Support Measure Data Description for group anomaly detection

Jorge Guevara, Stéphane Canu, R Hirata

Support Measure Data Description for group anomaly detection

Jorge Guevara
University of Sao Paulo
IME-USP
Sao Paulo, Brazil
jorjasso@ime.usp.br

Stéphane Canu
Normandie Université
St Etienne du Rouvray,France
INSA de Rouen - LITIS
scanu@insa-rouen.fr

R. Hirata Jr.
University of Sao Paulo
IME-USP
Sao Paulo, Brazil
hirata@ime.usp.br

ABSTRACT

We address the problem of learning a data description model from a dataset containing probability measures as observations. We estimate the data description model by optimizing volume-sets of probability measures where each volume-set is defined as a set of probability measures whose representative functions in a reproducing kernel Hilbert space (RKHS) belong to an enclosing ball. We present three data description models, which are functions in a RKHS depending only on some probability measures, named *support measures* in analogy to support vectors. An advantage of the method is that we do not consider any particular form for the probability measures. We validate our method in the task of group anomaly detection, with artificial and real datasets.

Keywords

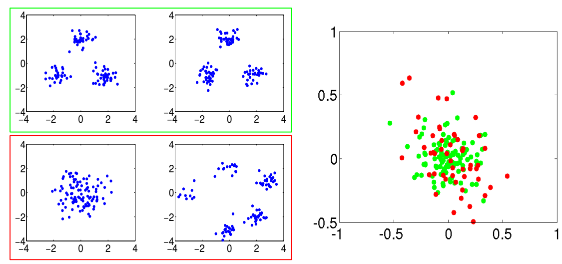
Kernel on distributions, One-class classification, support vector data description, embedding of probability measures, mean map, group anomaly detection. MV-set

1. INTRODUCTION

Data description (DD) is the task of building models to depict the common characteristics of objects in some data set aiming to perform machine learning tasks such as anomaly and novelty detection, clustering and classification [21, 24, 23, 30, 4]. The main idea of DD methods is to assume that observations are generated by an underlying unknown distribution. Consequently, a valid approach is to estimate some distribution information from a training dataset. For example, an empirical probability density function, or a density level set, or some information about the density support set.

Very often, DD methods are defined for 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 $\mathbb{R}^D$ [32, 33, 20, 19, 25, 35, 10, 34, 26, 15, 18]. Such

ODDx3'15, August 10th, 2015, Sydney, AU.
.



(a) (b)

Figure 1: a) The distributions of points of the two groups in the red box are considered anomalous with respect to distribution of the ones in the blue box. b) Overlap between the means of anomalous groups (red points) and the means of non-anomalous groups (blue points). Note how hard it is to find group anomalies using only one representative value (in this case, a mean) per group.

datasets are sets of the form:

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

where N is the number of observations, each s_i is a set $\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{L_i}^{(i)}\}$ with cardinality L_i, and $\mathbf{x}_i\mathbb{R}^D$. Practical examples of observations taking the form of s_i are: sets of image features in an image dataset [17]; sets of spatio-temporal features [16]; sets of replicate values in a measurement process [31]; sets describing point wise uncertainty [25, 35]; or sets describing the invariance of some particular object [10].

1.1 Group anomaly detection

Group anomalies can be given by [32]: 1) *point-based* anomalies, defined as being an aggregation of anomalous points; 2) *distribution-based* anomalies, defined as being an anomalous aggregation of non-anomalous points. In order to construct robust DD models for datasets given by (1) and detect group anomalies, DD methods must take into account the distribution information provided by each s_i. Figure 1 shows how the information provided by each local distribution of points is crucial to perform a right description of (1).

1.1.1 Related work

Several solutions have been proposed to this kind of estimation, including representing groups by sets of features [3, 14], or estimating group anomalies by clustering point anomalies [5]. However, such procedures heavily rely on the feature engineering process or ignores the fact that anomalous groups can be formed by non-anomalous points. Recently, a generative approach based on hierarchical probabilistic models was proposed to identify group anomalies [32, 33] Furthermore, one-class support measure machine was proposed as a discriminative approach to perform group anomaly detection [19]. Both approaches (generative, or discriminative), give state of the art results.

1.2 Contributions

This work presents three novel discriminative and non-parametric DD models for datasets of the form given by (1), named *support measure data description* (SMMD) models in analogy to support vectors in kernel methods. The potential applications of the models presented here are: clustering, classification and other related machine learning tasks for datasets of the form of (1). In this paper, we focus in the use of the SMDD model, as one-class classification models, to the task of detecting group anomaly.

The estimation of the SMDD models is based on the following assumptions: 1) the observations s_i are distributed according to an unknown probability measure $\mathbb{P}_i$; 2) the empirical measure $\widehat{\mathbb{P}}_i$ is obtained from s_i in (1), i.e. $\widehat{\mathbb{P}}_i = \frac{1}{L_i}\sum_{\ell=1}^{L_i} \delta_{\mathbf{x}_\ell}(s_i)$ approximates well the true unknown probability measure $\mathbb{P}_i$; 3) the representative functions, $\mu_{\mathbb{P}_i}$, of the probability measures in a RKHS can be used to find the description of the set $\{\mathbb{P}_i\}_{i=1}^N$ and hence a DD model for (1); 4) the DD model of the set $\{\mathbb{P}_i\}_{i=1}^N$ can be estimated by optimizing volume-sets of probability measures, where each volume-set is constructed using the information provided by enclosing balls of the representative functions in a RKHS; 5) the DD model will be a function in a RKHS relying only on a subset of representative functions of probability measures.

The paper is organized as follows: Section 2 gives some background in Hilbert space embedding for probability measures. All the SMDD's 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 variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as a Borel measurable map: $\Omega \to \mathbb{R}^D$, satisfying $X(\omega) = \omega$, $\forall \omega \in \Omega$, i.e, X is an identity map. We use $\mathcal{B}(\mathbb{R}^D)$ to denote the Borel σ-algebra of $\mathbb{R}^D$. In addition, we always assume that $\Omega = \mathbb{R}^D$ and $\mathcal{F} = \mathcal{B}(\mathbb{R}^D)$, implying, for all $B \in \mathcal{B}(\mathbb{R}^D)$, the induced probability measure by X given by $\mathbb{P}_X(B) = \mathbb{P}\{\omega : X(\omega) \in B\}$ is equal to $\mathbb{P}(B)$, i.e., $\mathbb{P}_X = \mathbb{P}$. We always write $\mathbb{P}(a < X \leq b)$ instead of $\mathbb{P}\{\omega : a < X(\omega) \leq b\}$. Letter $\mathcal{H}$ denotes a RKHS of functions $f : \mathbb{R}^D \to \mathbb{R}$, with reproducing kernel $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, norm $\|.\|_{\mathcal{H}}$, and inner product $\langle\cdot\,,\cdot\rangle_{\mathcal{H}}$. We denote by $k(., s)$ the mapping $t \to k(t, s)$ with fixed s. Notation $\mathbb{E}_{X\sim\mathbb{P}}[f(X)]$ means the expectation of $f(X)$, where X is distributed according to $\mathbb{P}$

Remark. We do not include any proofs in this paper because of size restrictions. They are available upon request.

2. KERNEL HILBERT SPACE EMBEDDING

This section gives a little background on Hilbert space embedding for probability measures. The main concepts presented here are *mean map*, *Hilbert space embedding* and *kernel on probability measures.*

Hilbert space embedding of probability measures [13, 29, 1, 26], gives a way to represent probability measures $\mathbb{P}_i$ as functions in a RKHS. Such functions are commonly named as *representative functions*, *mean functions* or *mean maps.* We present them in the following definition.

Definition 1 (Mean map). *Let $\mathbb{P}$ be a probability measure and $X \sim \mathbb{P}$. The mean map in $\mathcal{H}$ is the function:*

$$\begin{aligned} \mu_{\mathbb{P}} : \mathbb{R}^D &\to \mathbb{R} \\ t &\mapsto \mu_{\mathbb{P}}(t) = \mathbb{E}_{\mathbb{P}}[k(X, t)] = \int_{\mathbf{x}\in\mathbb{R}^D} k(\mathbf{x}, t) d\mathbb{P}(\mathbf{x}), \end{aligned} \tag{2}$$

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 [12, 26, 28]. 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 Hilbert space embedding for probability measures is given in the following definition.

Definition 2. *The embedding of probability measures $\mathbb{P} \in \mathcal{P}$ in $\mathcal{H}$ is given by the mapping*

$$\begin{aligned} \mu : \mathcal{P} &\to \mathcal{H} \\ \mathbb{P} &\mapsto \mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[k(X, .)] = \int_{\mathbf{x}\in\mathbb{R}^D} k(\mathbf{x}, .) d\mathbb{P}(\mathbf{x}). \end{aligned}$$

Hence, $\mu_{\mathbb{P}}$ acts as the representative function in $\mathcal{H}$ for $\mathbb{P}$. Choosing *characteristic kernels* [8, 27, 28] for k, makes the embedding μ injective. Some examples of characteristic kernels are the Gaussian, Laplacian, inverse multiquadratics, B_{2n+1}-splines kernels. See [28] for details. 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 an empirical estimator of $\mu_{\mathbb{P}}$, is bounded [26].

2.1 Kernel on probability measures

The mapping

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

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

$$\begin{aligned} \tilde{k}(\mathbb{P}, \mathbb{Q}) &= \langle \mathbb{P}, \mathbb{Q}\rangle_{\mathcal{P}} = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} \\ &= \int_{\mathbf{x}\in\mathbb{R}^D}\int_{\mathbf{x}'\in\mathbb{R}^D} k(\mathbf{x}, \mathbf{x}')d\mathbb{P}(\mathbf{x})d\mathbb{Q}(\mathbf{x}') \end{aligned} \tag{3}$$

is positive definite [1].

3. SMDD MODELS

In this section, we introduce three DD models, for datasets given by sets of probability measures. We call these models *Support Measure Data Description Models* (SMDD's). Those models are based on the concept of *minimum volume-set* and enclosing balls for the representative functions $\mu_{\mathbb{P}_i}$

of probability measures. SMDD is a data description model given by a function in a RKHS relying only in some subset from the training set: the *support measures.*

3.1 Minimum Volume Sets

Volume-sets are widely used to find a description of datasets of the form $\{\mathbf{x}_i\}_{i=1}^N,\ \mathbf{x}_i \in \mathbb{R}^D$ [21, 24, 23]. A minimum volume-set (MV-set) is a volume-set satisfying some optimization criteria over all the possible volume-sets. The class of sets used in DD methods ranges from convex sets [21] to sets implicitly defined in a RKHS via positive definite kernels [23, 19, 30].

We assume that the points in s_i, are i.i.d. [1] realizations of a random variable $X \sim \mathbb{P}_i$. A generalization of the definition of MV-set given in [21, 24, 23] to the case of probability measures is stated below.

Definition 3 (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*

$$G_\alpha^* = \underset{G\in\mathcal{A}}{\operatorname{argmin}}\{\rho(G) | \mathcal{E}(G) \geq \alpha\}, \tag{4}$$

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 $\mathcal{E}$.

To Compute a MV-set of a set of probability measures with the above procedure is very general. Therefore, we limit our attention to the class of sets $\mathcal{A}$ formed by sets of probability measures satisfying some certain criteria. We will assume that $\{\mathbb{P}_i\}_{i=1}^N$ is an i.i.d. sample distributed according to $\mathcal{E}$ (Def. 3), where each $\mathbb{P}_i$ is unknown. As $G \in \mathcal{A}$ is some set of probability measures, a first empirical[2] approximation for G in (4) is given by:

$$\hat{G}_0(R, \mathbf{c}) = \{\mathbb{P}_i \in \mathcal{P} \mid \|X_i - \mathbf{c}\|^2 \leq R^2\}, \tag{5}$$

where we consider a hypersphere of radius $R \in \mathbb{R}^+$ and center $\mathbf{c} \in \mathbb{R}^D$. A MV-set will be found optimizing over R and $\mathbf{c}$. However, (5) has two main drawbacks: it does not consider complex models, and some $\mathbb{P}_i$ will be in (5), if only if all possible realizations of $X_i \sim \mathbb{P}_i$ are inside the hypersphere $(R, \mathbf{c})$. Such limitations are overtaken considering the following three classes of sets described below.

The first class of volume-sets is defined by considering only the representative functions or mean maps $\mu_{\mathbb{P}_i}$ of each $\mathbb{P}_i$, and is given as follows:

$$\hat{G}_1(R, c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|_{\mathcal{H}}^2 \leq R^2\}, \tag{6}$$

The second class considers mean maps with norm one (we explain the motivation for this in Section 3.3).

$$\hat{G}_2(R, c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|_{\mathcal{H}}^2 \leq R^2,\ \|\mu_{\mathbb{P}}\|_{\mathcal{H}}^2 = 1\}. \tag{7}$$

The third class considers bounding values $\mathcal{K} = \{\kappa_i\}_{i=1}^N,\ \kappa_i \in [0,1]$. Thus, $\mathbb{P}_i$ is in the volume-set G, if a subset of the realizations of the random variable $k(X, \cdot), X \sim \mathbb{P}_i$ is inside the hypersphere $(R, \mathbf{c})$, with probability less than $1 - \kappa_i$.

$$\hat{G}_3(\mathcal{K}) = \{\mathbb{P}_i \in \mathcal{P} \mid \mathbb{P}_i(\|k(X_i, .) - c\|_{\mathcal{H}}^2 \leq R^2) \geq 1 - \kappa_i\}. \tag{8}$$

[1]Independent and identically distributed.
[2]Empirical in the sense of sample $\{\mathbb{P}_i\}_{i=1}^N$.

All three formulations use a Hilbert space embedding for probability measures, with the advantage that the knowledge of the the density $\mathbb{P}_i$ is not explicitly needed.

3.2 First model SMDD

The MV-set $\hat{G}_\alpha^*$ for volume-sets G of the form given by (6) can be computed solving the following optimization problem. Given the mean functions $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ of $\{\mathbb{P}_i\}_{i=1}^N$, the SMDD model is:

Problem 1.

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

Proposition 1 (Dual form). *The dual form of the previously problem is given by:*

Problem *2.*

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

where $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}\rangle_{\mathcal{H}}$ by (3)*, and $\boldsymbol{\alpha}$ is a Lagrange multiplier vector with non negative components α_i.*

Proposition 2 (Representer theorem). *The representer theorem for Problem 1 is:*

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

where $\mathcal{I} = \{1, 2, \ldots, N\}$. Furthermore, all $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ are inside the MV-set $\hat{G}_\alpha^*$. All $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i = \lambda\}$ are the training errors. All $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid 0 < \alpha_i < \lambda\}$ are the support measures.

Theorem 3. *Let η be the Lagrange multiplier of the constraint $\sum_{i=1}^N \alpha_i = 1$ of Problem 2, then $R^2 = -\eta + \|c\|_{\mathcal{H}}^2$.*

Consequently, to decide if some test probability measure $\mathbb{P}_t$ is in the SMDD model, we have to compute the score $\|\mu_{\mathbb{P}_t} - c\|_{\mathcal{H}}^2$, which, using Proposition 2 and Theorem 3 can be written in terms of the kernel $\tilde{k}$ by:

$$\tilde{k}(\mathbb{P}_t, \mathbb{P}_t) - 2\sum_i \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_t) + \sum_{i,j} \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j), \tag{9}$$

where indices i, j belongs to the support measure set. This score must be compared against the value R to decide if $\mathbb{P}_t$ is in the description of SMDD.

Note that, if the linear kernel: $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}'\rangle$ on $\mathbb{R}^D \times \mathbb{R}^D$ is used in (3), Problem 2 is equivalent to the dual problem of SVDD [30], 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 $\langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j\rangle$.

3.3 Second SMDD Model

This SMDD model considers mean maps with norm one, i.e., $\|\mu_{\mathbb{P}_i}\|^2 = 1$ and Stationary kernels[9], which are kernels of the form $k_I(\mathbf{x}, \mathbf{x}') = f(\mathbf{x} - \mathbf{x}')$, that is, they only depend on the difference $\mathbf{x} - \mathbf{x}'$.

Implicit feature maps of stationary kernels are functions $k_I(\mathbf{x}, \cdot)$ in a RKHS lying on a surface of a hypersphere because they have constant norm. To see that, note that stationary kernels satisfy:

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

where ϵ is a constant value. So $\|k_I(\mathbf{x}, \cdot)\|_{\mathcal{H}} = \sqrt{|\epsilon|}$, consequently, functions $k_I(\mathbf{x}, \cdot)$ lie on a surface of a hypersphere of radius $\sqrt{|\epsilon|}$. However, mean maps $\mu_{\mathbb{P}} = E_{\mathbb{P}}[k_I(X, .)]$, do 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 [19].

Theorem 4 (Spherical Normalization [19]). *If kernel $k(.,.)$ is characteristic and the examples are linearly independent in the RKHS $\mathcal{H}$, then the spherical normalization:*

$$\frac{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}}}{\sqrt{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}}\rangle_{\mathcal{H}} \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}}}}, \tag{10}$$

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

Basically, Theorem 4 says that all the information is preserved after performing spherical normalization of the data.

The MV-set $\hat{G}^*_\alpha$ for volume-sets G of the form given by (7) can be computed by solving the optimization problem similar as the one given in Problem 2 but with kernel:

$$\tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) = \frac{\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)}{\sqrt{\tilde{k}(\mathbb{P}_i, \mathbb{P}_i)\tilde{k}(\mathbb{P}_j, \mathbb{P}_j)}}, \tag{11}$$

because of Theorem 4. Furthermore, note that $\tilde{k}$ is given by (3) but with kernel k_I.

As $\sum_{i=1}^N \alpha_i \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_i)$ is constant in Problem 2 when a kernel $\tilde{\tilde{k}}$ is used, the MV-set $\hat{G}^*_\alpha$ can be computed by the following optimization problem:

Problem 3.

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

This formulation is very similar to the dual formulation of One-class Support Measures Machines [23, 19] but is not directly equivalent. We discuss this point in Section 4.

3.4 Third SMDD model

The MV-set $\hat{G}^*_\alpha$ for volume-sets G of the form given by (8) can be computed solving the chance-constrained optimization problem. 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 SMDD model is:

Problem 4.

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

The chance constraints of Problem 4 control the probability of constraint violation, allowing flexibility to the model. However, each constraint requires we deal with every possible realization of $k(X, \cdot), X \sim \mathbb{P}_i$. To implement this problem, it is necessary to turn probabilistic constraints into deterministic ones.

For a non negative random variable $X \sim \mathbb{P}$ and $t > 0$, this can be achieved by Markov's inequality which bounds $\mathbb{P}(X \geq t)$ by $\mathbb{E}_{\mathbb{P}}[X]/t$.

$$\mathbb{P}_i(\|k(X_i, .) - c(.)\|^2_{\mathcal{H}} \geq R^2 + \xi_i) \leq \frac{\mathbb{E}_{\mathbb{P}}[\|k(X_i, .) - c(.)\|^2_{\mathcal{H}}]}{R^2 + \xi_i}, \tag{12}$$

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

3.4.1 Trace of the Covariance Operator

The term $\mathbb{E}_{\mathbb{P}}[\|k(X_i, .) - c(.)\|^2_{\mathcal{H}}$ in the numerator of (12) 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 $\Sigma^{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$, 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 the reproducing property [3]. The covariance operator is subsequently the matrix:

$$\Sigma^{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[k(X, .)k(X, .)^\top] - \mathbb{E}_{\mathbb{P}}[k(X, .)]\mathbb{E}_{\mathbb{P}}[k(X, .)]^\top. \tag{13}$$

From this, the trace of $\Sigma^{\mathcal{H}}$ can be obtained as: [4]

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

where the last line is due to the reproducing property and Def. 1. Therefore, using (3), yields

$$tr(\Sigma^{\mathcal{H}}) = \mathbb{E}_{\mathbb{P}}[k(X, X)] - \tilde{k}(\mathbb{P}, \mathbb{P}), \tag{14}$$

that is, the trace of a possible infinite dimensional matrix can be computed in terms of kernel evaluations. We then have the following lemma.

Lemma 5.

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

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

[4] Because $\mu_{\mathbb{P}}(X) < \infty$, it follows that $tr(\Sigma^{\mathcal{H}}) < \infty$.

3.4.2 Deterministic Form

From Lemma (5), the deterministic form of the Problem 4 is the following optimization problem. 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 SMDD model is:

PROBLEM 5.

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

for all $i = 1,\ldots,N$, where $tr(\Sigma_i^{\mathcal{H}})$ is given by (14).

PROPOSITION 6 (DUAL FORM). *The dual form of Prob. 5 is given by*

PROBLEM *6.*

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

where $\langle\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}\rangle_{\mathcal{H}}$ is computed by $\tilde{k}(\mathbb{P}_i,\mathbb{P}_j)$, $\boldsymbol{\alpha}$ is a Lagrange multiplier vector with α_i non negative components; and $tr(\Sigma_i^{\mathcal{H}})$ is given by (14).

A remark about the nature of that problem it that it is a fractional programming problem [7].

PROPOSITION 7 (REPRESENTER THEOREM). *The representer theorem for Problem 5 is:*

$$c(.) = \frac{\sum_i \alpha_i\mu_{\mathbb{P}_i}}{\sum_i\alpha_i},\ i\in\{i\in\mathcal{I}\mid 0<\alpha_i\kappa_i\le\lambda\}, \tag{15}$$

where $\mathcal{I} = \{1,2,\ldots,N\}$. Furthermore, all $\mathbb{P}_i$, $i\in\{i\in\mathcal{I}\mid\alpha_i=0\}$ are inside the MV-set $\hat{G}_\alpha^*$. All $\mathbb{P}_i$, $i\in\{i\in\mathcal{I}\mid\alpha_i\kappa_i=\lambda\}$ are the training errors. All $\mathbb{P}_i$, $i\in\{i\in\mathcal{I}\mid 0<\alpha_i\kappa_i<\lambda\}$ are the support measures *and, from this, the radius is computed by*

$$R^2 = \frac{\|\mu_{\mathbb{P}_i} - c(.)\|^2 + tr(\Sigma_i^{\mathcal{H}})}{\kappa_i}, \tag{16}$$

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

Alternatively, we have the following result to compute R.

THEOREM 8. *Let η be the Lagrange multiplier of the constraint $\sum_{i=1}^N\alpha_i\kappa_i = 1$ of the Lagrangian of Problem 6, then $R^2 = -\eta$.*

As a consequence, to test if some test probability measure $\mathbb{P}_t$ is in this SMMD model, we have to compute the score $\|\mu_{\mathbb{P}_t} - c\|_{\mathcal{H}}^2 + tr(\Sigma_t^{\mathcal{H}})$. Using Prop. 7, Theorem 8, and Eq. (14), the score can be written in terms of the kernel $\tilde{k}$ by:

$$\tilde{k}(\mathbb{P}_t,\mathbb{P}_t) - 2\sum_i\alpha_i\tilde{k}(\mathbb{P}_i,\mathbb{P}_t) + \sum_{i,j}\alpha_i\alpha_j\tilde{k}(\mathbb{P}_i,\mathbb{P}_j) + tr(\Sigma_t^{\mathcal{H}}) \tag{17}$$

where indices i,j belong to the support measure set. This score must be compared against the value R to decide if $\mathbb{P}_t$ is in the description of SMDD.

4. EQUIVALENCES AMONG MODELS

In this section, we describe the relationship among SMDD models and the equivalence between SMDD models and One-Class Support Measure Machine (OCSMM) [23, 19]. For this purpose, we use the notation given in Table 1. We start showing how M1 can be formulated if we restrict it only to the case of joint constraints and a sharing covariance matrix. We then use this formulation to compare the restricted M1 with the original M1 and M2.

THEOREM 9. *The Primal form of M1 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 *7.*

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

where

$$\xi_i' = \frac{1}{2}\kappa\xi_i - \frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2} \tag{18}$$

Problem 7 is a less flexible formulation of M1 because it considers the same local covariance and the same κ values for all points. Using optimal $c\in\mathcal{H}$ and ρ' values from Problem 7, the radius is computed by:

$$R = \sqrt{(tr(\Sigma) + \|c\|^2 - 2\rho'))/\kappa}, \tag{19}$$

or equivalently, solving Problem 5 for $\kappa_i = \kappa$ and $\Sigma_i = \Sigma$, for all $i = 1,2\ldots,N$, we can retrieve ρ' of Problem 7 as follows:

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

COROLLARY 10 (DUAL FORM). *Using the kernel between probability measures given by* (3)*, the dual of Prob. 7 is given by:*

PROBLEM *8.*

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

LEMMA 11. *Let η be the Lagrange multiplier of constraint $\sum_{i=1}^N\alpha_i = 1$ of the Lagrangian of Prob. 8, then $\rho = \eta$.*

From this, we can solve Prob. 8 and apply Lemma 11 to retrieve ρ, the center by $c = \sum_i \alpha_i \mu_{\mathbb{P}_i},\ i \in \{i | 0 < \alpha_i \leq \lambda\}$, and the radius R from (19).

Under this *particular setting for M1*, we have the following equivalence among the SMDD models:

- **M1 vs M2**, M1 is almost the same as M2 but with a difference of a scaling factor of 0.5 in the dual objective function;
- **M1 vs M3**, after spherical normalization on data, the dual objective function of M1 as is given by Prob. 8, becomes $-0.5 \sum_{i,j=1}^{N} \alpha_i \alpha_j \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j)$, where $\tilde{\tilde{k}}$ is the kernel given by (11), because the other term in the objective function is constant. Therefore, M1 is equivalent to M3, with a difference of a scaling factor of 0.5 in the dual objective function.

We conclude this section describing how SMDD models are equivalent to OCSMM. It is widely known that SVDD [30] and One-Class Support Vector Machines (OCSVM) [23] are similar if stationary kernels are used [23, 30]. Although, Prob. 7 is similar to OCSMM, SMDD is not directly equivalently to it because mean maps under stationary kernels do not have constant norm. However, under a spherical normalization on data, there is the following equivalence:

Corollary 12. *M2, M3 and OCSMM [19] are equivalent under a spherical normalization of the training set* $\{\mathbb{P}_i\}_{i=1}^{N}$ *by (4).*

Consequently, M1 under the assumptions given by Prob. 8 is equivalent to OCSMM, with a difference of a scaling factor of 0.5 in the dual objective function.

5. SUPERVISED GROUP ANOMALY DETECTION EXPERIMENTS

In this section, we present an experimental evaluation of SMDD models for the task of group anomaly detection using artificial and real datasets. In the experiments, we consider two types of group anomalies: Point-based anomaly detection, described in Section 5.3 and Distribution-based anomaly detection described in Section 5.4. Finally, in Section 5.5, we use real data from the *Sloan Digital Sky Survey* (SDSS) project to detect anomalous groups of galaxies.

5.1 Kernel and covariance estimation

The kernel between probability measures given by (3) was estimated via the empirical estimator:

$$\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \approx \frac{1}{L_i L_j} \sum_{l=1}^{L_i} \sum_{l'=1}^{L_j} k(\mathbf{x}_l^{(i)}, \mathbf{x}_{l'}^{(j)}), \tag{20}$$

from a training set given by (1). Furthermore, the trace of the covariance operator in the RKHS given by (14) was estimated by:

$$\begin{aligned} tr(\Sigma_i^{\mathcal{H}}) \approx & \frac{1}{L_i - 1} \sum_{l=1}^{L_i} k(\mathbf{x}_l^{(i)}, \mathbf{x}_l^{(i)}) \\ & - \frac{1}{L_i(L_i - 1)} \sum_{l=1}^{L_i} \sum_{l=1}^{L_i} k(\mathbf{x}_l^{(i)}, \mathbf{x}_l^{(i)}). \end{aligned} \tag{21}$$

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

Model	Problem	Section/Ref.
M1	6	3.4
M2	2	3.2
M3	3	3.3
OCSMM	-	[19]
SVDD	-	[30]

Table 1: Models used in the experiments

5.2 Experimental settings

The notations for the DD models used throughout this section are given by Table 1. For comparison purposes, we use SVDD, trained using only the empirical group means, as the baseline. Because some experimental results for group anomaly detection between OCSMM and other approaches, including a state of the art method proposed in [33] were reported in [19], we only compare SMDD models against the OCSMM model.

We used *CVX*, a package for specifying and solving convex programs [11] to solve M1. To solve M2, M3, OCSMM and SVDD, we used the *SVM and Kernel Methods Matlab Toolbox* (SVM-KM) [2] [5].

Remark Because we model anomaly detection as a one-class classification problem (only the non-anomalous class is labeled), it is difficult to build a confusion matrix to get statistics. However, an approach based on testing with artificial group anomalies will reflect the power of the presented models.

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

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 challenging problem. In *Point-Based Group Anomaly* detection [32], anomalous groups are given by aggregating individually anomalous points. For this experiment, we generated 50 non-anomalous groups of points and 30 groups for test. From the 30 groups in the test set, 20 groups correspond to anomalous groups. The number of points by group for all non-anomalous and anomalous groups was randomly chosen from a Poisson distribution with parameter $\beta = 10$.

The points for non-anomalous groups were randomly sampled from a *Multimodal Gaussian Mixture Distribution* or GMD. We considered *two types* of non-anomalous groups, following the same experimental setting described in references [33, 19]. The first type was given by groups sampled 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 denotes the 2×2 identity matrix. The second type was given by groups of points sampled from a GMD with the same parameters, but with mixture weights: $(0.33, 0.03, 0.64)$. The probability of chosen each group was $\pi = (0.48, 0.52)$, respectively. The green box in Fig. 2 shows three non-anomalous

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

groups for $\pi = 0.48$ and the yellow box shows two non-anomalous groups for $\pi = 0.52$.

We generated three different types of anomalous groups. The first type of group anomalies was given by 10 groups of points sampled from the normal distribution: $\mathcal{N}((-0.4, 1), I_2)$. Figure 2 shows five anomalous groups of this type (magenta box). The second type of group anomalies was given by five groups of points sampled from a two-dimensional GMD with four components, with the following parameters: weights: $(0.1, 0.08, 0.07, 0.75)$; means: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$, $(0.6, -1)$; and a sharing covariance matrix given by $0.2 * I_2$. Figure 2 shows five anomalous groups of this type (blue box). The third type of group anomalies was given by five groups of points sampled from a two-dimensional GMD with four components with parameters: 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. Figure 2 shows five anomalous groups of this type (red box).

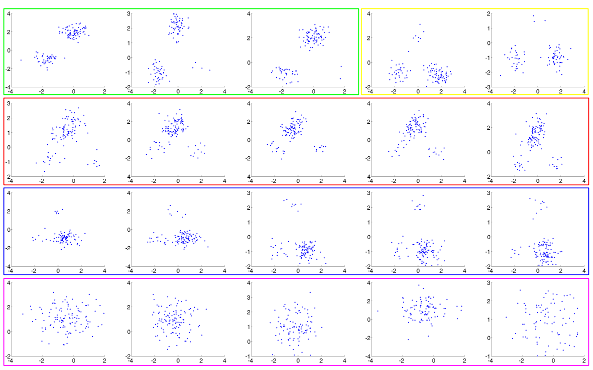


Figure 2: Group anomaly detection dataset. Green and yellow boxes contain non-anomalous groups of points. Red, blue, and magenta boxes contain anomalous groups of points.

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

As it is usual in one-class classification tasks, it is not possible to have a validation set for model selection because the data (training or test) have no labels. We follow the same methodology used in literature, that is, we choose arbitrarily a value for the regularization parameter λ of the SMDD model and computed the kernel parameters using some heuristic on the available data. In this way we avoid to employ the training or the test set for model selection (See Chap. 7, pp. 215-219 [22] and Sec. 5 in [19].).

A regularization parameter $\lambda = 1$ was considered and the kernel in (3), implemented by a RBF kernel with bandwidth parameter γ, was computed by:

$$\gamma = 1/s(\|\mathbf{x}_k^{(i)} - \mathbf{x}_l^{(j)}\|^2), \tag{22}$$

where i, j are the indices of the groups, k, l are the indices of the points , and s is the 0.1 quantile of the Euclidean distance between all possible pairs of points in the dataset.

Figure 3 shows the AUC, the ACC for non-anomalous groups, the ACC for anomalous groups, and the plot of the means of the non-anomalous groups (green points) vs. the means of anomalous groups (Red, blue, and magenta points corresponding to the red, blue, and magenta boxes in Fig. 2.). This experiment shows that all the SMDD models (M1, M2, and M3) can detect well such anomalies. The AUC values close to one indicate that the SMDD models and also OCSMM (M4) detected group anomalies with few false positives and false negatives. On the other hand, SVDD (M5) could not detect those group anomalies using only the group means as the training set. Because the means of the non-anomalous groups overlap with the anomalous groups, methods such as OCSMM and SVDD will not perform well. The reason for this is that, for such methods, anomalies are points far away from the mean of the description of the data.

5.4 Distribution-Based Group Anomaly Detection over a Gaussian Mixture Distribution dataset

Distribution-Based Group Anomalies [32] are anomalous groups of points that individually are non-anomalous but together form anomalous groups. In this experiment, 50 non-anomalous groups of points were generated to form the training set and 15 anomalous groups of points plus 15 non-anomalous groups of points were generated to form the test set. The number of points per group was the same as in the last experiment.

Points in each non-anomalous group were sampled from a two-dimensional *GMD* with three components and the following parameters: mixture weights: $p = \{1/3, 1/3, 1/3\}$; means: $(-1.7, 1), (1.7, -1), (0, 2)$ and sharing the same covariance matrix $0.2 * I_2$.

To build the group anomalies, groups of points were sampled from the same GMD used to generate non-anomalous groups. Next, we rotated all the points belonging to the set containing all the non-anomalous groups by 45 degrees, that is, $\mathbf{x}_i^R = \mathbf{x}_i^\top R$, where R is a rotation matrix of 45 degrees. Further, we estimated the covariance matrix of those rotated points. Finally, group anomalies were sampled from the same GMD of the non-anomalous groups but with two of their covariance matrices given by the covariance matrices of the rotated points. That is, individually, points are non-anomalous, but an aggregation of them is anomalous.

For this experiment, we used a kernel given by (3) implemented by a Gaussian kernel with parameter given by (22) but with s given by the median of the Euclidean distance between all possible pairs of points in the dataset. Furthermore, we used a regularization $\lambda = 1$ for all the models.

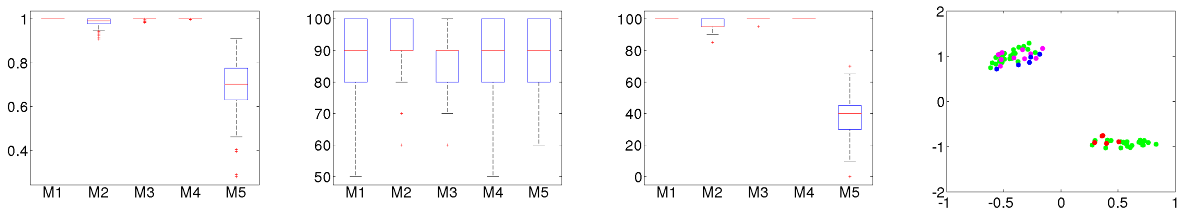
Figure 4 shows the performance metrics AUC, ACC for non-anomalous groups, and ACC for anomalous groups. In addition, it is shown the group means of non anomalous groups (green points), and the group means of anomalous groups (red points).

All the three SMDD models and also the OCSMM (M4) presented good performance in terms of AUC metric. On the other hand, the results show that SVDD (M5) was the model with the worst performance.

5.5 Group Anomaly Detection in Astronomical Data

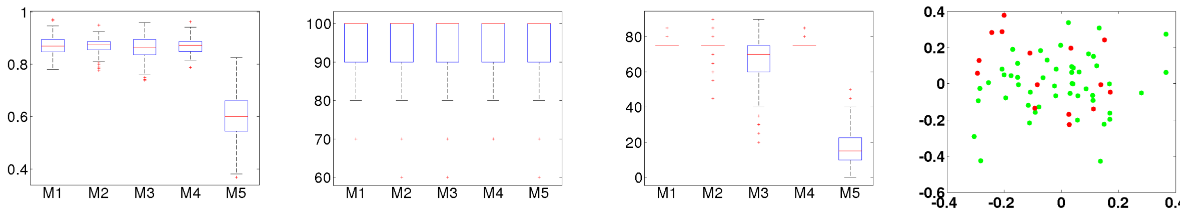
In this section, we tested the SMDD models with real data: *The Sloan Digital Sky Survey*[6] *(SDSS)* project, previ-

[6]http://www.sdss3.org/



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

Figure 3: Experimental results and a plot of the group means for the point-based group anomaly detection experiment.



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

Figure 4: Experimental results and plot of the group means for the two distribution-based anomaly detection experiment.

ously used for comparison in [20, 33, 19].This dataset contains massive spectroscopic surveys of the Milky Way galaxy and extra solar planetary systems. The idea is to use the dataset to detect anomalous clusters of galaxies. The dataset contains about 7×10^5 galaxies, each of them represented by a 4000-dimensional vector denoting spectral information. Following [20], each vector was down-sampled to a 500-dimensional vector and clusters of galaxies were obtained analyzing the spatial neighborhood of galaxies. The analysis returns 505 clusters of galaxies of a total of 7530 galaxies. Thus, each cluster of galaxies corresponds to one group of about $10 - 15$ galaxies. Finally, PCA was applied to the vectors to get a four-dimensional dataset, preserving about 85% of the variance of the data.

The training set was formed by randomly choosing 455 groups of galaxies among the first 505 groups. Furthermore, two test datasets, each of them containing the remaining 50 non-anomalous groups, from the original 505 groups, plus 50 anomalous groups, were generated.

In the first test dataset, each anomalous group was generated by randomly selecting about n_i galaxies from the 7530 galaxies, where n_i is distributed according to a Poisson distribution with parameter $\beta = 15$. As galaxies were randomly chosen, the aggregation itself of such galaxies is anomalous.

Anomalous groups for the second test dataset were generated as follows: first, we empirically estimated the covariance of the 7530 observations (galaxies) and, then, we selected randomly three sets of galaxies from the 7530 galaxies, each one containing about n_i galaxies (the same n_i of the last experiment). We estimated the empirical means of the three sets and using them and the empirical covariance matrix Σ, we constructed a GMD with three components and weights: $p = \{0.33, 0.33, 0.33\}$ and a covariance matrix $5 * \Sigma$. Finally, we generated anomalous groups of points for the second test dataset from the above GMD with about n_i points per group.

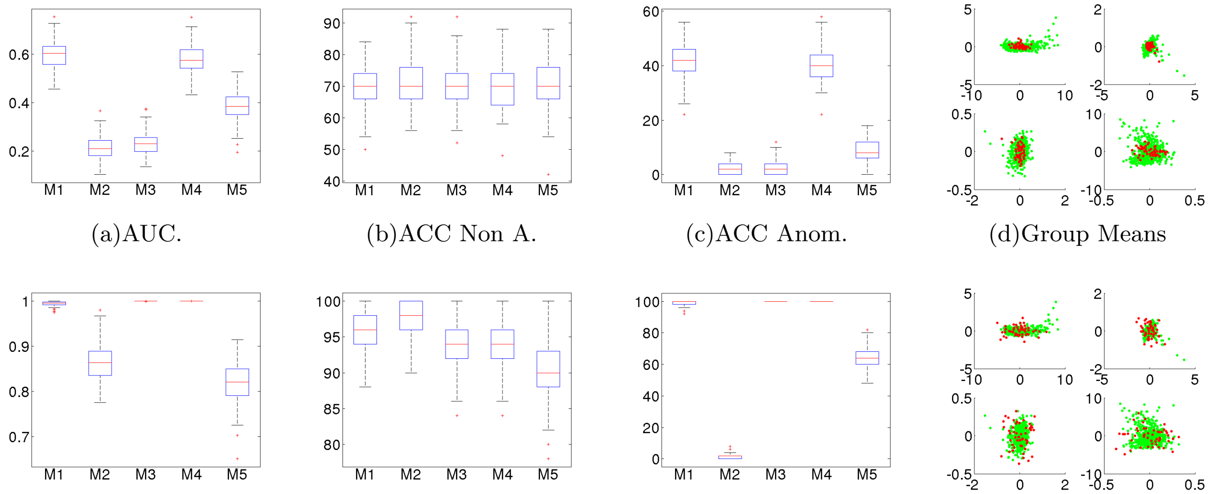
We show in Sub-figures 5d and 5h, the group means of the PCA vectors. Green points are the non-anomalous group means, and red points are the anomalous group means. Each sub-figure shows four plots: upper-left: the plot of the first vs. second dimensions, upper-right: the plot of the second vs. third dimensions, bottom-left: the plot of the third vs. four dimensions, bottom-right: the plot of the four vs. first dimensions. Moreover, because the overlapping of the group means, group anomalies for this experiment are hard to be detect by common methods.

We carried out 200 runs to get reliable statistics. Figure 5 shows that performance metrics for the first test set (top), and the second test set (bottom).

It is important to emphasize that M4 was compared against other group anomaly generative method detectors [19] and obtained equivalent performance. Therefore, we compare only SMDD models against M4 and M5 models.

For the first test set, we computed the RBF kernel parameter using (22) but with s being the median. We considered a regularization parameter letting about 30% of the non-anomalous groups to be the errors allowed in the training set. Models M2 and M3 performed a little worse than SVDD (M5) for this choice of parameters when detecting group anomalies. However, the AUC metric for SVDD shows that the performance of this model is no more than chance. On the other hand, M1 and OCSMM (M4) perform better than the baseline for detecting group anomalies. Note that the ACC for the non-anomalous groups is about 70% because the choice of the regularization parameter.

Results for the second test set are shown in the bottom part of Fig. 5. The experimental setup is the same as before but now we considered a regularization parameter $\lambda = 1$ and



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

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

Figure 5: The results of the experiment for the group anomaly detection task over a SDSS III dataset.

a kernel parameter given by (22). The ACC for anomalous groups shows that M2 is the worst for detecting the group anomalies. The AUC metric shows that all the models performed well. Furthermore, we note that a spherical normalization has a positive effect, increasing M3 AUC value close to one.

6. CONCLUSION

In this work, we presented a data description method named SMDD for datasets given as set of points, that is, each observation is considered to be a set of points distributed according to an unknown probability measure. SMDD models describe datasets of probability measures by optimizing volume-sets of probability measures. Such volume-sets are constructed using the information provided by the representative functions or mean maps of probability measures in a RKHS. In this work, we considered the class of sets of probability measures given by enclosing hyperspheres of mean functions in a RKHS. The main advantage of our approach is that it does not require a density estimation for $\mathbb{P}_i$. However, the description will be dependent in the choice of the kernel.

We formulated and described three SMDD models. The first is a direct extension of the SVDD method for the case of probability measures. This model also uses the mean map embedding of probability measures technique. The second SMDD model is almost the same as the first one but it considers a scaling of data and stationary kernels. The reason behind this, is that mean maps under stationary kernels do not have a constant norm in the RKHS. The third model uses information of covariance matrices and mean maps. This model is formulated as a chance constrained program, which is further transformed into a deterministic problem by Markov's inequality. We also compared the relationship among models, showing the cases where the SMDD models are equivalent.

The SMDD models were tested in the challenging group anomaly detection task. We showed empirically that they perform well for such a task, showing that the SMDD method is an alternative methodology to deal with group anomaly detection. Experimental evaluation, using those datasets, shows that SMDD model M1 is better than SMDD models M2 and M3, and performs similarly to OCSMM. However SMMD model M1 is more flexible than OCSMM. Also SMMD model M3 performs better than M2 showing a positive effect of the spherical normalization of the data. Future work includes applications in novelty detection, clustering and classification, for datasets of probability measures.

7. ACKNOWLEDGMENTS

This work was started when the first author was in LITIS, INSA Rouen, France. The authors would like to thank to FAPESP grant # 2011/50761-2, CNPq, CAPES, NAP eScience - PRP - USP.

8. REFERENCES

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*, volume 3. Kluwer Academic Boston, 2004.

[2] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. Svm and kernel methods matlab toolbox. Perception Systemes et Information, INSA de Rouen, Rouen, France, 2005.

[3] P. Chan and M. Mahoney. Modeling multiple time series for anomaly detection. In *Data Mining, Fifth IEEE International Conference on*, pages 8 pp.–, Nov 2005.

[4] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

[5] K. Das, J. Schneider, and D. B. Neill. Anomaly pattern detection in categorical datasets. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 169–176, New York, NY, USA, 2008. ACM.

[6] L. Debnath and P. Mikusiński. *Hilbert Spaces with Applications*. Elsevier Academic Press, 2005.

[7] C. A. Floudas and P. M. Pardalos. *Encyclopedia of optimization*, volume 1. Springer Science & Business Media, 2008.

[8] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496. MIT Press, Cambridge, MA, 2008.

[9] M. G. Genton. Classes of kernels for machine learning: A statistics perspective. *J. Mach. Learn. Res.*, 2:299–312, Mar. 2002.

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

[11] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, Mar. 2014.

[12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[13] C. Guilbart. Produits scalaires sur l'espace des mesures. In *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, volume 15, pages 333–354. Gauthier-Villars, 1979.

[14] E. Keogh, J. Lin, and A. Fu. Hot sax: efficiently finding the most unusual time series subsequence. In *Data Mining, Fifth IEEE International Conference on*, pages 8 pp.–, Nov 2005.

[15] R. Kondor and T. Jebara. A kernel between sets of vectors. In *ICML*, pages 361–368, 2003.

[16] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[18] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. SchÃűlkopf. Learning from distributions via support measure machines. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 10–18. 2012.

[19] 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)*, pages 449–458, Corvallis, Oregon, 2013. AUAI Press.

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

[21] W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1 – 24, 1997.

[22] B. Scholkopf. The kernel trick for distances. *Advances in neural information processing systems*, pages 301–307, 2001.

[23] 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*, 13(7):1443–1471, 2001.

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

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

[26] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[27] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. SchÃűlkopf. Injective hilbert space embeddings of probability measures. In *In COLT*, 2008.

[28] 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*, 99:1517–1561, 2010.

[29] 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*, 2(2):161–182, 1995.

[30] D. M. Tax and R. P. Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[31] 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*, 19(1):53–61, 2003.

[32] L. Xiong, B. Póczos, and J. G. Schneider. Group anomaly detection using flexible genre models. In *NIPS*, pages 1071–1079, 2011.

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

[34] J. Yang and S. Gunn. Exploiting uncertain data in support vector classification. In B. Apolloni, R. Howlett, and L. Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 4694 of *Lecture Notes in Computer Science*, pages 148–155. Springer Berlin Heidelberg, 2007.

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