Input Space Regularization Stabilizes Pre-images for Kernel PCA De-noising

Abrahamsen, Trine Julie; Hansen, Lars Kai

Published in:
IEEE International Workshop on Machine Learning for Signal Processing, 2009. MLSP 2009.

Link to article, DOI:
10.1109/MLSP.2009.5306191

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Abrahamsen, T. J., & Hansen, L. K. (2009). Input Space Regularization Stabilizes Pre-images for Kernel PCA De-noising. In IEEE International Workshop on Machine Learning for Signal Processing, 2009. MLSP 2009. (pp. 1-6). IEEE. DOI: 10.1109/MLSP.2009.5306191

INPUT SPACE REGULARIZATION STABILIZES PRE-IMAGES FOR KERNEL PCA DE-NOISING

Trine Julie Abrahamsen   *Lars Kai Hansen*

DTU Informatics
Technical University of Denmark
DK-2800 Kgs. Lyngby, DENMARK

ABSTRACT

Solution of the pre-image problem is key to efficient non-linear de-noising using kernel Principal Component Analysis. Pre-image estimation is inherently ill-posed for typical kernels used in applications and consequently the most widely used estimation schemes lack stability. For de-noising applications we propose input space distance regularization as a stabilizer for pre-image estimation. We perform extensive experiments on the USPS digit modeling problem to evaluate the stability of three widely used pre-image estimators. We show that the previous methods lack stability when the feature mapping is non-linear, however, by applying a simple input space distance regularizer we can reduce variability with very limited sacrifice in terms of de-noising efficiency.

Index Terms— Kernel PCA, Pre-image, De-noising

1. INTRODUCTION

We are interested in unsupervised learning methods for de-noising, i.e., in the projection of noisy or distorted observational data onto a ’clean’ signal manifold and, if necessary, we will use non-linear maps to implement the projection. Kernel PCA and similar methods are widely used candidates for such projection beyond conventional linear unsupervised learning schemes like PCA principal component analysis, ICA independent component analysis, and NMF non-negative matrix factorization. The basic idea is to implement the projection in three steps, in the first step we map the original data referred to as in input space, into a feature space and then in the second step we use a conventional linear algorithm, like PCA, to identify the signal manifold by linear projection in feature space. Finally, in the third step we estimate the input space - de-noised - points that best correspond to the projected feature space points. The latter step is referred to as the *pre-image problem*. Unfortunately, finding a reliable pre-image is entirely non-trivial and has given rise to several algorithms [1, 2, 3, 4]. *In this work we analyze the stability of the estimated pre-images from the most used of these algorithms, we suggest a new regularizer appropriate for de-noising applications, and we show that the new pre-image algorithm improves the stability relative to the existing approaches.*

To understand the pre-image problem, let us recapitulate some basic aspects of de-noising with kernel PCA. Let $\mathcal{F}$ be the Reproducing Kernel Hilbert Space (RKHS) associated with the kernel function $k(\mathbf{x},\mathbf{x}') = \varphi(\mathbf{x})^T\varphi(\mathbf{x}')$, where $\varphi : \mathcal{X} \mapsto \mathcal{F}$ is a possibly nonlinear map from the D-dimensional input space $\mathcal{X}$ to the high dimensional (possibly infinite) feature space $\mathcal{F}$ (see notation[1]). In de-noising and a number of other applications it is of interest to reconstruct a data point in input space from a point in feature space, i.e. applying the inverse map of φ. As mentioned, in de-noising by kernel PCA we map a noisy input point $\mathbf{x}$ into feature space, $\varphi(\mathbf{x}) \in \mathcal{F}$, project it onto q principal components in feature space giving $P_q\varphi(\mathbf{x})$. By mapping the projection back into input space a new and hopefully less noisy point $\mathbf{z} = \varphi^{-1}(P_q\varphi(\mathbf{x}))$ is obtained. Given a point in feature space Ψ, the pre-image problem thus consists of finding a point $\mathbf{z} \in \mathcal{X}$ in the input space such that $\varphi(\mathbf{z}) = \Psi$. $\mathbf{z}$ is then called the pre-image of Ψ. For many non-linear kernels dim$(\mathcal{F}) \gg$ dim$(\mathcal{X})$ and φ is not surjective. Furthermore, whether φ is injective depends on the choice of kernel function. As a function $f : X \mapsto Y$ has an inverse iff it is bijective, we do not expect φ to have an inverse. When φ is not surjective, it follows that not all points in $\mathcal{F}$ or even the span of $\{\varphi(\mathcal{X})\}$ is the image of some $\mathbf{x} \in \mathcal{X}$. Finally, when φ is not injective, any recovered pre-image might not be unique. Thus the pre-image problem is ill-posed [1, 2, 3, 4, 5, 6, 7]. As we can not expect an exact pre-image, we follow [1] and relax the quest to find an *approximate pre-image*, i.e., a point in input space which maps into a point in feature space ‘as close as possible’ to Ψ .

2. KERNEL PCA

Kernel Principal Component Analysis is a nonlinear generalization of linear PCA, in which PCA is carried out in the fea-

This work is supported in part by the Lundbeckfonden through the Center for Integrative Molecular Brain Imaging - CIMBI www.cimbi.org

[1] Bold uppercase letters denote matrices, bold lowercase letters represent column vectors, and non-bold letters denote scalars. $\mathbf{a}_j$ denotes the *j’th* column of $\mathbf{A}$, while a_{ij} denotes the scalar in the *i’th* row and *j’th* column of $\mathbf{A}$. Finally $\mathbf{1}_{NN}$ is a $N \times N$ matrix of ones

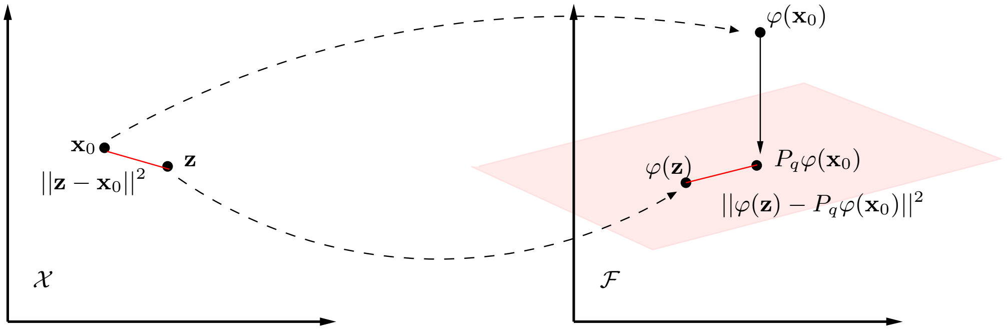
$\mathbf{x}_0$
$\mathbf{z}$
$||\mathbf{z}-\mathbf{x}_0||^2$
$\mathcal{X}$
$\varphi(\mathbf{x}_0)$
$\varphi(\mathbf{z})$
$P_q\varphi(\mathbf{x}_0)$
$||\varphi(\mathbf{z})-P_q\varphi(\mathbf{x}_0)||^2$
$\mathcal{F}$

Fig. 1. The pre-image problem in kernel PCA de-noising concerns estimating $\mathbf{z}$ from $\mathbf{x}_0$, through the projection of the image onto the principal subspace. Presently available methods for pre-image estimation lead to unstable pre-images because the inverse is ill-posed. We show that simple input space regularization, with a penalty based on the distance $||\mathbf{z}-\mathbf{x}_0||$ leads to a stable pre-image.

ture space $\mathcal{F}$ mapped data [8]. However, as $\mathcal{F}$ can be infinite dimensional we can not work directly with the feature space covariance matrix. Fortunately, the so-called kernel trick allows us to formulate nonlinear extensions of linear algorithms when these are expressed in terms of inner-products.

Let $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ be N training data points in $\mathcal{X}$ and $\{\varphi(\mathbf{x}_1),\ldots,\varphi(\mathbf{x}_N)\}$ be the corresponding images in $\mathcal{F}$. The mean of the φ-mapped data points is denoted $\bar{\varphi} = \frac{1}{N}\sum_{n=1}^{N}\varphi(\mathbf{x}_n)$ and the 'centered' images are given by $\tilde{\varphi}(\mathbf{x}) = \varphi(\mathbf{x}) - \bar{\varphi}$. Now kernel PCA is performed by solving the eigenvalue problem

$$\widetilde{\mathbf{K}}\boldsymbol{\alpha}_i = \lambda_i\boldsymbol{\alpha}_i \tag{1}$$

where $\widetilde{\mathbf{K}}$ is the centered kernel matrix defined as $\widetilde{\mathbf{K}} = \mathbf{K} - \frac{1}{N}\mathbf{1}_{NN}\mathbf{K} - \frac{1}{N}\mathbf{K}\mathbf{1}_{NN} + \frac{1}{N^2}\mathbf{1}_{NN}\mathbf{K}\mathbf{1}_{NN}$.

The projection of a φ-mapped test point onto the *i'th* component is

$$\beta_i = \tilde{\varphi}(\mathbf{x})^T\mathbf{v}_i = \sum_{n=1}^{N}\alpha_{in}\tilde{\varphi}(\mathbf{x})^T\tilde{\varphi}(\mathbf{x}_n) = \sum_{n=1}^{N}\alpha_{in}\tilde{k}(\mathbf{x},\mathbf{x}_n) \tag{2}$$

where $\mathbf{v}_i$ is the *i'th* eigenvector of the feature space covariance matrix and the $\boldsymbol{\alpha}_i$'s have been normalized. The centered kernel function can be found as $\tilde{k}(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - \frac{1}{N}\mathbf{1}_{1\times N}\mathbf{k}_{\mathbf{x}} - \frac{1}{N}\mathbf{1}_{1\times N}\mathbf{k}_{\mathbf{x}'} + \frac{1}{N^2}\mathbf{1}_{1\times N}\mathbf{K}\mathbf{1}_{N\times 1}$, where $\mathbf{k}_{\mathbf{x}} = [k(\mathbf{x},\mathbf{x}_1),\ldots,k(\mathbf{x},\mathbf{x}_N)]^T$. The projection of $\varphi(\mathbf{x})$ onto the subspace spanned by the first q eigenvectors will be denoted $P_q\varphi(\mathbf{x})$ and can be found as

$$\begin{aligned} P_q\varphi(\mathbf{x}) &= \sum_{i=1}^{q}\beta_i\mathbf{v}_i + \bar{\varphi} = \sum_{i=1}^{q}\beta_i\sum_{n=1}^{N}\alpha_{in}\tilde{\varphi}(\mathbf{x}_n) + \bar{\varphi} \\ &= \sum_{n=1}^{N}\tilde{\gamma}_n\tilde{\varphi}(\mathbf{x}_n) + \bar{\varphi} \end{aligned} \tag{3}$$

where $\tilde{\gamma}_n = \sum_{i=1}^{q}\beta_i\alpha_{in}$. Kernel PCA satisfies properties similar to those for linear PCA, namely that the squared reconstruction error is minimal and the retained variance is maximal. However, these proporties holds in $\mathcal{F}$ not $\mathcal{X}$.

3. APPROXIMATE PRE-IMAGES

Several optimality criteria can be used for the pre-image approximation, see e.g., [5],

$$\text{Distance:} \qquad \mathbf{z} = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{X}} ||\varphi(\mathbf{z}) - \Psi||^2 \tag{4}$$

$$\text{Co-linearity:} \qquad \mathbf{z} = \operatorname*{argmax}_{\mathbf{z}\in\mathcal{X}} \left\langle \frac{\varphi(\mathbf{z})}{||\varphi(\mathbf{z})||}, \frac{\Psi}{||\Psi||} \right\rangle \tag{5}$$

For RBF kernels of the form $k(\mathbf{x}_i,\mathbf{x}_j) = \kappa(||\mathbf{x}_i - \mathbf{x}_j||^2)$ the co-linearity criteria and the distance criteria coincide

$$\begin{aligned} ||\varphi(\mathbf{z}) - \Psi||^2 &= \langle\varphi(\mathbf{z}),\varphi(\mathbf{z})\rangle + \langle\Psi,\Psi\rangle - 2\langle\varphi(\mathbf{z}),\Psi\rangle \\ &= k(\mathbf{z},\mathbf{z}) + ||\Psi||^2 - 2\langle\varphi(\mathbf{z}),\Psi\rangle \end{aligned} \tag{6}$$

As $k(\mathbf{z},\mathbf{z})$ is constant for RBF kernels and $||\Psi||^2$ is independent of $\mathbf{z}$, minimizing $||\varphi(\mathbf{z}) - \Psi||^2$ is equivalent to maximizing the co-linearity. As $\mathcal{F}$ is a RKHS, the distance will be the same before and after centering. However, the expression gets a bit more tedious when using explicit centering as will be shown later, even though the result is the same: Minimizing the distance is identical to maximizing the inner-product.

Thus we seek to minimize the distance between $\varphi(\mathbf{z})$ and Ψ w.r.t $\mathbf{z}$. When it is assumed that Ψ lies in (or close to) the span of $\{\varphi(\mathbf{x}_i)\}$, Ψ can be represented as a linear combination of the training images, i.e. $P_q\varphi(\mathbf{x})$, without loss of generality. When $q = N$ this will translate to projecting Ψ onto the span of $\{\varphi(\mathbf{x}_i)\}$. We are interested in an expression for

$$\begin{aligned} ||\varphi(\mathbf{z}) - P_q\varphi(\mathbf{x})||^2 = ||\varphi(\mathbf{z})||^2 + ||P_q\varphi(\mathbf{x})||^2 \\ - 2\varphi(\mathbf{z})^T P_q\varphi(\mathbf{x}). \end{aligned} \tag{7}$$

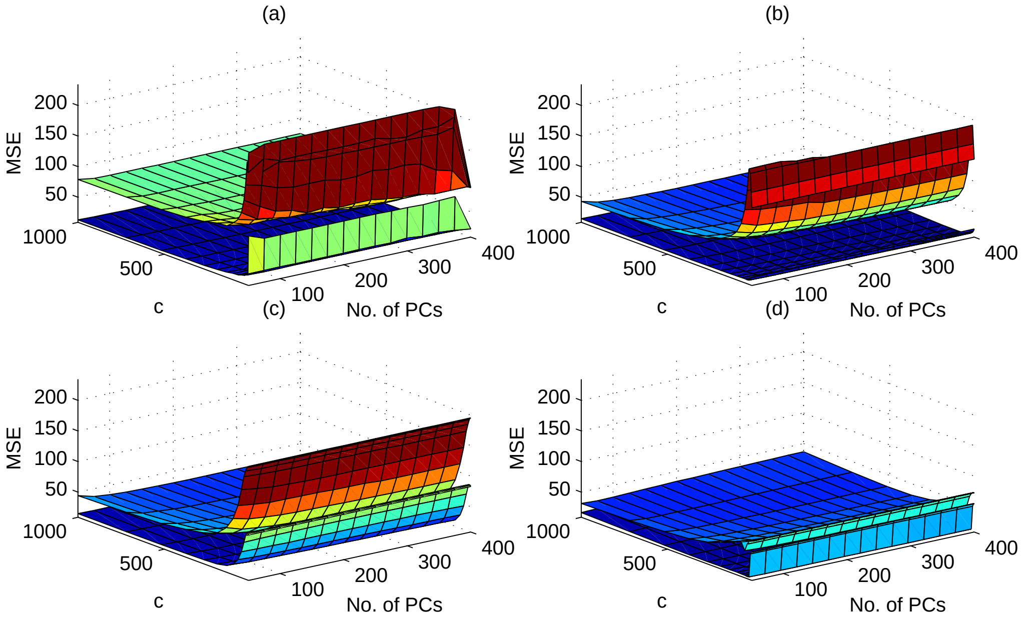

Fig. 2. Experiment to illustrate the stability of pre-image based de-noising of USPS digits. A training set of 400 digits ($100@0, 2, 4, 9$) is used to define the signal manifold. We show the confidence intervals (5th and the 95th percentile) for the mean square error (MSE) in different combinations of kPCA subspace dimension and non-linearity. MSE computed for 400 de-noised test samples for (a) Kwok-Tsang, (b) Mika et al., (c) Dambreville et al., and (d) the new input space distance regularization approach. The previous schemes are seen to deteriorate in the non-linear regime (small c).

The terms will in the following be expanded separately, starting with the first term

$$||\varphi(\mathbf{z})||^2 = \varphi(\mathbf{z})^T\varphi(\mathbf{z}) = k(\mathbf{z},\mathbf{z}) \tag{8}$$

From (3) and the definition of centering and mean in feature space, we have

$$\begin{aligned}||P_q\varphi(\mathbf{x})||^2 &= \left(\sum_{i=1}^{q}\beta_i\mathbf{v}_i + \bar{\varphi}\right)^T\left(\sum_{i=1}^{q}\beta_i\mathbf{v}_i + \bar{\varphi}\right)\\ &= \sum_{i=1}^{q}\beta_i^2 + \bar{\varphi}^T\bar{\varphi} + 2\bar{\varphi}^T\sum_{n=1}^{N}\tilde{\gamma}_n\tilde{\varphi}(\mathbf{x}_n)\\ &= \sum_{i=1}^{q}\left(\sum_{n=1}^{N}\alpha_{in}\tilde{k}(\mathbf{x},\mathbf{x}_n)\right)^2 + \frac{1}{N^2}\sum_{n,m=1}^{N}k(\mathbf{x}_n,\mathbf{x}_m)\\ &+ \frac{2}{N}\sum_{n=1}^{N}\left(\tilde{\gamma}_n\sum_{m=1}^{N}k(\mathbf{x}_m,\mathbf{x}_n) - \frac{\tilde{\gamma}_n}{N}\sum_{m,l=1}^{N}k(\mathbf{x}_m,\mathbf{x}_l)\right)\end{aligned} \tag{9}$$

Finally the last term can be expanded using the same properties as above

$$\begin{aligned}\varphi(\mathbf{z})^TP_q\varphi(\mathbf{x}) &= \varphi(\mathbf{z})^T\left(\sum_{n=1}^{N}\tilde{\gamma}_n(\varphi(\mathbf{x}_n)-\bar{\varphi}) + \bar{\varphi}\right)\\ &= \sum_{n=1}^{N}\gamma_nk(\mathbf{z},\mathbf{x}_n)\end{aligned} \tag{10}$$

Where the last equality follows from letting $\gamma_n = \tilde{\gamma}_n + \frac{1}{N}(1 - \sum_{j=1}^{N}\tilde{\gamma}_j)$, where $\tilde{\gamma}_n = \sum_{i=1}^{q}\beta_i\alpha_{in}$ as defined in equation (3). Now combining the expressions gives

$$\begin{aligned}&||\varphi(\mathbf{z}) - P_q\varphi(\mathbf{x})||^2\\ &\quad= k(\mathbf{z},\mathbf{z}) - 2\sum_{n=1}^{N}\gamma_nk(\mathbf{z},\mathbf{x}_n) + \Omega\end{aligned} \tag{11}$$

where all the $\mathbf{z}$-independent terms originating from $||P_q\varphi(\mathbf{x})||^2$ have been collected in Ω.

3.1. Overview of existing algorithms

The non-linear optimization problem associated with finding the pre-image has been approached in a variety of ways. In

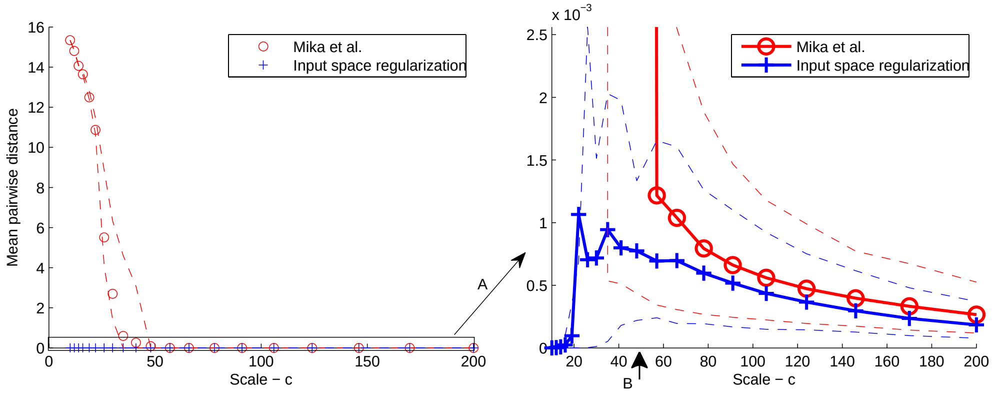

Fig. 3. The mean pairwise distances (mean, 5th and the 95th percentiles) for Mika et al. (red) and the new input space regularization approach (blue). We use 300 principal components in this study. The previous approach fails to provide a stable pre-image in the non-linear regime (small c). The right panel shows the box in the left panel, whereas arrow 'B' indicates the scale used in Figure 4.

the original work [1] and [2] proposed a fixed-point iteration method. It is a noted drawback of this method that it can be numerically unstable, sensitive to the initial starting point, and converge to a local extremum. To overcome this problem a more 'direct' approach was taken in [3], where the relationship between distance measures in feature space and input space as well as the idea of multidimensional scaling (MDS) were combined to produce a non-iterative constructive solution. These are the two approaches most widely used in applications. However, several modifications have already been proposed. In order to overcome possible numerical instabilities of the fixed-point approach, various ways of initializing the fixed-point iteration scheme have been suggested. The algorithm can be started in a 'random' input space point, but this can lead to slow convergence in real-life problems, since the cost-function can be very flat in regions away from data. Alternatively, for de-noising applications, it can be initialized in the point in input space, which we seek to de-noise. However, according to [9] this strategy will only work if the signal-to-noise ratio (SNR) is high. Instead [10] suggested to initialize the fixed-point iteration scheme in the solution found by the distance method in [3]. Later it was claimed that a more efficient starting point would be the mean of a certain number of neighbors of the point to be de-noised [11]. In [4] a modification of the method developed in [1], utilizing feature space distances was proposed. This method also minimizes the distance constraint in (4), but does so in a non-iterative approximation thereby avoiding numerical instabilities. In [12] kernel ridge regression was used to learn some inverse mapping of φ. While the formulation in [12] is in very general terms, the actual implementation is similar to [3]. The main issue is that we typically only have indirect access to feature space points, thus a learned pre-image needs to be formulated in terms of distances as in [3], rather than explicit input-output examples. It should be noted that with the relative general formulation the method of [12] in some cases can be applied beyond [3], e.g., to non-Euclidean input spaces. In lieu of the recognized ill-posed nature of the inverse problem attemps of more robust estimators have been pursued, in [13] a regularization was introduced that penalized the projection in feature space, while in [14] a ridge regression regularizer was used for the weights of a learned pre-image estimator as originally proposed in [12].

Returning to the iterative scheme of Mika et al., we work, as in most applications, with RBF kernels for which $k(\mathbf{z},\mathbf{z})$ is constant for all $\mathbf{z}$, hence minimizing the squared distance in (11) is identical to

$$\max_{\mathbf{z}} 2\sum_{n=1}^{N} \gamma_n k(\mathbf{z},\mathbf{x}_n) \tag{12}$$

Now in extrema of (12) the derivative with respect to $\mathbf{z}$ is zero, which leads to the following fixed-point iteration for a Gaussian kernel of the form $k(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{1}{c}||\mathbf{x}-\mathbf{x}'||^2\right)$ [1]

$$\mathbf{z}_{t+1} = \frac{\sum_{n=1}^{N} \gamma_n \exp(-||\mathbf{z}_t - \mathbf{x}_n||^2/c)\mathbf{x}_n}{\sum_{n=1}^{N} \gamma_n \exp(-||\mathbf{z}_t - \mathbf{x}_n||^2/c)} \tag{13}$$

As mentioned maximizing equation (12) is a non-linear optimization problem, and hence suffers from convergence to local minima and strong sensitivity to the initial point $\mathbf{z}$. As we shall see, this implies that the solutions are at times highly unstable.

3.2. The input space regularization approach

In-order to provide a more stable estimate of the pre-image we propose to augment the cost function with an input space distance penalty term (see Figure 1)

$$\begin{aligned}\rho_1(\mathbf{z}) &= ||\varphi(\mathbf{z}) - P_q\varphi(\mathbf{x})||^2 + \lambda||\mathbf{z} - \mathbf{x}_0||^2 \\ &= k(\mathbf{z},\mathbf{z}) - 2\sum_{n=1}^{N}\gamma_n k(\mathbf{z},\mathbf{x}_n) + \Omega \\ &+ \lambda(\mathbf{z}^T\mathbf{z} + \mathbf{x}_0^T\mathbf{x}_0 - 2\mathbf{z}\mathbf{x}_0) \qquad (14)\end{aligned}$$

λ is a non-negative regularization parameter and $\mathbf{x}_0$ is the noisy observation in $\mathcal{X}$. The main rationale is that among the solutions to the non-linear optimization problem we want the pre-image which is closest to the noisy input point, hence, hopefully reduce possible distortions of the signal. Thus we seek to minimize

$$\rho_2(\mathbf{z}) = k(\mathbf{z},\mathbf{z}) - 2\sum_{n=1}^{N}\gamma_n k(\mathbf{z},\mathbf{x}_n) + \lambda(\mathbf{z}^T\mathbf{z} - 2\mathbf{z}\mathbf{x}_0) \quad (15)$$

ignoring all $\mathbf{z}$-independent terms. This expression can be minimized for any kernel using a non-linear optimizer.

For RBF kernels the fixed-point iteration scheme can be regularized similarly, this typically leads to a faster evaluation than using an optimizer. Introducing regularization in the maximization problem given in (12) leads to the following objective function

$$\rho_3(\mathbf{z}) = 2\sum_{n=1}^{N}\gamma_n k(\mathbf{z},\mathbf{x}_n) - \lambda||\mathbf{z} - \mathbf{x}_0||^2 \quad (16)$$

which we seek to maximize w.r.t. $\mathbf{z}$. With straightforward algebra we get the regularized fixed-point iteration

$$\mathbf{z}_{t+1} = \frac{\frac{2}{c}\sum_{n=1}^{N}\gamma_n \exp\left(-\frac{1}{c}||\mathbf{z}_t - \mathbf{x}_n||^2\right)\mathbf{x}_n + \lambda\mathbf{x}_0}{\frac{2}{c}\sum_{n=1}^{N}\gamma_n \exp\left(-\frac{1}{c}||\mathbf{z}_t - \mathbf{x}_n||^2\right) + \lambda} \quad (17)$$

In this expression the denominator is given by $\frac{2}{c}\langle\varphi(\mathbf{z}_t), \Psi\rangle + \lambda$. As λ is a non-negative parameter, the denominator will always be non-zero in the neighborhood of a maximum because the inner-product will be positive in that same neighborhood.

4. EXPERIMENTS

In this section we compare the new regularized fixed-point iteration algorithm with the approaches proposed by: (a) Kwok-Tsang [3], (b) Mika et al. [1], and (c) Dambreville et al. [4]. The experiments are done on a subset of the USPS data consisting of 16×16 pixels handwritten digits[2]. For each of the digits $0, 2, 4,$ and 9 we chose 100 examples for training and another 100 examples for testing. We added gaussian noise $\mathcal{N}(0, 0.25)$ and set the regularization parameter $\lambda = 0.001$.

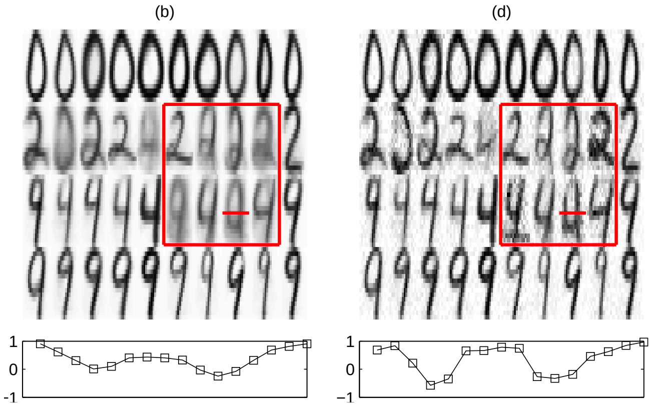

Fig. 4. Top: Example of de-noised digits using a very non-linear kernel ($c = 50$) and 100 principal components. (b) Mika et al and (d) our approach, note the visual improvement of the recovered pre-images in the red box. The colormap has been adjusted for better visualization. Bottom: The image intensity along the red line indicated above. Note the improvemed SNR in the result of the new method.

In-order to illustrate the stability and performance of the methods we vary both the number of principal components used to define the signal manifold and the scale parameter c of the Gaussian kernel. For each combination and pre-image estimator, the mean squared error (MSE) of the de-noised result for the 400 *test* examples is calculated. The iterative approaches are initialized in the noisy test point and for the Kwok-Tsang approach 10 neighbors were used for the approximation.

The results are summarized in Figure 2 where we show the lower 5th and upper 95th percentile confidence intervals for the MSE. As seen the confidence intervals blow up for the previous methods - panels (a-c) - in the non-linear regime in which the kernel has a relative small scale parameter, while the confidence interval points to a more stable de-noised solution for the new regularization based approach - as seen in panel (d).

To better understand the nature of the instability of the previous algorithms we have investigated the diversity of the solutions obtained when starting the iterative algorithms in different initial points. Specifically we compare the standard iterative solution of Mika et al. and the new regularized version. For each of the 400 test examples the two algorithms are initialized in 40 randomly chosen training examples. This leads to 40 (potentially different) pre-image solution for each test sample. We measure the stability of these solution sets as the mean pairwise distance between them 40 pre-images, and report the mean across the 400 test examples This mean and its confidence intervals are presented in Figure 3 as function of the non-linearity scale parameter c. As seen, the new method produces a stable pre-image even for very non-linear models (small c), where the un-regularized iterative scheme

[2] The USPS data set is described in [15] and can be downloaded from www.kernel-machines.org

fails to reproduce.

Finally Figure 4 shows examples of the de-noised images obtained with Mika et al.'s and the new input space regularization approach, respectively. For the images which are successfully de-noised by Mika et al.'s method, e.g., 'zeros' and 'nines', the input distance regularization has very little effect, while a clear improvement can be seen for the images for which Mika et al.'s algorithm fails to recover good visual solution, see, e.g., the red box with the blurred 'twos' and 'fours'. For these digits the input space regularization method do reconstruct the correct digit, albeit with a price paid in terms of a slightly less de-noised result. However, the image intensity, as shown in the lower part of Figure 4, clearly illustrates the increased SNR achieved by the input space regularization.

5. CONCLUSION

In this contribution we addressed the problem of pre-image instability for kernel PCA de-noising. The recognized concerns of current methods, e.g., the sensitivity to local minima and large variability was demonstrated found for the most widely used methods including Mika et al.'s iterative scheme, the Kwok-Tsang local linear approximation and the method of Dambreville et al. By introducing simple input space distance regularization in the existing pre-image approaximation cost functions, we achieved a more stable pre-image, with very little sacrifice of the de-noising ability. Experimental results on the USPS data illustrated how our method provides a more stable pre-image; both in the sense of variability between test points and by reducing the sensitivity to starting conditions, hence convergence to local minima.

We thus recommend the use of input space distance contraints as it provides a reliable pre-image in cases where current methods fail to recover a meaningful result.

In future work we aim to further improve pre-image estimation by introducing other types of regularization appropriate for specific de-noising tasks, this can, e.g., be in the form of sparsity of the sought pre-image, which would be very relevant for, e.g., digit de-noising.

6. REFERENCES

[1] S. Mika, B. Schölkopf, A. Smola, K. R. Müller, M. Scholz, and G. Rätsch, "Kernel pca and de-noising in feature spaces," in *Advances in Neural Information Processing Systems 11*. 1999, pp. 536–542, MIT Press.

[2] B. Schölkopf, A. J. Smola, P. Knirsch, and C. J. C. Burges, "Fast approximation of support vector kernel expansions, and an interpretation of clustering as approximation in feature spaces," in *Mustererkennung 1998, 20. DAGM-Symposium*. 1998a, pp. 125–132, Springer-Verlag.

[3] J. Tin-Yau Kwok and I. Wai-Hung Tsang, "The pre-image problem in kernel methods," *IEEE transactions on neural networks*, vol. 15, no. 6, pp. 1517–1525, 2004.

[4] S. Dambreville, Y. Rathi, and A. Tannenbaum, "Statistical shape analysis using kernel pca," in *IS&T/SPIE Symposium on Electrical Imaging*, 2006.

[5] P. Arias, G. Randall, and G. Sapiro, "Connecting the out-of-sample and pre-image problems in kernel methods," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 18-23 jun 2007.

[6] B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Müller, G. Rtsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions On Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[7] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[8] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998b.

[9] T. Takahashi and T. Kurita, "Robust de-noising by kernel pca," in *ICANN '02: Proceedings of the International Conference on Artificial Neural Networks*. 2002, pp. 739–744, Springer-Verlag.

[10] K. In Kim, M. O. Franz, and B. Schölkopf, "Iterative kernel principal component analysis for image modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 9, pp. 1351–1366, 2005.

[11] A. R. Teixeira, A. M. Tomé, K. Stadlthanner, and E. W. Lang, "Kpca denoising and the pre-image problem revisited," *Digit. Signal Process.*, vol. 18, no. 4, pp. 568–580, 2008.

[12] G. H. Bakir, J. Weston, and B. Schölkopf, "Learning to find pre-images," in *Advances in Neural Information Processing Systems 16*, pp. 449–456. MIT Press, 2004.

[13] M. H. Nguyen and F. De la Torre Frade, "Robust kernel principal component analysis," in *Advances in Neural Information Processing Systems*, December 2008.

[14] Wei-Shi Zheng and Jian huang Lai, "Regularized locality preserving learning of pre-image problem in kernel principal component analysis," *Pattern Recognition, International Conference on*, vol. 2, pp. 456–459, 2006.

[15] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, 1994.