frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 08 November 2017
doi: 10.3389/fams.2017.00023

# Regularized Kernel Algorithms for Support Estimation

*Alessandro Rudi*[1,2], *Ernesto De Vito*[3]*, *Alessandro Verri*[4] *and Francesca Odone*[4]

[1] INRIA—Sierra team—École Normale Supérieure, Paris, France, [2] Laboratory for Computational and Statistical Learning, Istituto Italiano di Tecnologia, Genova, Italy, [3] Dipartimento di Matematica, Università di Genova, Genova, Italy, [4] DIBRIS, Università di Genova, Genova, Italy

In the framework of non-parametric support estimation, we study the statistical properties of a set estimator defined by means of Kernel Principal Component Analysis. Under a suitable assumption on the kernel, we prove that the algorithm is strongly consistent with respect to the Hausdorff distance. We also extend the above analysis to a larger class of set estimators defined in terms of a low-pass filter function. We finally provide numerical simulations on synthetic data to highlight the role of the hyper parameters, which affect the algorithm.

**Keywords: support estimation, Kernel PCA, novelty detection, dimensionality reduction, regularized Kernel methods**

OPEN ACCESS

***Edited by:***
*Ding-Xuan Zhou,*
*City University of Hong Kong,*
*Hong Kong*

***Reviewed by:***
*Ting Hu,*
*Wuhan University, China*
*Sergiy Pereverzyev,*
*University of Innsbruck, Austria*
*Qiang Wu,*
*Middle Tennessee State University,*
*United States*

***Correspondence:**
*Ernesto De Vito*
*devito@dima.unige.it*

***Specialty section:***
*This article was submitted to Mathematics of Computation and Data Science, a section of the journal Frontiers in Applied Mathematics and Statistics*

***Received:*** *11 August 2017*
***Accepted:*** *25 October 2017*
***Published:*** *08 November 2017*

***Citation:***
*Rudi A, De Vito E, Verri A and Odone F (2017) Regularized Kernel Algorithms for Support Estimation. Front. Appl. Math. Stat. 3:23. doi: 10.3389/fams.2017.00023*

## 1. INTRODUCTION

A classical issue in statistics is support estimation, i.e., the problem of learning the support of a probability distribution from a set of points identically sampled according to the distribution. For example, the Devroye-Wise algorithm [1] estimates the support with the union of suitable balls centered in the training points. In the last two decades, many algorithms have been proposed and their statistical properties analyzed [1–14] and references therein.

An instance of the above setting, which plays an important role in applications, is the problem of novelty/anomaly detection, see Campos et al. [15] for an updated review. In this context, in Hoffmann [16] the author proposed an estimator based on Kernel Principal Component Analysis (KPCA), first introduced in Schölkopf et al. [17] in the context of dimensionality reduction. The algorithm was successfully tested in many applications from computer vision to biochemistry [18–24]. In many of these examples the data are often represented by high dimensional vectors, but they actually live close to a nonlinear low dimensional submanifold of the original space, and the proposed estimator takes advantage of the fact that KPCA provides an efficient compression/dimensionality reduction of the original data [16, 17], whereas many classical set estimators refer to the dimension of the original space, as it happens for the Devroye-Wise algorithm.

In this paper we prove that KPCA is a consistent estimator of the support of the distribution with respect to the Hausdorff distance. The result is based on an intriguing property of the reproducing kernel, called *separating condition*, first introduced in De Vito et al. [25]. This assumption ensures that any closed subset of the original space is represented in the feature space by a linear subspace. We show that this property remains true if the data are recentered to have zero mean in the feature space. Together with the results in De Vito et al. [25], we conclude that the consistency of KPCA algorithm is preserved by recentering of the data, which can be regarded as a degree of freedom to improve the empirical performance of the algorithm in a specific application.

Our main contribution is sketched in the next subsection together with some basic properties of KPCA and some relevant previous works. In section 2, we describe the mathematical framework and the related notations. Section 3 introduces the spectral support estimator and informally discusses its main features, whereas its statistical properties and the meaning of the *separating condition* for the kernel are analyzed in section 4. Finally section 5 presents the effective algorithm to compute the decision function and discusses the role of the two meta-parameters based on the previous theoretical analysis. In the Appendix (Supplementary Material), we collect some technical results.

## 1.1. Sketch of the Main Result and Previous Works

In this section we sketch our main result by first recalling the construction of the KPCA estimator introduced in Hoffmann [16]. We have at disposal a training set $\{x_1, \ldots, x_n\} \in \mathcal{D} \subset \mathbb{R}^d$ of $n$ points independently sampled according to some probability distribution $P$. The input space $\mathcal{D}$ is a known compact subset of $\mathbb{R}^d$, but the probability distribution $P$ is unknown and the goal is to estimate the support $C$ of $P$ from the empirical data. We recall that $C$ is the smallest closed subset of $\mathcal{D}$ such that $\mathbb{P}[C] = 1$ and we stress that $C$ is in general a proper subset of $\mathcal{D}$, possibly of low dimension.

Classical Principal Component Analysis (PCA) is based on the construction of the vector space $V$ spanned by the first $m$ eigenvectors associated with the largest eigenvalues of the empirical covariance matrix

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \otimes (x_i - \overline{x}),$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the empirical mean. However, if the data do not live on an affine subspace, the set $V$ is not a consistent estimator of the support. In order to take into account non-linear models, following the idea introduced in Schölkopf et al. [17] we consider a feature map $\Phi$ from the input space $\mathcal{D}$ into the corresponding feature space $\mathcal{H}$, which is assumed to be a Hilbert space, and we replace the empirical covariance matrix with the empirical covariance operator

$$\widehat{T}_n^c = \frac{1}{n}\sum_{i=1}^{n}(\Phi(x_i) - \widehat{\mu}_n) \otimes (\Phi(x_i) - \widehat{\mu}_n),$$

where $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)$ is the empirical mean in the feature space. As it happens in PCA, we consider the subspace $\widehat{V}_{m,n}$ of $\mathcal{H}$ spanned by the first $m$- eigenvectors $\hat{f}_1, \ldots, \hat{f}_m$ of $\widehat{T}_n^c$. According to the proposal in Hoffmann [16], we consider the following estimator of the support of the probability distribution $P$

$$\widehat{C}_n = \left\{ x \in \mathcal{D} \mid \left\| \widehat{\mu}_n + \sum_{j=1}^{m}\langle \Phi(x) - \widehat{\mu}_n, \hat{f}_j\rangle\, \hat{f}_j - \Phi(x) \right\| \leq \tau_n \right\},$$

where $\tau_n$ is a suitable threshold depending on the number of examples and

$$\widehat{\mu}_n + \sum_{j=1}^{m}\langle \Phi(x) - \widehat{\mu}_n, \hat{f}_j\rangle\, \hat{f}_j$$

is the projection of an arbitrary point $x \in \mathcal{D}$ onto the affine subspace $\widehat{\mu}_n + \widehat{V}_{m,n}$. We show that, under a suitable assumption on the feature map, called *separating property*, $\widehat{C}_n$ is a consistent estimator of $C$ with respect to the Hausdorff distance between compact sets, see Theorem 3 of section 4.2.

The separating property was introduced in De Vito et al. [25] and it ensures that the feature space is rich enough to learn any closed subset of $\mathcal{D}$. This assumption plays the same role of the notion of universal kernel [26] in supervised learning.

Moreover, following [25, 27] we extend the KPCA estimator to a class of learning algorithms defined in terms of a low-pass filter function $r_m(\sigma)$ acting on the spectrum of the covariance matrix and depending on a regularization parameter $m \in \mathbb{N}$. The projection of $\widehat{\Phi}_n(x)$ onto $\widehat{V}_{m,n}$ is replaced by the vector

$$\widehat{\Phi}_{n,m}(x) = \sum_{j=1}^{+\infty} r_m(\hat{\sigma}_j)\langle \Phi(x) - \widehat{\mu}_n, \hat{f}_j\rangle\, \hat{f}_j,$$

where $\{\hat{f}_j\}_j$ is the family of eigenvectors of $\widehat{T}_n^c$ and $\{\hat{\sigma}_j\}_j$ is the corresponding family of eigenvalues. The support is then estimated by the set

$$\left\{ x \in \mathcal{D} \mid \left\| \widehat{\mu}_n + \widehat{\Phi}_{n,m}(x) - \Phi(x) \right\| \leq \tau_n \right\}.$$

Note that KPCA corresponds to the choice of the hard-cut off filter

$$r_m(\hat{\sigma}_j) = \begin{cases} 1 & i \leq m \\ 0 & i > m \end{cases}.$$

However, other filter functions can be considered, inspired by the theory of regularization for inverse problems [28] and by supervised learning algorithms [29, 30]. In this paper we show that the explicit computation of these spectral estimators reduces to a finite dimensional problem depending only on the kernel $K(x, w) = \langle \Phi(x), \Phi(w)\rangle$ associated with the feature map, as for KPCA. The computational properties of each learning algorithm depend on the choice of the low-pass filter $r_m(\sigma)$, which can be tuned to out-perform of some specific data set, see the discussion in Rudi et al. [31].

We conclude this section with two considerations. First, in De Vito et al. [25, 27] it is proven a consistency result for a similar estimator, where the subspace $\widehat{V}_{n,m}$ is computed with respect to the non-centered covariance matrix in the feature space $\mathcal{H}$, instead of the covariance matrix. In this paper we analyze the impact of recentering the data in the feature space $\mathcal{H}$ on the support estimation problem, see Theorem 1 below. This point of view is further analyzed in Rudi et al. [32, 33].

Finally note that, our consistency results are based on convergence rates of empirical subspaces to true subspaces of the covariance operator, see Theorem 2 below. The main difference

between our result and the result in Blanchard et al. [34], is that we prove the consistency for the case when the dimension $m = m_n$ of the subspace $\widehat{V}_{m,n}$ goes to infinity slowly enough. On the contrary, in their seminal paper [34] the authors analyze the most specific case when the dimension of the projection space is fixed.

## 2. MATHEMATICAL ASSUMPTIONS

In this section we introduce the statistical model generating the data, the notion of separating feature map and the properties of the filter function. Furthermore, we show that KPCA can be seen as a filter function and we recall the main properties of the covariance operators.

We assume that the input space $\mathcal{D}$ is a bounded closed subset of $\mathbb{R}^d$. However, our results also hold true by replacing $\mathcal{D}$ with any compact metric space. We denote by $\mathrm{d}(x, w)$ the Euclidean distance $|x - w|$ between two points $x, w \in \mathbb{R}^d$ and by $\mathrm{d}_H(A, B)$ the Hausdorff distance between two compact subsets $A, B \subset \mathcal{D}$, explicitly given by

$$\mathrm{d}_H(A, B) = \max\left\{\sup_{x\in A} \mathrm{d}(x, B), \sup_{x\in B} \mathrm{d}(x, A)\right\},$$

where $\mathrm{d}(x, A) = \inf_{w\in A} \mathrm{d}(x, w)$.

### 2.1. Statistical Model

The statistical model is described by a random vector $X$ taking value in $\mathcal{D}$. We denote by $P$ the probability distribution of $X$, defined on the Borel $\sigma$-algebra of $\mathcal{D}$, and by $C$ the support of $P$.

Since the probability distribution $P$ is unknown, so is its support. We aim to estimate $C$ from a training set of empirical data, which are described by a family $X_1, \ldots, X_n$ of random vectors, which are independent and identically distributed as $X$. More precisely, we are looking for a closed subset $\widehat{C}_n = \widehat{C}_{X_1,\ldots,X_n} \subset \mathcal{D}$, depending only on $X_1, \ldots, X_n$, but independent of $P$, such that

$$\mathbb{P}\left[\lim_{n\to+\infty} \mathrm{d}_H(\widehat{C}_n, C) = 0\right] = 1$$

for all probability distributions $P$. In the context of regression estimate, the above convergence is usually called universal strong consistency [35].

### 2.2. Mercer Feature Maps and Separating Condition

To define the estimator $\widehat{C}_n$ we first map the data into a suitable feature space, so that the support $C$ is represented by a linear subspace.

Assumption 1. *Given a Hilbert space $\mathcal{H}$, take $\Phi : \mathcal{D} \to \mathcal{H}$ satisfying the following properties:*

*(H1) the set $\Phi(\mathcal{D})$ is total in $\mathcal{H}$,* i.e.,

$$\overline{\mathrm{span}}\{\Phi(x) \mid x \in \mathcal{D}\} = \mathcal{H},$$

*where $\overline{\mathrm{span}}\{\cdot\}$ denotes the closure of the linear span;*

*(H2) the map $\Phi$ is continuous.*

*The space $\mathcal{H}$ is called the feature space and the map $\Phi$ is called* a Mercer feature map.

In the following the norm and scalar product of $\mathcal{H}$ are denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively.

Assumptions (H1) and (H2) are standard for kernel methods, see Steinwart and Christmann [36]. We now briefly recall some basic consequences. First of all, the map $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$

$$K(x, w) = \langle\Phi(x), \Phi(w)\rangle$$

is a Mercer kernel and we denote by $\mathcal{H}_K$ the corresponding (separable) reproducing kernel Hilbert space, whose elements are continuous functions on $\mathcal{D}$. Moreover, each element $f \in \mathcal{H}$ defines a function $f_\Phi \in \mathcal{H}_K$ by setting $f_\Phi(x) = \langle f, \Phi(x)\rangle$ for all $x \in \mathcal{D}$. Since $\Phi(\mathcal{D})$ is total in $\mathcal{H}$, the linear map $f \mapsto f_\Phi$ is an isometry from $\mathcal{H}$ onto $\mathcal{H}_K$. In the following, with slight abuse of notation, we write $f$ instead of $f_\Phi$, so that the elements $f \in \mathcal{H}$ are viewed as functions on $\mathcal{D}$ satisfying the reproducing property

$$f(x) = \langle f, \Phi(x)\rangle \qquad x \in \mathcal{D}.$$

Finally, since $\mathcal{D}$ is compact and $\Phi$ is continuous, it holds that

$$R = \sup_{x\in\mathcal{D}} \left\|\Phi(x)\right\|^2 = \sup_{x\in\mathcal{D}} K(x, x) < +\infty. \tag{1}$$

Following De Vito et al. [27], we call $\Phi$ a *separating Mercer feature map* if the following *the separating property* also holds true.

*(H3)* The map $\Phi$ is injective and for all closed subsets $\mathcal{C} \subset \mathcal{D}$

$$\Phi(\mathcal{C}) = \Phi(\mathcal{D}) \cap \overline{\mathrm{span}}\{\Phi(x) \mid x \in \mathcal{C}\}. \tag{2}$$

It states that any closed subset $\mathcal{C} \subset \mathcal{D}$ is mapped by $\Phi$ onto the intersection of $\Phi(\mathcal{D})$ and the closed subspace $\overline{\mathrm{span}}\{\Phi(x) \mid x \in \mathcal{C}\} \subset \mathcal{H}$. Examples of kernels satisfying the separating property are for $\mathcal{D} \subset \mathbb{R}^d$ [27]:

- Sobolev kernels with smoothness index $s > \frac{d}{2}$;
- the Abel/Laplacian kernel $K(x, w) = e^{-\gamma|x-w|}$ with $\gamma > 0$;
- the $\ell_1$-kernel $K(x, w) = e^{-\gamma|x-w|_1}$, where $|\cdot|_1$ is the $\ell_1$-norm and $\gamma > 0$.

As shown in De Vito et al. [25], given a closed set $\mathcal{C}$ the equality (2) is equivalent to the condition that for every $x_0 \notin \mathcal{C}$ there exists $f \in \mathcal{H}$ such that

$$f(x_0) \neq 0 \qquad \text{and} \qquad f(x) = 0 \quad \forall x \in \mathcal{C}. \tag{3}$$

Clearly, an arbitrary Mercer feature map is not able to separate all the closed subsets, but only few of them. To better describe these sets, we introduce the *elementary learnable* sets, namely

$$\mathcal{C}_f = \left\{x \in \mathcal{D} \mid f(x) = \langle f, \Phi(x)\rangle = 0\right\},$$

where $f \in \mathcal{H}$. Clearly, $\mathcal{C}_f$ is closed and the equality (3) holds true. Furthermore the intersection of an arbitrary family of elementary

learnable sets $\cap_{f\in\mathcal{F}} C_f$ with $\mathcal{F}\subset\mathcal{H}$ satisfies (3), too. Conversely, if $\mathcal{C}$ is a set satisfying (2), select a maximal family $\{f_j\}_{j\in J}$ of orthonormal functions in $\mathcal{H}$ such that

$$f_j(x) = \langle f_j, \Phi(x)\rangle = 0 \qquad \forall x\in\mathcal{C},\ j\in J,$$

i.e., a basis of the orthogonal complement of $\{\Phi(x) \mid x\in\mathcal{C}\}$, then it is easy to prove that

$$\mathcal{C} = \{x\in\mathcal{D} \mid \langle f_j, \Phi(x)\rangle = 0\ \forall j\in J\} = \bigcap_{j\in J} C_{f_j}, \tag{4}$$

so that any set which is separating by $\Phi$ is the (possibly denumerable) intersection of elementary sets. Assumption (H3) is hence a requirement that the family of the elementary learnable sets, labeled by the elements of $\mathcal{H}$, is rich enough to parameterize all the closed subsets of $\mathcal{D}$ by means of (4). In section 4.3 we present some examples.

The Gaussian kernel $K(x,w) = e^{-\gamma|x-w|^2}$ is a popular choice in machine learning, however it is not separating. Indeed, since $K$ is analytic, the elements of the corresponding reproducing kernel Hilbert space are analytic functions, too [36]. It is known that, given an analytic function $f\neq 0$, the corresponding elementary learnable set $C_f = \{x\in\mathcal{D} \mid f(x)=0\}$ is a closed set whose interior is the empty set. Hence also the denumerable intersections have empty interior, so that $K$ can not separate a support with not-empty interior. In **Figure 1** we compare the decay behavior of the eigenvalues of the Laplacian and the Gaussian kernels.

## 2.3. Filter Function

The second building block is a low pass filter, we introduce to avoid that the estimator overfits the empirical data. The filter functions were first introduced in the context of inverse problem, see Engl et al. [28] and references therein, and in the context of supervised learning, see Lo Gerfo et al. [29] and Blanchard and Mucke [30].

We now fix some notations. For any $f\in\mathcal{H}$, we denote by $f\otimes f$ the rank one operator $(f\otimes f)g = \langle g,f\rangle f$. We recall that a bounded operator $A$ on $\mathcal{H}$ is a Hilbert-Schmidt operator if for some (any) basis $\{f_j\}_j$ the series $\|A\|_2^2 := \sum_j \|Af_j\|^2$ is finite, $\|A\|_2$ is called the Hilbert-Schmidt norm and $\|A\|_\infty \le \|A\|_2$, where $\|\cdot\|_\infty$ is the spectral norm. We denote by $\mathcal{S}_2$ the space of Hilbert-Schmidt operators, which is a separable Hilbert space under the scalar product $\langle A,B\rangle_2 = \sum_j \langle Af_j, Bf_j\rangle$.

Assumption 2. *A filter function is a sequence of functions* $r_m : [0,R]\to[0,1]$, *with* $m\in\mathbb{N}$, *satisfying*

*(H4) for any* $m\in\mathbb{N}$, $r_m(0)=0$;
*(H5) for all* $\sigma>0$, $\lim_{m\to+\infty} r_m(\sigma) = 1$;
*(H6) for all* $m\in\mathbb{N}$, *there is* $L_m>0$ *such that*

$$|r_m(\sigma') - r_m(\sigma)| \le L_m |\sigma' - \sigma|,$$

i.e., $r_m$ *is a Lipschitz function with Lipschitz constant* $L_m$.

For fixed $m$, $r_m$ is a filter cutting the smallest eigenvalues (high frequencies). Indeed, (H4) and (H6) with $\sigma' = 0$ give

$$|r_m(\sigma)| \le L_m\ |\sigma|. \tag{5}$$

On the contrary, if $m$ goes to infinity, by (H5) $r_m$ converges point-wisely to the Heaviside function

$$\Theta(\sigma) = \begin{cases} 1 & \sigma > 0 \\ 0 & \sigma = 0 \end{cases}.$$

Since $r_m(\sigma)$ converges to $\Theta(\sigma)$, which does not satisfy (5), we have that $\lim_{m\to+\infty} L_m = +\infty$.

We fix the interval $[0,R]$ as domain of the filter functions $r_m$ since the eigenvalues of the operators we are interested belong to $[0,R]$, see (23).

Examples of filter functions are

- Tikhonov filter

$$r_m(\sigma) = \frac{m\sigma}{m\sigma + R} \qquad L_m = \frac{m}{R}.$$

- Soft cut-off

$$r_m(\sigma) = \begin{cases} 1 & \sigma \ge \frac{R}{m} \\ \frac{m\sigma}{R} & \sigma < \frac{R}{m}. \end{cases} \qquad L_m = \frac{m}{R}.$$

- Landweber iteration

$$r_m(\sigma) = \frac{\sigma}{R}\sum_{k=0}^{m}\left(1-\frac{\sigma}{R}\right)^k = 1-\left(1-\frac{\sigma}{R}\right)^{m+1} \qquad L_m = \frac{m+1}{R}.$$

We recall a technical result, which is based on functional calculus for compact operators. If $A$ is a positive Hilbert-Schmidt operator, Hilbert-Schmidt theorem (for compact self-adjoint operators) gives that there exist a basis $\{f_j\}_j$ of $\mathcal{H}$ and a family $\{\sigma_j\}_j$ of positive numbers such that

$$A = \sum_j \sigma_j f_j\otimes f_j \qquad \Longleftrightarrow \qquad Af_j = \sigma_j f_j. \tag{6}$$

If the spectral norm $\|A\|_\infty \le R$, then all the eigenvalues $\sigma_j$ belong to $[0,R]$ and the spectral calculus defines $r_m(A)$ as the operator on $\mathcal{H}$ given by

$$r_m(A) = \sum_j r_m(\sigma_j) f_j\otimes f_j \qquad \Longleftrightarrow \qquad r_m(A)f_j = r_m(\sigma_j)f_j.$$

With this definition each $f_j$ is still an eigenvector of $r_m(A)$, but the corresponding eigenvalue is shrunk to $r_m(\sigma_j)$. Proposition 1 in the Appendix in Supplementary Material summarizes the main properties of $r_m(A)$.

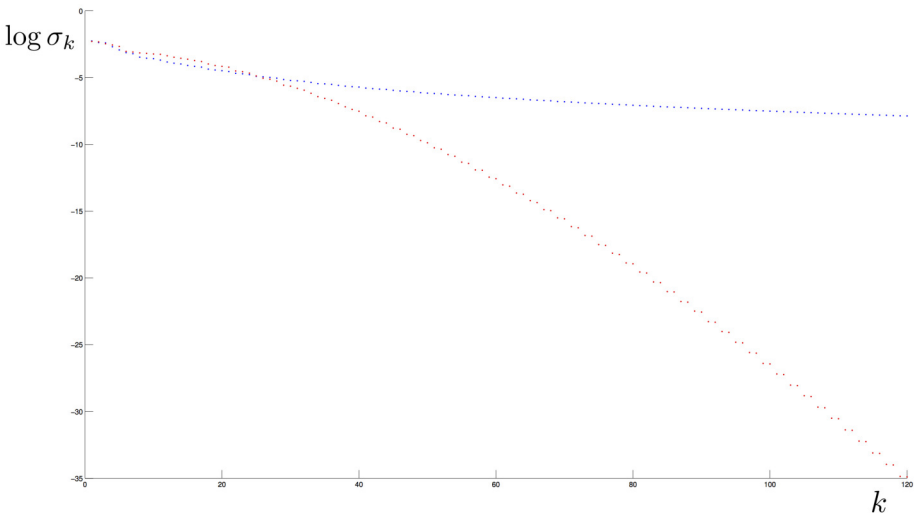


**FIGURE 1 |** Eigenvalues in logarithmic scale of the Covariance operator when the kernel is Abel (blue) and Gaussian (red) and the distribution is uniformly supported on the "8" curve in **Figure 2**. Note that the eigenvalue decay rate of the first operator has a polynomial behavior while the second has an exponential one.

## 2.4. Kernel Principal Component Analysis

As anticipated in the introduction, the estimators we propose are a generalization of KPCA suggested by Hoffmann [16] in the context of novelty detection. In our framework this corresponds to the hard cut-off filter, i.e., by labeling the different eigenvalues of $A$ in a decreasing order[1] $\sigma_1 > \sigma_2 > \ldots > \sigma_m > \sigma_{m+1} > \ldots$, the filter function is

$$r_m(\sigma) = \begin{cases} 1 & \sigma \geq \sigma_m \\ 0 & \sigma < \sigma_m \end{cases}.$$

Clearly, $r_m$ satisfies (H4) and (H5), but (H6) does not hold. However, the Lipschitz assumption is needed only to prove the bound (21e) and, for the hard cut-off filter, $r_m(A)$ is simply the orthogonal projector onto the linear space spanned by the eigenvectors whose eigenvalues are bigger than $\sigma_{m+1}$. For such projections [37] proves the following bound

$$\left\| r_m(A') - r_m(A) \right\|_2 \leq \frac{2}{\sigma_{m+1} - \sigma_m} \left\| A' - A \right\|_2,$$

so that (21e) holds true with $L_m = \frac{2}{\sigma_{m+1} - \sigma_m}$. Hence, our results also hold for hard cut-off filter at the price to have a Lipschitz constant $L_m$ depending on the eigenvalues of $A$.

## 2.5. Covariance Operators

The third building block is made of the eigenvectors of the distribution dependent covariance operator and of its empirical version. The covariance operators are computed by first mapping the data in the feature space $\mathcal{H}$.

As usual, we introduce two random variables $\Phi(X)$ and $\Phi(X) \otimes \Phi(X)$, taking value in $\mathcal{H}$ and in $\mathcal{S}_2$, respectively. Since $\Phi$ is continuous and $X$ belongs to the compact subset $\mathcal{D}$, both random variables are bounded. We set

$$\mu = \mathbb{E}\left[\Phi(X)\right] = \int_{\mathcal{D}} \Phi(x)\, dP(x), \tag{7a}$$

$$T = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right] = \int_{\mathcal{D}} \Phi(x) \otimes \Phi(x)\, dP(x), \tag{7b}$$

$$T^c = T - \mu \otimes \mu, \tag{7c}$$

where the integrals are in the Bochner sense.

We denote by $\widehat{\mu}_n$ and $\widehat{T}_n$ the empirical mean of $\Phi(X)$ and $\Phi(X) \otimes \Phi(X)$, respectively, and by $\widehat{T}^c_n$ the empirical covariance operator, respectively. Explicitly,

$$\widehat{\mu}_n = \frac{1}{n} \sum_i \Phi(X_i), \tag{8a}$$

$$\widehat{T}_n = \frac{1}{n} \sum_i \Phi(X_i) \otimes \Phi(X_i), \tag{8b}$$

$$\widehat{T}^c_n = \widehat{T}_n - \widehat{\mu}_n \otimes \widehat{\mu}_n. \tag{8c}$$

The main properties of the covariance operator and its empirical version are summarized in Proposition 2 in the Appendix in Supplementary Material.

# 3. THE ESTIMATOR

Now we are ready to construct the estimator, whose computational aspects are discussed in section 5. The set $\widehat{C}_n$ is defined by the following three steps:

a) the points $x \in \mathcal{D}$ are mapped into the corresponding centered vectors $\Phi(x) - \widehat{\mu}_n \in \mathcal{H}$, where the center is the empirical mean;

[1]Here, the labeling is different from the one in (6), where the eigenvalues are repeated according to their multiplicity.

b) the operator $r_m(\widehat{T}_n^c)$ is applied to each vector $\Phi(x) - \widehat{\mu}_n$;
c) the point $x \in \mathcal{D}$ is assigned to $\widehat{C}_n$ if the distance between $r_{m_n}(\widehat{T}_n^c)(\Phi(x) - \widehat{\mu}_n)$ and $\Phi(x) - \widehat{\mu}_n$ is smaller than a threshold $\tau$.

Explicitly we have that

$$\widehat{C}_n = \left\{x \in \mathcal{D} \mid \left\| r_{m_n}(\widehat{T}_n^c)(\Phi(x) - \widehat{\mu}_n) - (\Phi(x) - \widehat{\mu}_n) \right\| \leq \tau_n \right\}, \tag{9}$$

where $\tau = \tau_n$ and $m = m_n$ are chosen as a function of the number $n$ of training data.

With the choice of the hard cut-off filter, this reduces to the KPCA algorithm [16, 17]. Indeed, $r_m(\widehat{T}_n^c)$ is the projection $Q^m$ onto the vector space spanned by the first $m$ eigenvectors. Hence $\widehat{C}_n$ is the set of points $x$ whose image $\Phi(x) - \widehat{\mu}_n$ is close to $Q^m$. For an arbitrary filter function $r_m$, $Q^m$ is replaced by $r_m(\widehat{T}_n^c)$, which can be interpreted as a smooth version of $Q^m$. Note that, in general, $r_m(\widehat{T}_n^c)$ is not a projection.

In De Vito et al. [27] a different estimator is defined. In that paper the data are mapped in the feature space $\mathcal{H}$ without centering the points with respect to the empirical mean and the estimator is given by

$$\widetilde{C}_n = \left\{x \in \mathcal{D} \mid \left| \langle \Phi(x) - r_{m_n}(\widehat{T}_n)\Phi(x), \Phi(x) \rangle \right| \leq \tau_n^2 \right\},$$

where the filter function $r_m$ is as in the present work, but $r_{m_n}(\widehat{T}_n)$ is defined in terms of the eigenvectors of the non-centered second momentum $\widehat{T}_n$. To compare the two estimators note that

$$\begin{aligned} &\left\| r_m(\widehat{T}_n^c)(\Phi(x) - \widehat{\mu}_n) - (\Phi(x) - \widehat{\mu}_n) \right\|^2 \\ &\quad = \langle \left(I - r_m(\widehat{T}_n^c)\right)^2 (\Phi(x) - \widehat{\mu}_n), \Phi(x) - \widehat{\mu}_n \rangle \\ &\quad = \langle \left(I - r_m^*(\widehat{T}_n^c)\right) (\Phi(x) - \widehat{\mu}_n), \Phi(x) - \widehat{\mu}_n \rangle, \end{aligned}$$

where $r_m^*(\sigma) = 2r_m(\sigma) - r_m(\sigma)^2$, which is a filter function too, possibly with a Lipschitz constant $L_m^* \leq 2L_m$. Note that for the hard cut-off filter $r_m^*(\sigma) = r_m(\sigma)$.

Though $r_{m_n}(\widehat{T}_n)$ and $r_{m_n}(\widehat{T}_n^c)$ are different, both $\widehat{C}_n$ and $\widetilde{C}_n$ converge to the support of the probability distribution $P$, provided that the separating property (H3) holds true. Hence, one has the freedom to choose if the empirical data have or not zero mean in the feature space.

# 4. MAIN RESULTS

In this section, we prove that the estimator $\widehat{C}_n$ we introduce is strongly consistent. To state our results, for each $n \in \mathbb{N}$, we fix an integer $m_n \in \mathbb{N}$ and set $\widehat{F}_n : \mathcal{D} \to \mathcal{H}$ to be

$$\widehat{F}_n(x) = (I - r_{m_n}(\widehat{T}_n^c))(\Phi(x) - \widehat{\mu}_n),$$

so that Equation (9) becomes

$$\widehat{C}_n = \left\{x \in \mathcal{D} \mid \left\| \widehat{F}_n(x) \right\| \leq \tau_n \right\}. \tag{10}$$

## 4.1. Spectral Characterization

First of all, we characterize the support of $P$ by means of $Q^c$, the orthogonal projector onto the null space of the distribution dependence covariance operator $T^c$. The following theorem will show that the centered feature map

$$\Phi^c : \mathcal{D} \to \mathcal{H} \qquad \Phi^c(x) = \Phi(x) - \mu$$

sends the support $C$ onto the intersection of $\Phi^c(\mathcal{D})$ and the closed subspace $(I - Q^c)\mathcal{H}$, *i.e,*

$$\Phi^c(C) = \Phi^c(\mathcal{D}) \cap (I - Q^c)\mathcal{H}.$$

**Theorem 1.** *Assume that $\Phi$ is a separating Mercer feature map, then*

$$C = \left\{x \in \mathcal{D} \mid Q^c(\Phi(x) - \mu) = 0 \right\}, \tag{11}$$

*where $Q^c$ is the orthogonal projector onto the null space of the covariance operator $T^c$.*

*Proof:* To prove the result we need some technical lemmas, we state and prove in the Appendix in Supplementary Material. Assume first that $x \in \mathcal{D}$ is such that $Q^c(\Phi(x) - \mu) = 0$. Denoted by $Q$ the orthogonal projection onto the null space of $T$, by Lemma 2 $QQ^c = Q$ and $Q\mu = 0$, so that

$$Q\Phi(x) = Q(\Phi(x) - \mu) = QQ^c(\Phi(x) - \mu) = 0.$$

Hence Lemma 1 implies that $x \in C$.

Conversely, if $x \in C$, then as above $Q(\Phi(x) - \mu) = 0$. By Lemma 2 we have that $Q^c(1 - Q) = \|Q^c\mu\|^{-2}\, Q^c\mu \otimes Q^c\mu$. Hence it is enough to prove that

$$\langle Q^c\mu, \Phi(x) - \mu \rangle = 0 \iff Q^c\Phi(x) = Q^c\mu,$$

which holds true by Lemma 3.

## 4.2. Consistency

Our first result is about the convergence of $\widehat{F}_n$.

**Theorem 2.** *Assume that $\Phi$ is a Mercer feature map. Take the sequence $\{m_n\}_n$ such that*

$$\lim_{n \to \infty} m_n = +\infty, \tag{12a}$$

$$L_{m_n} \leq \kappa \frac{\sqrt{n}}{\ln n}, \tag{12b}$$

*for some constant $\kappa > 0$, then*

$$\mathbb{P}\left[ \lim_{n \to \infty} \sup_{x \in \mathcal{D}} \left\| \widehat{F}_n(x) - Q^c(\Phi(x) - \mu) \right\| = 0 \right] = 1. \tag{13}$$

*Proof:* We first prove Equation (13). Set $A_n = I - r_{m_n}(\widehat{T}_n^c)$. Given $x \in \mathcal{D}$,

$$\begin{aligned}
&\left\|A_n(\Phi(x) - \widehat{\mu}_n) - Q^c(\Phi(x) - \mu)\right\| \\
&\quad \le \left\|(A_n - Q^c)(\Phi(x) - \mu)\right\| + \left\|A_n(\mu - \widehat{\mu}_n)\right\| \\
&\quad \le \left\|r_{m_n}(\widehat{T}_n^c) - r_{m_n}(T^c)\right\|_2 \left\|\Phi(x) - \mu\right\| \\
&\qquad + \left\|(r_{m_n}(T^c) - (I - Q^c))(\Phi(x) - \mu)\right\| + \left\|A_n(\mu - \widehat{\mu}_n)\right\| \\
&\quad \le 2\sqrt{R} L_{m_n} \left\|\widehat{T}_n^c - T^c\right\|_2 \\
&\qquad + \left\|(r_{m_n}(T^c) - (I - Q^c))(\Phi(x) - \mu)\right\| + \quad \|\mu - \widehat{\mu}_n\|,
\end{aligned}$$

where the fourth line is due to (21e), the bound $\|A_n\|_\infty = \sup_{\sigma\in[0,1]} \left|1 - r_m(\sigma)\right| \le 1$, and the fact that both $\Phi(x)$ and $\mu$ are bounded by $\sqrt{R}$. By (12b) it follows that

$$\begin{aligned}
\sup_{x\in\mathcal{D}} \left\|A_n(\Phi(x) - \widehat{\mu}_n) - Q^c(\Phi - \mu)\right\| &\le 2\sqrt{R}\kappa \frac{\sqrt{n}}{\ln n} \left\|\widehat{T}_n^c - T^c\right\|_2 \\
&+ \sup_{x\in\mathcal{D}} \left\|(r_{m_n}(T^c) - (I - Q^c))(\Phi(x) - \mu)\right\| - \|\mu - \widehat{\mu}_n\|,
\end{aligned}$$

so that, taking into account (24a) and (24c), it holds that

$$\lim_{n\to+\infty} \sup_{x\in\mathcal{D}} \left\|A_n(\Phi(x) - \widehat{\mu}_n) - Q^c(\Phi - \mu)\right\| = 0$$

almost surely, provided that $\lim_{n\to+\infty} \|(r_{m_n}(T^c) - (I - Q^c))(\Phi(x) - \mu)\| = 0$. This last limit is a consequence of (21d) observing that $\left\{\Phi(x) - \mu \mid x \in \mathcal{D}\right\}$ is compact since $\mathcal{D}$ is compact and $\Phi$ is continuous.

We add some comments. Theorem 1 suggests that the consistency depends on the fact that the vector $(I - r_{m_n}(\widehat{T}_n^c))(\Phi(x) - \widehat{\mu}_n)$ is a good approximation of $Q^c(\Phi(x) - \mu)$. By the law of large numbers, $\widehat{T}_n$ and $\widehat{\mu}_n$ converge to $T$ and $\mu$, respectively, and Equation (21d) implies that, if $m$ is large enough, $(I - r_m(T))(\Phi(x) - \mu)$ is closed to $Q^c(\Phi(x) - \mu)$. Hence, if $m_n$ is large enough, see condition (12a), we expect that $r_{m_n}(\widehat{T}_n^c)$ is close to $r_{m_n}(T^c)$. However, this is true only if $m_n$ goes to infinity slowly enough, see condition (12b). The rate depends on the behavior of the Lipschitz constant $L_m$, which goes to infinity if $m$ goes to infinity. For example, for Tikhonov filter a sufficient condition is that $m_n \sim n^{\frac{1}{2}-\epsilon}$ with $\epsilon > 0$. With the right choice of $m_n$, the empirical decision function $\widehat{F}_n$ converges uniformly to the function $F(x) = Q^c(\Phi(x) - \mu)$, see Equation (13).

If the map $\Phi$ is separating, Theorem 1 gives that the zero level set of $F$ is precisely the support $C$. However, if $C$ is not learnable by $\Phi$, i.e., the equality (2) does not hold, then the zero level set of $F$ is bigger than $C$. For example, if $\mathcal{D}$ is connected, $C$ has not-empty interior and $\Phi$ is the feature map associated with the Gaussian kernel, it is possible to prove that $F$ is an analytic function, which is zero on an open set, hence it is zero on the whole space $\mathcal{D}$. We note that, in real applications the difference between Gaussian and Abel kernel, which is separating, is not so big and in our experience the Gaussian kernel provides a reasonable estimator.

From now on we assume that $\Phi$ is separating, so that Theorem 1 holds true. However, the uniform convergence of $\widehat{F}_n$ to $F$ does not imply that the zero level sets of $\widehat{F}_n$ converges to $C = F^{-1}(0)$ with respect to the Hausdorff distance. For example, with the Tikhonov filter $\widehat{F}_n^{-1}(0)$ is always the empty set. To overcome the problem, $\widehat{C}_n$ is defined as the $\tau_n$-neighborhood of the zero level set of $\widehat{F}_n$, where the threshold $\tau_m$ goes to zero slowly enough.

Define the data dependent parameter $\widehat{\tau}_n$ as

$$\widehat{\tau}_n = \max_{1\le i\le n} \left\|\widehat{F}_n(X_i)\right\|. \tag{14}$$

Since $\widehat{F}_n \in [0, 1]$, clearly $\widehat{\tau}_n \in [0, 1]$ and the set estimator becomes

$$\widehat{C}_n = \left\{x \in \mathcal{D} \mid \left\|\widehat{F}_n(x)\right\| \le \widehat{\tau}_n\right\}.$$

The following result shows that $\widehat{C}_n$ is a universal strongly consistent estimator of the support of the probability distribution $P$. Note that for KPCA the consistency is not universal since the choice of $m_n$ depends on some a-priori information about the decay of the eigenvalues of the covariance operator $T^c$, which depends on $P$.

**Theorem 3.** *Assume that $\Phi$ is a separating Mercer feature map. Take the sequence $\{m_n\}_n$ satisfying (12a)-(12b) and define $\widehat{\tau}_n$ by (14). Then*

$$\mathbb{P}\left[\lim_{n\to\infty} \widehat{\tau}_n = 0\right] = 1, \tag{15a}$$

$$\mathbb{P}\left[\lim_{n\to\infty} \mathrm{d}_H(\widehat{C}_n, C) = 0\right] = 1. \tag{15b}$$

*Proof:* We first show Equation (15a). Set $F(x) = Q^c(\Phi(x) - \mu)$ and let $E$ be the event on which $\widehat{F}_n$ converges uniformly to $F(x)$, and $F$ be the event such that $X_i \in C$ for all $i \ge 1$. Theorem 2 shows that $\mathbb{P}[E] = 1$ and, since $C$ is the support, then $\mathbb{P}[F] = 1$. Take $\omega \in E \cap F$ and fix $\epsilon > 0$, then there exists $n_0 > 0$ (possibly depending on $\omega$ and $\epsilon$) such that for all $n \ge n_0$ $|\widehat{F}_n(x) - F(x)| \le \epsilon$ for all $x \in \mathcal{D}$. By Theorem 1 $F(x) = 0$ for all $x \in C$ and $X_1(\omega), \ldots, X_n(\omega) \in C$, it follows that $|\widehat{F}_n(X_i(\omega))| \le \epsilon$ for all $1 \le i \le n$ so that $0 \le \widehat{\tau}_n(\omega) \le \epsilon$, so that the sequence $\widehat{\tau}_n(\omega)$ goes to zero. Since $\mathbb{P}[E \cap F] = 1$ Equation (15a) holds true.

We split the proof of Equation (15b) in two steps. We first show that with probability one $\lim_{n\to+\infty} \sup_{x\in\widehat{C}_n} d(x, C) = 0$. On the event $E \cap F$, suppose, by contraction, that the sequence $\left\{\sup_{x\in\widehat{C}_n} \mathrm{d}(x, C)\right\}_n$ does not converge to zero. Possibly passing to a subsequence, for all $n \in \mathbb{N}$ there exists $x_n \in \widehat{C}_n$ such that $d(x_n, \widehat{C}_n) \ge \epsilon_0$ for some fixed $\epsilon_0 > 0$. Since $\mathcal{D}$ is compact, possibly passing to a subsequence, $\{x_n\}_n$ converges to $x_0 \in \mathcal{D}$ with $\mathrm{d}(x_0, C) \ge \epsilon_0$. We claim that $x_0 \in C$. Indeed,

$$\begin{aligned}
\left\|Q^c(\Phi(x_0) - \mu)\right\| &\le \left\|Q^c(\Phi(x_0) - \Phi(x_n))\right\| \\
&\quad + \left\|\widehat{F}_n(x_n) - Q^c(\Phi(x_n) - \mu)\right\| + \left\|\widehat{F}_n(x_n)\right\| \\
&\le \left\|Q^c(\Phi(x_0) - \Phi(x_n))\right\| \\
&\quad + \sup_{x\in\mathcal{D}} \left\|\widehat{F}_n(x) - Q^c(\Phi(x) - \mu)\right\| + \tau_n,
\end{aligned}$$

since $x_n \in \widehat{C}_n$ means that $\left\|\widehat{F}_n(x_n)\right\| \le \tau_n$. If $n$ goes to infinity, since $\Phi$ is continuous and by the definition of $E$ and $F$, the right side of the above inequality goes to zero, so that

$\left\|Q^c(\Phi(x_0)-\mu)\right\|=0$, *i.e.,* by Theorem 1 we get $x_0 \in C$, which is a contraction since by construction $d(x_0, C) \geq \epsilon_0 > 0$.
We now prove that

$$\lim_{n\to\infty} \sup_{x\in C} \mathrm{d}(x, \widehat{C}_n).$$

For any $x \in \mathcal{D}$, set $X_{1,n}(x)$ to be a first neighbor of $x$ in the training set $\{X_1, \ldots, X_n\}$. It is known that for all $x \in C$,

$$\mathbb{P}\left[\lim_{n\to+\infty} \mathrm{d}(X_{1,n}(x), x) = 0\right] = 1, \tag{16}$$

see for example Lemma 6.1 of Györfi et al. [35].

Choose a denumerable family $\left\{z_j\right\}_{j\in J}$ in $C$ such that is dense in $C$. By Equation (16) there exists an event $G$ with such that $\mathbb{P}[G] = 1$ and, on $G$, for all $j \in J$

$$\lim_{n\to+\infty} \mathrm{d}(X_{1,n}(z_j), z_j) = 0.$$

Fix $\omega \in G$, we claim that $\lim_n \sup_{x\in C} \mathrm{d}(x, \widehat{C}_n) = 0$. Observe that, by definition of $\widehat{\tau}_n$, $X_i \in \widehat{C}_n$ for all $1 \leq i \leq n$ and

$$\sup_{x\in C} \mathrm{d}(x, \widehat{C}_n) \leq \sup_{x\in C} \min_{1\leq i\leq n} \mathrm{d}(x, X_i) = \sup_{x\in C} \mathrm{d}(X_{1,n}(x), x),$$

so that it is enough to show that $\lim_{n\to+\infty} \sup_{x\in C} \mathrm{d}(X_{1,n}(x), x) = 0$.
Fix $\epsilon > 0$. Since $C$ is compact, there is a finite subset $J_\epsilon \subset J$ such that $\left\{B(z_j, \epsilon)\right\}_{j\in J_\epsilon}$ is a finite covering of $C$. Furthermore,

$$\sup_{x\in C} \mathrm{d}(X_{1,n}(x), x) \leq \max_{j\in J_\epsilon} \mathrm{d}(X_{1,n}(z_j), z_j) + \epsilon. \tag{17}$$

Indeed, fix $x \in C$, there exists an index $j \in J_\epsilon$ such that $x \in B(z_j, \epsilon)$. By definition of first neighbor, clearly

$$\mathrm{d}(X_{1,n}(x), x) \leq \mathrm{d}(X_{1,n}(z_j), x),$$

so that by triangular inequality we get

$$\begin{aligned}\mathrm{d}(X_{1,n}(x), x) &\leq \mathrm{d}(X_{1,n}(z_j), x) \leq \mathrm{d}(X_{1,n}(z_j), z_j) + \mathrm{d}(z_j, x)\\ &\leq \mathrm{d}(X_{1,n}(z_j), z_j) + \epsilon\\ &\leq \max_{j\in J_\epsilon} \mathrm{d}(X_{1,n}(z_j), z_j) + \epsilon.\end{aligned}$$

Taking the supremum over $C$ we get the claim. Since $\omega \in G$ and $J_\epsilon$ is finite,

$$\lim_{n\to+\infty} \max_{j\in J_\epsilon} \mathrm{d}(X_{1,n}(z_j), z_j) = 0,$$

so that by Equation (17)

$$\operatorname*{limsup}_{n\to+\infty} \sup_{x\in C} \mathrm{d}(X_{1,n}(x), x) \leq \epsilon.$$

Since $\epsilon$ is arbitrary, we get $\lim_{n\to+\infty} \sup_{x\in C} \mathrm{d}(X_{1,n}(x), x) = 0$, which implies that

$$\lim_{n\to+\infty} \sup_{x\in C} \mathrm{d}(x, \widehat{C}_n) = 0.$$

Theorem 3 is an asymptotic result. Up to now, we are not able to provide finite sample bounds on $\mathrm{d}_H(\widehat{C}_n, C)$. It is possible to have finite sample bounds on $\left\|\widehat{F}_n(x) - Q^c(\Phi(x)-\mu)\right\|$, as in Theorem 7 of De Vito et al. [25] with the same kind of proof.

## 4.3. The Separating Condition

The following two examples clarify the notion of the separating condition.

Example 1. Let $\mathcal{D}$ be a compact subset of $\mathbb{R}^2$, $\mathcal{H} = \mathbb{R}^6$ with the euclidean scalar product, and $\Phi : \mathcal{D} \to \mathbb{R}^6$ be the feature map

$$\Phi((x, y)) = (x^2, y^2, \sqrt{2}xy, \sqrt{2}x, \sqrt{2}y, 1),$$

whose corresponding Mercer kernel is a polynomial kernel of degree two, explicitly given by

$$K(x_1, y_1; x_2, y_2) = (x_1x_2 + y_1y_2 + 1)^2. \tag{18}$$

Given a vector $f = (f_1, \ldots, f_6)^\top$, the corresponding elementary set is the conic

$$\begin{aligned}C_{f,c} = \Big\{(x, y) \in \mathcal{D} \mid f_1x^2 + f_2y^2 + f_3\sqrt{2}xy\\ + f_4\sqrt{2}x + f_5\sqrt{2}y + f_6 = 0\Big\},\end{aligned}$$

Conversely, all the conics are elementary sets. The family of all the intersections of at most five conics, i.e., the sets whose cartesian equation is a system of the form

$$\begin{cases} f_{11}x^2 + f_{12}y^2 + f_{13}\sqrt{2}xy + f_{14}\sqrt{2}x + f_{15}\sqrt{2}y + f_{16} = 0\\ \vdots\\ f_{51}x^2 + f_{52}y^2 + f_{53}\sqrt{2}xy + f_{54}\sqrt{2}x + f_{55}\sqrt{2}y + f_{56} = 0\end{cases},$$

where $f_{11}, \ldots, f_{56} \in \mathbb{R}$.

Example 2. The data are the random vectors in $\mathbb{R}^2$

$$X_i = (\sin(a\Theta_i + b), \sin(c\Theta_i + d)),$$

where $a, c \in \mathbb{N}$, $b, d \in [0, 2\pi]$ and $\Theta_1, \ldots, \Theta_n$ are independent random variables, each of them uniformly distributed on $[0, 2\pi]$. Setting $\mathcal{D} = [-1, 1]^2$, clearly $X_i \in \mathcal{D}$ and the support of their common probability distribution is the Lissajous curve

$$C = \mathrm{Lis}_{a,b,c,d} = \{(\sin(a\theta + b), \sin(c\theta + d)) \in \mathcal{D} \mid \theta \in [0, 2\pi]\}.$$

**Figure 2** shows two examples of Lissajous curves. As a filter function $r_m$, we fix the hard cut-off filter where $m$ is the number of eigenvectors corresponding to the highest eigenvalues we keep. We denote by $\hat{C}_n^{m,\tau}$ the corresponding estimator given by (10).

In the first two tests we use the polynomial kernel (18), so that the elementary learnable sets are conics. One can check that the rank of $T^c$ is less or equal than 5. More precisely, if $\mathrm{Lis}_{a,b,c,d}$ is a conic, the rank of $T^c$ is 4 and we need to estimate five parameters, whereas if $\mathrm{Lis}_{a,b,c,d}$ is not a conic, $\mathrm{Lis}_{a,b,c,d}$ is not a learnable set and the rank of $T^c$ is 5.

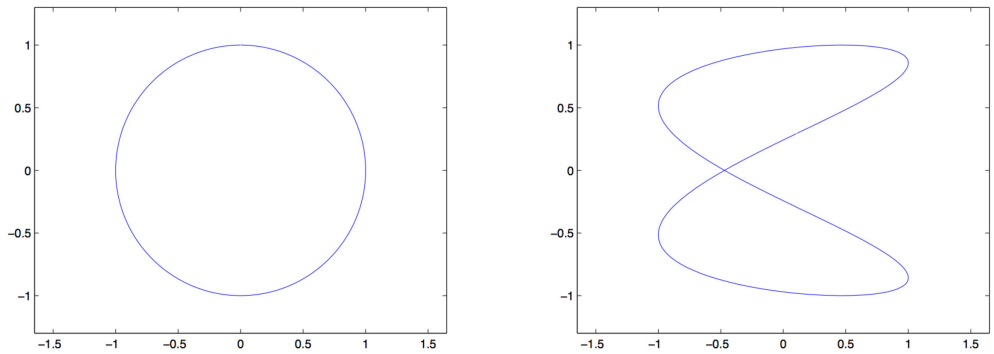

**FIGURE 2 |** Examples of Lissajous curves for different values of the parameters. **(Left)** $\text{Lis}_{1,0,1,\frac{\pi}{2}}$. **(Right)** $\text{Lis}_{2,0.11,1,0.3}$.

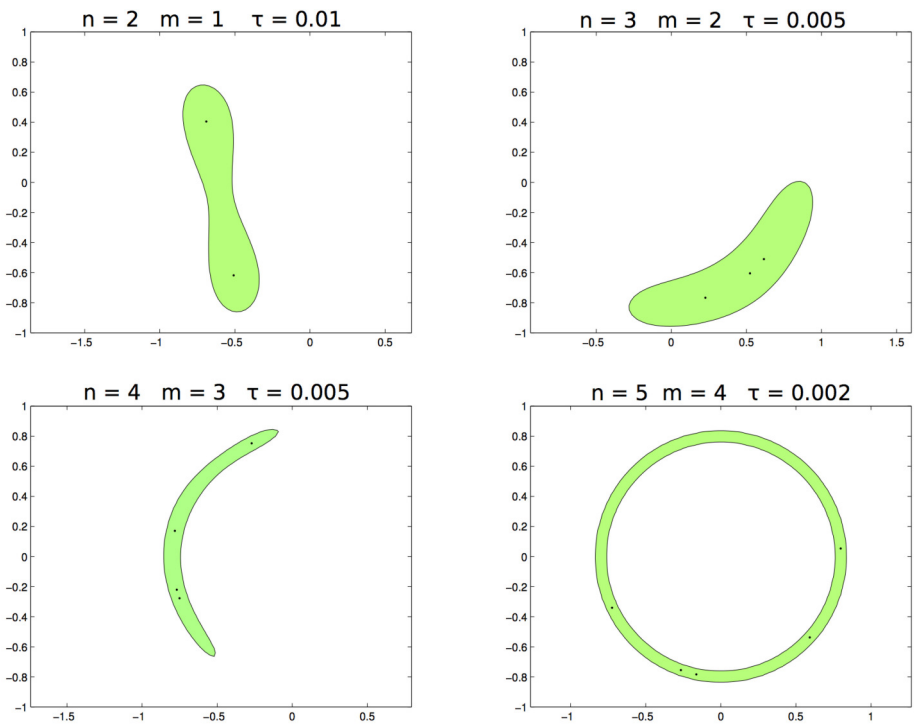


**FIGURE 3 |** From left to right, top to bottom. The set $\hat{C}_n^{m,\tau}$ with $n$, respectively $2, 3, 4, 5$, $m = 1, 2, 3, 4$ and $\tau = 0.1, 0.005, 0.005, 0.002$.

In the first test the data are sampled from the distribution supported on the circumference $\text{Lis}_{1,0,1,\frac{\pi}{2}}$ (see panel left of **Figure 2**). In **Figure 3** we draw the set $\hat{C}_n^{m,\tau}$ for different values of $m$ and $\tau$ when $n$ varies. In this toy example $n = 5$ is enough to learn exactly the support, hence for each $n = 2, \ldots, 5$ the corresponding values of $m_n$ and $\tau_n$ are $m_n = 1, 2, 3, 4$ and $\tau_n = 0.01, 0.005, 0.005, 0.002$.

In the second test the data are sampled from the distribution supported on the curve $\text{Lis}_{2,0.11,1,0.3}$, which is not a conic (see panel right of **Figure 2**). In **Figure 4** we draw the set $\hat{C}_n^{m,\tau}$ for $n = 10, 20, 50, 100$, $m = 4$, and $\tau = 0.01$. Clearly, $\widehat{C}_n$ is not able to estimate $\text{Lis}_{2,0.11,1,0.3}$.

In the third test, we use the Abel kernel with the data sampled from the distribution supported on the curve $\text{Lis}_{2,0.11,1,0.3}$ (see panel right of **Figure 2**). In **Figure 5** we show the set $\hat{C}_n^{m,\tau}$ for $n = 20, 50, 100, 500$, $m = 5, 20, 30, 50$, and $\tau = 0.4, 0.35, 0.3, 0.2$. According the fact that the kernel is separating, $\widehat{C}_n$ is able to estimate $\text{Lis}_{2,0.11,1,0.3}$ correctly.

We now briefly discuss how to select the parameter $m_n$ and $\tau_n$ from the data. The goal of set-learning problem is to recover the support of the probability distribution generating the data by using the given input observations. Since no output is present, set-learning belongs to the category of unsupervised learning problems, for which there is not a general framework

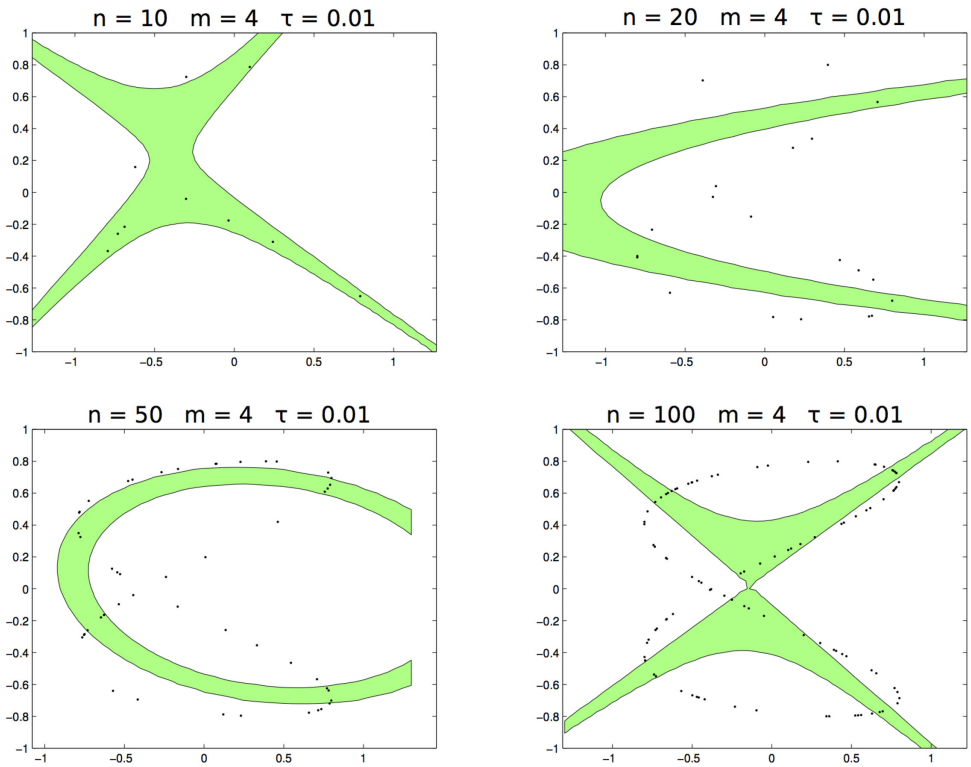


**FIGURE 4 |** The set $\hat{C}_n^{4,0.01}$ with $n$ respectively 10, 20, 50, 100 .

accounting for model selection. However there are some possible strategies (whose analysis is out of the scope of this paper). A first approach, we used in our simulations, is based on the monotonicity properties of $\hat{C}_n^{m,\tau}$ with respect to $m, \tau$. More precisely, given $f \in (0,1)$, we select (the smallest) $m$ and (the biggest) $\tau$ such that at most $nf$ observed points belong to the the estimated set. It is possible to prove that this method is consistent when $f$ tends to 1 as the number of observations increases. Another way to select the parameters consists in transforming the set-learning problem is a supervised one and then performing standard model selection techniques like cross validation. In particular set-learning can be casted in a classification problem by associating the observed example to the class $+1$ and by defining an auxiliary measure $\mu$ (e.g., uniform on a ball of interest in $\mathcal{D}$) associated to $-1$, from which $n$ i.i.d. points are drawn. It is possible to prove that this last method is consistent when $\mu(\text{supp}\rho) = 0$.

## 4.4. The Role of the Regularization

We now explain the role of the filter function. Given a training set $X_1, \ldots, X_n$ of size $n$, the separating property (3) applied to the support of the empirical distribution gives that

$$\{X_1, \ldots, X_n\} = \left\{x \in \mathcal{D} \mid \widehat{Q}_n^c(\Phi(x) - \widehat{\mu}_n) = 0\right\},$$

where $\widehat{\mu}_n$ is the empirical mean and $I - \widehat{Q}_n^c$ is the orthogonal projection onto the linear space spanned by the family $\left\{\Phi(X_1) - \widehat{\mu}_n, \ldots, \Phi(X_n) - \widehat{\mu}_n\right\}$, which are the centered images of the examples. Hence, given a new point $x \in D$ the condition $\left\|\widehat{Q}_n^c(\Phi(x) - \widehat{\mu}_n)\right\| \leq \tau$ with $\tau \ll 1$ is satisfied if only if $x$ is close to one of the examples in the training set. Hence the naive estimator $\left\{x \in \mathcal{D} \mid \left\|\widehat{Q}_n^c(\Phi(x) - \widehat{\mu}_n)\right\| \leq \tau\right\}$ overfits the data. Hence we would like to replace $\widehat{Q}_n^c$ with an operator, which should be close to the identity on the linear subspace spanned by $\left\{\Phi(X_1) - \widehat{\mu}_n, \ldots, \Phi(X_n) - \widehat{\mu}_n\right\}$ and it should have a small range. To modulate the two requests, one can consider the following optimization problem

$$\min_{A \in \mathcal{S}_2} \left( \frac{1}{n} \sum_{i=1}^{n} \left\|(I - A)(\Phi(X_i) - \widehat{\mu}_n)\right\|^2 + \lambda \|A\|_2^2 \right).$$

We note that if $A$ is a projection its Hilbert-Schmidt norm $\|A\|_2$ is the square root of the dimension of the range of $A$. Since

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \left\|(I - A)(\Phi(X_i) - \widehat{\mu}_n)\right\|^2 \\ &= \text{Tr}\left( (I - A^\top)(I - A) \frac{1}{n} \sum_{i=1}^{n} (\Phi(X_i) - \widehat{\mu}_n) \otimes (\Phi(X_i) - \widehat{\mu}_n) \right) \\ &= \text{Tr}\left( (I - A^\top)(I - A) \widehat{T}_n^c \right), \end{aligned}$$

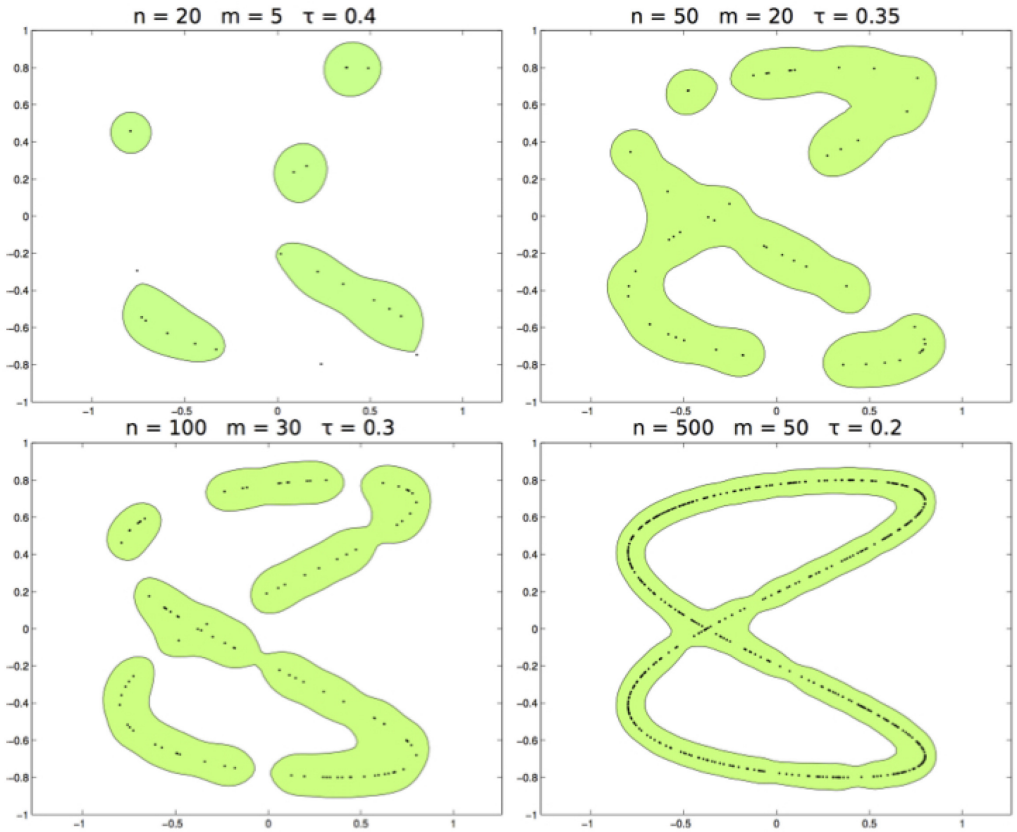


**FIGURE 5 |** From left to right and top to bottom. The learned set $\hat{C}_n^{m,\tau}$ with $n$ respectively of 20, 50, 100, 500, and $m = 5, 20, 30, 50$, $\tau = 0.4, 0.35, 0.3, 0.2$.

where $\mathrm{Tr}(A)$ is the trace, $A^\top$ is the transpose and $\|A\|_2 = \sqrt{\mathrm{Tr}(A^\top A)}$ is the Hilbert-Schmidt norm, then

$$\frac{1}{n}\sum_{i=1}^{n}\left\|(I-A)(\Phi(X_i)-\widehat{\mu}_n)\right\|^2+\lambda\,\|A\|_2^2 = \mathrm{Tr}\left((I-A^\top)(I-A)\widehat{T}_n^c\;+\;\lambda_m A^\top A\right),$$

and the optimal solution is given by

$$A_{\mathrm{opt}} = \widehat{T}_n^c(\widehat{T}_n^c+\lambda_m)^{-1},$$

i.e., $A_{\mathrm{opt}}$ is precisely the operator $r_m(\widehat{T}_n^c)$ with the Tikhonov filter $r_m(\sigma) = \frac{\sigma}{\sigma+\lambda}$ and $\lambda = \frac{R}{m}$. A different choice of the filter function $r_m$ corresponds to a different regularization of the least-square problem

$$\min\frac{1}{n}\sum_{i=1}^{n}\left\|(I-A)(\Phi(X_i)-\widehat{\mu}_n)\right\|^2.$$

## 5. THE KERNEL MACHINE

In this section we show that the computation of $\left\|\widehat{F}_n(x)\right\|$, in terms of which is defined the estimator $\widehat{C}_n$, reduces to a finite dimensional problem, depending only on the Mercer kernel $K$, associated with the feature map. We introduce the centered sampling operator

$$S_n^c:\mathcal{H}\to\mathbb{R}^n \qquad (S_n^c f)_i = \langle f, \Phi(X_i)-\widehat{\mu}_n\rangle,$$

whose transpose is given by

$$S_n^{c\,\top}:\mathbb{R}^n\to\mathcal{H} \qquad S_n^{c\,\top}\mathbf{v} = \sum_{i=1} v_i\,(\Phi(X_i)-\widehat{\mu}_n),$$

where $v_i$ is the $i$-th entry of the column vector $\mathbf{v}\in\mathbb{R}^n$. Hence, it holds that

$$\frac{1}{n}S_n^{c\,\top}S_n^c = \widehat{T}_n^c,$$
$$\frac{1}{n}S_n^c S_n^{c\,\top} = (\mathrm{I}_n-\frac{\mathbf{1}\mathbf{1}^\top}{n})\frac{K_n}{n}(\mathrm{I}_n-\frac{\mathbf{1}\mathbf{1}^\top}{n}),$$

where $K_n$ is the $n\times n$ matrix whose $(i,j)$-entry is $K(X_i,X_j)$ and $\mathrm{I}_n$ is the identity $n\times n$ matrix, so that the $(i,j)$-entry of $S_n^c S_n^{c\,\top}$ is

$$K(X_i,X_j)-\frac{1}{n}\sum_b K(X_i,X_b)-\frac{1}{n}\sum_a K(X_a,X_j)+\frac{1}{n^2}\sum_{a,b}K(X_a,X_b).$$

Denoted by $\ell$ the rank of $S_n^c S_n^{c\top}$, take the singular value decomposition of $S_n^c S_n^{c\top}/n$, i.e.,

$$\frac{S_n^c S_n^{c\top}}{n} = V\Sigma V^\top,$$

where $V$ is an $n \times \ell$ matrix whose columns $\mathbf{v}_j \in \mathbb{R}^n$ are the normalized eigenvectors, $V^\top V = \mathrm{I}_\ell$, and $\Sigma$ is a diagonal $\ell \times \ell$ matrix with the strictly positive eigenvalues on the diagonal, i.e., $\Sigma = \mathrm{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_\ell)$. Set $U = S_n^{c\top} V \Sigma^{-\frac{1}{2}}$, regarded as operator from $\mathbb{R}^\ell$ to $\mathcal{H}$, then a simple calculation shows that

$$\widehat{T}_n^c = U\Sigma U^\top \qquad r_m(\widehat{T}_n^c) = U r_m(\Sigma) U^\top,$$

where $r_m(\Sigma)$ is the diagonal $\ell \times \ell$ matrix

$$r_m(\Sigma) = \mathrm{diag}(r_m(\widehat{\sigma}_1), \ldots, r_m(\widehat{\sigma}_\ell)),$$

and the equation for $r_m(\widehat{T}_n^c)$ holds true since by assumption $r_m(0) = 0$. Hence

$$\begin{aligned}
\left\|\widehat{F}_n(x)\right\|^2 &= \langle \left(I - r_m(\widehat{T}_n^c)\right)(\Phi(x) - \widehat{\mu}_n), \left(I - r_m(\widehat{T}_n^c)\right)(\Phi(x) - \widehat{\mu}_n)\rangle \\
&= \langle \Phi(x) - \widehat{\mu}_n, \Phi(x) - \widehat{\mu}_n\rangle - \langle (2r_m(\widehat{T}_n^c) \\
&\quad - r_m(\widehat{T}_n^c)^2)(\Phi(x) - \widehat{\mu}_n)\rangle \Phi(x) - \widehat{\mu}_n \\
&= \langle \Phi(x) - \widehat{\mu}_n, \Phi(x) - \widehat{\mu}_n\rangle - \langle U(2r_m(\Sigma) \\
&\quad - r_m(\Sigma)^2) U^\top (\Phi(x) - \widehat{\mu}_n)\rangle \Phi(x) - \widehat{\mu}_n \\
&= \langle \Phi(x) - \widehat{\mu}_n, \Phi(x) - \widehat{\mu}_n\rangle - \langle V\Sigma^{-\frac{1}{2}}(2r_m(\Sigma) \\
&\quad - r_m(\Sigma)^2)\Sigma^{-\frac{1}{2}} V^\top S_n^c (\Phi(x) - \widehat{\mu}_n)\rangle S_n^c (\Phi(x) - \widehat{\mu}_n) \\
&= w(x) - \mathbf{v}(x)^\top G_m \mathbf{v}(x), \qquad (19)
\end{aligned}$$

where the real number $w(x) = \langle \Phi(x) - \widehat{\mu}_n, \Phi(x) - \widehat{\mu}_n\rangle$ is

$$w(x) = K(x,x) - \frac{2}{n}\sum_b K(x, X_b) + \frac{1}{n^2}\sum_{a,b} K(X_a, X_b),$$

the $i$-th entry of the column vector $\mathbf{v}(x) \in \mathbb{R}^n$ is

$$\begin{aligned}
\mathbf{v}(x)_i = (S_n^c(\Phi(x) - \widehat{\mu}_n))_i &= K(x, X_i) - \frac{1}{n}\sum_a K(X_a, x) \\
&\quad - \frac{1}{n}\sum_b K(X_i, X_b) + \frac{1}{n^2}\sum_{a,b} K(X_a, X_b),
\end{aligned}$$

the diagonal $\ell \times \ell$ matrix $R_m(\Sigma) = \Sigma^{-1}(2r_m(\Sigma) - r_m(\Sigma)^2)$ is

$$R_m(\Sigma) = \mathrm{diag}\left(\frac{2r_m(\widehat{\sigma}_1) - r_m(\widehat{\sigma}_1)^2}{\widehat{\sigma}_1}, \ldots, \frac{2r_m(\widehat{\sigma}_\ell) - r_m(\widehat{\sigma}_\ell)^2}{\widehat{\sigma}_\ell}\right),$$

and the $n \times n$-matrix $G_m$ is

$$G_m = V R_m(\Sigma) V^\top. \qquad (20)$$

In **Algorithm 1** we list the corresponding MatLab Code.

The above equations make clear that both $\widehat{F}_n$ and $\widehat{C}_n$ can be computed in terms of the singular value decomposition $(V, \Sigma)$ of the $n \times n$ Gram matrix $K_n$ and of the filter function $r_m$, so that $\widehat{F}_n$ belongs to the class of kernel methods and $\widehat{C}_n$ is a plug-in estimator. For the hard cut-off filter, one simply has

$$R_m(\Sigma)_{ii} = \begin{cases} \dfrac{1}{\widehat{\sigma}_i} & \widehat{\sigma}_i \geq \widehat{\sigma}_m \\ 0 & \widehat{\sigma}_i < \widehat{\sigma}_m \end{cases}.$$

For real applications, a delicate issue is the choice of the parameters $m$ and $\tau$, we refer to Rudi et al. [31] for a detailed discussion. Here, we add some simple remarks.

We first discuss the role of $\tau$. According to (10), $\hat{C}_n^{m,\tau} \subseteq \hat{C}_n^{m,\tau'}$ whenever $\tau < \tau'$. We exemplify this behavior with the dataset

**Algorithm 1** Matlab code for Set Learning.

```
function [Gm, mu, s] = learnSet(X, k,
rm, m)
% X: n x d matrix of training data
% k: kernel type
% rm: spectral filter
% m: regularization parameter
   n = size(X,1);
   K = gram(X, X, k); %computes the Gram
   matrix with the kernel k
   mu = sum(K,2)/n;
   s = sum(sum(K))/n^2;
   ScScT = (K - ones(n,1)*mu' - mu*ones(1,
   n) + s);
   [Rm, V]= rm(ScScT/n, m);
   Gm = V * Rm * V'; % see Equation  (24)
end

function y = testSet(X, Gm, mu, s, Y)
% Y: t x d -matrix of test data
% y: t-column vector, each entry is (23)
%    for the corresponding test point
   n = size(X,1);
   W = gram(Y,X,k);
   vx = W - repmat(sum(W,2)/n,1,size(Y,
   2)) ...
    - repmat(mu', size(Y,1),
   1)+ s;
   w = diag(gram(Y,Y,k)) - 2*sum(W,2)/
   n + s;
   y = w - diag(vx*Gm*vx');
end

%----------------
% main script
  ...% creation of training set X and test
  set Y and kernel width c
  [Gm, mu, s] = learnSet(X, @abel(c),
   @hardcutoff, m);
  y = testSet(X, Gm, mu, s, Y);
  y <= tau; % membership of test data
```

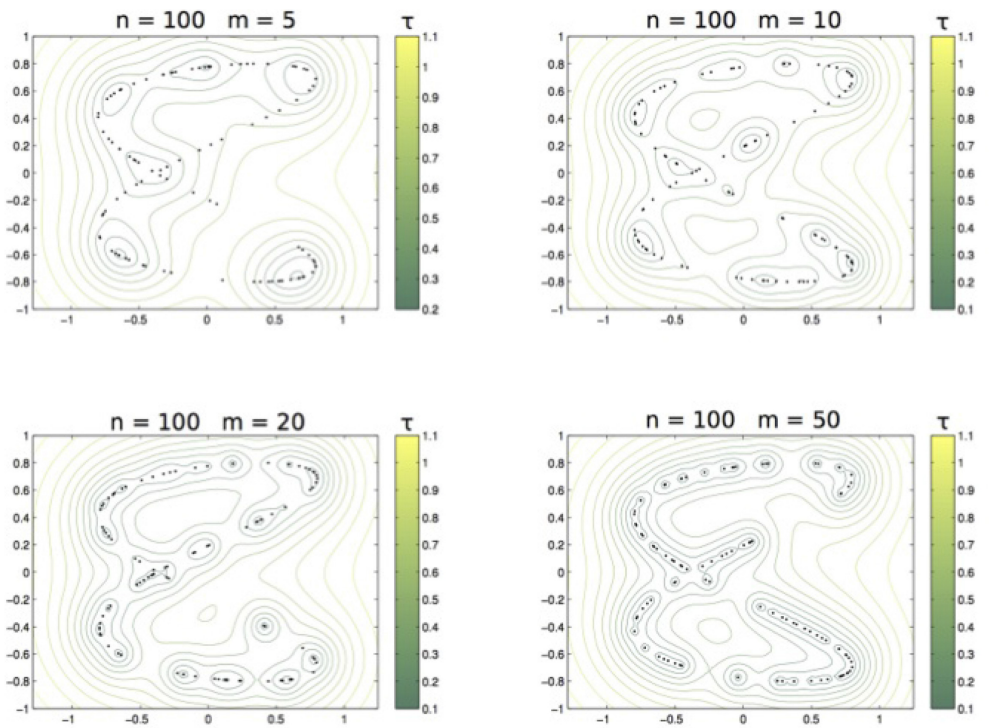


**FIGURE 6 |** From left to right and top to bottom: The family of sets $\hat{C}_{100}^{5,\tau}, \hat{C}_{100}^{10,\tau}, \hat{C}_{100}^{20,\tau}, \hat{C}_{100}^{50,\tau}$ with $\tau$ varying as in the related colorbar.

of Example 2. The training set is sampled from the distribution supported on the curve $\mathrm{Lis}_{2,0.11,1,0.3}$ (see panel right of **Figure 2**) and we compute $\widehat{C}_n$ with the Abel kernel, $n = 100$ and $m$ ranging over $5, 10, 20, 50$. **Figure 6** shows the nested sets when $\tau$ runs in the associated color-bar.

Analyzing the role of $m$, we now show that, for the the hard cut-off filter, $\hat{C}_n^{m',\tau} \subseteq \hat{C}_n^{m,\tau}$ whenever $m' \leq m$. Indeed, this filter satisfies $r_{m'}(\sigma) \leq r_m(\sigma)$ and, since $0 \leq r_m(\sigma) \leq 1$, one has $(1 - r_m(\sigma))^2 \leq (1 - r_{m'}(\sigma))^2$. Hence, denoted by $\{\widehat{u}_j\}_j$ a base of eigenvectors of $\widehat{T}_n^c$, it holds that

$$\begin{aligned}\left\|(I - r_m(\widehat{T}_n^c))(\Phi(x) - \widehat{\mu}_n)\right\|^2 &= \sum_j (1 - r_m(\widehat{\sigma}_j))^2 \langle \Phi(x) - \widehat{\mu}_n, \widehat{u}_j \rangle^2 \\ &\leq \sum_j (1 - r_{m'}(\widehat{\sigma}_j))^2 \langle \Phi(x) - \widehat{\mu}_n, \widehat{u}_j \rangle^2 \\ &= \left\|(I - r_{m'}(\widehat{T}_n^c))(\Phi(x) - \widehat{\mu}_n)\right\|^2.\end{aligned}$$

Hence, for any point in $x \in \hat{C}_n^{m',\tau}$,

$$\begin{aligned}&\left\|(I - r_m(\widehat{T}_n^c))(\Phi(x) - \widehat{\mu}_n)\right\|_{\mathcal{H}}^2 \\ &\quad \leq \left\|(I - r_{m'}(\widehat{T}_n^c))(\Phi(x) - \widehat{\mu}_n)\right\|_{\mathcal{H}}^2 \leq \tau^2,\end{aligned}$$

so that $x \in \hat{C}_n^{m,\tau}$.

As above, we illustrate the different choices of $m$ with the data sampled from the curve $\mathrm{Lis}_{2,0.11,1,0.3}$ and the Abel kernel where $n = 100$ and $\tau$ ranges over $0.25, 0.3, 0.4, 0.5$. **Figure 7** shows the nested sets when $m$ runs in the associated color-bar.

## 6. DISCUSSION

We presented a new class of set estimators, which are able to learn the support of an unknown probability distribution from a training set of random data. The set estimator is defined through a decision function, which can be seen as a novelty/anomality detection algorithm as in Schölkopf et al. [6].

The decision function we defined is a kernel machine. It is computed by the singular value decomposition of the empirical (kernel)-covariance matrix and by a low pass filter. An example of filter is the hard cut-off function and the corresponding decision function reduces to KPCA algorithm for novelty detection first introduced by Hoffmann [16]. However, we showed that it is possible to choose other low pass filters, as it was done for a class of supervised algorithms in the regression/classification setting [38].

Under some weak assumptions on the low pass filter, we proved that the corresponding set estimator is strongly consistent with respect to the Hausdorff distance, provided that the kernel satisfies a suitable separating condition, as it happens, for example, for the Abel kernel. Furthermore, by comparing Theorem 2 with a similar consistency result in De Vito et al. [27], it appears clear that the algorithm correctly learns the support both if the data have zero mean, as in our paper, and if the data are not centered, as in De Vito et al. [27]. On the contrary, if

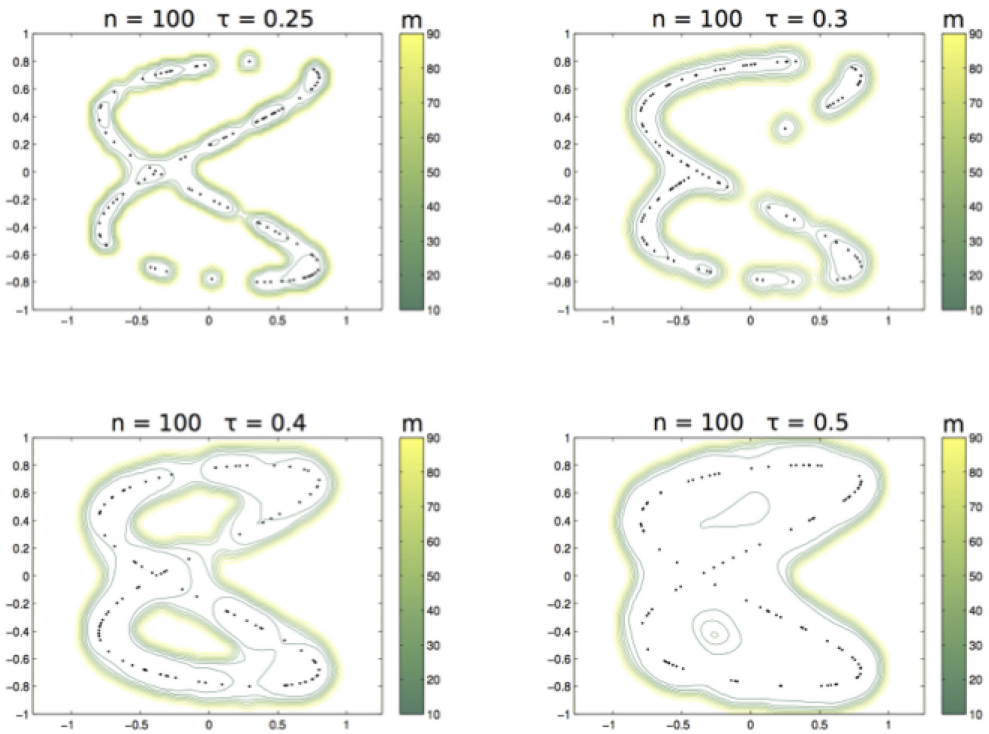


**FIGURE 7 |** From left to right and top to bottom: The family of sets $\hat{C}_{100}^{m,0.25}, \hat{C}_{100}^{m,0.3}, \hat{C}_{100}^{m,0.4}, \hat{C}_{100}^{m,0.5}$ with $m$ varying as in the related colorbar.

the separating property does not hold, the algorithm learns only the supports that are mapped into linear subspaces by the feature map defined by the kernel.

The set estimator we introduced depends on two parameters: the *effective* number $m$ of eigenvectors defining the decision function and the thickness $\tau$ of the region estimating the support. The role of these parameters and of the separating property was briefly discussed by a few tests on toy data.

We finally observe that our class of set learning algorithms is very similar to classical kernel machines in supervised learning. So, in order to reduce both the computational cost and the memory requirements, there is the possibility to successfully implement some new advanced approximation techniques, for which there exist theoretical guarantees for the statistical learning setting. For example random features [39, 40], Nyström projections [41, 42] or mixed approaches with iterative regularization and preconditioning [43, 44].

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

## ACKNOWLEDGMENTS

ED is member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2017.00023/full#supplementary-material

## REFERENCES

1. Devroye L, Wise GL. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J Appl Math.* (1980) **38**:480–8. doi: 10.1137/0138038
2. Korostelëv AP, Tsybakov AB. *Minimax Theory of Image Reconstruction*. New York, NY: Springer-Verlag (1993).
3. Dümbgen L, Walther G. Rates of convergence for random approximations of convex sets. *Adv Appl Probab.* (1996) **28**:384–93. doi: 10.2307/1428063
4. Cuevas A, Fraiman R. A plug-in approach to support estimation. *Ann Stat.* (1997) **25**:2300–12. doi: 10.1214/aos/1030741073
5. Tsybakov AB. On nonparametric estimation of density level sets. *Ann Stat.* (1997) **25**:948–69. doi: 10.1214/aos/1069362732
6. Schölkopf B, Platt J, Shawe-Taylor J, Smola A, Williamson R. Estimating the support of a high-dimensional distribution. *Neural Comput.* (2001) **13**:1443–71. doi: 10.1162/089976601750264965
7. Cuevas A, Rodríguez-Casal A. Set estimation: an overview and some recent developments. In: *Recent Advances and Trends in Nonparametric Statistics*. Elsevier: Amsterdam (2003). p. 251–64.

8. Reitzner M. Random polytopes and the Efron-Stein jackknife inequality. *Ann Probab.* (2003) **31**:2136–66. doi: 10.1214/aop/1068646381
9. Steinwart I, Hush D, Scovel C. A classification framework for anomaly detection. *J Mach Learn Res*. (2005) **6**:211–32.
10. Vert R, Vert JP. Consistency and convergence rates of one-class SVMs and related algorithms. *J Mach Learn Res*. (2006) **7**:817–54.
11. Scott CD, Nowak RD. Learning minimum volume sets. *J Mach Learn Res*. (2006) **7**:665–704.
12. Biau G, Cadre B, Mason D, Pelletier B. Asymptotic normality in density support estimation. *Electron J Probab*. (2009) **91**:2617–35.
13. Cuevas A, Fraiman R. Set estimation. In: W. Kendall and I. Molchanov, editors. *New Perspectives in Stochastic Geometry*. Oxford: Oxford University Press (2010). p. 374–97.
14. Bobrowski O, Mukherjee S, and Taylor JE. Topological consistency via kernel estimation. *Bernoulli* (2017) **23**:288–328. doi: 10.3150/15-BEJ744
15. Campos GO, Zimek A, Sander J, Campello RJ, Micenková B, Schubert E, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Min Knowl Discov*. (2016) **30**:891–927. doi: 10.1007/s10618-015-0444-8
16. Hoffmann H. Kernel PCA for novelty detection. *Pattern Recognit*. (2007) **40**:863–74. doi: 10.1016/j.patcog.2006.07.009
17. Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput*. (1998) **10**:1299–319.
18. Ristic B, La Scala B, Morelande M, Gordon N. Statistical analysis of motion patterns in AIS data: anomaly detection and motion prediction. In: *2008 11th International Conference on Information Fusion* (2008). p. 1–7.
19. Lee HJ, Cho S, Shin MS. Supporting diagnosis of attention-deficit hyperactive disorder with novelty detection. *Artif Intell Med*. (2008) **42**:199–212. doi: 10.1016/j.artmed.2007.11.001
20. Valero-Cuevas FJ, Hoffmann H, Kurse MU, Kutch JJ, Theodorou EA. Computational models for neuromuscular function. *IEEE Rev Biomed Eng*. (2009) **2**:110–35. doi: 10.1109/RBME.2009.2034981
21. He F, Yang JH, Li M, Xu JW. Research on nonlinear process monitoring and fault diagnosis based on kernel principal component analysis. *Key Eng Mater*. (2009) **413**:583–90. doi: 10.4028/www.scientific.net/KEM.413-414.583
22. Maestri ML, Cassanello MC, Horowitz GI. Kernel PCA performance in processes with multiple operation modes. *Chem Prod Process Model*. (2009) **4**:1934–2659. doi: 10.2202/1934-2659.1383
23. Cheng P, Li W, Ogunbona P. Kernel PCA of HOG features for posture detection. In: *VCNZ'09. 24th International Conference on Image and Vision Computing New Zealand, 2009*. Wellington (2009). p. 415–20.
24. Sofman B, Bagnell JA, Stentz A. Anytime online novelty detection for vehicle safeguarding. In: *2010 IEEE International Conference on Robotics and Automation (ICRA)*. Pittsburgh, PA (2010). p. 1247–54.
25. De Vito E, Rosasco L, Toigo A. A universally consistent spectral estimator for the support of a distribution. *Appl Comput Harmonic Anal*. (2014) **37**:185–217. doi: 10.1016/j.acha.2013.11.003
26. Steinwart I. On the influence of the kernel on the consistency of support vector machines. *J Mach Learn Res*. (2002) **2**:67–93. doi: 10.1162/153244302760185252
27. De Vito E, Rosasco L, Toigo A. Spectral regularization for support estimation. In: *NIPS*. Vancouver, BC (2010). p. 1–9.
28. Engl HW, Hanke M, Neubauer A. *Regularization of Inverse Problems. Vol. 375 of Mathematics and its Applications*. Dordrecht: Kluwer Academic Publishers Group (1996).
29. Lo Gerfo L, Rosasco L, Odone F, De Vito E, Verri A. Spectral algorithms for supervised learning. *Neural Comput*. (2008) **20**:1873–97. doi: 10.1162/neco.2008.05-07-517
30. Blanchard G, Mücke N. Optimal rates for regularization of statistical inverse learning problems. In: *Foundations of Computational Mathematics*. (2017). Available online at: https://arxiv.org/abs/1604.04054
31. Rudi A, Odone F, De Vito E. Geometrical and computational aspects of Spectral Support Estimation for novelty detection. *Pattern Recognit Lett*. (2014) **36**:107–16. doi: 10.1016/j.patrec.2013.09.025
32. Rudi A, Canas GD, Rosasco L. On the sample complexity of subspace learning. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. *Advances in Neural Information Processing Systems*. Lake Tahoe: Neural Information Processing Systems Conference (2013). p. 2067–75.
33. Rudi A, Canas GD, De Vito E, Rosasco L. Learning Sets and Subspaces. In: Suykens JAK, Signoretto M, and Argyriou A, editors. *Regularization, Optimization, Kernels, and Support Vector Machines*. Boca Raton, FL: Chapman and Hall/CRC (2014). p. 337.
34. Blanchard G, Bousquet O, Zwald L. Statistical properties of kernel principal component analysis. *Machine Learn*. (2007) **66**:259–94. doi: 10.1007/s10994-006-8886-2
35. Györfi L, Kohler M, Krzy zak A, Walk H. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. New York, NY: Springer-Verlag (2002).
36. Steinwart I, Christmann A. Support Vector Machines. Information Science and Statistics. New York, NY: Springer (2008).
37. Zwald L, Blanchard G. On the Convergence of eigenspaces in kernel principal component analysis. In: Weiss Y, Schölkopf B, Platt J, editors. *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press (2006). p. 1649–56.
38. De Vito E, Rosasco L, Caponnetto A, De Giovannini U, Odone F. Learning from examples as an inverse problem. *J Machine Learn Res*. (2005) **6**:883–904.
39. Rahimi A, Recht B. Random features for large-scale kernel machines. In: Koller D, Schuurmans D, Bengio, Y, Bottou L, editors. *Advances in Neural Information Processing Systems*. Vancouver, BC: Neural Information Processing Systems Conference (2008). p. 1177–84.
40. Rudi A, Camoriano R, Rosasco L. Generalization properties of learning with random features. *arXiv preprint arXiv:160204474* (2016).
41. Smola AJ, Schölkopf B. Sparse greedy matrix approximation for machine learning. In: *Proceeding ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*. Stanford, CA: Morgan Kaufmann (2000). p. 911–18.
42. Rudi A, Camoriano R, Rosasco L. Less is more: nyström computational regularization. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. *Advances in Neural Information Processing Systems*. Montreal, QC: Neural Information Processing Systems Conference (2015). p. 1657–1665.
43. Camoriano R, Angles T, Rudi A, Rosasco L. NYTRO: when subsampling meets early stopping. In: Gretton A, Robert CC, editors. *Artificial Intelligence and Statistics*. Cadiz: Proceedings of Machine Learning Research (2016). p. 1403–11.
44. Rudi A, Carratino L, Rosasco L. FALKON: an optimal large scale Kernel method. *arXiv preprint arXiv:170510958* (2017).
45. Folland G. *A Course in Abstract Harmonic Analysis. Studies in Advanced Mathematics*. Boca Raton, FL: CRC Press (1995).
46. Birman MS, Solomyak M. Double operator integrals in a Hilbert space. *Integr Equat Oper Theor*. (2003) **47**:131–68. doi: 10.1007/s00020-003-1157-8
47. De Vito E, Umanità V, Villa S. A consistent algorithm to solve Lasso, elastic-net and Tikhonov regularization. *J Complex*. (2011) **27**:188–200. doi: 10.1016/j.jco.2011.01.003


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.