Chapter 1

Learning Sets and Subspaces

Alessandro Rudi
DIBRIS, Università degli Studi di Genova
LCSL, Massachusetts Institute of Technology & Istituto Italiano di Tecnologia
alessandro.rudi@unige.it

Guillermo D. Canas
Massachusetts Institute of Technology
guilledc@mit.edu

Ernesto De Vito
DIMA, Università degli Studi di Genova
devito@dima.unige.it

Lorenzo Rosasco
DIBRIS, Università degli Studi di Genova
LCSL, Massachusetts Institute of Technology & Istituto Italiano di Tecnologia
lrosasco@mit.edu

We consider here the classic problem of support estimation, or learning a set from random samples, and propose a natural but novel approach to address it. We do this by investigating its connection with a seemingly distinct problem, namely subspace learning.

The problem of learning the smallest set containing the data distribution is often called support estimation and it is a fundamental problem in statistics and machine learning. As discussed in the following, its applications range from surface estimation, to novelty detection, to name a few. In the following we discuss how a suitable family of positive definite kernels, called separating kernels, allows to relate the problem of learning a set to the problem of learning an appropriate linear subspace of a Hilbert space. More precisely, we reduce the set learning problem to that of learning the smallest subspace that contains the support of the distribution after a kernel (feature) embedding. This connection between learning sets and learning subspaces allows on the one hand to design natural spectral estimators for this problem, and on the other hand to use analytic and probabilistic tools to derive generalization guarantees for them.

Besides establishing this novel connection, the goal of this work is to introduce novel sharp sample complexity estimates for subspace and set learning. The theoretical results are illustrated and complemented through some numerical experiments.

The Chapter is structured as follows. We begin by briefly discussing some concepts from the statistical analysis of unsupervised learning algorithms (Section 1.1). We then develop our analysis of the subspace learning problem, and discuss set learning in Section 1.3. Finally, we conclude in Section 1.4 with some numerical results.

1.1 Unsupervised Statistical Learning

The present work can be more broadly framed in the context of unsupervised learning, a term typically used to describe the general problem of extracting *patterns* from data [23, 19]. Here, the term pattern refers to some geometric property of the data distribution. Specifically, in the sequel we will be interested in recovering the following: 1) the smallest (closed) set containing the data distribution, and 2) the smallest subspace spanned by the data distribution. As we will discuss, these two problems are indeed tightly connected. After formally describing this connection, our focus will be in introducing a class of spectral estimators for this problem, and deriving sharp generalization error estimates for them.

Given a probability space (X, ρ) from which data X_n are drawn identically and independently, we let $\mathcal{S}$ be a set endowed with a (pseudo) metric d. We view $\mathcal{S}$ as the collection of possible patterns/structures in the data distribution (for instance the set of possible supports of a distribution). In many circumstances, the true distribution ρ identifies an element S_ρ in the space of structures (for instance the true support), and the goal of an (unsupervised) learning algorithm is to estimate an approximation $\hat{S}_n$ given the data. For

example, in the context of set learning, $\mathcal{S}$ may be defined as the collection of all closed subsets of X endowed with the Hausdorff distance, and S_ρ to be the support of ρ. In the context of subspace learning, $\mathcal{S}$ is the collection of all linear subspaces of X, with some suitable pseudo-metric such as the reconstruction criterion in (1.3) and S_ρ is the smallest subspace spanned by points drawn from ρ.

Since S_ρ is estimated from random samples, we characterize the learning error of an algorithm through non asymptotic bounds of the form

$$\mathbb{P}\left[d(\hat{S}_n, S_\rho) \leq R_\rho\left(\delta, n\right)\right] \geq 1 - \delta \tag{1.1}$$

for $0 < \delta \leq 1$, where the *learning error* $R_\rho(\delta, n)$ typically depends on n and δ, but also on ρ. Once a bound of the form of (1.1), with an asymptotically vanishing learning error R is obtained, almost sure convergence of $d(\hat{S}_n, S_\rho) \to 0$ as $n \to \infty$ follows from the Borel-Cantelli Lemma [29].

1.2 Subspace learning

Subspace learning is the problem of finding the smallest linear space supporting data drawn from an unknown distribution. It is a classical problem in machine learning and statistics and is at the core of a number of spectral methods for data analysis, most notably PCA [26], but also multidimensional scaling (MDS) [8, 59]. While traditional methods, such as PCA and MDS, perform subspace learning in the original data space, more recent manifold learning methods, such as isomap [51], Hessian eigenmaps [18], maximum-variance unfolding [57, 58, 50], locally-linear embedding [39, 42], and Laplacian eigenmaps [2] (but also kernel PCA [44]), begin by embedding the data in a *feature space*, in which subspace estimation is carried out. As pointed out in [22, 4, 3], all the algorithms in this family have a common structure. They embed the data in a suitable Hilbert space $\mathcal{F}$, and compute a linear subspace that best approximates the embedded data. The local coordinates in this subspace then become the new representation space.

The analysis in this paper applies to learning subspaces both in the data and in a feature space. In the following, we introduce a general formulation of the subspace learning problem and derive novel learning error estimates. Our results rely on natural assumptions on the spectral properties of the covariance operator associated to the data distribution, and hold for a wide class of metrics between subspaces. As a special case, we discuss sharp error estimates for the reconstruction properties of PCA. Key to our analysis is an operator theoretic approach that has broad applicability to the analysis of spectral learning methods.

1.2.1 Problem definition and notation

Given a measure ρ with support M in the unit ball of a separable Hilbert space $\mathcal{F}$, we consider in this work the problem of estimating, from n i.i.d. samples $X_n = \{x_i\}_{1 \leq i \leq n}$, the smallest linear subspace $S_\rho := \overline{\text{span}(M)}$ that contains M. In the framework introduced in Section 1.1, the above problem corresponds to a choice of input space $\mathcal{F}$, and the space of candidate structures is the collection of all linear subspaces of $\mathcal{F}$. The target of the learning problem is S_ρ, the smallest linear subspace that contains the support of ρ. As described in Section 1.1, the quality of an estimate $\hat{S}_n$ of S_ρ, for a given metric (or error criterion) d, is characterized in terms of probabilistic bounds of the form of Equation (1.1).

In the following the metric projection operator onto a subspace S is denoted by P_S, where $P_S^2 = P_S^* = P_S$ (every P is idempotent and self-adjoint). We denote by $\|\cdot\|_{\mathcal{F}}$ the norm induced by the dot product $< \cdot, \cdot >_{\mathcal{F}}$ in $\mathcal{F}$, and by $\|A\|_p := \sqrt[p]{\text{Tr}(|A|^p)}$ the p-Schatten, or p-class norm of a linear bounded operator A [37, p. 84].

1.2.2 Subspace estimators

Spectral estimators can be naturally derived from the characterization of S_ρ in terms of the covariance operator C associated to ρ. Indeed, if $C := \mathbb{E}_{x \sim \rho} x \otimes x$ is the (uncentered) covariance operator associated to ρ, it is easy to show that $S_\rho = \overline{\text{Ran}\, C}$. Similarly, given the empirical covariance $C_n := \frac{1}{n} \sum_{i=1}^n x \otimes x$, we define the *empirical subspace estimate*,

$$\hat{S}_n := \text{span}(X_n) = \text{Ran}\, C_n,$$

where the closure is not needed because $\hat{S}_n$ is finite-dimensional. We also define the *k-truncated (kernel) PCA subspace estimate* $\hat{S}_n^k := \text{Ran}\, C_n^k$, where C_n^k is obtained from C_n by keeping only its k top eigenvalues, see also Section 1.2.5. Note that, since the PCA estimate $\hat{S}_n^k$ is spanned by the top k eigenvectors of C_n, then clearly $\hat{S}_n^k \subseteq \hat{S}_n^{k'}$ for $k < k'$, and therefore $\{\hat{S}_n^k\}_{k=1}^n$ is a nested family of subspaces (all of which are contained in S_ρ). As discussed in Section 1.2.5, since kernel-PCA reduces to regular PCA in a feature space [44] (and can be computed with knowledge of the kernel alone), the following discussion applies equally to kernel-PCA estimates.

1.2.3 Performance criteria

We define the pseudo-metric

$$d_{\alpha,p}(U, V) := \|(P_U - P_V) C^\alpha\|_p \tag{1.2}$$

between subspaces U, V, which is a metric over the collection of subspaces contained in S_ρ, for $0 \leq \alpha \leq \frac{1}{2}$ and $1 \leq p \leq \infty$. Note that $d_{\alpha,p}$ depends on

ρ through C but this dependence is omitted in the notation. A number of important performance criteria can be recovered as particular cases of $d_{\alpha,p}$. In particular, the so-called reconstruction error [46, 7],

$$d_R(S_\rho, \hat{S}) := \mathbb{E}_{x\sim\rho}\|P_{S_\rho}(x) - P_{\hat{S}}(x)\|_{\mathcal{F}}^2 \tag{1.3}$$

is $d_R(S_\rho, \cdot) = d_{1/2,2}(S_\rho, \cdot)^2$. Note that d_R is a natural criterion because a k-truncated PCA estimate minimizes a suitable error d_R over all subspaces of dimension k. Clearly, $d_R(S_\rho, \hat{S})$ vanishes whenever $\hat{S}$ contains S_ρ and, because the family $\{\hat{S}_n^k\}_{k=1}^n$ of PCA estimates is nested, then $d_R(S_\rho, \hat{S}_n^k)$ is non-increasing with k. As shown in [32], a number of unsupervised learning algorithms, including (kernel) PCA, k-means, k-flats, sparse coding, and non-negative matrix factorization, can be written as a minimization of d_R over an algorithm-specific class of sets (e.g. over the set of linear subspaces of a fixed dimension in the case of PCA).

1.2.4 Summary of results

Our main technical contribution is a bound of the form of Eq. (1.1), for the k-truncated PCA estimate $\hat{S}_n^k$ (with the empirical estimate $\hat{S}_n := \hat{S}_n^n$ being a particular case), whose proof is postponed to Sec. 1.5.

We begin by bounding the distance $d_{\alpha,p}$ between S_ρ and the k-truncated PCA estimate $\hat{S}_n^k$, given a known covariance C.

Theorem 1 *Let $\{x_i\}_{1\leq i\leq n}$ be drawn i.i.d. according to a probability measure ρ supported on the unit ball of a separable Hilbert space $\mathcal{F}$, with covariance C. Assuming $n > 3$, $0 < \delta < 1$, $0 \leq \alpha \leq \frac{1}{2}$, $1 \leq p \leq \infty$, the following holds for each $k \in \{1, \ldots, n\}$:*

$$\mathbb{P}\left[d_{\alpha,p}(S_\rho, \hat{S}_n^k) \leq 3t_k^\alpha \left\|C^\alpha(C + t_k I)^{-\alpha}\right\|_p\right] \geq 1 - \delta \tag{1.4}$$

where $t_k = \max\{\sigma_k, \frac{9}{n}\log\frac{n}{\delta}\}$, and σ_k is the k-th top eigenvalue of C.

We say that C has *eigenvalue decay rate of order r* if there are constants $q, Q > 0$ such that $qj^{-r} \leq \sigma_j \leq Qj^{-r}$, where σ_j are the (decreasingly ordered) eigenvalues of C, and $r > 1$. From Equation (1.2) it is clear that, in order for the subspace learning problem to be well-defined, it must be $\|C^\alpha\|_p < \infty$, or alternatively: $\alpha p > 1/r$. Note that this condition is always met for $p = \infty$, and also holds in the reconstruction error case ($\alpha = 1/2, p = 2$), for any decay rate $r > 1$.

Knowledge of an eigenvalue decay rate can be incorporated into Theorem 1 to obtain explicit learning rates, as follows.

Theorem 2 (Polynomial eigenvalue decay) *Let C have eigenvalue decay rate of order r. Under the assumptions of Theorem 1, it holds, with probability*

$1-\delta$*:*

$$d_{\alpha,p}(S_\rho, \hat{S}_n^k) \leq \begin{cases} Q' k^{-r\alpha+\frac{1}{p}} & \text{if } k < k_n^* \quad \textit{(polynomial decay)} \\ Q' k_n^{*-r\alpha+\frac{1}{p}} & \text{if } k \geq k_n^* \quad \textit{(plateau)} \end{cases} \tag{1.5}$$

where it is $k_n^* = \left(\frac{qn}{9\log(n/\delta)}\right)^{1/r}$*, and*

$$Q' = 3\left(Q^{\frac{1}{r}} \frac{\Gamma\left(\alpha p - \frac{1}{r}\right)\Gamma\left(1+\frac{1}{r}\right)}{\Gamma\left(\frac{1}{r}\right)}\right)^{\frac{1}{p}}. \tag{1.6}$$

The above theorem guarantees a decay of $d_{\alpha,p}$ with increasing k, at a rate of $k^{-r\alpha+1/p}$, up to $k = k_n^*$, after which the bound remains constant. The estimated plateau threshold k^* is thus the value of truncation past which the upper bound does not improve. Note that, as described in Section 1.4, this error decay and plateau behavior is observed in practice.

The proofs of Theorems 1 and 2 rely on recent non-commutative Bernstein-type inequalities on operators [5, 52], and a novel analytical decomposition. Note that classical Bernstein inequalities in Hilbert spaces (e.g. [34]) could also be used instead of [52]. While this approach would simplify the analysis, it produces looser bounds, as described in Section 1.5.

If we consider an algorithm that produces, for each set of n samples, an estimate $\hat{S}_n^k$ with $k \geq k_n^*$ then, by plugging the definition of k_n^* into Eq. 1.5, we obtain an upper bound on $d_{\alpha,p}$ as a function of n.

Corollary 3 *Let C have eigenvalue decay rate of order r, and Q', k_n^* be as in Theorem 2. Let $\hat{S}_n^*$ be a truncated subspace estimate $\hat{S}_n^k$ with $k \geq k_n^*$. It is, with probability $1-\delta$,*

$$d_{\alpha,p}(S_\rho, \hat{S}_n^*) \leq Q' \left(\frac{9\left(\log n - \log\delta\right)}{qn}\right)^{\alpha-\frac{1}{rp}}$$

Remark 4 *Note that, by setting $k = n$, the above corollary also provides guarantees on the rate of convergence of the empirical estimate $\hat{S}_n = span(X_n)$ to S_ρ, of order*

$$d_{\alpha,p}(S_\rho, \hat{S}_n) = O\left(\left(\frac{\log n - \log\delta}{n}\right)^{\alpha-\frac{1}{rp}}\right).$$

Corollary 5 and remark 4 are valid for all n such that $k_n^* \leq n$ (or equivalently such that $n^{r-1}(\log n - \log\delta) \geq q/9$). Note that, because ρ is supported on the unit ball, its covariance has eigenvalues no greater than one, and therefore it must be $q < 1$. It thus suffices to require that $n > 3$ to ensure the condition $k_n^* \leq n$ to hold.

1.2.5 Kernel PCA and embedding methods

One of the main applications of subspace learning is to perform dimensionality reduction. In particular, one may find nested subspaces of dimensions $1 \leq k \leq n$ that minimize the distances from the original to the projected samples. This procedure is known as the Karhunen-Loève, PCA, or Hotelling transform [26], and has been generalized to Reproducing-Kernel Hilbert Spaces (RKHS) [44].
In particular, the above procedure amounts to computing an eigendecomposition of the empirical covariance

$$C_n = \sum_{i=1}^{n} \sigma_i u_i \otimes u_i,$$

where the k-th subspace estimate is $\hat{S}_n^k := \operatorname{Ran} C_n^k = \operatorname{span}\{u_i : 1 \leq i \leq k\}$. In the case of kernel PCA, the samples $\{x_i\}_{1 \leq i \leq n}$ belong to some RKHS $\mathcal{F}$, and we can think of them as the embedding $x_i := \phi(z_i)$ of some original data $(z_1, \ldots, z_n) \in Z^n$, where e.g. $Z = \mathbb{R}^D$. The measure ρ can be seen as the measure induced by the embedding and the original data distribution. Interestingly, in practice we may only have indirect information about ϕ in the form a kernel function $K : Z \times Z \to \mathbb{R}$: a symmetric, positive definite function satisfying $K(z, w) = \langle \phi(z), \phi(w) \rangle_{\mathcal{F}}$ [48] (for technical reasons, we also assume K to be continuous). Recall that every such K has a unique associated RKHS, and viceversa [48, p. 120–121], whereas, given K, the embedding ϕ is only unique up to an inner product-preserving transformation. The following reproducing property $f(x) = \langle f, K(z, \cdot) \rangle_{\mathcal{F}}$ holds for all $z \in Z$, $f \in \mathcal{F}$.

If the embedding is defined through a kernel K, it easy to see that the k-truncated kernel PCA can be computed considering the n by n kernel matrix K_n, where $(K_n)_{i,j} = K(x_i, x_j)$ [44]. It is easy to see that the k-truncated kernel PCA subspace $\hat{S}_n^k$ minimizes the empirical reconstruction error $d_R(\hat{S}_n, \hat{S})$, among all subspaces $\hat{S}$ of dimension k. Indeed, it is

$$\begin{aligned} d_R(\hat{S}_n, \hat{S}) &= \mathbb{E}_{x \sim \hat{\rho}} \|x - P_{\hat{S}}(x)\|_{\mathcal{F}}^2 = \mathbb{E}_{x \sim \hat{\rho}} \left\langle (I - P_{\hat{S}})x, (I - P_{\hat{S}})x \right\rangle_{\mathcal{F}} \\ &= \mathbb{E}_{x \sim \hat{\rho}} \left\langle I - P_{\hat{S}}, x \otimes x \right\rangle_{HS} = \left\langle I - P_{\hat{S}}, C_n \right\rangle_{HS}, \end{aligned} \tag{1.7}$$

where $\langle \cdot, \cdot \rangle_{HS}$ is the Hilbert-Schmidt inner product. From this, it clearly follows that the k-dimensional subspace minimizing Equation 1.7 (maximizing $\langle P_{\hat{S}}, C_n \rangle$) is spanned by the k top eigenvectors of C_n. Since we are interested in the expected error $d_R(S_\rho, \hat{S}_n^k)$ of the kernel PCA estimate (rather than the empirical error $d_R(\hat{S}_n, \hat{S})$), we may obtain a learning rate for Equation 1.7 by specializing Theorem 2 to the reconstruction error, for all k (Theorem 2), and for $k \geq k^*$ with a suitable choice of k^* (Corollary 5). In particular, recalling that $d_R(S_\rho, \cdot) = d_{\alpha,p}(S_\rho, \cdot)^2$ with $\alpha = 1/2$ and $p = 2$, and choosing a value of $k \geq k_n^*$ that minimizes the bound of Theorem 2, we obtain the following result.

Corollary 5 (Performance of PCA / Reconstruction error) *Let C have eigenvalue decay rate of order r, and $\hat{S}_n^*$ be as in Corollary 3. Then it holds, with probability $1-\delta$,*

$$d_R(S_\rho, \hat{S}_n^*) = O\left(\left(\frac{\log n - \log \delta}{n}\right)^{1-1/r}\right).$$

1.2.6 Comparison with previous results in the literature

Figure 1.1 shows a comparison of our learning rates with existing rates in the literature [7, 46]. The plot shows the polynomial decay rate c of the high probability bound $d_R(S_\rho, \hat{S}_n^k) = O(n^{-c})$, as a function of the eigenvalue decay rate r of the covariance C, computed at the best value k_n^* (which minimizes the bound).

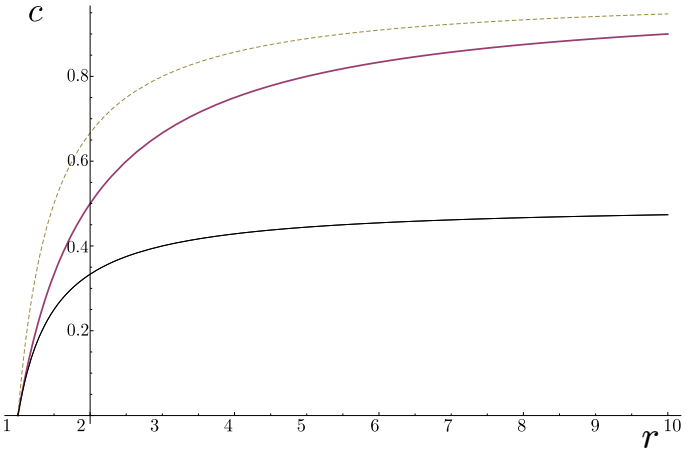

FIGURE 1.1: Known upper bounds for the polynomial decay rate c (for the best choice of k), for the expected distance from a random sample to the empirical k-truncated kernel-PCA estimate, as a function of the covariance eigenvalue decay rate (higher is better). Our bound (purple line), consistently outperforms previous ones [46] (black line). The top (dashed) line [7], has significantly stronger assumptions, and is only included for completeness.

The learning rate exponent c, under a polynomial eigenvalue decay assumption of the data covariance C, is $c = \frac{s(r-1)}{r-s+sr}$ for [7] and $c = \frac{r-1}{2r-1}$ for [46], where s is related to the fourth moment. Note that, among the two (purple and black) that operate under the same assumptions, our bound (purple line) is the best by a wide margin. The top, best performing, dashed line [7] is obtained for the best possible fourth-order moment constraint $s = 2r$, and is therefore not a fair comparison. However, it is worth noting that our bounds

perform almost as well as the most restrictive one, even when we do not include any fourth-order moment constraints.

Choice of truncation parameter k. Since, as pointed out in Section 1.2.2, the subspace estimates $\hat{S}_n^k$ are nested for increasing k (i.e. $\hat{S}_n^k \subseteq \hat{S}_n^{k'}$ for $k < k'$), the distance $d_{\alpha,p}(S_\rho, \hat{S}_n^k)$, and in particular the reconstruction error $d_R(S_\rho, \hat{S}_n^k)$, is a non-increasing function of k. As discussed [7], this suggests that there is no bias-variace trade- off in the choice of k. Indeed, the fact that the estimates $\hat{S}_n^k$ become increasingly close to S_ρ as k increases indicates that, when minimizing $d_{\alpha,p}(S_\rho, \hat{S}_n^k)$, the best choice is simply $k = n$.

Interestingly, however, both in practice (Section 1.4), and in theory (Section 1.2.4), we observe that a typical behavior for the subspace learning problem in high dimensions (e.g. kernel PCA) is that there is a certain value of $k = k_n^*$, past which performance plateaus. For problems such as spectral embedding methods [51, 18, 58], in which a degree of dimensionality reduction is desirable, producing an estimate $\hat{S}_n^k$ where k is close to the plateau threshold may be a natural parameter choice: it leads to an estimate of the lowest dimension ($k = k_n^*$), whose distance to the true S_ρ is almost as low as the best-performing one ($k = n$).

1.3 Set learning

The problem of set, or support estimation has received a great deal of attention in the Statistics community since the sixties [36, 21], and since then a number of practical approaches have been proposed to address it [17, 27, 20, 11, 53, 43, 13, 35, 49, 55, 45, 6, 12]. Support estimation is often considered in machine learning in situations in which it is difficult to gather negative examples (as it often happens in biological and biomedical problems) or when the negative class is not well defined (as in object detection problems in computer vision), as is the case in one class estimation [43], and novelty and anomaly detection [31, 9].

In this section, we describe an approach that is based on reducing the set learning problem to that of learning a subspace. The results in this section largely draw from [40, 14, 41].

1.3.1 Set Learning via Subspace Learning

We begin by recalling how support estimation can be reduced to subspace learning, and discuss how our results specialize to this setting. From an algorithmic perspective, the approach we discuss is closely related to the one in [25] and has been successfully applied in several practical domains [38, 28, 54, 24, 30, 10, 47].

Central to the connection between set and subspace learning is the notion of *separating kernel* and *separating feature map*, which was introduced in [15].

Let K be a reproducing kernel on some space Z, and $(\phi, \mathcal{F})$ an associated feature map and feature space pair (see section 1.2.5). For simplicity, we assume that $\|\phi(z)\|_{\mathcal{F}} = 1$ for all $z \in Z$. This assumption is without loss of generality because a kernel with non-zeros in its diagonal can always be normalized. Given a non-empty set $C \subseteq Z$, let $\mathcal{F}_C = \overline{\text{span}\{\phi(z) \mid z \in C\}}$ be the closure of all finite linear combinations of points in the range $\phi(C)$ of C. The distance from any given point $\phi(z)$, with $z \in Z$, to the linear subspace $\mathcal{F}_C$ is

$$d_{\mathcal{F}_C}(\phi(z)) := \inf_{f \in \mathcal{F}_C} \|\phi(z) - f\|_{\mathcal{F}}.$$

The following, key definition is equivalent to the *separating property* in Definition 1 of [56].

Definition 1 *We say that a feature map ϕ (and hence the corresponding kernel) separates a set $C \subset Z$ if for all $z \in Z$ it holds:*

$$d_{\mathcal{F}_C}(\phi(z)) = 0 \quad \textit{iff} \quad z \in C.$$

An example of separating kernel for $\mathbb{R}^d$ is the exponential kernel $K(x, x') = e^{-\|x-x'\|}$. The proof of this fact, see [14], crucially depends on the fact that for each compact subset of $\mathbb{R}^d$ the associated reproducing kernel Hilbert space contains functions that are zero on the set and non-zero outside. Interestingly, the Gaussian kernel is not separating, because the associated Hilbert space contains only analytic functions, and the only function that is zero on a compact subset (with non-empty interior) is the zero function.

The separating property has a clear geometric interpretation in the feature space: the set $\phi(S_\rho)$ is the intersection of the closed subspace $\mathcal{F}_{S_\rho}$ (the smallest linear subspace containing $\phi(S_\rho)$), and $\phi(Z)$ (see Figure 1.2).

Using the notion of separating kernel, the support S_ρ can be characterized in terms of the subspace $\mathcal{F}_\rho = \overline{\text{span } \phi(S_\rho)} \subseteq \mathcal{F}$. More precisely, it can be shown (see the next subsection) that, if the feature map ϕ *separates* S_ρ, then it is

$$S_\rho = \{z \in Z \mid d_{\mathcal{F}_\rho}(\phi(z)) = 0\}.$$

The above discussion naturally leads to an empirical estimate $\hat{S}_n = \{z \in Z \mid d_{\mathcal{F}_n}(\phi(z)) \leq \tau\}$ of S_ρ, where $\hat{\mathcal{F}}_n = \overline{\text{span } \phi(Z_n)}$, and $\tau > 0$. Given a training set $z_1, \ldots, z_n$, the estimator $\hat{S}_n$ is therefore the set of points $z \in Z$ whose associated distance from $\phi(z)$ to the linear space spanned by $\{\phi(z_1), \ldots, \phi(z_n)\}$ is sufficiently small, according to some tolerance τ. Any point with distance greater than τ will be considered to be *outside of the support* by this estimator.

With the above choice of estimator, it can be shown that almost sure convergence $\lim_{n\to\infty} d_H(S_\rho, \hat{S}_n) = 0$ in the Hausdorff distance [1] is related to the convergence of $\hat{\mathcal{F}}_n$ to $\mathcal{F}_\rho$ [15]. More precisely, if the eigenfunctions of the covariance operator $C = \mathbb{E}_{z\sim\rho}[\phi(z) \otimes \phi(z)]$ are uniformly bounded, then it

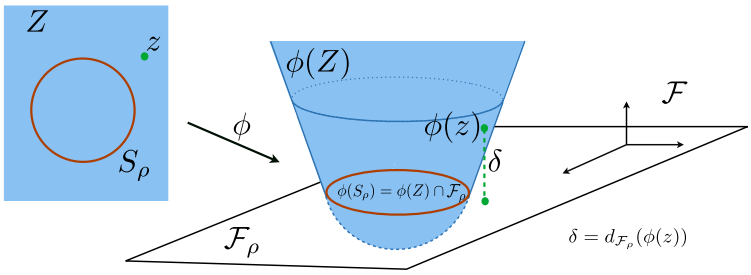

FIGURE 1.2: The input space Z and the support S_ρ are mapped into the feature space $\mathcal{F}$ by the feature map ϕ. Letting $\mathcal{F}_\rho := \mathcal{F}_{S_\rho}$ be the smallest linear subspace containing $\phi(S_\rho)$ then, if the kernel is separable, the image of the support $\phi(S_\rho)$ is given by the intersection between $\phi(Z)$ and $\mathcal{F}_\rho$. By the separating property, a point z belongs to the support if and only the distance between $\phi(z)$ and $\mathcal{F}_\rho$ is zero.

suffices for Hausdorff convergence to bound from above $d_{\frac{r-1}{2r},\infty}$ (where $r > 1$ is the eigenvalue decay rate of C) as shown in Section 1.3.2.

1.3.2 Consistency results

Before proving the consistency of the set estimator $\hat{S}_n$ we show an improved learning rate for the associated subspace $\hat{\mathcal{F}}_n$. In particular we study the linear subspace $\hat{\mathcal{F}}_n^*$ that is the one spanned by the first k components of the empirical covariance matrix $\hat{C}_n$, where $k \geq k_n^*$ (see Subsection 1.2.5 and Theorem 2). Note that $\hat{\mathcal{F}}_n^* = \hat{\mathcal{F}}_n$ when $k = n$. The following result specializes Corollary 3 to this setting.

Corollary 6 (Performance of KPCA with the set learning metric) *If* $0 \leq \alpha \leq \frac{1}{2}$*, then it holds, with probability* $1 - \delta$*,*

$$d_{\alpha,\infty}(\mathcal{F}_\rho, \hat{\mathcal{F}}_n^*) = O\left(\left(\frac{\log n - \log \delta}{n}\right)^\alpha\right)$$

where the constant in the Landau symbol does not depend on δ*.*

Letting $\alpha = \frac{r-1}{2r}$ above yields a high probability bound of order $O\left(n^{-\frac{r-1}{2r}}\right)$ (up to logarithmic factors), which is considerably sharper than the bound $O\left(n^{-\frac{r-1}{2(3r-1)}}\right)$ found in [16] (Theorem 7). Note that these are upper bounds for the best possible choice of k (which minimizes the bound). While the optima of both bounds vanish with $n \to \infty$, their behavior is qualitatively different. In particular, the bound of [16] is U-shaped, and diverges for $k = n$, while ours is L-shaped (no trade-off), and thus also convergent for $k = n$. Therefore, when compared with [16], our results suggest that no regularization is required from a statistical point of view though, as discussed in the following, it may be required for numerical stability. With the above tools at hand, we are now in a position to prove the consistency of $\hat{S}_n$.

Theorem 7 (Consistency of Set Learning) *Let the input space Z be metrizable, K be a kernel on Z with the separating property [15], let the dimension k of the empirical subspace $\hat{\mathcal{F}}_n^*$, be $k_n^* \leq k \leq n$ and the threshold parameter $\tau = \max_{1 \leq i \leq n} d_{\hat{\mathcal{F}}_n^k}(\phi(z_i))$, then*

$$\hat{S}_n^* = \left\{ z \in Z \;\middle|\; d_{\hat{\mathcal{F}}_n^*}(\phi(z)) \leq \tau \right\} \tag{1.8}$$

is a universally consistent unsupervised learning algorithm.

Proof. By Theorem 6 of [56] and our Corollary 6, the estimator $\hat{S}_n^*$ satisfies the universal consistency conditions given in Section 1.1, under the given hypotheses. □

We note that the above result is an example of how kernel embedding techniques can used to provably estimate geometric invariants of the the original data distribution. Note that the considered estimator achieve this without having to explicitly solve a pre-image problem [33].

We end this section noting that, while, as proven in Corollary 6, regularization is not needed from a statistical perspective, it can play a role in ensuring numerical stability in practice. Indeed, in order to find $\hat{S}$, we compute $d_{\hat{\mathcal{F}}_n}(\phi(z))$ with $z \in Z$. Using the reproducing property of K, it can be shown that, for $z \in Z$, it is $d_{\hat{\mathcal{F}}_n^k}(\phi(z)) = K(z,z) - \left\langle t_z, (\hat{K}_n^k)^\dagger t_z \right\rangle$ where $(t_z)_i = K(z, z_i)$, $\hat{K}_n$ is the Gram matrix $(\hat{K}_n)_{ij} = K(z_i, z_j)$, $\hat{K}_n^k$ is the rank-k approximation of $\hat{K}_n$, and $(\hat{K}_n^k)^\dagger$ is the pseudo-inverse of $\hat{K}_n^k$. The computation of $\hat{S}$ therefore requires a matrix inversion, which is prone to instability for high condition numbers. Figure 1.3 shows the behavior of the error that results from replacing $\hat{\mathcal{F}}_n$ by its k-truncated approximation $\hat{\mathcal{F}}_n^k$. For large values of k, the small eigenvalues of $\hat{\mathcal{F}}_n$ are used in the inversion, leading to numerical instability.

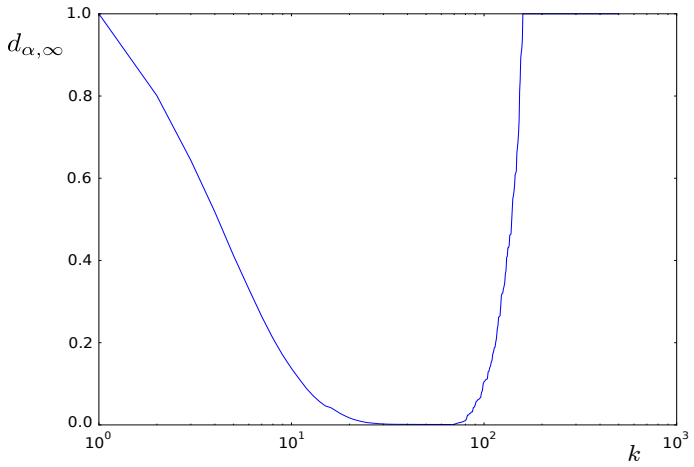

FIGURE 1.3: The experimental behavior of the distance $d_{\alpha,\infty}(\hat{S}^k, S_\rho)$ between the empirical and the actual support subspaces, with respect to the regularization parameter. The setting is the one of section 1.4. Here the actual subspace is analytically computed, while the empirical one is computed on a dataset with $n = 1000$ and 32bit floating point precision. Note the numerical instability as k tends to 1000.

1.4 Numerical experiments

In order to validate our analysis empirically, we consider the following experiment. Let ρ be a uniform one-dimensional distribution in the unit interval. We embed ρ into a reproducing-kernel Hilbert space $\mathcal{F}$ using the exponential of the ℓ_1 distance ($k(u, v) = \exp\{-\|u - v\|_1\}$) as kernel. Given n samples drawn from ρ, we compute its empirical covariance in $\mathcal{F}$ (whose spectrum is plotted in Figure 1.4 (top)), and truncate its eigen-decomposition to obtain a subspace estimate $\hat{\mathcal{F}}_n^k$, as described in Section 1.2.2.

Figure 1.4 (right) is a box plot of reconstruction error $d_R(\mathcal{F}_\rho, \hat{\mathcal{F}}_n^k)$ associated with the k-truncated kernel-PCA estimate $\hat{\mathcal{F}}_n^k$ (the expected distance in $\mathcal{F}$ of samples to $\hat{\mathcal{F}}_n^k$), with $n = 1000$ and varying k. While d_R is computed analytically in this example, and $\mathcal{F}_\rho$ is fixed, the estimate $\hat{\mathcal{F}}_n^k$ is a random variable, and hence the variability in the graph. Notice from the figure that, as pointed out in [7] and discussed in Section 1.2.6, the reconstruction error $d_R(\mathcal{F}_\rho, \hat{\mathcal{F}}_n^k)$ is always a non-increasing function of k, due to the fact that the kernel-PCA estimates are nested: $\hat{\mathcal{F}}_n^k \subset \hat{\mathcal{F}}_n^{k'}$ for $k < k'$ (see Section 1.2.2). The graph is highly concentrated around a curve with a steep intial decay, until reaching some sufficiently high k, past which the reconstruction (pseudo) dis-

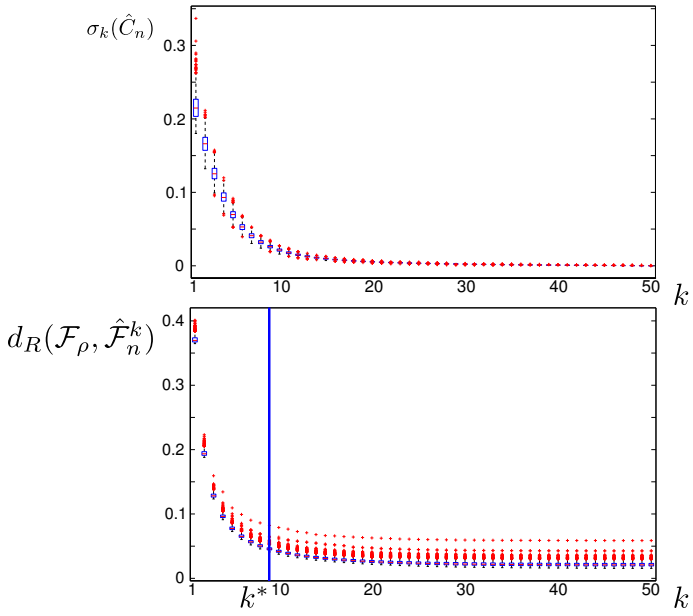

FIGURE 1.4: The spectrum of the empirical covariance (top), and the expected distance from a random sample to the empirical k-truncated kernel-PCA subspace estimate (bottom), as a function of k ($n = 1000$, 1000 trials shown in a boxplot). Our predicted plateau threshold k_n^* (Theorem 2) is a good estimate of the value k past which the distance stabilizes.

tance becomes stable, and does not vanish. In our experiments, this behavior is typical for the reconstruction distance and high-dimensional problems.

Due to the simple form of this example, we are able to compute analytically the spectrum of the true covariance C. In this case, the eigenvalues of C decay as $2\gamma/((k\pi)^2 + \gamma^2)$, with $k \in \mathbb{N}$, and therefore they have a polynomial decay rate $r = 2$ (see Section 1.2.4). Given the known spectrum decay rate, we can estimate the plateau threshold $k = k_n^*$ in the bound of Theorem 2, which can be seen to be a good approximation of the observed start of a plateau in $d_R(\mathcal{F}_\rho, \hat{\mathcal{F}}_n^k)$ (Figure 1.4, right). Notice that our bound for this case (Corollary 5) similarly predicts a steep error decay until the threshold $k = k_n^*$ (indicated in the figure by the vertical blue line), and a plateau afterwards.

1.5 Sketch of the proofs

For the sake of completeness we sketch the main step in the proof of our main theoretical result, Theorem 1, with some details omitted in the interest of conciseness.

For each $\lambda > 0$, we denote by $r^\lambda(x) := \mathbf{1}\{x > \lambda\}$ the step function with a cut-off at λ. Given an empirical covariance operator C_n, we will consider the truncated version $r^\lambda(C_n)$ where, in this notation, r^λ is applied to the eigenvalues of C_n, that is, $r^\lambda(C_n)$ has the same eigen-structure as C_n, but its eigenvalues that are less or equal to λ are clamped to zero.

In order to prove the bound of Equation (1.4), we begin by proving a more general upper bound of $d_{\alpha,p}(S_\rho, \hat{S}_n^k)$, which is split into a random ($\mathcal{A}$), and a deterministic part ($\mathcal{B}, \mathcal{C}$). The bound holds for all values of a free parameter $t > 0$, which is then constrained and optimized in order to find the (close to) tightest version of the bound.

Lemma 8 *Let $t > 0$, $0 \leq \alpha \leq 1/2$, and $\lambda = \sigma_k(C)$ be the k-th top eigenvalue of C, it is,*

$$d_{\alpha,p}(S_\rho, \hat{S}_n^k) \leq \underbrace{\|(C+tI)^{\frac{1}{2}}(C_n+tI)^{-\frac{1}{2}}\|_\infty^{2\alpha}}_{\mathcal{A}} \cdot \tag{1.9}$$

$$\cdot \underbrace{\{3/2(\lambda+t)\}^\alpha}_{\mathcal{B}} \cdot \underbrace{\|C^\alpha(C+tI)^{-\alpha}\|_p}_{\mathcal{C}} \tag{1.10}$$

Note that the right-hand side of Equation (1.9) is the product of three terms, the left of which ($\mathcal{A}$) involves the empirical covariance operator C_n, which is a random variable, and the right two ($\mathcal{B}, \mathcal{C}$) are entirely deterministic. While the term $\mathcal{B}$ has already been reduced to the known quantities t, α, λ, the remaining terms are bounded next. We bound the random term $\mathcal{A}$ in the next Lemma, whose proof makes use of recent concentration results [52].

Lemma 9 (Term $\mathcal{A}$) *Let $0 \leq \alpha \leq 1/2$, for each $\frac{9}{n}\log\frac{n}{\delta} \leq t \leq \|C\|_\infty$, with probability $1-\delta$ it is*

$$(2/3)^\alpha \leq \|(C+tI)^{\frac{1}{2}}(C_n+tI)^{-\frac{1}{2}}\|_\infty^{2\alpha} \leq 2^\alpha$$

Lemma 10 (Term $\mathcal{C}$) *Let C be a symmetric, bounded, positive semidefinite linear operator on $\mathcal{F}$. If $\sigma_k(C) \leq f(k)$ for $k \in \mathbb{N}$, where f is a decreasing function then, for all $t > 0$ and $\alpha \geq 0$, it holds*

$$\left\|C^\alpha(C+tI)^{-\alpha}\right\|_p \leq \inf_{0\leq u\leq 1} g_{u\alpha} t^{-u\alpha} \tag{1.11}$$

where $g_{u\alpha} = \left(f(1)^{u\alpha p} + \int_1^\infty f(x)^{u\alpha p} dx\right)^{1/p}$. Furthermore, if $f(k) = gk^{-1/\gamma}$, with $0 < \gamma < 1$ and $\alpha p > \gamma$, then it holds

$$\left\|C^\alpha(C+tI)^{-\alpha}\right\|_p \leq Q t^{-\gamma/p} \tag{1.12}$$

where $Q = (g^{\gamma}\Gamma(\alpha p - \gamma)\Gamma(1+\gamma)/\Gamma(\gamma))^{1/p}$.

The combination of Lemmas 8 and 9 leads to the main theorem 1, which is a probabilistic bound, holding for every $k \in \{1, \ldots, n\}$, with a deterministic term $\|C^{\alpha}(C+tI)^{-\alpha}\|_p$ that depends on knowledge of the covariance C. In cases in which some knowledge of the decay rate of C is available, Lemma 10 can be applied to obtain Theorem 2 and Corollary 3. Finally, Corollary 5 is simply a particular case for the reconstruction error $d_R(S_\rho, \cdot) = d_{\alpha,p}(S_\rho, \cdot)^2$, with $\alpha = 1/2, p = 2$.

As noted in Section 1.2.4, looser bounds would be obtained if classical Bernstein inequalities in Hilbert spaces [34] were used instead. In particular, Lemma 9 would result in a range for t of $qn^{-r/(r+1)} \leq t \leq \|C\|_{\infty}$, implying $k^* = O(n^{1/(r+1)})$ rather than $O(n^{1/r})$, and thus Theorem 2 would become (for $k \geq k^*$) $d_{\alpha,p}(S_\rho, S_n^k) = O(n^{-\alpha r/(r+1)+1/(p(r+1))})$ (compared with the sharper $O(n^{-\alpha+1/rp})$ of Theorem 2). For instance, for $p = 2$, $\alpha = 1/2$, and a decay rate $r = 2$ (as in the example of Section 1.4), it would be: $d_{1/2,2}(S_\rho, S_n) = O(n^{-1/4})$ using Theorem 2, and $d_{1/2,2}(S_\rho, S_n) = O(n^{-1/6})$ using classical Bernstein inequalities.

1.6 Conclusions

The problem of set learning consists in estimating the smallest subset of the input space containing the data distribution. In this chapter the problem has been investigated by analyzing its relations with subspace learning, that consists in estimating the smallest linear subspace containing the distribution. In particular we showed that, given a suitable feature map, the set learning problem can be cast as a subspace learning problem in the associated feature space. In order to analyze the theoretical properties of the first problem, the statistical analysis for the second has been developed obtaining novel and sharper sample complexity upper bounds. Finally, by exploiting such results, the consistency of set learning has been established. The chapter is concluded by numerical examples that show the effectiveness of our analysis.

Acknowledgments L. R. acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC.

Bibliography

[1] G. Beer. *Topologies on Closed and Closed Convex Sets*. Springer, 1993.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Y. Bengio, O. Delalleau, N.L. Roux, J.F. Paiement, P. Vincent, and M. Ouimet. Learning eigenfunctions links spectral embedding and kernel pca. *Neural Computation*, 16(10):2197–2219, 2004.

[4] Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, 16:177–184, 2004.

[5] S. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

[6] G. Biau, B. Cadre, D. Mason, and Bruno Pelletier. Asymptotic normality in density support estimation. *Electron. J. Probab.*, 14:no. 91, 2617–2635, 2009.

[7] G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294, 2007.

[8] I. Borg and P.J.F. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[9] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[10] Peng Cheng, Wanqing Li, and Philip Ogunbona. Kernel pca of hog features for posture detection. In *Image and Vision Computing New Zealand, 2009. IVCNZ'09. 24th International Conference*, pages 415–420. IEEE, 2009.

[11] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *Ann. Statist.*, 25(6):2300–2312, 1997.

[12] A. Cuevas and R. Fraiman. Set estimation. In *New perspectives in stochastic geometry*, pages 374–397. Oxford Univ. Press, Oxford, 2010.

[13] A. Cuevas and A. Rodríguez-Casal. Set estimation: an overview and some recent developments. In *Recent advances and trends in nonparametric statistics*, pages 251–264. Elsevier B. V., Amsterdam, 2003.

[14] E. De Vito, Lorenzo Rosasco, and Alessandro Toigo. A universally consistent spectral estimator for the support of a distribution. *Applied Computational Harmonic Analysis*, 2014. in press, DOI 10.1016/j.acha.2013.11.003.

[15] Ernesto De Vito, Lorenzo Rosasco, and Alessandro Toigo. Spectral regularization for support estimation. *Advances in Neural Information Processing Systems, NIPS Foundation*, pages 1–9, 2010.

[16] Ernesto De Vito, Lorenzo Rosasco, and Alessandro Toigo. Learning sets with separating kernels. *arXiv:1204.3573*, 2012.

[17] L. Devroye and G. L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, 38(3):480–488, 1980.

[18] D.L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[19] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[20] L. Dümbgen and G. Walther. Rates of convergence for random approximations of convex sets. *Adv. in Appl. Probab.*, 28(2):384–393, 1996.

[21] J. Geffroy. Sur un probleme d'estimation géométrique. *Publ. Inst. Statist. Univ. Paris*, 13:191–210, 1964.

[22] J. Ham, D.D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, page 47. ACM, 2004.

[23] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[24] Fei He, Jian Hong Yang, Min Li, and Jin Wu Xu. Research on nonlinear process monitoring and fault diagnosis based on kernel principal component analysis. *Key Engineering Materials*, 413:583–590, 2009.

[25] H. Hoffmann. Kernel pca for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007.

[26] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[27] A. P. Korostelëv and A. B. Tsybakov. *Minimax theory of image reconstruction*. Springer-Verlag, New York, 1993.

[28] Hyoung-Joo Lee, Sungzoon Cho, and Min-Sup Shin. Supporting diagnosis of attention-deficit hyperactive disorder with novelty detection. *Artificial intelligence in medicine*, 42(3):199–212, 2008.

[29] Michel Loève. *Probability theory*, volume 45. Springer, 1963.

[30] Mauricio L Maestri, Miryan C Cassanello, and Gabriel I Horowitz. Kernel pca performance in processes with multiple operation modes. *Chemical Product and Process Modeling*, 4(5), 2009.

[31] M. Markou and S. Singh. Novelty detection: a review–part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[32] Andreas Maurer and Massimiliano Pontil. K–dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

[33] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *NIPS*, volume 11, pages 536–542, 1998.

[34] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[35] M. Reitzner. Random polytopes and the Efron-Stein jackknife inequality. *Ann. Probab.*, 31(4):2136–2166, 2003.

[36] A. Rényi and R. Sulanke. Über die konvexe hülle von n zufällig gewählten punkten. *Probability Theory and Related Fields*, 2(1):75–84, 1963.

[37] J.R. Retherford. *Hilbert Space: Compact Operators and the Trace Theorem*. London Mathematical Society Student Texts. Cambridge University Press, 1993.

[38] Branko Ristic, Barbara La Scala, Mark Morelande, and Neil Gordon. Statistical analysis of motion patterns in ais data: Anomaly detection and motion prediction. In *Information Fusion, 2008 11th International Conference on*, pages 1–7. IEEE, 2008.

[39] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[40] Alessandro Rudi, Guillermo D Canas, and Lorenzo Rosasco. On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, pages 2067–2075, 2013.

[41] Alessandro Rudi, Francesca Odone, and Ernesto De Vito. Geometrical and computational aspects of spectral support estimation for novelty detection. *Pattern Recognition Letters*, 36:107–116, 2014.

[42] L.K. Saul and S.T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, 4:119–155, 2003.

[43] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, 2001.

[44] B. Schölkopf, A. Smola, and K.R. Müller. Kernel principal component analysis. *Artificial Neural Networks-ICANN'97*, pages 583–588, 1997.

[45] C. D. Scott and R. D. Nowak. Learning minimum volume sets. *J. Mach. Learn. Res.*, 7:665–704, 2006.

[46] J. Shawe-Taylor, C. K. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *Information Theory, IEEE Transactions on*, 51(7), 2005.

[47] Boris Sofman, James A Bagnell, and Anthony Stentz. Anytime online novelty detection for vehicle safeguarding. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 1247–1254. IEEE, 2010.

[48] I. Steinwart and A. Christmann. *Support vector machines*. Information science and statistics. Springer-Verlag. New York, 2008.

[49] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *J. Mach. Learn. Res.*, 6:211–232 (electronic), 2005.

[50] J. Sun, S. Boyd, L. Xiao, and P. Diaconis. The fastest mixing markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM review*, 48(4):681–699, 2006.

[51] J.B. Tenenbaum, V. De Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[52] J.A. Tropp. User-friendly tools for random matrices: An introduction. 2012.

[53] A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969, 1997.

[54] Francisco J Valero-Cuevas, Heiko Hoffmann, Manish U Kurse, Jason J Kutch, and Evangelos A Theodorou. Computational models for neuromuscular function. *Biomedical Engineering, IEEE Reviews in*, 2:110–135, 2009.

[55] R. Vert and J.-P. Vert. Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research*, 7:817–854, 2006.

[56] Ernesto De Vito, Lorenzo Rosasco, and Alessandro Toigo. Learning sets with separating kernels. *Applied and Computational Harmonic Analysis*, 2013.

[57] K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–988. IEEE, 2004.

[58] K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[59] C.K.I. Williams. On a connection between kernel pca and metric multidimensional scaling. *Machine Learning*, 46(1):11–19, 2002.

Index