Support Measure Data Description

Jorge Guevara, Stephane Canu, Roberto Hirata Jr

Support Measure Data Description

Jorge Guevara,Stéphane Canu,and Roberto Hirata,

Abstract—We address the problem of learning a data description model for datasets containing examples given by groups, clusters or sets of points. Specifically, we assume each example containing values drawing from some unknown local probability measure. We found such a description by empirically approximating a minimum volume set in the space of probability measures by means of a minimum enclosing ball in a Reproducing Kernel Hilbert Space of the representer functions of such measures. As a result, the data description model is a function that only depends on some probability measures called support measures. We formulated three data description models for such datasets. The optimization problem for the first one is a chance constrained program. The second and the third models are quadratic programs. We validate our method in the challenging setting of group anomaly detection task.

Index Terms—Kernel on distributions, One-class classification, support vector data description, embedding of probability measures, mean map, group anomaly detection.

1 Introduction

Data description (DD) or One-Class Classification is the task of building models to depict the common characteristics of objects in some data set, with the aim of performing machine learning tasks such as anomaly and novelty detection, clustering and classification [1]–[5] . The main idea of DD methods is to assume an underlying distribution generating the points in the dataset, consequently, most of them extract from training data some distribution information, for instance, an empirically probability density function, a density level set or information of the density support set.

Usually, DD methods work on datasets whose examples, are individual points in some space. However, there is a growing interest in machine learning methods for datasets whose individual examples are clusters, groups or sets of points [6]–[20].

Formally:

$$\mathcal{T} = \{\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{L_i}^{(i)}\}\}_{i=1}^{N},\ \mathbf{x}_{1\leq l\leq L_i}^{(i)} \in \mathbb{R}^D, \quad (1)$$

is a dataset with N examples, where the i example with L_i elements is the set $\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{L_i}^{(i)}\}$. For instance, each example of $\mathcal{T}$ could be: a set of image features in a image dataset [21], a set of spatio-temporal features [22], a set of replicates values for a measurement process [23], a Ellipsoidal or interval set describing point wise uncertainty [12]–[14], a set describing subjective judgments [15], a set describing the invariance of some particular object [16].

- *Jorge Guevara is with the Department of Computer Science, Institute of Mathematics and Statistics, University of Sao Paulo, Sao Paulo, Brazil. E-mail: see http://www.vision.ime.usp.br/ jorjasso/*
- *Stéphane Canu is with the Department of Computer Science, Normandie Université, INSA de Rouen - LITIS, St Etienne du Rouvray,France. E-mail: scanu@insa-rouen.fr*
- *Roberto Hirata is with the Department of Computer Science, Institute of Mathematics and Statistics, University of Sao Paulo, Sao Paulo, Brazil. E-mail: see hirata@ime.usp.br*

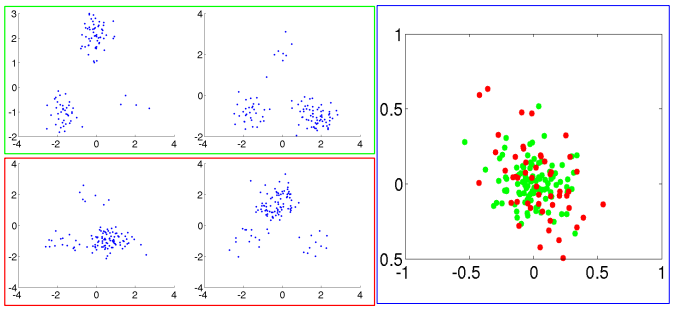


Fig. 1: The Green box contains 2 groups (examples) of non-anomalous groups. The red box contains two anomalous groups. Blue box on the right shows that anomalous group means (red points) are highly mixed with non-anomalous group means (green points).

Group anomaly detection [6], [7] illustrate the importance of finding the description of datasets given by (1). Anomalous groups to be detected are formed by the aggregation of anomalous points, or inclusively, by anomalous aggregation of non-anomalous points. Figure 1 shows two anomalous groups of points (red box) and two non-anomalous groups of points (green box), it is easy to see that both anomalous and non-anomalous groups of points overlap each other. Also, left part of Figure 1 shows an overlapping between the means of anomalous (red points) and non-anomalous groups (green points) for this example. As conventional DD methods often consider anomalies to points far away from the description of the data, such methods can fail if the data description is found only using some representative values per each group, for instance the group mean or the group median. What's more important, if the nature of the data is to contain examples given by sets of points, a preprocessing step for reducing each set of points to a single value will turn conventional

anomaly detection methods, highly depended of such a procedure, moreover, by doing that, useful information could be discarded.

The aim of this work is to find a description of datasets given by (1). We follow the assumption that elements of each example, i.e., $\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{L_i}^{(i)}\}$, are i.i.d [1] realizations: of a random variable X distributed according to some unknown local probability measure $\mathbb{P}_i$ defined on the measurable space $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$ with $\mathcal{B}(\mathbb{R}^D)$ denoting the Borel σ-algebra of $\mathbb{R}^D$. Then the problem of describing datasets given by (1) is posed as finding a description for the (unknown) local probability measures $\{\mathbb{P}_i\}_{i=1}^N$, such a description is found it by an empirical minimum volume set estimator for $\{\mathbb{P}_i\}_{i=1}^N$. We generalize the definition of minimum volume set [1]–[4] to the case of probability measures as follows.

Definition 1.1 (Minimum Volume Set)**.** Let $(\mathcal{P}, \mathcal{A}, \mathcal{E})$ be a probability space, where $\mathcal{P}$ is the space of all probability measures $\mathbb{P}$ on $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$, $\mathcal{A}$ is some suitable σ-algebra of $\mathcal{P}$, and $\mathcal{E}$ is a probability measure on $(\mathcal{P}, \mathcal{A})$. The minimum volume set is the set[2] [3]:

$$G_\alpha^* = \inf_G \{\rho(G) | \mathcal{E}(G) \geq \alpha, G \in \mathcal{A}\}, \tag{2}$$

where ρ is a reference measure on $\mathcal{A}$, for instance the Lebesgue measure, and $\alpha \in [0, 1]$. The minimum volume set G_α^*, describes a fraction α of the mass concentration of $\mathcal{E}$ [4].

Here, the class $\mathcal{A}$, can be given by regions or spaces formed by polynomials, ellipsoids, half-spaces [4], [9], enclosing balls [5], etc., see for example [1] and references therein. Using a kernel on probability measures, we consider the class $\mathcal{A}$ implicitly defined by such a kernel as the set of enclosing balls in a Reproducing Kernel Hilbert Space or RKHS.

Consequently, given the i.i.d sample $\{\mathbb{P}_i\}_{i=1}^N$ distributed according to $\mathcal{E}$, we empirically estimate (2) by computing a minimum enclosing ball in a Reproducing Kernel Hilbert Space or RKHS for the set of *representer functions* of the local probability measures $\{\mathbb{P}_i\}_{i=1}^N$. Such representer functions in a RKHS are given by the embedding of probability measures into a proper Hilbert space [18], [26]–[28]. As the local probability measures $\{\mathbb{P}_i\}_{i=1}^N$ are unknown, Hilbert space embedding of probability measures gives a way to compute a inner product in $\mathcal{P}$ *without* computing the density of such local distributions, such a inner product is computed by means of a real-valued positive definite kernel defined on $\mathcal{P} \times \mathcal{P}$. Moreover, a good approximation is assured by an empirically estimation of the kernel $\mathcal{P} \times \mathcal{P}$ using (1) [18]. Consequently, in the same way of kernel methods, the description of $\{\mathbb{P}_i\}_{i=1}^N$ is a function that only depends on some training examples called as *support measures*.

We consider three *Support Measure Data Description* or SMDD Models. The first one found the description by solving an optimization problem with chance constraints in the RKHS. The second model only use representer functions of probability measures, and the third model uses an scaling of data and translation invariant kernels. We show through extensive experiments, the behavior of such models, also the models are tested in the group anomaly detection task using artificial and real world datasets. All the three models does not assume any particular form for the density of local probability measures.

We begin by formulate a Chance-constrained programm in the space of probability measures for the first SMMD model in Section 2. Such formulation is further kernelized in Section 3, this section present two more additional SMDD models as well. The relationship among all the SMDD models is presented in 4. Results of experiments are showed in Section 5. Finally, the conclusions are given in Section 6.

Notation. We consider a random vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as a Borel measurable map from $\Omega \to \mathbb{R}^D$, satisfying $X(\omega) = \omega, \ \forall \omega \in \Omega$, i.e, X is a identity map. Also, we always consider $\Omega = \mathbb{R}^D$ and $\mathcal{F} = \mathcal{B}(\mathbb{R}^D)$, implying that for $B \in \mathcal{B}(\mathbb{R}^D)$ the probability measure induced by X given by $\mathbb{P}_X(B) = \mathbb{P}\{\omega : X(\omega) \in B\}$ equals to the probability measure $\mathbb{P}(B)$, i.e., $\mathbb{P}_X = \mathbb{P}$. We always abbreviate $\{\omega : a < X(\omega) \leq b\}$ by $\{a < X \leq b\}$. Notation $\sim$ means *distributed according to*.

July 01, 2014

2 Support Measure Data Description in the Space of Probability Measures

Given the i.i.d sample $\{\mathbb{P}_i\}_{i=1}^N$, a first definition for a empirical version of the set G in (2) is given by enclosing balls of probability measures $\{\mathbb{P}_i\}_{i=1}^N$:

$$\hat{G} = \{\mathbb{P}_i \in \mathcal{P} \mid \|X_i - \mathbf{c}\|^2 \leq R^2\}, \tag{3}$$

where $R \in \mathbb{R}$, $\mathbf{c} \in \mathbb{R}^D$, and X_i is a random variable distributed according to $\mathbb{P}_i$. The empirical minimum volume set $\hat{G}_\alpha^*$ is found by estimating the minimum enclosing ball $(R^*, \mathbf{c}^*)$ of the sample $\{\mathbb{P}_i\}_{i=1}^N$. In this case, the optimal radius R^* is proportionally to the probability mass α in (2). However, (3) is very conservative, because $\mathbb{P}_i$ is in $\hat{G}$, only if all the possible realizations of $X_i \sim \mathbb{P}_i$ are in the sphere $(R, \mathbf{c})$.

Given the set $\mathcal{K} = \{\kappa_i\}_{i=1}^N, \ \kappa_i \in [0, 1]$, a more flexible formulation for $\hat{G}$ is:

$$\hat{G}(\mathcal{K}) = \{\mathbb{P}_i \in \mathcal{P} \mid \mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2) \geq 1 - \kappa_i\}, \tag{4}$$

because each probability measure is in $\hat{G}(\mathcal{K})$ depending on the associated value κ_i. It is possible to see that if all $\kappa_i = 0$, then (4) reduces to (3), and if for some $\mathbb{P}_i$,

1. Independently and identically distributed.

2. $\mathcal{A}$ is for instance the Borel σ-algebra with respect to the topology of weak convergence [9], [24]

3. Assuming that all Borel probability measures $\mathbb{P} \in \mathcal{P}$ have compact domain.

4. As density level sets [25] are minimum volume sets (the converse is not true [2], [3]), then alternatively (2) can be stated as estimating the *p-level set* of $\mathcal{E}$: $C_p = \{\mathbb{P} \in \mathcal{P} | \mathcal{E}(\mathbb{P}) \geq p\}, \ p \in [0, 1]$,, where the set C_p defines a minimum volume set satisfying that G_1^* correspond to the p-zero level set of $\mathcal{E}$, that is, the density support estimation set of $\mathcal{E}$.

$\kappa_i = 1$, then $\mathbb{P}_i$ is always in $\hat{G}(\mathcal{K})$. Probability measures not considered (or considered) to be part of $\hat{G}(\mathcal{K})$ are those for which the corresponding distribution function $\mathbb{P}_i(\|X_i - c\|^2 \leq R^2)$ is less (or greater) than κ_i.

Finding the empirical minimum volume set given by (4) means to define an optimization problem with chance constraints [29], [30]. Given the set of probability measures $\{\mathbb{P}_i\}_{i=1}^N$, and the probability levels $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0,1]$, such an optimization problem is the following *chance-constrained program*:

$$\begin{aligned} \min_{\mathbf{c}\in\mathbb{R}^D, R\in\mathbb{R}} \quad & R^2 \\ \text{subject to} \quad & \mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2) \geq 1 - \kappa_i, \end{aligned}$$

for all $i = 1, \ldots, N$, where R and $\mathbf{c}$ are the radius and the center of the hypersphere respectively and the random vector $X_i \sim \mathbb{P}_i$ is the uncertainty parameter for the chance-constrained model.

Allowing some probability measures from $\{\mathbb{P}_i\}_{i=1}^N$ not to be in the estimated minimum volume set, we have the following chance-constrained program:

Problem 2.1.

$$\begin{aligned} \min_{\mathbf{c}\in\mathbb{R}^D, R\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2 + \xi_i) \geq 1 - \kappa_i, \\ & \xi_i \geq 0. \end{aligned}$$

for all $i = 1, \ldots, N$, where R and $\mathbf{c}$ are the radius and the center of the hypersphere respectively, $\lambda > 0$ is a regularization parameter, $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_N)$ is a vector of slack variables and the random vector $X_i \sim \mathbb{P}_i$ is the uncertainty parameter for the chance-constrained model.

We call Problem 2.1 as the *Support Measure Data Description* or SMDD formulated as chance-constrained programming in the space of probability measures.

An intuitive interpretation for the chance constraints of Problem 2.1 is that *the probability that the random vector $X_i \sim \mathbb{P}_i$ takes its values outside the sphere of radius $(R, \mathbf{c})$ and error ξ_i is bounded by κ_i*, or in another words, we require that realizations of $X_i \sim \mathbb{P}_i$ lies in the sphere $(R, \mathbf{c})$ with probability greater than $1 - \kappa_i$. Equivalently, the left side of each probabilistic constraint of Problem (2.1) is the distribution function of the random variable[5] $Z_i = \|X_i - \mathbf{c}\|^2$ on the argument $R^2 + \xi_i$, that is, $F_{Z_i}(R^2 + \xi_i) = \mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2 + \xi_i)$, then, Problem 2.1 can be rewriting as:

$$\begin{aligned} \min_{\mathbf{c}\in\mathbb{R}^D, R\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & F_{Z_i}(R^2 + \xi_i) \geq 1 - \kappa_i, \ i = 1, \ldots, N \\ & \xi_i \geq 0, \ i = 1, \ldots, N, \end{aligned}$$

in this formulation, the $1 - \kappa_i$-values are *lower bounds* for the distribution function F_{Z_i}. Then, all $\mathbb{P}_i$, such that $\xi_i = 0$ and $F_{Z_i}(R) = 1 - \kappa_i$, are the support measures (in analogy with support vectors), all $\mathbb{P}_i$, such that $\xi_i > 0$ and $F_{Z_i}(R + \xi_i) = 1 - \kappa_i$, are errors allowed in the training set and, all $\mathbb{P}_i$, such that $\xi_i = 0$ and $F_{Z_i}(R) > 1 - \kappa_i$, are non critical points because they are in the minimum volume set.

Example 2.1. Figure (2a) plots the probability density functions for the training set $\{\mathbb{P}_i\}_{i=1}^5$, where $\mathbb{P}_i = \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$, $\boldsymbol{\mu}_i \in \mathbb{R}^2$, and $\Sigma_i \in \mathbb{R}^{2\times 2}$. Also, the empirical minimum volume set of $\mathcal{E}$ is shown as the red enclosing sphere. As $\mathbb{P}_i$ is normal, then $Z_i \sim \chi^2$ (Chi-square distribution) with one degree of freedom. The particular cases are: probability measures $\mathbb{P}_1$ and $\mathbb{P}_4$ are the *support measures* in $\mathcal{P}$. $\mathbb{P}_5$ is the allowed *error* associated to the slack variable ξ_5. Probability measures $\mathbb{P}_2$ and $\mathbb{P}_3$ are not critical measures. Figure 2b shows the cumulative Chi-square distribution F_{Z_i}, we observe that decreasing κ_i has the effect to increase the radius R to cover a particular $\mathbb{P}_i$, then κ_i values are directly related to the probability mass α in (2). Figure 2c shows five different kappa values for $\mathbb{P}_4$, the values are: $\{1, 00.8, 0.6, 0.4, 0.2\}$, we can see that as κ_i tends to zero, the radius tends to cover $\mathbb{P}_4$. The lower bounds $1 - \kappa_i$ for F_{Z_i} allow to have a more (if κ_i goes to one) or a less conservative (if κ_i goes to zero) model to estimate the minimum volume set.

It is worth to note that Problem 2.1 is is equivalent to SVDD [5] when the probability measures are the probability Dirac measures, i.e., $\mathbb{P}_i = \delta_{\mathbf{x}_i}$, where $\delta_{\mathbf{x}_i}(X_i) = 1$ iff $X_i = \mathbf{x}_i$ and zero otherwise. Then there is certainty with probability one that the only possible realization of $X_i \sim \delta_{\mathbf{x}_i}$ is $\mathbf{x}_i$, this allow us to eliminate the probabilistic constraints and to formulate the problem as the usual SVDD, that is, in this case finding the solution in the input space equals to finding the solution in the space of all probability Dirac measures.

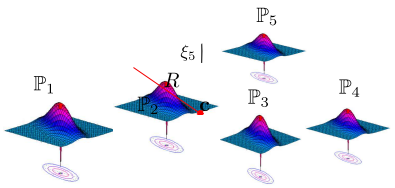
2.1 Formulation by Markov's Inequality

Chance constraints of Problem 2.1 control the probability of constraint violation, achieving flexibility in the model, however, each chance constraints requires to deal with every possible realization of $X \sim \mathbb{P}_i$, then, it is necessary to transform this problem into another one with deterministic constraints, this can be achieved, by using the Markov's inequality, which for a nonnegative random variable $X \sim \mathbb{P}$ and for some $t > 0$, bounds $P(X \geq t)$ by $\mathbb{E}_{\mathbb{P}}[X]/t$.

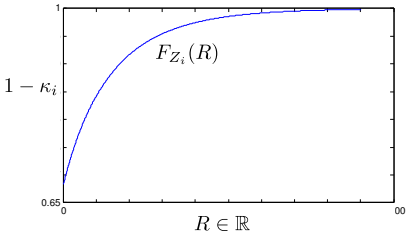
Each chance constraint of Problem 2.1 can be written in equivalent form as $\mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \geq R^2 + \xi_i) \leq \kappa_i$. Assuming that each probability measure $\mathbb{P}_i$ has mean $\boldsymbol{\mu}_i \in \mathbb{R}^D$ and covariance $\Sigma_i \in \mathbb{R}^{D\times D}$, and noting that $\|X_i - \mathbf{c}\|^2 \geq 0$ and $(R^2 + \xi_i) \geq 0$ are satisfied, Markov's inequality bounds each chance constraint as:

$$\mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \geq R^2 + \xi_i) \leq \frac{\mathbb{E}_{\mathbb{P}_i}[\|X_i - \mathbf{c}\|^2]}{R^2 + \xi_i}, \ i = 1, 2 \ldots, N,$$

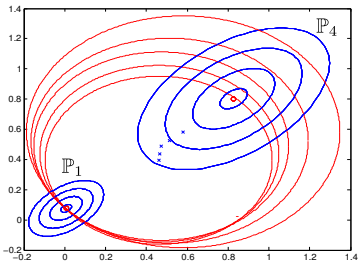
5. A distribution function of a random variable X is the function F_X from $\mathbb{R}$ to $[0,1]$ given by $F_X(x) = \mathbb{P}(\{\omega : X(\omega) \leq x\}) \equiv \mathbb{P}(X \leq x)$



(a) Empirical minimum volume set for $\{\mathbb{P}_i\}_{i=1}^5$.



(b) κ value vs. the radius R.



(c) increasing (decreasing) κ_4 decreases (increases) R to cover $\mathbb{P}_4$.

Fig. 2

where $\mathbb{E}_{\mathbb{P}_i}$ denotes the expectation for a random variable distributed according $\mathbb{P}_i$.

Lema 2.1. *Let $\mathbb{P}$ be a probability measure with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then for $X \sim \mathbb{P}$*

$$\mathbb{E}_{\mathbb{P}}[\|X-\mathbf{c}\|^2] = tr(\Sigma) + \|\boldsymbol{\mu}-\mathbf{c}\|^2.$$

The proof is the Appendix A. Applying Lemma (2.1) and Markov's inequality to the chance constraints of Problem (2.1) yields:

$$\mathbb{P}_i(\|X_i-\mathbf{c}\|^2 \geq R^2+\xi_i) \leq \frac{tr(\Sigma_i)+\|\boldsymbol{\mu}_i-\mathbf{c}\|^2}{R^2+\xi_i},$$

$\forall i = 1, 2 \ldots, N$. As κ_i is the upper bound for the chance constraint i, it is necessary to ensure that

$$\frac{tr(\Sigma_i)+\|\boldsymbol{\mu}_i-\mathbf{c}\|^2}{R^2+\xi_i} \leq \kappa_i. \tag{5}$$

Using (5) and given the set of probability measures $\{\mathbb{P}_i\}_{i=1}^N$, the probability levels $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in (0,1]$, the deterministic form of Problem (2.1) is the following:

Problem 2.2.

$$\begin{aligned} \min_{\mathbf{c}\in\mathbb{R}^D, R\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N} \quad & R^2+\lambda\sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \|\boldsymbol{\mu}_i-\mathbf{c}\|^2 \leq (R^2+\xi_i)\kappa_i - tr(\Sigma_i), \\ & \xi_i \geq 0, \end{aligned}$$

for all $i = 1, \ldots, N$, where R and $\mathbf{c}$ are the radius and the center of the hypersphere respectively, $\lambda > 0$ is a regularization parameter, $\boldsymbol{\xi} = (\xi_1, \xi_1, \ldots, \xi_N)^\top$ is a vector of slack variables, $\boldsymbol{\mu}_i \in \mathbb{R}^D$ and $\Sigma_i \in \mathbb{R}^D \times \mathbb{R}^D$ are the mean and covariance of $\mathbb{P}_i$ respectively. Problem (2.2) is named as *SMDD with joint constraints* if $\kappa_i = \kappa$ for all $i \in 1, 2, \ldots, N$.

Lema 2.2. *If there is no information about Σ_i, and $\kappa_i = 1$, $\forall i$ then, SMDD (Problem (2.2)) is equivalent to a SVDD with $\boldsymbol{\mu}_i$ instead of $\mathbf{x}_i$.*

Proof: By hypothesis, $tr(\Sigma_i) = 0$, replacing $\kappa_i = 1, \forall i$ in Problem 2.2 we get the SVDD [5] with $\boldsymbol{\mu}_i$ instead of $\mathbf{x}_i$. □

2.1.1 Geometric Interpretation

Assuming $\xi_i = 0$, $1, 2, \ldots, N$, for each constraint, we have:

$$\begin{aligned} R\sqrt{\kappa_i} &\geq \sqrt{\|\boldsymbol{\mu}_i-\mathbf{c}\|^2 + tr(\Sigma_i)} \\ &= \|\boldsymbol{\mu}_i-\mathbf{c}\| + \sqrt{tr(\Sigma_i)} - \gamma_i,\ \gamma_i \in \mathbb{R}^+ \end{aligned}$$

where the last equation comes from the fact that for all $a, b \in \mathbb{R}^+ \cup \{0\}$, $\sqrt{a^2+b^2} = \sqrt{(a+b)^2-2ab} \leq a+b$, then follows: $\exists \gamma \in \mathbb{R}^+$ such $\sqrt{a^2+b^2} = a+b-\gamma$ and replacing $a = \|\boldsymbol{\mu}-\mathbf{c}\|$ and $b = \sqrt{tr(\Sigma)}$. See Figure (3).

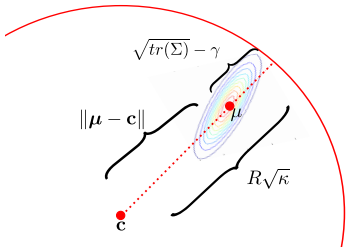


Fig. 3: Geometric interpretation of the deterministic constraints.

2.1.2 Dual Formulation

Denote by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the Lagrange multiplier vectors with nonnegative components α_i and β_i, $i = 1, 2, \ldots, N$, respectively. The Lagrangian for the SMDD is:

$$\begin{aligned} \mathcal{L}(R, \mathbf{c}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) =& R^2 + \lambda\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i\{(R^2+\xi_i)\kappa_i \\ & -\|\boldsymbol{\mu}_i-\mathbf{c}\|^2\} - tr(\Sigma_i)\} - \sum_{i=1}^N \beta\xi_i \end{aligned}$$

The stationarity (6) and complementarity (7) Karush-Kuhn-Tucker (KKT) conditions for this problem are:

$$\left.\begin{array}{rcc} \partial_R\mathcal{L} = 0 & : & \sum_{i=1}^N \alpha_i\kappa_i = 1 \\ \nabla_{\mathbf{c}}\mathcal{L} = 0 & : & -2\sum_{i=1}^N \alpha_i\boldsymbol{\mu}_i + 2\sum_{i=1}^N \alpha_i\mathbf{c} = 0 \\ \partial_{\xi_i}\mathcal{L} = 0 & : & \lambda - \alpha_i\kappa_i - \beta_i = 0 \end{array}\right\} \tag{6}$$

$$\left.\begin{array}{c} \alpha_i\{(R^2+\xi_i)\kappa_i - \|\boldsymbol{\mu}_i-\mathbf{c}\|^2 - tr(\Sigma_i)\} = 0 \\ \beta_i\xi_i = 0 \end{array}\right\} \tag{7}$$

Replacing, the KKT's condition in the Lagrangian, we obtain the dual problem as follows:

Given the set of probability measures $\{\mathbb{P}_i\}_{i=1}^N$, and the probability levels $\{\kappa_i\}_{i=1}^N,\ \kappa_i \in [0,1]$, the dual form of Problem (2.2) is given by

Problem 2.3.

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i\langle\boldsymbol{\mu}_i,\boldsymbol{\mu}_i\rangle - \frac{\sum_{i,j=1}^N \alpha_i\alpha_j\langle\boldsymbol{\mu}_i,\boldsymbol{\mu}_j\rangle}{\sum_{i=1}^N\alpha_i} \\ & + \sum_{i=1}^N \alpha_i tr(\Sigma_i) \\ \text{subject to} \quad & 0\le\alpha_i\kappa_i\le\lambda,\ i=1,\dots,N \\ & \sum_{i=1}^N\alpha_i\kappa_i = 1\end{aligned}$$

where $\lambda > 0$ is a regularization parameter, $\boldsymbol{\alpha}\in\mathbb{R}^N$ are the Lagrangian multipliers $\boldsymbol{\mu}_i\in\mathbb{R}^D$ is the mean of $\mathbb{P}_i$ and $\Sigma_i\in\mathbb{R}^D\times\mathbb{R}^D$ is the covariance matrix of $\mathbb{P}_i$.

From stationary conditions (6), the Representer Theorem [31] for $\mathbf{c}$ is:

$$\mathbf{c} = \frac{\sum_i\alpha_i\boldsymbol{\mu}_i}{\sum_i\alpha_i},\ i\in\{i|0<\alpha_i\kappa_i\le\lambda\}. \tag{8}$$

Analyzing the complementarity conditions (7) we identify the following cases for all $i=1,2,\dots,N$. See Table (1) for a summary.

- $\alpha_i=0,\beta_i>0 \implies \xi_i=0$, that yields $\|\boldsymbol{\mu}_i-\mathbf{c}\|^2 + tr(\Sigma_i)\le R^2\kappa_i$. All the realizations $\mathbf{x}'$ of $X_i\sim\mathbb{P}_i$. satisfying $\|\mathbf{x}'-\mathbf{c}\|^2 = (\|\boldsymbol{\mu}_i-\mathbf{c}\|^2+tr(\Sigma_i))/\kappa_i,\ \kappa_i\neq 0$ for , $i\in\{i|\alpha_i=0\}$ will be *inside* the hypersphere *no matters the value for* κ_i. All $\mathbb{P}_i,\ i\in\{i|\alpha_i=0\}$ are considered to be in the minimum volume set of $\mathcal{E}$.

- $\alpha_i>0,\beta_i=0 \implies \xi_i>0$, that yields $\|\boldsymbol{\mu}_i-\mathbf{c}\|^2 + tr(\Sigma_i) = (R^2+\xi_i)\kappa_i$. All the realizations $\mathbf{x}'$ of $X_i\sim\mathbb{P}_i$ satisfying $\|\mathbf{x}'-\mathbf{c}\|^2 = (\|\boldsymbol{\mu}_i-\mathbf{c}\|^2+tr(\Sigma_i))/\kappa_i,\ \kappa_i\neq 0$ for $i\in\{i|\alpha_i\kappa_i=\lambda\}$ will be *outside* the hypersphere with probability κ_i. All $\mathbb{P}_i,\ i\in\{i|\alpha_i\kappa_i=\lambda\}$ are considered to be errors allowed in the training set.

- $\alpha_i>0,\beta_i>0 \implies \xi_i=0$ and $0<\alpha_i\kappa_i<\lambda$. From this and (7) we can retrieve the radius

$$R^2 = \frac{\|\boldsymbol{\mu}_i-\mathbf{c}\|^2+tr(\Sigma_i)}{\kappa_i},\ i\in\{i|0<\alpha_i\kappa_i<\lambda\}. \tag{9}$$

All the realizations $\mathbf{x}'$ of $X_i\sim\mathbb{P}_i$ satisfying $\|\mathbf{x}'-\mathbf{c}\|^2 = (\|\boldsymbol{\mu}_i-\mathbf{c}\|^2+tr(\Sigma_i))/\kappa_i,\ \kappa_i\neq 0$ for $i\in\{i|0<\alpha_i\kappa_i<\lambda\}$ will be *on* the surface of the hypersphere. All $\mathbb{P}_i,\ i\in\{i|0<\alpha_i\kappa_i<\lambda\}$ are considered to be in the minimum volume set of $\mathcal{E}$ and are called *support measures*.

Teorema 2.3. *Let η be the Lagrange multiplier of the constraint $\sum_{i=1}^N\alpha_i\kappa_i=1$ of Lagrangian of Problem* (2.3)*, then $R=\sqrt{\eta}$.*

Optimization models with chance constraints in kernel methods were previously studied for the case of input uncertainty. For example, to model input uncertainties, [14] considers bounded uncertainty model for data with additive noise, in [12], [13] is assumed some a priori distribution for uncertainties. All such models are formulated as Second Order Cone Programs. In [17] is considered a Taylor approximation in the RKHS for input uncertainty sets. The main disadvantage of such models is the kernelization step, also some of them require assumptions in the bounds of the moments of the probabilistic constraints or its support.

$\alpha_i=0$	$\beta_i>0$	$\xi_i=0$ $\|\boldsymbol{\mu}_i-\mathbf{c}\|^2+tr(\Sigma_i)\le R^2\kappa_i$ $i\in\{i\|\alpha_i=0\}$
$\alpha_i>0$	$\beta_i=0$	$\xi_i>0$ $\|\boldsymbol{\mu}_i-\mathbf{c}\|^2+tr(\Sigma_i)=(R^2+\xi_i)\kappa_i$ $i\in\{i\|\alpha_i\kappa_i=\lambda\}$
	$\beta_i>0$	$\xi_i=0$ $\|\boldsymbol{\mu}_i-\mathbf{c}\|^2+tr(\Sigma_i)=R^2\kappa_i$ $i\in\{i\|0<\alpha_i\kappa_i<\lambda\}$

TABLE 1: Summarizing table for the analysis of KKT's condition

3 Support Measure Data Description in Reproducing Kernel Hilbert Spaces

In this section we present three different formulations of SMMD in Reproducing Kernel Hilbert spaces or RKHS. All the three formulations are minimum enclosing balls in the RKHS, corresponding to no linear descriptions of the training set $\{\mathbb{P}_i\}_{i=1}^N$. The embedding of probability measures into a RKHS, discussed in Section 3.1, is given by representer functions in the RKHS called *mean maps*, then empirical minimum volume sets in the RKHS are formulated as minimum enclosing balls for mean maps.

The first SMDD model in a RKHS, described in Section 3.2, is the kernelization of the SMDD model presented in the last section, that is, such a SMDD model is formulated as a chance constrained programming, and the deterministic form is found by using the Markov's inequality. This model assumes that each probability measure $\mathbb{P}_i$ has the two first moments.

The second SMMD model in a RKHS, described in Section 3.3, is formulated using the mean map embedding of probability measures of $X\sim\mathbb{P}$ into the RKHS.

The third model, described in Section 3.4, is an extension of the second model requiring that for invariant translation kernels, mean maps have norm one.

Also we present in this section the case when SMDD can be formulated as a quadratic programming problem and a discussion about a connection with One-Class Support Vector Machines.

Notation. Letter $\mathcal{H}$ denotes a Reproducing Kernel Hilbert Space (RKHS) of functions $f:\mathbb{R}^D\to\mathbb{R}$, with positive definite kernel $k:\mathbb{R}^D\times\mathbb{R}^D\to\mathbb{R}$, and norm $\|.\|_{\mathcal{H}}$. Also, notation $k(X_i,.)$, means the mapping $t\to k(X_i,t)$,

with fixed value $X_i \sim \mathbb{P}_i$. Inner products in $\mathcal{H}$ are denoted by $\langle .,. \rangle_{\mathcal{H}}$.

3.1 Hilbert Space Embedding of Probability Measures

Definition 3.1 (Mean map)**.** Let $\mathbb{P}$ be a probability measure and $X \sim \mathbb{P}$. The mean map in $\mathcal{H}$ is the function:

$$\begin{aligned} \mu_{\mathbb{P}} : \mathbb{R}^D &\rightarrow \mathbb{R} \\ t &\mapsto \mu_{\mathbb{P}}(t) = \mathbb{E}_{\mathbb{P}}[k(X,t)], \end{aligned} \tag{10}$$

where $\mathbb{E}_{\mathbb{P}}[k(X,t)] = \int_{\mathbf{x}\in\mathbb{R}^D} k(\mathbf{x},t)d\mathbb{P}(\mathbf{x})$, A sufficient condition guaranteeing the existence of $\mu_{\mathbb{P}}$ in $\mathcal{H}$ is given by assuring that $\mu_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[k(X,X)] < \infty$ and $k(.,.)$ being a measurable function [18], [32], [33], as consequence, the reproducing property $\langle f, \mu_{\mathbb{P}} \rangle = \langle f, \mathbb{E}_{\mathbb{P}}[k(X,.)] \rangle = \mathbb{E}_{\mathbb{P}}[f(X)]$ holds for all $f \in \mathcal{H}$.

Definition 3.2 (Embedding of probability measures on $\mathcal{H}$)**.** The embedding of probability measures $\mathbb{P} \in \mathcal{P}$ on $\mathcal{H}$ is given by the mapping

$$\begin{aligned} \mu : \mathcal{P} &\rightarrow \mathcal{H} \\ \mathbb{P} &\mapsto \mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[k(X,.)] \end{aligned} \tag{11}$$

where $\mathbb{E}_{\mathbb{P}}[k(X,.)] = \int_{\mathbf{x}\in\mathbb{R}^D} k(\mathbf{x},.)d\mathbb{P}(\mathbf{x})$ and $\mathcal{P}$ is the space of probability measures.

Mean maps $\mu_{\mathbb{P}}$ are the representer functions in the RKHS $\mathcal{H}$ for probability measures $\mathbb{P}$. Choosing *characteristic kernels* [33]–[35] for $k(.,.)$, the embedding μ is injective, that is, $\langle \mu_{\mathbb{P}}, f \rangle = \langle \mu_{\mathbb{Q}}, f \rangle$ for all $f \in \mathcal{H}$ implies $\mathbb{P} = \mathbb{Q}$, or equivalently, a positive definite kernel is characteristic if $d(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$, where d is a metric on $\mathcal{P}$. Some examples of characteristic kernels are the Gaussian, Laplacian, inverse multiquadratics, B_{2n+1}-splines kernels, etc [33]. Further, an empirical estimator of $\mu_{\mathbb{P}}$ from the sample $\{x_i\}_{i=1}^M$ drawn i.d.d. from $\mathbb{P}$ assure a good approximation for $\mu_{\mathbb{P}}$, i.e., the term $\|\mu_{\mathbb{P}} - \mu_{emp}\|$, where μ_{emp} is a empirical estimator of $\mu_{\mathbb{P}}$, is bounded [18].

The mapping

$$\begin{aligned} \mathcal{P} \times \mathcal{P} &\rightarrow \mathbb{R} \\ (\mathbb{P},\mathbb{Q}) &\mapsto \langle \mathbb{P},\mathbb{Q} \rangle_{\mathcal{P}} = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \end{aligned} \tag{12}$$

$$\tag{13}$$

defines an inner product on $\mathcal{P}$, where from Fubini's theorem follows that $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \int_{\mathbf{x}\in\mathbb{R}^D} \int_{\mathbf{x}'\in\mathbb{R}^D} k(\mathbf{x},\mathbf{x}')d\mathbb{P}(\mathbf{x})d\mathbb{Q}(\mathbf{x}')$, consequently, the real-valued kernel on $\mathcal{P} \times \mathcal{P}$

$$\begin{aligned} \tilde{k}(\mathbb{P},\mathbb{Q}) =& \langle \mathbb{P},\mathbb{Q} \rangle_{\mathcal{P}} = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ =& \int_{\mathbf{x}\in\mathbb{R}^D} \int_{\mathbf{x}'\in\mathbb{R}^D} k(\mathbf{x},\mathbf{x}')d\mathbb{P}(\mathbf{x})d\mathbb{Q}(\mathbf{x}') \end{aligned} \tag{14}$$

is positive definite [28]. Note that, $\tilde{k}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[\mathbb{E}_{\mathbb{Q}}[k(X,X')]]$, $X \sim \mathbb{P}, X' \sim \mathbb{Q}$, by virtue of the reproducing property.

Hilbert space embedding of signed measures was introduced in [26] and studied by [27], [28], and by [18] when the measures are probability measures. Some applications in machine learning include, dimensionality reduction [36], measuring independence of random variables [37], two-sample test [32], embeddings of Hidden markov Models into RKHS [38], Bayes rule [39], support vector machines [9], [20] among others [33], [35], [40]. The kernel on probability measures can be estimated using (1) without requiring fitting some probabilistic models to such data examples. Another related kernels on distributions which assume probabilistic models for data examples are the Fisher kernel [41], the kernel based on the symmetrized Kullback-Leibler (KL) divergence on distributions [42], the Bhattacharyya kernel [19], and the probability product kernel [43].

3.2 SMMD model in a RKHS as Chance Constrained Problem

Using positive definite kernel functions, a kernelized formulation for the set $\hat{G}$ given by (4) is:

$$\hat{G}(\mathcal{K}) = \{\mathbb{P}_i \in \mathcal{P} \mid \mathbb{P}_i(\|k(X_i,.) - c(.)\|_{\mathcal{H}}^2 \leq R^2) \geq 1 - \kappa_i\},$$

where $c(.) \in \mathcal{H}$, and $R \in \mathbb{R}$. Enclosing balls $(R, c(.))$ are now in the RKHS, which in the input space correspond to nonlinear descriptions of $\{\mathbb{P}\}_{i=1}^N$.

Given the set of probability measures $\{\mathbb{P}_i\}_{i=1}^N$ and the probability levels $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0,1]$, SMDD in the RKHS as chance-constrained program is the following:

$$\begin{aligned} \min_{c(.),R,\boldsymbol{\xi}} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \mathbb{P}_i(\|k(X_i,.) - c(.)\|_{\mathcal{H}}^2 \leq R^2 + \xi_i) \geq 1 - \kappa_i, \\ & \xi_i \geq 0, \end{aligned}$$

for all $i = 1, \ldots, N$, where $\lambda > 0$ is a regularization parameter, $\boldsymbol{\xi} = (\xi_1, \xi_1, \ldots, \xi_N)^\top \in \mathbb{R}^N$ is a vector of slack variables. $\mathcal{H}$ is a RKHS with kernel k and norm $\|.\|_{\mathcal{H}}$. Function $c(.) \in \mathcal{H}$ is the center. Value $R \in \mathbb{R}$ is the radius in $\mathcal{H}$. and the $X_i \sim \mathbb{P}_i$ is a random vector.

Solving the above chance constrained program requires that constrains must satisfy all possible realizations of $X_i \sim \mathbb{P}_i$, which is hard to compute. Instead, it is possible to transform it into a deterministic one by performing the embeddings of probability measures into a RKHS and using Markov's inequality.

Using the same argument of Section 2.1, Markov's inequality also holds in the RKHS:

$$\mathbb{P}_i(\|k(X_i,.) - c(.)\|_{\mathcal{H}}^2 \geq R^2 + \xi_i) \leq \frac{\mathbb{E}_{\mathbb{P}}[\|k(X_i,.) - c(.)\|_{\mathcal{H}}^2]}{R^2 + \xi_i},$$

for all $i = 1, 2, \ldots, N$.

The term $\mathbb{E}_{\mathbb{P}}[\|k(X_i,.) - c(.)\|_{\mathcal{H}}^2$ can be computed using the trace of the covariance operator in $\mathcal{H}$ and mean maps $\mu_{\mathbb{P}}$. The covariance operator in the RKHS $\mathcal{H}$ with kernel k is the mapping $\Sigma^{\mathcal{H}} : \mathcal{H} \rightarrow \mathcal{H}$, such for all $f, g \in \mathcal{H}$ satisfy:

$$\langle f, \Sigma^{\mathcal{H}} g \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[f(X)g(X)] - \mathbb{E}_{\mathbb{P}}[f(X)]\mathbb{E}_{\mathbb{P}}[g(X)],$$

because reproducing property [6]. The covariance operator is then the possible infinite dimensional matrix:

$$\Sigma^{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[k(X,.)k(X,.)^{\top}] - \mathbb{E}_{\mathbb{P}}[k(X,.)]\mathbb{E}_{\mathbb{P}}[k(X,.)]^{\top}. \quad (15)$$

From this, the trace of $\Sigma^{\mathcal{H}}$ can be obtained as: [7]

$$\begin{aligned} tr(\Sigma^{\mathcal{H}}) &= \int_{t\in\mathbb{R}^D} \mathbb{E}_{\mathbb{P}}[k(X,t)k(X,t)^{\top}] \\ &\quad -\mathbb{E}_{\mathbb{P}}[k(X,t)]\mathbb{E}_{\mathbb{P}}[k(X,t)]^{\top}dt \\ &= \mathbb{E}_{\mathbb{P}}[\langle k(X,.)k(X,.)\rangle_{\mathcal{H}}] \\ &\quad -\langle \mathbb{E}_{\mathbb{P}}[k(X,.)], \mathbb{E}_{\mathbb{P}}[k(X,.)]\rangle_{\mathcal{H}} \\ &= \mathbb{E}_{\mathbb{P}}[k(X,X)] - \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}}\rangle_{\mathcal{H}} \end{aligned}$$

Then, using (14), yields

$$tr(\Sigma^{\mathcal{H}}) = \mathbb{E}_{\mathbb{P}}[k(X,X)] - \tilde{k}(\mathbb{P},\mathbb{P}), \quad (16)$$

that is, the trace of a possible infinite dimensional matrix can be computed in terms of kernels evaluations. Consequently, in the same way of Section 2.1, Lemma 2.1 becomes:

Lema 3.1.

$$\mathbb{E}_{\mathbb{P}}[\|k(X,.) - c(.)\|^2_{\mathcal{H}}] = tr(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}} - c(.)\|^2_{\mathcal{H}}.$$

The proof is in the Apendix A.

By a similar analysis of Section 2.1, given the set of probability measures $\{\mathbb{P}_i\}_{i=1}^N$, the probability levels $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in (0,1]$ and using Markov's inequality, the deterministic form of SMDD in the RKHS is the following:

Problem 3.1.

$$\begin{aligned} \min_{c(.)\in\mathcal{H}, R\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N} \quad & R^2 + \lambda\sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \|\mu_{\mathbb{P}_i} - c(.)\|^2_{\mathcal{H}} \leq (R^2+\xi_i)\kappa_i - tr(\Sigma^{\mathcal{H}}_i), \\ & \xi_i \geq 0, \end{aligned}$$

for all $i = 1,\ldots,N$, where $\mu_{\mathbb{P}_i} \in \mathcal{H}$ is the mean map in the RKHS, $tr(\Sigma^{\mathcal{H}}_i)$ is given by Equation (16), $\lambda > 0$ is a regularization parameter, $\boldsymbol{\xi} = (\xi_1, \xi_1, \ldots, \xi_N)^{\top} \in \mathbb{R}^N$ is a vector of slack variables. Function $c(.) \in \mathcal{H}$ is the center and, $R \in \mathbb{R}$ is the radius in $\mathcal{H}$.

The Lagrangian of Problem (3.1) is

$$\begin{aligned} \mathcal{L}(R, c(.), \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = R^2 + \lambda\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i\{(R^2+\xi_i)\kappa_i \\ -\|\mu_{\mathbb{P}_i} - c(.)\|^2_{\mathcal{H}}\} - tr(\Sigma^{\mathcal{H}}_i)\} - \sum_{i=1}^N \beta\xi_i \end{aligned} \quad (17)$$

The stationarity (6) and complementarity (7) Karush-Kuhn-Tucker (KKT) conditions for this problem are:

$$\left.\begin{array}{rcl} \partial_R\mathcal{L} = 0 & : & \sum_{i=1}^N \alpha_i\kappa_i = 1 \\ \nabla_{c(.)}\mathcal{L} = 0 & : & -2\sum_{i=1}^N \alpha_i\mu_{\mathbb{P}_i} + 2\sum_{i=1}^N \alpha_i c(.) = 0 \\ \partial_{\xi_i}\mathcal{L} = 0 & : & \lambda - \alpha_i\kappa_i - \beta_i = 0 \end{array}\right\} (18)$$

$$\left.\begin{array}{rcl} \alpha_i\{(R^2+\xi_i)\kappa_i - \|\mu_{\mathbb{P}_i} - c(.)\|^2_{\mathcal{H}} - tr(\Sigma^{\mathcal{H}}_i)\} &=& 0 \\ \beta_i\xi_i &=& 0 \end{array}\right\} (19)$$

Replacing (18) into (17), the dual problem is obtained as follows: Given the set of probability measures $\{\mathbb{P}_i\}_{i=1}^N$, and the probability levels $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0,1]$, the dual form of Problem (3.1) is given by

Problem 3.2.

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i\langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_i}\rangle_{\mathcal{H}} - \frac{\sum_{i,j=1}^N \alpha_i\alpha_j\langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}\rangle_{\mathcal{H}}}{\sum_{i=1}^N \alpha_i} \\ & + \sum_{i=1}^N \alpha_i tr(\Sigma^{\mathcal{H}}_i) \\ \text{subject to} \quad & 0 \leq \alpha_i\kappa_i \leq \lambda,\ i = 1,\ldots,N \\ & \sum_{i=1}^N \alpha_i\kappa_i = 1 \end{aligned}$$

where $\lambda > 0$ is a regularization parameter, $\boldsymbol{\alpha} = (\alpha_1, \alpha_1, \ldots, \alpha_N)^{\top} \in \mathbb{R}^N$ is a vector of Lagrange multipliers, $\mu_{\mathbb{P}_i} \in \mathcal{H}$ is the mean embedding of $\mathbb{P}_i$ and $tr(\Sigma^{\mathcal{H}}_i)$ is given by (16).

By virtue of (14), $\tilde{k}(\mathbb{P}_i, \mathbb{P}_i) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_i}\rangle_{\mathcal{H}}$, then the dual objective function of Problem 3.2 becomes:

$$\sum_{i=1}^N \alpha_i\tilde{k}(\mathbb{P}_i,\mathbb{P}_i) - \frac{\sum_{i,j=1}^N \alpha_i\alpha_j\tilde{k}(\mathbb{P}_i,\mathbb{P}_j)}{\sum_{i=1}^N \alpha_i} + \sum_{i=1}^N \alpha_i tr(\Sigma^{\mathcal{H}}_i)$$

From KKT's conditions, the Representer Theorem in the RKHS is:

$$c(.) = \frac{\sum_i \alpha_i\mu_{\mathbb{P}_i}}{\sum_i \alpha_i} = \frac{\sum_i \alpha_i\mathbb{E}_{\mathbb{P}_i}[k(X,.)]}{\sum_i \alpha_i}, \quad (20)$$

for all $i \in \{i | 0 < \alpha_i\kappa_i \leq \lambda\}$.

By a similar analysis of Section 2.1.2:

- all $\mathbb{P}_i$, $i \in \{i|\alpha_i = 0\}$ are in the minimum volume set of $\mathcal{E}$,
- all $\mathbb{P}_i$, $i \in \{i|\alpha_i\kappa_i = \lambda\}$ are considered to be the errors allowed in the training set.
- all $\mathbb{P}_i$, $i \in \{i|0 < \alpha_i\kappa_i < \lambda\}$ are the *support measures*.

The radius is retrieved as

$$R^2 = \frac{\|\mu_{\mathbb{P}_i} - c(.)\|^2 + tr(\Sigma^{\mathcal{H}}_i)}{\kappa_i},\ i \in \{i|0 < \alpha_i\kappa_i < \lambda\}. \quad (21)$$

Alternatively, R can be computed by Theorem (2.3) as $R = \sqrt{\eta}$. It is worth to note that using the linear kernel: $k(\mathbf{x},\mathbf{x}') = \langle \mathbf{x},\mathbf{x}'\rangle$, Problem 3.2 is equivalent to Problem 2.3, because, $\tilde{k}(\mathbb{P}_i,\mathbb{P}_j) = \mathbb{E}_{\mathbb{P}_i}[\mathbb{E}_{\mathbb{P}_j}[\langle \mathbf{x},\mathbf{x}'\rangle]] = \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j\rangle$.

6. $\Sigma^{\mathcal{H}}$ is a bounded operator on a separable infinite dimensional Hilbert space and can be represented by an infinite matrix [44].

7. Note that as $\mu_{\mathbb{P}}(X) < \infty$, follows that $tr(\Sigma^{\mathcal{H}}) < \infty$.

3.3 SMDD Model in a RKHS a Direct Approach using Mean Maps

Differently of the SMDD model of the last section that uses mean maps and covariance operators, the SMDD model presented in this section only uses mean maps. This model is a direct extension of the SVDD to deal with probability measures. Using the mean maps $\mu_{\mathbb{P}_i}$, it is possible to define $\hat{G}$ as:

$$\hat{G} = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c(.)\|^2_{\mathcal{H}} \leq R^2\}, \tag{22}$$

then, the empirical minimum volume set $\hat{G}^*_\alpha$ is computed by a minimum enclosing ball for mean maps $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$.

Given the set of probability measures $\{\mathbb{P}_i\}_{i=1}^N$, the SMDD model is the following:

$$\begin{aligned} \min_{c(.)\in\mathcal{H}, R\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \|\mu_{\mathbb{P}_i} - c(.)\|^2_{\mathcal{H}} \leq R^2 + \xi_i, i = 1, \ldots, N \\ & \xi_i \geq 0,\ i = 1, \ldots, N. \end{aligned}$$

The Lagrangian for the above Problem is:

$$\begin{aligned} \mathcal{L}(R, c(.), \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) =& R^2 + \lambda \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i \{(R^2 + \xi_i) \\ & - \|\mu_{\mathbb{P}_i} - c(.)\|^2\} - \sum_{i=1}^N \beta \xi_i \end{aligned} \tag{23}$$

The optimality (KKT) conditions for this problem are:

$$\left.\begin{array}{lcl} \partial_R \mathcal{L} = 0 & : & \sum_{i=1}^N \alpha_i = 1 \\ \nabla_{c(.)} \mathcal{L} = 0 & : & -2\sum_{i=1}^N \alpha_i \mu_{\mathbb{P}_i} + 2\sum_{i=1}^N \alpha_i c(.) = 0 \\ \partial_{\xi_i} \mathcal{L} = 0 & : & \lambda - \alpha_i - \beta_i = 0 \end{array}\right\} \tag{24}$$

$$\left.\begin{array}{rcl} \alpha_i\{(R^2 + \xi_i) - \|\mu_{\mathbb{P}_i} - c(.)\|^2\} & = & 0 \\ \beta_i \xi_i & = & 0 \end{array}\right\} \tag{25}$$

Replacing, (24) into (23), we obtain the dual problem as follows:

Given the set of probability measures $\{\mathbb{P}_i\}_{i=1}^N$, the dual form of the previously Problem is given by

Problem 3.3.

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda,\ i = 1, \ldots, N \\ & \sum_{i=1}^N \alpha_i = 1 \end{aligned}$$

where it was used (14) to replace $\langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle$ by the kernel $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)$. From (24) follows that the Representer Theorem is:

$$c(.) = \sum_i \alpha_i \mu_{\mathbb{P}_i}, \quad i \in \{i | 0 < \alpha_i \leq \lambda\}.$$

Analyzing (25) follows that all $\mathbb{P}_i,\ i \in \{i|\alpha_i = 0\}$ are probability measures in the minimum volume set. All $\mathbb{P}_i,\ i \in \{i|\alpha_i = \lambda\}$ are errors. All $\mathbb{P}_i,\ i \in \{i|0 < \alpha_i < \lambda\}$ are support measures.

If the linear kernel: $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$ is used, Problem (3.3) is equivalent to the dual problem of SVDD [5], because, $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \mathbb{E}_{\mathbb{P}_i}[\mathbb{E}_{\mathbb{P}_j}[\langle X, X' \rangle]]$ will be $\langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle$.

3.4 SMDD Model in a RKHS using Mean Maps with Norm One and Invariant Translation Kernels

Translation invariant kernels satisfy:

$$k(x, x) = \langle k(x, .), k(x, .) \rangle_{\mathcal{H}} = \vartheta,\ \forall x \in \mathbb{R}^D$$

where ϑ is a constant value, then immediately follows that $\|k(x, .)\|_{\mathcal{H}} = \sqrt{|\vartheta|}$, that is, functions $k(x, .)$ lie on a surface or radio $\sqrt{|\vartheta|}$. However, this is not the case for mean maps $\mu_{\mathbb{P}} = E_{\mathbb{P}}[k(X, .)]$, because means maps have norm:

$$\|\mu_{\mathbb{P}}\|_{\mathcal{H}} = \|\mathbb{E}_{\mathbb{P}}[k(X.,)]\|_{\mathcal{H}} \leq \mathbb{E}_{\mathbb{P}}[\|k(X.,)\|_{\mathcal{H}}] = \sqrt{|\vartheta|},$$

by convexity of $\|.\|_{\mathcal{H}}$ and Jensen's inequality.

A possible solution to prevent small values for the radius, is to scale mean maps $\mu_{\mathbb{P}}$ to have norm one. to lie on the surface of the hypersphere $(R, c(.))$. The following Theorem is due to Muandet et al [9].

Teorema 3.2 (Spherical Normalization [9])**.** *If kernel $k(., .)$ is characteristic and the examples are linearly independent in the RKHS $\mathcal{H}$, then the spherical normalization :*

$$\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \frac{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}{\sqrt{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}}, \tag{26}$$

preserves the injectivity of the mapping $\mu : \mathcal{P} \to \mathcal{H}$.

Theorem 3.2 says that all the information is preserved after performing spherical normalization on the data.

Consequently, the set $\hat{G}$ is defined as:

$$\hat{G} = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c(.)\|^2_{\mathcal{H}} \leq R^2,\ \|\mu_{\mathbb{P}}\|^2_{\mathcal{H}} = 1\}$$

then, the empirical minimum volume set $\hat{G}^*_\alpha$ is the minimum enclosing ball of the mean maps $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ satisfying $\|\mu_{\mathbb{P}}\|^2_{\mathcal{H}} = 1$, such the kernel used in $\mathbb{E}_{\mathbb{P}_i}[k(X, .)] = \mu_{\mathbb{P}_i}$ is translation invariant.

The optimization problem for this model is

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_i) - \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda,\ i = 1, \ldots, N \\ & \sum_{i=1}^N \alpha_i = 1, \end{aligned}$$

which is the same as Problem 3.3 but with kernel

$$\tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) = \frac{\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)}{\sqrt{\tilde{k}(\mathbb{P}_i, \mathbb{P}_i)\tilde{k}(\mathbb{P}_j, \mathbb{P}_j)}}. \tag{27}$$

As $\sum_{i=1}^N \alpha_i \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_i)$ is constant, formerly Problem can be written as

Problem 3.4.

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & -\sum_{i,j=1}^{N}\alpha_i\alpha_j\tilde{\tilde{k}}(\mathbb{P}_i,\mathbb{P}_j)\\ \text{subject to} \quad & 0\leq\alpha_i\leq\lambda,\ i=1,\dots,N\\ & \sum_{i=1}^{N}\alpha_i=1,\end{aligned}$$

This formulation looks like the dual formulation of One-class Support Vector Machine, but is not directly equivalent. We discuss this point in the next section.

4 Relationship among SMDD models

In this section we point out the relationship among SMDD models, also we discuss when SMDD models are equivalent to a One-class support measure machine (OCSMM) [4], [9]. For this, we call the SMDD presented in Section 3.2 as Model M1 (Problems 3.1 and 3.2), the SMDD of Section 3.3 as Model M2 (Problem 3.3), and the SMDD of Section 3.4 as Model M3 (Problem 3.4).

Teorema 4.1. *The Primal form of Model M1 (Problem 3.1) with joint constraints sharing the same covariance matrix, i.e, $\kappa_i=\kappa$ and $\Sigma_i=\Sigma$ for all $i=1,2\dots,N$ and $\lambda>0$, could be written as*

Problem 4.1.

$$\begin{aligned}\min_{c(.)\in\mathcal{H},\rho'\in\mathbb{R},\boldsymbol{\xi}'\in\mathbb{R}^N} \quad & \frac{\|c(.)\|^2_{\mathcal{H}}}{2}-\rho'+\lambda\sum_{i=1}^{N}\xi'_i\\ \text{subject to} \quad & \langle\mu_{\mathbb{P}_i},c(.)\rangle_{\mathcal{H}}\geq\rho'-\xi'_i,\ i=1,\dots,N\\ & \xi'_i\geq-\frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2},\ i=1,\dots,N.\end{aligned}$$

where

$$\xi'_i=\frac{1}{2}\kappa\xi_i-\frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2} \tag{28}$$

Proof is in the Apendix §A. Problem 4.1 is a less flexible formulation of Model M1 (Problem 3.1), because it considers the same local covariance for all points, and the same κ values. It is easy to verify that, for optimal $c(.)$ and ρ' values from 4.1. [8].

$$R=\sqrt{(tr(\Sigma)+\|c\|^2-2\rho'))/\kappa}, \tag{29}$$

or equivalently, solving Problem (3.1) for $\kappa_i=\kappa$ and $\Sigma_i=\Sigma$ for all $i=1,2\dots,N$, we can retrieve ρ' of Problem 4.1 as follows:

$$\rho'=-\frac{1}{2}(R^2\kappa-tr(\Sigma)-\|c\|^2).$$

Teorema 4.2. *Using the kernel between probability measures given by* (14)*, the dual of Problem* (4.1) *is given by:*

8. See proof of Theorem 4.1 in the Apendix §A as an example

Problem 4.2.

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \frac{1}{2}\sum_{i=1}^{N}\alpha_i\tilde{k}(\mathbb{P}_i,\mathbb{P}_i)-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j\tilde{k}(\mathbb{P}_i,\mathbb{P}_j)\\ \text{subject to} \quad & 0\leq\alpha_i\leq\lambda,\ i=1,\dots,N\\ & \sum_{i=1}^{N}\alpha_i=1.\end{aligned}$$

The proof is in the Appendix A.

Lema 4.3. *Let η be the Lagrange multiplier of constraint $\sum_{i=1}^{N}\alpha_i=1$ of Lagrangian of problem* (4.2)*, then $\rho=\eta$.*

From this, we can solve Problem (4.2) and apply Lemma 4.3 to retrieve ρ, the center via the Representer Theorem given by (35) as $c(.)=\sum_i\alpha_i\mu_{\mathbb{P}_i},\ i\in\{i|0<\alpha_i\leq\lambda\}$, and the radius R from (29).

From this, follows that Model M1 with joint constraints sharing the covariance matrix (Problem 4.2) is equivalent to Model M2 (Problem 3.3) with the only difference of a scaling factor of 0.5 of the dual objective function. If we apply spherical normalization on data, then the dual objective of Problem 4.2 becomes: $-0.5\sum_{i,j=1}^{N}\alpha_i\alpha_j\tilde{\tilde{k}}(\mathbb{P}_i,\mathbb{P}_j)$, where $\tilde{\tilde{k}}$ is the kernel given by (27), consequently, model M1 with this setting is equivalent to Model M3 (Problem 3.4) also with the difference of a scaling factor of 0.5 in the dual objective function.

4.1 Connection with One-Class Support Vector Machines

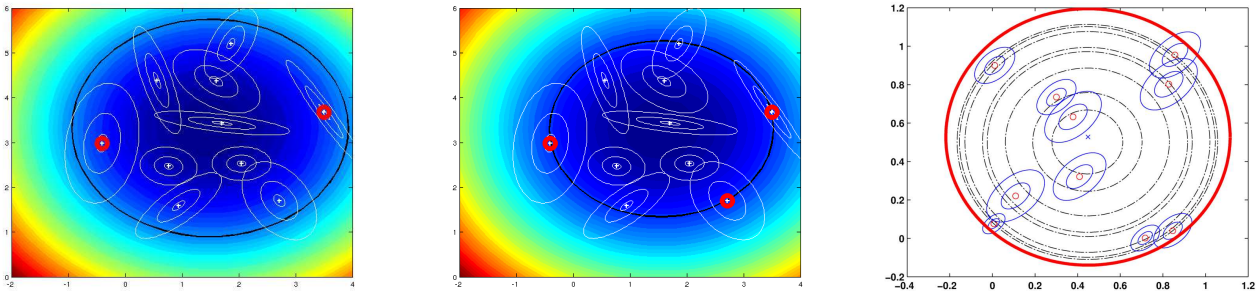
SVDD [5] and one-class support vector machines (OCSVM) [4] are equivalent if translation invariant kernels are used [4], [5]. Although, Problem 4.1 is pretty similar to OCSVM with probability measures [4], [9], SMDD is not directly equivalently with one-class support vector machines, because even if an invariant translation kernel is used, norms of mean maps are not constant. However, if is performed spherical normalization on data there is the following equivalence:

Corollary 4.4. *If it is performed spherical normalization on the training set $\{\mathbb{P}_i\}_{i=1}^{N}$ by* (3.2) *then, Problems 4.2, 3.3, 3.4, and the OCSMM [9] are equivalents.*

Proof: After a spherical normalization $\|\mu_{\mathbb{P}_i}\|^2=1$ holds, then, if a translation invariant kernel is used, then $\sum_{i=1}^{N}\alpha_i\tilde{k}(\mathbb{P}_i,\mathbb{P}_i)$ is constant, consequently such problems are equivalent. □

5 Experiments

Through this section, we called Model M1 to the SMDD defined by 3.2, Model M2 by 3.3, and Model M3 by Problem 3.4. Also, for comparison purposes we called model M4 to the OCSMM [9], and Model M5 to the SVDD [5]. We treat Model M5 as the baseline for the experiments. Table 2 shows a summary of all those models. Also we called as *group* or *cluster* to each set i of



(a) (b) (c)

Fig. 4: Minimum enclosing balls from Models M1 and M2 with linear kernel for (14) and regularization parameter $\lambda = 1$ showed in (a) and (b). Minimum enclosing balls for several choices of λ for Model M5 (dashdot black circles) and a minimum enclosing ball from Model M5 with $\lambda = 1$ (red circle) showed in (c). Small Red points indicate support measures.

points of (1), i.e., training sets has N groups, clusters or set of points. As Model M5 is not designed to deal with probability measures, it was trained using the empirical group means.

The kernel between probability measures given by (14) was estimated via the empirical estimator:

$$\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \approx \frac{1}{L_i L_j} \sum_{l=1}^{L_i} \sum_{l'=1}^{L_j} k(\mathbf{x}_l^{(i)}, \mathbf{x}_{l'}^{(j)}), \tag{30}$$

on the training set given by (1). Also, the trace of the covariance in the RKHS: $tr(\Sigma^{\mathcal{H}})$ given by (16) was estimated by:

$$\begin{aligned} tr(\Sigma_i^{\mathcal{H}}) \approx & \frac{1}{L_i - 1} \sum_{l=1}^{L_i} k(\mathbf{x}_l^{(i)}, \mathbf{x}_l^{(i)}) \\ & - \frac{1}{L_i(L_i - 1)} \sum_{l=1}^{L_i} \sum_{l=1}^{L_i} k(\mathbf{x}_l^{(i)}, \mathbf{x}_l^{(i)}). \end{aligned} \tag{31}$$

where $\mathbb{R}^D \times \mathbb{R}^D$ is a positive definite kernel.

Model	Problem	Section/Ref.
M1	3.2	3.2
M2	3.3	3.3
M3	3.4	3.4
M4	OCSMM	[9]
M5	SVDD	[5]

TABLE 2: Models used in experiments

To solve Model M1, we used *CVX*, a package for specifying and solving convex programs [45], [46]. To solve Models M2, M3, M4 and M5 we used the *SVM and Kernel Methods Matlab Toolbox* (SVM-KM) [47]. The Matlab code and datasets for experiments can be found here: http://www.vision.ime.usp.br/~jorjasso/SMDD.html

5.1 SMDD models with linear kernel

We start by exploring the behaviour of SMDD models when the linear kernel is used. From (14) with linear kernel, follows, that $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle = \langle \mathbb{E}_{\mathbb{P}_i}[X], \mathbb{E}_{\mathbb{P}_j}[X] \rangle = \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle$, where $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$ are the means of groups i and j in the training set respectively, then it is easy to see that Model M2 is the same as a SVDD trained on the group means. As Model M3 is defined only for translation invariant kernels, such a model it is not discussed here.

Figures 4a and 4b shows Model M1 and Model M2 respectively, with regularization parameter $\lambda = 1$, and linear kernel to implement (14). The training set was given by 10 groups of points, each of them artificially generated from a two dimensional Gaussian distribution. The individual ellipsoids plotted are the contours of the estimated probability density functions of the distributions. such the inner and outer ellipses cover about the 35% and 85% of the probability mass of each local probability measure respectively. Small red circles designate the means of the support measures. Figure 4a shows a minimum enclosing ball found by Model M1, the radius does not cross the means of the support measures. because this model beside to use mean maps, also uses the trace of the covariance operator. Figure 4b shows a minimum enclosing ball found by a Model M2, this model uses only information given by mean maps, consequently, the radius cross the means of the support measures. Generally, the radius found by Model M1 is greater than the radius found by Model M2. The only case, where both radius are equal is where the space of probability Dirac measures are considered, that is, no information about the covariance of $\mathbb{P}_i$.

Dashdot black circles in Figure 4c denote several minimum enclosing balls found by a Model M5 (SVDD), each of them obtained by varying the regularization parameter λ to allow $\{0\%, 10\%, 20\%, 30\%, 40 \ldots, 90\%\}$ of the training data to be errors. This was done by set λ proportionally to $1/(N * s\%)$, where N is the number of

examples in the training set and $s\%$ is the percentage of training examples to be considered errors. Red circle in Figure 4c denote the minimum enclosing ball found by Model M1 when the regularization parameter is set to one. that is, all the training set is in the ball and no errors are allowed. The κ-values for Model M1 were all ones. The minimum enclosing ball obtained from Model M1 is bigger than the one obtained from Model M5 in the case of 0% errors allowed, because Model M1 considers information of the trace of local covariances in the training set. The equivalence between Models M1 and M5 is described by Lemma 2.2. That is if all the κ values are ones, and there is no information of local covariance matrices.

5.2 Role of κ-values in Model M1

We explored in this experiment the role of κ-values in the estimation of minimum enclosing balls by Model M1. The training set was generated as before but with the same covariance matrix for all the groups for an easy visualization. It was used the RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$ with $\gamma = 2^0$. The kernel on probability measures (14) was estimated by (30) and the trace by (31). The regularization parameter was $\lambda = 1$.

Left to right of top part of Figure 5 shows three minimum enclosing balls from Model M1 with $\kappa_i = 0.8$, $\kappa_i = 0.9$ and $\kappa_i = 1.0$ for all $i = 1, \ldots, 10$, respectively. Left to right of bottom part of Figure 5 shows another three minimum enclosing balls from Model M1 with all the $\kappa_i = 1$ except that $\kappa_1 = \kappa_2 = 0.8$, $\kappa_1 = \kappa_2 = 0.9$, and $\kappa_1 = \kappa_2 = 1.0$ for each one of the models. Note that depending on particular κ values, the corresponding probability measures, belongs to the minimum volume set in some degree.

As it was point out in Section 3.2, κ-values increases the radius allowing associated probability measures to be in the minimum volume set in some degree. SMDD Models with κ-values close to one are more conservatives in the description of the training set than models with κ-values close to zero.

5.3 Point-Based Group Anomaly Detection over a Gaussian Mixture Distribution data set

The goal of group anomaly detection is to find groups of points with unexpected behavior from datasets given by (1). Differently to usual anomaly detection, points of anomalous groups can be highly mixed with points of non-anomalous groups turning group anomaly detection a challenge problem.

In *Point-Based Group Anomaly* detection [6], anomalous groups are given by aggregating individually anomalous points. Points of non-anomalous groups were randomly sampled from a *Multimodal Gaussian Mixture Distribution* or GMD [7], [9], where the group type distribution was $\pi = (0.48, 0.52)$. The first 48% of non-anomalous groups of points were generated from a two dimensional GMD with three components, mixture weights: $(0.33, 0.64, 0.03)$, means: $(-1.7, -1), (1.7, -1), (0, 2)$, and $0.2 * I_2$ as the sharing covariance matrix, where I_2 denote the 2×2 identity matrix. The another 52% of non-anomalous groups were generated from another GMD with the same parameters but with mixture weights: $(0.33, 0.03, 0.64)$. In total it were generated 50 non-anomalous groups of points as the training set. The red box in Figure 6 shows three non-anomalous groups for $\pi = 0.48$ and the yellow box shows two non-anomalous groups for $\pi = 0.52$.

To perform group anomaly detection, it was generated a test set as follows: we generated three different types of anomalous groups. the first type of group anomalies was given by 10 groups of points randomly generated from the normal distribution: $\mathcal{N}((-0.4, 1), I_2)$. The magenta box in Figure 6 shows five anomalous groups of this type. The second type of group anomalies was given by 5 group of points sampled from a GMM with four components, weights: $(0.1, 0.08, 0.07, 0.75)$, means: $(-1.7, -1), (1.7, -1), (0, 2), (0.6, -1)$, and $0.2 * I_2$ as the sharing covariance matrix. The blue box in Figure 6 shows five anomalous groups of this type. The third type of group anomalies was given by 5 group of points sampled from a GMM with four components, weights: $(0.14, 0.1, 0.28, 0.48)$, means. $(-1.7, -1), (1.7, -1), (0, 2), (-0.5, 1)$, and $0.2 * I_2$ as the sharing covariance matrix. The red box in Figure 6 shows five anomalous groups of this type.

The number of points by group for all non-anomalous and anomalous groups was randomly chosen from a Poisson distribution: $n_i \sim$ Poisson(100).

To see why group anomaly detection is a hard problem, the plot of the means of all the non-anomalous and anomalous groups is shown in Figure 7d. Green points are the means of non-anomalous groups. Red, blue, and magenta points are the means of anomalous groups of points corresponding to the red, blue, and magenta boxes in Figure 6. Because the non-anomalous group means overlap the anomalous group means, methods as One-class support vector machines and SVDD will no perform well, because such methods consider anomalies to points far away from the mean of the description of the data.

To get reliable statistics, we performed 200 runs, over a training set of 50 non-anomalous groups and a test set of 20 anomalous groups, plus 10 non-anomalous groups, totalizing a test set of 30 groups. The performance metrics were: the area under the ROC curve (AUC), and the accuracy (ACC).

It was considered a regularization parameter $\lambda = 1$, and a kernel between probability measures (14) implemented by a RBF kernel with bandwidth parameter γ computed as the inverse of the 0.1 quantile of the euclidean distance between all possible pair of points in the dataset. i.e.,

$$\gamma = 1/s(\|\mathbf{x}_k^{(i)} - \mathbf{x}_l^{(j)}\|^2), \tag{32}$$

where s is the 0.1 quantile, i, j are the groups indices,

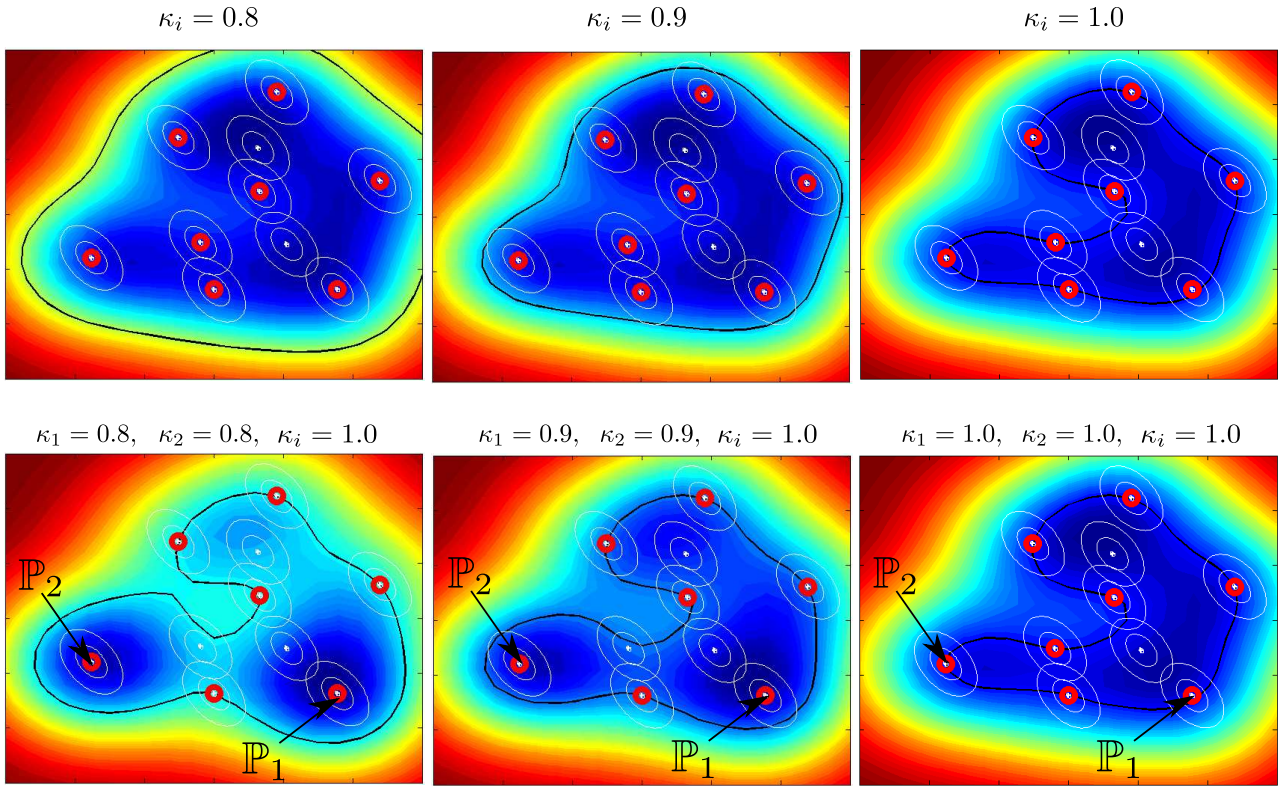


Fig. 5: The effect of different κ-values of Model M1 in the description of a dataset of probability measures.

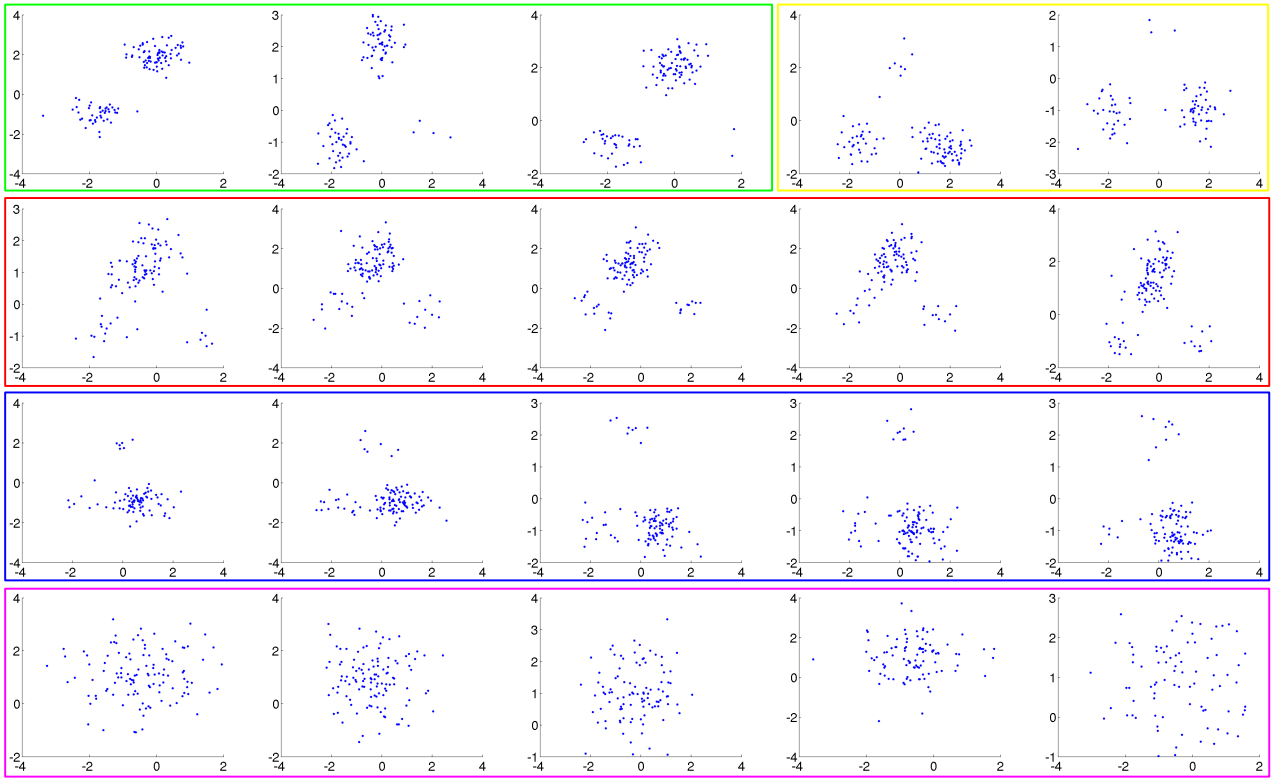


Fig. 6: Group anomaly detection dataset. Green and yellow boxes contains non-anomalous groups. Red, blue, and magenta boxes contains anomalous groups.

and k, l are the points indices.

Figures 7a, 7b, 7c show in boxplots the AUC, the ACC for non-anomalous groups, and the ACC for anomalous groups respectively. The red mark in each boxplot is the median and the edges of each boxplot are the 25th and 75th percentiles, the height of each boxplot is the inter quartile range. This experiment shows that all the SMDD models: Models M1, M2, and M3 detect such anomalies very well. The AUC value close to one of those models indicate that the SMDD models detect group anomalies with few false positives and false negatives. On the other hand, Model M5 (SVDD) can not detect such group anomalies using only the group means as the training set.

5.4 Distribution-Based Group Anomaly Detection over a Gaussian Mixture Distribution data set

Distribution-Based Group Anomalies [6] are anomalous groups of points that are individually non-anomalous but together form anomalous groups. For this experiment, points in each non-anomalous group, were sampled from a two dimensional *unimodal GMD* with three components, mixture weights: $p = \{0.33, 0.33, 0.33\}$, means: $(-1.7, 1), (1.7, -1), (0, 2)$, and sharing the same covariance matrix: $0.2 * I_2$. It was generated 50 non-anomalous groups of points to form the training set (1).

Anomalous groups were generated from the same GMD than non-anomalous groups, but with two of their covariance matrices rotated 45 degrees. that is, individually the points are relatively normal but together as a group are anomalous. The number of points by all the groups was randomly chosen from a Poisson distribution: $n_i \sim$ Poisson(100). In total it was generated 15 anomalous groups for the test set.

The performance of all the models was obtained using the same setup of the last experiment.

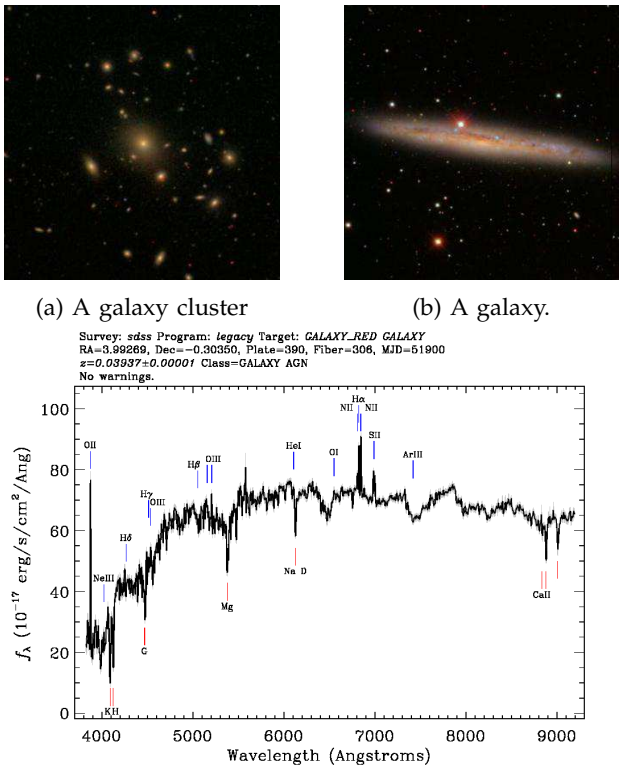
It was considered a regularization parameter $\lambda = 1$, and a kernel between probability measures (14) implemented by a RBF kernel with bandwidth parameter γ computed as the inverse of the median of the euclidean distance between all possible pair of points in the dataset.

Figure 7h shows the group means, green points are the non-anomalous groups means and red points are the anomalous groups means. As in the last experiment, to find such group anomalies is hard because the overlapping between the group means of the non-anomalous and anomalous groups.

Figures 7e, 7f, 7g show in boxplots the AUC, the ACC for non-anomalous groups, and the ACC for anomalous groups respectively. Also, for this type of group anomalies all the SMDD models performs very well. As it was expected, performance of Model M5 is the worst because the overlapping of group means.

Finally, for the same training set, we considered another distribution-based group anomalies, ten group anomalies were generated from a GMD with the same parameters of the training set, but with weights: $p = \{0.85, 0.08, 0.07\}$ and another ten group anomalies with weights: $p = \{0.04, 0.48, 0.48\}$. Figure 7l shows the means of the anomalous (red) and non-anomalous groups (green). For this particular setting, it is possible to see that classical methods such as SVDD and one-class support vector machine will perform well based only on the means information. Figures 7i, 7j, 7k show the AUC, the ACC for non-anomalous groups, and the ACC for anomalous groups respectively. Models M2 and M3 perform as SVDD but are more computationally expensive. Performance of Model M1 is affected because it uses first and second moment information for a problem that is easily solved using only information of the group means. However, because the dimensionality of the data, it is very hard to to know beforehand such a information.

5.5 Group Anomaly Detection in Astronomical Data



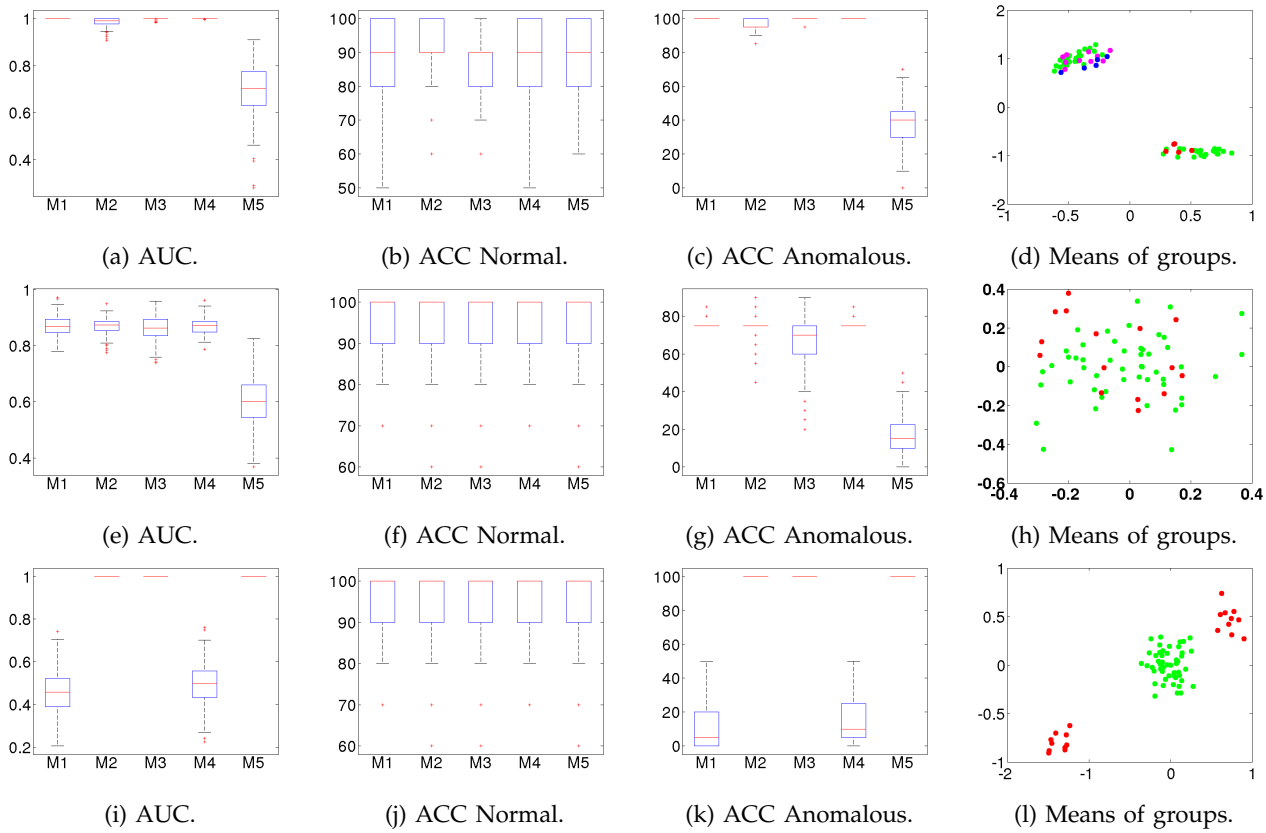
(a) A galaxy cluster (b) A galaxy.

(c) Spectrum of a galaxy. Images from http://www.sdss3.org/

Fig. 8

In this section we tested the SMDD models with real data, for this we used the data from *The Sloan Digital Sky Survey*[9] *(SDSS)* project. SDSS contains massive spectroscopy surveys of the universe, the milky way galaxy, and extrasolar planetary systems. Figure 8 shows a cluster of galaxies, a galaxy, and the measure of light at

9. http://www.sdss3.org/



(a) AUC. (b) ACC Normal. (c) ACC Anomalous. (d) Means of groups.

(e) AUC. (f) ACC Normal. (g) ACC Anomalous. (h) Means of groups.

(i) AUC. (j) ACC Normal. (k) ACC Anomalous. (l) Means of groups.

Fig. 7: Boxplots show the ACC and AUC statistics for group anomaly detection for three experiments, each row is a experiment. Marks on the x-axis of each boxplot represents the models: SMDDCCP (M1), SMDDDA (M2), SMDDDA with spherical normalization (M3), OCSMM (M4), and SVDD (M5). Also in the last column (Figures 7d, 7h, and 7l) plots the means of non-anomalous groups vs. the means of anomalous groups for the experiments considered.

different wavelengths or spectra of a galaxy from SDSS data [10].

In this experiment we use a SDSS dataset of galaxies to detect anomalous clusters of galaxies, such problem, using this SDSS dataset, had been previously studied in [7]–[9] as a group anomaly detection problem. The dataset contains about 7×10^5 galaxies, where each galaxy is represented by a 4000-dimensional feature vector representing spectral information. Features vectors were processed as follows [8]: each feature was down sampled to get a 500-dimensional feature vector for galaxy, then, clusters of galaxies were obtained analysing the spatial neighbourhood of galaxies. This procedure gives 505 clusters of galaxies of a total of 7530 galaxies. Each cluster of galaxies is a group of about $10 - 15$ galaxies.

Finally it was applied PCA to the feature vectors of galaxies, to get a 4-dimensional dataset, preserving about 85% of the variance of the data.

To perform group anomaly detection, it was generated five test datasets. The first test dataset contains 50 anomalous and 50 non-anomalous groups. Each anomalous group was constructed by selecting randomly about $n_i \sim$ Poisson(15) galaxies from the galaxies of the non-anomalous groups. Note that the aggregation of such galaxies are anomalous. The 50 non-anomalous groups were randomly chosen from the 505 non-anomalous groups of galaxies from the original dataset. The remaining 455 groups of galaxies were set as the training set.

The second, third, fourth and fifth test sets, contains 100 groups (50 non-anomalous and 50 anomalous groups). The 50 non-anomalous groups in each test set were also randomly chosen from the 505 original non-anomalous groups. To form the 50 anomalous groups it was proceeded as follows: It was empirical estimated the covariance Σ of all the examples (galaxies), It was selected randomly 3 sets of points, each one containing about $n_i \sim$ Poisson (15) galaxies from the non-anomalous groups. Finally, it was computed the mean of each set. With all those values (the three mean values and the covariance matrix) it was constructed a Gaussian Mixture distribution with weights: $p = \{0.33, 0.33, 0.33\}$. Each anomalous groups for the second test set has points from the above Gaussian Mixture Distribution, sharing the covariance matrix Σ with about $n_i \sim$ Poisson (15) points per group Each anomalous group of the third,

10. Galaxies are huge collections of stars, and they come in spiral, elliptical and irregular shapes.

fourth and fifth, test sets were generated also from the same Gaussian Mixture Distribution but with sharing covariance matrix $5 * \Sigma$, $10 * \Sigma$, and $100 * \Sigma$ respectively.

We plotted in Figures 9d, 9h, 9l, 9p,and 9t the group means of the PCA features. Green points are the non-anomalous group means and red points are the anomalous group means. Each individual figure shows the plot of the first vs second dimension (upper-left), second vs third dimension (upper-right), third vs four dimension (bottom-left), and four vs first dimension (bottom-right). Group anomalies for this experiment are hard to detect because the overlapping of group means of non-anomalous groups and anomalous groups.

For all the SMDD models, it was used a kernel between probability measures implemented with the RBF kernel. To get reliable statistics, it was performed 200 runs for each test set to get the AUC, the ACC of non-anomalous and anomalous groups. It was tested all the SMDD models. Figures 9a, 9b, and 9c show the AUC, the ACC for non-anomalous groups (normal groups), and the ACC for the anomalous groups in the first test set. The kernel parameter was computed with (32) but with s being the median. It was considered a regularization parameter allowing about 30% of the non-anomalous groups to be the errors allowed in the training set. Models M2 and M3 performs a little worst detecting group anomalies than Model M5 for this choice of parameters. However, the AUC metric for Model M5 shows that performance for this model is not better than chance. On the another hand Model M1 and M4 performs better than the baseline and both in similar way detecting group anomalies. Note that the ACC for the non-anomalous groups is about 70% because the choice of the regularization parameter. Plot of the group means in 9d shows the hardness of the problem.

Figures 9e, 9f, and 9g show the AUC, the ACC for non-anomalous groups, and the ACC for the anomalous ones in the second test set. The RBF kernel parameter was computed with (32). It was considered a regularization parameter allowing about 20% of the non-anomalous groups to be the errors allowed in the training set. The AUC metric shows that Model M5 performs worst than the another models, and spherical normalization on data increases the performance as it can be seen by the AUC value close to one of Model M3. On the another hand, the accuracy of normal groups is about 80% because the choice of λ. The plot of the group means is shown in Figure 9h.

Figures 9i, 9j, and 9k show the AUC, the ACC for non-anomalous groups (normal groups), and the ACC for the anomalous groups in the third test set. The RBF kernel parameter was computed with (32) and the regularization parameter was set as $\lambda = 1$. The ACC for anomalous groups shows that Model M2 performs worst detecting the group anomalies, however, such a metric is only based on a threshold of 0 for the output of the models (models outputs greater than zero are anomalies, otherwise are considered non-anomalies). The AUC metric shows that for several choices of thresholds all the models performs pretty well as it is shown in Figure 9i. Again the Models with worst performance are Model M5 and Model M2, and spherical normalization has a positive effect, increasing the AUC value close to one of model M3.

Performance metrics for the fourth test set, has similar characteristics than the third set set as it can be seen in Figures 9m, 9n, and 9o, where the AUC, the ACC for non-anomalous groups, and the ACC for the anomalous groups are shown. The RBF kernel parameter and the regularization parameter were the same as the above experiment. Again, spherical normalization has a positive effect in the AUC metric.

As the group means becomes more spread because characteristics of the five test set, AUC metric shows that all the models performs pretty well, nevertheless, model M5 is the model with worst performance. Figures 9q, 9r, and 9s show the AUC, the ACC for non-anomalous groups, and the ACC for the anomalous groups. The regularization paramter was $\lambda = 1$ and the RBF kernel parameter was chosen as (32) but with s being the 0.9 quantile.

6 Conclusion

In this work we presented a data description method for datasets whose individual examples are set of objects. Such a method was given by computing a minimum volume set for the local underlying distributions (probability measures) generating the examples of the dataset. The minimum volume set of such local probability measures was computed by a minimum enclosing ball of the mean maps embeddings of those measures, without requiring an estimation of a probabilistic model for each example. Based on that, we formulate three models, the first one is given by a chance constrained programming problem, which is transformed in an optimization problem with deterministic constraints via the markov's inequality. This model uses mean maps embeddings and the trace of a covariance matrix embedding. The second model is a direct extension of the SVDD method to the case of the mean map embeddings. The third models uses a kernel on probability measures based on a normalization of data. We compared such three models and we showed the case when such models are equivalents.

The presented SMDD models were tested in the challenging group anomaly detection task. We showed empirically that such models performs pretty well for such a task, then SMDD is an alternative methodology to deal with group anomaly detection. All the three SMDD models perform group anomaly detection by describing a region representing normal behavior in the input space given as an enclosing ball in the RKHS for the non-anomalous groups. Groups not belonging to such a region are considered anomalies.

Another important tasks to be addressed include novelty detection, clustering and classification, for datasets

(a) AUC. (b) ACC Normal. (c) ACC Anomalous. (d) Means of groups.

(e) AUC. (f) ACC Normal. (g) ACC Anomalous. (h) Means of groups.

(i) AUC. (j) ACC Normal. (k) ACC Anomalous. (l) Means of groups.

(m) AUC. (n) ACC Normal. (o) ACC Anomalous. (p) Means of groups.

(q) AUC. (r) ACC Normal. (s) ACC Anomalous. (t) Means of groups.

Fig. 9: .

of probability measures, as well as research other possible practical applications on noisy datasets, image datasets, test for multimodality, bioinformatics and considering the possibility of information given by the embedding of another moments.

APPENDIX A
PROOFS

Lema A.1. *Let $\mathbb{P}$ be a probability measure with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then for $X \sim \mathbb{P}$*

$$\mathbb{E}_{\mathbb{P}}[\|X - \mathbf{c}\|^2] = tr(\Sigma) + \|\boldsymbol{\mu} - \mathbf{c}\|^2$$

Proof: Let $X = (X_1, \ldots, X_j, \ldots, X_D)^\top$ and $\mathbf{c} =$

$(c_1, \ldots, c_j \ldots, c_D)^\top$, follows

$$\begin{aligned} E[\|X-\mathbf{c}\|^2] &= E[X^\top X] - 2E[X^\top \mathbf{c}] + \|\mathbf{c}\|^2 \\ &\quad \text{By covariance formula} \\ &= \sum_j^D (cov(X_j, X_j) - \mathbb{E}[X_j]\mathbb{E}[X_j]) \\ &\quad -2\sum_{j=1}^D E[X_j]c_j + \|\mathbf{c}\|^2 \\ &= \sum_j^D (\Sigma)_{jj} + \boldsymbol{\mu}^\top\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\mathbf{c} + \|\mathbf{c}\|^2 \\ &= tr(\Sigma) + \|\boldsymbol{\mu}-\mathbf{c}\|^2 \end{aligned}$$

□

Alternatively, using the expectation of a quadratic form $\mathbb{E}[X^\top \mathbf{A} X] = \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu} + tr(\Sigma)$, for the $N \times N$ matrix $\mathbf{A}$, and replacing $\mathbf{A}$ by the identify matrix $\mathbf{I}$ we have:

$$\begin{aligned} \mathbb{E}[\|X-\mathbf{c}\|^2] &= \mathbb{E}[X^\top X] - 2\mathbb{E}[X^\top\mathbf{c}] + \|\mathbf{c}\|^2 \\ &= \mathbb{E}[X^\top \mathbf{I} X] - 2\boldsymbol{\mu}^\top\mathbf{c} + \|\mathbf{c}\|^2 \\ &= \boldsymbol{\mu}^\top\boldsymbol{\mu} + tr(\Sigma) - 2\boldsymbol{\mu}^\top\mathbf{c} + \|\mathbf{c}\|^2 \\ &= tr(\Sigma) + \|\boldsymbol{\mu}-\mathbf{c}\|^2 \end{aligned}$$

Teorema A.2. *Let η be the Lagrange multiplier of the constraint $\sum_{i=1}^N \alpha_i\kappa_i = 1$ of Lagrangian of Problem* (2.3)*, then $R = \sqrt{\eta}$.*

Proof: The Lagrangian of Problem (2.3) is

$$\begin{aligned} \mathcal{L}(\alpha,\eta,\nu) = &-\sum_{i=1}^N \alpha_i\langle\boldsymbol{\mu}_i,\boldsymbol{\mu}_i\rangle + \frac{\sum_{i,j=1}^N \alpha_i\alpha_j\langle\boldsymbol{\mu}_i,\boldsymbol{\mu}_j\rangle}{\sum_{i=1}^N\alpha_i} \\ &-\sum_{i=1}^N \alpha_i tr(\Sigma_i) - \eta(\sum_{i=1}^N \alpha_i\kappa_i - 1) \\ &-\sum_{i=1}^N \nu_i(\alpha_i\kappa_i - \lambda) \end{aligned}$$

where

$$\begin{aligned} \partial_{\alpha_i}\mathcal{L} = &\frac{(\sum_{j=1}^N\alpha_j)2\langle\boldsymbol{\mu}_i, \sum_{j=1}^N\alpha_j\boldsymbol{\mu}_j\rangle}{(\sum_{j=1}^N\alpha_j)^2} \\ &- \frac{\sum_{i=1}^N\sum_{j=1}^N\alpha_i\alpha_j\langle\boldsymbol{\mu}_i,\boldsymbol{\mu}_j\rangle}{(\sum_{j=1}^N\alpha_j)^2} \\ &- \|\boldsymbol{\mu}_i\|^2 - tr(\Sigma_i) - \eta\kappa_i \\ &- \nu(\kappa_i - c) = 0 \end{aligned} \tag{33}$$

but for complementarity condition of (2.3) follows that $\nu_i = 0$ implies $\alpha_i > 0$, then

$$\partial_{\alpha_i}\mathcal{L} = 2\langle\boldsymbol{\mu}_i,\boldsymbol{c}\rangle - \|\boldsymbol{c}\|^2 - \|\boldsymbol{\mu}_i\|^2 - tr(\Sigma_i) - \eta\kappa_i = 0$$

Using (9), follows that $\eta = R^2$, then $R = \sqrt{\eta}$. □

Lema A.3.

$$\mathbb{E}_{\mathbb{P}}[\|k(X,.) - c(.)\|^2] = tr(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}} - c(.)\|^2_{\mathcal{H}}.$$

Proof:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}[\|k(X,.)-c(.)\|^2] &= \mathbb{E}_{\mathbb{P}}[\langle k(X,.),k(X,.)\rangle_{\mathcal{H}}] \\ &\quad -2\langle\mu_{\mathbb{P}},c(.)\rangle_{\mathcal{H}} + \|c(.)\|^2_{\mathcal{H}} \\ &= tr(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}}\|^2_{\mathcal{H}} \\ &\quad -2\langle\mu_{\mathbb{P}},.c(.)\rangle_{\mathcal{H}} + \|c(.)\|^2_{\mathcal{H}} \\ &= tr(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}} - c(.)\|^2_{\mathcal{H}} \end{aligned}$$

□

Teorema A.4. *SMDD 3.1 with joint constraints sharing the same covariance matrix, i.e, $\kappa_i = \kappa$ and $\Sigma_i = \Sigma$ for all $i = 1, 2 \ldots, N$ and $\lambda > 0$, could be written as*

Problem A.1.

$$\begin{aligned} \min_{c(.)\in\mathcal{H},\rho'\in\mathbb{R},\boldsymbol{\xi}'\in\mathbb{R}^N} \quad & \frac{\|c(.)\|^2_{\mathcal{H}}}{2} - \rho' + \lambda\sum_{i=1}^N \xi'_i \\ \text{subject to} \quad & \langle\mu_{\mathbb{P}_i}, c(.)\rangle_{\mathcal{H}} \geq \rho' - \xi'_i,\ i = 1,\ldots,N \\ & \xi'_i \geq -\frac{\|\mu_{\mathbb{P}_i}\|^2}{2},\ i = 1,\ldots,N. \end{aligned}$$

Proof: Changing variables in Problem 3.1 by $-\rho = \frac{1}{2}(R^2\kappa - tr(\Sigma^{\mathcal{H}}) - \|c(.)\|^2_{\mathcal{H}})$, implies $R^2 = (tr(\Sigma^{\mathcal{H}}) + \|c(.)\|^2_{\mathcal{H}} - 2\rho))/k$, Problem 3.1 becomes:

$$\begin{aligned} \min_{c(.)\in\mathcal{H},R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^N} \quad & \frac{tr(\Sigma^{\mathcal{H}}) + \|c(.)\|^2_{\mathcal{H}} - 2\rho}{k} + \lambda\sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \langle\mu_{\mathbb{P}_i}, c(.)\rangle_{\mathcal{H}} \geq \rho - \frac{1}{2}(\kappa\xi_i - \|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}), \\ & \xi_i \geq 0, \end{aligned}$$

for all $i = 1, \ldots, N$. Setting $\xi'_i = \frac{1}{2}(\kappa\xi_i - \|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}})$ implies $\xi_i = \frac{2\xi'_i + \|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{k}$ and multiplying by $\frac{k}{2}$ the objective function gives:

$$\begin{aligned} \min_{c(.)\in\mathcal{H},R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^N} \quad & \frac{1}{2}(tr(\Sigma^{\mathcal{H}}) + \|c(.)\|^2_{\mathcal{H}} - 2\rho) \\ & + \frac{k}{2}\lambda\sum_{i=1}^N \frac{2\xi'_i + \|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{k} \\ \text{subject to} \quad & \langle\mu_{\mathbb{P}_i}, c(.)\rangle_{\mathcal{H}} \geq \rho - \xi'_i,\ i = 1,\ldots,N \\ & \frac{2\xi'_i + \|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{k} \geq 0,\ i = 1,\ldots,N. \end{aligned}$$

this is simplified to

$$\begin{aligned} \min_{c(.)\in\mathcal{H},R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^N} \quad & \frac{1}{2}(tr(\Sigma^{\mathcal{H}}) + \|c(.)\|^2_{\mathcal{H}} - 2\rho) + \\ & \lambda\sum_{i=1}^N \xi'_i + \frac{1}{2}\lambda\sum_{i=1}^N \|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}} \\ \text{subject to} \quad & \langle\mu_{\mathbb{P}_i}, c(.)\rangle_{\mathcal{H}} \geq \rho - \xi'_i,\ i = 1,\ldots,N \\ & \xi'_i \geq -\frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2},\ i = 1,\ldots,N. \end{aligned}$$

dropping the constant terms, we arrive at

$$\begin{aligned}\min_{c(.)\in\mathcal{H},R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^N} \quad & \frac{\|c(.)\|^2_{\mathcal{H}}}{2} - \rho + \lambda\sum_{i=1}^{N}\xi_i' \\ \text{subject to} \quad & \langle\mu_{\mathbb{P}_i}, c(.)\rangle_{\mathcal{H}} \geq \rho - \xi_i',\ i=1,\ldots,N \\ & \xi_i' \geq -\frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2},\ i=1,\ldots,N.\end{aligned}$$

□

Teorema A.5. *Using the kernel between probability measures given by* (14)*, the dual of Problem* (4.1) *is given by:*

Problem A.2.

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \frac{1}{2}\sum_{i=1}^{N}\alpha_i\tilde{k}(\mathbb{P}_i,\mathbb{P}_i) - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j\tilde{k}(\mathbb{P}_i,\mathbb{P}_j) \\ \text{subject to} \quad & 0\leq\alpha_i\leq\lambda,\ i=1,\ldots,N \\ & \sum_{i=1}^{N}\alpha_i = 1,\end{aligned}$$

Proof: From (28), if:

$$\xi'' = \frac{1}{2}\kappa\xi \implies \xi'' = \xi' + \frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2},$$

then, Problem 4.1 could be written as:

$$\begin{aligned}\min_{c(.)\in\mathcal{H},\rho'\in\mathbb{R},\boldsymbol{\xi}'\in\mathbb{R}^N} \quad & \frac{\|c(.)\|^2_{\mathcal{H}}}{2} - \rho' + \lambda\sum_{i=1}^{N}(\xi_i'' - \frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2}) \\ \text{subject to} \quad & \langle\mu_{\mathbb{P}_i}, c(.)\rangle_{\mathcal{H}} \geq \rho' - \xi_i'' + \frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2}, \\ & \xi_i'' \geq 0,\end{aligned}$$

for all $i=1,\ldots,N$.

The Lagrangian for the previously Problem is:

$$\begin{aligned}\mathcal{L}(c(.),\rho,\boldsymbol{\xi},\boldsymbol{\alpha},-\boldsymbol{\beta}) = & \frac{\|c(.)\|^2_{\mathcal{H}}}{2} - \rho' + \lambda\sum_{i=1}^{N}(\xi_i'' - \frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2}) \\ & - \sum_{i=1}^{N}\alpha_i\{\langle\mu_{\mathbb{P}_i}, c(.)\rangle_{\mathcal{H}} - \rho' + \xi_i'' - \frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2}\} \\ & - \sum_{i=1}^{N}\beta_i\xi''\end{aligned} \tag{34}$$

The optimality (KKT) conditions for this problem are:

$$\left.\begin{array}{rcccc}\partial_\rho\mathcal{L} = 0 & : & \sum_{i=1}^{N}\alpha_i & = & 1 \\ \nabla_{c(.)}\mathcal{L} = 0 & : & c(.) - \sum_{i=1}^{N}\alpha_i\mu_{\mathbb{P}_i} & = & 0 \\ \partial_{\xi_i''}\mathcal{L} = 0 & : & \lambda - \alpha_i - \beta_i & = & 0\end{array}\right\} \tag{35}$$

$$\left.\begin{array}{c}\alpha_i\{\langle\mu_{\mathbb{P}_i}, c(.)\rangle_{\mathcal{H}} - \rho' + \xi_i'' - \frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2}\} = 0 \\ \beta_i\xi_i'' = 0\end{array}\right\} \tag{36}$$

Replacing, (35) into (34) yields $\frac{1}{2}\sum_{i=1}^{N}(\alpha_i-\lambda)\tilde{k}(\mathbb{P}_i,\mathbb{P}_i) - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j\tilde{k}(\mathbb{P}_i,\mathbb{P}_j)$, but $-\frac{\lambda}{2}\sum_{i=1}^{N}\tilde{k}(\mathbb{P}_i,\mathbb{P}_i)$ is constant, then, the dual form of the above Problem is given by:

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \frac{1}{2}\sum_{i=1}^{N}\alpha_i\tilde{k}(\mathbb{P}_i,\mathbb{P}_i) - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j\tilde{k}(\mathbb{P}_i,\mathbb{P}_j) \\ \text{subject to} \quad & 0\leq\alpha_i\leq\lambda,\ i=1,\ldots,N \\ & \sum_{i=1}^{N}\alpha_i = 1,\end{aligned}$$

where we used the kernel between probability measures given by (14). □

ACKNOWLEDGMENTS

This work was partly done while the first author was visiting the Institute National of Science Apliques, Rouen-France. The authors would like to thank to FAPESP grant # 2011/50761-2, CNPq, CAPES, NAP eScience - PRP - USP.

REFERENCES

[1] W. Polonik, "Minimum volume sets and generalized quantile processes," *Stochastic Processes and their Applications*, vol. 69, no. 1, pp. 1 – 24, 1997.

[2] J. N. Garcia, Z. Kutalik, K.-H. Cho, and O. Wolkenhauer, "Level sets and minimum volume sets of probability density functions," *International Journal of Approximate Reasoning*, vol. 34, no. 1, pp. 25 – 47, 2003.

[3] C. Scott and R. D. Nowak, "Learning minimum volume sets." *Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.

[4] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[5] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[6] L. Xiong, B. Póczos, and J. G. Schneider, "Group anomaly detection using flexible genre models," in *NIPS*, 2011, pp. 1071–1079.

[7] L. Xiong, B. Póczos, J. G. Schneider, A. J. Connolly, and J. VanderPlas, "Hierarchical probabilistic models for group anomaly detection," in *AISTATS*, 2011, pp. 789–797.

[8] B. Póczos, L. Xiong, and J. G. Schneider, "Nonparametric divergence estimation with applications to machine learning on distributions," *CoRR*, vol. abs/1202.3758, 2012.

[9] K. Muandet and B. Schoelkopf, "One-class support measure machines for group anomaly detection," in *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*. Corvallis, Oregon: AUAI Press, 2013, pp. 449–458.

[10] J. Guevara, R. Hirata, and S. Canu, "Kernel functions in takagi-sugeno-kang fuzzy system with nonsingleton fuzzy input," in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, 2013, pp. 1–8.

[11] ——, "Positive definite kernel functions on fuzzy sets," in *Fuzzy Systems (FUZZ), 2014 IEEE International Conference on*, 2014.

[12] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *J. Mach. Learn. Res.*, vol. 7, pp. 1283–1314, Dec. 2006.

[13] A. Ben-Tal, S. Bhadra, C. Bhattacharyya, and J. S. Nath, "Chance constrained uncertain classification via robust optimization," *Mathematical programming*, vol. 127, no. 1, pp. 145–173, 2011.

[14] J. B. T. Zhang, "Support vector classification with input data uncertainty," in *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, vol. 17. MIT Press, 2005, p. 161.

[15] R. Viertl, *Statistical Methods for Fuzzy Data*, ser. Wiley Series in Probability and Statistics. Wiley, 2011.

[16] T. Graepel and R. Herbrich, "Invariant pattern recognition by semidefinite programming machines," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2003, p. 2004.

[17] J. Yang and S. Gunn, "Exploiting uncertain data in support vector classification," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, B. Apolloni, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2007, vol. 4694, pp. 148–155.
[18] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A hilbert space embedding for distributions," in *Algorithmic Learning Theory*. Springer, 2007, pp. 13–31.
[19] R. Kondor and T. Jebara, "A kernel between sets of vectors," in *ICML*, 2003, pp. 361–368.
[20] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schlkopf, "Learning from distributions via support measure machines," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 10–18.
[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
[22] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.
[23] L. Wernisch, S. L. Kendall, S. Soneji, A. Wietzorrek, T. Parish, J. Hinds, P. D. Butcher, and N. G. Stoker, "Analysis of whole-genome microarray replicates using mixed models," *Bioinformatics*, vol. 19, no. 1, pp. 53–61, 2003.
[24] T. S. Ferguson, "Prior distributions on spaces of probability measures," *The Annals of Statistics*, pp. 615–629, 1974.
[25] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Machine Learning Research*, vol. 6, pp. 211–232, 2005.
[26] C. Guilbart, "Produits scalaires sur l'espace des mesures," in *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, vol. 15, no. 4. Gauthier-Villars, 1979, pp. 333–354.
[27] C. Suquet *et al.*, "Distances euclidiennes sur les mesures signees et applications a des theoremes de berry-esseen." *Bulletin of the Belgian Mathematical Society Simon Stevin*, vol. 2, no. 2, pp. 161–182, 1995.
[28] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Boston, 2004, vol. 3.
[29] A. Shapiro, D. Dentcheva, and A. P. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2009, vol. 9.
[30] S. W. Wallace and W. T. Ziemba, *Applications of stochastic programming*. Siam, 2005.
[31] B. Schlkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," in *Computational Learning Theory*, ser. Lecture Notes in Computer Science, D. Helmbold and B. Williamson, Eds. Springer Berlin Heidelberg, 2001, vol. 2111, pp. 416–426.
[32] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.
[33] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *The Journal of Machine Learning Research*, vol. 99, pp. 1517–1561, 2010.
[34] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 489–496.
[35] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schlkopf, "Injective hilbert space embeddings of probability measures," in *In COLT*, 2008.
[36] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces," *The Journal of Machine Learning Research*, vol. 5, pp. 73–99, 2004.
[37] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *J. Mach. Learn. Res.*, vol. 6, pp. 2075–2129, Dec. 2005.
[38] L. Song, B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola, "Hilbert space embeddings of hidden markov models," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 991–998.
[39] K. Fukumizu, L. Song, and A. Gretton, "Kernel bayes' rule," *arXiv preprint arXiv:1009.5736*, 2010.
[40] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift."
[41] T. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *Advances in neural information processing systems*, pp. 487–493, 1999.
[42] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Advances in neural information processing systems*, 2003, p. None.
[43] T. Jebara and R. Kondor, "Bhattacharyya and expected likelihood kernels," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 57–71.
[44] L. Debnath and P. Mikusiński, *Hilbert Spaces with Applications*. Elsevier Academic Press, 2005.
[45] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.
[46] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.
[47] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "Svm and kernel methods matlab toolbox," Perception Systmes et Information, INSA de Rouen, Rouen, France, 2005.