A closed-form solution for the pre-image problem in kernel-based machines

Paul Honeine, Cédric Richard

Noname manuscript No.
(will be inserted by the editor)

A closed-form solution for the pre-image problem in kernel-based machines

Paul Honeine · Cédric Richard

Received: date / Accepted: date

Abstract The pre-image problem is a challenging research subject pursued by many researchers in machine learning. Kernel-based machines seek some relevant feature in a reproducing kernel Hilbert space (RKHS), optimized in a given sense, such as kernel-PCA algorithms. Operating the latter for denoising requires solving the pre-image problem, i.e. estimating a pattern in the input space whose image in the RKHS is approximately a given feature. Solving the pre-image problem is pioneered by Mika's fixed-point iterative optimization technique. Recent approaches take advantage of prior knowledge provided by the training data, whose coordinates are known in the input space and implicitly in the RKHS, a first step in this direction made by Kwok's algorithm based on multidimensional scaling (MDS). Using such prior knowledge, we propose in this paper a new technique to learn the pre-image, with the elegance that only linear algebra is involved. This is achieved by establishing a coordinate system in the RKHS with an isometry with the input space, i.e. the inner products of training data are preserved using both representations. We suggest representing any feature in this coordinate system, which gives us information regarding its pre-image in the input space. We show that this approach provides a natural pre-image technique in kernel-based machines since, on one hand it involves only linear algebra operations, and on the other it can be written directly using the kernel values, without the need to evaluate distances as with the MDS approach. The performance of the proposed approach is illustrated for denoising with kernel-PCA, and compared to state-of-the-art methods on both synthetic datasets and realdata handwritten digits.

Keywords kernel-based machines · pre-image problem · linear algebra · kernel-PCA · nonlinear denoising

1 Introduction

In the last decade or so, kernel-based machines have enjoyed increasing popularity, providing a breakthrough in both statistical learning theory and low computational complexity of nonlinear algorithms. Pioneered by Vapnik's Support Vector Machines (SVM) [20], this concept attracted significant attention due to the ever-expanding challenges in machine learning. Since then, many nonlinear algorithms have been developed, for supervised learning (or classification) such as kernel Fisher discriminant analysis [13] and least-squares SVM [18], and for unsupervised learning (with unlabelled data) with kernel principal component analysis (kernel-PCA) [17] and support vector domain description [19]. The main idea behind nonlinear algorithms in kernel-based machines is the *kernel trick* [1]. This concept gives rise to nonlinear algorithms based on classical linear ones, under the only requirement that the algorithm can be expressed only in terms of inner products between data. Then, data from the input space are (nonlinearly) mapped into a feature space. This mapping is achieved implicitly by substituting the inner product operator by a positive definite kernel, thus without

This is an extended version of the paper [8], winner of the Best Paper Award at the IEEE Machine Learning For Signal Processing workshop.

Paul Honeine
Institut Charles Delaunay (FRE CNRS 2848), LM2S, Université de technologie de Troyes, 10010 Troyes, France
E-mail: paul.honeine@utt.fr

Cédric Richard
Laboratoire Fizeau (UMR CNRS 6525), Observatoire de la Côte d'Azur, Université de Nice Sophia-Antipolis, 06108 Nice, France
E-mail: cedric.richard@unice.fr

much additional computational cost. This is the essence of the kernel trick. In order to provide the unified functional framework, common to many communities, this kernel is called the *reproducing kernel* while the induced feature space is the *reproducing kernel Hilbert space* (RKHS).

With the ever-increasing demands in machine learning, new challenges require computing the inverse map. For instance, while the kernel trick provides an elegant approach to apply denoising or compression techniques in the RKHS, we need to go back into the input space for the final result. This is the case in denoising an image (or a signal), the reconstructed image belongs to the input space of training images. However, getting back to the input space from the RKHS is not obvious in general, as most features of the latter may not have an exact pre-image in the former. This is the pre-image problem in kernel-based machines, as one seeks an approximate solution. Solving this problem has received a growing amount of attention, with the most breakthrough given in [14] and [11]. In the former work, Mika *et al.* present the problem and its ill-posedness, and derive a fixed-point iterative scheme to find an approximate solution. Hence, there is no guarantee that this leads to a global optimum, and may be unstable. In the latter work, Kwok *et al.* determine a relationship between the distances in the RKHS and the distances in the input data, based on a set of training data. Applying a multidimensional scaling technique (MDS) leads to an inverse map estimate and thus to the pre-image. This approach opens the door to a range of other techniques taking prior knowledge from training data in both spaces, such as manifold learning [6] and out-of-sample methods [5,2].

In this paper, we propose a novel approach to solve the pre-image problem. To achieve this, we learn a coordinate system, not necessarily orthogonal, in the RKHS having an isometry with the input space. In other words, the inner products of the training data are (approximately) equal in both representations. Thus, by representing any feature of the RKHS in this coordinate system, we get an estimate of the inner products between the training data and its counterpart in the input space. It turns out that this approach is natural to kernel-based machines, and essentially requires only linear algebra, with any off-the-shelf linear solver. The proposed method is universal in the sense of being independent, in its formulation, of both the type of the adopted kernel and of the feature under investigation. Moreover, it extends naturally to get the pre-images of a set of features, since the coordinate system is computed only once.

The rest of the paper is organized as follows. In the next Section, we briefly present the framework behind kernel-based machines, with an illustration on kernel-PCA for denoising. In Section 3, the pre-image problem is described, and previous work on solving the problem are examined. The proposed method is described in Section 4, with connections to other methods. Experiments on synthetic and real datasets are presented in Section 5. Section 6 ends this paper with a brief conclusion.

2 Kernel-based machines, with application to nonlinear denoising using kernel-PCA

2.1 Kernel-based machines

Let $\mathcal{X}$ be a compact of $\mathbb{R}^p$, equipped with the natural Euclidean inner product defined for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ by $\boldsymbol{x}_i^\top \boldsymbol{x}_j = \sum_{\ell=1}^p \boldsymbol{x}_{i,\ell}\boldsymbol{x}_{j,\ell}$, with $\boldsymbol{x}_{\cdot,\ell}$ the ℓ-th entry of vector $\boldsymbol{x}_\cdot$. Let $\kappa(\cdot,\cdot)$ be a positive (semi-)definite kernel on $\mathcal{X}\times\mathcal{X}$, where the positive (semi-)definitness is defined by the property

$$\sum_{i,j} \alpha_i\alpha_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) \geq 0$$

for all $\alpha_i, \alpha_j \in \mathbb{R}$ and $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$. The Moore-Aronszajn theorem [3] states that for every positive definite kernel, there exists a unique reproducing kernel Hilbert space (RKHS), and viceversa. With this one-to-one correspondence between RKHS and positive definite kernels, the latter will be called reproducing kernels hereafter. Let $\mathcal{H}$ be the RKHS associated with κ, and let $\langle\cdot\,,\cdot\rangle_{\mathcal{H}}$ be the endowed inner product. This means that any arbitrary function $\psi(\cdot)$ in $\mathcal{H}$ can be evaluated at any $\boldsymbol{x}_j \in \mathcal{X}$ with

$$\psi(\boldsymbol{x}_j) = \langle\psi(\cdot), \kappa(\cdot,\boldsymbol{x}_j)\rangle_{\mathcal{H}}. \tag{1}$$

This expression shows that the kernel is the representer of evaluation. Moreover, replacing in this expression $\psi(\cdot)$ by $\kappa(\cdot,\boldsymbol{x}_i)$ yields to the popular property

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \langle\kappa(\cdot,\boldsymbol{x}_i), \kappa(\cdot,\boldsymbol{x}_j)\rangle_{\mathcal{H}}, \tag{2}$$

for all $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$. This is the reproducing property from which the name of reproducing kernel is derived. Denoting by $\phi(\cdot)$ the map that assigns to each input $\boldsymbol{x} \in \mathcal{X}$ the kernel function $\kappa(\cdot,\boldsymbol{x})$, the reproducing property (2) implies that $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle_{\mathcal{H}}$. The kernel then evaluates the inner product of any pair of elements of $\mathcal{X}$ mapped into $\mathcal{H}$, without any explicit knowledge of either the mapping function $\phi(\cdot)$ or the RKHS $\mathcal{H}$. This is the well-known kernel trick. Examples of commonly used reproducing kernels are given in Table 1.

Table 1 Commonly used reproducing kernels in machine learning, with parameters $\beta > 0$, $q \in \mathbb{N}$, and $\sigma > 0$.

Polynomial	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + \beta)^q$
Laplace	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|/\sigma\right)$
Gaussian	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2\right)$

In combination with the kernel trick, the representer theorem provides a powerful theoretical foundation for kernel-based machines. Initially derived in [10] and recently generalized in [16], results of this theorem include SVM and kernel-PCA, where one seeks to maximize the separating margin between classes or the variance of projected data, respectively. This theorem states that any function $\varphi^*(\cdot)$ of a RKHS $\mathcal{H}$ minimizing a regularized cost functional of the form

$$\sum_{i=1}^{n} J(\varphi(\boldsymbol{x}_i), y_i) + g(\|\varphi\|^2_{\mathcal{H}}),$$

with predicted output $\varphi(\boldsymbol{x}_i)$ for input $\boldsymbol{x}_i$, and eventually the desired output y_i, and $g(\cdot)$ a strictly monotonically increasing function on $\mathbb{R}_+$, can be written as a kernel expansion in terms of available data

$$\varphi^*(\cdot) = \sum_{i=1}^{n} \gamma_i \, \kappa(\cdot, \boldsymbol{x}_i). \tag{3}$$

This theorem shows that even in an infinite dimensional RKHS, as with the Gaussian kernel, we only need to work in the subspace spanned by the n kernel functions of the training data, $\kappa(\cdot, \boldsymbol{x}_1), \ldots, \kappa(\cdot, \boldsymbol{x}_n)$.

2.2 Kernel-PCA for denoising

An elegant kernel-based machine is the kernel-PCA [17], a nonlinear extension of one of the most used dimension reduction and denoising technique, the principal component analysis (PCA).

With PCA, one seeks principal axes that capture the highest variance in the data, that is, useful information as opposed to noise, and thus projecting data onto the space spanned by these relevant axes yields a denoising scheme. These principal axes are the eigenvectors φ_k associated with the largest eigenvalues λ_k of the covariance matrix $\boldsymbol{R}$ of data, i.e. solving the eigen-problem $\boldsymbol{R}\varphi_k = \lambda_k \varphi_k$. There exists another formulation of the PCA algorithm, using only inner products of the training data. By substituting the inner product operator with any valid reproducing kernel, we get an implicit nonlinear mapping of the data into a RKHS. This is the kernel-PCA, where each of the resulting principal functions takes the representer form (3), with

$$\varphi^*_k(\cdot) = \sum_{i=1}^{n} \gamma_{i,k} \, \kappa(\cdot, \boldsymbol{x}_i).$$

The weighting coefficients $\gamma_{i,k}$ are obtained from the eigen-decomposition of the so-called Gram matrix $\boldsymbol{K}$, whose entries are $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, for $i, j = 1, \ldots, n$, by solving

$$\boldsymbol{K}\boldsymbol{\gamma} = n\lambda\boldsymbol{\gamma}.$$

In order to have a PCA interpretation in feature space, two issues should be carried out. First, data is implicitly centered in feature space by substituting in this expression $\boldsymbol{K}$ with $(\mathbf{1} - \mathbf{1}_n)\boldsymbol{K}(\mathbf{1} - \mathbf{1}_n)$, with $\mathbf{1}_n$ the n-by-n matrix of entries $1/n$ and $\mathbf{1}$ the identity matrix; second, principal functions are normalized to 1, by scaling expansion coefficients such that $\sum_{i=1}^{n} \gamma_{i,k}^2 = 1/\lambda_k$.

In the same spirit of the conventional PCA, one constructs a subspace of $\mathcal{H}$ spanned by the most relevant principal functions. Using kernel-PCA for denoising any given $\boldsymbol{x} \in \mathcal{X}$, we project the associated kernel function $\kappa(\cdot, \boldsymbol{x})$ onto that subspace. Since this subspace is spanned by most relevant principal functions, each of the form (3), any function from this subspace takes the same form, i.e. can be written as a kernel expansion in terms of available data. Let $\varphi^*(\cdot)$ be this projection, with

$$\varphi^*(\cdot) = \sum_{i=1}^{n} \gamma_i \, \kappa(\cdot, \boldsymbol{x}_i),$$

which is assumed to be noise-free by virtue of the PCA interpretation. From this denoised feature, we need to get its counterpart in the input space, e.g. a denoised image in the image space. As illustrated in Figure 1, this requires the estimation of the pattern $\boldsymbol{x}^*$ from $\kappa(\cdot, \boldsymbol{x})$, by solving the pre-image problem.

3 A brief review of the pre-image problem

For supervised learning, one seeks a prediction value associated to any input such as in regression problems, while in classification this value is compared to a threshold, which yields a decision rule. While every optimal function $\varphi^*(\cdot)$ takes the form (3), we obtain its evaluation at any $\boldsymbol{x}$ with $\sum_{i=1}^{n} \gamma_i \, \kappa(\boldsymbol{x}_i, \boldsymbol{x})$, thus requiring only computing values of the kernel. For pattern recognition with unsupervised learning, one is often interested in the feature in the feature space, or more precisely in its counterpart in the input space.

Estimating the input whose map is an arbitrary function in the RKHS is an ill-posed problem. To show

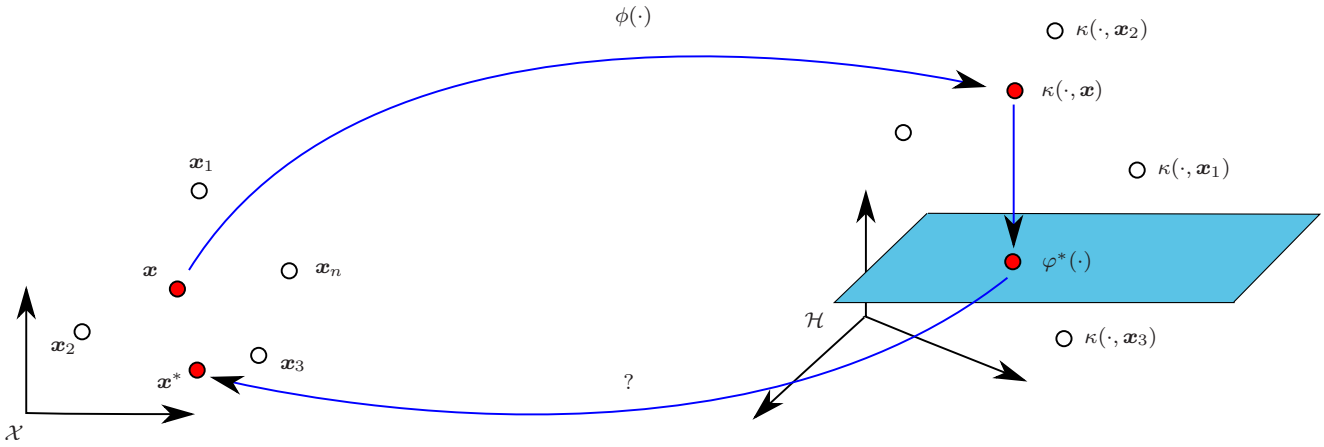

Fig. 1 Illustration of the pre-image problem in kernel-based machines.

this, recall that the dimensionality of the feature space can be very high, and even infinite with some kernels such as the Gaussian kernel. Thus, (most) features $\varphi^*(\cdot) \in \mathcal{H}$ might not have an existing pre-image in $\mathcal{X}$, i.e. a $\boldsymbol{x}^*$ such that $\kappa(\cdot, \boldsymbol{x}^*) = \varphi^*(\cdot)$. In order to circumvent this difficulty, one seeks an approximate solution, i.e. $\boldsymbol{x}^* \in \mathcal{X}$ whose map $\kappa(\cdot, \boldsymbol{x}^*)$ is as close as possible to $\varphi^*(\cdot)$. This is the pre-image problem in kernel-based machines. Methods for solving the pre-image problem are roughly classified into two categories: Fixed-point iterative methods and methods based on learning the *inverse map*.

The pre-image problem was initially studied by Mika *et al.* in [14]. They proposed to solve the optimization problem

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{X}} \|\varphi^*(\cdot) - \kappa(\cdot, \boldsymbol{x})\|^2_{\mathcal{H}}, \tag{4}$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the RKHS. For this purpose, a fixed-point iterative scheme is used to solve the pre-image problem. However, since this optimization problem is highly non-convex, such iterative technique suffers from numerical instabilities and local minima. The pre-image will highly depend on the initial guess and is likely to get stuck in a local minimum. A further improvement of the fixed-point iterative scheme is presented in [15], where the authors operate additional approximations by, roughly speaking, substituting the mapping $\kappa(\cdot, \boldsymbol{x})$ in (4) with its projection onto the subspace. It is worth noting that as an alternative to Mika's distance minimization, one may consider a collinearity maximization problem [2], with

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in \mathcal{X}} \left\langle \frac{\varphi^*(\cdot)}{\|\varphi^*(\cdot)\|_{\mathcal{H}}}, \frac{\kappa(\cdot, \boldsymbol{x})}{\|\kappa(\cdot, \boldsymbol{x})\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}}.$$

A key ingredient of these methods is a high dependence on the kernel type, since the fixed-point can only be applied to some specific kernels such as the Gaussian kernel, and only more recently extended to polynomial kernels in [11].

Recent approaches take advantage of prior knowledge provided by some available training data, whose *coordinates* are available in both the input and the feature spaces. This approach is initiated by an algorithm based on multidimensional scaling (MDS), presented by Kwok *et al.* [11]. This is achieved by computing distances between every pair of training data, in both spaces. For each pair, the classical Euclidean distance is used in the input space, as well as the distance in the RKHS which can be computed using kernel values. With these pairs of distances, a MDS technique is considered by performing a singular-value-decomposition[1]. This yields an *inverse map*, in the same spirit of the out-of-sample extension [2]. In order to make this method tractable in practice, only the neighboring data affect the pre-image estimation. Learning the inverse map is studied in [4] by solving a regression problem, while alternative approaches can be based on the manifold learning [6]. All these methods take advantage of prior knowledge, i.e. training data with information available in both input and feature spaces.

Using such prior knowledge, we propose in this paper to learn the inverse map without the need to compute distances, and does not require sophisticated optimization schemes. Only conventional linear algebra are needed. Furthermore, it is universe, in the sense that it does not depend on the kernel type, as opposed to fixed-point iterative techniques.

[1] This is done by operating on the distances, transforming them into inner products, and then apply eigen-decomposition into the resulting Gram matrix to get the coordinates. This nicely captures our guiding intuition of the problem in contrast with the MDS : we propose to work exclusively on the inner products, without the need to compute distances.

4 The proposed pre-image method

Given a set of training data $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, we seek the pre-image in $\mathcal{X}$ of some arbitrary $\varphi^*(\cdot)$ of the RKHS $\mathcal{H}$, denoted $\boldsymbol{x}^*$. The proposed method can be organized into two stages: learning the inverse map and operating a pre-image. To learn the inverse map, a coordinate system is constructed in the RKHS, having an isometry with the input space coordinates, where the isometry is defined with respect to the training data. In order to operate a pre-image, we represent $\varphi^*(\cdot)$ in this coordinate system which, by virtue of the isometry, gives the values of the inner products of its pre-image with the training data in the input space. From these values we obtain the pre-image $\boldsymbol{x}^*$.

Stage 1: Learn the inverse map

In this stage, we provide a coordinate system in the RKHS that is isometric with the input space. In order to achieve such isometry, we consider a set of n training data $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \in \mathcal{X}$. By virtue of the representer theorem, we only need to consider the subspace spanned by their kernel functions $\{\kappa(\cdot, \boldsymbol{x}_1), \kappa(\cdot, \boldsymbol{x}_2), \ldots, \kappa(\cdot, \boldsymbol{x}_n)\}$. Within this subspace, we define the set of ℓ coordinate functions, denoted $\{\psi_1(\cdot), \psi_2(\cdot), \ldots, \psi_\ell(\cdot)\}$ with $\ell \leq n$, and write

$$\psi_k(\cdot) = \sum_{i=1}^{n} \alpha_{k,i}\, \kappa(\cdot, \boldsymbol{x}_i),$$

for $k = 1, 2, \ldots, \ell$. For any kernel function $\kappa(\cdot, \boldsymbol{x})$, its coordinate on $\psi_k(\cdot)$ is given by

$$\langle \psi_k(\cdot), \kappa(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}} = \psi_k(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_{k,i}\, \kappa(\boldsymbol{x}_i, \boldsymbol{x}),$$

where (1) is used. Therefore, its representation in this coordinate system is obtained by the ℓ coordinates, written vector-wise as

$$\boldsymbol{\Psi}_{\boldsymbol{x}} = [\psi_1(\boldsymbol{x}) \;\; \psi_2(\boldsymbol{x}) \;\; \cdots \;\; \psi_\ell(\boldsymbol{x})]^\top,$$

where the k-th entry depends on the $\alpha_{k,i}$, for $i = 1, \ldots, n$.

In order to estimate the coordinate functions, we propose an equivalence, between the inner products in this coordinate system and their counterparts in the canonic input space, using the model

$$\boldsymbol{\Psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\Psi}_{\boldsymbol{x}_j} = \boldsymbol{x}_i^\top \boldsymbol{x}_j + \epsilon_{ij}, \tag{5}$$

for all the training set, i.e. $i, j = 1, 2, \ldots, n$, and where ϵ_{ij} corresponds to the lack-of-fit of the model. We insist on the fact that this model is not coupled with any constraint on the coordinate functions, as opposed to the orthogonality between the functions resulting from the kernel-PCA. The only requirement we impose is the isometry defined in (5). The minimization of the variance of ϵ_{ij}, a lack-of-fit criterion, consists of solving the optimization problem

$$\min_{\psi_1, \ldots, \psi_\ell} \frac{1}{2} \sum_{i,j=1}^{n} \left(\boldsymbol{x}_i^\top \boldsymbol{x}_j - \boldsymbol{\Psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\Psi}_{\boldsymbol{x}_j} \right)^2 + \lambda\, R(\psi_1, \ldots, \psi_\ell).$$

As suggested in machine learning literature, we include in this expression a regularization term, where λ a tunable parameter controlling the tradeoff between the fitness to the model (5) and the smoothness of the solution. In order to penalize high norm functions, the regularization $R(\psi_1, \ldots, \psi_\ell) = \sum_{k=1}^{\ell} \|\psi_k\|_{\mathcal{H}}^2$ is used in this paper.

This optimization problem can be written in matrix form. This is done by a factorization of $\boldsymbol{\Psi}_{\boldsymbol{x}}$ into a matrix of unknowns and a vector of available information, with

$$\boldsymbol{\Psi}_{\boldsymbol{x}} = \boldsymbol{A}\, \boldsymbol{\kappa}_{\boldsymbol{x}},$$

where $\boldsymbol{\kappa}_{\boldsymbol{x}} = [\kappa(\boldsymbol{x}_1, \boldsymbol{x}) \;\; \kappa(\boldsymbol{x}_2, \boldsymbol{x}) \;\; \cdots \;\; \kappa(\boldsymbol{x}_n, \boldsymbol{x})]^\top$ and $\boldsymbol{A}$ is a $\ell \times n$ matrix of unknowns whose (k, i)-th entry is $\alpha_{k,i}$. This leads to the optimization problem

$$\widehat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \; \frac{1}{2} \sum_{i,j=1}^{n} \left(\boldsymbol{x}_i^\top \boldsymbol{x}_j - \boldsymbol{\kappa}_{\boldsymbol{x}_i}^\top \boldsymbol{A}^\top \boldsymbol{A}\, \boldsymbol{\kappa}_{\boldsymbol{x}_j} \right)^2 + \lambda \sum_{k=1}^{\ell} \sum_{i,j=1}^{n} \alpha_{k,i} \alpha_{k,j}\, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

By denoting $\|\cdot\|_F$ the Frobenius norm[2] of a matrix and $\mathrm{tr}(\cdot)$ its trace, this yields

$$\widehat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \frac{1}{2} \|\boldsymbol{P} - \boldsymbol{K}\, \boldsymbol{A}^\top \boldsymbol{A}\, \boldsymbol{K}\|_F^2 + \lambda\, \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{K}),$$

where $\boldsymbol{P}$ and $\boldsymbol{K}$ are the Gram matrices with entries $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ and $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, respectively. Taking the derivative of the above cost function with respect to $\boldsymbol{A}^\top \boldsymbol{A}$, rather than $\boldsymbol{A}$, and setting it to zero, we get

$$\widehat{\boldsymbol{A}}^\top \widehat{\boldsymbol{A}} = \boldsymbol{K}^{-1} \left(\boldsymbol{P} - \lambda\, \boldsymbol{K}^{-1} \right) \boldsymbol{K}^{-1}. \tag{6}$$

In what follows, we show that only $\boldsymbol{A}^\top \boldsymbol{A}$ is required to find the pre-image, rather than $\boldsymbol{A}$. Fortunately, we do not need to compute the coefficients $\alpha_{k,j}$ to generate the coordinate system in the RKHS; only their inner products are required.

[2] The Frobenius norm of a matrix is the root of sum of squared (absolute) values of all its elements, or equivalently $\|\boldsymbol{M}\|_F^2 = \mathrm{tr}(\boldsymbol{M}^\top \boldsymbol{M})$.

Stage 2: Operate a pre-image

Since the model (5) is valid for all the training data, we apply it to do the pre-image, as discussed in this stage. Let $\varphi^*(\cdot)$ be any optimal function resulting from a kernel-based machine, with $\varphi^*(\cdot) = \sum_{i=1}^n \gamma_i\, \kappa(\cdot, \boldsymbol{x}_i)$ as given in (3). By virtue of the representer theorem, it belongs to the subspace spanned by the training kernel functions, and therefore can be expressed in terms of the provided coordinate system. The coordinate of $\varphi^*(\cdot)$ associated to the coordinate function $\psi_k(\cdot)$ is

$$\langle \varphi^*(\cdot), \psi_k(\cdot) \rangle_{\mathcal{H}} = \sum_{i,j=1}^{n} \alpha_{k,i}\, \gamma_j\, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Each of these ℓ coordinates are computed and collected into one vector, denoted Ψ_{φ^*} with some abuse of notation. Thus, we extend the model (5), and write

$$\Psi_{\boldsymbol{x}_i}^\top \Psi_{\varphi^*} = \boldsymbol{x}_i^\top \boldsymbol{x}^*,$$

for $i = 1, 2, \ldots, n$, where $\boldsymbol{x}^*$ is the pre-image to be estimated. This identity can be expressed matrix-wise with

$$\boldsymbol{K}\, \widehat{\boldsymbol{A}}^\top\, \widehat{\boldsymbol{A}}\, \boldsymbol{K}\boldsymbol{\gamma} = \boldsymbol{X}^\top \boldsymbol{x}^*$$

where $\boldsymbol{\gamma} = [\gamma_1\ \gamma_2\ \cdots\ \gamma_n]^\top$ and $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]$. By injecting the provided system (6) into this expression, we get

$$\boldsymbol{X}^\top \boldsymbol{x}^* = (\boldsymbol{P} - \lambda\, \boldsymbol{K}^{-1})\, \boldsymbol{\gamma}. \tag{7}$$

This is a classical system of linear equations. Thus, the pre-image can be estimated by applying any off-the-shelf solver. For instance, one can solve the linear least-squares optimization problem

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \|\boldsymbol{X}^\top \boldsymbol{x} - (\boldsymbol{P} - \lambda\, \boldsymbol{K}^{-1})\, \boldsymbol{\gamma}\|^2, \tag{8}$$

where any iterative or non-iterative technique can be used, such as the pseudo-inverse or the eigen-decomposition[3], in the spirit of the Nyström method. It is worth noting that the optimization scheme is applied here to the input space, as opposed to high dimensional RKHS with the fixed-point iteration schemes. Moreover, one needs only to consider solution from the span of the training data, in coherence with previous work on the pre-image problem [14,11]. The proposed method is universal in the sense of being independent, in its formulation, of both the type of the adopted kernel and of the feature under investigation.

[3] Doing eigen-decomposition gives the pre-image relative to the eigen-basis in the input space. A post-processing is required to set the pre-image relative to the training data; this is called the procrustes problem.

In order to better understand this result, consider the potential theoretical setting of linear independent training data. In this case, the minimization problem (8) has a unique solution, given by solving the normal equations $\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{x}^* = \boldsymbol{X}\big(\boldsymbol{P} - \lambda\, \boldsymbol{K}^{-1}\big)\, \boldsymbol{\gamma}$. By using the pseudo-inverse matrix algebra with the identity $(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, we get

$$\boldsymbol{x}^* = \boldsymbol{X}\boldsymbol{P}^{-1}\big(\boldsymbol{P} - \lambda\, \boldsymbol{K}^{-1}\big)\, \boldsymbol{\gamma}. \tag{9}$$

Extension to a set of features

These expressions can be applied readily to a set of features in the RKHS to get their pre-images in the input space. This can be done straightforwardly by writing (7) as

$$\boldsymbol{X}^\top \boldsymbol{X}^* = (\boldsymbol{P} - \lambda\, \boldsymbol{K}^{-1})\, \boldsymbol{\Gamma},$$

where each column of matrix $\boldsymbol{\Gamma}$ represents the coefficient vector $\boldsymbol{\gamma}$, and each column of $\boldsymbol{X}^*$ the corresponding pre-image. From the solution (9), we see that the matrix

$$\boldsymbol{M} = \boldsymbol{X}\boldsymbol{P}^{-1}\big(\boldsymbol{P} - \lambda\, \boldsymbol{K}^{-1}\big)$$

needs to be computed only once, and then applied with

$$\boldsymbol{X}^* = \boldsymbol{M}\, \boldsymbol{\Gamma}.$$

This corresponds to a matrix completion scheme, or more specifically the kernel matrix regression approach, as given in [21,9].

5 Experiments

In this section, we compare the proposed method with two state-of-the art methods[4]: the fixed-point iterative technique [14] and the MDS-based approach [11]. For this purpose, the kernel-PCA for denoising is applied on synthetic and real datasets. The Gaussian kernel $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 / 2\sigma^2)$ is used, its bandwidth σ fixed to the same value for the three methods.

5.1 Synthetic datasets

We consider a family of four datasets in 2-D, each having a geometric form corrupted by a noise of bandwidth parameter ν. The data are sampled uniformly randomly within this area. We generate n_{train} data to train the

[4] Matlab codes for these algorithms are available from the Statistical Pattern Recognition Toolbox `http://cmp.felk.cvut.cz/cmp/software/stprtool/`

Table 2 Values of the parameters for the synthetic datasets.

	n_{train}	$n_{\text{pre-image}}$	ν	n_{eigen}	σ
frame	350	850	0.1	5	0.4
banana	300	200	0.2	3	0.5
spiral	70	250	0.3	10	0.3
sine	420	330	0.5	10	0.4

n_{eigen} eigenfunctions and to construct the coordinate system. Then, we apply these results on another set of $n_{\text{pre-image}}$ generated data, in order to denoise using the pre-image techniques. For each dataset, the parameters values are summarized in Table 2.

The frame dataset consists of a square with sides of length 2. Data are generated uniformly randomly on each side and corrupted by a noise uniformly distributed on the interval $[-\nu, \nu]$ normal to the side. The banana dataset is given by the parabola defined by the coordinates $(x, x^2 + \xi)$, with x on the x-axis uniformly distributed on the interval $[-1, 1]$, and ξ normally distributed with a standard deviation of ν. The spiral is defined by the coordinates $(A(\varphi)\cos(\varphi), A(\varphi)\sin(\varphi))$, with $A(\varphi) = 0.07\varphi + \xi$, where φ and ξ are generated uniformly on the intervals $[0, 6\pi]$ and $[0, \nu]$, respectively. The sine dataset is defined by the coordinates $(\varphi, 0.8\sin(2\varphi))$, where φ is generated uniformly on the interval $[0, 2\pi]$, and corrupted with an additive uniformly distributed noise in the range $[0, \nu]^2$. See [7] for more information.

The fixed-point iterative algorithm is set with a stopping criterion of maximum 100 iterations, reaching the limit of reasonable cpu time. The initial estimate is chosen from the valid model $\boldsymbol{x}^* = \sum_i \gamma_i \boldsymbol{x}_i$, with the weighting coefficients γ_i generated uniformly on the interval $[-1, 1]$. The MDS-based algorithm operates using a global optimization scheme, which gives better results than the neighborhood setting. Since this algorithm is based on an eigen-decomposition technique, it results in a new coordinate system in the input space. Hence, we consider a procrustes technique to align it with the initial canonical one, by minimizing the mean-squares error.

In Figure 2, we show the four datasets, with on the one hand the training data (blue dots), and on the other the denoised estimates (red dots) obtained from another set of noisy data (not shown here, yet given by the unmarked ends of green lines). Green lines show the distance between the denoised and the initial noisy data.

The fixed-point iterative method suffers on one side from numerical instabilities, illustrated through many estimates falling outside the bounds of the images (following the long green lines), and on the other from local minima, illustrated with improper denoising (for instance, the upper border of the frame dataset (y-axis close to 1) are not denoised to the same area). It is obvious that the MDS-based approach is clearly inappropriate to any of the given datasets. The method presented in this paper gives good results with the four proposed datasets, with the smallest reconstruction error of all algorithms. It seems less sharper in denoising than the fixed-point iterative algorithm, without suffering from the drawbacks of the latter. However, it causes the estimates to fold over itself, in the same sense of manifold learning. This is illustrated for instance with the banana data, yet much less pronounced than the MDS-based results.

5.2 Real datasets

We illustrated the efficiency of our method on denoising real datasets. We consider the handwritten digit "2", obtained from the MNIST database of handwritten digits [12]. The images are (almost) binary images of 28-by-28 pixels. Hence, from a machine learning point of view, each image is simply a point in the 784-dimensional space. The original images were corrupted by adding a zero-mean white Gaussian noise with variance $\nu = 0.1$. A set of $n_{\text{train}} = 1\,000$ images are used to train the kernel-PCA with the $n_{\text{eigen}} = 100$ leading principal functions retained. We apply the Gaussian kernel to all three algorithms, with bandwidth set to $\sigma = 10^5$. The parameter settings are summarized in Table 3.

To illustrate the denoised ability of each algorithm, another set of $n_{\text{pre-image}} = 10$ images is considered under the same noise conditions. These images are illustrated in Figure 3 (first row), with results from the fixed-point iterative (second row), the MDS-based (third row) and the proposed (fourth row) algorithms. It is obvious that fixed-point iterative algorithm is inappropriate for such application, even with the number of maximum iterations set to 10 000 corresponding to an average total CPU time of up to 1 hour and a half. To take advantage of prior knowledge, the same training set is used for learning the inverse map. Realistic results can be obtained using the MDS-based algorithm, with 5 minutes and a half. The algorithm proposed in this paper achieves better denoised results, as illustrated in Figure 3. For this simulation, the regularization parameter was set to $\lambda = 10^{-9}$, and the resulting average total CPU time is 1.3 seconds[5].

[5] CPU times are given only as an indication of the computations required for the various algorithms.

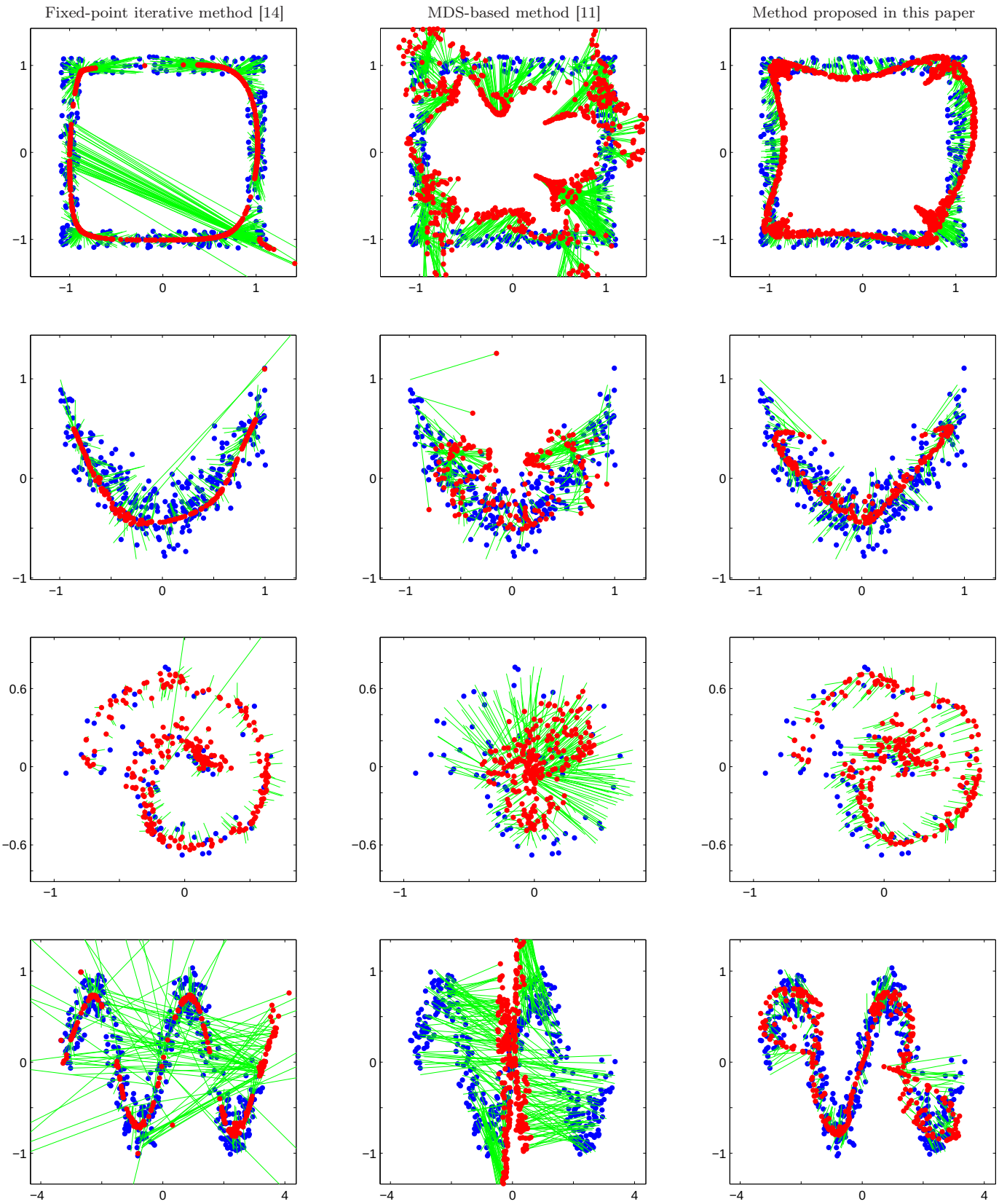

Fig. 2 Experimental results for the frame (first row), the banana (second row), the spiral (third row), and the sine (fourth row) datasets, using the fixed-point iterative (left), the MDS-based (middle), and the proposed (right) algorithms. Training data are represented by blue dots, estimated pre-images by red dots, and green lines illustrate the distance between these estimates and the initial noisy data (not shown).

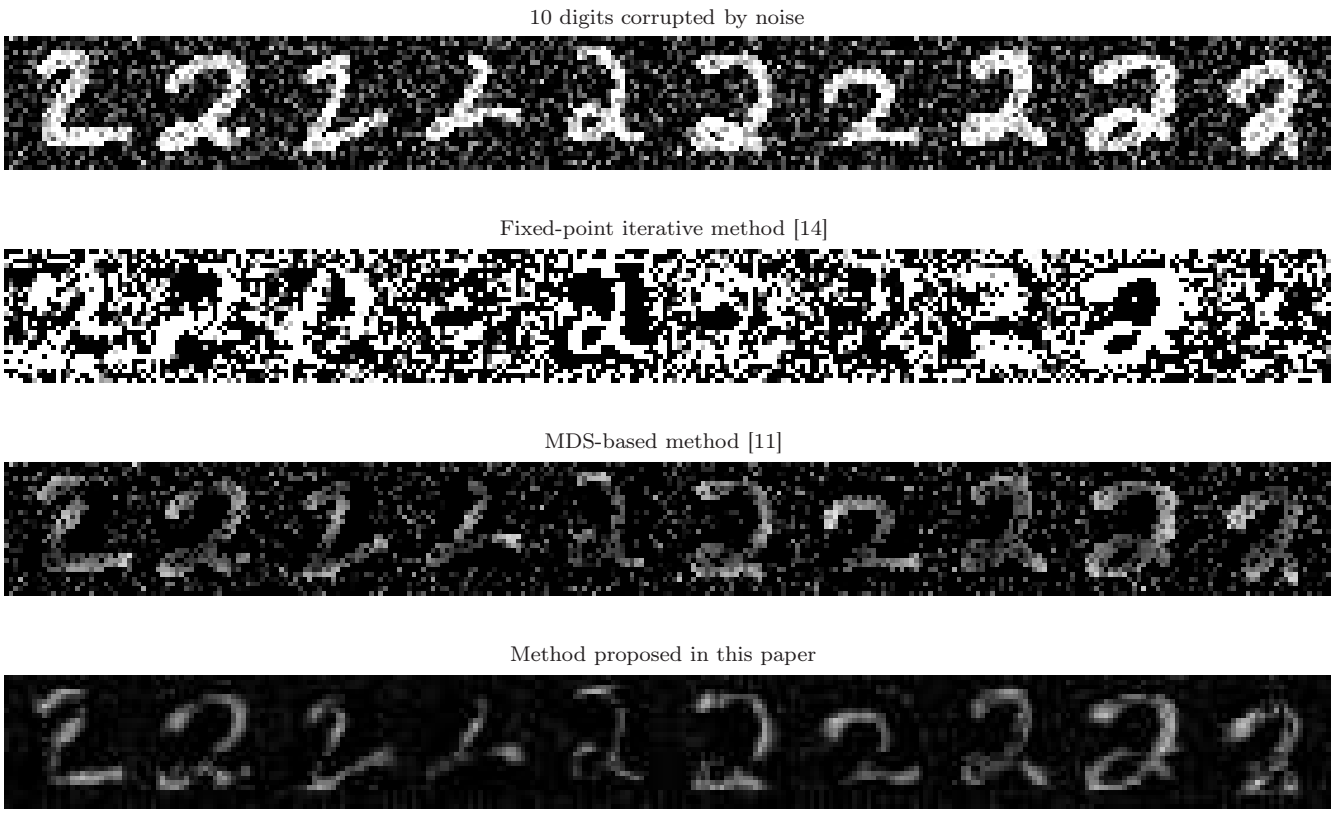

Fig. 3 Comparative analysis for denoising a set of ten "2" digits (first row), with the denoised images from the fixed-point iterative (second row), the MDS-based (third row), and the proposed (fourth row) algorithms.

Table 3 Values of the parameters for the real digit dataset.

n_{train}	$n_{\text{pre-image}}$	ν	n_{eigen}	σ
1 000	10	0.1	100	10^5

6 Conclusion

In this paper, we presented a new method to solve the pre-image problem. As opposed to previous work, the proposed method neither suffers from numerical instability, nor requires computing the distances in the input and the RKHS spaces. We showed that using the inner product information in both spaces, we can provide a coordinate system in the RKHS to learn the inverse map. The efficiency of the proposed method were studied with experiments on both synthetic data and real handwritten digits, and compared to state-of-the-art methods. The major advantage of the proposed method resides on its simplicity in dealing with the optimization issue, thanks to conventional linear algebra.

References

1. Aizerman, M., Braverman, E., Rozonoer, L.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control **25**, 821–837 (1964)
2. Arias, P., Randall, G., Sapiro, G.: Connecting the out-of-sample and pre-image problems in kernel methods. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2007). URL http://ampere.iie.edu.uy/publicaciones/2007/ARS07
3. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337–404 (1950)
4. Bakir, G., Weston, J., Schölkopf, B.: Learning to find pre-images. In: Thrun S., L.S., Schölkopf, B. (eds.) NIPS 2003, vol. 16, pp. 449–456. MIT Press, Cambridge, MA, USA (2004)
5. Bengio, Y., Paiement, J., Vincent, P., Delalleau, O., Roux, N.L., Ouimet, M.: Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA (2004)
6. Etyngier, P., SÈgonne, F., Keriven, R.: Shape priors using manifold learning techniques. In: 11th IEEE Interna-

tional Conference on Computer Vision. Rio de Janeiro, Brazil (2007)
7. Hoffmann, H.: Kernel PCA for novelty detection. Pattern Recognition **40**, 863–874 (2007)
8. Honeine, P., Richard, C.: Solving the pre-image problem in kernel machines: a direct method. In: IEEE Workshop on Machine Learning for Signal Processing. Grenoble, France (2009)
9. Honeine, P., Richard, C., Essoloh, M., Snoussi, H.: Localization in sensor networks - a matrix regression approach. In: 5^{th} IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM). Darmstadt, Germany (2008)
10. Kimeldorf, G., Wahba, G.: Some results on tchebycheffian spline functions. Journal of Mathematical Analysis and Applications **33**, 82–95 (1971)
11. Kwok, J.T., Tsang, I.W.: The pre-image problem in kernel methods. In: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), pp. 408–415. AAAI Press, Washington, DC, USA (2003)
12. Lecun, Y., Cortes, C.: The mnist database of handwritten digits (1998). URL http://yann.lecun.com/exdb/mnist/
13. Mika, S.: Kernel fisher discriminants. Phd thesis, University of Technology, Berlin (2002)
14. Mika, S., Schölkopf, B., Smola, A., Müller, K., Scholz, M., Rätsch, G.: Kernel pca and de-noising in feature spaces. In: Proceedings of the 1998 conference on advances in neural information processing systems II, pp. 536–542. MIT Press, Cambridge, MA, USA (1999)
15. Rathi, Y., Dambreville, S., Tannenbaum, A.: Statistical shape analysis using kernel pca. In: IS&T/SPIE Symposium on Electronic Imaging (2006)
16. Schölkopf, B., Herbrich, R., Williamson, R.: A generalized representer theorem. Tech. Rep. NC2-TR-2000-81, Royal Holloway College, Univ. of London, UK (2000)
17. Schölkopf, B., Smola, A., Müller, K.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10**(5), 1299–1319 (1998)
18. Suykens, J., Gestel, T.V., Brabanter, J.D., Moor, B.D., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific Pub. Co., Singapore (2002)
19. Tax, D.: One-class classification; concept-learning in the absence of counter-examples. Phd thesis, Advanced School for Computing and Imaging – Delft University of Technology (2001)
20. Vapnik, V.: Statistical Learning Theory. Wiley, New York, NY, USA (1998)
21. Yamanishi, Y., Vert, J.P.: Kernel matrix regression. Tech. Rep. http://arxiv.org/abs/q-bio/0702054v1 (2007)