

NORMALIZATION AND PREIMAGE PROBLEM IN GAUSSIAN KERNEL PCA

Nicolas Thorstensen and Florent Segonne and Renaud Keriven

CERTIS - Ecole des Ponts
19, rue Alfred Nobel - Cité Descartes
77455 Marne-la-Vallée - France

ABSTRACT

Kernel PCA has received a lot of attention over the past years and showed useful for many image processing problems. In this paper we analyse the issue of normalization in Kernel PCA for the pre-image problem. We present a geometric interpretation of the normalization process for the gaussian kernel. As a consequence, we could formulate a correct normalization criterion in centered feature space. Furthermore, we show how the proposed normalization criterion improves previous pre-image methods for the task of image denoising.

Index Terms— Kernel PCA, Out-of-Sample, Image Denoising

1. INTRODUCTION

1.1. Kernel Methods

Kernel methods are a class of powerful techniques that have been widely used in the field of pattern recognition, with applications ranging from clustering, classification and recognition to image denoising, signal reconstruction and shape priors [1, 2]. The key idea of these methods is to map the training data (such as vectors, images, graphs, ...) from the input space χ into a high-dimensional Hilbert space \mathbb{H} that is better suited for analysis than the original input space. To do so, a mapping, denoted $\Phi^\circ : \chi \mapsto \mathbb{H}$, is implicitly defined by the property $\langle \Phi^\circ(s_i), \Phi^\circ(s_j) \rangle_{\mathbb{H}} = W_{i,j}$, where $W_{i,j} = w(s_i, s_j)$ gives the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ between two points in the feature space and is a measure for similarity. In practice, the mapping does not have to be computed explicitly as most techniques only require the computation of dot products that can be evaluated directly using the kernel $w(\cdot, \cdot)$. This is called the kernel trick.

The high-dimensional, possibly infinite-dimensional, space \mathbb{H} is better suited for analysis because data may then be processed by linear methods such as Principal Component Analysis (PCA). PCA is a widely used method to compute second order statistics in data sets. The principal axis found by PCA reflect the main modes of variation present in the data set. Kernel PCA refers to the generalization of linear PCA to its nonlinear counterpart. It was introduced by Schoelkopf [2]

and is one amongst the most prominent kernel methods. It has received a lot of attention in the data analysis and computer vision community. Using this methodology, it is possible to extract efficiently meaningful structure present in non-linear data, thereby significantly improving PCA results [3, 4, 5, 6].

In general, the mapping Φ° , also referred to as an embedding, is only known over the training set. The extension of the mapping to new input points is of primary importance for kernel based methods whose success depends crucially on the “accuracy” of the extension. This problem, referred to as the *out-of-sample* problem, is often solved using the popular Nyström extension method [6, 7, 8]. In addition, the reverse mapping from the feature space back to the input space is often required. After operations are performed in feature space (these operations often necessitate the extension of the mapping), corresponding data points in input space often needs to be estimated. This problem is known as the *pre-image* problem.

The pre-image problem has received a lot of attention in kernel methods [6, 3, 5, 4]. Recently, Arias and coworkers [6] have shown its close connection with the out-of-sample problem. They also carefully considered the issue of normalization in feature space, thereby improving the “accuracy” of the out-of-sample extension and the pre-image estimation.

1.2. Contributions

Kernel PCA is achieved by applying a principal component analysis on the mapped training samples. PCA computes an eigen-decomposition of a kernel matrix deduced from the adjacency matrix W . Before applying PCA, the data is centered at the origin. In Kernel PCA the mean of the mapped input points is not known. Therefore, to simplify, one often assumes that the mapped training points $\Phi(s_i)$ are already centered in the feature space \mathbb{H} and incorrectly diagonalize the adjacency matrix W [6, 4]. Although simpler to understand, the resulting presentation of kernel methods misses some important points.

Our analysis of the kernel PCA methods studies in detail the centering of the data and underlines some important properties of the geometry of the mapped data induced by the kernel. We focus on the Gaussian kernel $w(s_i, s_j) =$

$\exp(-d_\chi^2(s_i, s_j)/2\sigma^2)$, with σ estimated as the median of all the distances between all training points [6, 9]. In accordance with the geometry induced by the Gaussian kernel, we highlight some non-trivial elements and rephrase some pre-image methods in a centered feature space[6]. A comparison based on numerical experiments demonstrates the superiority of our pre-image methods using a careful normalization in a centered feature space.

The remainder of this paper is organized as follows. Section 2 reviews Kernel PCA and the out-of-sample problem. Section 3 states the pre-image problem and insists on the issue of normalization in centered feature space. Numerical experiments on real data are reported in section 4 and section 5 concludes.

2. KERNEL PCA

Let $\{s_1, \dots, s_p\}$ be a set of training data in the input space χ . Kernel PCA computes the principal components of mapped features in the feature space \mathbb{H} . The mapping can be explicitly computed by the eigen-decomposition of a kernel matrix deduced from the adjacency matrix W . The coefficients of the adjacency matrix W are a measure of similarity between samples. Typically, the kernel function $w(\cdot, \cdot)$ is a decreasing function of the distance d_χ between training points s_i and s_j .

In this work, we focus on the Gaussian kernel. The Gaussian kernel has the important property of implicitly mapping the training points onto the unit sphere of \mathbb{H} , since $\|\Phi^\circ(s_i)\|^2 = \langle \Phi^\circ(s_i), \Phi^\circ(s_i) \rangle_{\mathbb{H}} = W_{i,i} = 1$. This important normalization property has been extensively used by Arias and coworkers [6] to improve the ‘‘accuracy’’ of previous pre-image methods [3, 5, 4]. In this work, we state the Kernel PCA methodology in centered space and shows that a finer degree of normalization can be achieved by considering the geometry of the mapped features.

Let $\bar{\Phi}^\circ = \frac{1}{p} \sum_{x_k \in \Gamma} \Phi^\circ(s_k)$ and $\bar{\Phi}^*$ denote the centered mapping, i.e. $\bar{\Phi}^*(s_i) = \Phi^\circ(s_i) - \bar{\Phi}^\circ$. The mapping $\bar{\Phi}^*$ can be computed by the eigen-decomposition of a centered kernel P^* [2]:

$$P^* = HWH = \Psi^* \Lambda^* \Psi^{*T} = \Psi^* \sqrt{\Lambda^*} (\Psi^* \sqrt{\Lambda^*})^T,$$

where H is the centering matrix $H = \mathbb{I} - \frac{1}{p} \mathbb{1}_p \mathbb{1}_p^T$ and $\Lambda^* = \text{diag}\{\lambda_1^*, \dots, \lambda_p^*\}$ with $\lambda_1^* \geq \dots \geq \lambda_{p-1}^* > \lambda_p^* = 0$. We denote $\hat{\Lambda} = \text{diag}\{\lambda_1^*, \dots, \lambda_{p-1}^*\}$ and $\hat{\Psi} = (\Psi_1^*, \dots, \Psi_{p-1}^*)$, the mapping is obtained as:

$$\bar{\Phi}^* : \chi \rightarrow \mathbb{R}^{p-1}, s_i \mapsto \sqrt{\hat{\Lambda}} \hat{\Psi}^T e_i^*. \quad (1)$$

The canonical basis $\{e_1^*, \dots, e_{p-1}^*\}$ of \mathbb{R}^{p-1} , defined formally by $e_k^* = \frac{1}{\sqrt{\lambda_k^*}} \sum_{s_i \in \Gamma} \Psi_k^*(s_i) \Phi^*(s_i)$, captures the variability of the point cloud of training samples. Projection of a

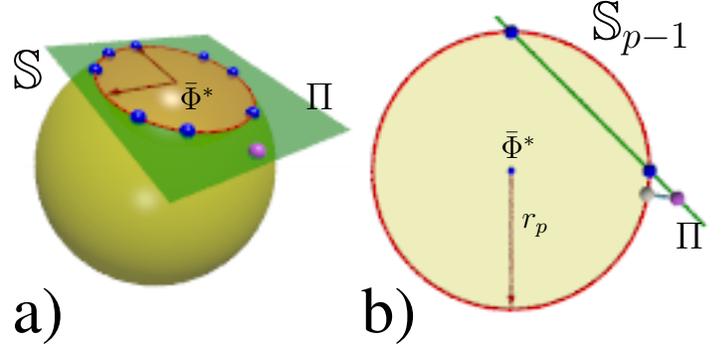


Fig. 1. a) Visualization of the feature points(blue) geometry in \mathbb{H} and the affine subspace(red circle); b) Affine subspace \mathbb{S}_{p-1}

new test point $s \in \chi$ onto the k^{th} -canonical vector e_k^* in the feature space can be shown to be:

$$\beta_k(s) = \langle e_k^*, \bar{\Phi}^*(s) \rangle = e_k^{*T} \frac{1}{\sqrt{\hat{\Lambda}}} \hat{\Psi}^T p_s^*, \quad (2)$$

$$\text{where } p_s^*(s_j) = H(w_s - \frac{1}{p} W \mathbb{1}_p)(s, s_j). \quad (3)$$

$p_s^*(s_j)$ is the extended mapping in centered feature space computed by centering the kernel vector w_s . This way of extending embedding coordinates to new test points has been used implicitly[3, 5, 4] or explicitly[6] in Kernel methods[10]. Projecting a new test point $s \in \chi$ onto the subspace spanned by the first m^* vectors $\{e_1^*, \dots, e_{m^*}^*\}$ (i.e. $P_{m^*}(\bar{\Phi}^*(s)) = \sum_{1 \leq k \leq m^*} \beta_k(s) e_k^*$) does not require the explicit computation of the mapping $\bar{\Phi}^*(s)$ since Eq.2 can be written only in terms of the kernel.

Working in a centered feature space, some important (often mistakenly ignored) comments follow. We show that the fundamental property of the mapped input points $\|\Phi^\circ(s_i)\|^2 = w(s_i, s_i) = 1$ can be greatly improved in a centered feature space. To do so, we define the mean in feature space $\bar{\Phi}^*(\in \mathbb{R}^{p-1}) = \frac{1}{p} \frac{1}{\sqrt{\Lambda^*}} \Psi^{*T} H W \mathbb{1}_p$ and consider some properties of the feature points mapped under:

$$\tilde{\Phi}^* : \chi \rightarrow \mathbb{R}^{p-1}, s \mapsto \bar{\Phi}^* + \Phi^*(s). \quad (4)$$

Under this mapping, the training samples verify: $\langle \tilde{\Phi}^*(s_i), \tilde{\Phi}^*(s_j) \rangle = w(s_i, s_j) - \bar{\Phi}_p^{*2}$, with $0 \leq \bar{\Phi}_p^* \leq 1$. The adjacency matrix W therefore gives (up to an additional factor $\bar{\Phi}_p^{*2}$) the inner product between two points in the feature space under the mapping $\tilde{\Phi}^*$. The constant $\bar{\Phi}_p^*$ has a simple geometric interpretation. In the feature space, the p non-centered training points, which belong to the unit sphere, define an affine space that is isomorphic to \mathbb{R}^{p-1} . This affine space, spanned by the vectors $\{e_1^*, \dots, e_{p-1}^*\}$, is at distance $\bar{\Phi}_p^*$ from the origin $\mathbf{0}$. Consequently, feature points mapped under $\tilde{\Phi}^* : s \mapsto \bar{\Phi}^* + \Phi^*(s)$ all belong to an hypersphere of \mathbb{R}^{p-1} of radius $r_p = \sqrt{1 - \bar{\Phi}_p^{*2}}$,

i.e. $\mathbb{S}_{p-1}(\mathbf{0}, r_p)$. This implies that, for all training sample $s_i \in \Gamma$, we have $\|\tilde{\Phi}^*(s_i)\| = r_p$. This normalization property of training samples is stronger than the usual property $\|\Phi^\circ(s_i)\| = 1$ and will prove important in the next section¹. In particular, this allows us to rephrase some pre-image methods, such as [6], in a centered feature space, leading to better results (sect 4). Finally, we note that the mapping $\tilde{\Phi}^\circ$ can be deduced from $\tilde{\Phi}^*$ by $\tilde{\Phi}^\circ : \chi \rightarrow \mathbb{R}^p, s \mapsto (\tilde{\Phi}^*(s)^T, \bar{\Phi}_p^*)^T$.

3. PRE-IMAGE

Given a point in the feature space ψ , the pre-image problem consists in finding a point $s \in \chi$ in the input space such that $\Phi(s) = \psi$, i.e. the pre-image of ψ . The exact pre-image might not exist (when it exists, it might also not be unique) and the pre-image problem is ill-posed [6, 3, 5, 4]. To circumvent this problem, one usually settles for an approximate solution and search for a pre-image that optimizes a given optimality criterion in the feature space. The pre-image problem has received a lot of attention in kernel methods [6, 3, 5, 4] and different optimality criteria have been proposed. Although most of those are based on the property $\|\Phi^\circ(s_i)\|^2 = 1$, significant improvement can be attained by considering that the mapped feature points $\tilde{\Phi}^*(s_i)$ belong to the hypersphere $\mathbb{S}_{p-1}(\mathbf{0}, r_p)$ (or equivalently stated that $\|\tilde{\Phi}^*(s_i)\| = r_p$). In particular, we insist on the fact that the popular normalization $\frac{\Phi^\circ(s)}{\|\Phi^\circ(s)\|}$ is not equivalent to the normalization $\frac{\tilde{\Phi}^*(s)}{\|\tilde{\Phi}^*(s)\|}$. In more detail, note that after normalization by the former criterion, a feature point does not belong any longer to the affine space defined by the p -training points. This behavior can also be seen in Figure 1b), which is the two dimensional visualization of the affine subspace (red circle) in Figure 1a). Figure 1a) shows the sphere \mathbb{S} and the layout of feature points on \mathbb{S} . The extended mapping of a new input point does not lie on the sphere (visualized as a purple point). As can be clearly seen the normalization as proposed in [6] projects the feature point (purple) onto the sphere (white). But the projected point does not lie in the span. This is clearly problematic as the principal modes of variations span only this affine space. The later normalization is the correct one and should be advantageously used. Therefore, we capitalize on our careful analysis of KPCA and define the different optimality criteria in centered feature space:

$$\text{Distance: } s = \arg \min_{z \in \chi} \|\tilde{\Phi}^*(z) - \tilde{\psi}^*\|^2, \quad (5)$$

$$\text{Collinearity: } s = \arg \max_{z \in \chi} \left\langle \frac{\tilde{\Phi}^*(z)}{\|\tilde{\Phi}^*(z)\|}, \frac{\tilde{\psi}^*}{\|\tilde{\psi}^*\|} \right\rangle, \quad (6)$$

where $\tilde{\psi}^* = \tilde{\Phi}^* + \psi^*$. Recently, Arias and coworkers [6] have shown the connections between the out-of-sample and

¹Note that to compute the radius value r_p (or, equivalently, the distance $\bar{\Phi}_p^*$), it is sufficient to compute $\|\tilde{\Phi}^*(s_i)\|$ for only one of the training samples $s_i \in \Gamma$.

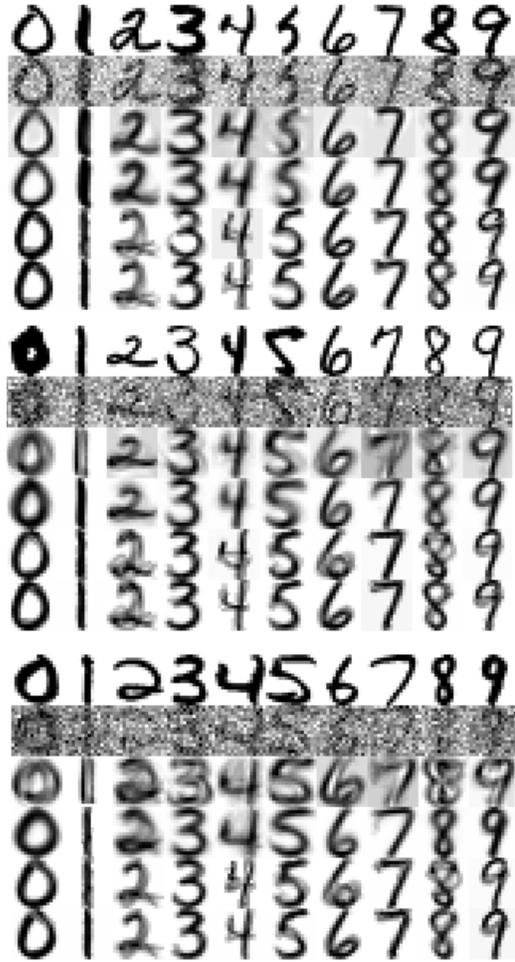


Fig. 2. Digit images corrupted by additive Gaussian noise (from top to bottom, $\sigma = 0.25, 0.45, 0.65$). The different rows respectively represent: the original digits and corrupted digits; different reconstruction methods: [3]; [3] with normalization; [5]; [5] with normalization.

the pre-image problems and proposed a normalized optimality criterion addressing the important lack of normalization in kernel methods:

$$s = \arg \min_{z \in \chi} \|\tilde{\Phi}^*(z) - \bar{\psi}\|^2 \text{ with } \bar{\psi} = r_p \frac{\tilde{\psi}^*}{\|\tilde{\psi}^*\|}. \quad (7)$$

Instead of solving directly for the pre-image in Eq.7, they first estimate the optimal kernel vector p_ψ^* as a standard least-squares problem $p_\psi^* = \hat{\Psi} \sqrt{\hat{\Lambda}} (\bar{\psi} - \bar{\Phi}^*)$ and then use previous methods [3, 5] to estimate the optimal pre-image.

4. APPLICATION IN IMAGE DENOISING

In order to validate the proposed algorithm, we run experiments on real world data. We test our pre-image algorithm on

the denoising of noisy images and compare our approach to previous methods. The computation of Kernel PCA is done using the Gaussian kernel $\exp(-d_{\chi}^2(s_i, s_j)/2\sigma^2)$ where σ is the median over all distances between points[6].

To test the performance of our approach on the task of image denoising, we apply the algorithm on the USPS dataset of handwritten digits². We show that our normalization method improves two recent state-of-the-art algorithms [3], [5]. Therefore, we form two training sets composed of randomly selected samples (60 and 200 respectively) for each of the ten digits. The test set is composed of 60 images randomly selected and corrupted by some additive Gaussian noise at different noise levels. The process of denoising simply amounts to estimating the pre-images of the feature vectors given by the Nyström extension of the noisy samples. In the case of Kernel PCA, we use the first $m^* = 8$ eigenvectors $\{e_1^*, \dots, e_{m^*}^*\}$ to compute projections in feature space.

σ^2	[3]	[3]	[5]	[5]
0.25	10.39	11.71	15.88	16.18
0.45	10.22	12.54	15.80	16.35
0.65	9.95	12.72	15.54	16.32
0.85	9.52	12.58	15.31	16.28
0.25	12.11	12.14	15.83	15.89
0.45	10.22	12.54	15.80	16.35
0.65	9.95	12.72	15.54	16.32
0.85	9,24	12.59	15.31	16.28

Table 1. Average PSNR (in dB) of the denoised images corrupted by different noise level. Training set is composed of 60 samples (first 4 rows) and 200 samples (last 4 rows). The first and third column show the denoising results without and the second and last columns with the normalization as proposed in this paper

Figure 2 displays some of the computed pre-images using different methods. Table 1 shows a quantitative comparison between different methods based on the pixel-signal-to-noise ratio(PSNR). Our normalisation method improves visually and quantitatively both pre-image methods. The results confirm that the new normalisation criterion in centered features space (second and fourth column) yields better results than previous pre-image methods (first and third column).

5. CONCLUSION

In this paper, we focused on the pre-image problem in kernel methods such as Kernel PCA. We especially focussed on the issue of correctly normalizing in centered feature space. A geometric interpretation eased the understanding of operations involved when working with centered data in feature space. As a consequence, we deduced a new normalization

criterion for previous proposed pre-image methods. The theoretical results could be nicely verified at hand of computed examples.

6. REFERENCES

- [1] M. Leventon, E. Grimson, and O. Faugeras, “Statistical shape influence in geodesic active contours,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 316–323.
- [2] B. Schölkopf, A.-J. Smola, and K.-R. Müller, “Kernel principal component analysis,” *Advances in kernel methods: support vector learning*, pp. 327–352, 1999.
- [3] S. Dambreville, Y. Rathi, and A. Tannenbeau, “Statistical shape analysis using kernel pca,” *IS&T/SPIE Symposium on Electronic Imaging*, 2006.
- [4] S. Mika, B. Schölkopf, A. J. 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*, M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds. 1999, MIT Press.
- [5] James T. Kwok and Ivor W. Tsang, “The pre-image problem in kernel methods.,” *IEEE Transaction in Neural Network*, vol. 15, no. 6, pp. 1517–1525, 2004.
- [6] Pablo Arias, Gregory Randall, and Guillermo 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.
- [7] Yoshua Bengio, Jean-Francois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet, “Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering,” in *Advances in Neural Information Processing Systems 16*, Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, Eds. MIT Press, Cambridge, MA, 2004.
- [8] S. Lafon and A. B. Lee, “Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 9, pp. 1393–1403, 2006.
- [9] Stephane Lafon, Yosi Keller, and Ronald R. Coifman, “Data fusion and multicue data matching by diffusion maps,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1784–1797, 2006.
- [10] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, “A kernel view of the dimensionality reduction of manifolds,” Tech. Rep. 110, Max-Planck-Institut für Biologische Kybernetik, Tübingen, Germany, 2003.

²The USPS dataset is available from <http://www.kernel-machines.org>.