Kernel Principal Components Are Maximum Entropy Projections*

António R.C. Paiva, Jian-Wu Xu, and José C. Príncipe

Computational NeuroEngineering Laboratory,
Dept. of Electrical and Computer Engineering,
University of Florida, Gainesville, FL 32611, USA
{arpaiva, jianwu, principe}@cnel.ufl.edu

Abstract. Principal Component Analysis (PCA) is a very well known statistical tool. Kernel PCA is a nonlinear extension to PCA based on the kernel paradigm. In this paper we characterize the projections found by Kernel PCA from a information theoretic perspective. We prove that Kernel PCA provides optimum entropy projections in the input space when the Gaussian kernel is used for the mapping and a sample estimate of Renyi's entropy based on the Parzen window method is employed. The information theoretic interpretation motivates the choice and specifices the kernel used for the transformation to feature space.

Keywords: Kernel PCA, information-theoretic learning, entropy projections.

1 Introduction

Many real world problems deal with a very high number of signals not all equally important for the application. Therefore, a simplification of the problem is often desirable, and sometimes imperative. The goal is to obtain a smaller number of projections that describes the data and minimize the loss of information in the projection. A very well known statistical tool for data projection is Principal Component Analysis (PCA) [1]. PCA searches for the projections of maximum variance. If the process that generated the data is Gaussian this projection is optimum. This is because Gaussian processes are totally described by their mean and variance. The same is not true, however, for other data distributions.

PCA can be formulated in terms of inner (or dot) products. Following a recent trend, a kernel-based extension named Kernel PCA was proposed by Schölkopf et al. [2,3]. In fact, it has been pointed out that any algorithm that can be formulated using only dot products can be immediately *kernelized*, yielding an easily trackable nonlinear formulation. Kernel PCA performs PCA in feature space. It has been verified, that by selecting the kernel appropriately, it is possible to find a projection in the input space that is more descriptive of the data, even

* This work was supported in part by Fundação para a Ciência e a Tecnologia (FCT) grant SFRH/BD/18217/2004 and NSF grant ECS-0300340.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 846–853, 2006.

if the data is described by a non-Gaussian distribution. Recently, Williams [4] pointed out that KERNEL PCA algorithm can be interpreted as a form of multidimensional scaling provided that the kernel function $\kappa(\mathbf{x}, \mathbf{y})$ is isotropic, i.e. it depends only on $\| \mathbf{x} - \mathbf{y} \|$. This connection provides a metric multidimensional scaling algorithm to solve KERNEL PCA instead of a eigendecomposition of the Gram matrix. Bengio et al. [5] pointed out the link between KERNEL PCA and spectral embedding. The direct relation resides in a more general learning problem: learning the principal eigenfunctions of operators defined from a kernel and the unknown data-generating density function.

In this paper we take an information-theoretic perspective to KERNEL PCA. We show a direct connection between KERNEL PCA and maximization of entropy, and prove mathematically why this happens. As Bach and Jordan [6] pointed out, this insight is also highly valuable to ICA, since ICA can be viewed as a generalization of PCA, one that depends on high order moments. Although a relation between KERNEL PCA and ICA is not made here, the demonstration we make inherently connects both concepts.

2 Kernel PCA

Let $\mathbf{x}_i$, $i = 1, \ldots, M$ be a set of M sample vectors in a N-dimensional (input) space, and $\mathbf{\Phi}(\cdot) : \mathbb{R}^N \longrightarrow \mathcal{F}$ be the mapping to the feature space. KERNEL PCA is simply PCA applied in feature space. Hence, the goal of KERNEL PCA is to find variance maximizing projections of the vectors $\mathbf{\Phi}(\mathbf{x}_i)$. If the vectors $\mathbf{\Phi}(\mathbf{x}_i)$, $i = 1, \ldots, M$ have zero mean KERNEL PCA can be stated as the following optimization problem: we wish to maximize the cost function

$$J(\mathbf{w}) = E\left\{(\mathbf{w}^T \mathbf{\Phi}(\mathbf{x}))^2\right\}. \tag{1}$$

Because the above equation depends on the norm of the projection vector, the Lagrange multiplier method is used to force the vectors to unit norm. Thus, the following cost function is maximized instead

$$\begin{aligned} J(\mathbf{w}) &= E\left\{(\mathbf{w}^T \mathbf{\Phi}(\mathbf{x}))^2\right\} - \lambda(\mathbf{w}^T \mathbf{w} - 1) \\ &= \mathbf{w}^T E\left\{\mathbf{\Phi}(\mathbf{x})\mathbf{\Phi}(\mathbf{x})^T\right\} \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1). \end{aligned} \tag{2}$$

Notice that $\mathbf{C} = E\left\{\mathbf{\Phi}(\mathbf{x})\mathbf{\Phi}(\mathbf{x})^T\right\}$ is the covariance matrix of the vectors in the feature space. The solution of (2) is found by solving

$$\mathbf{C}\mathbf{w} = \lambda \mathbf{w}. \tag{3}$$

As for PCA, the solutions to this equation are well known to be the eigenvectors and eigenvalues of the covariance matrix, although in this situation computed in feature space. Solving this problem directly in feature space is very complicated. Fortunately, this equation can be restated in terms of dot products, for which a solution can be easily found, as we shown next.

As all solutions $\mathbf{w}$ of (3) for which $\lambda \geq 0$ lie in the span of the transformed vectors we can write,

$$\mathbf{w} = \sum_{i=1}^{M} \alpha_i \mathbf{\Phi}(\mathbf{x}_i). \tag{4}$$

Also, the covariance matrix of the transformed vectors can be estimated from the vectors as

$$C = \frac{1}{M} \sum_{i=1}^{M} \mathbf{\Phi}(\mathbf{x}_i)\mathbf{\Phi}(\mathbf{x}_i)^T. \tag{5}$$

Returning to the problem of the eigendecomposition of the covariance matrix of the feature vectors, we have that (3) is equivalent to

$$\langle \mathbf{\Phi}(\mathbf{x}_k), \mathbf{C}\mathbf{w} \rangle = \lambda \langle \mathbf{\Phi}(\mathbf{x}_k), \mathbf{w} \rangle, \text{for all } k = 1, \ldots, M. \tag{6}$$

Then, substituting (4) and (5) yields

$$\frac{1}{M} \sum_{i=1}^{M} \mathbf{\Phi}^T(\mathbf{x}_k)\mathbf{\Phi}(\mathbf{x}_i) \sum_{j=1}^{M} \alpha_j \mathbf{\Phi}^T(\mathbf{x}_i)\mathbf{\Phi}(\mathbf{x}_j) = \lambda \sum_{j=1}^{M} \alpha_j \mathbf{\Phi}^T(\mathbf{x}_k)\mathbf{\Phi}(\mathbf{x}_j), \quad \text{for all } k = 1, \ldots, M. \tag{7}$$

Defining the Gram matrix $\mathbf{K}$, as $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle$, $i, j = 1, \ldots, M$, we can rewrite (7) in matrix form as

$$\mathbf{K}^2 \alpha = M\lambda \mathbf{K}\alpha. \tag{8}$$

where $\alpha = [\alpha_1, \ldots, \alpha_M]^T$. This equation has solutions found by the eigendecomposition of $\mathbf{K}$ but, most important of all, is that tells us that the eigenvectors of the Gram Matrix are the coefficients the decomposition of the eigenvectors of $\mathbf{C}$. Consequently, the projection of a feature vector is

$$\langle \mathbf{\Phi}(\mathbf{x}), \mathbf{w}_j \rangle = \sum_{i=1}^{M} \alpha_i^j \langle \mathbf{\Phi}(\mathbf{x}), \mathbf{\Phi}(\mathbf{x}_i) \rangle, \tag{9}$$

where $\mathbf{w}_j$ denotes the j-th *positive* eigenvector of $\mathbf{C}$.

3 Information Theoretic Concepts

In this section we briefly introduce some of the information theoretic core concepts needed to later establish its connection to Kernel PCA.

The key information measure in information theoretic applications is Rényi's quadratic entropy [7], defined for the pdf, $f(\mathbf{x})$, of a random variable $\mathbf{X}$ as

$$H(\mathbf{x}) = -\log \int_{-\infty}^{\infty} f^2(\mathbf{x}) dx = -\log E\{f(\mathbf{x})\}. \tag{10}$$

The argument of the logarithm,

$$V(\mathbf{x}) = \int_{-\infty}^{\infty} f^2(\mathbf{x})dx = E\left\{f(\mathbf{x})\right\}, \tag{11}$$

is what it is called the *information potential* (IP), so named due to a similarity with the potential energy field in physics [8]. Notice that the information potential depends directly on the pdf of $\mathbf{X}$, which is normally unknown. Luckily, we can circumvent the explicit estimation of the pdf because entropy is a "moment" of the pdf. In fact, using the Parzen window method [9], written as

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \kappa_{\sigma/\sqrt{2}}(\mathbf{x}, \mathbf{x}_i), \tag{12}$$

where $\kappa_{\sigma/\sqrt{2}}(\mathbf{x}, \mathbf{x}_i)$ is the estimation kernel, commonly taken as a Gaussian,with bandwidth $\sigma/\sqrt{2}$, although other kernels may be used [9]. This kernel must be a valid pdf, i.e. be positive and integrate to one. Then, substituting this estimator in the IP we do not need to explicitly compute the integral because the integral of a product of Gaussians is a Gaussian (with twice the variance), yielding directly

$$\hat{V}(\mathbf{x}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \kappa_{\sigma}(\mathbf{x}_i, \mathbf{x}_j). \tag{13}$$

Although the information potential as given by the previous equation is an approximation, this is only to the extent of the error in the pdf estimation. In other words, if $\hat{f}(\mathbf{x})$ from (12) equals the true pdf then the estimator given by (13) also has no error.

For any Mercer kernel, one can employ *Mercer's theorem*,

$$\kappa_{\sigma}(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \right\rangle, \tag{14}$$

to rewrite the information potential of (13) as [10]

$$\hat{V}(\mathbf{x}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \left\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \right\rangle = \left\langle \frac{1}{N}\sum_{i=1}^{N} \mathbf{\Phi}(\mathbf{x}_i), \frac{1}{N}\sum_{j=1}^{N} \mathbf{\Phi}(\mathbf{x}_j) \right\rangle = \| \mu_{\mathbf{\Phi}} \|^2, \tag{15}$$

where $\mu_{\mathbf{\Phi}}$ is the mean of the vectors in feature space. That is, the information potential is the squared norm of the mean vector of the data in kernel space. This equation shows exactly the duality existing between the information potential and second order statistics computed in feature space on the transformed data.

Finally, we remark that extremization (maximization or minimization) of $H(\mathbf{x})$ can be alternatively achieved by extremizing the information potential in the opposite direction, because of the minus signs in (10) and the fact that the logarithm is a monotonic function. Hence, if we wish to maximize the entropy we can simply minimize the information potential. Conversely, maximizing the information potential yields minimum entropy.

4 Characterization of Kernel PCA Projections in Input Space

In section 2 explained the fundamentals of Kernel PCA. The most important point was to explicitly formulate Kernel PCA as a tool for finding projections of maximum variance in feature space, as (2) states. On the other hand, (15) shows that a relationship between second order statistics in the feature space and quadratic Renyi's entropy in the input space exists.

Let us analyze in detail what is the meaning of the variance of the feature vectors. The variance of the feature vectors is

$$\mathrm{var}(\mathbf{\Phi}(\mathbf{x})) = E\left\{\mathbf{\Phi}(\mathbf{x})^T\mathbf{\Phi}(\mathbf{x})\right\} - E\left\{\mathbf{\Phi}(\mathbf{x})\right\}^T E\left\{\mathbf{\Phi}(\mathbf{x})\right\}. \tag{16}$$

Expressing the inner product as a kernel operation and using identity (15),

$$\mathrm{var}(\mathbf{\Phi}(\mathbf{x})) = E\left\{\kappa(\mathbf{x},\mathbf{x})\right\} - V(\mathbf{x}). \tag{17}$$

The quantity $E\left\{\kappa(\mathbf{x},\mathbf{x})\right\}$ is the information potential at the origin, $V(0)$. This is a constant value representing the zero entropy situation, for which the maximum value of the information potential is achieved. From (17), maximizing the variance of the feature vectors corresponds therefore to the minimization of the information potential, $V(\mathbf{x})$, in the input space.

The fact that Kernel PCA finds projections that minimize the information potential in input space, together with the remarks made in section 3 on the relationship between the information potential and entropy prove the statement that kernel principal components are maximum entropy projections. Since entropy is associated with information [11], maximum entropy projections are the directions more informative to explore for machine learning algorithms. Furthermore, notice that at no point in our proof of this connection an assumption of a specific kernel was made, other than it has to be able to accurately provide an estimation to the input sample vectors pdf.

5 Example

In this section we illustrate what was just proved in the previous section. We will use a small example, in which the goal is to obtain the maximum informative projection of a mixture of two Gaussian distributions. The overall pdf is specified by

$$p(\mathbf{x}) = \frac{1}{2}\left(N(\mathbf{x},\mu_1,\mathbf{\Sigma}_1) + N(\mathbf{x},\mu_2,\mathbf{\Sigma}_2)\right), \tag{18}$$

where $N(\mathbf{x},\mu,\mathbf{\Sigma})$ is a Gaussian distribution with mean μ and covariance matrix $\mathbf{\Sigma}$. In this case,

$$\mu_1 = \begin{bmatrix}-1\\-1\end{bmatrix}, \quad \mathbf{\Sigma}_1 = \begin{bmatrix}1 & 0\\0 & 0.1\end{bmatrix}, \quad \mu_2 = \begin{bmatrix}1\\1\end{bmatrix}, \quad \mathbf{\Sigma}_2 = \begin{bmatrix}0.1 & 0\\0 & 1\end{bmatrix}.$$

For reference, we show the contours of constant projection (constant eigenvalue) of standard PCA in Fig. 1(a). Recall that the projection is made along a line orthogonal to the contours. The contours for KERNEL PCA using a Gaussian kernel are a little more difficult to construct, since KERNEL PCA has as many principal directions as the size of the Gram matrix. In exploratory data analysis, what matters are the directions in the input space, and it is not clear how they are related. In this case we decided to plot in Fig. 1(b)-(c) the direction corresponding to the maximum eigenvalue in kernel space, using a kernel size (variance) $\sigma^2 = 1$ and $\sigma^2 = 10$, respectively. Note how the contours bend

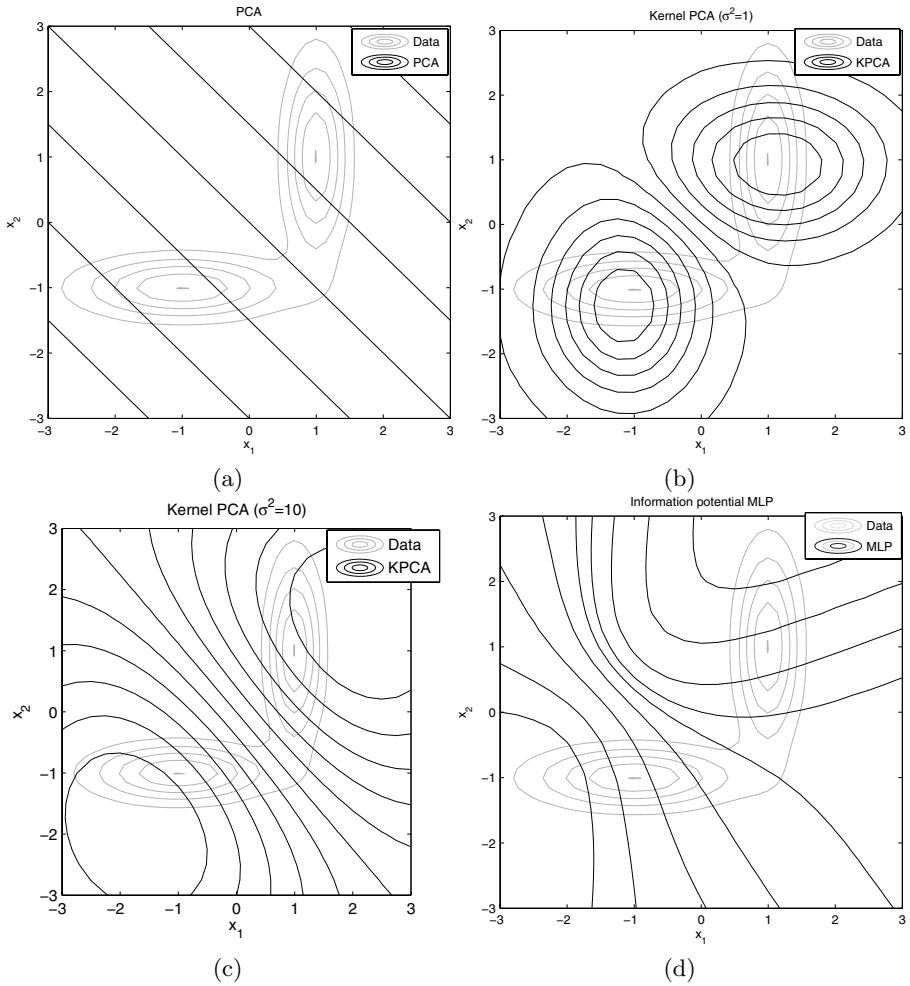

Fig. 1. Contours of constant projection and the pdf for the example of Sec. 5. From (a) to (d), the contours are for standard PCA, KERNEL PCA with $\sigma^2 = 1$, KERNEL PCA with $\sigma^2 = 10$, and MLP output.

themselves and wrap around the distribution to make the projection as uniform as possible. Also, when the kernel size is increased (Fig. 1(c)) the contours tend to those of standard PCA (Fig. 1(a)).

With our information theoretic interpretation there is another alternative to create the maximum entropy direction that uses always the input space dimension. In fact, we can train with backpropagation a MLP with architecture 2-4-1 (2 inputs, 4 hidden PEs and 1 output PE) and instead of using the conventional MSE criterion, substitute it for the maximum entropy cost [8]. The MLP was trained for 200 epochs to minimize the information potential of the outputs, as evaluated by (13), with a kernel size of 0.2. The nonlinearity used at the PEs is the hyperbolic tangent function. The contours of the surface generated by the neural network are shown in Fig. 1(d). The contours are obviously different from the ones for Kernel PCA since the basis functions are different and the method uses gradient descent learning, but is remarkable how they bend so that a projection to a line orthogonal to these contours would have maximum entropy. Although in this example we are only interested in the first projection, the neural network framework can also be used to obtain as many projections as needed up to the dimensionality of the space by using concepts of orthogonalizing the outputs [12].

6 Conclusions

Kernel PCA was proposed as an nonlinear extension of PCA. Despite this simple motivation, in this paper we prove that the principal components determined by Kernel PCA provides optimum entropy mappings when the Gaussian kernel is used both for the mapping and in Parzen window pdf estimation method. The use of the Gaussian kernel is not restrictive since the same result holds for any Mercer theorem, although the connection between the pdf estimation and IP becomes more difficult to express. This motivates the choice for the kernel and, considering the implicit pdf estimation, how to select its parameters.

The main contribution of this work is the understanding of the underlying properties of the projections found by Kernel PCA in feature space. This insight becomes especially important if we intend to use Kernel PCA as a data exploratory tool. Notice how the projections of Kernel PCA and maximum entropy achieve fundamentally the same result, although they are different due to the differences in the basis functions used (Gaussians in kernel methods, ridge functions in the MLP). This is a very interesting result given that Kernel PCA has an analytical solution, while the MLP requires adaptation. Yet, the Kernel PCA looses the intuition of the meaning of PCA in the input space. Indeed, Schölkopf et al. [2] mention about the possibility of finding more eigenvectors than the dimension of the input space which is clearly misleading in data analysis. The maximum entropy projection brings the insight that effectively Kernel PCA is projecting the data in informative directions using local bases. Therefore, Kernel PCA will require many such projections to cover the full data space. However, it is still not clear how to distinguish a minor component from a major component since the bases are local.

References

1. Diamantaras, K.I., Kung, S.Y.: Principal Component Neural Networks: Theory and Applications. John Wiley & Sons (1996)
2. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10**(5) (1998) 1299–1319
3. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. In Gerstner, W., Germond, A., Hasler, M., Nicoud, J.D., eds.: Proc. Artificial Neural Networks ICANN'97, Berlin, Springer Lecture Notes in Computer Science, Vol. 1327 (1997) 583–588
4. Williams, C.K.I.: On a connection between kernel pca and metric multidimensional scaling. Machine Learning **46**(1–3) (2002) 11–19
5. Bengio, Y., Delalleau, O., Roux, N.L., Paiement, J.F., Vincent, P., Ouimet, M.: Learning eigenfunctions links spectral embedding and kernel pca. Neural Computation **16**(10) (2004) 2197–2219
6. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. Journal of Machine Learning Research **3** (2002) 1–48
7. Rényi, A.: On measures of entropy and information. In: Selected paper of Alfréd Rényi. Volume 2. Akademiai Kiado, Budapest, Hungary (1976) 565–580
8. Príncipe, J.C., Xu, D., Fisher, J.W.: Information theoretic learning. In Haykin, S., ed.: Unsupervised Adaptive Filtering. Volume 2. John Wiley & Sons (2000) 265–319
9. Parzen, E.: On the estimation of a probability density function and the mode. The Annals of Mathematical Statistics **33**(2) (1962) 1065–1076
10. Jenssen, R., Erdogmus, D., Príncipe, J.C., Eltoft, T.: Towards a unification of information theoretic learning and kernel methods. In: Proc. MLSP'04, São Luís, Brazil (2004)
11. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27** (1948) 379–423, 623–656
12. Sanger, T.D.: Optimal unsupervised learning in a single layer linear feedforward neural network. Neural Networks **2**(7) (1989) 459–473