Support Measure Data Description

Jorge Guevara, S Ephane Canu, R Hirata

Support Measure Data Description

Jorge Guevara, Stéphane Canu, and R. Hirata Jr.

Abstract—We address the problem of learning a data description model for a dataset whose elements or observations are itself a set of points in $\mathbb{R}^D$. Modeling each observation as a probability measure, we describe such dataset by computing a minimum volume set for the probability measures, as means of a minimum enclosing ball of the representer functions of the probability measures in a Reproducing Kernel Hilbert Space (RKHS). The advantage is that we do not consider any particular form for the probability measures, instead, we use the embedding of such measures into a RKHS given by a positive definite kernel on probability measures. As a result, the data description model is a function that only depends on some probability measures: the *support measures*. We formulated three support measure data description models for such datasets: the optimization problem for the first one is a chance constrained program; the second is a direct extension of the support vector data description method to the case of probability measures; the third is the same as the second one, but defined for stationary kernels and scaling on data. We validate our method in the challenging task of group anomaly detection, with artificial and real datasets.

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]–[6] . 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 empirical probability density function, a density level set or information about the density support set.

Usually, DD methods work on datasets given by sets of the form: $\{\mathbf{x}_i\}_{i=1}^N,\ \mathbf{x}_i \in \mathbb{R}^D$, where N is the number of observations in the dataset. However, there is a growing interest in machine learning methods for datasets whose individual observations are clusters, groups or, sets of points in $\mathbb{R}^D$ [7]–[20]. Such datasets are sets of the form:

$$\mathcal{T} = \{s_i\}_{i=1}^N, \tag{1}$$

where N is the number of observations and the observation s_i is the set $\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{L_i}^{(i)}\}$ with cardinality L_i, and each element of s_i is a point in $\mathbb{R}^D$. Practical examples of observations taking the form of s_i are: a set of image features in an image dataset [21], a set of spatio-temporal features [22], a set of replicates values for a measurement process [23], a set of points describing point wise uncertainty [12]–[14], a set describing subjective judgments [15], a set describing the invariance of some particular object [16].

Data description of datasets given by (1) is of crucial importance for possible practical applications as we illustrate taking as example the *group anomaly detection task* [7], [8]. The aim of group anomaly detection is to detect anomalous sets of points from datasets taking the form of (1), i.e., to detect *unusual* observations s_i from (1). Each anomalous set of points, or group anomaly, could be given by [7]: 1) *point-based* anomalies, that is, the aggregation of anomalous points, i.e., all the elements in s_i are anomalous, or 2) *distribution-based* anomalies, that is, the anomalous aggregation of non-anomalous points, i.e., all the elements in s_i are non-anomalous but the aggregation itself is unusual or anomalous.

In classical anomaly detection[1], a point is anomalous if it differs from the majority of points in the dataset, for example, a point far away from the empirical mean, or from the support of the generating distribution of points in the dataset. However in group anomaly detection, a group or set of points is considered anomalous, if the local distribution of those points is not similar to the local distributions of the non-anomalous groups of points. Consequently, the information provided by each local distribution of points is crucial for a right description of datasets given by (1). Figure 1 shows how taking a statistic from s_i, and describing such representative values with conventional DD methods, is not enough to guarantee group anomaly detection, because reducing the information provided by s_i to a single value will turn conventional anomaly detection methods, highly depended of such a procedure. Moreover, by doing so, useful information could be discarded.

Minimum Volume Sets (MV-sets) for describing

- *Jorge Guevara is with the Department of Computer Science, Institute of Mathematics and Statistics, University of Sao Paulo, Sao Paulo, Brazil. E-mail: jorge.jorjasso@gmail.com*
- *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*
- *R. Hirata Jr is with the Department of Computer Science, Institute of Mathematics and Statistics, University of Sao Paulo, Sao Paulo, Brazil. E-mail: hirata@ime.usp.br*

1. Detecting anomalous $\mathbf{x}_i \in \mathbb{R}^D$ from a set $\{\mathbf{x}_i\}_{i=1}^N$

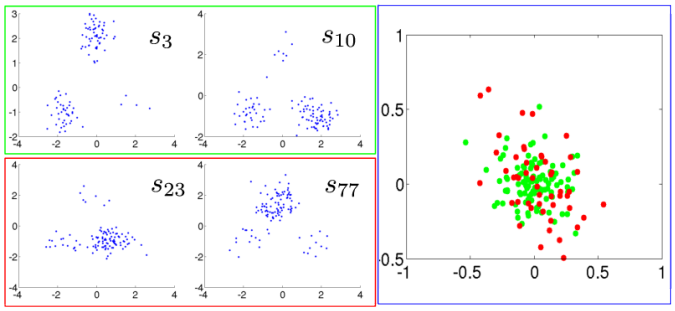


Fig. 1: Left part: four observations: $s_1, s_{10}, s_{23}, s_{77}$, from a dataset given by (1), containing one hundred observations. Observations s_3, s_{10} are two non-anomalous groups of points. Observations s_{23}, s_{77} are two anomalous groups of points. The plotting axes suggest an overlapping between points from anomalous and non-anomalous groups. Right: red and green points are the group means of anomalous and non-anomalous groups of all the observations in the dataset, respectively. The overlapping between the group means of the anomalous and non-anomalous groups will turn classical DD methods not suitable to perform group anomaly detection in this dataset.

datasets. MV-sets are widely used to find a description of datasets of the form $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$ [1]–[4]. A volume set is a set of points belonging to some region in $\mathbb{R}^D$. A MV-set is then computed following some optimization criteria, over all the possible volume sets. Instead of consider such a general case, DD methods consider only the class of sets formed by sets of points belonging to some specific geometric form, as for example, classes of sets given by ellipsoids and convex sets [1], half-spaces in a Reproducing Kernel Hilbert Spaces (RKHS) [4], [10] or enclosing balls in a RKHS [5].

This work aims to find a description of datasets given by (1). To do that, we assume that points in s_i are generated from a unknown local generating distribution $\mathbb{P}_i$. Doing that, the description of (1) can be posed as estimating a MV-set for $\{\mathbb{P}_i\}_{i=1}^N$. Formally, we assume that the points in each observation: $\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{L_i}^{(i)} \in s_i$, are i.i.d [2] 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 $\mathbb{R}^D$. A generalization of the definition of MV-set [1]–[4] to the case of probability measures is given by the next definition.

Definition 1.1 (MV-set for probability measures)**.** 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 MV-set is the set[3] [4]:

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

where ρ is a reference measure on $\mathcal{A}$, and $\alpha \in [0, 1]$. The MV-set G_α^*, describes a fraction α of the mass concentration of $\mathcal{E}$ [5].

That is, given $\alpha \in [0, 1]$ and a reference measure ρ, a MV-set for a set of probability measures is found as follows: 1) measuring each possible set of probability measures $G \in \mathcal{A}$, by $\mathcal{E}(G)$, 2) keeping all the sets of probability measures $G \in \mathcal{A}$ satisfying $\mathcal{E}(G) \geq \alpha$ and, 3) selecting the set $G_\alpha^* \in \mathcal{A}$, such that $\rho(G_\alpha^*)$ is the infimum $\rho(G)$ over are all the sets $G \in \mathcal{A}$ satisfying $\mathcal{E}(G) \geq \alpha$. Finding a MV-set of a set of probability measures with the above procedure is very general, instead, we limit our attention to the class of sets $\mathcal{A}$ formed by sets of probability measures satisfying that the realizations of the random variables following such probability measures are inside of a hypersphere or ball.

Enclosing balls in Reproducing Kernel Hilbert Spaces. Considering $\{\mathbb{P}_i\}_{i=1}^N$ as an i.i.d sample distributed according to $\mathcal{E}$, such that each $\mathbb{P}_i$ is unknown, we use an implicit feature mapping given by a real-valued positive definite kernel defined on $\mathcal{P} \times \mathcal{P}$. We consider then the class $\mathcal{A}$ in (2) implicitly defined by such a kernel, as the set of enclosing balls of the implicit *representer functions* of $\{\mathbb{P}_i\}_{i=1}^N$ in a RKHS. Consequently, the MV-set given by (2) is given by the minimum enclosing ball (MEB) of such functions in the RKHS.

The representer functions of $\{\mathbb{P}_i\}_{i=1}^N$ in a RKHS are given by performing the embedding of $\{\mathbb{P}_i\}_{i=1}^N$ into a RKHS [18], [26]–[28]. Such embedding provides a way to compute inner products in $\mathcal{P}$ using a kernel defined on $\mathcal{P} \times \mathcal{P}$, *without* computing the density of such local distributions as an intermediate step. Moreover, a good approximation for the kernel is assured by an empirically estimation of it using (1) [18]. Consequently, in the same way of kernel methods, the description of $\{\mathbb{P}_i\}_{i=1}^N$ will be a function depending only on some training examples, $\mathbb{P}_i$, which in analogy to support vectors in kernel methods are called as *support measures*. Consequently, we call the DD method presented in this work as *Support Measure Data Description* or SMDD.

We present three SMDD models through the paper. In Section 2 is presented the first one as an optimization problem with chance constraints [29], [30] in the space of probability measures, which is further extended to a RKHS in Section 3. The second and third models are direct extensions of the SVDD model [5] to the case of

2. Independently and identically distributed.

3. $\mathcal{A}$ is for instance the Borel σ-algebra with respect to the topology of weak convergence [10], [24].

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

5. As density level sets [25] are MV-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 MV-set satisfying that G_1^* correspond to the p-zero level set of $\mathcal{E}$, that is, the density support estimation set of $\mathcal{E}$.

datasets given by $\{\mathbb{P}_i\}_{i=1}^N$. The third SMDD model is almost the same as the second model with the only difference that it considers a scaling of data and translation invariant kernels. Both SMMD models are presented in Section 3. The relationship among all the SMDD models is presented in Section 4. We show through a set of experiments in Section 5, the behavior of such models in the group anomaly detection task using artificial and real world datasets. Finally, some 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: $X : \Omega \to \mathbb{R}^D$, 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 $\mathbb{P}\{\omega : a < X(\omega) \leq b\}$ by $\mathbb{P}(a < X \leq b)$.

2 Support Measure Data Description in the Space of Probability Measures

In this section we estimate a empirical MV-set for $\{\mathbb{P}_i\}_{i=1}^N$, as a MEB for the realizations of random variables distributed according $\{\mathbb{P}_i\}_{i=1}^N$. We name such approach as *SMDD model in the space of probability measures*, which is presented as an optimization problem with chance constraints [29], [30]. Further, using Markov's inequality [31], we transform such an optimization problem into another one with deterministic constraints. We also present its dual formulation as well. All the results in the present section are further kernelized and consequently extended to a RKHS in Section 3.

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

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

where $(R, \mathbf{c}) \in \mathbb{R} \times \mathbb{R}^D$ are named as the *enclosing balls* for $\{\mathbb{P}_i\}_{i=1}^N$, and $X_i \sim \mathbb{P}_i$[6]. The empirical MV-set $\hat{G}_\alpha^*$ is then found by estimating the MEB (optimal R and $\mathbf{c}$ values, following some optimization criteria) using the sample $\{\mathbb{P}_i\}_{i=1}^N$. In this case, the optimal radius will be proportional to the value α in (2). However, (3) is very conservative definition, because $\mathbb{P}_i$ will be in (3), only if all the possible realizations of $X_i \sim \mathbb{P}_i$ are inside the enclosing ball $(R, \mathbf{c})$.

A more flexible formulation is given by allowing some realizations of $X_i \sim \mathbb{P}_i$ not be part of the set $\hat{G}$, this can be done by setting some arbitrary threshold values: $\mathcal{K} = \{\kappa_i\}_{i=1}^N,\ \kappa_i \in [0, 1]$ to control the probability for which the realizations of $X_i \sim \mathbb{P}_i$ are inside of the ball $(R, \mathbf{c})$. That is, given the set $\mathcal{K} = \{\kappa_i\}_{i=1}^N,\ \kappa_i \in [0, 1]$, a more flexible formulation for an empirical version of G in (2) is given by:

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

6. Notation $\sim$ means *distributed according to*.

As each probability measure $\mathbb{P}_i$ is in $\hat{G}(\mathcal{K}, R, \mathbf{c})$ depending on its associated value κ_i, it is possible to see that if all $\kappa_i = 0$, then (4) reduces to (3), and if for some i, $\kappa_i = 1$, 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 MV-set given by (4) means to define an optimization problem with chance constraints. That is, given $\{\mathbb{P}_i\}_{i=1}^N$, and $\{\kappa_i\}_{i=1}^N,\ \kappa_i \in [0, 1]$, the optimization problem is given by

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

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

In the same way of kernel methods, we introduce the slack variables: $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_N) \in \mathbb{R}^N$ to allow some probability measures from $\{\mathbb{P}_i\}_{i=1}^N$ not be part of the estimated MEB, then 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 $\lambda > 0$ is a regularization parameter.

As the κ_i-values are thresholds, 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 ball $(\sqrt{R^2 + \xi_i}, \mathbf{c})$ is bounded by κ_i*. Equivalently, the left side of each probabilistic constraint of Problem 2.1 is the distribution function of the random variable[7] $Z_i = \|X_i - \mathbf{c}\|^2$ 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 posed 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 *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 MV-set.

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

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 [28]. 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, allowing flexibility in the model. However, each constraint requires we 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 [31], which for a nonnegative random variable $X \sim \mathbb{P}$ and for some $t > 0$, bounds $\mathbb{P}(X \geq t)$ 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:

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

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

Lemma 2.1. [8] *Let $\mathbb{P}$ be a probability measure with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then for $X \sim \mathbb{P}$*

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

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 $\{\boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}^{D\times D}$ estimated from $\{\mathbb{P}_i\}_{i=1}^N$ and $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in (0, 1]$, 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 $i = 1, \ldots, N$. Problem 2.2 is named as *SMDD with joint constraints* if $\kappa_i = \kappa$ for all $i \in 1, 2, \ldots, N$.

Lemma 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 [5] 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 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 Problem 2.2 is:

$$\begin{aligned} \mathcal{L}(R, \mathbf{c}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) =& R^2 + \lambda \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i \{(R^2 + \xi_i)\kappa_i \\ & - \|\boldsymbol{\mu}_i - \mathbf{c}\|^2\} - tr(\Sigma_i)\} - \sum_{i=1}^N \beta\xi_i \end{aligned} \tag{6}$$

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

$$\left.\begin{array}{lll} \partial_R \mathcal{L} = 0 & : & \sum_{i=1}^N \alpha_i\kappa_i = 1 \\ \nabla_{\mathbf{c}} \mathcal{L} = 0 & : & -2\sum_{i=1}^N \alpha_i\boldsymbol{\mu}_i + 2\sum_{i=1}^N \alpha_i\mathbf{c} = 0 \\ \partial_{\xi_i} \mathcal{L} = 0 & : & \lambda - \alpha_i\kappa_i - \beta_i = 0 \end{array}\right\} \tag{7}$$

$$\left.\begin{array}{l} \alpha_i\{(R^2 + \xi_i)\kappa_i - \|\boldsymbol{\mu}_i - \mathbf{c}\|^2 - tr(\Sigma_i)\} = 0 \\ \beta_i\xi_i = 0 \end{array}\right\} \tag{8}$$

Replacing, the KKT's condition in (6), we obtain the dual problem, i.e., given $\{\boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}^{D\times D}$ estimated from $\{\mathbb{P}_i\}_{i=1}^N$ and $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0, 1]$, the dual form of Problem 2.2 is given by

Problem 2.3.

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

2.1.2 Representer Theorem and Analysis of KKT's

From stationary conditions (7), the Representer Theorem [32] for $\mathbf{c}$ is:

$$\mathbf{c} = \frac{\sum_i \alpha_i\boldsymbol{\mu}_i}{\sum_i \alpha_i},\ i \in \{i \in \mathcal{I} \mid 0 < \alpha_i\kappa_i \leq \lambda\}, \tag{9}$$

where $\mathcal{I} = \{1, 2, \ldots, N\}$.

Analyzing the complementarity conditions (8) we identify the following cases for all $i \in \mathcal{I}$. See Table (1) for a summary.

8. The proof is found in the supplemental material.

- $\alpha_i = 0, \beta_i > 0 \implies \xi_i = 0$, that yields $\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + tr(\Sigma_i) \leq R^2\kappa_i$. 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 $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ will be *inside* the ball $(R, \mathbf{c})$ *no matters the value for* κ_i. Consequently, all $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ are considered to be described by the ball.

- $\alpha_i > 0, \beta_i = 0 \implies \xi_i > 0$, 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 $i \in \{i \in \mathcal{I} \mid \alpha_i\kappa_i = \lambda\}$ will be *outside* the ball $(R, \mathbf{c})$ with probability κ_i. Consequently, all $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i\kappa_i = \lambda\}$ are considered to be errors allowed in the training set $\{\mathbb{P}_i\}_{i=1}^N$.

- $\alpha_i > 0, \beta_i > 0 \implies \xi_i = 0$ and $0 < \alpha_i\kappa_i < \lambda$. From this and (8) we can retrieve the radius

$$R^2 = \frac{\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + tr(\Sigma_i)}{\kappa_i}, \tag{10}$$

where $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i\kappa_i < \lambda\}$. Notice that 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 $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i\kappa_i < \lambda\}$ will be *on* the surface of the ball $(R, \mathbf{c})$. Consequently, all $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid 0 < \alpha_i\kappa_i < \lambda\}$ are considered described by the ball $(R, \mathbf{c})$ and are called *support measures*.

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

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

The radius value given by (10) can also be obtained alternatively using information from a Lagrange multipliers of the Lagrangian of Problem 2.3, as we point out in the following result.

Teorema 2.3. [9] *Let η be the Lagrange multiplier of the constraint $\sum_{i=1}^N \alpha_i\kappa_i = 1$ of the 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. Zhang et al [14], model input uncertainties with a bounded uncertainty model for data with additive noise. In [12], [13] is assumed some a priori distribution for model uncertainties resulting in an optimization model formulated as a Second Order Cone Program [33]. In [17] is considered a Taylor approximation in the RKHS to deal with input uncertainties. The main disadvantage of all those models is the kernelization step and the a priori assumptions for $\mathbb{P}_i$.

9. The proof is found in the supplemental material.

3 Support Measure Data Description in Reproducing Kernel Hilbert Spaces

In this section we present three formulations of SMDD in a RKHS. Those formulations are MEB's in the RKHS, which correspond to nonlinear descriptions of the training set $\{\mathbb{P}_i\}_{i=1}^N$. The first SMDD model in a RKHS, described in Section 3.2, is the kernelization of the SMDD model presented in the last section. The second and third SMDD models in a RKHS, described in Sections 3.3 and 3.4, respectively, are direct extensions of the SVDD [5] to the case of training sets of probability measures. with the only difference that the third model uses only invariant translation kernels and a scaling of data. We start our discussion showing the main facts of the Hilbert space embeddings of probability measures.

Notation. Letter $\mathcal{H}$ denotes a 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 $\langle ., . \rangle_{\mathcal{H}}$ and $\|.\|_{\mathcal{H}}$ denotes the norm in the RKHS.

3.1 Hilbert Space Embedding of Probability Measures

Hilbert space embedding of probability measures [18], [26]–[28], gives a way to represent probability measures $\mathbb{P}_i$ as functions in a RKHS. Such functions are commonly named as *representer functions*, *mean functions* or *mean maps*. We present them in the following definition.

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:

$$\begin{aligned} \mu_{\mathbb{P}} : \mathbb{R}^D &\to \mathbb{R} \\ t &\mapsto \mu_{\mathbb{P}}(t) = \mathbb{E}_{\mathbb{P}}[k(X, t)], \end{aligned} \tag{11}$$

where $\mathbb{E}_{\mathbb{P}}[k(X, t)] = \int_{\mathbf{x} \in \mathbb{R}^D} k(\mathbf{x}, t) d\mathbb{P}(\mathbf{x})$.

Thus, $\mu_{\mathbb{P}}$ is the representer function in $\mathcal{H}$ for $\mathbb{P}$. 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], [34], [35]. As a 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 $f \in \mathcal{H}$.

The embedding of probability measures $\mathbb{P} \in \mathcal{P}$ to mean maps $\mu_{\mathbb{P}} \in \mathcal{H}$, is given in the following definition.

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

$$\begin{aligned} \mu : \mathcal{P} &\to \mathcal{H} \\ \mathbb{P} &\mapsto \mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[k(X, .)] \end{aligned} \tag{12}$$

where $\mathbb{E}_{\mathbb{P}}[k(X,.)] = \int_{\mathbf{x}\in\mathbb{R}^D} k(\mathbf{x},.)d\mathbb{P}(\mathbf{x})$.

Choosing *characteristic kernels* [35]–[37] 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 [35]. Furthermore, an empirical estimator of $\mu_{\mathbb{P}}$ from the sample $\{x_i\}_{i=1}^M$ drawn i.i.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].

3.1.1 Kernel on probability measures

The mapping

$$\begin{aligned}\mathcal{P}\times\mathcal{P} &\rightarrow \mathbb{R}\\ (\mathbb{P},\mathbb{Q}) &\mapsto \langle\mathbb{P},\mathbb{Q}\rangle_{\mathcal{P}} = \langle\mu_{\mathbb{P}},\mu_{\mathbb{Q}}\rangle_{\mathcal{H}},\end{aligned} \tag{13}$$

defines an inner product on $\mathcal{P}$, where from Fubini's theorem [38] 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 $\mathcal{P}\times\mathcal{P}$, defined by

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

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

Hilbert space embedding of signed measures was introduced in [26] and later by [27], [28], and by [18] when the measures are probability measures. Some applications in machine learning include, dimensionality reduction [39], measuring independence of random variables [40], two-sample test [34], embeddings of Hidden Markov Models into RKHS [41], Bayes rule [42], support vector machines [10], [20] among others [35], [37], [43]. The kernel on probability measures can be estimated using (1) without requiring fitting some probabilistic models to the observations. Another related kernels on distributions which assume probabilistic models for observations are the Fisher kernel [44], the kernel based on the symmetrized Kullback-Leibler (KL) divergence on distributions [45], the Bhattacharyya kernel [19], and the probability product kernel [46].

3.2 SMDD model in a RKHS as Chance Constrained Problem

Applying the theory of Hilbert space embeddings of probability measures, in this section we generalize the SMDD presented in Section 2 to the case of RKHS. For example, Markov's inequality and lemmas of Section 2 are extended to the case of mean functions and covariance operators in RKHS.

Using a positive definite real-valued kernel k, defined on $\mathbb{R}^D\times\mathbb{R}^D$, the set $\hat{G}$ from (4) is given by:

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

where $k(X,.) \in \mathcal{H}$, $X \sim \mathbb{P}$, and the center and radius of the ball are $c \in \mathcal{H}$, and $R \in \mathbb{R}$, respectively. Enclosing balls $(R, c(.))$, are defined in the RKHS, which in the input space correspond to nonlinear descriptions of $\{\mathbb{P}\}_{i=1}^N$. The MEB is then formulated as follows: given $\{\mathbb{P}_i\}_{i=1}^N$ and $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0,1]$, SMDD in the RKHS as chance-constrained program is the following:

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

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

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 embedding the 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 $i = 1,2,\ldots,N$.

3.2.1 Trace of the Covariance Operator

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 $\mathcal{H}$ with kernel k is the mapping $\Sigma^{\mathcal{H}} : \mathcal{H} \rightarrow \mathcal{H}$, such that for all $f, g \in \mathcal{H}$ it satisfies:

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

because reproducing property [10]. The covariance operator is then the possible infinite dimensional matrix:

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

From this, the trace of $\Sigma^{\mathcal{H}}$ can be obtained as: [11]

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

where the last line is due to the reproducing property and Definition 3.1. Then, using (14), yields

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

10. $\Sigma^{\mathcal{H}}$ is a bounded operator on a separable infinite dimensional Hilbert space and can be represented by an infinite matrix [47].

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

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

Lemma 3.1. [12]

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

3.2.2 Deterministic Form in the RKHS

From Lemma (3.1), Markov's inequality and a similar analysis of Section 2.1, given the mean functions $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ of $\{\mathbb{P}_i\}_{i=1}^N$ and $\{\kappa_i\}_{i=1}^N,\ \kappa_i \in (0,1]$, the deterministic form of SMDD in the RKHS is the following:

Problem 3.1.

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

for all $i = 1,\ldots,N$, where $tr(\Sigma_i^{\mathcal{H}})$ is given by (16).

3.2.3 Dual Formulation of Problem 3.1

Denoting by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the Lagrange multipliers vectors with nonnegative components α_i and β_i, $i = 1,2,\ldots,N$, respectively, the Lagrangian of Problem 3.1 is

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

The stationarity and complementarity KKT conditions for this problem are:

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

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

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

Problem 3.2.

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

12. The proof is found in the supplemental material.

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:

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

3.2.4 Representer Theorem and Analysis of KKT's

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

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

for all $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i\kappa_i \leq \lambda\}$, where $\mathcal{I} = \{1,2,\ldots,N\}$

Analyzing stationary (18) and complementary (19) conditions we have the same information provided by Table 1 but with $\mu_{\mathbb{P}_i} \in \mathcal{H}$ instead of $\boldsymbol{\mu}_i \in \mathbb{R}^D$, $c \in \mathcal{H}$ instead $\mathbf{c} \in \mathbb{R}^D$ and $\Sigma_i^{\mathcal{H}}$ instead of Σ_i. From this we have

- all $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ are described by (R, c), because $\mu_{\mathbb{P}_i}$ are inside (R, c)
- all $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i\kappa_i = \lambda\}$ are considered to be the errors allowed in the training set.
- all $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid 0 < \alpha_i\kappa_i < \lambda\}$ are the *support measures*, from this the radius is computed as

$$R^2 = \frac{\|\mu_{\mathbb{P}_i} - c(.)\|^2 + tr(\Sigma_i^{\mathcal{H}})}{\kappa_i}, \tag{21}$$

for all $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i\kappa_i < \lambda\}$.

Alternatively, R can be computed by Theorem (2.3) as $R = \sqrt{\eta}$, where η is the Lagrange multiplier of the constraint $\sum_i^N \alpha_i\kappa_i$ of the Lagrangian of Problem 3.2. It is worth to note that using the linear kernel: $k(\mathbf{x}, \mathbf{x}') = \langle\mathbf{x}, \mathbf{x}'\rangle$ in (14), Problem 3.2 is equivalent to Problem 2.3, because, $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \mathbb{E}_{\mathbb{P}_i}[\mathbb{E}_{\mathbb{P}_j}[\langle X, X'\rangle]] = \langle\boldsymbol{\mu}_i, \boldsymbol{\mu}_j\rangle$.

3.2.5 Role of κ-values

As it was point out in Section 2, particular values for κ in the constraint i will increment or decrement the radius covering the realizations of $X \sim \mathbb{P}_i$. We illustrate this behavior in Figure 2, which shows a dataset of ten Gaussian distributions, and six MEB's for them, given by solving Problem 3.2, with six different settings for the κ values. We used the RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$ and $\gamma = 1$ to implement (14).

Left to right of top part of Figure 2 shows three MV-sets in the input space which are three MEB's in RKHS, given by solving three SMDD (Problem 3.2) with $\kappa_i = 0.8$, $\kappa_i = 0.9$ and $\kappa_i = 1.0$ for all the constraints i of the three models, respectively.

Left to right of bottom part of Figure 2 shows another three MV-set. All the three SMDD models have $\kappa_i = 1$ in their constraints, excepting $\kappa_1 = \kappa_2 = 0.8$ for the first, $\kappa_1 = \kappa_2 = 0.9$ for the second, and $\kappa_1 = \kappa_2 = 1.0$ for the third SMDD. Depending on particular κ values, the corresponding probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$, belongs to the MV-set in some degree.

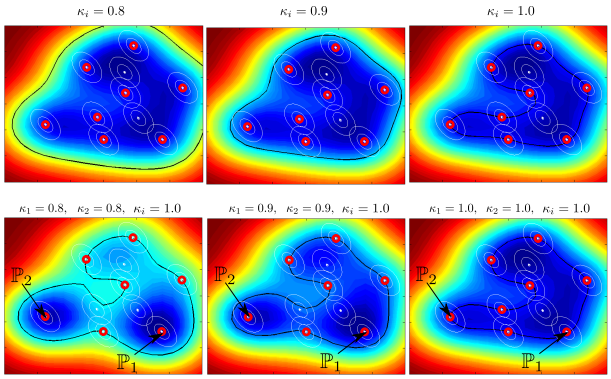


Fig. 2: Six MEB's describing a dataset of Gaussian distributions, showing the effect of choosing particular κ values for their constraints. Top part: all the three models share the same κ-value, which is incremented from left to right. Bottom part: all the three models have $\kappa = 1$ for all their constraints, except its constraints $i = 1$ and $i = 2$, which are incremented from left to right. Decreasing a κ-value for some specific constraint, tend to cover the respectively $\mathbb{P}_i$

3.3 SMDD in a RKHS as a direct extension of SVDD

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 $\mu_{\mathbb{P}_i}$, an empirical version of G from (4) is given by:

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

then, the empirical MV-set $\hat{G}^*_\alpha$ is computed by a MEB for the mean maps $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$. This can be formulated as follows: given the mean functions $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ of $\{\mathbb{P}_i\}_{i=1}^N$, 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:

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

The optimality (KKT) conditions for this problem are:

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

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

3.3.1 Dual Formulation

Given the mean functions $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ of $\{\mathbb{P}_i\}_{i=1}^N$, the dual form of the previously Problem is given by replacing, (24) into (23) as follows:

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:

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

Analyzing (25) follows that all $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ are probability measures described by (R, c), because $\mu_{\mathbb{P}_i}$ are inside the ball (R, c). All $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i = \lambda\}$ are errors. All $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid 0 < \alpha_i < \lambda\}$ are support measures.

If the linear kernel: $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$ is used in (14), 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 or stationary kernels [48], are kernels of the form $k_I(x, x) = k'(x - x')$, that is, such kernels only depend on the difference $x - x'$, and not of the observations x, x' themselves. The kernel k' is positive definite only if it can be written as $\int_{\mathbb{R}^D} cos(\omega^\top (x - x'))dF(\omega)$, where F is a positive measure.

Implicit feature maps of translation invariant kernels, are functions $k_I(x, .)$ in a RKHS lying on a surface of a hypersphere, that is, have constant norm. To see that, note that translation invariant kernels satisfy:

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

where ϵ is a constant value, then immediately follows that $\|k_I(x, .)\|_{\mathcal{H}} = \sqrt{|\epsilon|}$, that is, functions $k_I(x, .)$ lie on a surface of a sphere of radius $\sqrt{|\epsilon|}$.

However, mean maps given as $\mu_{\mathbb{P}} = E_{\mathbb{P}}[k_I(X, .)]$, does not have constant norm, because

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

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 some hypersphere. The following theorem is due to Muandet et al [10].

Teorema 3.2 (Spherical Normalization [10])**.** *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 $\mu : \mathcal{P} \to \mathcal{H}$.

Or in another words, Theorem 3.2 says that all the information is preserved after performing spherical normalization on the data.

Consequently, an empirical version for the set $\hat{G}$ from (4) is given by:

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

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

The optimization problem for this model is almost the same as given in Problem 3.3:

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

but with kernel:

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

which is due to Theorem 3.2. Note that $\tilde{k}$ is given by (14) but with kernel k_I.

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.

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

This formulation looks like the dual formulation of One-class Support Vector Machine [4], 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 the equivalence between SMDD models and One-Class Support Measure machine (OCSMM) [4], [10]. Thorough this Section, we denote the SMDD presented in Section 3.2 as M1 (Problems 3.1 and 3.2), the SMDD of Section 3.3 as M2 (Problem 3.3), and the SMDD of Section 3.4 as M3 (Problem 3.4).

We start showing how M1 can be formulated if we restrict it only to the case of joint constraints and same covariance matrix. Thus, we use this formulation to compare it with M1 and M2.

Teorema 4.1. [13] *The Primal form of M1 (Problem 3.1) with joint constraints sharing the same covariance matrix, i.e, $\kappa_i = \kappa$ and $\Sigma_i^{\mathcal{H}} = \Sigma^{\mathcal{H}}$ for all $i = 1, 2 \ldots, N$ and $\lambda > 0$, can be written as*

Problem 4.1.

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

where

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

Notice that Problem 4.1 is a less flexible formulation of M1 (Problem 3.1), because it considers the same local covariance and the same κ values for all points. It is easy to verify that, for optimal $c \in \mathcal{H}$ and ρ' values from Problem 4.1. [14], we retrieve the radius as:

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

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

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

Teorema 4.2. [15] *Using the kernel between probability measures given by* (14)*, the dual of Problem 4.1 is given by:*

Problem 4.2.

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

Lemma 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 (24) as $c = \sum_i \alpha_i \mu_{\mathbb{P}_i},\ i \in \{i | 0 < \alpha_i \leq \lambda\}$, and the radius R from (29).

13. The proof is found in the supplemental material.

14. See proof of Theorem 4.1 in the supplemental material as an example

15. The proof is found in the supplemental material.

4.1 Equivalence among SMDD models

4.1.1 M1 vs M2

Problem 4.2 is the dual of M1 with joint constraints sharing the same covariance matrix. Under this setting M1 is almost the same as M2 but with a difference of a scaling factor of 0.5 in the dual objective function.

4.1.2 M1 vs M3

After spherical normalization on data, the dual objective function of Problem 4.2, becomes $-0.5\sum_{i,j=1}^{N}\alpha_i\alpha_j\tilde{\tilde{k}}(\mathbb{P}_i,\mathbb{P}_j)$, because the other term in the objetive function is constant where $\tilde{\tilde{k}}$ is the kernel given by (27). Therefor, M1 with joint constraints sharing the same covariance matrix and spherical normalization on data is equivalent to model M3, with a difference of a scaling factor of 0.5 in the dual objective function, as well.

4.2 Connection with One-Class Support Vector Machines

It is widely known that 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], [10], 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. *M2, M3 and OCSMM [10] are equivalents if it is performed spherical normalization on the training set* $\{\mathbb{P}_i\}_{i=1}^{N}$ *by* (3.2).

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. □

Consequently, it is possible to deduce that M1 under the setting given by Problem 4.2 is equivalent to OCSMM, with a difference of a scaling factor of 0.5 in the dual objective function.

5 Experiments

In this section, following the works [7], [8], [10], we tested the SMDD models in the challenging task of group anomaly detection for the case of Point-based anomaly detection in 5.1 and, for the case of Distribution-based anomaly detection in Section5.2. Finally, in Section 5.3, we use real data from the *Sloan Digital Sky Survey* (SDSS) project, to detect anomalous groups of galaxies. We employed the same notation as the ones given in Section 4 to denote the SMDD models, i.e., M1 is the SMDD given by Problem 3.2, M2 by Problem 3.3, and M3 by Problem 3.4. Also, for comparison purposes, M4 will be the OCSMM [10], and M5 will be the SVDD [5]. We used M5 as the baseline for our experiments. This is shown in Table 2.

Kernel and covariance estimation. The kernel between probability measures given by (14) was estimated via the empirical estimator:

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

on the training set given by (1). Also, the trace of the covariance operator in the RKHS 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	[10]
M5	SVDD	[5]

TABLE 2: Models used in experiments

We used *CVX*, a package for specifying and solving convex programs [49], [50] to solve M1. To solve M2, M3, M4 and M5 we used the *SVM and Kernel Methods Matlab Toolbox* (SVM-KM) [51]. [16]

Thorough the experiments, for each observation s_i from (1) we used the terms: *group*, *cluster* or set of points, interchangeably, i.e., we say training sets given by (1) have N groups, clusters or set of points. As M5 was not originally designed to deal with probability measures it was trained using only the empirical group means.

5.1 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 from 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 [7], anomalous groups are given by aggregating individually anomalous points. For this experiment, we generated 50 non-anomalous groups of points and 20 anomalous groups

16. The Matlab code and datasets for experiments can be found at http://www.vision.ime.usp.br/~jorjasso/SMDD.html

of points as we will detail in the next paragraph. 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).

Points of non-anomalous groups were randomly sampled from a *Multimodal Gaussian Mixture Distribution* or GMD [8], [10], with two different group type distributions: $\pi = (0.48, 0.52)$. That is, 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 other 52% of non-anomalous groups were generated from another GMD with the same parameters but with mixture weights: $(0.33, 0.03, 0.64)$. The green box in Figure 3 shows three non-anomalous groups for $\pi = 0.48$ and the yellow box shows two non-anomalous groups for $\pi = 0.52$.

Three different types of anomalous groups have been generated. the first type 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 3 shows five anomalous groups of this type. The second type was given by 5 group of points sampled from a GMD 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 3 shows five anomalous groups of this type. The third type of group anomalies was given by 5 group of points sampled from a GMD 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 3 shows five anomalous groups of this type.

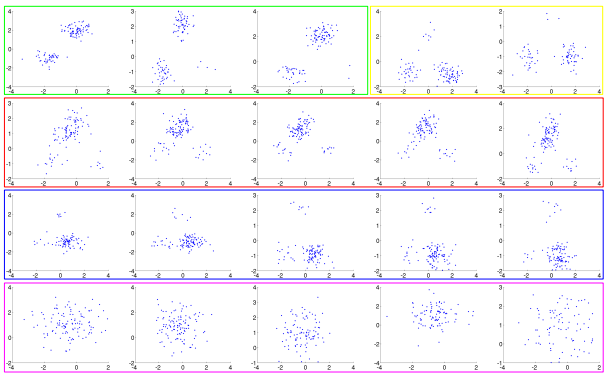


Fig. 3: Group anomaly detection dataset. Green and yellow boxes contains non-anomalous groups of points. Red, blue, and magenta boxes contains anomalous groups of points.

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

As usually, in one-class classification, training data and test data has no labels, turning impossible to have a validation set for model selection. Also, as the problem of describing a dataset is not trivial, because it may exists many models describing well such dataset, we follows the same methodology used in literature, that is, we choose arbitrarily a value for the regularization parameter λ of the SMDD model and, the kernel parameters are computed using some heuristic on the available data, avoiding to use the training or the test set for perform model selection.

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.,

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

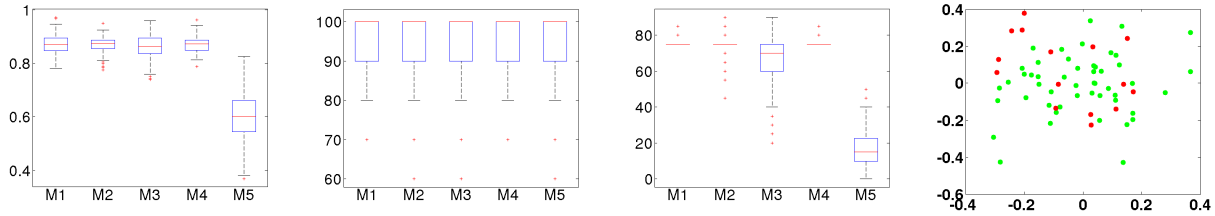
where s is the 0.1 quantile, i, j are the groups indices, and k, l are the points indices.

Figures 4a, 4b, 4c show the results in boxplots for this experiment. The boxplots shows 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: 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, M5 (SVDD) can not detect such group anomalies using only the group means as the training set.

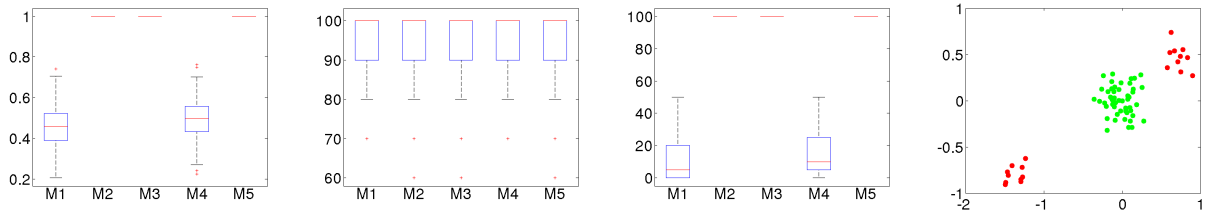
To see why group anomaly detection is a hard problem, the plot of the means of all the non-anomalous and anomalous groups is shown in Figure 4d. The green points are the means of non-anomalous groups. Red, blue, and magenta points are the means of anomalous groups of points corresponding to the red, blue, and magenta boxes in Figure 3. Because the non-anomalous group means overlap the anomalous group means, methods as One-class support vector machines and SVDD will not perform well, because such methods consider anomalies the points far away from the mean of the description of the data.

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

Distribution-Based Group Anomalies [7] are anomalous groups of points that individually are non-anomalous but together form anomalous groups. For this experiment, it was generated 50 non-anomalous groups of points to form the training set and 15 anomalous groups of points plus 15 non-anomalous points to form the test

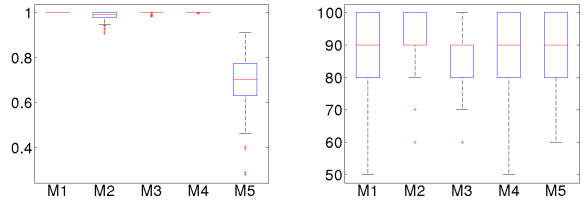


(a) AUC. (b) ACC Non A. (c) ACC Anom. (d)

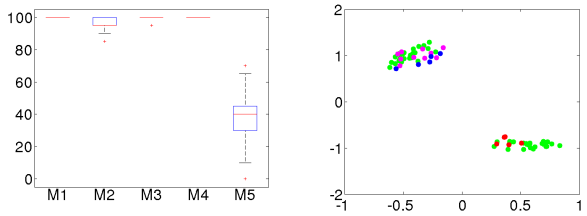


(e) AUC. (f) ACC Non A. (g) ACC Anom. (h)

Fig. 5: Experimental results and plot of the group means for the two Distribution-based anomaly detection experiments. Each row of figures represent one experiment. Boxplots show the ACC and AUC statistics. Marks on the x-axis of each boxplot are the DD models and y-axis is the performace measure.



(a) AUC. (b) ACC Non A.



(c) ACC Anom. (d) Group Means

Fig. 4: Experimental results and a plot of the group means for the Point-based group anomaly detection experiment. Boxplots show the ACC and AUC statistics. Marks on the x-axis of each boxplot are the DD models and y-axis is the performace measure. Figure 4d plots the means of non-anomalous groups vs. the means of anomalous groups. Red, magenta and blue points are the group means of anomalous groups. Green points are the means of non-anomalous groups

set, as we will explain in the next paragraph. Also, the number of points per group was given as the previous experiment, that is: $n_i \sim$ Poisson(100).

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: $0.2 * I_2$.

Anomalous groups were generated from the same GMD of 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.

As the last experiment, it was used the AUC and ACC metrics for performance measure, 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 5d 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 5a, 5b, 5c 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 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 5h 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 5e, 5f, 5g show the

AUC, the ACC for non-anomalous groups, and the ACC for anomalous groups respectively. Models M2 and M3 perform as M5 (SVDD). Performance of 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 information.

5.3 Group Anomaly Detection in Astronomical Data

In this section, we tested the SMDD models with real data: *The Sloan Digital Sky Survey*[17] *(SDSS)* project. This dataset contains massive spectroscopy surveys of the universe, the Milky Way galaxy, and extrasolar planetary systems. The idea of the experiment is to use the dataset to detect anomalous clusters of galaxies. Such a problem, using the same dataset, had been previously studied in [8]–[10] 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 [9]: each vector was down-sampled to get a 500-dimensional feature vector to represent a galaxy. Clusters of galaxies were obtained analyzing the spatial neighborhood of galaxies, see [9] for details. This procedure returns 505 clusters of galaxies of a total of 7530 galaxies. Thus, each cluster of galaxies is a group of about $10 - 15$ galaxies. Finally, PCA was applied to the feature vectors of galaxies, to get a four-dimensional dataset, preserving about 85% of the variance of the data.

To perform group anomaly detection, a training set was formed by randomly choosing 455 group of galaxies among the original 505 groups. Also, it was generated five test datasets, each of them containing the remaining 50 non-anomalous groups from the original 505 groups plus 50 anomalous groups.

In the first test dataset, each anomalous group was constructed by randomly selecting about $n_i \sim$ Poisson(15) galaxies from the 7530 galaxies, i.e, galaxies from the 505 non-anomalous groups. Note that, because galaxies were randomly chosen, the aggregation itself of such galaxies are anomalous.

Anomalous groups for the second, third, fourth and fifth test sets were generated as follows: First, the covariance of the 7530 observations (galaxies) was empirically estimated. After that, were selected randomly three sets of galaxies from the 7530 galaxies, each one containing about $n_i \sim$ Poisson (15) galaxies. Were computed the empirical means of the three sets. With the three empirical mean values and the empirical covariance matrix, a GMD with three components and weights: $p = \{0.33, 0.33, 0.33\}$ was constructed.

Anomalous groups of points for the second test data set were generated from the above GMD, with about $n_i \sim$ Poisson (15) points per group. Anomalous groups of points for the third, fourth and fifth test set were generated as the second test set, but using the covariance matrices $5 * \Sigma$, $10 * \Sigma$, and $100 * \Sigma$, respectively.

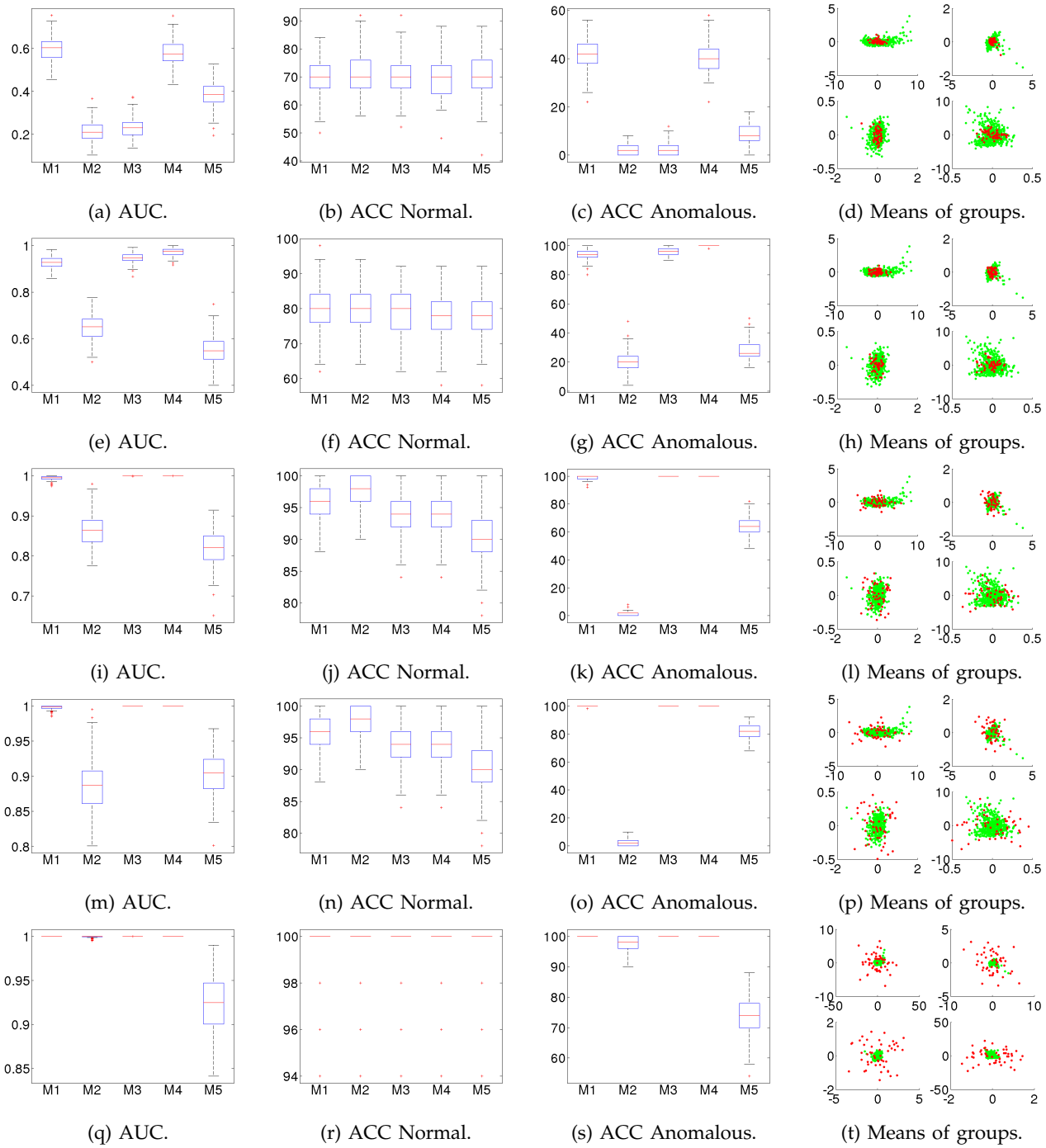
We plotted in Figures 6d, 6h, 6l, 6p,and 6t the group means of the PCA features. Green points are the non-anomalous group means and red points are the anomalous group means. Each figure shows four plots: upper-left: the plot of the first vs second dimension, upper-right: the plot of the second vs third dimension, bottom-left: the plot of the third vs four dimension, bottom-right: the plot of the four vs first dimension. Note that, because the overlapping of group means of non-anomalous groups and anomalous groups, group anomalies for this experiment are hard to detect.

For all the SMDD models, the kernel between probability measures was implemented via 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 the ACC of anomalous groups. Figures 6a, 6b, and 6c show the AUC, the ACC for non-anomalous groups, and the ACC for the anomalous groups in the first test set. The kernel parameter was computed by (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 M5 for this choice of parameters. However, the AUC metric for M5 shows that performance for this model is not better than chance. On the another hand M1 and M4 perform 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 Figure 6d shows the hardness of the problem.

Figures 6e, 6f, and 6g 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 by (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 M5 performs worst than the other models, and spherical normalization on data increases the performance as it can be seen by the AUC value of M3. On the other hand, the accuracy of normal groups is about 80% because the choice of λ. The plot of the group means is shown in Figure 6h.

Figures 6i, 6j, and 6k show the AUC, the ACC for non-anomalous groups, and the ACC for the anomalous groups in the third test set. The RBF kernel parameter was computed by (32) and the regularization parameter was set to $\lambda = 1$. The ACC for anomalous groups shows that 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

17. 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.

(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. 6: Experimental results and plots of the means of non-anomalous groups vs. the means of anomalous group for the group anomaly detection task over a SDSS III dataset. Each row of figures represent one experiment.

pretty well as it is shown in Figure 6i. Again the models with worst performance are M5 and M2, and spherical normalization has a positive effect, increasing the AUC value close to one of M3.

Performance metrics for the fourth test set has similar characteristics than the third set as it can be seen in Figures 6m, 6n, and 6o, 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 fifth test set, AUC metric shows that all the models performs pretty well, nevertheless, M5 is the model with worst performance. Figures 6q, 6r, and 6s 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 called SMDD for datasets, where each observation is a set of points in $\mathbb{R}^D$. To do that, we considered each observation as a probability measure, thus, the SMDD method finds the MEB of the representer functions of a set of probability measures in a RKHS. The main advantages of our approach is that it does not required an estimation of probabilistic models for each observations, that is, it does not assumed any particular form for the probability density functions of each $\mathbb{P}_i$. Instead, everything is done by the embedding of probability measures into RKHS, using a real-valued positive definite kernel on probability measures.

Through the paper, we formulated three SMDD models. The first one uses information of the trace of the covariance operator in a RKHS and mean maps. This model is formulated as a chance constrained program which is further transformed into a deterministic one by means of Markov's inequality. The second SMDD model is a direct extension of the SVDD method to the case of probability measures. This model also uses the mean map embedding of probability measures technique. The third SMDD model is almost the same as the second one, but it considers an scaling of data and translation invariant kernels. The reason behind this, is that mean maps under translation invariant kernels do not have a constant norm in the RKHS. We compared the relationship of the three models, showing the cases when the SMDD models are equivalents, as well.

The presented SMDD models were tested in the challenging group anomaly detection task. We showed empirically that they perform pretty well for such a task, showing that the SMDD method is an alternative methodology to deal with group anomaly detection. SMDD gives a way to perform group anomalous detection by describing a region in the RKHS, given by a MEB of the mean maps of the probability measures used for represent the non-anomalous groups of points. Therefore, mean maps not belonging to the MEB are considered anomalies. As possible practical task to be addressed by SMDD models include novelty detection, clustering and classification, for datasets of probability measures.

Acknowledgments

This work was partly done while the first author was visiting the Institute National of Science Appliqués, 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] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.

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

[8] 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.

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

[10] 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.

[11] 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.

[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] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.
[32] 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.
[33] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear algebra and its applications*, vol. 284, no. 1, pp. 193–228, 1998.
[34] 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.
[35] 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.
[36] 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.
[37] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schlkopf, "Injective hilbert space embeddings of probability measures," in *In COLT*, 2008.
[38] R. B. Ash and C. A. Doléans-Dade, *Probability and measure theory*. Access Online via Elsevier, 2000.
[39] 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.
[40] 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.
[41] 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.
[42] K. Fukumizu, L. Song, and A. Gretton, "Kernel bayes' rule," *arXiv preprint arXiv:1009.5736*, 2010.
[43] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift."
[44] T. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *Advances in neural information processing systems*, pp. 487–493, 1999.
[45] 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.
[46] T. Jebara and R. Kondor, "Bhattacharyya and expected likelihood kernels," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 57–71.
[47] L. Debnath and P. Mikusiński, *Hilbert Spaces with Applications*. Elsevier Academic Press, 2005.
[48] M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," *J. Mach. Learn. Res.*, vol. 2, pp. 299–312, Mar. 2002.
[49] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.
[50] ——, "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.
[51] 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.

Jorge Guevara Jorge Guevara is a Ph.D. student of the Institute of Mathematics (IME) and Statistics of the University of São Paulo (USP), under advise of Roberto Hirata Jr. He did an internship at the National institute of applied science in Rouen (INSA) under advise of Stéphane Canu . He received a Master and a Bsc degree in Computer Science at University National of Trujillo, Perú. His research interest are kernel methods, machine learning and optimization theory.

Stéphane Canu Stéphane Canu is a Professor of the LITIS research laboratory and of the information technology department, at the National institute of applied science in Rouen (INSA). He received a Ph.D. degree in System Command from Comigne University of Technology in 1986. He joined the faculty department of Computer Science at Compiegne University of Technology in 1987. He received the French habilitation degree from Paris 6 University. In 1997, he joined the Rouen Applied Sciences National Institute (INSA) as a full professor, where he created the information engineering department. He has been the dean of this department until 2002 when he was named director of the computing service and facilities unit. In 2004 he join for one sabbatical year the machine learning group at ANU/NICTA (Canberra) with Alex Smola and Bob Williamson. In the last five years, he has published approximately thirty papers in refereed conference proceedings or journals in the areas of theory, algorithms and applications using kernel machines learning algorithm and other flexible regression methods. His research interests includes kernels machines, regularization, machine learning applied to signal processing, pattern classification, factorization for recommander systems and learning for context aware applications.

R. Hirata Jr. Roberto Hirata Jr. is an Associate Professor of the Institute of Mathematics and Statistics (IME) of the University of São Paulo (USP). He received a degree in Physics (Institute of Physics - USP) and one in Mathematics (IME-USP). He received a MSc degree in Computer Science (IME-USP-1997) under Prof. Dr. Junior Barrera for his work on morphological segmentation and fast algorithms for basic morphological operators and a PhD degree in Computer Science (IME-USP-2001) also under Dr. Barrera for his work on Aperture operators. As part of his PhD's training, he has worked under Prof. Dr. Edward Russel Dougherty at the Texas A&M University. His last works during the PhD were related to bioinformatics. He was hired in 2001 as a researcher and professor for the SENAC College of Computer Science and Technology where he co-lead the Computer Vision group. In 2004 he joined the department of Computer Science (IME - USP) and since then he has advised seven MSc and three PhD projects. He is author or co-author in more than twenty papers in journals or conferences in the last five years and his main research interests includes mathematical morphology, machine learning and computer vision.

Support Measure Data Description:Supplemental Material

Jorge Guevara, Stéphane Canu, and R. Hirata Jr.

Abstract—This note contains supplementary materials of the paper *Support Measure Data Description*.

1 Introduction

In this notes, we provide further details of the paper *Support measure data description*. We start in Section 2 using an example of a dataset whose observations are Gaussian distribution to understand the formulation of SMDD in the space of probability measures. This section could be read along Section 2 of the paper. In Section 3, we give a list of al the SMDD models in the RKHS presented in the Section 3 of the paper. Section 4 presents a formulation when it is used joint constraints for a SMDD model with chance constraints. In Section 5, we use the linear kernel in two SMDD models and, compare them with the SVDD method. Finally, all the proofs of the paper are presented in Section 6.

Table 1 shows the notation used in this note.

2 Interpretation of SMDD in the Space of Probability Measures

The following chance-constrained program is the SMDD in the space of probability measures:

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, and the random vector $X_i \sim \mathbb{P}_i$ is the uncertainty parameter for the chance-constrained model.

Interpretation of Problem 2.1 thorough an example Figure (1a) plots the probability density functions for a 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 $\Sigma_i \in \mathbb{R}^{2\times 2}$. The empirical minimum volume set of $\mathcal{E}$ given by a enclosing ball is the red circle. 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 1b 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 Equation 2 of the paper. Figure 1c 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 describe a set of probability measures.

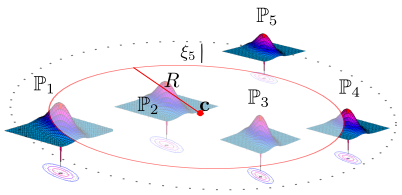
Problem 2.1 is futher simplified using Markov's inequality:

$$\frac{tr(\Sigma_i) + \|\boldsymbol{\mu}_i - \mathbf{c}\|^2}{R^2 + \xi_i} \leq \kappa_i, \tag{1}$$

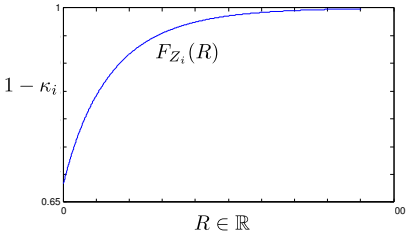
Symbol	Description
$D \in \mathbb{N}$	dimension
$N \in \mathbb{N}$	number of elements of a dataset
$R \in \mathbb{R}$	the radius of the ball
$\mathbf{c} \in \mathbb{R}^D$	the center of the ball
$\lambda \in \mathbb{R}^+$	regularization parameter
$\boldsymbol{\xi} \in \mathbb{R}^D$	vector of slack variables
$\mathcal{P}$	space of probability measures
$\mathbb{P}$	probability measure
X	random variable taking values in $\mathbb{R}^D$
κ	threshold values
$tr(A)$	trace of a matrix A
$\boldsymbol{\mu} \in \mathbb{R}^D$	mean of a set of points in $\mathbb{R}^D$
Σ	covariance of a set of points in $\mathbb{R}^D$
$\boldsymbol{\alpha} \in \mathbb{R}^N$	vector of Lagrange multipliers
$\mathcal{H}$	a Reproducing kernel Hilbert Space (RKHS)
k	a real valued positive definite on $\mathbb{R}^D \times \mathbb{R}^D$
$k(x,.) \in \mathcal{H}$	evaluation funtion at the point x in $\mathcal{H}$
$c \in \mathcal{H}$ or $c(.) \in \mathcal{H}$	the center of the ball in a RKHS
$\mu_{\mathbb{P}}$	mean function of $\mathbb{P}$ in $\mathcal{H}$
$\langle .,. \rangle_{\mathcal{H}}$	inner product on the RKHS $\mathcal{H}$
$\Sigma^{\mathcal{H}}$	covariance operator on the RKHS $\mathcal{H}$
$\tilde{k}$	real valued positive definite kernel on $\mathcal{P} \times \mathcal{P}$
$\|.\|_{\mathcal{H}}$	norm in the RKHS

TABLE 1: Symbols used in the paper.

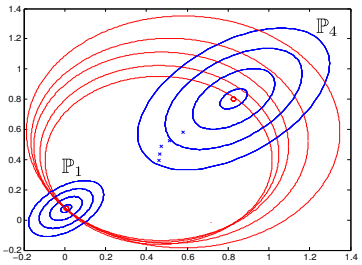
- *Jorge Guevara is with the Department of Computer Science, Institute of Mathematics and Statistics, University of Sao Paulo, Sao Paulo, Brazil. E-mail: see http://www.vision.ime.usp.br/ jorjasso/*
- *Stéphane Canu is with the Department of Computer Science, Normandie Université, INSA de Rouen - LITIS,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



(a) Red circle: empirical minimum volume set for $\{\mathbb{P}_i\}_{i=1}^5$ given by a minimum enclosing ball in the space of probability measures.



(b) The cumulative Chi-square distribution (blue curve) shows that decreasing κ_i has the effect to increase the radius R to cover some $\mathbb{P}_i$.



(c) Decreasing κ_4 increases the radius R to cover the observation $\mathbb{P}_4$.

Fig. 1

to the following optimization Problem:

Problem 2.2.

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

whose dual is

Problem 2.3.

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

Geometric intepretation of Problem 2.2 Assuming $\xi_i = 0, \ 1, 2, \ldots, N$, for each constraint, we have from (1):

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

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 $b = \sqrt{tr(\Sigma)}$ we have the equality. Figure (2) shows this interpretation.

3 List of SMDD models in a RKHS

We summarize the SMDD models in a RKHS presented in the paper.

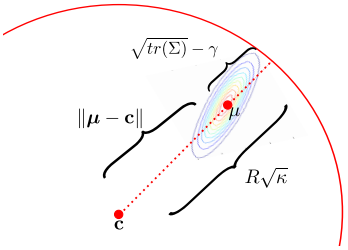


Fig. 2: Geometric interpretation of the deterministic constraints.

3.1 M1: SMDD model in a RKHS as Chance Constrained Problem

3.1.1 CCP Problem

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

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

3.1.2 Primal

Problem 3.1.

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

3.1.3 Dual

Problem 3.2.

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

3.2 M2: SMDD in a RKHS as a direct extension of SVDD

3.2.1 Primal

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

3.2.2 Dual

Problem 3.3.

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

3.3 M3: SMDD Model in a RKHS using Mean Maps with Norm One and Invariant Translation Kernels

Problem 3.4.

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

but with kernel:

$$\tilde{\tilde{k}}(\mathbb{P}_i,\mathbb{P}_j)=\frac{\tilde{k}(\mathbb{P}_i,\mathbb{P}_j)}{\sqrt{\tilde{k}(\mathbb{P}_i,\mathbb{P}_i)\tilde{k}(\mathbb{P}_j,\mathbb{P}_j)}}, \tag{2}$$

4 M1 WITH JOINT CONSTRAINTS

4.1 Primal

Problem 4.1.

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

where

$$\xi_i'=\frac{1}{2}\kappa\xi_i-\frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2} \tag{3}$$

4.2 Dual

Problem 4.2.

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

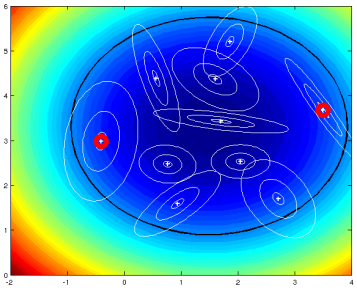
5 SMDD MODELS WITH LINEAR KERNEL

We examined here the behaviour of M1, M2 and M5 when the linear kernel is used. As we point out in the end of Section 2 of the paper, if the linear kernel is used, M2 is equivalent to a SVDD (M5) trained on the group means, because under that setting, the RKHS will be the same as the input space, consequently mean functions $\mu_{\mathbb{P}}\in\mathcal{H}$ will be $\boldsymbol{\mu}\in\mathbb{R}^D$.

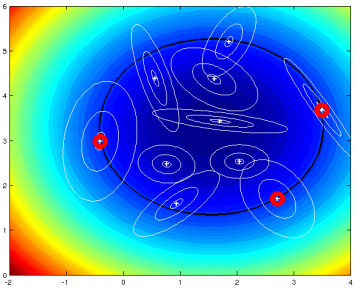
The training set was given by 10 groups of points, artificially generated from ten two-dimensional Gaussian distributions. To do that, we set arbitrarily the mean and covariance matrix for each two-dimensional distribution. Elements of the training set are plotted in 3a, 3b and 3c, as several ellipsoids, which are the contours of the probability density function of the local Gaussian distribution, such inner and outer ellipses cover about the 35% and 85% of the probability mass of each local Gaussian distribution, respectively.

Setting the regularization parameter to $\lambda=1$ for M1 and M5, Figure 3a shows a minimum enclosing ball found by M1. The radius does not cross the means of the support measures (small red circles). because M1 beside to use mean maps, also uses the trace of the covariance operator. Figure 3b shows a minimum enclosing ball found by M2. This model uses only information given by mean maps, consequently, the radius cross the means of the support measures. Therefore, the radius found by M1 is greater than the radius found by 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 $\mathbb{P}_i$, as we pointed out in Section 2 of the paper.

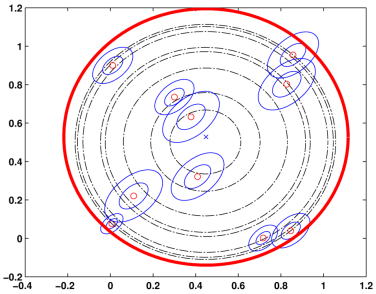
Figure 3c shows how by varying the regularization parameter in a SVDD solution is not equal to the solution provided by M1. Dashdot black circles denote several minimum enclosing balls found by M5 (SVDD), each of them obtained by varying the regularization parameter λ to allow $\{0\%,10\%,20\%,30\%,40\ \ldots,90\%\}$ of the training data to be errors. This was done by set λ proportionally to $1/(N*q)$, where N is the number of examples in the training set and q is the percentage of training examples to be considered errors. Red circle in Figure 3c denote the minimum enclosing ball found by M1 with $\lambda=1$ and all κ-values set to one. The minimum enclosing ball obtained from M1 is bigger than the one obtained from M5 in the case of 0% errors allowed, because M1 considers the information given by the trace of local covariances in the training set. We pointed out in the paper that M1 is equivalent to M5 only if all the κ values are ones, and there is no information of local covariance matrices.



(a) A MV-set given by a minimum enclosing ball from M1 with linear kernel for a kernel on probability measures, joint constraints and $\lambda = 1$, for a dataset of Gaussian distributions. Small red circles are the support measures. The means are not on the surface of the ball, because M1 considers the trace of the covariance.



(b) A minimum enclosing ball from M2 with linear kernel for a kernel on probability measures and $\lambda = 1$, for a dataset of Gaussian distributions. Small red circles are the support measures. Model M2 with this setting is equivalent to a SVDD (M5), trained using only the group means. M2 differs from M1 inthe length of the radius.



(c) Dash-dot black circles: Minimum enclosing balls for the group means of a dataset of Gaussian distributions, for several choices of λ in M5 with linear kernel. Red circle: a minimum enclosing ball from M1 with $\lambda = 1$ with linear kernel for a kernel on probability measures and $\lambda = 1$, for a dataset of Gaussian distributions.

Fig. 3

6 Proofs

6.1 Proof of Lemma 2.1

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:

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

6.2 Proof of Theorem 2.3

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

$$\begin{aligned} \partial_{\alpha_i}\mathcal{L} = &\frac{(\sum_{j=1}^N \alpha_j) 2\langle\boldsymbol{\mu}_i, \sum_{j=1}^N \alpha_j\boldsymbol{\mu}_j\rangle}{(\sum_{j=1}^N \alpha_j)^2} \\ &- \frac{\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j\langle\boldsymbol{\mu}_i,\boldsymbol{\mu}_j\rangle}{(\sum_{j=1}^N \alpha_j)^2} \\ &- \|\boldsymbol{\mu}_i\|^2 - tr(\Sigma_i) - \eta\kappa_i \\ &- \nu(\kappa_i - c) = 0 \end{aligned} \tag{4}$$

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

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

Analyzing the KKT's condition of Problem 2.3, we have that $\alpha_i > 0, \beta_i > 0 \implies \xi_i = 0$ and $0 < \alpha_i\kappa_i < \lambda$. From this and KKT's complementarity conditions we can retrieve the radius

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

where $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i\kappa_i < \lambda\}$. Thus, follows that $\eta = R^2$, then $R = \sqrt{\eta}$.

6.3 Proof of Lema 3.1

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

6.4 Proof of Theorem 4.1

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

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

6.5 Proof ot Theorem 4.2

From (3), 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:

$$\begin{aligned}\mathcal{L}(c(.),\rho,\boldsymbol{\xi},\boldsymbol{\alpha},-\boldsymbol{\beta}) &= \frac{\|c(.)\|^2_{\mathcal{H}}}{2}-\rho'+\lambda\sum_{i=1}^N(\xi''_i-\frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2}) \\ &\quad -\sum_{i=1}^N\alpha_i\{\langle\mu_{\mathbb{P}_i},c(.)\rangle_{\mathcal{H}}-\rho'+\xi''_i-\frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2}\} \\ &\quad -\sum_{i=1}^N\beta_i\xi''\end{aligned} \tag{6}$$

The optimality (KKT) conditions for this problem are:

$$\left.\begin{array}{rcrcl}\partial_\rho\mathcal{L}=0 &:& \sum_{i=1}^N\alpha_i &=& 1\\ \nabla_{c(.)}\mathcal{L}=0 &:& c(.)-\sum_{i=1}^N\alpha_i\mu_{\mathbb{P}_i} &=& 0\\ \partial_{\xi''_i}\mathcal{L}=0 &:& \lambda-\alpha_i-\beta_i &=& 0\end{array}\right\} \tag{7}$$

$$\left.\begin{array}{rcl}\alpha_i\{\langle\mu_{\mathbb{P}_i},c(.)\rangle_{\mathcal{H}}-\rho'+\xi''_i-\frac{\|\mu_{\mathbb{P}_i}\|^2_{\mathcal{H}}}{2}\} &=& 0\\ \beta_i\xi''_i &=& 0\end{array}\right\} \tag{8}$$

Replacing, (7) into (6) 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:

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