Level Sets

We choose $\mathcal{D}$ as $C_c(X)$, the space of all compactly supported, piecewise continuous functions on X (compact), as test functions (remember that $C_c(X)$ is dense in L_p). Given two PMs $\mathbb{P}$ and $\mathbb{Q}$, we consider a family of test functions $\{\phi_i\}_{i\in I} \subseteq \mathcal{D}$ and then define distances between $\mathbb{P}$ and $\mathbb{Q}$ by weighting terms of the type $d\left(\langle \mathbb{P}, \phi_i\rangle, \langle \mathbb{Q}, \phi_i\rangle\right)$ for $i \in I$, where d is some distance function. Our test functions will be indicator functions of α-level sets, described below.

Given a PM $\mathbb{P}$ with density function $f_{\mathbb{P}}$, minimum volume sets are defined by $S_\alpha(f_{\mathbb{P}}) = \{x \in X |\, f_{\mathbb{P}}(x) \geq \alpha\}$, such that $P(S_\alpha(f_{\mathbb{P}})) = 1 - \nu$, where $0 < \nu < 1$. If we consider an ordered sequence $0 \leq \alpha_1 < \ldots < \alpha_m$, then $S_{\alpha_{i+1}}(f_{\mathbb{P}}) \subseteq S_{\alpha_i}(f_{\mathbb{P}})$. Let us define the α_i-level set: $A_i(\mathbb{P}) = S_{\alpha_i}(f_{\mathbb{P}}) - S_{\alpha_{i+1}}(f_{\mathbb{P}})$, $i \in \{1, \ldots, m-1\}$. We can choose $\alpha_1 \simeq 0$ and $\alpha_m \geq \max_{x\in X} f_{\mathbb{P}}(x)$ (which exists, given that X is compact and $f_{\mathbb{P}}$ piecewise continuous); then $\bigcup_i A_i(\mathbb{P}) \simeq \mathrm{Supp}(\mathbb{P}) = \{x \in X |\, f_{\mathbb{P}}(x) \neq 0\}$ (equality takes place when $m \to \infty$, $\alpha_1 \to 0$ and $\alpha_m \to \max_{x\in X} f_{\mathbb{P}}(x)$). Given the definition of the A_i, if $A_i(\mathbb{P}) = A_i(\mathbb{Q})$ for every i when $m \to \infty$, then $\mathbb{P} = \mathbb{Q}$. We formally prove this proposition with the aid of the following theorem.

Definition 5.2. $\boldsymbol{\alpha}^m_{\mathbb{P}}$ sequence. Given a PM $\mathbb{P}$ defined on the measure space $(X, \mathcal{F}, \mu)$, with density function $f_{\mathbb{P}}$ and $m \in \mathbb{N}$, define $\boldsymbol{\alpha}^m_{\mathbb{P}} = \{\alpha_1, \ldots, \alpha_m\}$ where $0 = \alpha_1 < \ldots < \alpha_m = \max_x f_{\mathbb{P}}(x)$ and the elements of the sequence $\boldsymbol{\alpha}^m_{\mathbb{P}}$ constitutes an asymptotically dense set in $[0, \max_x f_{\mathbb{P}}(x)]$.

Theorem 5.1. ***α-level set representation of a PM.*** *Given a PM $\mathbb{P}$ defined on the measure space $(X, \mathcal{F}, \mu)$, with density function $f_{\mathbb{P}}$ and a sequence $\boldsymbol{\alpha}^m_{\mathbb{P}}$, consider the set of indicator functions $\phi_{i,\mathbb{P}} = \mathbb{1}_{[A_i(\mathbb{P})]} : X \to \{0, 1\}$ of the α-level sets $A_i(\mathbb{P}) = S_{\alpha_i}(f_{\mathbb{P}}) - S_{\alpha_{i+1}}(f_{\mathbb{P}})$ for $i \in \{1, \ldots, m-1\}$. Define $f_m(x) = \sum_{i=1}^m \alpha_i \phi_{i,\mathbb{P}}(x)$. Then:*

$$\lim_{m\to\infty} f_m(x) = f_{\mathbb{P}}(x),$$

where the convergence is pointwise almost everywhere. Moreover, as the sequence f_m is monotonically increasing ($f_{m-1} \leq f_m$), by Dini's Theorem, the convergence is also uniform (converge uniformly almost everywhere).

Proof. Consider $x \in \mathrm{Supp}(\mathbb{P})$; given m and a sequence $\boldsymbol{\alpha}^m_{\mathbb{P}}$, $x \in A_i(\mathbb{P}) = S_{\alpha_i}(f_{\mathbb{P}}) - S_{\alpha_{i+1}}(f_{\mathbb{P}})$ for one (and only one) $i \in \{1, \ldots, m-1\}$, that is $\alpha_i \leq f_{\mathbb{P}}(x) \leq \alpha_{i+1}$. Then $\phi_{i,\mathbb{P}}(x) = \mathbb{1}_{[A_i(\mathbb{P})]}(x) = 1$

in the region $A_i(\mathbb{P})$ and zero elsewhere. Given $\varepsilon > 0$, choose $m > \frac{1}{\varepsilon}$ and $\alpha_{i+1} = \alpha_i + \frac{1}{m}$. Given that $\alpha_i \leq f_{\mathbb{P}}(x) \leq \alpha_{i+1}$, then $|\alpha_i - f_{\mathbb{P}}(x)| \leq \frac{1}{m}$, and thus:

$$|f_{m-1}(x) - f_{\mathbb{P}}(x)| = \left| \sum_{j=1}^{m-1} \alpha_j \phi_{j,\mathbb{P}}(x) - f_{\mathbb{P}}(x) \right| = |\alpha_i - f_{\mathbb{P}}(x)| \leq \frac{1}{m} < \varepsilon.$$

That is $\lim_{m\to\infty} f_{m-1}(x) = f_{\mathbb{P}}(x)$ pointwise and also uniformly by Dini's Theorem. Therefore we can approximate (by fixing $m \gg 0$) the density function as a simple function, made up of linear combination of indicator functions weighted by coefficients that represents the density value of the α-level sets of the density at hand. □

Corollary 5.1. ***α-level sets identification of PMs.*** *If the set of test functions $\mathcal{D}$ contains the indicator functions of the α-level sets, then $\mathcal{D}$ is rich enough to discriminate among PMs.*

Proof. By Theorem 5.1 we can approximate (by fixing $m \gg 0$) the density function as: $f_{\mathbb{P}}(x) \approx \sum_{j=1}^{m-1} \alpha_j \phi_j(x)$, where $\alpha_j = \langle \phi_j, f_{\mathbb{P}} \rangle$ and ϕ_j is the indicator function of the α_j-level set of $\mathbb{P}$. Then if $\langle \phi_j, f_{\mathbb{P}} \rangle \xrightarrow[m\to\infty]{} \langle \phi_j, f_{\mathbb{Q}} \rangle$, for all the indicator functions ϕ_j, then $f_{\mathbb{P}} = f_{\mathbb{Q}}$. □

Now we elaborate on the construction of a metric that is able to identify PM. Denote by $\mathscr{D}_X$ to the set of probability distributions on X and given a suitable sequence of non-decreasing values $\{\alpha_i\}_{i=1}^m$, define: $\mathscr{D}_X \xrightarrow{\phi_i} \mathcal{D} : \phi_i(\mathbb{P}) = \mathbb{1}_{[A_i(\mathbb{P})]}$. We propose distances of the form $\sum_{i=1}^{m-1} w_i d\left(\phi_i(\mathbb{P}), \phi_i(\mathbb{Q})\right)$. Consider, as an example, the measure of the standardized symmetric difference:

$$d\left(\phi_i(\mathbb{P}), \phi_i(\mathbb{Q})\right) = \frac{\mu\left(A_i(\mathbb{P}) \triangle A_i(\mathbb{Q})\right)}{\mu\left(A_i(\mathbb{P}) \cup A_i(\mathbb{Q})\right)}.$$

This motivates the definition of the α-level set semi-metric as follows.

Definition 5.3. Weighted α-level set semi-metric. Given $m \in \mathbb{N}$, consider two sequences: $\boldsymbol{\alpha}_{\mathbb{P}}^m$ and $\boldsymbol{\beta}_{\mathbb{Q}}^m$, for $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then define a family of weighted α-level set distances between $\mathbb{P}$ and $\mathbb{Q}$ by

$$d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^{m-1} w_i d\left(\phi_i(\mathbb{P}), \phi_i(\mathbb{Q})\right) = \sum_{i=1}^{m-1} w_i \frac{\mu\left(A_i(\mathbb{P}) \triangle A_i(\mathbb{Q})\right)}{\mu\left(A_i(\mathbb{P}) \cup A_i(\mathbb{Q})\right)}, \tag{5.5}$$

where $w_i \ldots, w_{m-1} \in \mathbb{R}^+$ and μ is the ambient measure.

Equation (5.5) can be interpreted as a weighted sum of Jaccard distances between the $A_i(\mathbb{P})$ and $A_i(\mathbb{Q})$ sets. For $m \gg 0$, when $\mathbb{P} \approx \mathbb{Q}$, then $d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P}, \mathbb{Q}) \approx 0$ since $\mu\left(A_i(\mathbb{P}) \triangle A_i(\mathbb{Q})\right) \approx 0$ for all $i \in \{1, \ldots, m\}$ (assume $|f_{\mathbb{P}}(x) - f_{\mathbb{Q}}(x)| \leq \varepsilon$ for all x, since $f_{\mathbb{P}} \overset{\varepsilon\to 0}{=} f_{\mathbb{Q}}$ then $\mu\left(A_i(\mathbb{P}) \triangle A_i(\mathbb{Q})\right) \xrightarrow{\varepsilon\to 0} 0 \ \forall i$, because otherwise contradicts the fact that $f_{\mathbb{P}} \overset{\varepsilon\to 0}{=} f_{\mathbb{Q}}$).

Proposition 5.1. ***Convergence of the α-level set semi-metric to a metric.*** *$d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q})$ converges to a metric when $m \to \infty$.*

Proof. If $\lim_{m\to\infty} d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q}) = 0$, then $A_i(\mathbb{P}) \overset{m\to\infty}{=} A_i(\mathbb{Q})\ \forall i$. Thus $f_m^{\mathbb{P}}(x) = \sum_{j=1}^{m} \alpha_j \phi_j(x) = f_m^{\mathbb{Q}}(x)\ \forall m$, and by Theorem 5.1: $f_{\mathbb{P}} = f_{\mathbb{Q}}$. In the other way around if $\mathbb{P} = \mathbb{Q}$ ($f_{\mathbb{P}} = f_{\mathbb{Q}}$) then it is certain that $\lim_{m\to\infty} d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q}) = 0$. □

The semi-metric proposed in Equation (5.5) obeys the following properties: is non-negative, that is $d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q}) \geq 0$ and $\lim_{m\to\infty} d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. For fixed pairs $(\boldsymbol{\alpha},\mathbb{P})$ and $(\boldsymbol{\beta},\mathbb{Q})$ it is symmetric $d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q}) = d_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbb{Q},\mathbb{P})$. Therefore constitutes a proper metric when $m \to \infty$. The semi-metric proposed in Equation (5.5) is invariant under affine transformations (see the Appendix B for a formal proof). In Section 5.3.2 we will propose a weighting scheme for setting the weights $\{w_i\}_{i=1}^{m-1}$.

Of course, we can calculate $d_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ in Equation (5.5) only when we know the distribution function for both PMs $\mathbb{P}$ and $\mathbb{Q}$. In practice there will be available two data samples generated from $\mathbb{P}$ and $\mathbb{Q}$, and we need to define some plug in estimator: Consider estimators $\hat{A}_i(\mathbb{P}) = \hat{S}_{\alpha_i}(f_{\mathbb{P}}) - \hat{S}_{\alpha_{i+1}}(f_{\mathbb{P}})$ (details in subsection 5.3.1), then we can estimate $d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q})$ by

$$\hat{d}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q}) = \sum_{i=1}^{m-1} w_i \frac{\mu\left(\hat{A}_i(\mathbb{P}) \triangle \hat{A}_i(\mathbb{Q})\right)}{\mu\left(\hat{A}_i(\mathbb{P}) \cup \hat{A}_i(\mathbb{Q})\right)}. \tag{5.6}$$

It is clear that $\mu\left(\hat{A}_i(\mathbb{P}) \cup \hat{A}_i(\mathbb{Q})\right)$ equals the total number of points in $\hat{A}_i(\mathbb{P}) \cup \hat{A}_i(\mathbb{Q})$, say $\#\left(\hat{A}_i(\mathbb{P}) \cup \hat{A}_i(\mathbb{Q})\right)$. Regarding the numerator in Equation (5.6), given two level sets, say A and B to facilitate the notation, and the corresponding sample estimates $\hat{A}$ and $\hat{B}$, one is tempted to estimate $\mu(A \triangle B)$, the area of region $A \triangle B$, by $\mu(\widehat{A \triangle B}) = \#(\hat{A} - \hat{B}) \cup \#(\hat{B} - \hat{A}) = \#(A \cup B) - \#(A \cap B)$. However this is incorrect since probably there will be no points in common between $\hat{A}$ and $\hat{B}$ (which implies $\widehat{A \triangle B} = \widehat{A \cup B}$).

In our particular case, the algorithm in Table 1 shows that $\hat{A}_i(\mathbb{P})$ is always a subset of the sample $s_{\mathbb{P}}$ drawn from the density function $f_{\mathbb{P}}$, and we will denote this estimation by $s_{\hat{A}_i(\mathbb{P})}$ from now on. We will reserve the notation $\hat{A}_i(\mathbb{P})$ for the covering estimation of $A_i(\mathbb{P})$ defined by $\cup_j^n B(x_j, r_A)$ where $x_j \in s_{\hat{A}_i(\mathbb{P})}$, $B(x_j, r_A)$ are closed balls with centres at x_j and (fixed) radius r_A Devroye and Wise (1980). The radius is chosen to be constant (for data points in $\hat{A}_i(\mathbb{P})$) because we can assume that density is approximately constant inside region $\hat{A}_i(\mathbb{P})$, if the partition $\{\alpha_i\}_{i=1}^{m}$ of the set is fine enough. For example, in the experimental section, we fix r_A as

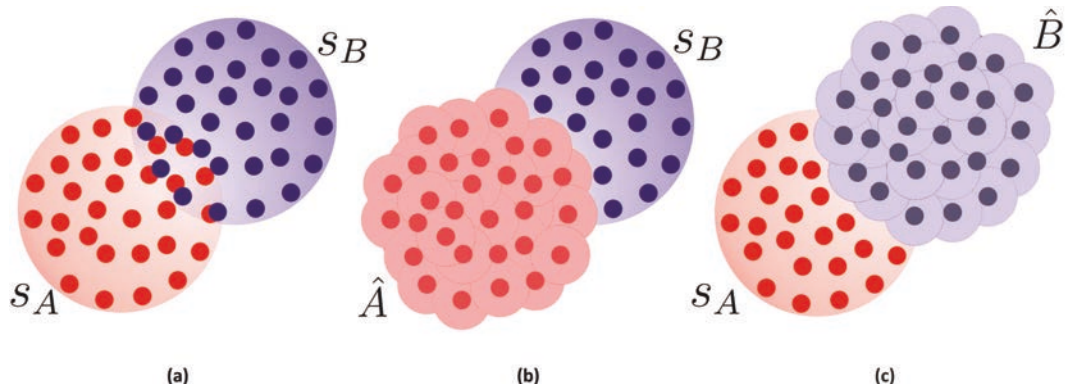


Figure 5.1: Set estimate of the symmetric difference. (a) Data samples s_A (red) and s_B (blue). (b) s_B - Covering $\hat{A}$: blue points. (c) s_A - Covering $\hat{B}$: red points. Blue points in (b) plus red points in (c) are the estimate of $A \triangle B$.

the median distance between the points that belongs to the set $s_{\hat{A}_i(\mathbb{P})}$.

To illustrate this notation we include Figure 5.1. In Figure 5.1 (a) we show two data different samples from α-level sets A and B: s_A (red points) and s_B (blue points), respectively. In Figure 5.1(b) $\hat{A}$ is the covering estimation of set A made up of the union of balls centered in the data red points s_A. That is $\hat{A} = \cup_j^n B(x_j, r_A) \xrightarrow[n\to\infty]{r_A \to 0} A$. Figure 5.1 (c) can be interpreted equivalently regarding the covering of the sample s_B. The problem of calculating $\mu(\widehat{A \triangle B})$ thus reduces to estimate the number points in $\hat{B}$ not belonging to the covering estimate of A, plus the number points in $\hat{A}$ not belonging to the covering estimate of B. To make the computation explicit consider $x \in A, y \in B$ and define

$$I_{r_A, r_B}(x, y) = \mathbb{1}_{[B(x, r_A)]}(y) + \mathbb{1}_{[B(y, r_B)]}(x) - \mathbb{1}_{[B(x, r_A)]}(y)\mathbb{1}_{[B(y, r_B)]}(x),$$

where $I_{r_A, r_B}(x, y) = 1$ when y belongs to the covering $\hat{A}$, x belongs to the covering $\hat{B}$ or both events happen. Thus if we define

$$I(A, B) = \sum_{x \in A} \sum_{y \in B} I_{r_A, r_B}(x, y),$$

we are able to estimate the symmetric difference by

$$\mu(\widehat{A \triangle B}) = \mu(\widehat{A \cup B}) - \mu(\widehat{A \cap B}) = \#\mu(A \cup B) - I(A, B).$$

Table 5.1: Algorithm to estimate minimum volume sets ($S_\alpha(f)$) of a density f.

Estimation of $\mathbf{R_n} = \hat{\mathbf{S}}_\alpha(\mathbf{f})$:	
1	Choose a constant $\nu \in [0, 1]$.
2	Consider the order induced in the sample s_n by the sparsity measure $g_n(x)$, that is, $g_n(x_{(1)}) \leq \cdots \leq g_n(x_{(n)})$, where $x_{(i)}$ denotes the i^{th} sample, ordered after g.
3	Consider the value $\rho_n^* = g(x_{(\nu n)})$ if $\nu n \in \mathbb{N}$, $\rho_n^* = g_n(x_{([\nu n]+1)})$ otherwise, where $[x]$ stands for the largest integer not greater than x.
4	Define $h_n(x) = sign(\rho_n^* - g_n(x))$.

5.3.1 Estimation of level sets

To estimate level sets from a data sample, useful to obtain $\hat{S}_\alpha(f_{\mathbb{P}})$, we use the One-Class Neighbor Machine that solves the following optimization problem:

$$\begin{aligned} \max_{\rho,\xi} \quad & \nu n\rho - \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & g(x_i) \geq \rho - \xi_i \,, \\ & \xi_i \geq 0, \qquad\qquad i = 1, \ldots, n \,, \end{aligned} \tag{5.7}$$

where $g(x) = M(x, s_n)$ is a sparsity measure (see Chapter 3 Section 3.2.3 and reference therein for further details), $\nu \in [0, 1]$ such that $P(S_\alpha) = 1 - \nu$, ξ_i with $i = 1, \ldots, n$ are slack variables and ρ is a predefined constant.

With the aid of the Support Neighbor Machine, we estimate a density contour cluster $S_{\alpha_i}(f)$ around the mode for a suitable sequence of values $\{\nu_i\}_{i=1}^m$ (note that the sequence $0 \geq \nu_1, \ldots, \nu_m = 1$ it is in a one-to-one correspondence with the sequence $0 \leq \alpha_1 < \ldots < \alpha_m = \max_{x \in X} f_{\mathbb{P}}(x)$). In Table 1 we present the algorithm to estimate $S_\alpha(f)$ of a density function f. Hence, we take $s_{\hat{A}_i(\mathbb{P})} = \hat{S}_{\alpha_i}(f_{\mathbb{P}}) - \hat{S}_{\alpha_{i+1}}(f_{\mathbb{P}})$, where $\hat{S}_{\alpha_i}(f_{\mathbb{P}})$ is estimated by R_n defined in Table 5.1 (the same estimation procedure applies for $s_{\hat{A}_i(\mathbb{Q})}$).

The computational complexity of the algorithm of Table 5.1 and more details on the estimation of the regions $\hat{\mathbf{S}}_\alpha(\mathbf{f})$ can be seen in Muñoz and Moguerza (2004, 2005, 2006). The execution time required to compute $\hat{\mathbf{S}}_\alpha(\mathbf{f})$ grows at a rate of order $\mathcal{O}(dn^2)$, where d represent the dimension and n the sample size of the data at hand. We present on the Appendix C a brief explanation about the computational complexity of the algorithm of Table 5.1. We also compare in the Appendix the computational times of the proposed distance against other metrics used in the Experimental section.

Hence following the procedure given in Section 3.2.3, we take $\hat{A}_i(\mathbb{P}) = \hat{S}_{\alpha_i}(f_{\mathbb{P}}) - \hat{S}_{\alpha_{i+1}}(f_{\mathbb{P}})$, where $\hat{S}_{\alpha_i}(f_{\mathbb{P}})$ is estimated by R_n defined in Table 5.1. Theorem 3.1 ensures the convergence of the empirical estimation of the proposed distance. When the sample size increases, we are able to determine with more precision the sets $A_i(\mathbb{P})$ and $A_i(\mathbb{Q})$ and therefore $\hat{d}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q}) \to d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P},\mathbb{Q})$.

5.3.2 Choice of weights for α-level set distances

In this section we explore different weighting schemes for the family of distances defined by Equation (5.5). Denote by $s_{\mathbb{P}}$ and $s_{\mathbb{Q}}$ the data samples corresponding to PMs $\mathbb{P}$ and $\mathbb{Q}$ respectively, and denote by $s_{\hat{A}_i(\mathbb{P})}$ and $s_{\hat{A}_i(\mathbb{Q})}$ the data samples that estimate $A_i(\mathbb{P})$ and $A_i(\mathbb{Q})$, respectively. Remember that we can estimate these sets by coverings $\hat{A}_i(\mathbb{P}) = \cup_{x \in s_{\hat{A}_i(\mathbb{P})}} B(x, r_{\hat{A}_i(\mathbb{P})})$, $\hat{A}_i(\mathbb{Q}) = \cup_{x \in s_{\hat{A}_i(\mathbb{Q})}} B(x, r_{\hat{A}_i(\mathbb{Q})})$.

Let m denote the number of levels in partition $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^m$. Denote by $n_{\hat{A}_i(\mathbb{P})}$ the number of data points in $s_{\hat{A}_i(\mathbb{P})}$, $n_{\hat{A}_i(\mathbb{Q})}$ the number of data points in $s_{\hat{A}_i(\mathbb{Q})}$, $r_{\hat{A}_i(\mathbb{P})}$ the (fixed) radius for the covering $\hat{A}_i(\mathbb{P})$ and $r_{\hat{A}_i(\mathbb{Q})}$ the (fixed) radius for the covering $\hat{A}_i(\mathbb{Q})$, usually the mean or the median distance inside the region $\hat{A}_i(\mathbb{P})$ and $\hat{A}_i(\mathbb{Q})$ respectively. We define the following schemes:

Weighting Scheme 1 : Choose w_i in (5.5) by:

$$w_i = \frac{1}{m} \sum_{x \in s_{\hat{A}_i(\mathbb{P})}}^{n_{\hat{A}_i(\mathbb{P})}} \sum_{y \in s_{\hat{A}_i(\mathbb{Q})}}^{n_{\hat{A}_i(\mathbb{Q})}} \left(1 - I_{r_{\hat{A}_i(\mathbb{P})}, r_{\hat{A}_i(\mathbb{Q})}}(x,y)\right) \frac{\| x - y \|_2}{(s_{\hat{A}_i(\mathbb{Q})} - \hat{A}_i(\mathbb{P})) \sqcup (s_{\hat{A}_i(\mathbb{P})} - \hat{A}_i(\mathbb{Q}))}. \tag{5.8}$$

Weighting Scheme 2 : Choose w_i in (5.5) by:

$$w_i = \frac{1}{m} \max_{x \in s_{\hat{A}_i(\mathbb{P})}, y \in s_{\hat{A}_i(\mathbb{Q})}} \left\{ (1 - I_{r_{\hat{A}_i(\mathbb{P})}, r_{\hat{A}_i(\mathbb{Q})}}(x,y)) \| x - y \|_2 \right\}. \tag{5.9}$$

Weighting Scheme 3 : Choose w_i in (5.5) by:

$$w_i = \frac{1}{m} \hat{H}\left(s_{\hat{A}_i(\mathbb{Q})} - \hat{A}_i(\mathbb{P}), s_{\hat{A}_i(\mathbb{P})} - \hat{A}_i(\mathbb{Q})\right), \tag{5.10}$$

where $\hat{H}(\hat{X}, \hat{Y})$ denotes the Hausdorff distance (finite size version) between finite sets $\hat{X}$ and $\hat{Y}$ (which estimates the "theoretical" Hausdorff distance between space regions X and Y). In this case $X = A_i(\mathbb{P}) - A_i(\mathbb{Q})$ and $Y = A_i(\mathbb{Q}) - A_i(\mathbb{P})$.

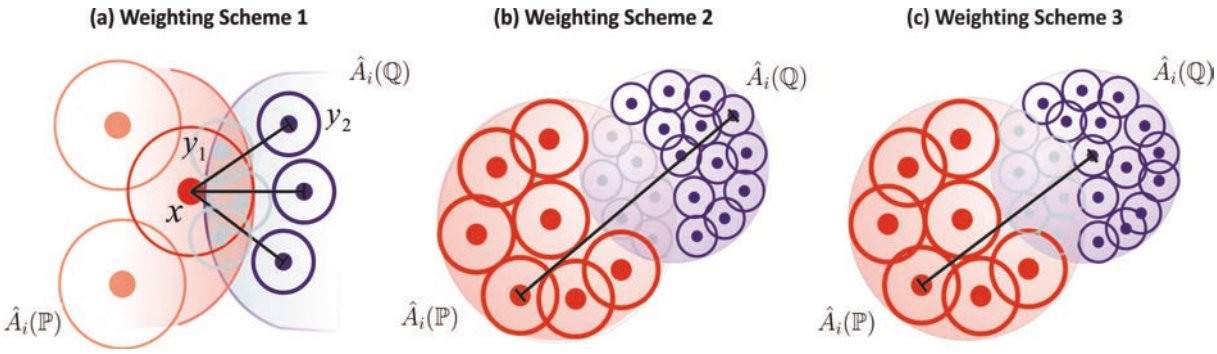


Figure 5.2: Calculation of weights in the distance defined by Equation (5.5).

The intuition behind the three weighting schemes is illustrated in Figure 5.2. In weighting scheme 1 the weight w_i is a weighted average of distances between a point of $s_{\hat{A}_i(\mathbb{P})}$ and a point of $s_{\hat{A}_i(\mathbb{Q})}$ where $\|x-y\|_2$ is taken into account only when $I_{r_{\hat{A}_i(\mathbb{P})}, r_{\hat{A}_i(\mathbb{Q})}}(x,y) = 0$. To illustrate this, consider $x \in s_{\hat{A}_i(\mathbb{P})}$ and $y_1, y_2 \in s_{\hat{A}_i(\mathbb{Q})}$ (Figure 5.2 (a)). The quantity $\|x-y_1\|_2$ does not contribute to calculation of the weight w_i because y_1 belongs to the (red) covering ball centered at x. That is, y_1 belongs to the cover estimation of $A_i(\mathbb{P})$ and therefore should not be taken into account for the calculation of the distance. On the other hand, $\|x-y_2\|_2$ contributes to the calculation of the weight w_i because y_2 does not belong to the (red) covering ball centered at x. In weighting scheme 2 w_i is proportional to the maximum distance between a point belonging to $\hat{A}_i(\mathbb{P})$ and a point belonging to $\hat{A}_i(\mathbb{Q})$, given that the covering balls centered at such points do not overlap. Figure 5.2 (c) illustrates the Hausdorff distance between the sets $s_{\hat{A}_i(\mathbb{Q})} - \hat{A}_i(\mathbb{P})$ and $s_{\hat{A}_i(\mathbb{P})} - \hat{A}_i(\mathbb{Q})$.

5.4 Experimental Section

Since the proposed distance is intrinsically nonparametric, there are no simple parameters on which we can concentrate our attention to do exhaustive benchmarking. The strategy will be to compare the proposed distance to other classical PM distances for some well known (and parametrized) distributions and for real data problems. Here we consider distances belonging to the main types of PMs metrics: Kullback-Leibler (KL) divergence Boltz et al. (2009); Nguyen et al. (2010) (f-divergence and also Bregman divergence), t-test (T) measure (Hotelling test in the multivariate case), Maximum Mean Discrepancy (MMD) distance Gretton et al. (2012); Sriperumbudur et al. (2010b) and Energy distance Székely and Rizzo (2004); Sejdinovic et al. (2013) (an Integral Probability Metric, as it is demonstrated in Sejdinovic et al. (2013)).

5.4.1 Artificial data

Discrimination between normal distributions

In this experiment we quantify the ability of the considered PM distances to test the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ when $\mathbb{P}$ and $\mathbb{Q}$ are multivariate normal distributions. To this end, we generate a data sample of size $100d$ from a normal distribution $N(\mathbf{0}, \mathbf{I}_d) = \mathbb{P}$, where d stands for dimension and then we generate 1000 *iid* data samples of size $100d$ from the same $N(\mathbf{0}, \mathbf{I}_d)$ distribution. Next we calculate the distances between each of these 1000 *iid* data samples and the first data sample to obtain the 95% distance percentile denoted as $d_{H_0}^{95\%}$.

Now define $\boldsymbol{\delta} = \delta\mathbf{1} = \delta(1, \ldots, 1) \in \mathbb{R}^d$ and increase δ by small amounts (starting from 0). For each $\boldsymbol{\delta}$ we generate a data sample of size $100d$ from a $N(\mathbf{0} + \boldsymbol{\delta}, \mathbf{I}_d) = \mathbb{Q}$ distribution. If $d(\mathbb{P}, \mathbb{Q}) > d_{\mathbb{P}}^{95\%}$ we conclude that the present distance is able to discriminate between both populations (we reject H_0) and this is the value δ^* referenced in Table 5.2. To track the power of the test, we repeat this process 1000 times and fix δ^* to the present δ value if the distance is above the percentile in 90% of the cases. Thus we are calculating the minimal value δ^* required for each metric in order to discriminate between populations with a 95% confidence level (type I error $= 5\%$) and a 90% sensitivity level (type II error $= 10\%$). In Table 5.2 we report the minimum distance ($\delta^*\sqrt{d}$) between distributions centers required to discriminate for each metric in several alternative dimensions, where small values implies better results. In the particular case of the T-distance for normal distributions we can use the Hotelling test to compute a p-value to fix the δ^* value.

Table 5.2: $\delta^*\sqrt{d}$ for a 5% type I and 10% type II errors.

Metric	d:	1	2	3	4	5	10	15	20	50	100
KL		0.870	0.636	0.433	0.430	0.402	0.474	0.542	0.536	0.495	0.470
T		0.490	0.297	0.286	0.256	0.246	0.231	0.201	0.182	0.153	0.110
Energy		0.460	0.287	0.284	0.256	0.250	0.234	0.203	0.183	0.158	0.121
MMD		0.980	0.850	0.650	0.630	0.590	0.500	0.250	0.210	0.170	0.130
LS(0)		0.490	0.298	0.289	0.252	0.241	0.237	0.220	0.215	0.179	0.131
LS(1)		**0.455**	**0.283**	**0.268**	**0.240**	**0.224**	**0.221**	**0.174**	**0.178**	**0.134**	**0.106**
LS(2)		**0.455**	**0.283**	**0.268**	**0.240**	0.229	0.231	0.232	0.223	0.212	0.134
LS(3)		0.470	0.284	0.288	0.300	0.291	0.237	0.240	0.225	0.219	0.141

The data chosen for this experiment are ideal for the use of the T statistics that, in fact,

outperforms KL and MMD. However, Energy distance works even better than T distance in dimensions 1 to 4. The LS(0) distance work similarly to T and Energy until dimension 10. LS(2) works similarly to SL(1), the best distance in discrimination power, until dimension 4.

In a second experiment we consider again normal populations but different variance-covariance matrices. Define as an expansion factor $\sigma \in \mathbb{R}$ and increase σ by small amounts (starting from 0) in order to determine the smallest σ^* required for each metric in order to discriminate between the $100d$ sampled data points generated for the two distributions: $N(\mathbf{0}, \mathbf{I}_d) = \mathbb{P}$ and $N(\mathbf{0}, (1+\sigma)\mathbf{I}_d) = \mathbb{Q}$. If $d(\mathbb{P}, \mathbb{Q}) > d_{\mathbb{P}}^{95\%}$ we conclude that the present distance is able to discriminate between both populations and this is the value $(1+\sigma^*)$ reported in Table 5.2. To make the process as independent as possible from randomness we repeat this process 1000 times and fix σ^* to the present σ value if the distance is above the 90% percentile of the cases, as it was done in the previous experiment.

Table 5.3: $(1 + \sigma^*)$ for a 5% type I and 10% type II errors.

Metric	dim:	1	2	3	4	5	10	15	20	50	100
KL		3.000	1.700	1.250	1.180	1.175	1.075	1.055	1.045	1.030	1.014
T		–	–	–	–	–	–	–	–	–	–
Energy		1.900	1.600	1.450	1.320	1.300	1.160	1.150	1.110	1.090	1.030
MMD		6.000	4.500	3.500	2.900	2.400	1.800	1.500	1.320	1.270	1.150
LS(0)		1.850	1.450	1.300	1.220	1.180	1.118	1.065	1.040	1.030	1.012
LS(1)		**1.700**	**1.350**	**1.150**	**1.120**	**1.080**	**1.050**	**1.033**	**1.025**	**1.015**	**1.009**
LS(2)		1.800	1.420	1.180	1.150	1.130	1.080	1.052	1.030	1.025	1.010
LS(3)		1.800	1.445	1.250	1.210	1.180	1.120	1.115	1.090	1.050	1.040

There are no entries in Table 5.3 for the T distance because it was not able to distinguish between the considered populations in none of the considered dimensions. The MMD distance do not show a good discrimination power in this experiment. We can see here again that the proposed LS(1) distance is better than the competitors in all the dimensions considered, having the LS(2), LS(3) similar performance in the second place and LS(0) and the KL similar performance in the third place among the metrics with best discrimination power.

We also perform the previous experiment by considering a permutation test. The results, not include here as we consider them redundant, can be seen in the appendix C.

Homogeneity tests

This experiment concerns a homogeneity test between two populations: a mixture between a Normal and a Uniform distribution ($\mathbb{P} = \alpha N(\mu = 1, \sigma = 1) + (1 - \alpha)U(a = 1, b = 8)$ where $\alpha = 0.7$) and a Gamma distribution ($\mathbb{Q} = \gamma(shape = 1, scale = 2)$). To test the null hypothesis: $H_0 : \mathbb{P} = \mathbb{Q}$ we generate two random i.i.d. samples of size 100 from $\mathbb{P}$ and $\mathbb{Q}$, respectively. Figure 5.3 shows the corresponding density functions for $\mathbb{P}$ and $\mathbb{Q}$.

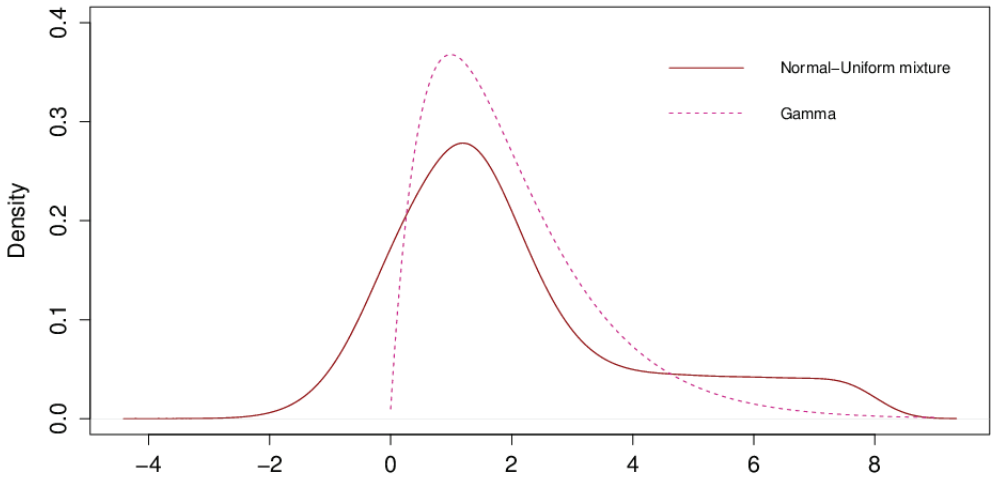


Figure 5.3: Mixture of a Normal and a Uniform Distribution and a Gamma distribution.

In the cases of KL-divergence, T, Energy MMD and LS distances we proceed as in the previous experiment, we run a permutation test based on 1000 random permutation of the original data in order to compute the p-value. In the case of Kolmogorov-Smirnov, χ^2 and Wilcoxon test we report the p-value given by these tests. Results are displayed in Table 5.4: Only the LS distances are able to distinguish between both distributions. Notice that first and second order moments for both distribution are quite similar in this case ($\mu_{\mathbb{P}} = 2.05 \simeq \mu_{\mathbb{Q}} = 2$ and $\sigma_{\mathbb{P}} = 4.5 \simeq \sigma_{\mathbb{Q}} = 4$) and additionally both distributions are strongly asymmetric, which also contributes to explain the failure of those metrics strongly based on the use of the first order moments.

Table 5.4: Hypothesis test (α-significance at 5%) between a mixture of Normal and Uniform distributions and a Gamma distribution.

Metric	Parameters	p-value	Reject?
Kolmogorov-Smirnov		0.281	No.
χ^2 test		0.993	No.
Wilcoxon test		0.992	No.
KL	$k = 10$	0.248	No.
T		0.342	No.
Energy		0.259	No.
MMD		0.177	No.
LS (0)	$m = 15$	0.050	**Yes.**
LS (1)	$m = 15$	**0.035**	**Yes.**
LS (2)	$m = 15$	0.040	**Yes.**
LS (3)	$m = 15$	0.050	**Yes.**

5.4.2 Real case-studies

Shape classification

As an application of the preceding theory to the field of pattern recognition problem we consider the MPEG7 CE-Shape-1 Latecki et al. (2000), a well known shape database. We select four different classes of objects/shapes from the database: hearts, coups, hammers and bones. For each object class we choose 3 images in the following way: 2 standard images plus an extra image that exhibit some distortion or rotation (12 images in total). In order to represent each shape we do not follow the usual approach in pattern recognition that consists in representing each image by a feature vector catching its relevant shape aspects; instead we will look at the image as a cloud of points in $\mathbb{R}^2$, according to the following procedure: Each image is transformed to a binary image where each pixel assumes the value 1 (white points region) or 0 (black points region) as in Figure 5.4 (a). For each image i of size $N_i \times M_i$ we generate a uniform sample of size $N_i M_i$ allocated in each position of the shape image i. To obtain the cloud of points as in Figure 5.4 (b) we retain only those points which fall into the white region (image body) whose intensity gray level are larger than a variable threshold fixed at 0.99 so as to yield around one thousand and two thousand points image representation depending on the image as can be seen in Figure 5.4 (b).

After rescaling and centering, we compute the 12×12 image distance matrices, using the LS(2) distance and the KL divergence, and then compute Euclidean coordinates for the images via MDS (results in Figure 5.5). It is apparent that the LS distance produces a MDS map coherent with human image perception (Figure 5.4 (a)). This does not happen for the rest of tested

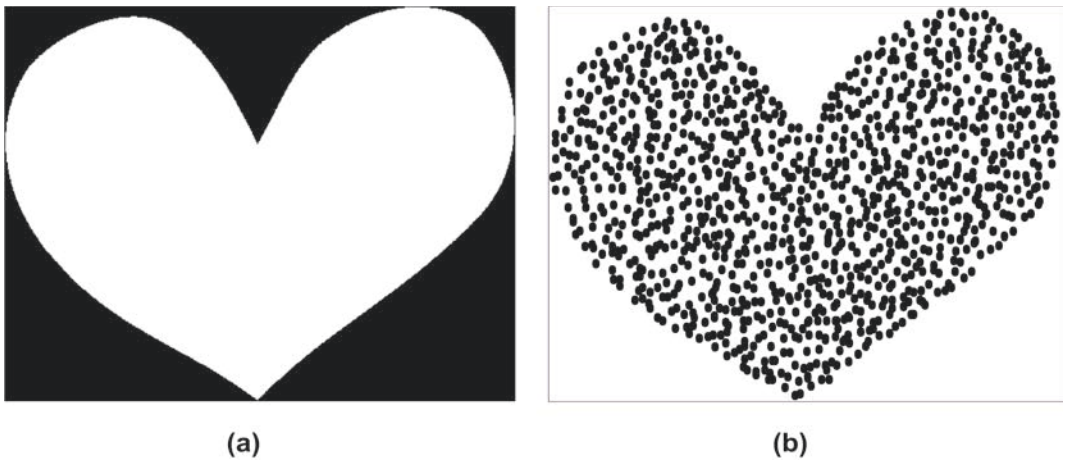


Figure 5.4: Real image (a) and sampled image (b) of a hart in the MPEG7 CE-Shape-1 database.

metrics, in particular for the KL divergence as it is shown in Figure 5.4 (b)).

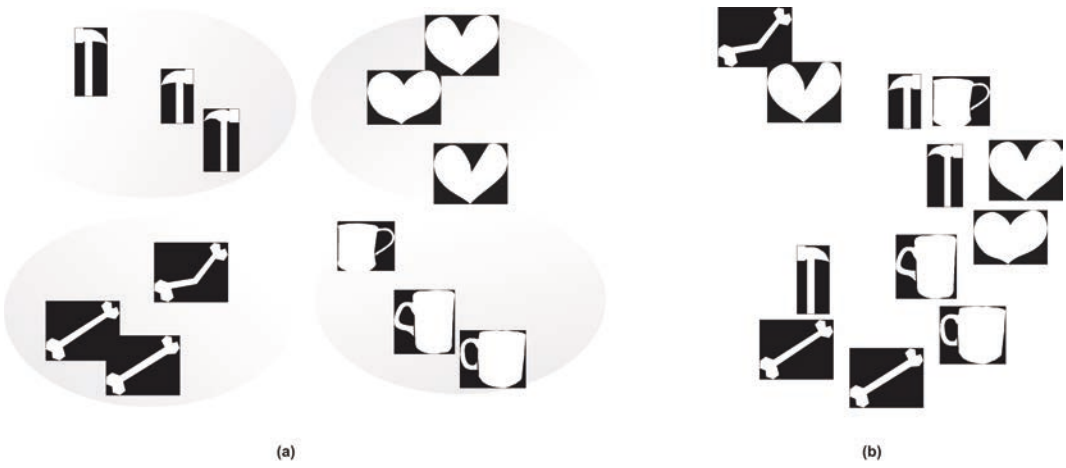


Figure 5.5: Multi Dimensional Scaling representation for objects based on (a) LS(2) and (b) KL divergence.

In order to compare our metric with other competitor metric we provide another similar examples, in this case by using the Tree Leaf Database of Information Theory and ASCR (2000). For this second experiment regarding shape classification, each leaf is represented by a cloud of points in $\mathbb{R}^2$, as in the previous experiment. As an example of the treatment given to a leaf consider the Figure 5.6.

Figure 5.6: Real image and sampled image of a leaf in the Tree Leaf Database.

After rescaling and centering, we computed the 10×10 distance matrix using the LS(1) distance and the Energy distance in this case. We project the shape images by using the Multidimensional Scaling as it is shown in Figure 5.7. It is clear that the LS(1) distance is able to better account for differences in shapes.

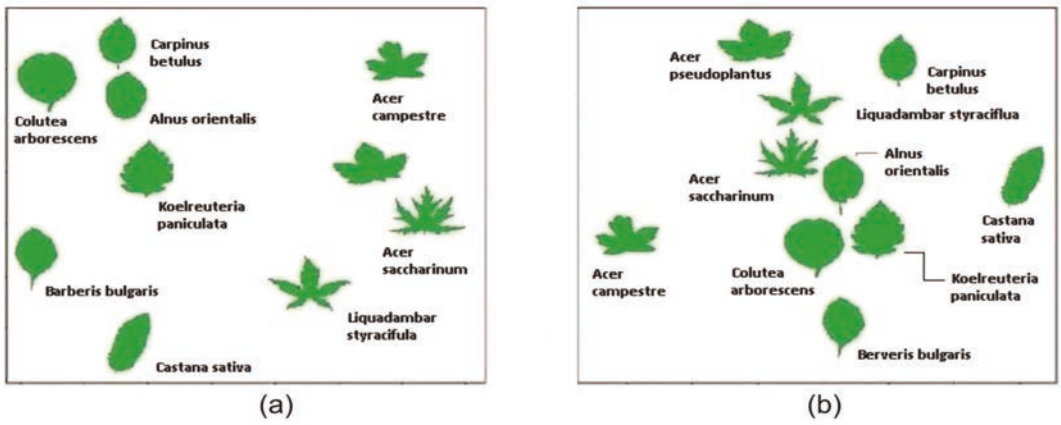


Figure 5.7: MDS representation for leaf database based on LS(1) (a); Energy distance (b).

Texture classification

To continue evaluating the performance of the proposed family of distances between PM's, we consider another application of distances in Image and Vision: texture classification.

To this aim, we consider 9 populations of textures from the Kylberg texture data set Kylberg (2011): 'blanket', 'canvas', 'seat', 'oatmeal', 'rice', 'lentils', 'linseeds', 'stone1', 'stone2'. There are $160 \times 9 = 1.440$ images of textures in total with a resolution of 576×576 pixels. We represent each image using the first 32 parameters of the wavelet representation proposed in Mallat (1989).

Next we calculate the distances between the populations of textures using the LS(1) distance and we obtain, via the multidimesional scaling, the 2D representation of the textures that it is shown in Figure 5.8. It is apparent that textures get organized in a very coherent way with the human criteria, what seems to indicate that the proposed distance is also suitable to solve real pattern recognition problems (high dimensional data and a small number of instances) in the context of texture recognition/classification.

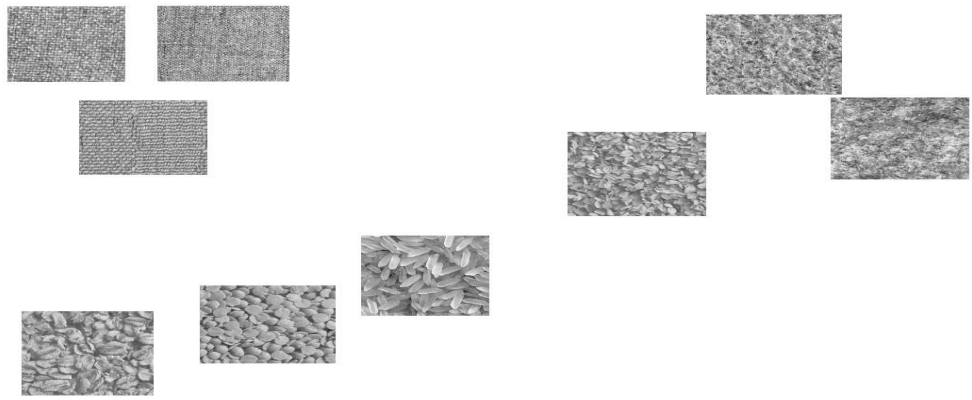


Figure 5.8: MDS plot for texture groups. A representer for each class is plotted in the map.

In Figure 5.9 we present the dendrogram with the cluster obtained for the Leaf data set by using a Hierarchical clustering algorithm combined with the proposed LS(1) metric.

Text Mining

In this experiment we consider a collection of 1774 documents, corresponding to 13 different topics, extracted from three bibliographic data bases: LISA, INSPEC and Sociological Abstracts. We present a brief summary list of topics considered in each data base:

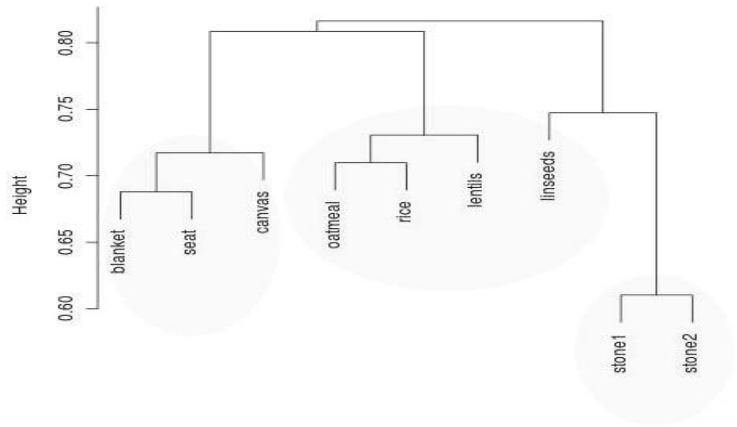


Figure 5.9: Dendrogram with shaded image texture groups.

```
LISA # Documents
---------------------------
business archives in de     137
lotka's law  90
biology in de 280
automatic abstracting 69
-----------------------------
INSPEC:
-------------------------------
Self organizing maps 83
dimensionality reduction 75
power semiconductor devices 170
optical cables 214
feature selection 236
---------------------------------
SOCIOLOGICAL ABSTRACTS:
---------------------------------
Intelligence tests 149
Retirement communities 74
Sociology of literature and discourse 106
Rural areas and rural poverty 91
```

Each document is converted into a vector into the Latent Semantic Space (see for example Landauer et al. (1998) for details) using the Singular Value Decomposition (SVD), and the documents corresponding to one topic are considered as a sample from the underlying distribution that generates the topic. Next we calculate the 13×13 distance matrix by using the LS(3) distance and project each document in the plane by using the Multidimensional Scaling, not on the individual documents, but on the document sets. The result is shown in Figure 5.10, where we can see that close groups correspond to close (in a semantic sense) topics, that indicates the distance is working properly in a nonparametric setting in high dimension.

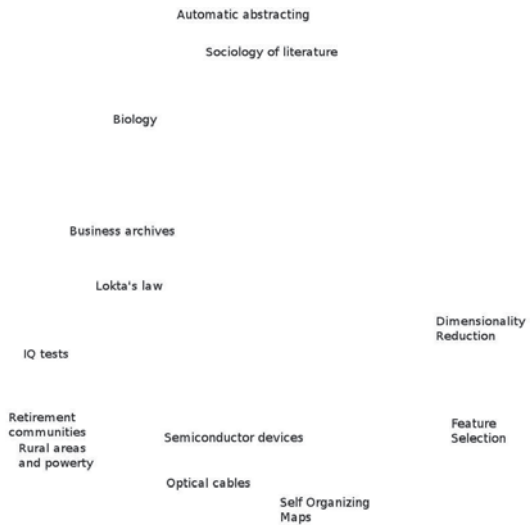


Figure 5.10: Multidimensional Scaling of the 13 groups of documents.

In Figure 5.11 we present the dendrogram with the cluster obtained by using a Hierarchical clustering algorithm combined with the proposed LS(3) metric.

Testing statistical significance in Microarray experients

Here we present an application of the proposed LS distance in the field of Bioinformatics. The data set we analyze comes from an experiment in which the time to respiratory recovery in ventilated post trauma patients is studied. Affymetrix U133+2 micro-arrays were prepared at days 0, 1, 4, 7, 14, 21 and 28. In this analysis, we focus on a subset of 48 patients which were originally divided into two groups: "early recovery patients" (group G_1) that recovered ventilation prior to day seven and "late recovery patients " (group G_2), those who recovered

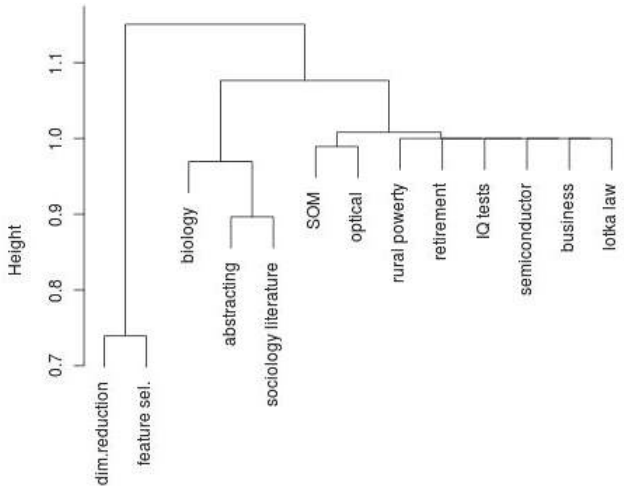


Figure 5.11: Dendrogram for the 13×13 document data set distance.

ventilation after day seven. The size of the groups is 22 and 26 respectively.

It is of clinical interest to find differences between the two groups of patients. In particular, the originally goal of this study was to test the association of inflammation on day one and subsequent respiratory recovery. In this experiment we will show how the proposed distance can be used in this context to test statistical differences between the groups and also to identify the genes with the largest effect in the post trauma recovery.

From the original data set [1] we select the sample of 675 probe sets corresponding to those genes whose GO annotation include the term "inflammatory". To do so we use a query (July 2012) on the Affymetrix web site (http://www.affymetrix.com/index.affx). The idea of this search is to obtain a pre-selection of the genes involved in post trauma recovery in order to avoid working with the whole human genome.

Figure 5.12 shows the heat map of day one gene expression for the 46 patients (columns) over the 675 probe-sets. By using a hierarchical procedure, it is apparent that the two main clusters we find do not correspond to the two groups of patients of the experiment. However, the first cluster (on the left hand of the plot) contains mainly patients form the "early recovery" group (approx. 65 %) whereas the second cluster (on the right) is mainly made up of patients from the "late recovery" group (approx 60%). This lack of balance suggests a different pattern

[1] Available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13488

of gene expression between the two groups of patients.

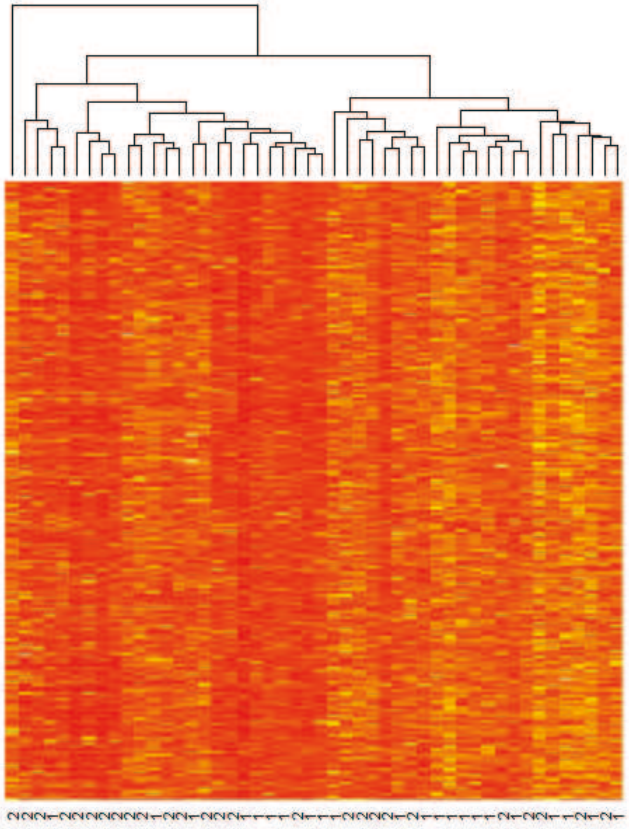


Figure 5.12: Affymetrix U133+2 micro-arrays data from the post trauma recovery experiment. On top, a hierarchical cluster of the patients using the Euclidean distance is included. At the bottom of the plot the grouping of the patients is shown: 1 for "early recovery" patients and 2 for "late recovery" patients.

In order to test if statistical differences exists between the groups G_1 and G_2 we define, inspired by Hayden et al. (2009), an statistical test based on the LS distance proposed in this work. To this end, we identify each patient i with a probability distribution $\mathbb{P}_i$. The expression of the 675 genes across the probe-sets are assumed to be samples of such distributions. Ideally, if the genes expression does not have any effect on the recovery speed then all distributions $\mathbb{P}_i$ should be equal (H_0). On the other hand, assume that expression of a gene or a group of genes effectively change between "early" and "late" recovery patients. Then, the distributions $\mathbb{P}_i$ will be different between patients belonging to groups G_1 and G_2 (H_1).

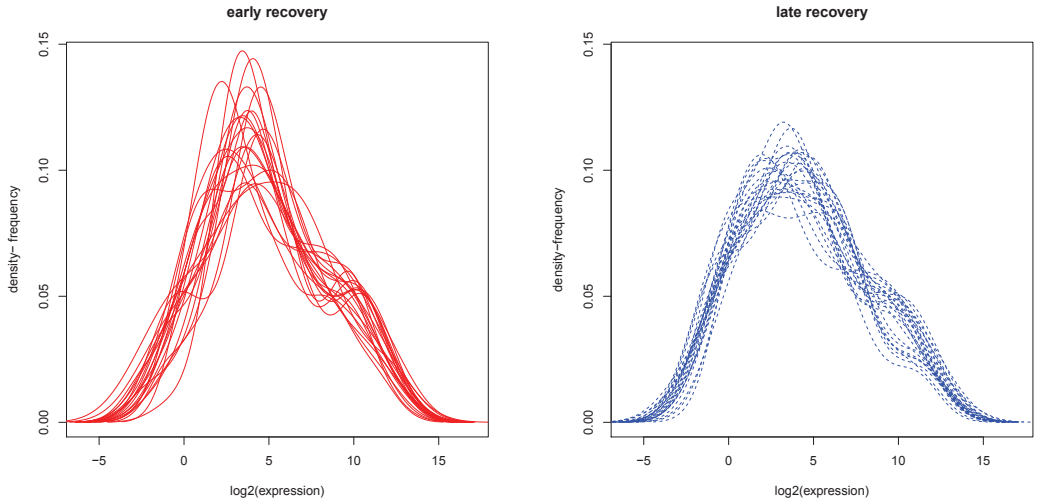


Figure 5.13: Gene density profiles (in logarithmic scale) of the two groups of patients in the sample. The 50 most significant genes were used to calculate the profiles with a kernel density estimator.

To validate or reject the previous hypothesis, consider the proposed LS distance $\hat{d}_{\alpha,\beta}(\mathbb{P}_i, \mathbb{P}_j)$ defined in (5.6) for two patients i and j. Denote by

$$\Delta_1 = \frac{1}{22(22-1)} \sum_{i,j \in G_1} \hat{d}_{\alpha,\beta}(\mathbb{P}_i, \mathbb{P}_j), \quad \Delta_2 = \frac{1}{26(26-1)} \sum_{i,j \in G_2} \hat{d}_{\alpha,\beta}(\mathbb{P}_i, \mathbb{P}_j) \tag{5.11}$$

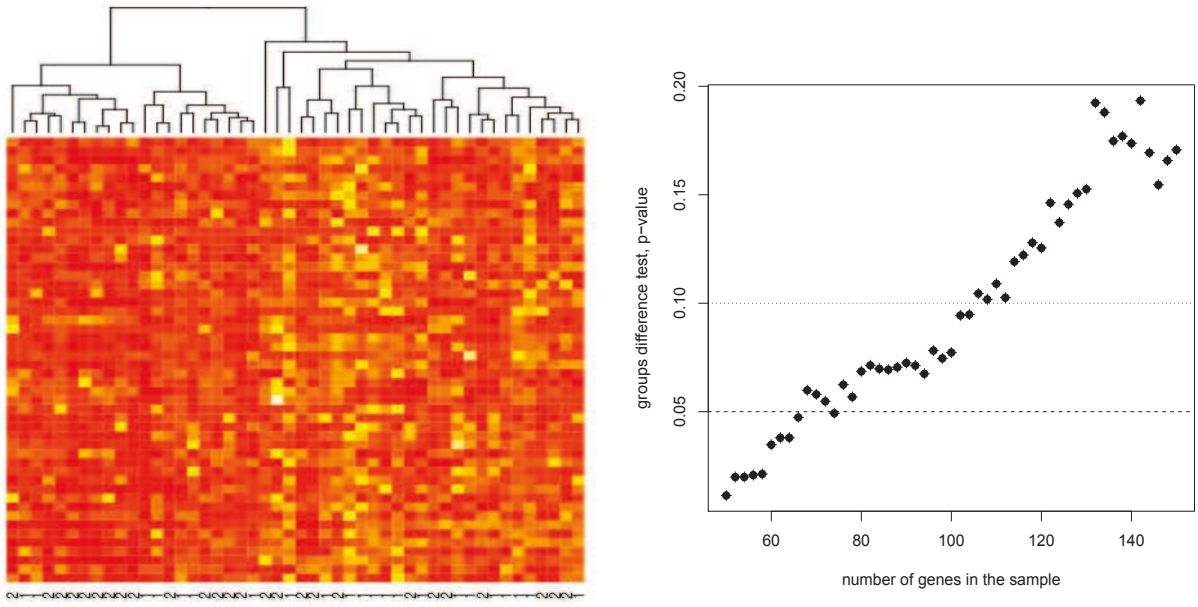
and

$$\Delta_{12} = \frac{1}{22 \cdot 26} \sum_{i \in G_1, j \in G_2} \hat{d}_{\alpha,\beta}(\mathbb{P}_i, \mathbb{P}_j), \tag{5.12}$$

the averaged α-level set distances within and between the groups of patients. Using the previous quantities we define a distance between the groups G_1 and G_2 as

$$\Delta^* = \Delta_{12} - \frac{\Delta_1 + \Delta_2}{2}. \tag{5.13}$$

Notice that if the distributions are equal between the groups then Δ^* will be close to zero. On the other hand, if the distributions are similar within the groups and different between them, then Δ^* will be large. To test if Δ^* is large enough to consider it statistically significant we need the distribution of Δ^* under the null hypothesis. Unfortunately, this distribution is unknown and some re-sampling technique must be used. In this work we approximate it by calculating a sequence of distances $\Delta^*_{(1)}, \ldots, \Delta^*_{(N)}$ where each $\Delta^*_{(k)}$ is the distance between the



(a) Heat-map of the top-50 ranked genes (rows). A hierarchical cluster of the patients is included on top. The labels of the patients regarding their recovery group are detailed at the bottom of the plot.

(b) P-values obtained by the proposed α-level set distance based test for different samples of increasing number of genes.

Figure 5.14: Heat-map of the 50-top ranked genes and p-values for different samples.

groups G_1 and G_2 under a random permutation of the patients. For a total of N permutations, then

$$p - value = \frac{\#[\Delta^*_{(k)} \geq \Delta^* : k = 1, \ldots, N]}{N}. \tag{5.14}$$

where $\#[\Theta]$ refers to the number of times the condition Θ is satisfied, is a one-side p-value of the test.

We apply the previous LS distance based test (weighting scheme 1 with 10000 permutations) using the values of the 675 probe-sets and we obtain a p-value $= 0.1893$. This result suggests that none differences exists between the groups exist. The test for micro-arrays proposed in Hayden et al. (2009) also confirms this result with a p-value of 0.2016. The reason to explain this -a priori- unexpected result is that, if differences between the groups exist, they are probably hidden by a main group of genes with similar behaviour between the groups. To validate this hypothesis, we first rank the set of 675 genes in terms of their individual variation between groups G_1 and G_2. To do so, we use the p-values of individual difference mean T-tests. Then, we consider the top-50 ranked genes and we apply the α-level set distance test. The obtained p-value is 0.010, indicating a significant difference in gene expression of the top-

50 ranked genes. In Figure 5.13 we show the estimated density profiles of the patients using a kernel estimator. It is apparent that the profiles between groups are different as it is reflected in the obtained results. In Figure 5.14, we show the heat-map calculated using the selection of 50 genes. Note that a hierarchical cluster using the Euclidean distance, which is the most used technique to study the existence of groups in micro-array data, is not able to accurately reflect the existence of the two groups even for the most influential genes.

To conclude the analysis we go further from the initial 50-genes analysis. We aim to obtain the whole set of genes in the original sample for which differences between groups remain significant. To do so, we sequentially include in the first 50-genes sample the next-highest ranked genes and we apply to the augmented data sets the LS distance based test. The p-values of such analysis are shown in Figure 5.14. Their value increases as soon as more genes are included in the sample. With a type-I error of 5%, statistical differences are found for the 75 first genes. For a 10% type-I error, with the first 110 genes we still are able to find differences between groups. This result shows that differences between "early" and "late" recovery trauma patients exist and they are caused by the top-110 ranked genes of the Affymetrix U133+2 micro-arrays (filtered by the query "inflammatory"). By considering each patient as a probability distribution the LS distance has been used to test differences between groups and to identify the most influential genes of the sample. This shows the ability of the new proposed distance to provide new insights in the analysis of biological data.

Chapter Summary

In this chapter we present a new family of distance measures for PM. The calculation of these PM distances does not require the use of either parametric assumptions or explicit probability estimations, which makes a clear advantage over most well established PM distances.

A battery of real and simulate examples have been used to study the performance of the new family of distances. Using synthetically generated data, we have shown their performance in the task of discriminating normally distributed data. Regarding the practical applications, the new PM distances have been proven to be competitive in shape recognition problems and text mining. Also they represent a novel way to identify genes and discriminate between groups of patients in micro-arrays.

Chapter 6

A Flexible and Affine Invariant k-Means Clustering Method for Sets of Points[1]

Chapter abstract

In this chapter we propose a novel k-mean clustering method for sets of points based on an affine invariant dissimilarity index. The dissimilarity index is computed with the aid of a kernel for sets of points that takes into account the distributional information of the data at hand. The proposed k-mean algorithm incorporates a process for finding optimal spatial transformation of the sets of point sets and makes the clustering process flexible. We present an application of the proposed method to brain spike train classification, a relevant problem in neural coding, but the given k-mean procedure is also suitable in more general contexts.

Keywords: Kernel for sets of points, distances for sets of points. Affine invariant metric. Matching functions. Adaptive k-Means algorithm. Spike trains.

6.1 Introduction

The development of new algorithms for clustering analysis is of fundamental importance in several research areas, for example: Neuroscience and Bioinformatics Orhan et al. (2011); Inza et al. (2010), Image and Vision Meyer-Baese and Schmid (2014); Comaniciu and Meer (2002),

[1] The work presented in this chapter is partially included in the Proceeding of the Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications conference (Muñoz et al., 2013).

Business, Marketing and Social Network Analysis Zeng et al. (2012); Jain (2010), just to mention a few.

The clustering problem consists in the task of grouping objects in clusters in such a way that the objects that belongs to the same cluster are similar. There exist several algorithms to solve this problem: k-means, k-medoids, hierarchical procedures, methods based on density among others (see Xu et al. (2005); Jain (2010), for a review on clustering methods and algorithms).

The clustering algorithms were originally developed to deal with a set of data points, where every point represents the features of the objects to cluster. To cluster complex objects as for instance 2D images, 3D surfaces, sets of documents, sets of time series that represents different neuronal patterns, etc, a distance (or a similarity) measure between the complex objects must be used. The distance matrix among the objects at hand is the input that when is used in combination with standard classification algorithm Jacques and Preda (2013) allow us to produce the desired clusters.

In several data analysis problems such as 3D Objects Classification Arbter et al. (1990), classification of Proteomic time series Listgarten and Emili (2005) or Neural Coding Sangalli et al. (2010a) to name a few, it is also important to consider the possible misalignment of the data. In these cases, it is of fundamental importance the incorporation of an alignment (or registration) step within the cluster algorithm in order to perform a correct data analysis.

In this chapter we propose a kernel function for data points with reference to a distribution function, that is extended later to a kernel (and to a dissimilarity index) for sets of points. By combining the dissimilarity index with an alignment procedure, we develop a flexible k-means clustering method for sets of points. The proposed algorithm is suited to solve clustering problems in general contexts. In this article we focus on the problem of brain spike trains clustering, a 1D case, as a possible application of the proposed method.

The Chapter is organized as follows: In Section 6.2 we introduce kernel functions for sets of points that induce dissimilarity indices for sets of points. Section 6.3 describes the alignment/matching procedure that we use in combination with the dissimilarity proposed in Section 6.2 in order to obtain a flexible k-mean clustering method. In Section 6.4 we present synthetic and real data experiments to show the performance of the proposed method.

6.2 A Density Based Dissimilarity Index for Sets of Points

Let $\mathbb{P}$ and $\mathbb{Q}$ be two σ-additive finite probability measures (PM) defined on the same measure space $(X, \mathcal{F}, \mu)$, where X is a compact set of a real vector space, $\mathcal{F}$ is a σ-algebra of measurable subsets of X and $\mu : \mathcal{F} \to \mathbb{R}^+$ is the Lebesgue measure. Both measures are assumed to be absolutely continuous w.r.t. μ and satisfy the three Kolmogorov axioms. By Radon-Nikodym theorem, there exists a measurable function $f_{\mathbb{P}} : X \to \mathbb{R}^+$ (the density function) such that $\mathbb{P}(A) = \int_A f_{\mathbb{P}} d\mu$, and $f_{\mathbb{P}} = \frac{d\mathbb{P}}{d\mu}$ is the Radon-Nikodym derivative (the same applies for $\mathbb{Q}$, with density function given by $f_{\mathbb{Q}} = \frac{d\mathbb{Q}}{d\mu}$).

Given two random samples $A = S^n_{\mathbb{P}} = \{x_i\}_{i=1}^n$ and $B = S^m_{\mathbb{Q}} = \{y_j\}_{j=1}^m$, both generated from the density functions $f_{\mathbb{P}}$ and $f_{\mathbb{Q}}$ respectively and defined on the same measure space. Define $r_A = \min d(x_l, x_s)$, where $x_l, x_s \in A$. Then r_A gives the minimum resolution level for the set A: If a point $z \in X$ is located at a distance smaller than r_A from a point $x \in A$ then, taken $\mathbb{P}$ as reference measure, it is impossible to differentiate z from x. That is, it is not possible to reject the hypothesis that z is generated from $\mathbb{P}$, given that z is closer to x than any other point from the same distribution. This suggest the following definition Muñoz et al. (2013):

Definition 6.1. Indistinguishability with respect to a distribution. Let $x \in A$, where A denotes a set of points generated from the probability measure $\mathbb{P}$, and $y \in X$. We say that y is *indistinguishable* from x with respect to the measure $\mathbb{P}$ in the set A when $d(x, y) \leq r_A = \min d(x_l, x_s)$, where $x_l, x_s \in A$. We will denote this relationship as: $y \overset{A(\mathbb{P})}{=} x$.

In this Chapter we start by assuming constant radius parameters for the sets of points A and B, in this way we are implicitly assuming that $\mathbb{P}$ and $\mathbb{Q}$ are two uniform distributions. This assumption can be lifted by simply doing the radius parameter dependent on the points $x \in A$ or $y \in B$ that we are analysing. For example we can define $r_A(x) = d_{A,k}(x)$, where $d_{A,k}(x)$ is the distance from the point x to its k-nearest neighbour in the data set A. In this way the resolution level of the data set A increase in the regions with high density and decreases in the regions with low density. In what follows we maintain fixed radius parameters r_A and r_B in order to avoid a systematic abuse of notation, but when the underlying distributions $\mathbb{P}$ and $\mathbb{Q}$ of the respective sets points A and B are not uniform, then the resolutions levels $r_A(x)$ and $r_B(y)$ should be understood as dependants of the points $x \in A$ and $y \in B$.

Given the sets of points $A = S^n_{\mathbb{P}}$ and $B = S^m_{\mathbb{Q}}$, we want to build kernel functions $K : X \times X \to [0, 1]$, such that $K(x, y) = 1$ when $y \overset{A(\mathbb{P})}{=} x$ or $x \overset{B(\mathbb{Q})}{=} y$, and $K(x, y) = 0$ if $y \overset{A(\mathbb{P})}{\neq} x$ and

$x \overset{B(\mathbb{Q})}{\neq} y$. For this purpose we can consider smooth indicator functions, for example:

Definition 6.2. Smooth indicator functions. Let $r > 0$ and $\gamma > 0$, define a family of smooth indicator functions with center in x as:

$$f_{x,r,\gamma}(y) = \begin{cases} e^{-\frac{1}{(\|x-y\|^{\gamma}-r^{\gamma})^2}+\frac{1}{r^{2\gamma}}} & \text{if } \|x-y\| \leq r \\ 0 & otherwise. \end{cases} \tag{6.1}$$

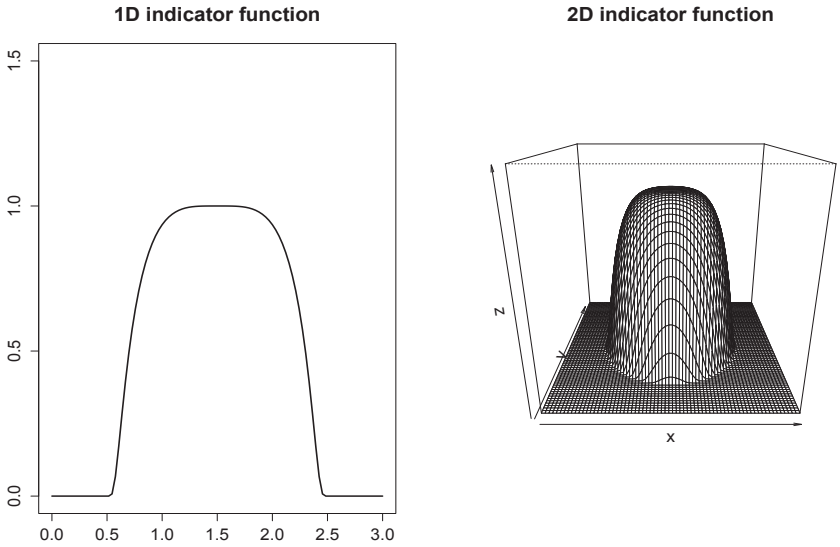


Figure 6.1: Smooth indicator functions. (a) 1D case. (b) 2D case.

In Figure 6.1 we give two examples of smooth indicator functions in dimension 1 and 2. The smooth function $f_{x,r,\gamma}(y)$ act as a bump function with center in the coordinate point given by x: $f_{x,r,\gamma}(y) \approx 1$ for $y \in B_r(x)$ (where $B_r(x)$ denotes the closed ball of radius r centered at the point x), and $f_{x,r,\gamma}(y)$ decays to zero out of $B_r(x)$, depending on the shape parameter γ.

Using the similarity relationship given in Definition 6.1, a distributional-indicator kernel function $K_{A,B} : X \times X \to [0,1]$ can be obtained:

Definition 6.3. Distributional indicator kernel. Let $A = S_{\mathbb{P}}^n$ and $B = S_{\mathbb{Q}}^m$ be two sets of points iid drawn from the probability measures $\mathbb{P}$ and $\mathbb{Q}$ respectively, define $K_{A,B} : X \times X \to [0,1]$

by:

$$K_{A,B}(x,y) = f_{x,r_A,\gamma}(y) + f_{y,r_B,\gamma}(x) - f_{x,r_A,\gamma}(y) f_{y,r_B,\gamma}(x), \tag{6.2}$$

where $r_A = \min d(x_l, x_s)$, with $x_l, x_s \in A$, $r_B = \min d(y_l, y_s)$, with $y_l, y_s \in B$ and γ it is a shape parameter.

Now, if $d(x,y) > r_A$ and $d(x,y) > r_B$ (see Figure 6.2-A) then $K_{A,B}(x,y) = 0$: $x \in A \backslash B$ w.r.t. $\mathbb{Q}$ and $y \in B \backslash A$ w.r.t. $\mathbb{P}$. If $d(x,y) > r_A$ but $d(x,y) < r_B$, then $y \in B \backslash A$ w.r.t. $\mathbb{P}$, but $x \overset{B(\mathbb{Q})}{=} y$ at radius r_B and $K_{A,B}(x,y) = 1$. If $d(x,y) < r_A$ but $d(x,y) > r_B$, then $x \in A \backslash B$ w.r.t. $\mathbb{Q}$, but $y \overset{A(\mathbb{P})}{=} x$ at radius r_A and $K_{A,B}(x,y) = 1$ (see Figure 6.2-B). Finally, if $d(x,y) < r_A$ and $d(x,y) < r_B$, then $K_{A,B}(x,y) = 1$ and $y \overset{A(\mathbb{P})}{=} x$ at radius r_A and $x \overset{B(\mathbb{Q})}{=} y$ at radius r_B (Figure 6.2-C).

The introduction of the smooth indicator functions paved the way to define a Kernel function for sets of points (in Muñoz et al. (2013) we give a more general definition of a Kernel for data sets).

Definition 6.4. A Kernel for sets of points. Let $A = S^n_{\mathbb{P}}$ and $B = S^m_{\mathbb{Q}}$ be two sets of points generated from the probability measures $\mathbb{P}$ and $\mathbb{Q}$ respectively, we consider kernels $K : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{N}$, where $\mathcal{P}(X)$ denotes the power set of X:

$$K(A,B) = \sum_{x \in A} \sum_{y \in B} K_{A,B}(x,y). \tag{6.3}$$

The kernel $K(A,B)$ is a measure for $A \cap B$ by counting, using as equality operators $\overset{A(\mathbb{P})}{=}$ and $\overset{B(\mathbb{Q})}{=}$, the points in common between the sets A and B: $\mu_{K_{A,B}}(A \cap B) = K(A,B)$.

Given the identity $A \cup B = \overbrace{(A-B) \cup (B-A)}^{A \Delta B} \cup (A \cap B)$, we will define $\mu_{K_{A,B}}(A \cup B) = N$, where $N = n + m = \#(A \cup B)$, is the counting measure of the set $A \cup B$. Therefore $\mu_{K_{A,B}}(A \Delta B) = N - \mu_{K_{A,B}}(A \cap B)$, and we can take this expression (dividing by N) as a definition for the distance between the sets A and B.

Kernel functions induce distance measures. By using Equation 2.5 given in Chapter 2:

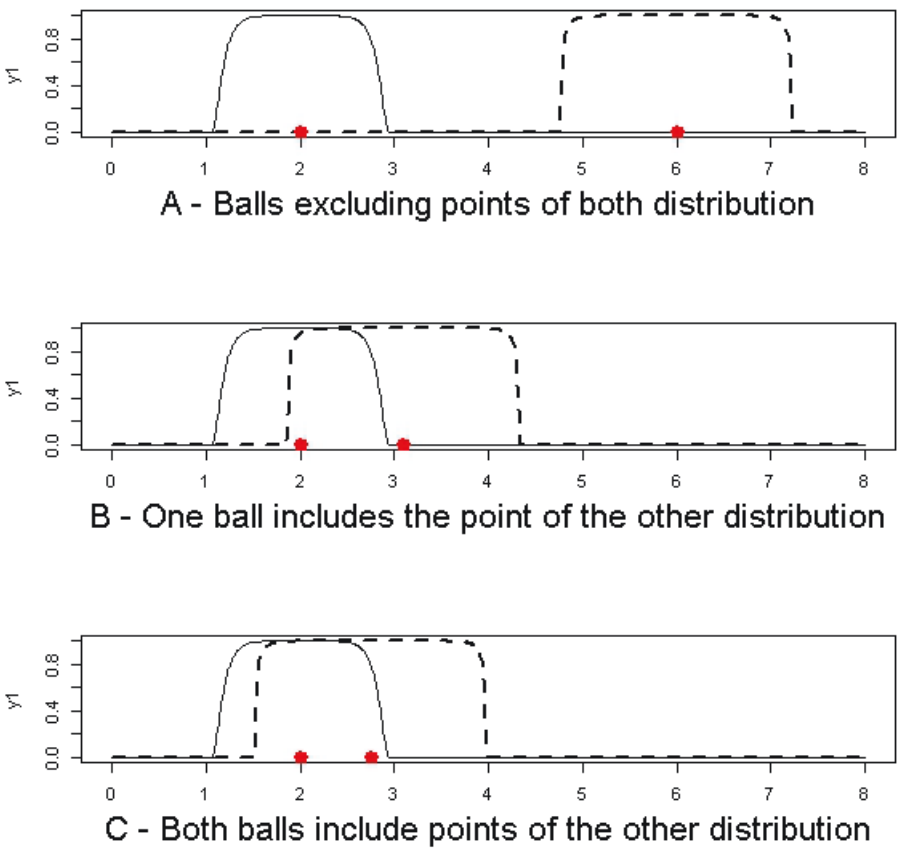


Figure 6.2: Illustration of the $\overset{A(\mathbb{P})}{=}$ and $\overset{B(\mathbb{Q})}{=}$ relationship using smooth indicator functions.

$$\begin{aligned} D_K^2(A,B) &= K(A,A) + K(B,B) - 2K(A,B), \\ &= n + m - 2K(A,B), \\ &= N - 2\sum_{x\in A}\sum_{y\in B} K_{A,B}(x,y). \end{aligned} \tag{6.4}$$

Note that this kernel distance is defined in terms of its square. We propose a slightly different distance induced by the kernel function of Definition 6.5 in order to maintain the spirit of the work done in Chapter 5. Given the kernel for sets of points in Definition 6.4, define a dissimilarity for sets of points in the following way:

Definition 6.5. Dissimilarity between sets of points. Given two random samples $A = S_{\mathbb{P}}^n$ and

$B = S_{\mathbb{Q}}^m$, generated from the probability measures $\mathbb{P}$ and $\mathbb{Q}$ respectively, we define the kernels dissimilarity for A and B in $\mathcal{P}(X)$ by:

$$d_K(A, B) = 1 - \frac{K(A, B)}{N}, \tag{6.5}$$

where $N = n_A + n_B = \#(A \cup B)$ represent the measure of the set $A \cup B$.

It is straightforward to check that $d_K(A, B)$ is a semi-metric (using the equality operators $y \overset{A(\mathbb{P})}{=} x$ or $y \overset{B(\mathbb{Q})}{=} x$ where it corresponds). When the size of the sets A and B increases, then: $\mu_{K_{A,B}}(A \cap B) \overset{n,m\to\infty}{\to} \mu(A \cap B)$ and $\mu_{K_{A,B}}(A \cup B) \overset{n,m\to\infty}{\to} \mu(A \cup B)$, therefore $\lim_{n,m\to\infty} d_K(A, B) = 1 - \frac{\mu(A\cap B)}{\mu(A\cup B)}$, that is the Jaccard Jaccard (1912) dissimilarity index for sets of points.

An important property of the proposed dissimilarity index is that is invariant to affine transformations. Let $\mathcal{T}$ be a class of translation, dilation and rotation transformations and $h \in \mathcal{T} : \mathcal{P}(X) \to \mathcal{P}(X)$ be an affine map, then: $d_K(A, B) = d'K(h \circ A, h \circ B)$ (see the appendix D for a formal proof). In next section we introduce the alignment procedure and the implementation of the adaptive k-mean clustering method, both based in the use of d_K.

We want to exemplify how the proposed dissimilarity works with a synthetic example. To this end, we generate two iid samples $S_{\mathbb{P}}^n = A$ and $S_{\mathbb{Q}} = B$ with $m = n = 500$, drawn from the bi-dimensional uniform density function inside a ball with center in zero and radius $r = 1$ ($f_{\mathbb{P}} = U[B_{r=1}(0)]$), and a bi-dimensional Normal distribution function with parameters $\boldsymbol{\mu} = (0, 0)$ and $\boldsymbol{\Sigma} = \mathbf{I}_2$ ($f_{\mathbb{Q}} = N((0, 0), \mathbf{I}_2)$), respectively. We generate new sets A' and B' by displacing all the points that belongs to the sets A and B a constant distance in the same direction. In Figure 6.3 we represent the sets A, A', B and B'.

Therefore by using the (semi)metric given in Definition 6.5, then the distance between sets B and B' should be greater compared with the distance between the sets A and A'. This is because as the intersection between the last two sets seems to be bigger: The sets A and A' are more "similar" in relation with the sets B and B'. To verify the similarity relations between the sets, we compute the distance matrix between the proposed sets A, A', B and B', according to Definition 6.5:

We can see that the proposed metric represent well the similarity relation between the hybrid data sets: $d_K(B, B') = 0.864 \geq d_K(A, A') = 0.792$, as we mention previously. Also notice

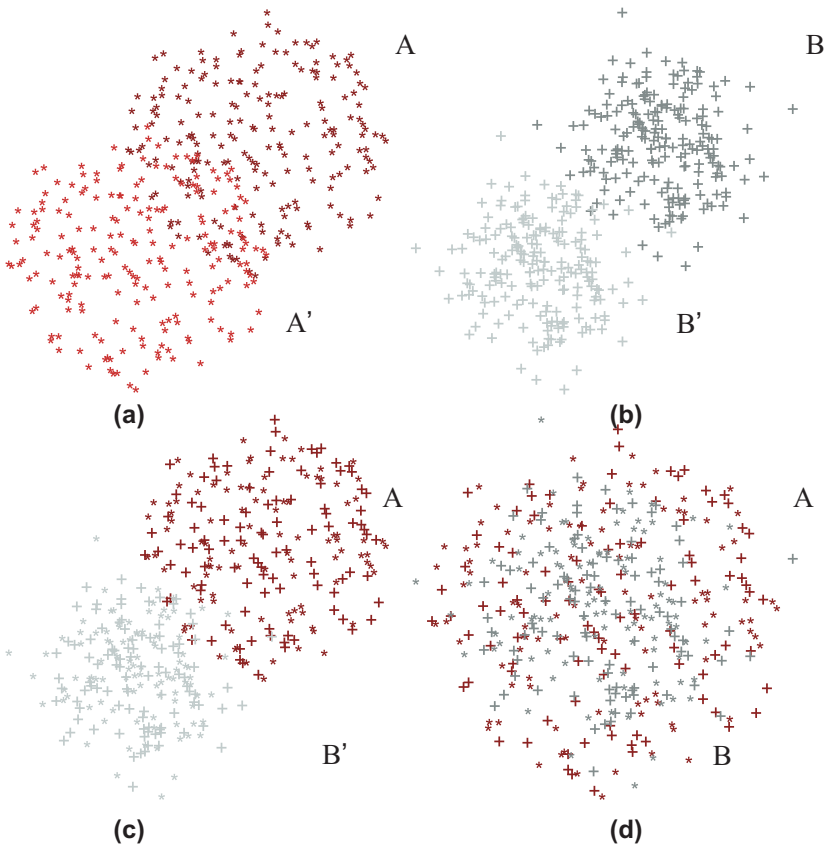


Figure 6.3: (a) A and A', (b) B and B', (c) A and B', (d) A and B

that

$$d_K(B, B') \geq d_K(A, B') \approx d_K(A', B) \geq d_K(A, A') \geq d_K(A, B) \approx d_K(A', B'),$$

as can be seen in Figure 6.3.

6.3 Registration Method for Sets of Points

We present in this section the details of the alignment method for sets of points. The work done in this section is based on Sangalli et al. (2010b,a).

6.3.1 Matching functions for sets of points

Given two sets of points $A = S^n_{\mathbb{P}} = \{x_i\}_{i=1}^n$ and $B = S^m_{\mathbb{Q}} = \{y_j\}_{j=1}^m$ generated from the probability measures $\mathbb{P}$ and $\mathbb{Q}$ respectively, the registration procedure is based on the work done by

Table 6.1: Matrix of distances between data sets: A, A', B and B'.

	A	A'	B	B'
A	0	0.792	0.104	0.821
A'		0	0.823	0.104
B			0	0.864
B'				0

Sangalli et al. (2010b,a) where the minimization of a dissimilarity index between the sets A and B is carried out to align the sets of points. This procedure is carried out by using a matching function $h : \mathcal{P}(X) \rightarrow \mathcal{P}(X)$, such that the two sets of points are the most similar among all the possible transformations $h \in \mathcal{T}$, where $\mathcal{T}$ is a family of possible matching functions. It is thus necessary to define the class $\mathcal{T}$ of admissible matching (or alignment) functions h, that combined with the dissimilarity measure d_K proposed in Equation (6.5), help us to find a suitable alignment of the set of points A with respect to the set of points B. That is: find $h^* \in \mathcal{T}$ such that $d_K(h \circ A, B)$ is minimized.

We are particularly interested in the class $\mathcal{T}$ of affine transformations. Therefore the class $\mathcal{T}$ is constrained to be $\mathcal{T} = \{h : h \circ A = \{h \circ x_i\}_{i=1}^n = \{t + sRx_i\}_{i=1}^n\}$ for all $A \in \mathcal{P}(X)$, where $t \in \mathbb{R}^d$ is a translation constant, $s \in \mathbb{R}^+$ is a scaling constant and R is a rotation (unitary transformation) matrix.

The dissimilarity index d_K and the class $\mathcal{T}$ of matching function satisfy the minimal requirements for the well posedness of the alignment problem as it is stated in Sangalli et al. (2010b,a):

- The dissimilarity index is bounded: $d_K(A, B) = 0$ when $A = B$ and increases (up to 1) when the similarity between A and B decreases.
- For all $A, B, C \in \mathcal{P}(X)$, the dissimilarity index d_K is symmetric: $d_K(A, B) = d_K(B, A)$, reflexive: $d_k(A, A) = 0$ and transitive: If $d_k(A, B) = 0$ and $d_K(B, C) = 0$ then $d_K(A, C) = 0$.
- The class of matching functions $\mathcal{T}$ constitutes a convex vector space and has a group structure with respect to the composition function (denoted by $\circ$), that is: If $h_1 \in \mathcal{T}$ and $h_2 \in \mathcal{T}$, then $h_1 \circ h_2 \in \mathcal{T}$.
- The choices of the dissimilarity index d_K and the class of matching functions $\mathcal{T}$ are consistent in the sense that, if two sets of points A and B are simultaneously aligned along

the same matching function $h \in \mathcal{T}$, the dissimilarity does not change:

$$d_K(A, B) = d_K(h \circ A, h \circ B) \quad \forall h \ \in \mathcal{T}.$$

The last requirement guarantees that it is not possible to obtain a fictitious decrement of the dissimilarity index between the sets of points by simply transforming them simultaneously in $h \circ A$ and $h \circ B$. The dissimilarity index d_K that we propose in Equation (6.5) and the class $\mathcal{T}$ of affine transformations fulfils these conditions.

6.3.2 An adaptive K-mean clustering algorithm for sets of points

Next we considers the simultaneous problem of clustering and alignment of sets of points following the work done in Sangalli et al. (2010b,a). To characterize in details this problem, consider first a collection of k-templates or centers: $\underline{\boldsymbol{\varphi}} = \{\varphi_1, \ldots, \varphi_k\}$ and a collection of N sets of points: $\mathbf{A} = \{A_1, \ldots, A_N\}$ to be assigned to these templates. For each template φ_j in $\underline{\boldsymbol{\varphi}}$ and a class $\mathcal{T}$ of matching functions, define the domain of attraction:

$$\triangle_j \left(\underline{\boldsymbol{\varphi}}\right) = \{A \in \mathcal{P}\left(X\right) : \inf_{h \in \mathcal{T}} d_K(\varphi_j, h \circ A) \leq \inf_{h \in \mathcal{T}} d_K(\varphi_r, h \circ A), \forall r \neq j\}, \ \forall j = 1, \ldots, k.$$

Define the labelling function $\lambda(\underline{\boldsymbol{\varphi}}, A)$, such that when $\lambda(\underline{\boldsymbol{\varphi}}, A) = j$ then the set of points A must be aligned and clustered to the template set $\varphi_j \in \underline{\boldsymbol{\varphi}}$, because the dissimilarity index d_K obtained by aligning A to φ_j is at most equal or lower than the dissimilarity index obtained by aligning A to any other template $\varphi_r \in \underline{\boldsymbol{\varphi}}$, with $r \neq j$.

When the k templates $\underline{\boldsymbol{\varphi}} = \{\varphi_1, \ldots, \varphi_k\}$ were known, then the clustering and aligning procedure of the N sets of points $\{A_1, \ldots, A_N\}$ with respect to $\underline{\boldsymbol{\varphi}}$ would simply mean to assign A_i to the cluster $\lambda(\underline{\boldsymbol{\varphi}}, A_i)$ and align A_i to the corresponding template $\varphi_{\lambda(\underline{\boldsymbol{\varphi}}, A_i)}$, for $i = 1, \ldots, N$.

In the most general case, we do not have information regarding the template set $\underline{\boldsymbol{\varphi}}$. In this case we need to solve two simultaneous problems:

(i) Find the set of centers $\underline{\boldsymbol{\varphi}} = \{\varphi_1, \ldots, \varphi_k\}$, such that:

$$\sum_{i=1}^{N} \inf_{h \in \mathcal{T}} d_K(\varphi_{\lambda(\underline{\boldsymbol{\varphi}}, A_i)}, h \circ A_i) \leq \sum_{i=1}^{N} \inf_{h \in \mathcal{T}} d_K(\psi_{\lambda(\underline{\boldsymbol{\psi}}, A_i)}, h \circ A_i),$$

for any other set of templates $\underline{\boldsymbol{\psi}}$.

(ii) Cluster and align the collection of N sets of points $A_1, \ldots, A_N$, to the set of k templates $\underline{\boldsymbol{\varphi}} = \{\varphi_1, \ldots, \varphi_k\}$.

To solve these two simultaneous problems, we develop a k-means algorithm that iteratively alternates between an update step, an assignment step and a normalization step. Let $\underline{\boldsymbol{\varphi}}^{[q-1]} = \{\varphi_1^{[q-1]}, \ldots, \varphi_k^{[q-1]}\}$ be the set of k templates after the iteration $[q-1]$, and let $\boldsymbol{A}^{[q-1]} = \{A_1^{[q-1]}, \ldots, A_N^{[q-1]}\}$ be the N sets of points aligned and clustered to the centers $\underline{\boldsymbol{\varphi}}^{[q-1]}$ at the iteration $q-1$. At the q^{th} iteration, the algorithm performs the following steps:

- **Update step**: The centers $\varphi_j^{[q]}$ for $j = 1, \ldots, k$ are re-estimated. Ideally the estimation of the new centers should be computed as:

$$\varphi_j^{[q]} = \min_{\varphi \in \mathcal{P}(X)} \sum_{\{i:\lambda(\underline{\boldsymbol{\varphi}}^{[q-1]}, A_i^{[q-1]})=j\}} d_K(\varphi, A_i^{[q-1]}) \ \ \forall j = 1, \ldots, k.$$

  Unfortunately this is not an easy solvable problem, therefore we approximate the solution using the medoid. Denote by $C_l^{[q-1]}$ to be the set of sets of points assigned to the cluster labeled as l in the $[q-1]$ iteration, then:

$$\varphi_j^{[q]} = \min_{\varphi \in C_j^{[q-1]}} \sum_{\{i:\lambda(\underline{\boldsymbol{\varphi}}^{[q-1]}, A_i^{[q-1]})=j\}} d_K(\varphi, A_i^{[q-1]}) \quad \text{for } j = 1, \ldots, k.$$

- **Assignment step**: The sets of points are clustered and aligned with respect the centers obtained in the update step. For $i = 1, \ldots, N$, the i^{th} set of points $A_i^{[q-1]}$ is aligned to $\varphi_{\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, A_i^{[q-1]})}$. The aligned set of points $\widetilde{A}_i^{[q]} = A_i^{[q-1]} \circ h_i^{[q]}$ is assigned to the cluster $\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \widetilde{A}_i^{[q]})$.

- **Normalization step**: The sets of points that belongs to the same cluster are normalized by using the (inverse) average matching function:

$$\overline{h}_l^{[q]} = \frac{1}{N_l^{[q]}} \sum_{i:\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \widetilde{A}^{[q]})=j} h_i^{[q]} \qquad \text{for } l = 1, \ldots, k,$$

  where $N_l^{[q]}$ stands for the number of sets of points that belongs to the cluster l at the iteration q. The average represents a rescaled composition of the similarity transformations

made on every cluster. Then we obtain that for $A_i^{[q]} \in C_l^{[q]}$:

$$A_i^{[q]} = \left(\overline{h}_l^{[q]}\right)^{-1} \circ \widetilde{A}_i^{[q]} = \left(\overline{h}_l^{[q]}\right)^{-1} \circ h_i^{[q]} \circ A_i^{[q-1]}.$$

With the implementation of the proposed procedure we solve, after a number of iterations, the problem of clustering sets of points. In the next section we present an application of the proposed method to the study of spike trend paths. Other applications can be considered, for example, clustering 2D images or 3D shapes represented as sets of points, the clustering of time series of DNA microarray information (in this case every time series is a 1D set of points), among other possible uses.

6.4 Experimental Section

In this section we present synthetic and real data experiments in order to demonstrate the ability of the proposed algorithm to produce clusters of sets of points. We are particularly interested in the problem of clustering brain spike trains because in the process of recording the data it is normal to observe registration problems Stevenson and Kording (2011); Buzsáki (2004); Brown et al. (2004). In this context, the use of a flexible k-means method that includes a registration step is adequate Srivastava et al. (2011); Sangalli et al. (2010a).

In Section 6.4.1 we test the adequacy of the alignment method and the performance of the clustering algorithm in four different synthetic scenarios. In Section 6.4.2 we present three real data experiments to demonstrate that proposed method is able to work well in real world cases.

6.4.1 Artificial Experiments

Classifying simulated spike brain paths:

In the first experiment we model the firing activity of a neuron as a one dimensional inhomogeneous Poisson process. We propose 4-scenarios to exemplify different situations: In the first scenario, denoted as A, there are 2 clusters of neurons (C_1 and C_2, respectively) and there is no recording problem in the data. We simulate 40 instances of the spike trains as realizations of two different Poisson processes with the following intensity rates:

$$\begin{aligned}
\rho(t)_i^{[C_1]} &= (1+\varepsilon_{1i})\sin(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right) \quad i=1,\ldots,20,\\
\rho(t)_i^{[C_2]} &= -(1+\varepsilon_{1i})\cos(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right) \quad i=21,\ldots,40,
\end{aligned}$$

where $t \in [0, \frac{3}{2}\pi]$, the set of random coefficients $\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4\}$ are independently and normally distributed with means: $\mu_{\varepsilon_1} = 2$, $\mu_{\varepsilon_2} = \mu_{\varepsilon_3} = \frac{1}{2}$, $\mu_{\varepsilon_4} = 1$ and variances $\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} = \sigma^2_{\varepsilon_3} = \sigma^2_{\varepsilon_4} = 0.05$. We maintain the specification in the distribution of these coefficients in all the considered scenarios.

In the scenario B, the cluster C_1 of spike trains incorporates amplitude variability in the following way:

$$\begin{aligned}
\rho(t)_i^{[C_1]} &= (1+\varepsilon_{1i})\sin(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right) \quad i=1,\ldots,10,\\
\rho(t)_i^{[C_1]} &= \frac{3}{4}\left((1+\varepsilon_{1i})\sin(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right)\right) \quad i=11,\ldots,20,\\
\rho(t)_i^{[C_2]} &= -(1+\varepsilon_{1i})\cos(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right) \quad i=21,\ldots,40,
\end{aligned}$$

In scenario C, the cluster C_1 incorporates phase variability:

$$\begin{aligned}
\rho(t)_i^{[C_1]} &= (1+\varepsilon_{1i})\sin(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right) \quad i=1,\ldots,10,\\
\rho(t)_i^{[C_1]} &= (1+\varepsilon_{1i})\sin(\varepsilon_{3i}+\varepsilon_{4i}(t-\frac{1}{2}))+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{(t-\frac{1}{2})^2}{2\pi}\right) \quad i=11,\ldots,20,\\
\rho(t)_i^{[C_2]} &= -(1+\varepsilon_{1i})\cos(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right) \quad i=21,\ldots,40,
\end{aligned}$$

Finally, scenario D is a combination of the scenarios B and C. The intensity rates are modeled in the following way:

$$
\begin{aligned}
\rho(t)_i^{[\mathrm{C}_1]} &= (1+\varepsilon_{1i})\sin(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right) \qquad i=1,\ldots,10,\\
\rho(t)_i^{[\mathrm{C}_1]} &= \frac{3}{4}\left((1+\varepsilon_{1i})\sin(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right)\right) \; i=11,\ldots,20,\\
\rho(t)_i^{[\mathrm{C}_1]} &= (1+\varepsilon_{1i})\sin(\varepsilon_{3i}+\varepsilon_{4i}(t-\frac{1}{2}))+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{(t-\frac{1}{2})^2}{2\pi}\right) \; i=21,\ldots,30,\\
\rho(t)_i^{[\mathrm{C}_2]} &= -(1+\varepsilon_{1i})\cos(\varepsilon_{3i}+\varepsilon_{4i}t)+(1+\varepsilon_{2i})\sin\left(\varepsilon_{3i}+\varepsilon_{4i}\frac{t^2}{2\pi}\right) \qquad i=31,\ldots,40,
\end{aligned}
$$

In Figure 6.4 we shown the intensity rate functions in all the considered scenarios.

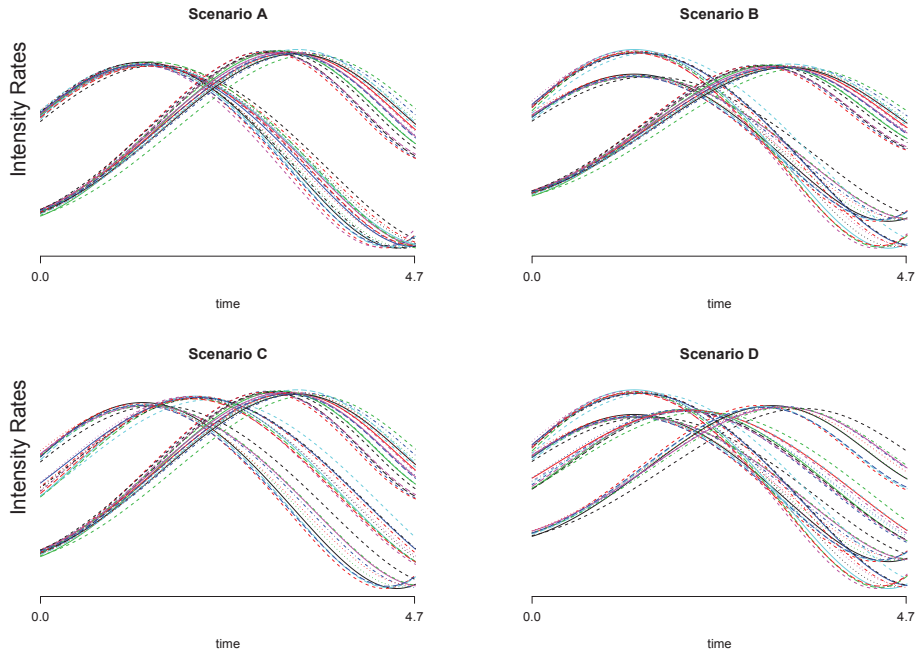


Figure 6.4: Intensity rates for the simulated faring activity (scenarios A to D).

The simulated spike train are generated by using the respective intensity rate functions created for the scenarios A to D. Each spike train can be represented as a discrete time series: $x(t) = 1$ if we observe a spike in the neuron x at time t and $x(t) = 0$ otherwise. In Figure 6.5, we present the raster-plots that contains the simulated instances of the firing activity of 40 neurons in each scenario.

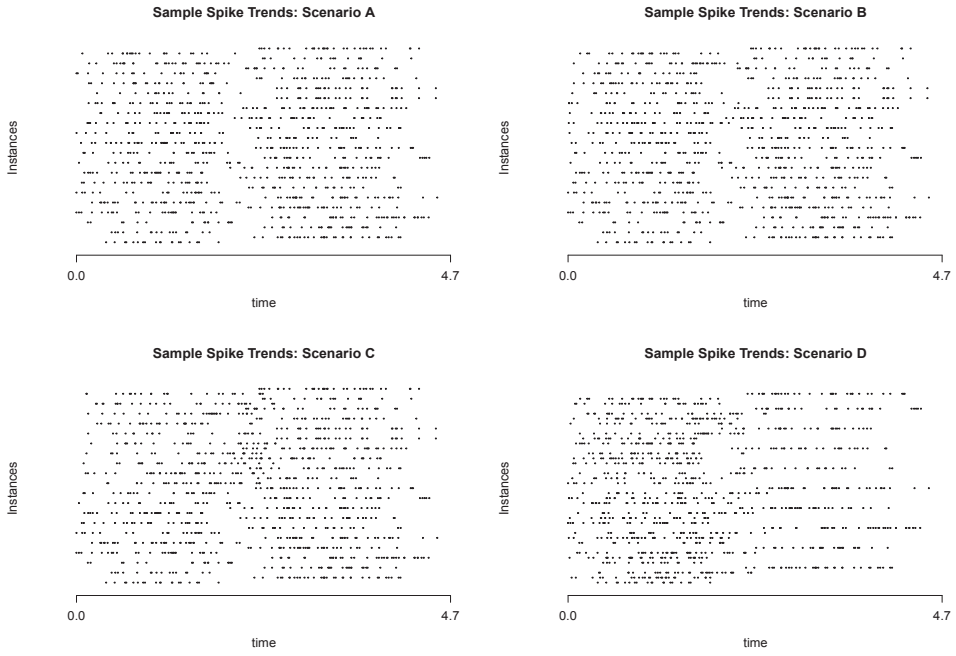


Figure 6.5: 40 instances of simulated spike trains: Scenario A to D.

In order to test the performance of the proposed clustering method, we introduce 4 different families of matching functions. Let $x(t)$ be a discrete time series representing the spike activity of a neuron, then we consider the following family of matching functions:

$$\begin{aligned}
\mathcal{T}_{\text{identity}} &= \{h : h \circ x(t) = x(t)\}, \\
\mathcal{T}_{\text{translation}} &= \{h : h \circ x(t) = x(\alpha + t), \forall \alpha \in \mathbb{R}\}, \\
\mathcal{T}_{\text{dilation}} &= \{h : h \circ x(t) = x(\beta t), \forall \beta \in \mathbb{R}^+\}, \\
\mathcal{T}_{\text{affine}} &= \{h : h \circ x(t) = x(\alpha + \beta t), \forall \alpha \in \mathbb{R} \text{ and } \forall \beta \in \mathbb{R}^+\}.
\end{aligned}$$

In this experiment the firing activity of a neuron is considered as a set of points. We choose the radii parameters (r_A in Section 6.2) as the median time between spikes for every simulated neuron, that is the sample median interarrival-time of the simulated inhomogeneous Poisson process. The implementation of the clustering procedure follows the details covered in Section 6.3, where the proposed algorithm is described in details.

Determining the optimal number of clusters, a parameter usually denoted as k, is an important problem in data clustering. The k-means clustering algorithms needs the parameter

k as an input in order to group the data into k groups. One of the most popular criteria to determine the number of clusters is an heuristic approach known as the elbow method. This method selects the parameter k as the minimum number of clusters such that the change in the within-groups variability[1] (WGV_k) at this cluster level is no longer significative. We compute the WGV_k by using the dissimilarity index d_K as follows:

$$WGV_k = \frac{\sum_{l=1}^{k} \sum_{j=1}^{N_l} d_K(x_j^{C_l}, \varphi_l)}{k-1},$$

where k denotes the number of clusters, $x_j^{C_l}$ is the j^{th} element of the cluster C_l, φ_l is the center of the cluster C_l and N_l stands for the number of elements in the cluster C_l (the size of C_l).

In Figure 6.6 we present the elbow-plots obtained after the implementation of the proposed clustering procedure in the four scenarios and for all the considered families of matching functions. As can be seen in Figure 6.6, the within-groups variability tends to decrease in all the scenarios (and for all the considered families of matching functions) when the number of clusters k increases. As we can foresee, the decrease in the unexplained variability is maximized when we allow affine transformation in the data.

In scenario A, all the considered families of matching functions give us the same clustering configuration, as is expected, because of the configuration of this scenario. In scenario B, we observe the same result as in A, but the within-groups variability is smaller when we allow affine transformations.

In scenario C, if we do not use the $\mathcal{T}_{\text{translation}}$, $\mathcal{T}_{\text{dilation}}$ or $\mathcal{T}_{\text{affine}}$ families of matching functions to align the input data, then the use of the elbow criterion induces to an inaccurate cluster configuration. It is not clear if 2, 3 or 4 clusters are necessary.

Finally, in scenario D, the use of the $\mathcal{T}_{\text{identity}}$ family of matching functions leads us to a wrong cluster solution: 3 clusters instead of 2. This situation is avoided when we allow affine transformations in the data. The performance of the clustering algorithm (after setting $k = 2$) it is outstanding: in the case of $\mathcal{T}_{\text{translation}}$, $\mathcal{T}_{\text{dilation}}$ and $\mathcal{T}_{\text{affine}}$ the misclassification error rates is 0% in all the considered scenarios. In the case of $\mathcal{T}_{\text{identity}}$ (no alignment) the misclassification error rate is 0% for scenarios A and B and 5% for scenarios C and D.

[1] The unexplained variability in the data.

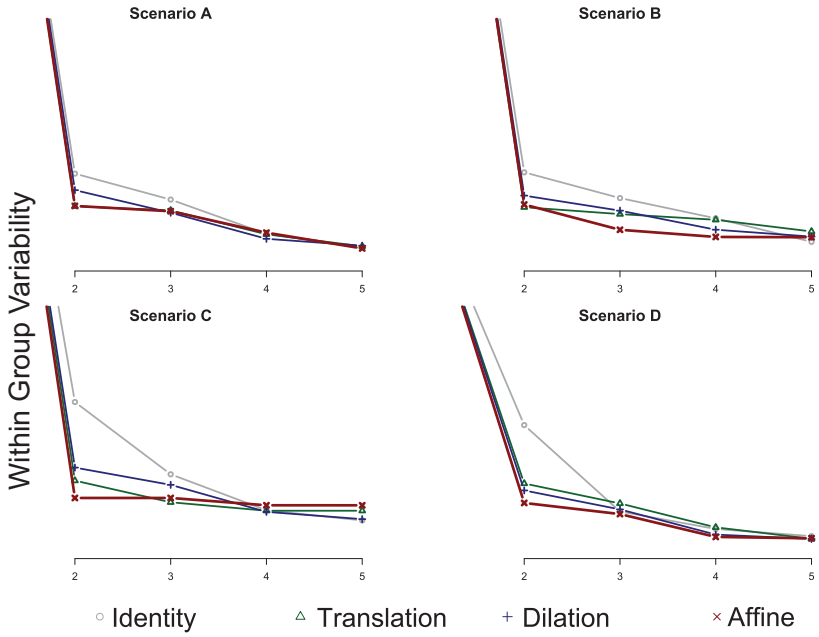


Figure 6.6: WGV_k (vertical axes) and Number of clusters (horizontal axes) for different matching functions: Scenarios A to D.

We conclude that the use of the proposed k-means clustering method improves the classification results when we deal with misaligned data, in particular when we deal with time series of brain spike train. Next we present three more examples, in this case using real data sets.

6.4.2 Real data experiments

fMCI data and Mice CA3 Hippocampus cells:

For this experiment, the data regarding spike trains are obtained by a functional imaging technique consisting in a multicell loading of calcium fluorophores (fMCI) (more details about the data extraction can bee seen in Ikegaya (2004)) on hippocampus CA3 pyramidal cells of mice. The mice carry out the task of go over a crossroad for food under different ambient conditions. We consider two clusters of neural paths according to the stimulus condition: A first cluster for cases when the ambient temperature is in the range of $[28^o, 32^o]$ Celsius degrees and a second cluster for the cases when the ambient temperature is in the range $[36^o, 40^o]$ Celsius degrees.

The firing paths of the two clusters of neurons are presented in Figure 6.7. As can be seen, in Figure 6.7-left, it is not easy to notice the clusters by simple visual inspection. In Figure

6.7-right, we reproduce again the neural data by using different colors for the two considered clusters.

In order to analyse the data, we run the proposed k-mean algorithm for $k \in \{2, 3, 4, 5, 6\}$ and by using the 4 families of matching functions introduced in the synthetic experiment. In Figure 6.8, we show the elbow plot. As can be seen, the correct choice of $k = 2$ clusters is only achieved when we use the $\mathcal{T}_{\text{affine}}$ family of matching functions to align the spike trains. We can also see that when we use the $\mathcal{T}_{\text{affine}}$ family of matching functions the unexplained variability between the clusters is minimized.

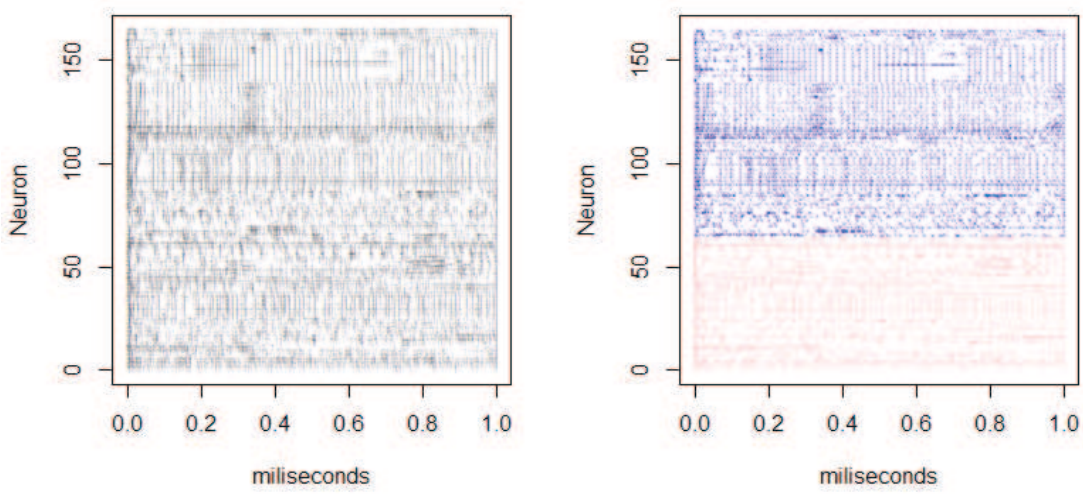


Figure 6.7: Spikes brain paths: fMCI data. The colors on the right show the two different clusters of firing paths.

In Table 6.2 we show the missclassification rates (for $k = 2$) for the 4 families of matching functions. As can be seen, the minimum error rate is acquired only for the affine family of matching functions.

Other standard alternative approaches to solve the clustering problem, for example by representing each time series of spike trains as a point in $\mathbb{R}^2$ where the first coordinate represent the mean spike time and the second coordinate represent the standard deviation of the spike time, gives poor classification results. By using the standard k-mean algorithm a missclassification error rate of 12.37% is obtained and using a hierarchical clustering procedure the error

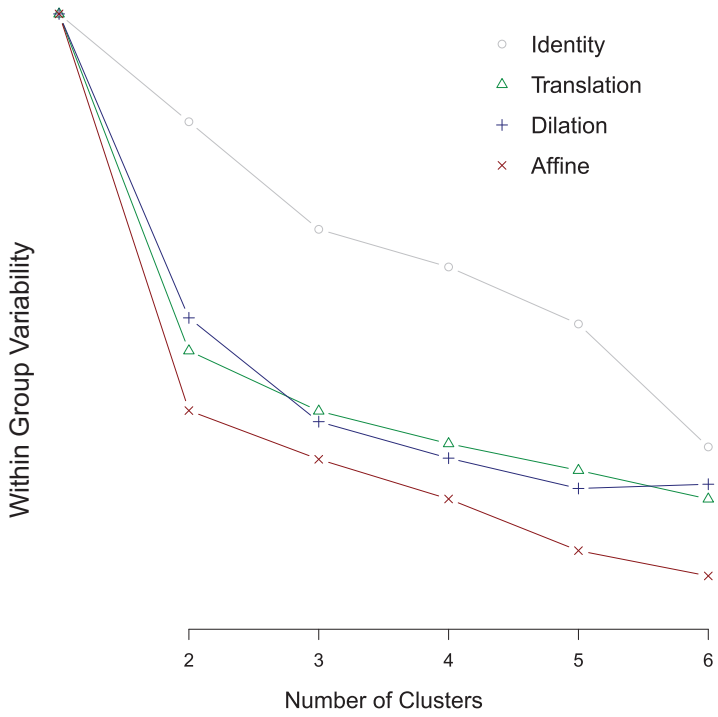


Figure 6.8: Normalized scree-plot for different matching functions.

rate achieved is 19.58%.

Table 6.2: Classification performance for different families of matching functions.

matching Family:	Identity	Translation	Dilation	Affine
Error Rate ($k = 2$)	5.15%	3.10%	3.10%	$\mathbf{2.06}\%$

With this real data example, we are able demonstrate the ability of the proposed method to adequately select the correct number of clusters in the data and to minimize the classification error rate.

Visual grating task on a monkey:

In this experiment a monkey is sitting with the head fixed and a grating pattern that moves in different direction was presented on a screen, blank and stimulus pattern were switched alternatively every 2 seconds (see Figure 6.9). The data were recorded by using an Electrocorticography (ECoG) , the details of the recording process can be seen in Laboratory for Adaptive Intelligence (2011).

Figure 6.9: Schema of the visual stimulus presented to the monkey.

In order to transform the electrocorticography data (signals composed of postsynaptic potentials) into spike trains we use a low-pass filter. The filtering process is carried out to obtain the peaks in the ECoG signals: if a peak is identified at the moment t in the neuron x then $x(t) = 1$ (otherwise $x(t) = 0$).

The first experiment consist in the clusterization of different zones in the monkey brain respect to a one visual stimulus. To this end, we select the data of visual grating at 45 degrees. The adaptive k-means algorithm is run for several cluster configurations: we use the 4 families of matching functions described in the first experiment and the parameter $k \in \{2, \ldots, 6\}$. In Figure 6.10 we present the elbow plot obtained for these clustering setups.

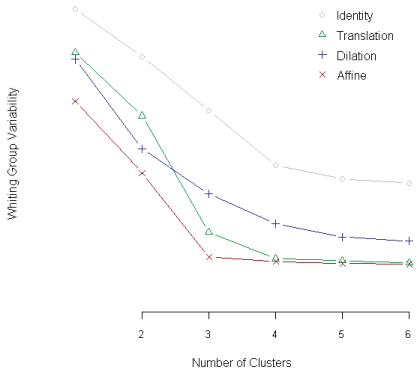


Figure 6.10: Elbow plot when clustering the monkey spike brain paths (visual stimulus: grating at 45 degrees).

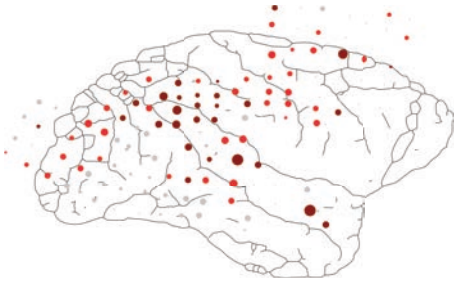


Figure 6.11: Monkey brain zones identified as clusters (size of the balls represents the average intensity rates and the colours the clusters labels).

As can be seen, we obtain different clustering results depending on which family of matching functions we use. When $h \in \mathcal{T}_{\text{identity}}$, the clustering algorithm suggest 4 groups. In the case of $h \in \mathcal{T}_{\text{affine}}$ or $h \in \mathcal{T}_{\text{translation}}$, the clustering algorithm suggest 3 clusters. Finally for $h \in \mathcal{T}_{\text{dilation}}$, the elbow criteria do not provide a clear rule to establish the number of clusters.

In Figure 6.11 we show the clusters obtained in the case of $h \in \mathcal{T}_{\text{affine}}$ ($k = 3$). The size of the points represents the average intensity rate computed for the neurons in the region and the color represents the clusters labels. We can see that the algorithm produces groups of neurons that has similar average intensity rates.

We also perform a third interesting experiment with the ECoG data set. We consider the sets of points in a raster plot represents the brain activity of the monkey during an stimulus: the horizontal axe represent the time dimension, that is: The evolution of the firing activity of the neurons during the 2 seconds that the experiment takes. And the vertical axe represents the spatial dimension, that is: The distribution of the neural firing process along the brain cortex.

For this experiment we select in total 9 raster-plots: 3 raster-plots for 3 different stimulus (grating at 45, 225 and 315 degrees). Every raster-plot it is a set of about 6.500 points in $\mathbb{R}^2$. We run the clustering algorithm for $k \in \{2, 3, 4, 5, 6\}$. The elbow plot obtained in this experiment is presented in Figure 6.12.

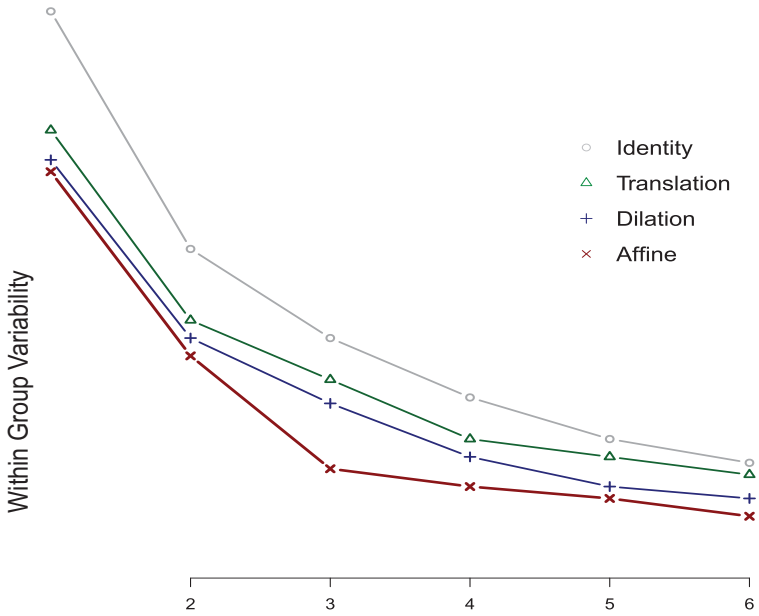


Figure 6.12: Scree-plot for the 4 families of matching functions.

As can be seen, only when we use $h \in \mathcal{T}_{\text{translation}}$ or $h \in \mathcal{T}_{\text{affine}}$ to align the raster-plots, the clustering algorithm suggest 3 clusters (the real number of clusters in the experiment).

Additionally, in Figure 6.13 we present the Multidimensional Scaling, a low dimensional representation of the neuronal firing activity, regarding of the 9 raster plots before the use of the proposed algorithm in Figure 6.13-left and after the use of the proposed method in Figure 6.13-right (using the family of affine transformations and $k = 3$). It can be seen, that the alignment procedure substantially improves the similarity between raster plots that belongs to the same cluster, organizing in this way the groups of brain signals according to the different exercised stimulus on the monkey.

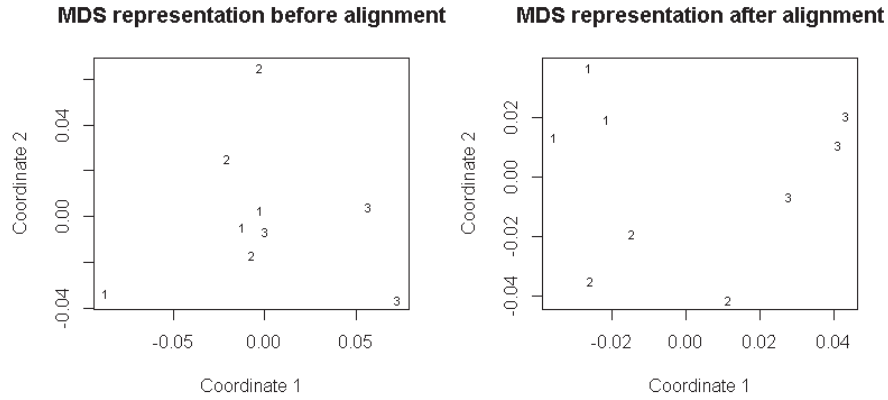


Figure 6.13: Multidimensional Scaling (MDS) representation of the brain activity before and after the alignment procedure. Numbers represent the labels of the raster-plot in the experiment.

Chapter Summary

In this chapter we propose a new affine invariant dissimilarity index for sets of points induced by a kernel for sets of points. The dissimilarity measure is combined with an alignment procedure to obtain a flexible k-means clustering algorithm. The proposed method is suitable to solve the clustering problem of sets of points that represent time series of brain spike trains.

Chapter 7

Conclusions and Future Work

7.1 Conclusions of the Thesis

In this thesis we propose distance measures that take into account the distributional properties of the statistical objects at hand in order to solve statistical and data analysis problems.

In Chapter 3 we define a Generalized Mahalanobis distance to the center of a probability measure for general unimodal probability distributions. We introduce a family of density kernels based on the underlying distribution of the data at hand. This family of density kernels induces distances that preserve the essential property of the Mahalanobis distance: "all the points that belong to the same probability curve, that is $L_c(f_{\mathbb{P}}) = \{\mathbf{x} | f_{\mathbb{P}}(\mathbf{x}) = c\}$ where $f_{\mathbb{P}}$ is the density function related to the probability measure $\mathbb{P}$, are equally distant from the center (the densest point) of the distribution". The family of density kernels introduced in Chapter 3 induces distances that generalize the Mahalanobis distance.

In Chapter 3 we also provide a computable version of the Generalized Mahalanobis distance that is based on the estimation of the level sets of the data at hand, avoiding in this way the need of explicitly estimate the density function. In the experimental section of Chapter 3, the proposed distance have been shown to be able to solve classification and outliers detection problems in a broad context.

Specific future research lines related to Chapter 3: In future work we will study how to generalize the distance induced by a density kernel for the general case, that is $d_{K_{\mathbb{P}}}(\mathbf{x}, \mathbf{y})$ where $\mathbf{y}$ does not to be the mode.

Another interesting applications of the proposed distance can also be considered. For instance, the density kernel distance proposed in Chapter 3 could be extended to be used in the context of functional data. In particular to compute the distance from a functional datum to the center of the distribution of the distribution. In this way, we will be able to solve the problem of outliers detection in the context of functional data.

In Chapter 4 we propose and study two distances for multivariate data that takes into account the probabilistic information of the data at hand. We provide estimation methods for the proposed distances and we study also its convergence. The proposed distances in this chapter are easily combined with standard classification methods in statistics and data analysis. We have shown that the proposed metrics work well in classification problems.

Specific future research lines related to Chapter 4: The study of alternative estimation methods of the proposed distances and its asymptotic properties is an important open research line for the near future. It is also interesting to study the underlying geometry induced by the proposed distances.

The distance proposed in this Chapter can be used to solve several practical problems in statistics and data analysis. In the experimental section of Chapter 4, we demonstrate that the proposed distance is able to solve classification problems in several contexts, but also can be implemented to solve regression problems. The extension of the use of the proposed metric to regression analysis problems and its applications is also an open research line.

The Chapter 5 of this thesis contains original contributions to the study of probability metrics. We consider, for the first time, probability measures as generalized functions, that is: continuous and linear functionals, that belong to a functional space endowed with an inner product. In this way, given two linear functionals (two probability measures) $\mathbb{P}_1$ and $\mathbb{P}_2$ will be identical (similar) if they act identically (similarly) for every function ϕ on the set of test functions $\mathcal{D}$.

In Chapter 5, we derive probability metrics from the metric structure inherited from the ambient inner product. We also demonstrate that when $\mathcal{D}$ is the set of indicator functions that

indicates the regions where the density remains constant, then $\mathcal{D}$ is rich enough to identify probability measures.

We propose 3 different weighting schemes for the family of distance measures proposed in Chapter 5 and we also provide the details about the computation of the proposed distances. A battery of real and simulated examples has been used to study the performance of the new family of distances. Using synthetically generated data, we have shown their performance in the task of discriminating data samples in several dimensions. Regarding the practical applications, the new family of probability metrics has been proven to be competitive in shape and textures recognition problems and text classification problems. The proposed distance also show to be useful to identify genes and discriminate between groups of patients by using DNA micro-arrays time series data.

Specific future research lines related to Chapter 5: In the near future we will treat the study of the asymptotic properties of the proposed family of probability metrics. We are also interested in the study of the geometry induced by the proposed family of probability metrics. Regarding the potential extension of the theory presented in Chapter 5, we are also interested in the study of distance measures for functional data; in particular for functional time series data. The potential list of applications of these research lines includes:

- **Control Theory and Time Series**: Linear and non linear systems can be studied observing the probability distributions of the response variables. The definition of suitable distance measures between these systems could be helpful in order to solve classification problems. This approach is particularly interesting to address several problems in Neuroscience, where the analysis of the data from neurophysiological investigations is challenging. The activity of the brain is modeled with complex systems of (usually non-linear) time series and the use of suitable distance measures is of fundamental importance in order to solve classification problems in this area.
- **Data analysis**: There exist a large list of algorithms that works using distance and/or similarity measures. Introducing new metrics that represent in a better way the distance between the data objects, in particular functional time series, will allow us to be able to improve the classification results in several practical areas as for example: Chemometrics Medicine or Computational biology to name just a few.
- **Information Fusion**: Kernel fusion and kernel approximation can be tackled by way

of developing a metric for kernels. The extension of the proposed theory to a distance measure for kernels could be helpful to implement kernel combinations in a better way.

In Chapter 6 we propose a kernel for data sets that takes into account the fundamental distributional properties of the data at hand. The proposed kernel for sets of points induces an affine invariant dissimilarity measure for sets of points. The dissimilarity measure for sets of points is combined with an alignment procedure in the context of a k-means clustering algorithm. The proposed clustering method is suitable to classify and to produce clusters of sets of points in several contexts. In the experimental section of Chapter 6, we use the proposed k-means method to classify spike train brain paths. Several artificial and real data experiments were carried out with outstanding results in classifying brain spike trains, a relevant problem in Neural Coding.

Specific future research lines related to Chapter 6: Future work includes the application of the proposed method in alternative contexts: clustering 2D images or 3D surfaces represented as sets of points, the classification of DNA micro-array time series, etc.

The inclusion of other families of matching functions, for example the reflexive transformations, useful when we work with images or shapes, is also part of the future work. The study of the rate of convergence and the properties of the proposed k-Means algorithm is also in the line of the future work plan.

7.2 General Future Research Lines

Usually the choice of a metric is based on the nature of the data: Euclidean distance for real data, Jaccard distance for binary data, Cosine distance for circular data and so on. In the case of functional data, the natural space is L_2 (the Hilbert space of all square integrable functions defined in a compact domain X). Therefore the ambient metric for functions in L_2 is simply $||f|| = \langle f, f \rangle^{\frac{1}{2}}$, that is the norm induced by the inner-product in the space. In general, this metric is relatively poor to describe similarity relationships for functional data.

We propose to continue studying the geometric properties and the probabilistic information of the data, in particular high-dimensional and functional data (FD), in order to define new distance and similarity measures between functional data. The aim to study new metrics

is the improvement of the performance in solving the typical statistical tasks such us classification and clustering with functional data, functional outlier detection and functional data representation and visualization.

7.2.1 A Mahalanobis-Bregman divergence for functional data

In the context of the study of distances for high dimensional data, we are working on the construction of a Mahalanobis-Bregman divergence for Functional Data. Our aim is to give a distance from a functional datum to the center (the most representative) functional datum in the population.

Let X be a compact domain and let μ be the Lebesgue measure in X. Let $\mathcal{H}$ be a Hilbert space of integrable functions ($\mathcal{H} \subset L^2_\mu(X)$). The covariance operator Σ_μ between two functional observations $f, g \in \mathcal{H}$ is defined as follows:

$$\Sigma_\mu(f,g) = \langle f, g \rangle_\mu = \int_X f(x)g(x)d\mu(x),$$

note that the **covariance** operator between f and $g \in \mathcal{H}$ it is well defined, a consequence of the Cauchy-Schwarz inequality:

$$\Big| \int_X f(x)g(x)d\mu(x) \Big|^2 \leq \int_X |f(x)|^2 d\mu(x) \int_X |g(x)|^2 d\mu(x) < \infty.$$

If $\mathcal{H}$ it is a Hilbert space of functions, by Moore-Aronszajn Theorem there exist a continuous, symmetric and positive definite function $K : X \times X \to \mathbb{R}$, the kernel function, such that for any $f \in \mathcal{H}$, then $f(x) = \langle f, K_x \rangle$ where $K_x : X \to \mathbb{R}$ is the evaluation functional at the point x, that is $K_x(t) = K(x,t)$.

By Mercer's theorem (refer to Chapter 2 for details) we can write $K(x;y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)$, where the corresponding set of eigenfunctions $\{\phi_i(x)\}_{i=1}^{\infty}$ form an orthonormal basis in $L^2_\mu(X)$. By using this kernel decomposition, we can write $f \in \mathcal{H}$ as follows:

$$\begin{aligned} f(x) = \langle f, K_x \rangle_\mu &= \langle f, \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(\cdot) \rangle_\mu \\ &= \int_X f(t) \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(t) d\mu(t), \\ &= \sum_{i=1}^{\infty} \lambda_i^* \phi_i(x), \end{aligned}$$

where $\lambda_i^* = \langle f, \lambda_i \phi_i \rangle_\mu$ (in what follows we use the notation λ_i instead of λ_i^* for simplicity). Then for two functions $f, g \in \mathcal{H}$ the covariance can be computed in the following way:

$$\Sigma_\mu(f, g) = \int_X \left(\sum_{i=1}^{\infty} \lambda_i \phi_i(x) \sum_{i=1}^{\infty} \alpha_i \phi_i(x) \right) d\mu(x),$$

By the orthonormality of the eigenfunctions associated to K:

$$\Sigma_\mu(f, g) = \sum_{i=1}^{\infty} \lambda_i \alpha_i,$$

wich is the expression of the **covariance** between the functions f and $g \in \mathcal{H}$.

The Functional Bregman divergence (refer to Chapter 2 Definition 2.2) between the functions f and $g \in \mathcal{H}$ associated to the strictly convex and differentiable function $\zeta : \mathcal{H} \to X$ is as follows:

$$BD_\zeta(f, g) = \int_X \Big(\zeta(f) - \zeta(g) - (f - g) \nabla\zeta(g) \Big) d\mu,$$

where μ is the Lebesgue measure in X and $\nabla\zeta(g)$ is the derivative of ζ evaluated at g.

If we choose now the strictly convex and differentiable function $\zeta : \mathcal{H} \to X$ as:

$$\zeta(t) = c \times t^2,$$

where the constant $c = \Sigma_\mu(f, g)^{-1}$, then for $f \in \mathcal{H}$:

$$\zeta(f) = \Sigma_\mu(f,g)^{-1} \times f^2.$$

By using the definition of the Bregman divergence (Chapter 2-2.2), the expression for the Mahalanobis-Bregman Divergence between f and g is:

$$BD_\zeta(f,g) = \frac{\sum_{i=1}^{\infty}(\lambda_i - \alpha_i)^2}{\sum_{i=1}^{\infty}\lambda_i\alpha_i}.$$

It is interesting to notice the following properties of the proposed metric:

- In the case when $\Sigma_\mu(f,g) \to 0$, that is f and g are orthogonal or independent, then $BD_\zeta(f,g) \to \infty$.
- When $g = a + bf$ for $a, b \in \mathbb{R}$, that is g and f are linear dependent, then $g(x) = a + b\sum_{i=1}^{\infty}\lambda_i\phi_i(x)$. In this case we have that $\alpha_i = b\lambda_i \; \forall i$. Therefore we have:

$$BD_\zeta(f,g) = \frac{(1-b)^2\sum_{i=1}^{\infty}\lambda_i^2 + a^2}{b\sum_{i=1}^{\infty}\lambda_i^2},$$

  then if
  - $a = 0 \longrightarrow BD_\zeta(f,g) = \frac{(1-b)^2}{b}$,
  - $b = 1 \longrightarrow BD_\zeta(f,g) = \frac{a^2}{\sum_{i=1}^{\infty}\lambda_i^2}$,
  - $a = 0$ and $b = 1 \longrightarrow BD_\zeta(f,g) = 0$.

Other important properties associated to this metric are:

(i) **Non-negativity**: $BD_\zeta(f,g) \geq 0$ and $BD_\zeta(f,g) = 0$ if and only if $f = g$,

(ii) **Convexity**: Bregman divergence is a convex function respect the first argument, i.e. $f \longmapsto BD_\zeta(f,g)$ is a convex function for all $g \in \mathcal{H}$.

(iii) **Linearity**: $BD_{\alpha\zeta_1+\beta\zeta_2} = \alpha BD_{\zeta_1} + \beta BD_{\zeta_2}$ for all ζ_1 and ζ_2 strictly convex functions and positive constants α and β.

(iv) **Affine Invariance**: let T be an affine function (i.e. $T \circ f(x)$ produces a combination of rotations, rescalings and translations on f), then $BD_\zeta(T \circ f, T \circ g) = BD_{\zeta\circ T}(f,g) = BD_\zeta(f,g)$.

The convexity of the Bregman Divergences is an important property for many Machine Learning algorithms. The following are interesting points that will be developed in the future research line:

- Let $F = \{f_1, ..., f_n\}$ and $G = \{g_1, ..., g_m\}$ be two different populations of functions where $\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$ and $\bar{g} = \frac{1}{m}\sum_{i=1}^{m} g_i$ are the respective "functional means" of every functional group. Then we can define the covariance between functional means as:

$$\Sigma_\mu(\bar{f}, \bar{g}) = \sum_{i=1}^{\infty} \boldsymbol{\lambda}_i \boldsymbol{\alpha}_i,$$

  where $\boldsymbol{\lambda}_i = \bar{\lambda_i} = \frac{1}{n}\sum_{j=1}^{n} \lambda_i^{[j]}$ and $\boldsymbol{\alpha}_i = \bar{\alpha_i} = \frac{1}{m}\sum_{j=1}^{m} \alpha_i^{[j]}$. Following the definition of the Mahalanobis Bregman Divergence for functional data, we can compute the distance between the functional populations F and G as follows:

$$BD_\zeta(F, G) = BD_\zeta(\bar{f}, \bar{g}),$$

  where $\zeta(t) = \Sigma_\mu(\bar{f}, \bar{g})^{-1} \times t^2$. We can use this measure in order to solve Hypotheses test for functional data, that is to accept $H_0 : F = G$ when $BD_\zeta(F, G) \approx 0$ and reject $H_0 : F = G$ otherwise.

- Functional depth: The depth associated to the functional datum $f_i \in F$ can be computed as the Mahalanobis Bregman divergence to the center:

$$BD_\zeta(\bar{f}, f_i),$$

  this measure could be helpful to identify "functional outliers" in a sample of functional data.

7.2.2 Pairwise distances for functional data

The study of pairwise distance and similarity measures for functional data, a more general case than the distance to a center, is also relevant in order to address several practical problems in Functional Data Analysis. The use of a RKHS to represent functional data provides a rich framework to define new distance measures for pairwise functional data but it is not the unique approach.

A different perspective is to consider functional data as points that belongs to a Riemannian manifold $\mathcal{M}$. A manifold $\mathcal{M}$ is a topological space with the important property that at every point $\mathbf{x} \in \mathcal{M}$, the geometry in a small neighborhood of $\mathbf{x}$ is Euclidean. A smooth Riemannian manifold carries the structure of a metric space where the distances between any two points in the manifold is defined as the length of the geodesic path that joints the two points. Let γ be a smooth curve parametrized by $t \in [0, 1]$ (we requires that $\gamma \in C^2[0, 1]$) in the manifold $\mathcal{M}$, then $\gamma'(t) \in T_{\gamma(t)}$, where $T_{\mathbf{x}}$ is the tangent space of $\mathcal{M}$ at the point $\gamma(t) = \mathbf{x}$. The length of the curve γ with respect to the Riemannian structure $\mathcal{M}$ is given by:

$$\text{arc lenght}\, \gamma = \int_0^1 \|\gamma'(t)||dt,$$

I propose to study the way to represent functional data in a Riemannian manifold $\mathcal{M}$ and study also its associated metric. The proposed research line also includes the study of different estimation methods for the proposed metrics and the study of the asymptotic properties of these estimators. This will allow the efficient computation of the distances when we use the metrics to solve different classification and regression problems in the context of Functional Data.

The list of potential applications of this research line includes:

- **Solving problems that involves complex functional data objects**: Functional data come from far ranging fields as for example Bioinformatics, Image and Vision, Medicine, Environmental Sciences, Marketing and Finance, among many other sources. Example of complex data objects in these areas are: images of the internal structure of the body, images from diagnostic medical scanners, sequences of DNA, time series representing the volatility in share prices, marketing networks, etc. The definition of suitable distance measures between these complex functional data objects could be helpful in order to solve several relevant problems in these areas. For example to provide a description of the variations in shapes of internal organs, or the evolution and subject-to-subject variability in the pattern of DNA sequences, or the variations in the connections of a marketing network, etc.

- **Registration and dynamics of functional data**: The process of functional data alignment is of crucial importance in order to solve several data analysis problems, as we have shown in Chapter 6 of this thesis. The study of distance measures that incorporates geometrical information of the complex functional objects at hand is of fundamental im-

portance in order to define proper registration procedures for these functional data.

- **Inferential problems in functional data analysis**: Most classical inferential techniques are inappropriate for functional data. The study of distances for functional data could produce important contributions to develop inferential techniques for functional data. As an example of the importance of this point, consider the case of the Hypothesis test between groups of functional populations treatised in Chapter 5 of this thesis.

- **Study the relationship on functional data objects**: Functional data techniques are increasingly being used in domains where the data are spatio-temporal in nature and hence is typical to consider correlated in time or space functional data. The development of reliable prediction and interpolation techniques for dependent functional data is crucial in these areas. This point can also be addressed by developing distance and similarity measures that incorporate the geometric and probabilistic information that characterize the complex functional data.

7.2.3 On the study of metrics for kernel functions

Several methods and algorithms in Statistics and Data Analysis make use of Kernel functions to solve classification, regression and density estimation problems, consider for example the case of Support Vector Machines that make use of a kernel function in order to solve the 3 tasks mentioned above. In this context, sometimes results convenient to merge different heterogeneous sources of information. Kernel fusion and kernel approximation can be tackled by way of developing a metric for kernels. If the metric takes into account the geometrical structure of the space where these functions lives, we are going to be able to produce better fusions and better approximations.

The study of representation methods for functional data is of fundamental importance in Functional Data Analysis. The use of a metric for kernel functions is also crucial when one needs to determine which is the best kernel function to represent a functional datum. In this way, the proposed future research line also includes the study of representation methods that make use of the proposed metrics for kernels.

Appendix A

Appendix to Chapter 3

Generally speaking, a function $K : X \times X \to \mathbb{R}^+$ is a Mercer kernel if it is a continuous, symmetric and positive semi-definite function.

Theorem A.1. *The Kernel proposed in Definition 3.3 it is a Mercer Kernel.*

Proof. The density kernel proposed in Definition 3.3 is continuous and symmetric by definition. To prove that it is also a semi-definite function consider an arbitrary collection of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in X. The matrix $\mathbf{K} = [K_{\mathbb{P}}(\mathbf{x}_i, \mathbf{x}_j)]_{i,j} \in \mathbb{R}^{n \times n}$ is a positive definite matrix:

$$\begin{aligned} \sum_{i,j}^{n} \mathbf{x}_i^T \mathbf{x}_j K_{\mathbb{P}}(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i,j}^{n} \mathbf{x}_i \phi_{\mathbb{P}}(\mathbf{x}_i) \mathbf{x}_j \phi_{\mathbb{P}}(\mathbf{x}_j), \\ &= \left(\sum_{i}^{n} \mathbf{x}_i \phi_{\mathbb{P}}(\mathbf{x}_i) \right)^2 \geq 0. \end{aligned}$$

□

Appendix B

Appendix of Chapter 4

The calculus of variations in a nutshell

This appendix provides a note on the methods of the Calculus of Variations, for further details about this topic refer to Fox (1987); Sagan (2012) and references therein. We assume here that X is a compact set (without loss of generality we assume that $X \subset \mathbb{R}^d$). The variational problem that we considers in this thesis consist in find a function $v(\mathbf{x}) : X \to \mathbb{R}$ which minimizes the following functional:

$$F(v) = \int \cdots \int_X \phi(\mathbf{x}, v, \nabla v)\, dx_1 \cdots dx_d,$$

where we can interpret ϕ as a loss-function that involves v and its gradient vector $\nabla v = (v_{x_1}, \cdots, v_{x_d})$, being $v_{x_i} = \frac{\partial v}{\partial x_i}$. Usually the integral on ϕ measures physical quantities: Distance, volume, time, etc. The necessary condition to minimize the functional is attained when the function v fulfills the Euler-Lagrange equation:

$$\sum_{i=1}^{d} \frac{d}{dx_i}\left(\frac{\partial \phi}{\partial v_{x_i}}\right) = \frac{\partial \phi}{\partial v},$$

where $v_{x_i} = \frac{\partial v}{\partial x_i}$. The variational problem stated above, can be naturally generalized to the case where we $\mathbf{v}$ is a vector valued minimizer. Denote by $\mathbf{v}(\mathbf{x}) : X^k \to \mathbb{R}$ to the vector function $\mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), \ldots, v_k(\mathbf{x}))$, and $\nabla \mathbf{v}(\mathbf{x}) = (\nabla v_1(\mathbf{x}), \ldots, \nabla v_k(\mathbf{x}))$ to its gradient, then the Euler-Lagrange necessary first-order conditions constitutes a system of differential equations:

$$\sum_{i=1}^{d} \frac{d}{dx_i}\left(\frac{\partial \phi}{\partial\left(\frac{\partial v_j}{\partial x_i}\right)}\right) = \frac{\partial \phi}{\partial v_j}, \quad j = 1, \ldots, k.$$

For further details and possible extensions to the variational problems refer to Fox (1987); Sagan (2012) and references therein. Next we give an example to clarify the concepts and to illustrates how the definition of distance can be obtained as a result of a (variational) minimization problem.

Example B.1. The shortest path between two points.
We considers the problem of find the shortest path between two points, say A and $B \in X$ (assume in this example that $X \subset \mathbb{R}^2$). It is well known that the answer to this question is the straight line (segment) that joints the points A and $B \in X$. Therefore the shortest distance between the points A and B is the length of the straight line between the points, that is the Euclidean distance between the points A and B. Next we give a justification based on the calculus of variations for this answer.

Considers a twice continuously differentiable curve $\gamma \in C^2[I]$ where I is an open interval in $\mathbb{R}$. Through this chapter we use a parametrization of curve γ defined in $I = [0, 1]$ as follows:

$$\gamma : [0, 1] \to \mathbb{R}^2 \quad t \longmapsto \gamma(t) = (x_1(t), x_2(t)) \,,$$

where (x_1, x_2) are the coordinates values in $\mathbb{R}^2$. If we are looking for curves that joints the points A and B, the additional condition $\gamma(0) = \left(x_1^A, x_2^A\right) = A$ and $\gamma(1) = \left(x_1^B, x_2^B\right) = B$ should also be fulfilled. In Figure B.1 we have represented several curves that fulfills these conditions. In this setting, the infinitesimal distance we cover when we move along the curve γ at any point t it can be approximated by: $d_\gamma(t, t + \Delta t)^2 \approx \Delta x_1^2 + \Delta x_2^2$. Dividing both sides by Δt and taking limits when $\Delta t \to 0$, we derive that:

$$\lim_{\Delta t \to 0} \frac{d_\gamma(t, t + \Delta t)^2}{\Delta^2 t} = x_1'(t)^2 + x_2'(t)^2 = ||\gamma'(t)||^2,$$

that can be interpreted as the "local" (infinitesimal) distance around the point t by moving through the curve γ. Therefore the total distance to move from A to B by using the curve γ is given by the sum of all the infinitesimal distances in the following way:

$$d_\gamma(A, B) = \int_0^1 ||\gamma'(t)||dt = \int_0^1 \sqrt{x_1'(t)^2 + x_2'(t)^2}dt.$$

Therefore, the minimum distance (variational) problem can be stated in the following way:

$$d(A, B) := \begin{cases} \min_{\gamma \in C^2} \int_0^1 ||\gamma'(t)||dt, \\ \text{s.t. } \gamma(t = 0) = A, \ \gamma(t = 1) = B. \end{cases}$$

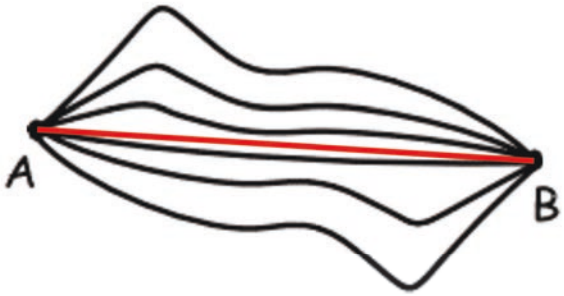


Figure B.1: Several smooth curves γ that joints the points A and B. The geodesic (red line) is the shortest path between the points A and B.

By using the Euler-Lagrange first order condition we obtain that:

$$\begin{array}{l} \frac{d}{dt}\left(2x_1'(t)\right) = 2x_1''(t) = 0, \\ \frac{d}{dt}\left(2x_2'(t)\right) = 2x_2''(t) = 0, \end{array} \quad s.t. \begin{cases} (x_1(0), x_2(0)) = \left(x_1^A, x_2^A\right), \\ (x_1(1), x_2(1)) = \left(x_1^B, x_2^B\right), \end{cases}$$

the first order condition implies that $x_1(t)$ and $x_2(t)$ are linear functions with respect to t. By using the constrains involved in this problem, the we arrive to a parametric expression for the optimal path:

$$\gamma(t) = (x_1(t), x_2(t)) = (x_1^A + t(x_1^B - x_1^A), x_2^A + t(x_2^B - x_2^A)).$$

Therefore, the optimal (in terms of minimal length) curve to go from point $A = \left(x_1^A, x_2^A\right)$ to the point $B = \left(x_1^B, x_2^B\right)$ in the plane it is the straight curve $\gamma(t) = (x_1^A + t(x_1^B - x_1^A), x_2^A + t(x_2^B - x_2^A))$. The optimal is represent in Figure B.1 with a red colored path. This example justify the use of the Euclidean distance in the case we are able to move freely through the space $\mathbb{R}^2$.

• • •

The calculus of variations and the Minimum Work Statistical Distance

The calculus of variations can also be used to determine a solution to the problem stated in the Definition 4.5, next we demonstrate how to compute the Euler-Lagrange condition for this case.

Let $f_{\mathbb{P}}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a multivariate normal density function parametrized by a vector of means $\boldsymbol{\mu} \in \mathbb{R}^d$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. Let $\gamma(t)$ be a smooth curve in $C^2[0, 1]$

parametrized in $t \in [0,1]$. Then the minimum work between two points $\mathbf{x}$ and $\mathbf{y}$ in X is obtained by minimizing the following functional:

$$\inf_{\gamma \in C^2} \int_0^1 f_{\mathbb{P}}(\mathbf{x}(t), \boldsymbol{\mu}, \boldsymbol{\Sigma})\, dt,$$

subject to the initial conditions $\gamma(t=0) = \mathbf{x}$ and $\gamma(t=1) = \mathbf{y}$. By using the Euler-Lagrange first order condition we obtain:

$$-\, f_{\mathbb{P}}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, (\mathbf{x}(t) - \boldsymbol{\mu})^T\, \boldsymbol{\Sigma}^{-1} \nabla \mathbf{x}(t) = 0, \tag{B.1}$$

where $\nabla \mathbf{x}(t) = \left(x_1'(t), \ldots, x_d'(t)\right)$ denotes the gradient of the vector valued function $\mathbf{x}(t) = (x_1(t), \ldots, x_d(t))$. It is possible to obtain a closed solution for the differential equation in (B.1) in some simple context. For example, in the particular case when $d = 2$, $\boldsymbol{\Sigma} = I_{2\times 2}$ (where I denotes the identity matrix), $\boldsymbol{\mu} = (0,0)$ and $f_{\mathbb{P}}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f_{\mathbb{P}}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$; then from the first order condition we obtain that:

$$\frac{x_1(t)}{x_2(t)} = -\frac{x_2'(t)}{x_1'(t)},$$

where $x_1(t)$ and $x_2(t)$ denotes the coordinates of the parametrized curve

$$\gamma(t) = (x_1(t), x_2(t))\,.$$

A solution for the differential Equation (B.1) is given by: $x_1(t) = r\cos(t)$ and $x_2(t) = r\sin(t)$. Therefore the resulting path between the points $\mathbf{x}$ and $\mathbf{y}$ is constituted by the level-set $L_c(f) = \{\mathbf{x} | f(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c\}$ between the two point. Then the minimum work statistical distance is constituted by the line integral of the density function that goes over the level set curve, as can be seen in Figure 4.6.

Distribution theory in a nutshell

This appendix constitutes a brief reference of the theory of Distributions (also known as Generalized Functions). For further details refer to Schwartz (1957); Strichartz (2003); Zemanian (1982) and references therein.

Let X be an open subset of $\mathbb{R}^d$, the space $C_c^\infty(X)$ consisting of the infinitely differentiable functions with compact support in X is called the space of *test* functions and it is denoted by $\mathcal{D}(X)$. We can now define the class of distributions on X, denoted as $\mathcal{D}'(X)$, to be all

continuous *linear functionals* on $\mathcal{D}(X)$. Thus the distribution $f \in \mathcal{D}'(X)$ is a linear map $f : \mathcal{D} \to \mathbb{R}$, that is:

$$f[\alpha_1\phi_1 + \alpha_2\phi_2] = \alpha_1 f[\phi_1] + \alpha_2 f[\phi_2],$$

for $\alpha_i \in \mathbb{R}$ and $\phi_i \in \mathcal{D}$ with $i = 1, 2$. The most natural way to define the distribution f is simply by ordinary integration, that is: $f[\phi] = \langle f, \phi \rangle = \int_{-\infty}^{\infty} f(x)\phi(x)\, dx$.

Although the integral of an ordinary function is one way to define a distribution, it is not the only way. We can define the linear and continuous map:

$$\delta[\alpha_1\phi_1 + \alpha_2\phi_2] = \alpha_1\phi_1(0) + \alpha_2\phi_2(0),$$

and δ is simply the distribution that assigns $\phi(0)$ to every $\phi \in \mathcal{D}$. The definition of the derivative of a distribution f is straightforward:

$$f'[\phi] = \langle f', \phi \rangle = \int_{-\infty}^{\infty} f'(x)\phi(x)\, dx = -\int_{-\infty}^{\infty} f(x)\phi'(x)\, dx = \langle f, -\phi' \rangle = f[-\phi'],$$

where we have integrated by parts and use the fact that ϕ vanishes outside a compact support. The function ϕ' is just the ordinary derivative of the test function ϕ. Since $\delta[\phi] = \phi(0)$, immediately follows that $\delta'[\phi] = \langle \delta, -\phi' \rangle = -\phi'(0)$, as we use in Equation 4.1 of Chapter 4.

Appendix C

Appendix of Chapter 5

Affine invariant property of the proposed metric

Lemma C.1. *Lebesgue measure is equivalent under affine transformations.*

Proof. Let X be a random variable that take values in $\mathbb{R}^d$ distributed according to $\mathbb{P}$, and let $f_{\mathbb{P}}$ be its density function. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be an affine transformation, define the r.v. $X^* = T(X) = a + bRX$, where $a \in \mathbb{R}^d$, $b \in \mathbb{R}^+$ and $R \in \mathbb{R}^{d\times d}$ is an orthogonal matrix with $\det(R) = 1$ (therefore R^{-1} exist and $R^{-1} = R^T$). Then X^* is distributed according to $f_{\mathbb{P}^*}$. Define $E^* = \{x^* | x^* = T(x) \text{ and } x \in E\}$, then:

$$\begin{aligned}\mu^*(E^*) &= \int_{E^*} d\mathbb{P}^* = \int_{E^*} f_{\mathbb{P}^*}(x^*)dx^* = \int_{E^*} f_{\mathbb{P}}(T^{-1}(x^*)) \left|\frac{\partial T^{-1}(x^*)}{\partial x^*}\right| dx^*, \\ &= \int_{E^*} f_{\mathbb{P}}\left(R^{-1}\left(\frac{x^*-a}{b}\right)\right)\frac{R^{-1}}{b}dx^* = \int_E f_{\mathbb{P}}(y)dy = \int_E d\mathbb{P} = \mu(E).\end{aligned}$$

□

Theorem C.1. ***Invariance under affine transformation*** *The metric proposed in Equation (5.5) is invariant under affine transformations.*

Proof. Let T be an affine transformation, we prove that the measure of the symmetric difference of any two α-level sets is invariant under affine transformation, that is: $\mu(A_i(\mathbb{P}) \triangle A_i(\mathbb{Q})) = \mu^*\left(T(A_i(\mathbb{P})) \triangle T(A_i(\mathbb{Q}))\right) = \mu^*(A_i(\mathbb{P}^*) \triangle A_i(\mathbb{Q}^*))$. By Lemma 1:

$$\begin{aligned}\mu^*(A_i(\mathbb{P}^*) \triangle A_i(\mathbb{Q}^*)) &= \int_{A_i(\mathbb{P}^*)-A_i(\mathbb{Q}^*)} d\mathbb{P}^* + \int_{A_i(\mathbb{Q}^*)-A_i(\mathbb{P}^*)} d\mathbb{Q}^* \\ &= \int_{A_i(\mathbb{P})-A_i(\mathbb{Q})} d\mathbb{P} + \int_{A_i(\mathbb{Q})-A_i(\mathbb{P})} d\mathbb{Q} = \mu(A_i(\mathbb{P}) \triangle A_i(\mathbb{Q})).\end{aligned}$$

The same argument can be applied to the denominator in the expression given in Equation (5.5), thus $\frac{w_i}{\mu^*(A_i(\mathbb{P}^*)\cup A_i(\mathbb{Q}^*))} = \frac{w_i}{\mu(A_i(\mathbb{P})\cup A_i(\mathbb{Q}))}$ for $i = 1, \ldots, m-1$. Therefore as this is true for all the α-level sets, then the distance proposed in Equation (5.5) is invariant under affine transformations:

$$d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P}^*, \mathbb{Q}^*) = \sum_{i=1}^{m-1} w_i d\left(\phi_i(\mathbb{P}^*), \phi_i(\mathbb{Q}^*)\right) = \sum_{i=1}^{m-1} \lambda_i d\left(\phi_i(\mathbb{P}), \phi_i(\mathbb{Q})\right) = d_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{P}, \mathbb{Q}).$$

□

Computational complexity of the proposed metric

We study theoretically the computational time of the proposed algorithm in Table X and find that the execution time required to compute $\hat{\mathbf{S}}_\alpha(\mathbf{f})$ grows at a rate of order $\mathcal{O}(dn^2)$, where d represent the dimension and n the sample size of the data at hand. This is because we need to compute (only once) a distance matrix.

In order to show empirically this effect, we simulate two data sets with $n = 100$ observations in dimensions $d = \{1, 5, 10, 15, 20, 25, 30, 50, 100, 200, 1000\}$ and we compute[1] the system execution time in every case. As can be seen in Figure C.1, the execution time increases linearly with respect the number of dimensions considered. In the other case, we simulate two data sets in dimension $d = 2$ with sample size $n = \{10, 100, 500, 1000, 2000, 5000\}$ and compute the system execution time in every case. As can be seen in Figure C.2, the execution time increases in a quadratic way in this case.

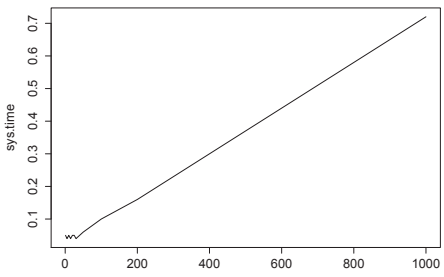


Figure C.1: The computational time as dimension increases.

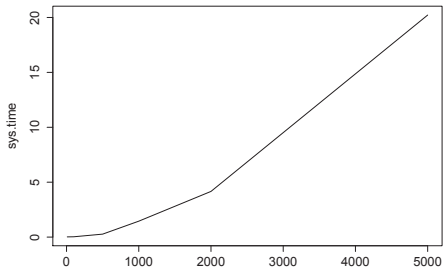


Figure C.2: The computational time as sample size increases.

[1] The computations were carried out with a $i5$-Intel processor computer with Windows 7 at $2.80GHz$ and 4GB of RAM.

Additionally we also compare in Table C.1 and Figure C.3, the computational times of the proposed distance with other referenced metrics in the chapter.

Table C.1: Computational time (sec) of main metrics in Experiment of Sec 4.1.1

Metric	dim:	1	2	3	4	5	10	15	20	50	100
Energy		0.01	0.02	0.05	0.25	0.44	1.12	2.13	3.57	40.12	118.3
MMD		0.01	0.04	0.07	0.12	0.20	0.92	1.24	2.02	19.84	89.5
LS(1)		0.02	0.06	0.10	0.19	0.31	1.07	2.06	3.67	35.36	112.9

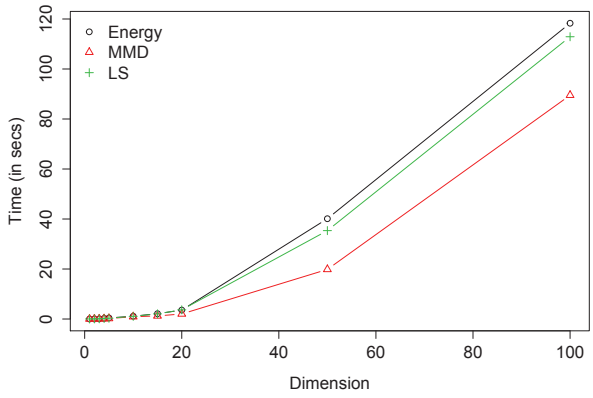


Figure C.3: Execution times of the main metrics in Experiment of Section 4.1.1.

As can be seen the computational time of the LS distance is in the same order of Energy and MMD distance.

An alternative approach to the first artificial experiment with the normal distributions

We originally develop the experiment of Section 5.4.1 by using a permutation test that do not allow us to track the power of test. In several discussions we find Statistical arguments to prefer this approach, therefore we include in this appendix the result of the experiment regarding the discrimination between normal distributions by using the permutation test procedure.

For this end we generate a data sample of size $150d$ from a $N(\mathbf{0}, \mathbf{I}_d)$ where d stands for dimension. We also considers a second data sample of a displaced distribution $N(\mathbf{0} + \boldsymbol{\delta}, \mathbf{I}_d)$

where $\boldsymbol{\delta} = \delta\mathbf{1} = \delta(1, \ldots, 1) \in \mathbb{R}^d$. Next we compute the distance between these two data samples and perform a permutation test, based on 1000 permutations, in order to determine the p-value of the discrimination test. The p-value is computed as the proportion of the times that the distance between the original samples is lower than the distance between the permuted samples. The number of level sets to consider, the parameter denoted as m in the previous section, where fixed as $25\sqrt{d}$, the radii parameter ($r_{\hat{A}_i(\mathbb{P})}$ and $r_{\hat{A}_i(\mathbb{Q})}$) has been chosen as the median distance among the elements inside the level set. In Table C.2 we report the minimum distance ($\delta^*\sqrt{d}$) between distributions centers (the means) required to discriminate for each distance whiting a fixed p-value of 0.05. The lower reported values indicates better discrimination power.

Table C.2: Minimum distance ($\delta^*\sqrt{d}$) to discriminate among the data samples with a 5% p-value.

Metric	d: 1	2	3	4	5	10	15	20	50	100
KL	0.285	0.375	0.312	0.309	0.306	0.305	0.302	0.298	0.297	0.295
T	0.255	0.332	0.286	0.280	0.279	0.284	0.257	0.255	0.254	0.204
Energy	0.235	0.318	0.286	0.280	0.279	0.269	0.255	0.250	0.247	0.202
MMD	0.315	0.321	0.302	0.295	0.287	0.278	0.265	0.255	0.249	0.223
LS(0)	0.255	0.346	0.294	0.290	0.287	0.284	0.278	0.264	0.268	0.242
LS(1)	**0.225**	**0.304**	**0.275**	**0.270**	**0.268**	**0.253**	**0.249**	**0.246**	**0.244**	**0.194**
LS(2)	0.235	0.310	0.277	0.275	0.270	0.263	0.251	0.250	0.245	0.200
LS(3)	0.238	0.318	0.282	0.278	0.275	0.269	0.257	0.255	0.254	0.232

In a second experiment we consider again normal populations but different variance-covariance matrices. Define as an expansion factor $\sigma \in \mathbb{R}$ and increase σ by small amounts (starting from 0) in order to determine the smallest σ^* required for each metric in order to discriminate between the $150d$ sampled data points generated for the two distributions: $N(\mathbf{0}, \mathbf{I}_d)$ and $N(\mathbf{0}, (1+\sigma)\mathbf{I}_d)$. In order to determine whether the computed distances are able to differentiate among the two simulated data samples, we repeat the process given in the first experiment. We first compute the distance between the two simulated data samples (considering $m = 25\sqrt{d}$ and the radii parameter chosen as the median distance among the elements inside the level set). Next we run a permutation test, based on 1000 permutations, in order to compute the p-values of the test. In Table C.3, we report the minimum $(1 + \sigma^*)$ to obtain a p-value of 5%. The lower reported values indicates better discrimination performance.

We do not include these results in Chapter 5 - Section 5.4.1 as we consider them redundant.

Table C.3: $(1 + \sigma^*)$ to discriminate among the data samples with a 5% p-value.

Metric	dim:	1	2	3	4	5	10	15	20	50	100
KL		1.850	1.820	1.810	1.785	1.750	1.700	1.690	1.620	1.565	1.430
T		–	–	–	–	–	–	–	–	–	–
Energy		1.630	1.550	1.530	1.500	1.480	1.420	1.400	1.390	1.340	1.300
MMD		1.980	1.755	1.650	1.580	1.520	1.490	1.430	1.410	1.390	1.340
LS(0)		1.690	1.580	1.530	1.510	1.490	1.450	1.410	1.390	1.350	1.310
LS(1)		**1.570**	**1.480**	**1.460**	**1.410**	**1.395**	**1.370**	**1.320**	**1.290**	**1.210**	**1.150**
LS(2)		**1.570**	1.490	1.480	1.450	1.420	1.390	1.360	1.310	1.290	1.220
LS(3)		1.580	1.520	1.510	1.480	1.460	1.410	1.390	1.370	1.340	1.290

Appendix D

Appendix of Chapter 6

Lemma D.1. *The counting measure defined on a finite set it is invariant under translation, scaling and rotation transformations.*

Proof. Let $A = S_{\mathbb{P}}^n = \{x_i\}_{i=1}^n$, and $B = S_{\mathbb{Q}}^m = \{y_j\}_{j=1}^m$ two finite sets of points, generated from the PMs $\mathbb{P}$ and $\mathbb{Q}$ respectively. Define the (finite) set $S = S_{\mathbb{P}}^n \cup S_{\mathbb{Q}}^m = A \cup B$ and denotes by μ_K the counting measure on S.

Let $\mathcal{T}$ be the class of affine transformation such that $\forall h \in \mathcal{T}$: $h(x) = a + bRx$, where $a \in \mathbb{R}^d$, $b \in \mathbb{R}^+$ and $R \in \mathbb{R}^{d\times d}$ is an orthogonal matrix with $\det(R) = 1$ ($R^{-1} = R^T$). As h it is a homeomorphism, then for all $x \in S$: $h(x) \in h \circ A \cap h \circ B$ if and only if $x \in A \cap B$. This last implies that $\mu_K(h \circ A \cap h \circ B) = \mu_K(A \cap B)$. □

Theorem D.1. ***Invariance under affine transformation*** *The dissimilarity index proposed in Definition 6.5 is invariant under translation, scaling and rotation transformations.*

Proof. Let $\mathcal{T}$ be the class of translation, scaling and rotation transformations and let $h \in \mathcal{T}$ be an affine map. We need to prove that $K(h \circ A, h \circ B) = K(A, B)$ for all sets of points $A,\ B \in X$, in order to demonstrate the theorem. By Lemma 1, we can write:

$$K(h \circ A, h \circ B) = \mu_K(h \circ A \cap h \circ B) = \mu_K(A \cap B) = K(A, B),$$

therefore $d_K(A, B) = d_K(h \circ A, h \circ B)$. □

References

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1988). Network flows. Technical report, DTIC Document.

Amari, S.-I. (2009a). Divergence is unique, belonging to both-divergence and bregman divergence classes. *Information Theory, IEEE Transactions on*, 55(11):4925–4931.

Amari, S.-i. (2009b). Information geometry and its applications: Convex function and dually flat manifold. In *Emerging Trends in Visual Computing*, pages 75–102. Springer.

Amari, S.-I., Barndorff-Nielsen, O. E., Kass, R., Lauritzen, S., and Rao, C. (1987). Differential geometry in statistical inference. *Lecture Notes-Monograph Series*, pages i–240.

Amari, S.-I. and Cichocki, A. (2010). Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195.

Amari, S.-i. and Nagaoka, H. (2007). *Methods of information geometry*, volume 191. American Mathematical Soc.

Arbter, K., Snyder, W. E., Burkhardt, H., and Hirzinger, G. (1990). Application of affine-invariant fourier descriptors to recognition of 3-d objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):640–647.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, pages 337–404.

Atkinson, C. and Mitchell, A. F. (1981). Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 345–365.

Bachman, D. (2012). *A geometric approach to differential forms*. Springer.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005). Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749.

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*, volume 3. Springer.

Boltz, S., Debreuve, E., and Barlaud, M. (2009). High-dimensional statistical measure for region-of-interest tracking. *Image Processing, IEEE Transactions on*, 18(6):1266–1283.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217.

Brown, E. N., Kass, R. E., and Mitra, P. P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456–461.

Burbea, J. and Rao, C. R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of Multivariate Analysis*, 12(4):575–596.

Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. *Nature neuroscience*, 7(5):446–451.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1.

Chenĥsov, N. N. (1982). *Statistical decision rules and optimal inference*. Number 53. American Mathematical Soc.

Cherkassky, B. V., Goldberg, A. V., and Radzik, T. (1996). Shortest paths algorithms: Theory and experimental evaluation. *Mathematical programming*, 73(2):129–174.

Cichocki, A. and Amari, S.-i. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619.

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68(1):161–186.

Csiszár, I. and Shields, P. C. (2004). Information theory and statistics: A tutorial. *Communications and Information Theory*, 1(4):417–528.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488.

Deza, M. M. and Deza, E. (2009). *Encyclopedia of distances*. Springer.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, pages 1102–1123.

Easterlin, R. A. (2005). Diminishing marginal utility of income? *Social Indicators Research*, 70(3):243–255.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711.

Flury, B. (1997). *A first course in multivariate statistics*. Springer.

Fox, C. (1987). *An introduction to the calculus of variations*. Courier Corporation.

Frey, P. W. and Slate, D. J. (1991). Letter recognition using holland-style adaptive classifiers. *Machine Learning*, 6(2):161–182.

Frigyik, B. A., Srivastava, S., and Gupta, M. R. (2008). Functional bregman divergence and bayesian estimation of distributions. *Information Theory, IEEE Transactions on*, 54(11):5130–5139.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435.

Goria, M. N., Leonenko, N. N., Mergel, V. V., and Novi Inverardi, P. L. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Nonparametric Statistics*, 17(3):277–297.

Gower, J. C. (2006). *Similarity, Dissimilarity and Distance Measures*. Wiley Online Library.

Gower, J. C. and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of classification*, 3(1):5–48.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. (2006). A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Hagedoorn, M. and Veltkamp, R. C. (1999). Reliable and efficient pattern matching using an affine invariant metric. *International Journal of Computer Vision*, 31(2-3):203–225.

Hayden, D., Lazar, P., Schoenfeld, D., Inflammation, the Host Response to Injury Investigators, et al. (2009). Assessing statistical significance in microarray experiments using the distance between microarrays. *PloS one*, 4(6):e5838.

Hovenkamp, H. (1990). Marginal utility and the coase theorem. *Cornell L. Rev.*, 75:783–1426.

Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K., and Sustik, M. A. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, pages 2330–2338.

Ikegaya, Y. (2004). Functional multineuron calcium imaging.

Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., and Lozano, J. A. (2010). Machine learning: an indispensable tool in bioinformatics. In *Bioinformatics methods in clinical research*, pages 25–48. Springer.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Jacques, J. and Preda, C. (2013). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, pages 1–25.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Jones, L. K. and Byrne, C. L. (1990). General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *Information Theory, IEEE Transactions on*, 36(1):23–30.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science*, pages 188–219.

Kylberg, G. (2011). The kylberg texture dataset v. 1.0. *External report (Blue series)*, 35.

Laboratory for Adaptive Intelligence, BSI, R. (2011). Visual grating task.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Latecki, L. J., Lakamper, R., and Eckhardt, T. (2000). Shape descriptors for non-rigid shapes with a single closed contour. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 424–429. IEEE.

Lebanon, G. (2006). Metric learning for text documents. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):497–508.

Listgarten, J. and Emili, A. (2005). Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics*, 4(4):419–434.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693.

Marriott, P. and Salmon, M. (2000). *Applications of differential geometry to econometrics*. Cambridge University Press.

Martos, G., Muñoz, A., and González, J. (2013). On the generalization of the mahalanobis distance. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 125–132. Springer.

Martos, G., Muñoz, A., and González, J. (2014). Generalizing the mahalanobis distance via density kernels. *Journal of Intelligent Data Analysis*, 18(6):19–31.

Meyer-Baese, A. and Schmid, V. J. (2014). *Pattern Recognition and Signal Analysis in Medical Imaging*. Elsevier.

Moguerza, J. M. and Muñoz, A. (2004). Solving the one-class problem using neighbourhood measures. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 680–688. Springer.

Moguerza, J. M. and Muñoz, A. (2006). Support vector machines with applications. *Statistical Science*, pages 322–336.

Moon, Y.-I., Rajagopalan, B., and Lall, U. (1995). Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443.

Muñoz, A., Martos, G., Arriero, J., and González, J. (2012). A new distance for probability measures based on the estimation of level sets. In *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 271–278. Springer.

Muñoz, A., Martos, G., and González, J. (2013). A new distance for data sets in a reproducing kernel hilbert space context. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 222–229. Springer.

Muñoz, A., Martos, G., and González, J. (2015). Level sets based distances for probability measures and ensembles with applications. *arXiv: http://arxiv.org/abs/1504.01664*.

Muñoz, A. and Moguerza, J. M. (2004). One-class support vector machines and density estimation: the precise relation. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 216–223. Springer.

Muñoz, A. and Moguerza, J. M. (2005). A naive solution to the one-class problem and its extension to kernel methods. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 193–204. Springer.

Muñoz, A. and Moguerza, J. M. (2006). Estimation of high-density regions using one-class neighbor machines. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):476–480.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *Information Theory, IEEE Transactions on*, 56(11):5847–5861.

of Information Theory, I. and ASCR, A. (2000). Leaf - tree leaf database.

Orhan, U., Hekim, M., and Ozer, M. (2011). Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, 38(10):13475–13481.

Pennec, X. (2006). Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154.

Phillips, J. M. and Venkatasubramanian, S. (2011). A gentle introduction to the kernel distance. *arXiv preprint arXiv:1103.1625*.

Poggio, T. and Shelton, C. (2002). On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49.

Poggio, T. and Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the AMS*, 50(5):537–544.

Rachev, S. T. and Rüschendorf, L. (1998). The monge-kantorovich problem. *Mass Transportation Problems: Volume I: Theory*, pages 57–106.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. Springer.

Rüschendorf, L. (1985). The wasserstein distance and approximation theorems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 70(1):117–129.

Sagan, H. (2012). *Introduction to the Calculus of Variations*. Courier Corporation.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010a). Functional clustering and alignment methods with applications. *Communications in Applied and Industrial Mathematics*, 1(1):205–224.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010b). K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233.

Scholkopf, B. (2001). The kernel trick for distances. *Advances in neural information processing systems*, pages 301–307.

Schwartz, L. (1957). Théorie des distributions à valeurs vectorielles. i. In *Annales de l'institut Fourier*, volume 7, pages 1–141. Institut Fourier.

Scott, D. W. (2009). *Multivariate density estimation: theory, practice, and visualization*, volume 383. John Wiley & Sons.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., et al. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291.

Si, S., Tao, D., and Geng, B. (2010). Bregman divergence-based regularization for transfer subspace learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(7):929–942.

Simard, P. Y., LeCun, Y. A., Denker, J. S., and Victorri, B. (2012). Transformation invariance in pattern recognition–tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 235–269. Springer.

Slate, D. J. (1991). Letter image recognition data base.

Smith, R., Tawn, J., and Yuen, H. (1990). Statistics of multivariate extremes. *International Statistical Review/Revue Internationale de Statistique*, pages 47–58.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Scholkopf, B., and Lanckriet, G. (2010a). Non-parametric estimation of integral probability metrics. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1428–1432. IEEE.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010b). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561.

Srivastava, A., Jermyn, I., and Joshi, S. (2007). Riemannian analysis of probability density functions with applications in vision. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. (2011). Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*.

Stevenson, I. H. and Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature neuroscience*, 14(2):139–142.

Strichartz, R. S. (2003). *A guide to distribution theory and Fourier transforms*. World Scientific.

Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5.

Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC.

Vakili, K. and Schmitt, E. (2014). Finding multivariate outliers with fastpcs. *Computational Statistics & Data Analysis*, 69:54–66.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Wang, Q., Kulkarni, S. R., and Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *Information Theory, IEEE Transactions on*, 51(9):3064–3074.

Xu, R., Wunsch, D., et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678.

Zemanian, A. H. (1982). Distribution theory and transform analysis.

Zeng, L., Li, L., and Duan, L. (2012). Business intelligence in enterprise computing environment. *Information Technology and Management*, 13(4):297–310.

Zhang, J., Olive, D. J., and Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability*, 1(2):p119.

Zhou, S. K. and Chellappa, R. (2006). From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(6):917–929.

Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387.

Zolotarev, V. M. (1983). Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28(2):264–287.