Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem

Mark Girolami
mark.girolami@paisley.ac.uk
Laboratory of Computing and Information Science,
Helsinki University of Technology
FIN-02015 HUT, Finland

Kernel principal component analysis has been introduced as a method of extracting a set of orthonormal nonlinear features from multivariate data, and many impressive applications are being reported within the literature. This article presents the view that the eigenvalue decomposition of a kernel matrix can also provide the discrete expansion coefficients required for a nonparametric orthogonal series density estimator. In addition to providing novel insights into nonparametric density estimation, this article provides an intuitively appealing interpretation for the nonlinear features extracted from data using kernel principal component analysis.

1 Introduction

Kernel principal component analysis (KPCA) is an elegant method of extracting nonlinear features from data, the number of which may exceed the dimensionality of the data (Schölkopf, Smola, & Müller, 1996, 1998). There have been many notable applications of KPCA for the denoising of images and extracting features for subsequent use in linear support vector classifiers (Schölkopf et al., 1996; Schölkopf, Bruges, & Smola, 1999). Computationally efficient methods have been proposed in Rosipal and Girolami (2001) for the extraction of nonlinear components from a Gram matrix, thus obviating the computationally burdensome requirement of diagonalizing a potentially high-dimensional Gram matrix.[1]

In KPCA, the implicit nonlinear mapping from input space to a possibly infinite-dimensional feature space often makes it difficult to interpret features extracted from the data. However, by considering the estimation of a probability density function from a finite data sample using an orthogonal

[1] The term *Gram matrix* refers to the $N \times N$ kernel matrix. The terms *kernel matrix* and *Gram matrix* may be used interchangeably.

series, some insights into the nature of the features extracted by KPCA can be provided.

Section 2 briefly reviews orthogonal series density estimation, and Section 3 introduces the notion of using KPCA to extract the orthonormal features required in constructing a finite series density estimator. Section 4 considers the important aspect of selecting the appropriate number of components that should appear in the series. Section 5 highlights the fact that the quadratic Renyi entropy of the data sample can be estimated using the associated Gram matrix. This strengthens the view that KPCA provides features that can be viewed as estimated components of the underlying data density. Section 6 provides some illustrative examples, and section 7 is devoted to conclusions and related discussion.

2 Finite Sequence Density Estimation

The estimation of a probability density by the construction of a finite series of orthogonal functions is briefly described in this section. (See Izenman, 1991, and the references in it for a complete exposition of this nonparametric method of density estimation.) We first consider density estimation using an infinite-length series expansion.

2.1 Infinite-Length Sequence Density Estimator. A probability density function that is square integrable can be represented by a convergent orthogonal series expansion (Izenman, 1991), such that

$$p(\mathbf{x}) = \sum_{k=1}^{\infty} c_k \Phi_k(\mathbf{x}), \tag{2.1}$$

where $\mathbf{x} \in \Re^D$ and the functions $\{\Phi_k(\mathbf{x})\}_{k=1}^{\infty}$ form an orthonormal system of functions. For any orthonormal series expansion in a Hilbert space (Kreyszig, 1989), for the case where $p(\mathbf{x})$ is a density function, the associated expansion coefficients follow as

$$c_k = \int p(\mathbf{x}) \Phi_k(\mathbf{x})\, d\mathbf{x} = E_p\{\Phi_k(\mathbf{x})\} \equiv \mu_k^{\Phi}, \tag{2.2}$$

where E_p denotes expectation with respect to the density function p. Thus, we can write $p(\mathbf{x}) = \sum_{k=1}^{\infty} \mu_k^{\Phi} \Phi_k(\mathbf{x}) = \langle \boldsymbol{\mu}^{\Phi} \cdot \boldsymbol{\Phi}(\mathbf{x}) \rangle$ where $\langle \cdot \rangle$ represents the canonical (Euclidean) inner product. We can also consider the elements of the series $\boldsymbol{\Phi}(\mathbf{x})$ as a nonlinear map from the data space to some feature space (Schölkopf et al., 1999), such that the inner product in feature space is computed directly using a kernel function $\langle \boldsymbol{\Phi}(\mathbf{x}') \cdot \boldsymbol{\Phi}(\mathbf{x}) \rangle = K(\mathbf{x}', \mathbf{x})$. Consider a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with associated kernel $K(\mathbf{x}', \mathbf{x}) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\mathbf{x}) \phi_k(\mathbf{x}')$ and inner product denoted by $\langle \cdot \rangle_{\mathcal{H}}$, therefore implicitly defining the elements of the mapping as $\Phi_k(\mathbf{x}) \equiv \sqrt{\lambda_k}\phi_k(\mathbf{x})$. The density function is given by the inner product of the mapped point in feature space $\boldsymbol{\Phi}(\mathbf{x}')$ and the mean of the distribution, $\boldsymbol{\mu}^{\Phi}$, in the defined feature space; thus, the reproducing property of the space yields the following density at the point $\mathbf{x}'$:

$$p(\mathbf{x}') = \langle p(\mathbf{x}) \cdot K(\mathbf{x}', \mathbf{x})\rangle_{\mathcal{H}} = \langle \boldsymbol{\mu}^{\Phi} \cdot \boldsymbol{\Phi}(\mathbf{x}')\rangle. \tag{2.3}$$

The form of the expression for the density function 2.3 gives an indication of a link between the fundamental problem of density estimation and unsupervised learning methods that employ the kernel trick (Schölkopf et al., 1998, 1999). If there is a finite sample of points $[\mathbf{x}_1, \cdots, \mathbf{x}_N]$ drawn from the true distribution $p(\mathbf{x})$, then an unbiased numerical estimate of the above expectation (coefficients of the expansion)[2] is $c_k \approx \hat{c}_k = \int \frac{1}{N}\sum_{n=1}^{N} \delta(\mathbf{x}_n - \mathbf{x})\Phi_k(\mathbf{x})d\mathbf{x} = \frac{1}{N}\sum_{n=1}^{N} \Phi_k(\mathbf{x}_n) \equiv \hat{\mu}_k^{\Phi}$. The estimated value of the probability density function for a point $\mathbf{x}'$, denoted by $\hat{p}(\mathbf{x}')$, is then given by the expression

$$\begin{aligned}\hat{p}(\mathbf{x}') &= \langle \hat{\boldsymbol{\mu}}^{\Phi} \cdot \boldsymbol{\Phi}(\mathbf{x}')\rangle = \frac{1}{N}\sum_{n=1}^{N}\langle \boldsymbol{\Phi}(\mathbf{x}_n) \cdot \boldsymbol{\Phi}(\mathbf{x}')\rangle \\ &= \frac{1}{N}\sum_{n=1}^{N} K(\mathbf{x}', \mathbf{x}_n) = \mathbf{1}_N^{\mathrm{T}}\mathbf{k}(\mathbf{x}'),\end{aligned}$$

where the $N \times 1$ vector $[K(\mathbf{x}', \mathbf{x}_1) \cdots K(\mathbf{x}', \mathbf{x}_N)]^{\mathrm{T}}$ is denoted by $\mathbf{k}(\mathbf{x}')$ and the vector $\mathbf{1}_N$ is an $N \times 1$ dimensional vector whose individual elements are each the value $1/N$. This expression is the familiar Parzen window density estimator with $K(.,.)$ being the smoothing kernel. For the case where a density-dependent weighting α_n is employed in estimating the series coefficients, then $\hat{p}(\mathbf{x}') = \sum_{n=1}^{N} \alpha_n K(\mathbf{x}', \mathbf{x}_n)$, and support vector methods can be employed in estimating the appropriate weighting coefficients (Mukherjee & Vapnik, 1999), which will yield a sparser representation of the density estimate than the Parzen window estimator.

2.2 Truncated Sequence Density Estimator. Turning to the truncated form of the infinite series estimate, the estimated value of the probability

[2] Note that the empirical density estimate based on the sum of delta functions $\frac{1}{N}\delta(\mathbf{x}_n - \mathbf{x})$ can also be exchanged for a weighting α_n such that $0 \leq \alpha_n \leq 1$ and $\sum_{n=1}^{N} \alpha_n = 1$, in which case $\hat{c}_k = \sum_{n=1}^{N} \alpha_n \Phi_k(\mathbf{x}_n)$, where each α_n will be estimated from the data using, for example, maximum likelihood.

density function for a point $\mathbf{x}'$ is then given by the expression

$$\begin{aligned}\hat{p}_M(\mathbf{x}') &= \sum_{k=1}^{M} \hat{c}_k \Phi_k(\mathbf{x}') \\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{M} \Phi_k(\mathbf{x}_n)\Phi_k(\mathbf{x}') = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{M} \lambda_k \phi_k(\mathbf{x}_n)\phi_k(\mathbf{x}'),\end{aligned} \tag{2.4}$$

where $\hat{p}_M(\mathbf{x}')$ denotes the density estimate yielded when M components of the series are retained. Many systems of orthogonal functions exist—for example, Legendre, Fourier, and Hermite systems have all been use in equation 2.4 for density estimation (Izenman, 1991). For the case where $p(\mathbf{x})$ has infinite support on the real line, then the system of orthogonal functions chosen is typically the normalized Hermite polynomials, and if the support is strictly positive, the Laguerre system may be used (Izenman, 1991; Tou & Gonzalez, 1974). Although density estimation using an orthogonal series is asymptotically unbiased, it has one distinct shortcoming: it can produce negative point values (Izenman, 1991).

A standard method for estimating a probability density function is the truncated orthogonal series density estimate, and this has been briefly introduced. The following section begins by considering the solution of the continuous Karhunen-Loève (KL) expansion from discrete data and points out that the nonlinear features provided by KPCA may be used in forming the basis functions for a series density estimator.

3 KPCA and Orthonormal Basis Functions

The solution of the continuous KL expansion from a discrete sample of data is a well-studied problem. Consider the integral equation that describes the KL expansion:

$$\int_{\mathcal{D}} K(\mathbf{x}, \mathbf{x}')\phi_i(\mathbf{x})\, d\mathbf{x} = \lambda_i \phi_i(\mathbf{x}'), \tag{3.1}$$

where $K(\mathbf{x}, \mathbf{x}')$ defines the covariance kernel of the associated stochastic process and the domain of integration is defined by $\mathcal{D}$. This integral can be generalized to take into account the variation over the domain of integration given by the underlying probability density function, in which case the integral in equation 3.1 becomes an expectation, and the eigenfunctions are then orthonormal with respect to the data density (Williams & Seeger, 2000). The estimation of the eigenfunctions of the continuous integral equation 3.1 from the associated discrete version, equation 3.2, generated by a finite sample of N data points was studied in Ogawa and Oja (1986):

$$\frac{1}{N}\sum_{n=1}^{N} K(\mathbf{x}_n, \mathbf{x}_m)\phi_i(\mathbf{x}_n) = \lambda_i\phi_i(\mathbf{x}_m). \tag{3.2}$$

Indeed, equation 3.1 is a homogeneous Fredholm integral equation of the second kind, and there are a number of quadrature methods available for its approximate numerical solution based on the discretized form of equation 3.2 (Delves & Mohamed, 1985). This is achieved by performing an eigenvalue decomposition on the $N \times N$-dimensional Gram matrix $\mathbf{K}$ whose elements are $\mathbf{K}_{in} = K(\mathbf{x}_i, \mathbf{x}_n)$, where $K(\mathbf{x}_i, \mathbf{x}_n)$ denotes the kernel function associated with the two points. The eigen-decomposition satisfies $\mathbf{KU} = \mathbf{US}$, where the columns of the $N \times N$ matrix $\mathbf{U}$ are the eigenvectors of the Gram matrix and the diagonal matrix $\mathbf{S}$ contains the elements $\tilde{\lambda}_k$, the corresponding eigenvalues. These eigenvectors form an estimate of the actual eigenfunctions such that $\phi_k(\mathbf{x}_n) \approx \hat{\phi}_k(\mathbf{x}_n) = \sqrt{N}u_{nk}$ and the eigenvalues are estimated by $\lambda_k \approx N^{-1}\tilde{\lambda}_k$ (Williams & Seeger, 2001). Estimates of the eigenfunctions at a new point $\mathbf{x}'$ can be made by simply using equation 3.2 as an interpolatory formula (Delves & Mohamed, 1985), in which case $\hat{\phi}_k(\mathbf{x}') = \frac{\sqrt{N}}{\tilde{\lambda}_k}\sum_{n=1}^{N} u_{nk}K(\mathbf{x}', \mathbf{x}_n)$. This approach is a form of the Nyström routine (Delves & Mohamed, 1985) and has recently been proposed in Williams and Seeger (2001) as a method for speeding up the inversion of the Gram matrix in kernel-based classification methods such as gaussian process classifiers.

In summary, the eigenvalue decomposition of a Gram matrix associated with a particular kernel $(K(\mathbf{x}_i, \mathbf{x}_n))$ provides estimates of the orthogonal system of eigenfunctions associated with the continuous integral equation (the KL expansion). It is now proposed that the extracted eigenvectors can be used to form the required orthogonal series for a probability density function estimate, and by doing so we gain some insights into the nature of the features extracted using KPCA.

3.1 Orthogonal Series Density Estimation from KPCA. It is clear that the eigenvalue decomposition of the Gram matrix $\mathbf{K}$ as in KPCA now provides estimates of the eigenfunctions associated with the kernel appearing in equation 3.1.[3]

Using the eigenvectors as finite sample estimates of the corresponding eigenfunctions, it is straightforward to see that the truncated estimate of the probability density function (see equation 2.4) at a point $\mathbf{x}'$

[3] In fact, kernel PCA is an eigenvalue decomposition performed on the centered kernel matrix $\tilde{\mathbf{K}}$, which is related to the original kernel matrix by $\tilde{\mathbf{K}} = (\mathbf{I} - \mathbf{1}_N)\mathbf{K}(\mathbf{I} - \mathbf{1}_N)$, where $\mathbf{I}$ is an $N \times N$ identity matrix and $\mathbf{1}_N$ is an $N \times N$ matrix whose elements are all 1/N.

follows as

$$
\begin{aligned}
\hat{p}_M(\mathbf{x}') = \sum_{k=1}^{M} \hat{c}_k \Phi_k(\mathbf{x}') &\approx \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{M} \frac{\tilde{\lambda}_k}{N} \hat{\phi}_k(\mathbf{x}_n) \hat{\phi}_k(\mathbf{x}') \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{M} \frac{\tilde{\lambda}_k}{N} \sqrt{N} u_{nk} \frac{\sqrt{N}}{\tilde{\lambda}_k} \sum_{l=1}^{N} u_{lk} K(\mathbf{x}', \mathbf{x}_l) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{M} \sqrt{\tilde{\lambda}_k} u_{nk} \sum_{l=1}^{N} \frac{u_{lk}}{\sqrt{\tilde{\lambda}_k}} K(\mathbf{x}', \mathbf{x}_l) \\
&= \mathbf{1}_N^{\mathrm{T}} \sum_{k=1}^{M} \sqrt{\tilde{\lambda}_k} \mathbf{u}_k \left\{ \sum_{l=1}^{N} \frac{1}{\sqrt{\tilde{\lambda}_k}} u_{lk} K(\mathbf{x}', \mathbf{x}_l) \right\},
\end{aligned}
$$

where the vector **1** is an $N \times 1$-dimensional vector whose individual elements are each the value one. It is worthy of note that the bracketed term appearing in the last line of the above equations is the projection of the kernel for the point $\mathbf{x}'$ onto the corresponding eigenvector—that is, the nonlinear principal component as in KPCA (Schölkopf et al., 1996). In other words, the density estimate is a weighted sum of the M significant (the sense of this significance will be defined in the following section) normalized eigenvectors of the Gram matrix. The weighting terms correspond to the associated nonlinear principal components of the point under question, so in this sense, the features extracted using KPCA can be viewed as components of the estimated data density. The implication is that the choice of kernel and any associated parameters[4] will have a profound effect on the quality of the associated density estimate. Therefore, good[5] features will be extracted by KPCA when the Gram matrix is such that a faithful reconstruction of the underlying data density can be made from the associated eigenfunctions.

The choice of kernel to use requires some consideration, and indeed it is this question that the focus of much research into kernel methods in machine learning is moving. From the viewpoint taken in this article, the choice of kernel is determined by the desire to model any density function. The gaussian, radial basis function (RBF) kernel has well-known universal approximation properties, and fitting a sufficient number of them to continuous data provides a means of estimating an arbitrary density function. This can be achieved either by semiparametric modeling using, for example, a mixture of gaussians, or by nonparametric modeling and using the gaussian smoothing window. The enhanced performance of support vector machines (Schölkopf et al., 1999) when employing features extracted by

[4] The width parameter, in the case of an RBF kernel.

[5] In the sense of providing discrimination power for a classifier, for example.

KPCA on difficult classification problems using RBF kernels (Schölkopf et al., 1998) can now be understood from the perspective given in this article. The principled selection of other kernels will be motivated by prior data and domain knowledge, and this is an open area of research investigation.

The eigenvectors extracted are estimates of the corresponding eigenfunctions of the continuous KL expansion; the accuracy of this estimate has been studied in Ogawa and Oja (1986). More recently in Williams and Seeger (2000), the significance of the eigenvectors of the Gram matrix for the purposes of classification has been discussed in detail.

A method for selecting the number of eigenvectors that should be retained in the expansion is presented in the following section.

4 Selecting the Length of Sequence

Now we see that this probability density function estimator can be written in compact format as

$$\begin{aligned}\hat{p}_M(\mathbf{x}') &= \mathbf{1}_N^{\mathrm{T}} \sum_{k=1}^{M} \sqrt{\tilde{\lambda}_k}\mathbf{u}_k \left\{ \sum_{l=1}^{N} \frac{1}{\sqrt{\tilde{\lambda}_k}} u_{lk} K(\mathbf{x}', \mathbf{x}_l) \right\} \\ &= \mathbf{1}_N^{\mathrm{T}} \mathbf{U}_M \mathbf{U}_M^{\mathrm{T}} \mathbf{k}(\mathbf{x}'),\end{aligned} \tag{4.1}$$

where $\mathbf{U}_M$ is the $N \times M$ matrix, which retains M of the eigenvectors of $\mathbf{K}$. For the case when $M = N$ as $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$, equation 4.1 reduces to the familiar form of $p(\mathbf{x}') = \mathbf{1}_N^{\mathrm{T}}\mathbf{k}(\mathbf{x}')$, the standard Parzen window density estimate. This representation of the orthogonal series estimator provides some insight into the kernel PCA approach to density estimation. It is clear from this that inserting the matrix $\mathbf{U}_M\mathbf{U}_M^{\mathrm{T}}$ in the standard Parzen window estimator has a smoothing effect on the estimated density. In other words, if, by way of an example, a series of noisy observations is available to estimate the density, then the orthogonal series estimator will potentially smooth out the effects of the noise by selectively removing $N - M$ eigenvectors from the expansion.

The length of the sequence of eigenvectors retained in the expansion requires consideration. The error generated by truncating the infinite expansion at M can be shown (Kreyszig, 1989; Tou & Gonzalez, 1974) to give an overall integrated-square truncation error of $\mathcal{E} = \sum_{k=M+1}^{\infty} c_k^2$, where each c_k is defined in equation 2.2. Therefore, the corresponding error associated with each element of the expansion is

$$\begin{aligned}\mathcal{E}_k = c_k^2 &= \left\{ \int p(\mathbf{x})\Phi_k(\mathbf{x})\, d\mathbf{x} \right\}^2 \approx \frac{1}{N^2} \left\{ \sum_{n=1}^{N} \Phi_k(\mathbf{x}_n) \right\}^2 \\ &\approx \frac{\lambda_k}{N} \left\{ \sum_{n=1}^{N} u_{nk} \right\}^2 \approx \tilde{\lambda}_k \left\{ \mathbf{1}_N^{\mathrm{T}} \mathbf{u}_k \right\}^2,\end{aligned}$$

where $\mathbf{u}_k$ and $\tilde{\lambda}_k$ are the kth eigenvector and eigenvalue of $\mathbf{K}$, respectively. The approximate truncation error associated with each element in the series based on the N-sample density estimate is $\hat{\mathcal{E}}_k = \tilde{\lambda}_k \left\{\mathbf{1}_N^{\mathrm{T}}\mathbf{u}_k\right\}^2$. Therefore, expansion coefficients that have a small squared-sum value of the components will have a negligible effect on the overall integrated-square error.[6] Although this error gives an indication of the contribution to the overall error from each component, it does not provide a stopping rule as such for the elimination of eigenvectors within the expansion. In Kronmal and Tarter (1968), the value of the mean integrated squared error is used in deriving a stopping criterion. This is one criterion that is appropriate for the application of kernel PCA in density estimation. In Diggle and Hall (1986), an alternative criterion is proposed; however, for the purposes of this work, the criterion of Kronmal and Tarter (1968) is adequate for the exposition of the ideas set forth. Details found in Kronmal and Tarter (1968) show that series elements that satisfy the following inequality should be considered for retention in the sequence:

$$\left\{\frac{1}{N}\sum\nolimits_{n=1}^{N}\Phi_k(\mathbf{x}_n)\right\}^2 > \frac{2}{1+N}\left\{\frac{1}{N}\sum\nolimits_{n=1}^{N}\Phi_k^2(\mathbf{x}_n))\}\right\}.$$

Substituting the eigenvector approximations for the eigenfunctions in the above equation gives the following cutoff threshold:

$$\left\{\mathbf{1}^{\mathrm{T}}\mathbf{u}_k\right\}^2 > \frac{2N}{1+N}(\mathbf{u}_k^{\mathrm{T}}\mathbf{u}_k) = \frac{2N}{1+N}. \tag{4.2}$$

If the sample size is large, then $\frac{2N}{1+N} \rightarrow 2$, and the stopping criterion is simply $\left\{\mathbf{1}^{\mathrm{T}}\mathbf{u}_k\right\}^2 > 2$. Much has been written in the literature of nonparametric statistics regarding the stopping criteria for orthogonal series density estimators. (See Diggle & Hall, 1986, for an extensive overview of these.)

This section has shown that the eigenvalue decomposition of the Gram matrix (KPCA) provides features that can be used in density function estimation based on a finite sample estimate of a truncated expansion of orthonormal basis functions. One criterion for the selection of the appropriate eigenvectors that will appear in the series has been considered.

The following section presents the nonparametric estimation of Renyi entropy from a data sample. The importance of the constructed Gram matrix along with the associated eigenspectrum is considered.

[6] This does not take into account the error in approximating the eigenfunctions using the estimated eigenvectors.

5 Nonparametric Estimation of Quadratic Renyi Entropy

Thus far, the discussion regarding the form of the kernel appearing in equation 3.1 has been general, and no specific kernel has been assumed. Let us now consider specifically a gaussian RBF kernel. Note that the quadratic Renyi entropy, defined as

$$\mathcal{H}_{R_2}(X) = -log \int p(\mathbf{x})^2 \, d\mathbf{x}, \tag{5.1}$$

can easily be estimated using a nonparametric Parzen estimator based on an RBF kernel. This above integral formed a measure of distribution compactness in (Friedman & Tukey, 1974) and has recently been used for certain forms of information-theoretic learning (Principe, Fisher, & Xu, 2000). Denoting an isotropic gaussian computed at $\mathbf{x}$ centered at $\boldsymbol{\mu}$ with covariance $\boldsymbol{\Lambda}$ as $\mathcal{N}_{\mathbf{x}}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, employing the standard result for a nonparametric density estimator using gaussian kernels and noting the convolution theorem for gaussians, the following holds:

$$\begin{aligned}\int p(\mathbf{x})^2 \, d\mathbf{x} \approx \int \hat{p}(\mathbf{x})^2 \, d\mathbf{x} &= \frac{1}{N^2} \int \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{N}_{\mathbf{x}}(\mathbf{x}_i, \boldsymbol{\Lambda}) \mathcal{N}_{\mathbf{x}}(\mathbf{x}_j, \boldsymbol{\Lambda}) \right\} d\mathbf{x} \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{N}_{\mathbf{x}_i}(\mathbf{x}_j, 2\boldsymbol{\Lambda}).\end{aligned}$$

For an RBF kernel with a common width of $2\boldsymbol{\Lambda}$, it is clear that the quadratic integral can be estimated from the sum of each element in the Gram matrix—in other words,

$$\int \hat{p}(\mathbf{x})^2 \, d\mathbf{x} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{1}_N^{\mathrm{T}} \mathbf{K} \mathbf{1}_N, \tag{5.2}$$

where each $N \times 1$ vector $\mathbf{1}_N$ has each element equal to $1/N$. Now we can see that the contribution to the overall estimated data entropy of each orthonormal component vector can be viewed using an eigenvalue decomposition of the Gram matrix

$$\int \hat{p}(\mathbf{x})^2 \, d\mathbf{x} = \sum_{k=1}^{N} \tilde{\lambda}_k \left\{ \mathbf{1}_N^{\mathrm{T}} \mathbf{u}_k \right\}^2 = \sum_{k=1}^{N} \hat{\mathcal{E}}_k.$$

It is clear that large contributions to the entropy will come from components that have small values of $\tilde{\lambda}_k \left\{ \mathbf{1}_N^{\mathrm{T}} \mathbf{u}_k \right\}^2$ and can be attributed to elements with little or no structure. This can be considered as the contribution caused by observation noise in some cases or diffuse regions in the data. Large values of $\tilde{\lambda}_k \left\{\mathbf{1}_N^{\mathrm{T}}\mathbf{u}_k\right\}^2$ therefore indicate regions of high density or compactness and are also indicative of possible modes of the density or underlying class and cluster structure. Interestingly, the integral considered in the computation of the quadratic Renyi entropy $\int \hat{p}(\mathbf{x})^2\, \mathbf{dx}$ also defines the squared norm of the functional form of $\hat{p}(\mathbf{x})$ such that

$$\int \hat{p}(\mathbf{x})^2\, d\mathbf{x} = \parallel \hat{p} \parallel_{\mathcal{H}}^2 = \langle \hat{\boldsymbol{\mu}}^{\Phi} \cdot \hat{\boldsymbol{\mu}}^{\Phi} \rangle = \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}_n) \phi_i(\mathbf{x}_m)$$

$$= \mathbf{1}_N^{\mathrm{T}} \mathbf{K} \mathbf{1}_N = \mathbf{1}_N^{\mathrm{T}} \mathbf{U} \mathbf{S} \mathbf{U}^{\mathrm{T}} \mathbf{1}_N = \sum_{k=1}^{N} \hat{\mathcal{E}}_k.$$

The above equation provides a more general result in that the kernel used in estimating the quadratic integral need not be restricted to an RBF form. The main point being made here is that the Gram matrix is fundamental to the estimation of the Renyi entropy, which is based on data density. When creating a Gram matrix for extraction of nonlinear features using, for example, KPCA, from the arguments presented in this and the previous section, the key is to choose a kernel that provides a reasonable estimate of the underlying data density. Because different types of kernel produce varying forms of associated eigenfunction, it is clear that the eigenfuctions should be appropriate for the density to be estimated. For example, an RBF kernel has eigenfunctions of the form of normalized Hermite polynomials (Zhu, Williams, Rohwer, & Morciniec, 1998; Williams & Seeger, 2001) and as such would be suitable for distributions with infinite support.

The following section provides a number of illustrative examples.

6 Simulation

The first simulation provides a two-dimensional illustration of the extracted features from the Gram matrix and how these can be interpreted. Analytic solutions to the one-dimensional form of equation 3.1 have been provided in Zhu et al. (1998) and Williams and Seeger (2001). An RBF kernel is used, and the weighting function is taken as a gaussian, in which case the eigenfunctions take the form of normalized Hermite polynomials (Kreyszig, 1989). Normalized Hermite polynomials form an orthonormal sequence such that in the univariate case,

$$\Phi_i(x) = \frac{1}{(2^i i! \sqrt{\pi})^{\frac{1}{2}}} \exp(-x^2/2) H_i(x), \tag{6.1}$$

where the Hermite polynomials $H_i(x)$ are defined by the recursion

$$H_0(x) = 1 \quad and \quad H_i(x) = (-1)^i exp(x^2) \frac{d^i}{dx^i} \exp(-x^2). \tag{6.2}$$

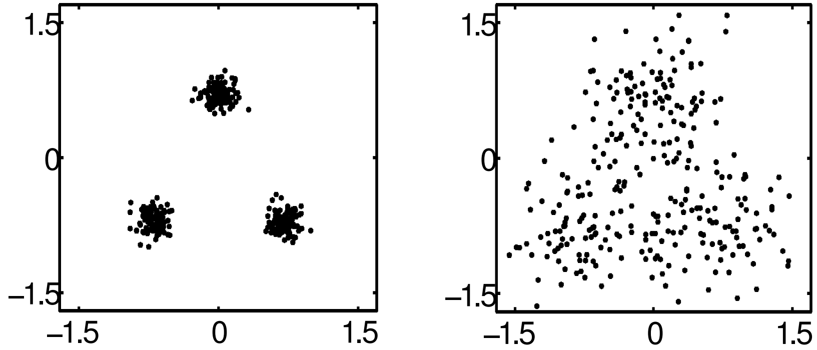

Figure 1: A sample of 300 two-dimensional points where 100 are drawn from each of three two-dimensional gaussian clusters. (Left) Scatterplot of the 300 points drawn from the gaussians with an isotropic variance of value 0.1. (Right) The same points with additive gaussian noise whose variance is 0.38.

The generalization of the univariate form to a multivariate representation is straightforward (Tou & Gonzalez, 1974).

Consider a sample of two-dimensional data points distributed in a similar manner to those presented in Schölkopf et al. (1998) for illustrative purposes. Three clusters of identical variance with value $\sigma = 0.1$ and centers $\boldsymbol{\mu} = [0.0, 0.7; 0.7, -0.7; -0.7 - 0.7]$ are generated. The left-hand plot of Figure 1 shows the scatter plot of the data. The density of the data corresponds to the general form of mixture such that $p(\mathbf{x}) = \sum_k^K \gamma_k \mathcal{N}_{\mathbf{x}}(\boldsymbol{\mu}_k, \sigma\mathbf{I})$, where the usual constraints $\sum_k \gamma_k = 1$ hold. Now note that an orthonormal set of basis functions with respect to the density of the data $p(\mathbf{x})$ must satisfy

$$\sum_{k=1}^{K} \gamma_k \int_{\mathbf{x}} \mathcal{N}_{\mathbf{x}}(\boldsymbol{\mu}_k, \sigma\mathbf{I}) \varphi_i(\mathbf{x} - \boldsymbol{\mu}_k) \varphi_j(\mathbf{x} - \boldsymbol{\mu}_k)\, d\mathbf{x} = \delta_{ij}. \tag{6.3}$$

Because the weighting function is a gaussian, two-dimensional Hermite polynomials of the form $H_i(\mathbf{x} - \boldsymbol{\mu}_k)$ will satisfy this requirement. This indicates that each of the components of the mixture (clusters in this instance) will have a set of orthonormal Hermite polynomial functions associated with them.

An RBF kernel of equal width to the individual cluster components was used, and 300 points were drawn from the distribution (see Figure 1). An eigenvalue decomposition was performed on the associated Gram matrix, and the eigenvectors were used in forming the series density estimate.

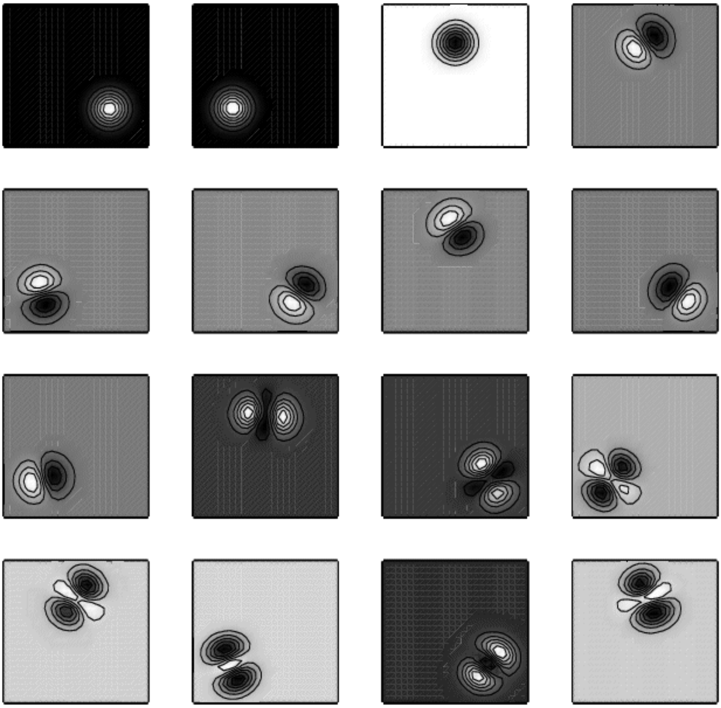

Figure 2
data of Figure 1 are shown starting from the top row. The characteristic structure of two-dimensional orthonormal Hermite polynomial functions is clear and most evident.

Figure 2 shows the nonlinear features associated with each of the first sixteen eigenvectors. These are clearly estimates of the orthonormal Hermite polynomial expansion coefficients associated with the density. The first four extracted features for one of the clusters are shown in Figure 3. It is clear that the extracted features are indeed, up to a rotation and scaling, estimates of the orthonormal Hermite polynomial functions (see Figure 3). In essence, what we see is that the nonlinear features extracted by kernel PCA using an RBF kernel are estimates of the orthonormal Hermite polynomial components associated with the underlying data distribution.

between the actual density
sity
estimated using the expectation maximization algorithm) and the KPCA method was computed for a range of the isotropic variance values ranging from 0.05 to 0.5, in 0.05 increments. The empirical KLD was computed

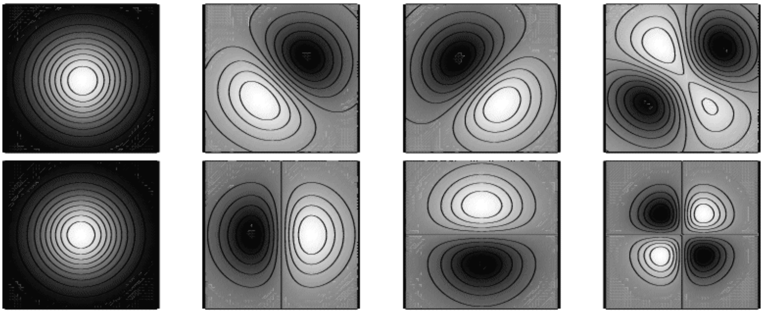

Figure 3
vectors associated with one of the clusters in the data. (Bottom) The first four two-dimensional orthonormal Hermite polynomial functions.

using the standard form

Three hundred sample points were used to create the Gram matrix and estimate the parameters of the gaussian mixture. Six hundred uniformly distributed points within the region defined in Figure 1 were used to compute the KLD for both methods. The mean KLD for the KPCA method was 0.036 compared to 0.043 for the mixture method over the range of values. The comparison shows similar performance for both methods on this particular two-dimensional data, as would be expected.

KPCA density estimation method. The 300 points from the clustered data in the previous simulation had gaussian noise of variance 0.38 added
ure 1 shows the original data and the noisy samples. From Figure 4, it is quite obvious that three components are all that is required to estimate the distribution corresponding to the noiseless data. However, in the case of the noisy data, a slowly decaying eigenspectrum can be seen (see Figure 5). However, when examining the contributions to the error of each eigenvector, it is still apparent that there are three significant generators of the data. The first three eigenvectors satisfy the Kronmal and Tarter criterion, after which a small number of eigenvectors satisfy the criterion. The right-hand plot of Figure 6 shows the estimated density contour plots when the first three eigenvectors are retained in the series expansion. This should be contrasted to that which a Parzen window estimator yields (middle plot of Figure 6). The smoothing effect can be noted due to the removal of series elements that capture the diffuse areas within the data and the sharpening of the three modes in the sample.

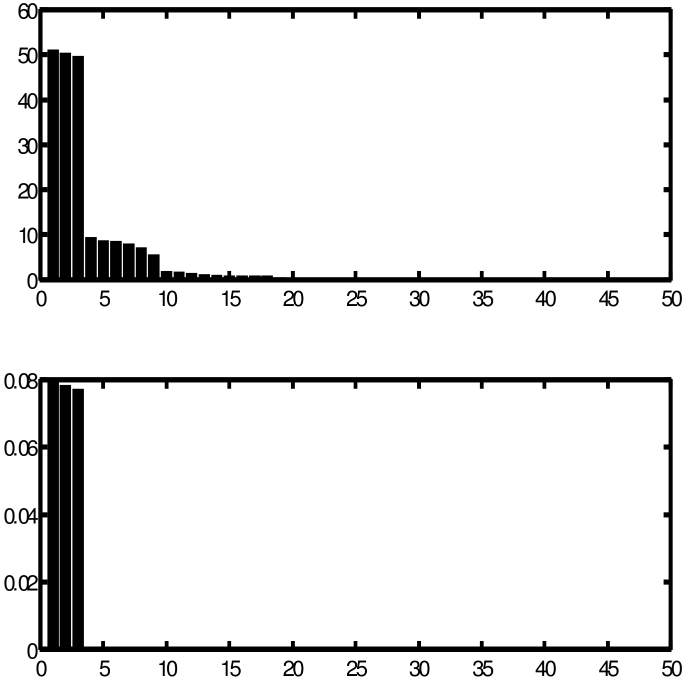

Figure 4: (Top) The first 50 eigenvalues of the Gram matrix created from the noiseless and well-separated clusters of Figure 1. Due to the distinct structure in the data, there are only three dominant eigenvalues. (Bottom) The contribution to the overall integrated square error of each of the first 50 eigenvectors. Again it is clear that only three eigenvectors satisfy the Kronmal and Tarter (1968) criterion, and it is these that are required for the density estimate.

The final illustrative simulation uses data drawn from a uniform distribution with finite support. The left-hand plot of Figure 7 shows the data drawn from a uniform distribution within the annular region that satisfies $9 \leq x^2 + y^2 \leq 25$. The Parzen window estimated density iso-contours are superimposed on these. The adjacent plot in Figure 7 gives the contour plot of the estimated density using the kernel PCA method. Figures 8 and 9 show the related nonlinear features and the relative importance of each.

7 Conclusion

Kernel PCA has proven to be an extremely useful method for extracting nonlinear features from a data set, and its utility has been demonstrated

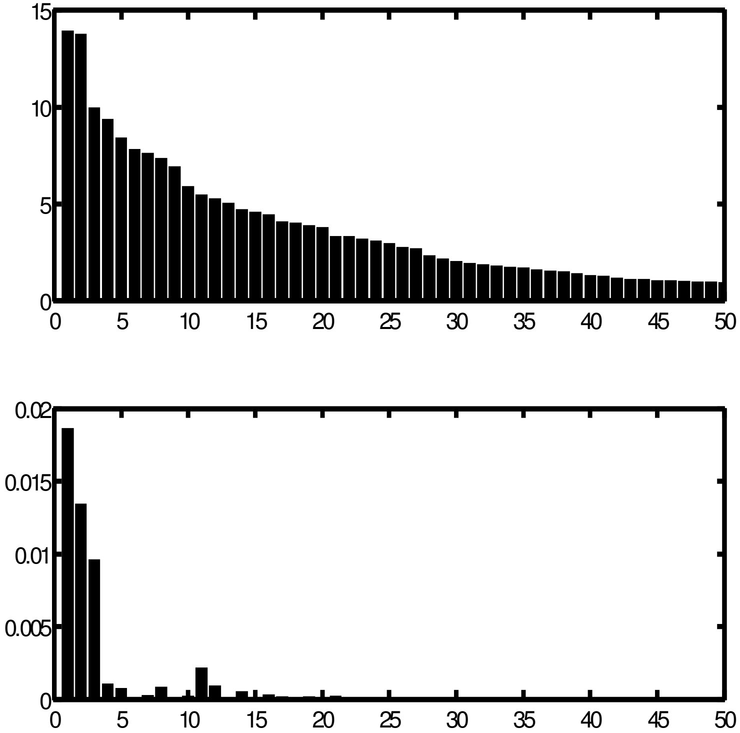

Figure 5: (Top) The first 50 eigenvalues of the Gram matrix created from the noisy clustered data of Figure 1. It is apparent that the slow exponential decay of the eigenvalues is attributed to the high level of additive noise on the finite number of observations. The fast decay of the previous example is now difficult to discern from the eigenvalues alone. (Bottom) The contribution to the overall integrated square error of each of the first 50 eigenvectors. It is apparent that there are only three dominant eigenvectors required for the majority of the density estimate. It is also clear that only the first three eigenvectors satisfy the Kronmal and Tarter (1968) criterion before it is violated, and it is these that are retained for the density estimate.

on, among other applications, many complex and demanding classification problems. An intuitive insight into the nature of these particular features has been somewhat lacking to date. This article has presented an argument that the nonlinear features extracted using KPCA (the eigendecomposition of a Gram matrix created using a specific kernel) provides features that can be considered as components of an orthogonal series density estimate. This follows somewhat from the observations made in Williams and Seeger (2001) regarding the effect of the data distribution on kernel-based classifiers.

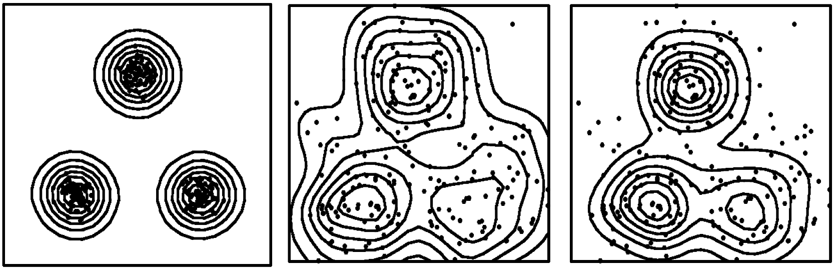

Figure 6: (Left) The estimated probability density contour plot of the data consisting of the three noiseless clusters using the kernel PCA-based orthogonal series approach. A Parzen window estimator using a gaussian kernel with a width of 0.1 gives an identical result. (Middle) The contour plots of the density estimate, using a Parzen window estimator, for the noisy data in Figure 1. (Right) The density for the noisy data using an orthogonal series estimator that consists of the first three eigenvectors of the Gram matrix that satisfied the Kronmal and Tarter (1968) criterion. A smoothing of the effects of the noise on the density estimate is apparent in this example.

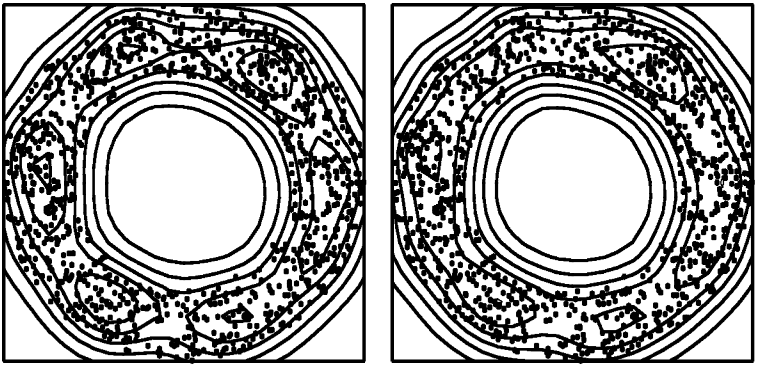

Figure 7: (Left) The scatterplot of 1000 points drawn from a uniform annular ring centered at the origin with uniform width. Superimposed on this are the iso-contours of the estimated probability density using Parzen window estimator. (Right) The iso-contours of the estimated probability density using kernel PCA method. An RBF kernel of unit width was used in this experiment. It is apparent that the uniform region of support has been extended in the density estimate due to the infinite support of the RBF kernel. Only eight eigenvectors satisfied the stopping criterion, and these were retained in the series expansion estimate, which amounts to a representation that uses 0.8% of the possible features.

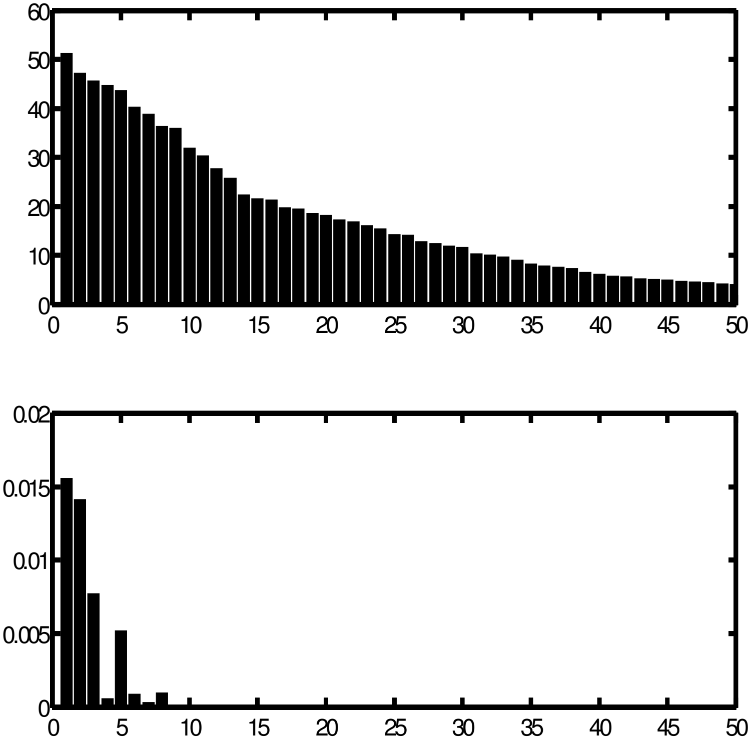

Figure 8: (Top) The eigenvalue spectrum for the first 50 eigenvalues of the kernel. (Bottom) The values of $\tilde{\lambda}_k \left\{\mathbf{1}_N^{\mathrm{T}}\mathbf{u}_k\right\}^2$ associated with each eigenvector. Eight of the eigenvectors satisfy the Kronmal and Tarter criterion (1968).

The probability density function estimate provided by the relevant M eigenvectors $\hat{p}_M(\mathbf{x}') = \mathbf{1}_N^{\mathrm{T}}\mathbf{U}_M\mathbf{U}_M^{\mathrm{T}}\mathbf{k}(\mathbf{x}')$ can be seen to be a smoothed Parzen window estimate where the matrix $\mathbf{U}_M\mathbf{U}_M^{\mathrm{T}}$ acts to smooth the estimate based on the data sample. In some sense, this can be seen as a reduced-set representation of the density function estimate based on the retained eigenvectors of the Gram matrix. The decomposition of the Gram matrix shows how each of the eigenvectors contributes to the overall data entropy (or norm of the functional form of the estimated density). Components that are related to the possible class structure (or modes) have a large sum-squared value, while those that are attributed to unstructured noise have low values. These particular values correspond to the induced error in the series density estimate when the related eigenvectors are discarded.

One point of note is the accuracy of the eigenvectors as estimates of the corresponding eigenfunctions and their effect on the density estimate. It is

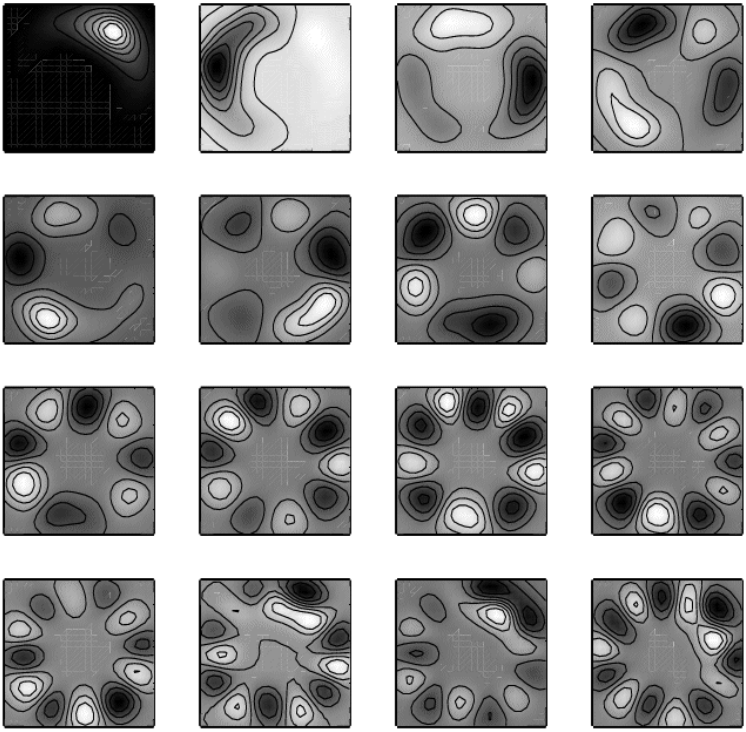

Figure 9
matrix. The characteristic Hermite polynomial structure of the features is most apparent.

noted that the series representaion is very sparse, with only eight components required to form the series density estimator for the data uniformly distributed within the annular region (see Figure 7). This amounts to the removal of 99.2% of the possible components in the series, with the large majority corresponding to the smaller eignvalues being discarded. Williams and Seeger (2001) show that for an RBF kernel, the accuracy of the estimates of the dominant eigenvalues is good, and this deteriorates for the estimation of the smaller values. The important components for the density function estimate are situated at the top end of the eigenspectrum, which do not suffer the effects of poor estimation.

sity estimation and provides an insight into the significance of the associated nonlinear features. This view may prove useful when considering kernel PCA as a means of nonlinear feature extraction for classifier design or data clustering and, of course, nonparametric density estimation.

Acknowledgments

This work was carried out under the support of the Finnish National Technology Agency TEKES. I am grateful to Harri Valpola and Erkki Oja for helpful discussions regarding this work. Many thanks to the anonymous reviewers for their helpful comments.

References

Delves, L. M., & Mohamed, J. L. (1985). *Computational methods for integral equations.* Cambridge: Cambridge University Press.

Diggle, P. J., & Hall, P. (1986). The selection of terms in an orthogonal series density estimator. *Journal of the American Statistical Association, 81*, 230–233.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computing, 23*, 881–890.

Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association, 86*, 205–224.

Kreyszig E. (1989). *Introductory functional analysis with applications.* New York: Wiley.

Kronmal, R., & Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *Journal of the American Statistical Association, 63*, 925–952.

Mukherjee, S., & Vapnik, V. (1999). *Support vector method for multivariate density estimation* (AI Memo 1653). Available at: ftp://publications.ai.mit.edu/ai-publications/1500-1999/AIM-1653.ps.

Ogawa, H., & Oja, E. (1986). Can we solve the continuous Karhunen-Loève eigenproblem from discrete data? *Transactions of the IECE of Japan, 69*(9), 1020–1029.

Principe, J., Fisher III, J., & Xu, D. (2000). Information theoretic learning. In S. Haykin (Ed.), *Unsupervised adaptive filtering.* New York: Wiley.

Rosipal, R., & Girolami, M. (2001). An expectation maximisation approach to nonlinear component analysis. *Neural Computation, 13*(3), 500–505.

Schölkopf, B., Bruges, C., & Smola, A. (Eds.). (1999). *Advances in kernel methods—Support vector learning.* Cambridge, MA: MIT Press.

Schölkopf, B., Smola, A., & Müller, K. R. (1996). Nonlinear component analysis as a kernel eigenvalue problem (Tech. Rep. MPI TR. 44). Available at: http://www.kernel-machines.org

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*(5), 1299–1319.

Tou, J. T., & Gonzalez, R. C. (1974). *Pattern recognition principles.* Reading, MA: Addison-Wesley.

Williams, C. K. I., & Seeger, M. (2000). The effect of the input density distribution on kernel-based classifiers. In P. Langley (Ed.), *Proceedings of the Seventeenth International Conference on Machine Learning, 2000* (pp. 1159–1166). San Mateo, CA: Morgan Kaufmann.

Williams, C. K. I., & Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13* (pp. 682–688). Cambridge, MA: MIT Press.

Zhu, H., Williams, C. K. I., Rohwer, R. J., & Morciniec, M. (1998). Gaussian regression and optimal finite dimensional linear models. In C. M. Bishop (Ed.), *Neural networks and machine learning.* Berlin: Springer-Verlag.

Received September 26, 2000; accepted May 30, 2001.