

# Accuracy of suboptimal solutions to kernel principal component analysis

Giorgio Gnecco · Marcello Sanguineti

Received: 23 May 2006 / Revised: 1 October 2007  
© Springer Science+Business Media, LLC 2007

**Abstract** For Principal Component Analysis in Reproducing Kernel Hilbert Spaces (KPCA), optimization over sets containing only linear combinations of all  $n$ -tuples of kernel functions is investigated, where  $n$  is a positive integer smaller than the number of data. Upper bounds on the accuracy in approximating the optimal solution, achievable without restrictions on the number of kernel functions, are derived. The rates of decrease of the upper bounds for increasing number  $n$  of kernel functions are given by the summation of two terms, one proportional to  $n^{-1/2}$  and the other to  $n^{-1}$ , and depend on the maximum eigenvalue of the Gram matrix of the kernel with respect to the data. Primal and dual formulations of KPCA are considered. The estimates provide insights into the effectiveness of sparse KPCA techniques, aimed at reducing the computational costs of expansions in terms of kernel units.

**Keywords** Principal component analysis (PCA) · Kernel methods · Suboptimal solutions · Primal and dual problems · Lagrangian · Regularized optimization problems

## 1 Introduction

*Principal Component Analysis* (PCA) [14] aims at representing in a compact way a set of data by extracting a certain number of so-called *features*, which contain “most of the information” conveyed by the data. When the data have a finite-dimensional

---

G. Gnecco · M. Sanguineti (✉)  
Department of Communications, Computer and System Sciences (DIST), University of Genova,  
Via Opera Pia 13, 16145 Genoa, Italy  
e-mail: marcello@dist.unige.it

G. Gnecco  
Department of Mathematics (DIMA), University of Genova, Via Dodecaneso 35, 16146 Genoa, Italy  
e-mail: giorgio.gnecco@dist.unige.it

representation, i.e., they are vectors of  $\mathbb{R}^d$ , for some positive integer  $d$ , PCA corresponds to an orthogonal transformation of the coordinate system, followed by the projection of the data onto a subset of the new coordinates, which are called *principal components* (hence the name of “Principal Component Analysis”). This technique is motivated by the fact that often only a small number of properly-chosen components contain nearly all the information needed to subsequent processing of the data.

The solution to PCA can be obtained by the eigenvectors of the sample covariance matrix of the data. By replacing such a matrix with the *Gram matrix of the data* (whose entries are the inner products between all pairs of data), the dependence of the PCA algorithm on the data can be expressed simply via their inner products, which represent a “measure of similarity” between data pairs [29, Chap. 14]. It is possible to consider generalizations of classical PCA by introducing a mapping from  $\mathbb{R}^d$  to another inner-product space, called *feature space*, which allows one to model more general measures of similarity between pairs of data. The choice of the mapping is usually based on *prior information* about the problem at hand.

A large class of feature spaces and mappings can be studied via the so-called *kernel trick* [29, p. 34]. When a kernel function can be used to compute inner products in the feature space, the corresponding generalization of PCA is called *Kernel Principal Component Analysis* (KPCA).

By exploiting the close relationship between kernel functions and the theory of Hilbert spaces of a special type, called *Reproducing Kernel Hilbert Spaces* (RKHSs), it is possible to restate KPCA in the RKHS setting. Reproducing Kernel Hilbert spaces were defined by Aronszajn [2], exploiting work from [32] introduced into applications related to learning by Parzen [23] and Wahba [39], and first used in learning theory by Cortes and Vapnik [5] and Girosi [12]. Roughly speaking, norms on RKHSs often play the role of measures of various types of oscillations of functions. Since the choice of a specific kernel in the feature space can be restated as the choice of a suitable RKHS, such a space offers an equivalent way to see how one can impose desired properties on the candidate solutions to PCA in the feature space (for more details, see Appendix 3).

If  $m$  is the cardinality of the sample of data, then computing the solutions to PCA and KPCA require one to store the  $m \times m$  Gram matrix of the data and to solve for such a matrix a symmetric eigenvalue problem ([8, 27, 34], [29, Chapter 14]). In the case of large data sets, both requirements pose serious problems in terms of increasing computational burden. Various approaches have been proposed to cope with these drawbacks; see [1], [29, Chap. 14], [30, Sect. 4.3], and the references therein.

In this paper, motivated by the above-mentioned computational issues, we study suboptimal solutions to KPCA (in the RKHS setting) by using tools from optimization, nonlinear approximation, and regularization. We start from the formulation of KPCA as an optimization problem in which a functional defined over a RKHS has to be maximized. We consider two ways in which the behavior of admissible solutions can be specified: (i) by restricting optimization to a subset of the RKHS and (ii) by replacing the functional to be optimized with its linear combination with another functional. The two formulations so obtained have a form similar to *Ivanov’s* and *Tikhonov’s* methods for regularization [36–38], respectively. However, in our context regularization is associated only to the choice of a kernel (or equivalently, of

a RKHS). Instead, in Ivanov's and Tikhonov's methods, there is also a tunable parameter, which controls, respectively, the radius of a ball, or the weight given to a stabilizer functional. We show that the same linear combination of kernel functions is a solution for both cases (i) and (ii), where the coefficients of the combination can be obtained by solving a symmetric eigenvalue problem.

Then we consider the approximating properties of certain suboptimal solutions to KPCA. We investigate the optimization of the functional that models KPCA, over sets containing for a fixed integer  $n$  only linear combinations of all  $n$ -tuples of kernel functions. As our main interest is in a reduced computational burden of algorithms searching for suboptimal solutions to KPCA, we focus on the case  $n \ll m$ , where  $m$  is the size of the sample of data. We estimate the quality of suboptimal solutions for KPCA independently of the choice of a particular algorithm used to obtain suboptimal solutions. More specifically, we derive upper bounds on the error in approximating the optimal solution, achievable without restrictions on the number of kernel functions, by suboptimal solutions to KPCA having the form of linear combinations of  $n$ -tuples of kernel functions. The upper bounds depend on the number  $n$  of kernel functions and the largest eigenvalue of the Gram matrix of the kernel with respect to the sample of data. Although the least upper bounds cannot be evaluated without finding the largest eigenvalue of the Gram matrix, its estimate from below can be used. Moreover, the interest of the bounds lies in their generality and in the form of their dependence on such an eigenvalue, whatever its value may be.

The paper is organized as follows. Section 2 introduces kernel PCA as an optimization problem in a feature space for which a kernel can be used (Problem KPCA in the feature space), discusses the role played by the choice of a kernel, and gives the explicit form of the solution. Section 3 restates KPCA in the RKHS setting (following case (i) discussed above) and describes tools used in Sect. 4 to obtain suboptimal solutions to this problem by optimization over hypothesis sets containing, for a positive integer  $n$ , linear combinations of  $n$ -tuples of kernel functions. Section 4 estimates the rates of approximation by suboptimal solutions to KPCA with increasing number  $n$  of kernel functions. In Sect. 5, following the approach used in [35], we consider another formulation of KPCA (following case (ii) discussed above) and we study its properties. Section 6 is a brief discussion. To make the paper self-contained, three appendices are included. Appendix 1 summarizes basic concepts of classical PCA. For completeness purposes, in Appendix 2 we report results regarding PCA in a generic feature space, which are independent of the existence of a kernel associate with the mapping of the data to the feature space. Finally, Appendix 3 shortly reviews RKHSs and the properties of some kernels commonly used in applications.

## 2 KPCA in the feature space and in the RKHS setting

By a normed linear space  $(X, \|\cdot\|)$  we mean a real normed linear space; we merely write  $X$  when the norm is clear from the context. When  $X$  is a Hilbert space, we denote by  $\langle \cdot, \cdot \rangle$  the inner product inducing the norm  $\|\cdot\|$ . For the sake of clarity, sometimes we specify the space  $X$  in the notations of the inner product and of the

norm, i.e., we write  $\langle \cdot, \cdot \rangle_X$  and  $\| \cdot \|_X$ , instead of simply  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \|$ , respectively. The closed ball of radius  $r$  centered at  $h \in X$  is denoted by  $B_r(h, \| \cdot \|) = \{f \in X : \|f - h\| \leq r\}$ . The sphere of radius  $r$  centered at  $h \in X$  is denoted by  $S_r(h, \| \cdot \|) = \{f \in X : \|f - h\| = r\}$ . We write shortly  $B_r(\| \cdot \|) = B_r(0, \| \cdot \|)$ ,  $S_r(\| \cdot \|) = S_r(0, \| \cdot \|)$ , and merely  $B_r(h) = B_r(h, \| \cdot \|)$ ,  $B_r = B_r(0)$ ,  $S_r(h) = S_r(h, \| \cdot \|)$ , and  $S_r = S_r(0)$ , when it is clear which norm is used.

By  $\mathbb{R}$  and  $\mathbb{N}_+$  we denote the sets of real numbers and positive integers, respectively. For a positive integer  $d$ ,  $\| \cdot \|_2$  denotes the  $l_2$ -norm on  $\mathbb{R}^d$ , induced by the standard inner product  $\langle \cdot, \cdot \rangle_2$ , and  $\| \cdot \|_1$  denotes the  $l_1$ -norm on  $\mathbb{R}^d$ . Sequences of elements are denoted merely by  $\{s_n\}$ , where  $n \in \mathbb{N}_+$  and each  $s_n$  can be a number, an element of a linear space, or a set. Vectors are considered as column vectors. Transposition of vectors and matrices is denoted by the superscript “ $T$ ”. We use the symbol  $\delta_{i,j}$  for the Kronecker’s delta.

Let a sample of  $m$  data be represented by the vectors  $x_1, \dots, x_m \in \Omega \subseteq \mathbb{R}^d$ . For simplicity, we consider the case of *centered data*, i.e., data such that  $\sum_{i=1}^m x_i = 0$ ; however, the analysis can be extended to uncentered data (see Sect. 6). The basic concepts of PCA are summarized in Appendix 1; a comprehensive reference book is [14]. PCA with  $p$  components can be formulated as the following constrained nonlinear programming problem:

$$\begin{cases} \min_{v_1, \dots, v_p \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^m \left\| x_i - \sum_{j=1}^p \langle x_i, v_j \rangle v_j \right\|_2^2 \\ \text{s.t. } \langle v_i, v_j \rangle = \delta_{i,j}, \quad i, j = 1, \dots, p. \end{cases} \quad (1)$$

One can formulate PCA in another (possibly infinite-dimensional) inner-product space, the so-called *feature space*  $F$ . This is achieved by mapping the data to  $F$  through a (possibly nonlinear) mapping  $\varphi$  (for more details, see Appendix 2 and [29, Chap. 14]). Mathematical modeling of this form of PCA exploits *a-priori information* about the behavior of potential solutions. This can be interpreted as a form of *regularization*, which in this case corresponds to the choice of a particular nonlinear mapping. A large class of nonlinear mappings can be studied by applying the kernel trick [29, p. 34], i.e., by assuming that there exists a symmetric positive semi-definite function<sup>1</sup>  $K : \Omega \times \Omega \rightarrow \mathbb{R}$ , called *kernel*, such that  $K(x, y) = \langle \varphi(x), \varphi(y) \rangle_F$ . In order to distinguish between a generic feature space  $F$  and a feature space for which there is a kernel associate with the mapping  $\varphi$ , we denote the latter by  $F_K$ . When  $F_K$  is used as a feature space, the corresponding PCA technique is called *Kernel PCA* or *KPCA* [29, Chap. 14]. We assume for simplicity that the data are centered in the feature space, i.e. that  $\sum_{i=1}^m \varphi(x_i) = 0$ . With this assumption, KPCA is formulated

<sup>1</sup>A function  $K : \Omega \times \Omega \rightarrow \mathbb{R}$  is *positive semi-definite* on  $\Omega$  if, for all positive integers  $l$ , all  $(a_1, \dots, a_l) \in \mathbb{R}^l$ , and all  $(x_1, \dots, x_l) \in \Omega^l$ ,  $\sum_{i,j=1}^l a_i a_j K(x_i, x_j) \geq 0$ .

as the following optimization problem:

$$\begin{cases} \min_{f_1, \dots, f_p \in F_K} \frac{1}{m} \sum_{i=1}^m \left\| \varphi(x_i) - \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_{F_K} f_j \right\|_{F_K}^2 \\ \text{s.t. } \langle f_i, f_j \rangle_{F_K} = \delta_{i,j}, \quad i, j = 1, \dots, p. \end{cases} \quad (2)$$

According to the properties of PCA in the feature space, the formulation (2) is equivalent to the following one:

$$\begin{cases} \max_{f_1, \dots, f_p \in F_K} \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_{F_K}^2 \\ \text{s.t. } \langle f_i, f_j \rangle_{F_K} = \delta_{i,j}, \quad i, j = 1, \dots, p. \end{cases} \quad (3)$$

Hence, in the feature space the  $p$  principal components can be obtained as follows.

- First principal component ( $p = 1$ ):  $\max_{f \in F_K} \frac{1}{m} \sum_{i=1}^m \langle f, \varphi(x_i) \rangle_{F_K}^2$ , under the constraint  $f \in S_1(\|\cdot\|_{F_K})$ .
- Other  $p - 1$  principal components: orthogonal to the preceding one(s) and obtained by solving  $p - 1$  problems analogous to the one for  $p = 1$ .

In the following, with a little abuse of terminology, we refer to problem (3) with  $p = 1$  as “Problem KPCA in the feature space”.

The constraint  $\|f\|_{F_K} = 1$  can be replaced by  $\|f\|_{F_K} \leq 1$  without affecting the solution.

*Problem KPCA (in the feature space)*

$$\begin{cases} \max_{f \in F_K} \frac{1}{m} \sum_{i=1}^m \langle f, \varphi(x_i) \rangle_{F_K}^2 \\ \text{s.t. } f \in B_1(\|\cdot\|_{F_K}). \end{cases} \quad (4)$$

Similar problems are associate with the determination of the other  $p - 1$  principal components, so from now on we focus on (4).

The role played by the kernel in computing inner products in the feature space is illustrated by the following theorem (which, for simplicity, is stated only for  $p = 1$  but can be generalized to the case  $p > 1$  [29, Chap. 14]). For a kernel  $K$ , a positive integer  $m$ , and a sample  $\mathbf{x} = (x_1, \dots, x_m) \in \Omega^m$  of data, we denote by  $\mathcal{K}[\mathbf{x}]$  the  $m \times m$  positive semi-definite *Gram matrix of the kernel  $K$  with respect to the data vector  $\mathbf{x}$* , whose  $(i, j)$ -th entry is given by

$$\mathcal{K}[\mathbf{x}]_{ij} = K(x_i, x_j).$$

**Theorem 2.1** (See, e.g., [29, Chap. 14]) *Let  $d$  and  $m$  be positive integers,  $\Omega \subseteq \mathbb{R}^d$ ,  $K : \Omega \times \Omega \rightarrow \mathbb{R}$  a kernel, and  $\mathbf{x} = (x_1, \dots, x_m) \in \Omega^m$  such that  $\sum_{i=1}^m \varphi(x_i) = 0$ .*

Then the solution to (4) is given by

$$f^o = \sum_{i=1}^m c_i \varphi(x_i), \quad (5)$$

where  $c = (c_1, \dots, c_m)$  solves the symmetric eigenvalue problem

$$m\lambda_{\max} c = \mathcal{K}[\mathbf{x}]c \quad (6)$$

for the maximum eigenvalue  $m\lambda_{\max}$  of  $\mathcal{K}[\mathbf{x}]$ .<sup>2</sup>

As an immediate consequence of Theorem 2.1 we have the following:

**Theorem 2.2** (Representer Theorem of KPCA; see, e.g., [29, Chap. 14]) *Under the same hypotheses and with the same notations of Theorem 2.1, the projection on  $f^o$  of the image  $\varphi(x)$  in the feature space of  $x \in \Omega$  is given by*

$$\langle f^o, \varphi(x) \rangle_{F_K} = \sum_{i=1}^m c_i K(x, x_i). \quad (7)$$

By Theorem 2.2, in order to compute the projections on  $f^o$  of the images of the data in the feature space, the explicit expression of the mapping  $\varphi$  is not needed: the knowledge of the kernel function is sufficient.

We call Theorem 2.2 “Representer Theorem,” by analogy with the case in which one has to learn, on the basis of a sample of  $m$  input-output data, an unknown mapping generating such data, by minimizing the so-called *regularized empirical error functional* with a convex loss function (see, e.g., [7, Chap. II, Sect. 6] and [21, Sect. 3]). In such a case, the theorem that gives the solution as a linear combination of kernel functions centered at the input data points is called the *Representer Theorem of learning theory* (see, e.g., [7, p. 42]); it also states that the coefficients of the linear combination are determined by the solution of a well-posed linear system of equations. Similarly, Theorem 2.2 gives the solution<sup>3</sup> to KPCA expressed as the linear combination (7) of the functions  $K(x, x_i)$  centered at the data points and the coefficients are determined by the solution to the eigenvalue problem (6). So we have a strong analogy between Theorem 2.2 and the Representer Theorem of learning theory. In general, “representer theorems” provide an explicit expression for the solution of a problem, in terms of a combination of a certain number of “basis” elements.

Theorems 2.1 and 2.2 have also another interesting interpretation. By Theorem 2.1, the solution to problem (4) does not change if one restricts the maximization to  $F_0 \cap B_1(\|\cdot\|_{F_K})$ , where  $F_0 = \{\sum_{i=1}^l a_i \varphi(\tilde{x}_i), l \in \mathbb{N}_+, \tilde{x}_i \in \Omega\}$ . By

<sup>2</sup>Refer to Appendix 1 for the reason why we use the notation  $m\lambda_{\max}$  for the maximum eigenvalue of the Gram matrix.

<sup>3</sup>To be more precise, this is an *implicit* solution, since  $f^o$  is not computed through the kernel, but one has an *explicit* expression for the subset of projections on  $f^o$  which one usually needs for subsequent data processing.

Theorem 2.2, the projections of the elements of  $F_0$  on the image in the feature space of  $x \in \Omega$  belong to  $H_0(\Omega) = \{\sum_{i=1}^l a_i K(x, \tilde{x}_i), l \in \mathbb{N}_+, \tilde{x}_i \in \Omega\}$ , so they can be computed through the kernel function, without explicitly considering the mapping  $\varphi$ . As done in [7, p. 35], let us define in  $H_0(\Omega)$  an inner product such that, if  $f(x) = \sum_{i=1}^s a_i K(x, \tilde{x}_i)$  and  $g(x) = \sum_{j=1}^r b_j K(x, \tilde{y}_j)$ , then

$$\langle f, g \rangle_{H_0(\Omega)} = \sum_{i=1}^s \sum_{j=1}^r a_i b_j K(\tilde{x}_i, \tilde{y}_j).$$

It can be seen that there is an Euclidean space isomorphism [16, p. 153] between  $F_0$  and  $H_0(\Omega)$  [6, p. 46]. This can be extended to an isomorphism between Hilbert spaces after completion of  $F_0$  and  $H_0(\Omega)$  in their respective norms. Let us denote by  $\mathcal{H}_K(\Omega)$  the Hilbert space obtained by completion of  $H_0(\Omega)$  in the norm  $\|\cdot\|_{H_0(\Omega)} = \|\cdot\|_{\mathcal{H}_K(\Omega)}$  (which for simplicity of notation in the following we simply denote by  $\|\cdot\|_K$ ). It can be seen ([2, p. 344] and [7, p. 35]) that  $\mathcal{H}_K(\Omega)$  is a Reproducing Kernel Hilbert Space (RKHS). Therefore, RKHSs represent an equivalent setting for Problem KPCA (having restricted the maximization in (4) to vectors belonging to the completion of  $F_0$ ). So we consider the following problem, where we let  $K_{x_i}(\cdot) := K(\cdot, x_i)$ :

*Problem KPCA (in the RKHS setting)*

$$\begin{cases} \max_{f \in H_K(\Omega)} \frac{1}{m} \sum_{i=1}^m \langle f, K_{x_i} \rangle_K^2 \\ \text{s.t. } f \in B_1(\|\cdot\|_K). \end{cases} \tag{8}$$

The reason why we choose Problem (8) as a starting point for our analysis in the next sections is that this formulation is more appropriate in order to use results from functional optimization theory, particularly those from [20]. For brevity, we refer to Problem (8) simply as “Problem KPCA.”

For regularization properties associated to specific choices of RKHSs, we refer the reader to Appendix 3, where it is shown that norms in RKHSs often play the role of measures of various types of oscillations of functions. Therefore, they can be used to model the requirement that admissible solutions to Problem KPCA should have to be “sufficiently smooth,” where the kind of smoothness is related to the choice of the kernel and reflects the a-priori knowledge about the problem at hand.

### 3 Tools for investigating approximate solutions to KPCA

In (7), the kernel functions  $K(\cdot, x_1), \dots, K(\cdot, x_m)$ , centered at the data  $x_1, \dots, x_m$ , play the role of so-called *representers*. If we let

$$G_{K,\mathbf{x}} = \{K(\cdot, x_1), \dots, K(\cdot, x_m)\} = \{K_{x_1}, \dots, K_{x_m}\},$$

then Theorem 2.2 implies that the projection on the solution  $f^0$  to Problem KPCA of the image  $\varphi(x)$  in the feature space of  $x \in \Omega$ , belongs to the subspace  $\text{span } G_{K,\mathbf{x}}$  of  $\mathcal{H}_K(\Omega)$  spanned by the kernel functions centered at the data points. Note that such

a subspace is the natural “ambient space” of Problem KPCA, which can be solved by a linear combination of  $K(x, x_1), \dots, K(x, x_m)$ , hence with as many elements as the number  $m$  of data. For example, for the Gaussian kernel the solution has the form of an input-output function of a Gaussian radial-basis-function network with  $m$  computational units [13]. However, this solution is practically useful only if  $m$  is not too large. When this is not the case, one has to look for suboptimal solutions to Problem PCA.

In searching for suboptimal solutions, we would like to meet two requirements: (1) they should be “parsimonious,” in the sense that they should be expressible as combinations of a small number  $n$  of representers, hopefully  $n \ll m$  (we refer to this requirement as *sparseness*); (2) they should be *sufficiently close* to the optimal solution and, when  $n$  increases, they should approximate it better and better (*approximation accuracy*). Let

$$G_K = \{K_x : x \in \Omega\}.$$

Since the RKHS  $\mathcal{H}_K(\Omega)$  is the completion of  $\text{span } G_K$  (which contains all linear combinations of the representers), we are motivated to search for approximate solutions to Problem KPCA by performing maximization over linear combinations of at most  $n$  representers, where the interest is in  $n \ll m$ . Such suboptimal solutions have the form

$$f_n(x) = \sum_{i=1}^n c_i K(x, y_i), \quad y_i \in \Omega, \quad (9)$$

where  $n < m$  and  $y_1, \dots, y_n$  are not necessarily data points. For a subset of the data points, representers correspond to the set  $G_{K,x}$  and the suboptimal solutions have the form

$$f_n(x) = \sum_{i=1}^n c_i K(x, x_i), \quad (10)$$

where  $x_1, \dots, x_m$  are the data points and  $n < m$ .

To address the trade off between approximation accuracy and sparseness, we investigate conditions guaranteeing that a good approximation can be obtained by using a small number  $n$  of representers. We first describe some mathematical tools used in the following. Suboptimal solutions to Problem KPCA expressed as linear combinations of  $n$  kernel functions can be studied in terms of optimization over nested (with respect to  $n$ ) families of subsets of RKHSs formed by linear combinations of all  $n$ -tuples of kernel functions chosen from the sets  $G_K$  or  $G_{K,x}$ . For a subset  $G$  of a linear space, we denote by  $\text{span}_n G$  the set of linear combinations of all  $n$ -tuples of elements of  $G$ , i.e.,

$$\text{span}_n G = \left\{ \sum_{i=1}^n w_i g_i : w_i \in \mathbb{R}, g_i \in G \right\}.$$

Sets of the form  $\text{span}_n G$  are called *variable-basis approximation schemes* and have been studied in mathematical theory of neural networks; see, e.g., [3, 15, 18, 19, 22].

In the following, we shall consider suboptimal solutions from

$$\text{span}_n G_{K,\mathbf{x}} = \text{span}_n \{K_{x_1}, \dots, K_{x_m}\}$$

or from

$$\text{span}_n G_K = \text{span}_n \{K_x : x \in \Omega\}.$$

Let us consider the following two “approximate” versions of Problem KPCA.

*Problem KPCA<sub>n</sub> (first version)*

$$\begin{cases} \max_{f \in \mathcal{H}_K(\Omega)} \frac{1}{m} \sum_{i=1}^m \langle f, K_{x_i} \rangle_K^2 \\ \text{s.t. } f \in B_1(\|\cdot\|_K) \cap \text{span}_n G_{K,\mathbf{x}}. \end{cases} \tag{11}$$

*Problem KPCA<sub>n</sub> (second version)*

$$\begin{cases} \max_{f \in \mathcal{H}_K(\Omega)} \frac{1}{m} \sum_{i=1}^m \langle f, K_{x_i} \rangle_K^2 \\ \text{s.t. } f \in B_1(\|\cdot\|_K) \cap \text{span}_n G_K. \end{cases} \tag{12}$$

To compare the optimal solution to Problem KPCA, given by Theorem 2.2, with suboptimal ones obtained by solving (both forms of) Problem KPCA<sub>n</sub> for each  $n < m$ , we shall employ a reformulation of Maurey–Jones–Barron’s theorem [3, 15, 24] in terms of a norm called *G-variation*. Such a norm, denoted by  $\|\cdot\|_G$ , was defined in [17] for a subset  $G$  of a normed linear space  $(X, \|\cdot\|)$  as the Minkowski functional<sup>4</sup> of the closure of the convex hull of the set  $G \cup -G$ . So for every  $f \in X$  we have  $\|f\|_G = \inf\{c > 0 : f/c \in \text{cl conv}(G \cup -G)\}$ . For properties of *G-variation*, see [17–19, 22].

Maurey–Jones–Barron’s theorem stated in terms of *G-variation* [18, 19, 22] gives for a Hilbert space  $(X, \|\cdot\|)$ , its bounded subset  $G$  with  $s_G = \sup_{g \in G} \|g\|$ , and every  $f \in X$ , the following upper bound on the rate of approximation of  $f$  by elements of  $\text{span}_n G$ :  $\|f - \text{span}_n G\| \leq \sqrt{\frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}}$ .

Maurey–Jones–Barron’s upper bound is stated in terms of two norms: the norm  $\|\cdot\|$  of the Hilbert space  $X$  and the *G-variation* norm  $\|\cdot\|_G$ . In the two versions of Problem KPCA<sub>n</sub>, the Hilbert space is the RKHS  $\mathcal{H}_K(\Omega)$  with the  $\|\cdot\|_K$ -norm and the variation norm is  $G_{K,\mathbf{x}}$ -variation or  $G_K$ -variation, so we need to evaluate or estimate  $\|f^\circ\|_{G_{K,\mathbf{x}}}$  or  $\|f^\circ\|_{G_K}$ . As  $f^\circ$  is the solution of Problem KPCA,  $\|f^\circ\|_K = 1$ . The next proposition gives upper bounds on the other two norms.

**Proposition 3.1** *Let  $d$  and  $m$  be positive integers,  $\Omega \subseteq \mathbb{R}^d$ ,  $K : \Omega \times \Omega \rightarrow \mathbb{R}$  a kernel,  $\mathbf{x} = (x_1, \dots, x_m) \in \Omega^m$  such that  $\sum_{i=1}^m \varphi(x_i) = 0$ ,  $f^\circ$  the solution to Problem KPCA,*

<sup>4</sup>The *Minkowski functional* of a subset  $M$  of a linear space  $X$ , denoted by  $p_M$ , is defined for every  $f \in X$  as  $p_M(f) = \inf\{\lambda \in \mathbb{R}_+ : f/\lambda \in M\}$ . If  $M$  is a subset of a normed linear space  $(X, \|\cdot\|)$ , we denote by  $\text{cl}M$  its *closure* with respect to the topology generated by  $\|\cdot\|$ , i.e.,  $\text{cl}M = \{f \in X : (\forall \varepsilon > 0) (\exists g \in M) \|f - g\| < \varepsilon\}$ .

and  $m\lambda_{\max}$  the maximum eigenvalue of the matrix  $\mathcal{K}[\mathbf{x}]$ . Then

$$\|f^0\|_{G_K} \leq \|f^0\|_{G_{K,\mathbf{x}}} \leq \sqrt{\frac{1}{\lambda_{\max}}}.$$

*Proof* The first inequality follows directly from the definition of  $G$ -variation, since  $G_{K,\mathbf{x}} \subseteq G_K$ . The normalization condition  $\|f^0\|_K = 1$  is translated into the normalization condition  $\|c\|_2 = \frac{1}{\sqrt{m\lambda_{\max}}}$  [29, p. 430], where  $\|\cdot\|_2$  denotes the Euclidean norm on  $\mathbb{R}^m$ . So by Theorem 2.1 and Hölder's inequality [16, p. 41], we get

$$\|f^0\|_{G_{K,\mathbf{x}}} \leq \sum_{i=1}^m |c_i| = \|c\|_1 \leq \sqrt{m}\|c\|_2 = \sqrt{\frac{1}{\lambda_{\max}}}. \quad \square$$

In the next section, we shall estimate the accuracy of suboptimal solutions to Problem KPCA by combining Maurey–Jones–Barron's upper bound with properties of the functional  $\frac{1}{m} \sum_{i=1}^m \langle f, K_{x_i} \rangle_K^2$  and properties of the ball  $B_1(\|\cdot\|_K)$  of the RKHS  $\mathcal{H}_K(\Omega)$ , over which it is maximized.

#### 4 Estimates of the accuracy of suboptimal solutions to Problem KPCA

Let us denote by  $\Psi_{m,K}$  the functional to be maximized in Problems KPCA and both versions of KPCA<sub>*n*</sub>, i.e.,

$$\Psi_{m,K}(f) = \frac{1}{m} \sum_{i=1}^m \langle f, K_{x_i} \rangle_K^2.$$

To estimate the accuracy of suboptimal solutions to Problem KPCA obtained by solving Problem KPCA<sub>*n*</sub>, in the next proposition we state continuity properties of  $\Psi_{m,K}$ . Recall that a functional  $\Upsilon : (X, \|\cdot\|) \rightarrow \mathbb{R}$  is *continuous* at  $f \in X$  if for any  $\varepsilon > 0$ , there exists  $\eta > 0$  such that  $\|f - g\| < \eta$  implies  $|\Upsilon(f) - \Upsilon(g)| < \varepsilon$ . A *modulus of continuity* of  $\Upsilon$  at  $f$  is a function  $\alpha_f : [0, +\infty) \rightarrow [0, +\infty)$  defined as  $\alpha_f(t) = \sup_{g \in X} \{|\Upsilon(f) - \Upsilon(g)| : \|f - g\| \leq t\}$ .

**Proposition 4.1** *Let  $d$  and  $m$  be positive integers,  $\Omega \subseteq \mathbb{R}^d$ ,  $K : \Omega \times \Omega \rightarrow \mathbb{R}$  a kernel,  $\mathbf{x} = (x_1, \dots, x_m) \in \Omega^m$  a sample of data, and  $s_{K,\mathbf{x}} = \max_{1 \leq i \leq m} \sqrt{K(x_i, x_i)}$ . Then, at every  $f \in \mathcal{H}_K(\Omega)$  the functional  $\Psi_{m,K}$  is continuous with a modulus of continuity bounded from above by the function  $\alpha_f(t) = a_2 t^2 + a_1 t$ , where  $a_2 = s_{K,\mathbf{x}}^2$  and  $a_1 = 2s_{K,\mathbf{x}}^2 \|f\|_K$ .*

*Proof* The functional  $\Psi_{m,K}$  is a weighted sum of the  $m$  functionals  $\langle \cdot, K_{x_i} \rangle_K^2$ ,  $i = 1, \dots, m$ . Each of them is continuous as it results from the composition of two continuous mappings (the square function and the inner product in the Hilbert space  $\mathcal{H}_K(\Omega)$ , whose continuity is implied by the Cauchy–Schwartz's inequality). Hence  $\Psi_{m,K}$  is continuous, too.

An upper bound on the modulus of continuity  $\alpha_f$  of  $\Psi_{m,K}$  at  $f \in \mathcal{H}_K(\Omega)$  can be obtained as follows. By the definition of  $\Psi_{m,K}$  and the Cauchy–Schwartz’s inequality

$$\begin{aligned}
 |\Psi_{m,K}(f) - \Psi_{m,K}(g)| &= \frac{1}{m} \left| \sum_{i=1}^m \langle f, K_{x_i} \rangle_K^2 - \sum_{i=1}^m \langle g, K_{x_i} \rangle_K^2 \right| \\
 &\leq \frac{1}{m} \sum_{i=1}^m |(\langle f, K_{x_i} \rangle_K + \langle g, K_{x_i} \rangle_K)(\langle f, K_{x_i} \rangle_K - \langle g, K_{x_i} \rangle_K)| \\
 &\leq \max_{i=1, \dots, m} |(\langle f, K_{x_i} \rangle_K + \langle g, K_{x_i} \rangle_K)(\langle f, K_{x_i} \rangle_K - \langle g, K_{x_i} \rangle_K)| \\
 &= \max_{i=1, \dots, m} |\langle f + g, K_{x_i} \rangle_K \langle f - g, K_{x_i} \rangle_K| \\
 &\leq \|f + g\|_K \|f - g\|_K \max_{i=1, \dots, m} \|K_{x_i}\|_K^2. \tag{13}
 \end{aligned}$$

From the definition of the inner product in  $\mathcal{H}_K(\Omega)$ , we have  $\|K_{x_i}\|_K^2 = K(x_i, x_i) \leq s_{K,x}^2$ . Hence, (13) implies

$$|\Psi_{m,K}(f) - \Psi_{m,K}(g)| \leq \|f + g\|_K \|f - g\|_K s_{K,x}^2. \tag{14}$$

Now, let  $\|f - g\|_K \leq t$ . As  $\|f + g\|_K \leq \|f\|_K + \|g\|_K = \|f\|_K + \|f + g - f\|_K \leq \|f\|_K + \|f - g\|_K + \|f\|_K = 2\|f\|_K + \|f - g\|_K$ , one has  $\|f + g\|_K \leq 2\|f\|_K + t$ . Hence, (14) gives

$$|\Psi_{m,K}(f) - \Psi_{m,K}(g)| \leq s_{K,x}^2 t (2\|f\|_K + t) = s_{K,x}^2 t^2 + 2s_{K,x}^2 \|f\|_K t.$$

Thus, an upper bound on the modulus of continuity of  $\Psi_{m,K}$  at  $f$  is  $\alpha_f(t) = a_2 t^2 + a_1 t$ , where  $a_2 = s_{K,x}^2$  and  $a_1 = 2s_{K,x}^2 \|f\|_K$ .  $\square$

The next theorem (which is stated for the first version of Problem  $KPCA_n$ ) estimates rates of approximation in terms of the  $G_{K,x}$ -variation norm  $\|f^0\|_{G_{K,x}}$  of  $f^0$ .

**Theorem 4.2** *Let  $d, m$  and  $n$  be positive integers,  $\Omega \subseteq \mathbb{R}^d$ ,  $K : \Omega \times \Omega \rightarrow \mathbb{R}$  a kernel,  $\mathbf{x} = (x_1, \dots, x_m) \in \Omega^m$  a sample of data such that  $\sum_{i=1}^m \varphi(x_i) = 0$ ,  $G_{K,x} = \{K_{x_1}, \dots, K_{x_m}\}$ ,  $s_{K,x} = \max_{1 \leq i \leq m} \sqrt{K(x_i, x_i)}$ ,  $f^0$  the solution of Problem  $KPCA$ ,  $m\lambda_{\max}$  be the maximum eigenvalue of the matrix  $\mathcal{K}[\mathbf{x}]$ ,  $b = 4(\frac{s_{K,x}^2}{\lambda_{\max}} - 1)$ ,  $c_2 = s_{K,x}^2$ , and  $c_1 = 2s_{K,x}^2$ . The following estimates hold:*

(i) if  $0 < n < m$ , then

$$\Psi_{m,K}(f^0) - \sup_{f \in B_1(\|\cdot\|_K) \cap \text{span}_n G_{K,x}} \Psi_{m,K}(f) \leq c_2 \frac{b}{n} + c_1 \sqrt{\frac{b}{n}};$$

(ii) if  $n \geq m$ , then  $\Psi_{m,K}(f^0) = \sup_{f \in B_1(\|\cdot\|_K) \cap \text{span}_n G_{K,x}} \Psi_{m,K}(f)$ .

*Proof* We apply [20, Theorem 4.2(i)]; we report it here for the readers' convenience. Given (1) two subsets  $M$  and  $G$  of a Hilbert space  $(X, \|\cdot\|)$  such that  $M$  is closed, convex, and  $0 \in \text{int } M$  (the topological interior of  $M$ ),  $G$  bounded, and  $s_G = \sup_{f \in G} \|f\|$ , and (2) a functional  $\Upsilon : X \rightarrow \mathbb{R}$  such that  $\Upsilon(f^0) = \inf_{f \in M} \Upsilon(f)$  and  $\Upsilon$  is continuous at  $f^0$  with a modulus of continuity  $\alpha_{f^0}$ , the following estimate holds:

$$\inf_{f \in M \cap \text{span}_n G} \Upsilon(f) - \Upsilon(f^0) \leq \alpha_{f^0} \left( (1 + L\|f^0\|) \sqrt{\frac{(s_G \|f^0\|_G)^2 - \|f^0\|^2}{n}} \right),$$

where  $L$  is the Lipschitz constant of the Minkowski functional of  $M$  (by the properties of  $G$ -variation [20, Sect. 3], one has  $(s_G \|f^0\|_G)^2 - \|f^0\|^2 \geq 0$ ).

To exploit this result, we proceed as follows. First we note that the set  $B_1(\|\cdot\|_K)$  is closed and convex and that by definition  $0$  belongs to its topological interior. On the other hand, by the properties of the Minkowski functional of convex sets containing zero (see, e.g., [20, Proposition 2.2(vi)]) and by the definition of  $B_1(\|\cdot\|_K)$ , the Lipschitz constant of the Minkowski functional of  $B_1(\|\cdot\|_K)$  is equal to 1. Moreover,  $\|f^0\|_K = 1$  and by Proposition 3.1 we have  $\|f^0\|_{G_{K,x}} \leq \sqrt{\frac{1}{\lambda_{\max}}}$ .

So, by applying the above-reported result from [20, Theorem 4.2(i)] with  $\Upsilon = -\Psi_{m,K}$ , as  $\inf(-\Upsilon) = -\sup \Upsilon$  we get

$$\begin{aligned} \Psi_{m,K}(f^0) - \sup_{f \in B_1(\|\cdot\|_K) \cap \text{span}_n G_{K,x}} \Psi_{m,K}(f) &\leq \alpha_{f^0} \left( \frac{1}{\sqrt{n}} \sqrt{4 \left( \frac{s_{K,x}^2}{\lambda_{\max}} - 1 \right)} \right) \\ &= \alpha_{f^0} \left( \sqrt{\frac{b}{n}} \right). \end{aligned} \tag{15}$$

By Proposition 4.1, an upper bound on the modulus of continuity  $\alpha_{f^0}$  of the functional  $\Psi_{m,K}$  at  $f^0$  is given by the function  $c_2 t^2 + c_1 t$ , where  $c_2 = s_{K,x}^2$  and  $c_1 = 2s_{K,x}^2 \|f^0\|_K = 2s_{K,x}^2$ , since  $\|f^0\|_K = 1$ . Together with (15), this proves (i).

The estimate (ii) follows from the Representer Theorem (Theorem 2.2), since for  $n \geq m$  the optimal solution of Problem KPCA, being an element of  $\text{span}_m G_{K,x}$ , is also an element of  $\text{span}_n G_{K,x}$ , therefore it is feasible for the first version of Problem  $\text{KPCA}_n$ . □

Since, for each  $n$ ,  $\text{span}_n G_{K,x} \subseteq \text{span}_n G_K$ , it easily follows that the estimates given in Theorem 4.2 hold also for the second version of Problem  $\text{KPCA}_n$ .

Note that the bounds from Theorem 4.2 are interesting for  $n < m$ , particularly when  $n \ll m$  (we have included the case  $n \geq m$  for completeness purposes), since here we are interested in approximation properties of sparse suboptimal solutions of Problem KPCA. In general, one is able to solve Problem  $\text{KPCA}_n$  only approximately. However the bounds hold, with minor modifications, also for its  $\varepsilon_n$ -near maximum

points.<sup>5</sup> Moreover, they are independent of the capability of finding the optimal subset of  $n < m$  kernel functions, which is a combinatorial problem.

To be exactly evaluated, the upper bounds from Theorem 4.2 require the determination of the largest eigenvalue of the Gram matrix. However, the interest of the bounds consists in their dependence on the largest eigenvalue, whatever its value is. Moreover, an estimate from below can be used for  $\lambda_{\max}$ . It has also to be remarked that, for particular choices of kernels, commonly used in the applications (see Appendix 3),  $s_K = \sup_{x \in \Omega} \sqrt{K(x, x)}$  is finite, so it can be used as an upper bound for  $s_{K, \mathbf{x}}$ , independently of the sample  $\mathbf{x}$  and the number  $m$  of data. With this substitution, the only quantity that depends on the particular sample is  $\lambda_{\max}$ , i.e., the maximum eigenvalue of the covariance matrix (or covariance operator, in the infinite-dimensional case, as explained in Appendix 2) of the data. When these are obtained by i.i.d. sampling the probability distribution associated to a random variable, one could expect that, as  $m$  increases, this estimate approaches the maximum eigenvalue of the covariance matrix of the random variable. A statistical analysis of this type is performed, in the non-asymptotical case, in [34] and may be combined with Theorem 4.2 (which refers to the approximation error) to give some insights into the effectiveness of sparse KPCA techniques.

### 5 On an alternative formulation of Problem KPCA

In [35], PCA and its kernel version were formalized in the style of Least Square Support Vector Machine (LS-SVM) classifiers. The starting point consists in considering PCA as a one-class modeling problem with a target value equal to zero, around which the variance of the data is maximized. Then, given a sample of  $m$  centered data represented by the vectors  $x_1 \dots, x_m \in \Omega \subseteq \mathbb{R}^d$  and  $\gamma > 0$ , in [35] the following constrained optimization problem is considered:

$$\max_{v \in \Omega} \left( \frac{1}{m} \sum_{i=1}^m \langle x_i, v \rangle^2 - \gamma \|v\|_2^2 \right). \tag{16}$$

The formulation (16) is interpreted as in [35] as a *primal* optimization problem. Then, in [35] a *dual problem* is derived that is the same symmetric eigenvalue problem associate with Problem PCA. Finally, after mapping the data from the input space to a feature space and applying the kernel trick, a primal formulation equivalent to Problem KPCA (in the sense that it gives the same solution) was obtained in [35]. As noted in [35, Introduction], the sparseness of such an interpretation of kernel PCA may be obtained by combining the analysis with a reduced set approach [31] or a Nyström approximation [9, 40].

Here we start observing that the kernel version of the primal problem formulated in [35] has a form similar to Tikhonov’s regularization. Given a normed linear

---

<sup>5</sup>For  $M \subseteq (X, \|\cdot\|)$ , a functional  $\Upsilon : M \rightarrow \mathbb{R}$ , and  $\varepsilon > 0$ ,  $f_\varepsilon$  is an  $\varepsilon$ -near maximum point of the problem  $\sup_{f \in M} \Upsilon(f)$  if one has  $\Upsilon(f_\varepsilon) > \sup_{f \in M} \Upsilon(f) - \varepsilon$ .

space  $(X, \|\cdot\|)$ , its subset  $M$ , a functional  $\Upsilon : M \rightarrow \mathbb{R}$ , and an optimization problem  $\sup_{f \in M} \Upsilon(f)$ , *Tikhonov's regularization* [36, 37], introduced into learning theory by Poggio and Girosi [13, 25, 26], yields to maximization over the whole space  $X$  of the functional  $\Upsilon - \gamma \Xi$ , where  $\Xi$ , called *stabilizer*, is a functional that expresses requirements for the global behavior of the input–output mapping. The *regularization parameter*  $\gamma$  controls the trade-off between maximizing the original functional and penalizing an undesired behavior.

In order to clarify the role of the parameter  $\gamma$  in (16) and its generalization obtained by applying the kernel trick, let us consider here only the case of a finite-dimensional feature space (for infinite-dimensional feature spaces, one can use tools from optimization in infinite dimensional spaces [41, Chap. 43]). Hence, using the notation in Sect. 2, assume that  $F_K$  is finite dimensional and consider the following problem [35]:

*Problem KPCA' (in the feature space)*

$$\max_{f \in F_K} \left( \frac{1}{m} \sum_{i=1}^m \langle f, \varphi(x_i) \rangle_{F_K}^2 - \gamma \|f\|_{F_K}^2 \right). \tag{17}$$

As done in Sect. 2, with a little abuse of terminology we call “Problem KPCA'” the problem of finding just the first principal component. Similar problems are associated with the determination of the other  $p - 1$  principal components.

In [35], Problem KPCA' was written in the following form, as a constrained optimization problem:

$$\begin{cases} \max_{f \in F_K, \pi_K \in \mathbb{R}^m} \left( \frac{1}{m} \sum_{i=1}^m \pi_{K,i}^2 - \gamma \|f\|_{F_K}^2 \right) \\ \text{s.t. } \pi_{K,i} = \langle f, \varphi(x_i) \rangle_{F_K}, \quad i = 1, \dots, m. \end{cases}$$

By using first order necessary conditions associate with the Lagrangian

$$\mathcal{L}(f, \pi_K, c) = \frac{1}{m} \sum_{i=1}^m \pi_{K,i}^2 - \gamma \|f\|_{F_K}^2 + \sum_{i=1}^m 2\gamma c_i (\pi_{K,i} - \langle f, \varphi(x_i) \rangle_{F_K}),$$

where  $2\gamma c_i$  are the Lagrange multipliers,  $i = 1, \dots, m$ , in [35] the following result was obtained:

**Theorem 5.1** (Representer Theorem for Problem KPCA' [35]) *Let  $d$  and  $m$  be positive integers,  $\Omega \subseteq \mathbb{R}^d$ ,  $K : \Omega \times \Omega \rightarrow \mathbb{R}$  a kernel, and  $\mathbf{x} = (x_1, \dots, x_m) \in \Omega^m$  a sample of data. Then a necessary condition for Problem KPCA' to have a solution is*

$$f^o = \sum_{i=1}^m c_i \varphi(x_i), \tag{18}$$

where  $c = (c_1, \dots, c_m)$  satisfies the equation

$$c_i = \frac{1}{\gamma m} \sum_{j=1}^m \alpha_j (\varphi(x_j), \varphi(x_i))_K = \frac{1}{\gamma m} \sum_{j=1}^m c_j \mathcal{K}_{ji}[\mathbf{x}], \tag{19}$$

so it is an eigenvector of  $\mathcal{K}[\mathbf{x}]$ .

Next, in [35] it was assumed that in Problem KPCA' the parameter  $\gamma$  is equal to  $\lambda_{\max}$ , where  $m\lambda_{\max}$  is the largest eigenvalue of the Gram matrix  $\mathcal{K}[\mathbf{x}]$ . In the remaining of this section, we show that for this choice of  $\gamma$ , (18, 19) give a sufficient condition for the existence of a solution to Problem KPCA' and that for every other choice of  $\gamma$ , no "interesting" solution exists. Therefore we argue that  $\gamma$  in Problem KPCA' should not be considered as a regularization parameter (in the Tikhonov sense).

**Theorem 5.2** (Existence of a solution for Problem KPCA') *Under the same hypotheses of Theorem 5.1, if in addition  $\gamma = \lambda_{\max}$ , where  $m\lambda_{\max}$  is the largest eigenvalue of the Gram matrix  $\mathcal{K}[\mathbf{x}]$ , then Problem KPCA' has a solution given by (18) and (19). If  $0 < \gamma < \lambda_{\max}$ , then no solution exists. For  $\gamma > \lambda_{\max}$ , the solution is given by  $f^0 = 0$ .*

*Proof* For simplicity of notation, let us consider only the case  $F = \mathbb{R}^d$  in which  $\varphi$  is the identity map. So we replace  $f$  with the vector  $v \in \mathbb{R}^d$ ,  $\varphi(x_i)$  with  $x_i$ , and  $\|\cdot\|_K$  with  $\|\cdot\|_2$ . Hence, (17) becomes

$$\max_{v \in \mathbb{R}^d} \left( \frac{1}{m} \sum_{i=1}^m \langle v, x_i \rangle_2^2 - \gamma \|v\|_2^2 \right). \tag{20}$$

The general case follows by the kernel trick.

By introducing the  $d \times d$  sample covariance matrix  $C[\mathbf{x}] = \frac{1}{m} \sum_{i=1}^m x_i x_i^T$  with entries  $C[\mathbf{x}]_{jl} = \frac{1}{m} \sum_{i=1}^m x_{i,j} x_{i,l}$  (see also Appendix 1), the following equivalent version of (20) is obtained:

$$\max_{v \in \mathbb{R}^d} (v^T C[\mathbf{x}] v - \gamma v^T v) = \max_{v \in \mathbb{R}^d} v^T (C[\mathbf{x}] - \gamma I) v = \max_{v \in \mathbb{R}^d} v^T S v,$$

where  $S = (C[\mathbf{x}] - \gamma I)$  is a symmetric  $d \times d$  matrix.

The existence of a solution of (20) depends on the eigenspectrum of  $S$ , which depends on the value of the parameter  $\gamma$ . In particular, if at least one of the eigenvalues of  $S$  is positive, then the function to be maximized in (20) is unbounded from above and there is no solution. If all the eigenvalues of  $S$  are negative one has only the trivial solution  $v = 0$ . So the only meaningful case occurs when all eigenvalues of the matrix  $S$  are negative, except for at least one of them, which is zero. This corresponds to taking  $\gamma$  equal to the maximum eigenvalue  $\lambda_{\max}$  of the sample covariance matrix  $C[\mathbf{x}]$ . In such a case, the solution of (20) is any eigenvector  $v$  of  $S$  that corresponds to such a value of  $\gamma$ . □

## 6 Discussion

One first remark is about possible generalizations of the results presented in this paper. For the sake of simplicity we have considered the case of centered data, which in the feature space corresponds to  $\sum_{i=1}^m \varphi(x_i) = 0$ . However, at the expense of a more complex notation, our analysis can be extended to the uncentered case. The uncentered version of KPCA can be dealt with as in [30, Appendix A].

The estimate of the accuracy of suboptimal solutions obtained by Problem  $KPCA_n$  (see Theorem 4.2) is the summation of two terms, one proportional to  $n^{-1/2}$  and the other to  $n^{-1}$ . Evaluating the bounds from Theorem 4.2 requires to find the largest eigenvalue of the Gram matrix or an estimate of it from below. Indeed, here we are interested in the type of dependence of the bounds on the largest eigenvalue, independently of its value. For example, lower bounds on the largest eigenvalue of the Gram matrix can be obtained at low computational cost, if a few Rayleigh quotients [8, Chap. 9] are computed. The eigenspectrum of the Gram matrix in relationship with the generalization error in KPCA was investigated in [34].

We conclude with some numerical remarks. Both PCA and KPCA require the solution of an eigenvalue problem with a symmetric  $m \times m$  matrix (see (6)). As the matrix is positive definite, the eigenvalue decomposition reduces to singular value decomposition. However, this becomes computationally cumbersome as the number  $m$  of data increases [27]. Hence, resorting to suboptimal solutions that work on smaller  $n \times n$  matrices often becomes mandatory. In these cases, estimating their maximum accuracy for increasing values of  $n$ , as done in Theorem 4.2, play an important role.

Numerical methods that can be found in the literature for the solution of the symmetric eigenvalue problem (6) fall into one of the following two classes: (1) methods that aim at finding *all* eigenvalues and eigenvectors; (2) methods that aim at finding *some* eigenvalues and eigenvectors. The first ones typically require  $O(m^3)$  operations to reach convergence [27]. They usually reduce the  $m \times m$  matrix to simpler forms (e.g., the so-called *tridiagonal* form), which in general are sparse. For matrices in such forms, eigenvalues and eigenvectors can be determined through efficient iterative algorithms (like the QL algorithm [27]). The second class of methods is more appropriate if one is interested only in a few number of eigenvalues and  $m$  is large enough to make the first class of methods computationally too expensive. For this reason, the second class is more useful for KPCA.

An example of the second class of algorithms is the Power Method [8, Chap. 9], which converges to an eigenvector associate with the eigenvalue of largest modulus, when it is non-degenerate. Extensions of the Power Method include, e.g., the application of the Aitken Acceleration Theorem (in order to improve the rate of convergence), the possibility of obtaining convergence to eigenvectors associated with eigenvalues different from the largest one (“Shifted Inverse Power Method”), and the possibility of generating a sequence of subspaces that converges to the subspace associate, for a positive integer  $l$ , with the  $l$  dominant eigenvalues of the matrix (the “Subspace Iteration Method” and its variants, such as the ones based on the so-called “Krylov subspaces” [8, Chap. 9]). An interesting feature of Subspace Iteration Methods is the fact that they are based on bounds, such as those based on the generalized

Rayleigh quotients [8, Chap. 9], which may be combined with our estimates from Sect. 4.

**Acknowledgements** M. Sanguineti was partially supported by a PRIN Grant of the Italian Ministry for University and Research (Project “Models and Algorithms for Robust Network Optimization”). The authors are grateful to J.A.K. Suykens (Katholieke Universiteit Leuven) for having called their attention to the paper [35] and to the problem of estimating the accuracy of suboptimal solutions to KPCA.

**Appendix 1: Some remarks on principal component analysis (PCA)**

Given  $m$  data represented by the vectors  $x_1, \dots, x_m \in \mathbb{R}^d$ , such that  $\sum_{i=1}^m x_i = 0$  (centered data), Principal Component Analysis (PCA) [14] searches for a small number  $p$  of orthonormal vectors  $v_j \in \mathbb{R}^d, j = 1, \dots, p$ , on which the data are projected in such a way to minimize the *mean-square reconstruction error* over the  $p$  directions  $v_1, \dots, v_p$ , defined as

$$e(v_1, \dots, v_p) = \frac{1}{m} \sum_{i=1}^m \left\| x_i - \sum_{j=1}^p \langle x_i, v_j \rangle v_j \right\|^2.$$

PCA can be interpreted in terms of an eigenvalue problem. If we define the  $d \times d$  positive semi-definite matrix

$$C[\mathbf{x}] = \frac{1}{m} \sum_{i=1}^m x_i x_i^T,$$

with entries  $C[\mathbf{x}]_{jl} = \frac{1}{m} \sum_{i=1}^m x_{i,j} x_{i,l}$ , then it can be proved [29, Proposition 14.1] that the solution to Problem PCA satisfies

$$C[\mathbf{x}]v_j = \lambda_j v_j, \quad j = 1, \dots, p,$$

i.e., the  $v_j$ 's are eigenvectors of the matrix  $C[\mathbf{x}]$  corresponding to its  $p$  largest eigenvalues  $\lambda_1, \dots, \lambda_p$  (taking into account their multiplicity). In order to reduce the mean-square reconstruction error, only the directions corresponding to strictly positive eigenvalues of  $C[\mathbf{x}]$  are meaningful for PCA, which can be regarded as the process of diagonalizing the matrix  $C[\mathbf{x}]$ . Moreover, there exist  $p$  vectors  $a_1, \dots, a_p \in \mathbb{R}^m$  with components  $a_{ij}$ , such that ([29, p. 430], [30, Sect. 2])

$$v_j = \sum_{i=1}^m a_{ij} x_i.$$

Hence, the directions  $v_1, \dots, v_p$  belong to the space spanned by the data vectors  $x_1, \dots, x_m$ . Note that, if the data  $x_1, \dots, x_m$  are  $m$  independent realizations of a random variable, then the matrix  $C[\mathbf{x}]$  is an estimate of the *covariance matrix* of such a variable [30]. One can expect that a new datum  $x_{m+1}$ , extracted independently of the previous ones and according to the same probability distribution, is “well-represented” by its projections on the  $p$  directions previously determined.

If we introduce the  $m \times m$  Gram matrix of the data

$$\mathcal{X}[\mathbf{x}]_{ij} = \langle x_i, x_j \rangle,$$

then the vectors  $a_1, \dots, a_p$  satisfy the equations ([29, p. 430], [30])

$$\mathcal{X}[\mathbf{x}]a_j = m\lambda_j a_j, \quad j = 1, \dots, p.$$

It can be easily proved that if  $\lambda$  is a positive eigenvalue of  $\mathcal{C}[\mathbf{x}]$ , then  $m\lambda$  is a positive eigenvalue of  $\mathcal{X}[\mathbf{x}]$ , and conversely, if  $m\lambda$  is a positive eigenvalue of  $\mathcal{X}[\mathbf{x}]$ , then  $\lambda$  is a positive eigenvalue of  $\mathcal{C}[\mathbf{x}]$ . Thus, PCA can be also regarded as the problem of determining the  $p$  largest positive eigenvalues  $m\lambda_1, \dots, m\lambda_p$  of the matrix  $\mathcal{X}[\mathbf{x}]$  and the corresponding eigenvectors  $a_1, \dots, a_p$ . Moreover, the normalization condition  $\|v_i\| = 1, i = 1, \dots, p$ , can be formulated as  $\|a_j\| = \frac{1}{\sqrt{m\lambda_j}}$  [29, p. 430].

## Appendix 2: Some remarks about PCA in the feature space

An extension of PCA consists in performing the same computations, as those made by PCA in the input space  $\mathbb{R}^d$ , in another (possibly infinite-dimensional) inner product space  $F$ , called *feature space* [29, Chap. 1], which is related to the input domain  $\Omega \subseteq \mathbb{R}^d$  by a (typically nonlinear) map

$$\varphi : \Omega \rightarrow F, \quad x \mapsto f = \varphi(x).$$

By means of the map  $\varphi$ , the data are represented as elements of the space  $F$  and via the inner product  $\langle \cdot, \cdot \rangle_F$  one can consider similarity measures between data, which are more general than those corresponding to the inner product in the input space  $\mathbb{R}^d$ . To avoid burdening the notation and without loss of generality, also in the feature space we consider the *centered case*  $\sum_{i=1}^m \varphi(x_i) = 0$  (see Sect. 6 for a short discussion on the extension to the uncentered case).

In the feature space  $F$ , for  $j = 1, \dots, p$  the normalization condition on  $f_j$  is given by  $\|f_j\|_F = 1$  and the *mean-square reconstruction error* to be minimized is

$$e_F(f_1, \dots, f_p) = \frac{1}{m} \sum_{i=1}^m \left\| \varphi(x_i) - \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_F f_j \right\|_F^2,$$

where  $\|\cdot\|_F$  is the norm induced on  $F$  by the inner product  $\langle \cdot, \cdot \rangle_F$ . Thus, we have the following formulation of *PCA in the feature space*  $F$ :

$$\begin{cases} \min_{f_1, \dots, f_p \in F} e_F(f_1, \dots, f_p) \\ \text{s.t. } \langle f_i, f_j \rangle_F = \delta_{i,j}, \quad i, j = 1, \dots, p. \end{cases} \quad (21)$$

We now derive a formulation equivalent to (21). Let

$$\text{var}_F(f_1, \dots, f_p) = \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_F^2$$

be the *sum of the variances* of the  $p$  projections on  $f_1, \dots, f_p$ . As  $\langle f_i, f_j \rangle_F = \delta_{i,j}$ ,  $i, j = 1, \dots, p$ , we get

$$\begin{aligned} & e_F(f_1, \dots, f_p) + \text{var}_F(f_1, \dots, f_p) \\ &= \frac{1}{m} \sum_{i=1}^m \left( \left\| \varphi(x_i) - \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_F f_j \right\|_F^2 + \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_F^2 \right) \\ &= \frac{1}{m} \sum_{i=1}^m \left( \|\varphi(x_i)\|_F^2 + \left\| \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_F f_j \right\|_F^2 \right. \\ &\quad \left. - 2 \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_F \langle f_j, \varphi(x_i) \rangle_F + \sum_{j=1}^p \langle f_j, \varphi(x_i) \rangle_F^2 \right) \\ &= \frac{1}{m} \sum_{i=1}^m \|\varphi(x_i)\|_F^2. \end{aligned}$$

Therefore, the quantity  $e_F + \text{var}_F = \frac{1}{m} \sum_{i=1}^m \|\varphi(x_i)\|_F^2$  does not depend on the  $p$ -tuple  $(f_1, \dots, f_p)$  of orthonormal functions. Thus, by minimizing  $e_F(f_1, \dots, f_p)$ , one also maximizes  $\text{var}_F(f_1, \dots, f_p)$ . So, (21) is equivalent<sup>6</sup> to

$$\begin{cases} \max_{f_1, \dots, f_p \in F} \text{var}_F(f_1, \dots, f_p) \\ \text{s.t. } \langle f_i, f_j \rangle_F = \delta_{i,j}, \quad i, j = 1, \dots, p. \end{cases} \tag{22}$$

Moreover, similarly to PCA, it can be proved [29, proof of Proposition 14.1] that  $\max_{f_1, \dots, f_p \in B_1(\|\cdot\|_F)} \text{var}_F(f_1, \dots, f_p) = \sum_{j=1}^p \lambda_j$ , where  $\lambda_1, \dots, \lambda_p$  are the  $p$  largest eigenvalues of the *sample covariance matrix of the data in  $F$* , defined as<sup>7</sup>

$$\mathcal{C}_F[\mathbf{x}] = \frac{1}{m} \sum_{i=1}^m \varphi(x_i) \varphi(x_i)^T.$$

So, PCA in the feature space  $F$  consists in finding the  $p$  eigenvectors of  $\mathcal{C}_F[\mathbf{x}]$  corresponding to its  $p$  largest eigenvalues; the first principal component is the direction of projection with maximum variance [29, p. 443].

<sup>6</sup>Note that this equivalence does not exploit the (possible) presence of a kernel, so it holds for both  $F$  and  $F_K$ , as defined in Sect. 2.

<sup>7</sup>If  $F$  is infinite-dimensional, then  $\varphi(x_i) \varphi(x_i)^T$  has to be considered as the linear operator in  $F$  that maps  $f$  to  $\varphi(x_i) \langle \varphi(x_i), f \rangle_F$ .

### Appendix 3: Reproducing kernel Hilbert spaces (RKHSs)

A Reproducing Kernel Hilbert Space (RKHS) is a Hilbert space  $X$  of functions defined on a nonempty set  $\Omega$  such that for every  $u \in \Omega$  the evaluation functional  $\mathcal{F}_u$ , defined for any  $f \in X$  as  $\mathcal{F}_u(f) = f(u)$ , is bounded [2, 4, 7].

RKHSs can be characterized in terms of symmetric positive semi-definite functions  $K : \Omega \times \Omega \rightarrow \mathbb{R}$ , called *kernels*. By the Riesz Representation Theorem [11, p. 200], for every  $u \in \Omega$  there exists a unique element  $K_u \in X$ , called the *representer* of  $u$ , such that  $\mathcal{F}_u(f) = \langle f, K_u \rangle$  for all  $f \in X$  (this property is called the *reproducing property*).

On the other hand, every kernel  $K : \Omega \times \Omega \rightarrow \mathbb{R}$  generates a RKHS  $\mathcal{H}_K(\Omega)$  that is the completion of the linear span of the set  $\{K_u : u \in \Omega\}$ , with the inner product defined as  $\langle K_u, K_v \rangle_K = K(u, v)$  and the induced norm  $\|\cdot\|_K$  (see, e.g., [2] and [4, p. 81]).

A paradigmatic example of a kernel on  $\mathbb{R}^d$  is the *Gaussian kernel*  $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ , defined as  $K(u, v) = \exp(-\|u - v\|_2^2)$ . Other examples of kernels are  $K(u, v) = \exp(-\|u - v\|_2)$ ,  $K(u, v) = \langle u, v \rangle^p$  (*homogeneous polynomial* of degree  $p$ ), where  $\langle \cdot, \cdot \rangle$  is any inner product on  $\mathbb{R}^d$ ,  $K(u, v) = (1 + \langle u, v \rangle)^p$  (*inhomogeneous polynomial* of degree  $p$ ), and  $K(u, v) = (a^2 + \|u - v\|_2^2)^{-\alpha}$ , with  $\alpha > 0$  [7, p. 38].

By the reproducing property and the Cauchy–Schwartz inequality, for every  $f \in \mathcal{H}_K(\Omega)$  and  $u \in \Omega$  one has  $|f(u)| = |\langle f, K_u \rangle_K| \leq \|f\|_K \|K_u\|_K$ . Therefore  $\sup_{u \in \Omega} |f(u)| \leq s_K \|f\|_K$ , where  $s_K = \sup_{u \in \Omega} \sqrt{K(u, u)}$ . Finiteness of  $s_K$  is guaranteed when  $\Omega$  is compact and the kernel  $K$  is continuous (which are common hypotheses in applications) [7], or simply it is guaranteed a priori, also for non compact domains, if one uses bounded kernels (like the Gaussian kernel).

The role played by norms on RKHSs in measuring various types of oscillations of input-output mappings can be illustrated by the following two examples of classes of kernels. The first one is formed by *Mercer kernels* ([2], [7], [29, Chap. 2]), i.e., continuous kernels  $K : \Omega \times \Omega \rightarrow \mathbb{R}$ , where  $\Omega \subset \mathbb{R}^d$  is compact. For a Mercer kernel  $K$ ,  $\|f\|_K^2$  can be expressed by using the eigenvectors and eigenvalues of the compact linear operator  $L_K : \mathcal{L}_2(\Omega) \rightarrow \mathcal{C}(\Omega)$  defined for every  $f \in \mathcal{L}_2(\Omega)$  as  $L_K(f)(x) = \int_{\Omega} K(x, u) f(u) du$ , where  $\mathcal{L}_2(\Omega)$  and  $\mathcal{C}(\Omega)$  denote the spaces of square integrable and continuous functions on  $\Omega$ , respectively. By the Mercer Theorem ([7, pp. 34–36], [29, p. 37])

$$\|f\|_K^2 = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i},$$

where the  $\lambda_i$ 's are the positive eigenvalues of  $L_K$ , ordered non-increasingly (and accounting for their multiplicity), the  $c_i$ 's are the coefficients of the representation  $f = \sum_{i=1}^{\infty} c_i \varphi_i$ , and  $\{\sqrt{\lambda_i} \varphi_i\}$  is the orthonormal basis<sup>8</sup> of  $\mathcal{H}_K(\Omega)$  formed by the eigenvectors associate with positive eigenvalues of  $L_K$  [11]. The sequence  $\{\lambda_i\}$  is either finite or convergent to zero and for  $K$  smooth enough, the convergence to

<sup>8</sup>It can be proved that  $\{\varphi_i\}$  is a subset of an orthonormal basis of  $\mathcal{L}_2(\Omega)$ .

zero is rather fast [10, p. 1119]. So restricting the minimization to a ball in  $\|\cdot\|_K$  means that only functions obtained by filtering out “high frequencies” are considered as admissible solutions to KPCA (where high frequencies are associated with large values of  $i$ ).

The second class of kernels illustrating the role of  $\|\cdot\|_K$  in measuring oscillations consists of *convolution kernels* (also called *translation-invariant kernels*), i.e., kernels defined on  $\mathbb{R}^d \times \mathbb{R}^d$  such that  $K(x, y) = k(x - y)$  and for which the Fourier transform<sup>9</sup>  $\tilde{k}$  of  $k$  is positive. For such kernels one has (see, e.g. [12] and [29, p. 97])

$$\|f\|_K^2 = (2\pi)^{d/2} \int_{\mathbb{R}^d} \frac{|\tilde{f}(\omega)|^2}{\tilde{k}(\omega)} d\omega.$$

So the function  $1/\tilde{k}$  plays a role analogous to that of the sequence  $\{1/\lambda_i\}$  in the case of a Mercer kernel. For example, the *Gaussian kernel* is a convolution kernel with a positive Fourier transform. Another example of a convolution kernel with a positive Fourier transform is  $K(x, y) = k(x - y)$ , where  $k(t) = \exp(-a \|t\|_2)$ ,  $\tilde{k}(\omega) = 2^{d/2} a \pi^{-1/2} \Gamma(d/2 + 1) (a^2 + \|\omega\|_2^2)^{-(d+1)/2}$  [29, p. 108] and  $\Gamma$  denotes the Gamma function, defined for  $g > 0$  as  $\Gamma(g) = \int_0^\infty \exp(-r) r^{g-1} dr$ . In this case, the rate of decay of  $\tilde{k}(\omega)$  is of order  $\|\omega\|_2^{-(d+1)}$ . In particular, for  $d = 1$  and  $a = 1$ , we get<sup>10</sup>  $K(u, v) = k(u - v) = \exp(-|u - v|)$  and  $\|f\|_K^2 = 2\sqrt{\pi}(\|f\|_{\mathcal{L}_2}^2 + \|f'\|_{\mathcal{L}_2}^2)$ , where  $\|\cdot\|_{\mathcal{L}_2}$  denotes the  $\mathcal{L}_2$ -norm for functions defined on  $\mathbb{R}$ . So, as pointed out in [12], in this case the norm on the RKHS is equal to a Sobolev norm, and restricting the minimization to a ball in  $\|\cdot\|_K$  means that only functions that “do not change too fast” are considered as admissible solutions to KPCA.

For more details on kernels and their role in learning theory, we refer the reader to [6, 29, 33].

## References

1. Achlioptas, D., McSherry, F., Schölkopf, B.: Sampling techniques for kernel methods. In: Proceedings of NIPS 2001, Vancouver, BC, Canada, 3–8 December 2001. Advances in Neural Information Processing Systems, vol. 14, pp. 335–342. MIT Press, Cambridge (2001)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. AMS **68**, 337–404 (1950)
3. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory **39**, 930–945 (1993)
4. Berg, C., Christensen, J.P.R., Ressel, P.: Harmonic Analysis on Semigroups. Springer, New York (1984)
5. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. **20**, 1–25 (1995)
6. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2003, first published in 2000)

<sup>9</sup>We use the definition  $\tilde{f}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x) \exp(-j\langle\omega, x\rangle) dx$ .

<sup>10</sup>As  $\Gamma(1) = 1$ ,  $\Gamma(1/2) = \sqrt{\pi}$ , and  $\Gamma(g + 1) = g \Gamma(g)$ , we have  $\tilde{k}(\omega) = (\sqrt{2}(1 + \omega^2))^{-1}$ . Thus  $\|f\|_K^2 = \sqrt{2\pi} \int_{\mathbb{R}} |\tilde{f}(\omega)|^2 (\sqrt{2}(1 + \omega^2)) d\omega = 2\sqrt{\pi} \int_{\mathbb{R}} |\tilde{f}(\omega)|^2 d\omega + 2\sqrt{\pi} \int_{\mathbb{R}} \omega^2 |\tilde{f}(\omega)|^2 d\omega$ . As  $\tilde{f}'(\omega) = j\omega \tilde{f}(\omega)$  and  $\int_{\mathbb{R}} f(t)^2 dt = \int_{\mathbb{R}} |\tilde{f}(\omega)|^2 d\omega$ , by Parseval’s formula [28, p. 189], we get  $\|f\|_K^2 = 2\sqrt{\pi}(\|f\|_{\mathcal{L}_2}^2 + \|f'\|_{\mathcal{L}_2}^2)$ .

7. Cucker, F., Smale, S.: On the mathematical foundations of learning. *Bull. AMS* **39**, 1–49 (2001)
8. Dahlquist, G., Björck, A.: *Numerical Methods in Scientific Computing*. SIAM, Philadelphia (to appear); <http://www.mai.liu.se/~akbjo/NMbook.html>
9. Drineas, P., Mahoney, M.W.: On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.* **6**, 2153–2175 (2005)
10. Dunford, N., Schwartz, J.T.: *Linear Operators. Part II: Spectral Theory*. Interscience, New York (1963)
11. Friedman, A.: *Foundations of Modern Analysis*. Dover, New York (1982)
12. Girosi, F.: An equivalence between sparse approximation and support vector machines. *Neural Comput.* **10**, 1455–1480 (1998)
13. Girosi, F., Jones, M., Poggio, T.: Regularization theory and neural networks architectures. *Neural Comput.* **7**, 219–269 (1995)
14. Jolliffe, I.T.: *Principal Component Analysis*. Springer Series in Statistics. Springer, New York (1986)
15. Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Stat.* **20**, 608–613 (1992)
16. Kolmogorov, A.N., Fomin, S.V.: *Introductory Real Analysis*. Dover, New York (1975)
17. Kůrková, V.: Dimension-independent rates of approximation by neural networks. In: Warwick, K., Kárný, M. (eds.) *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality*, pp. 261–270. Birkhäuser, Basel (1997)
18. Kůrková, V., Sanguineti, M.: Bounds on rates of variable-basis and neural-network approximation. *IEEE Trans. Inf. Theory* **47**, 2659–2665 (2001)
19. Kůrková, V., Sanguineti, M.: Comparison of worst case errors in linear and neural network approximation. *IEEE Trans. Inf. Theory* **48**, 264–275 (2002)
20. Kůrková, V., Sanguineti, M.: Error estimates for approximate optimization by the extended Ritz method. *SIAM J. Optim.* **15**, 461–487 (2005)
21. Kůrková, V., Sanguineti, M.: Learning with generalization capability by kernel methods of bounded complexity. *J. Complex.* **21**, 350–367 (2005)
22. Kůrková, V., Savický, P., Hlaváčková, K.: Representations and rates of approximation of real-valued Boolean functions by neural networks. *Neural Netw.* **11**, 651–659 (1998)
23. Parzen, E.: An approach to time series analysis. *Ann. Math. Stat.* **32**, 951–989 (1961)
24. Pisier, G.: Remarques sur un resultat non publié de B. Maurey. Séminaire d'Analyse Fonctionnelle 1980/81. École Polytechnique, Centre de Mathématiques, Palaiseau, France. Exposé no. V, V. 1–V. 12
25. Poggio, T., Girosi, F.: Networks for approximation and learning. *Proc. IEEE* **78**, 1481–1497 (1990)
26. Poggio, T., Girosi, F.: Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247**, 978–982 (1990)
27. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge (1992)
28. Rudin, W.: *Functional Analysis*. McGraw-Hill, New York (1973)
29. Schölkopf, B., Smola, A.: *Learning with Kernels—Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge (2002)
30. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299–1319 (1998)
31. Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K.-R., Rätsch, G., Smola, A.: Input space vs. feature space in kernel-based methods. *IEEE Trans. Neural Netw.* **10**, 1000–1017 (1999)
32. Schönberg, I.J.: Metric spaces and completely monotone functions. *Ann. Math.* **39**, 811–841 (1938)
33. Shawe-Taylor, J., Cristianini, N.: *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge (2004)
34. Shawe-Taylor, J., Williams, C.K.I., Cristianini, N., Kandola, J.: On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inf. Theory* **51**, 2510–2522 (2005)
35. Suykens, J.A.K., Van Gestel, T., Vandewalle, J., De Moor, B.: A support vector machine formulation to PCA analysis and its kernel version. *IEEE Trans. Neural Netw.* **14**, 447–450 (2003)
36. Tikhonov, A.N.: Solutions of incorrectly formulated problems and the regularization method. *Sov. Math. Dokl.* **4**, 1035–1038 (1963)
37. Tikhonov, A.N., Arsenin, V.Y.: *Solutions of Ill-Posed Problems*. Winston, Washington (1977)
38. Vasin, V.V.: Relationship of several variational methods for the approximate solution of ill-posed problems. *Math. Notes Acad. Sci. USSR* **7(3/4)**, 161–165 (1970) (Translated from *Matematicheskie Zametki* **7(3)**, 265–272 (1970))

39. Wahba, G.: *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. SIAM, Philadelphia (1990)
40. Williams, C.K.I., Seeger, M.: Using the Nyström method to speed up kernel machines. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) *Advances in Neural Information Processing Systems*, vol. 13, pp. 682–688. MIT Press, Cambridge (2001)
41. Zeidler, E.: *Nonlinear Functional Analysis and Its Applications III. Variational Methods and Optimization*. Springer, New York (1985)