On the Eigenspectrum of the Gram Matrix and the Generalisation Error of Kernel PCA

John Shawe-Taylor
School of Electronics and Computer Science, University of Southampton
Email: jst@ecs.soton.ac.uk
Christopher K.I. Williams
Division of Informatics, University of Edinburgh
Email: ckiw@dai.ed.ac.uk
Nello Cristianini
Department of Statistics, University of California at Davis
Email: nello@wald.ucdavis.edu
Jaz Kandola
Merrill Lynch Quantitative Analytics Division
2 King Edward Street London EC1A 1HQ
Email: Jasvinder_Kandola@ml.com

***Abstract*— In this paper we analyze the relationships between the eigenvalues of the $m \times m$ Gram matrix K for a kernel $\kappa(\cdot, \cdot)$ corresponding to a sample $\mathbf{x}_1, \ldots, \mathbf{x}_m$ drawn from a density $p(\mathbf{x})$ and the eigenvalues of the corresponding continuous eigenproblem. We bound the differences between the two spectra and provide a performance bound on kernel PCA showing that we can expect good performance even in very high dimensional feature spaces provided the sample eigenvalues fall sufficiently quickly.**

I. Introduction

Over recent years there has been a considerable amount of interest in kernel methods such as Support Vector Machines [1], Gaussian Processes *etc* in the machine learning area. In these methods the *Gram matrix* plays an important rôle. The $m \times m$ Gram matrix K has entries $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, m$, where $\{\mathbf{x}_i : i = 1, \ldots, m\}$ is a given dataset and $\kappa(\cdot, \cdot)$ is a kernel function. For Mercer kernels K is symmetric positive semi-definite. We denote its eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \ldots \geq \hat{\lambda}_m \geq 0$ and write its eigendecomposition as $K = V \hat{\Lambda} V'$ where $\hat{\Lambda}$ is a diagonal matrix of the eigenvalues and V' denotes the transpose of matrix V. The eigenvalues are also referred to as the spectrum of the Gram matrix, while the corresponding columns of V are their eigenvectors.

A number of learning algorithms rely on estimating spectral data on a sample of training points and using this data as input to further analyses. For example in Principal Component Analysis (PCA) the subspace spanned by the first k eigenvectors is used to give a k dimensional model of the data with minimal residual, hence forming a low dimensional representation of the data for analysis or clustering. Recently the approach has been applied in kernel defined feature spaces in what has become known as kernel-PCA [2]. This representation has also been related to an Information Retrieval algorithm known as latent semantic indexing, again with kernel defined feature spaces [3].

Furthermore eigenvectors have been used in the HITS [4] and Google's PageRank [5] algorithms. In both cases the entries in the eigenvector corresponding to the maximal eigenvalue are interpreted as authority weightings for individual articles or web pages.

The use of these techniques raises the question of how reliably these quantities can be estimated from a random sample of data, or phrased differently, how much data is required to obtain an accurate empirical estimate with high confidence. Ng *et al.* [6] have undertaken a study of the sensitivity of the estimate of the first eigenvector to perturbations of the connection matrix. They have also highlighted the potential instability that can arise when two eigenvalues are very close in value, so that their eigenspaces become very difficult to distinguish empirically.

Other authors have studied the concentration of linear functionals of the spectral measure or single eigenvalues of random matrices generated through distributions defined over their entries, see for example Guionnet and Zeitouni [7] and Alon *et al.* [8].

In this paper we shift the emphasis towards studying the concentration of sums of eigenvalues of a Gram matrix gained from a finite sample of vectors, so that the distribution over the matrices is defined implicitly by a distribution over vectors. In particular if we perform (kernel-) PCA on a random sample and project new data into the k-dimensional space spanned by the first k eigenvectors, how much of the data will be captured or in other words how large will the residuals be. It turns out that this accuracy is not sensitive to the eigenvalue separation, while at the same time being the quantity that is relevant in a practical application of dimensionality reduction using kernel PCA. The result shows that we can expect good performance even in very high dimensional feature spaces provided that the sample eigenvalues fall sufficiently quickly. In this sense the results give a dimension independent bound on the performance of kernel PCA.

The second question that motivated the research reported in this paper is the relation between the eigenvalues of the Gram matrix and those of the underlying process. For a given kernel function and density $p(\mathbf{x})$ on a space $\mathcal{X}$, we can also write down the eigenfunction problem

$$\int_{\mathcal{X}} \kappa(\mathbf{x},\mathbf{y})p(\mathbf{x})\phi_i(\mathbf{x})\, d\mathbf{x} = \lambda_i \phi_i(\mathbf{y}). \tag{1}$$

Note that the eigenfunctions are orthonormal with respect to $p(\mathbf{x})$, i.e.

$$\int_{\mathcal{X}} \phi_i(\mathbf{x})p(\mathbf{x})\phi_j(\mathbf{x})d\mathbf{x} = \delta_{ij}.$$

Let the eigenvalues of the underlying process be ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots$. This continuous eigenproblem can be approximated in the following way. Let $\{\mathbf{x}_i \colon i = 1, \ldots, m\}$ be a sample drawn according to $p(\mathbf{x})$. Then

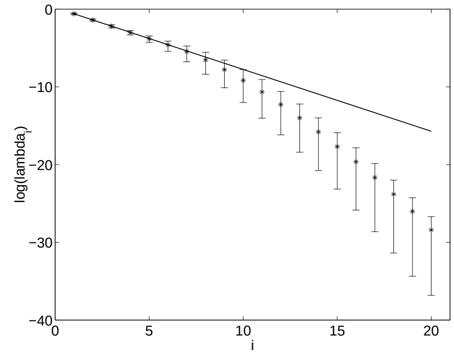
$$\int_{\mathcal{X}} \kappa(\mathbf{x},\mathbf{y})p(\mathbf{x})\phi_i(\mathbf{x})d\mathbf{x} \simeq \frac{1}{m}\sum_{k=1}^{m} \kappa(\mathbf{x}_k,\mathbf{y})\phi_i(\mathbf{x}_k) \tag{2}$$

As pointed out in [9], the standard numerical method (see, e.g., [10], chapter 3) for approximating the eigenfunctions and eigenvalues of equation (1) is to use a numerical approximation such as equation (2) to estimate the integral, and then plug in $\mathbf{y} = \mathbf{x}_j$ for $j = 1, \ldots, m$ to obtain a matrix eigenproblem

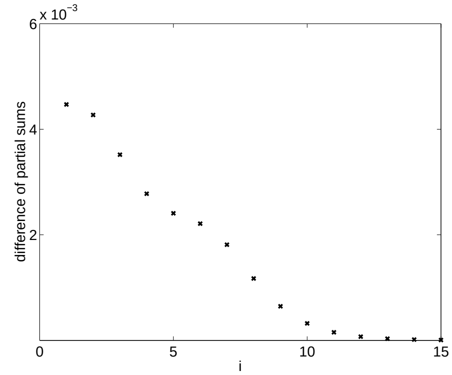
$$\sum_{k=1}^{m} \kappa(\mathbf{x}_k,\mathbf{x}_j)\phi_i(\mathbf{x}_k) = \hat{\lambda}_i \phi_i(\mathbf{x}_j).$$

Thus we see that $\mu_i \stackrel{def}{=} \frac{1}{m}\hat{\lambda}_i$ is an obvious estimator for the ith eigenvalue of the continuous problem. The theory of the numerical solution of eigenvalue problems ([10], Theorem 3.4) shows that for a fixed k, μ_k will converge to λ_k in the limit as $m \to \infty$.

For the case that $\mathcal{X}$ is one dimensional and $p(x)$ is Gaussian and $\kappa(x,y) = \exp -b(x-y)^2$ (the RBF kernel with length-scale $b^{-1/2}$), there are analytic results for the eigenvalues and eigenfunctions of equation (1) as given in section 4 of [11]. To compare the process eigenvalues with empirical eigenvalues 1000 samples of size $m = 100$ were used, with parameters $b = 3$ and $p(x) \sim \mathcal{N}(0, 1/4)$. The 1000 repetitions were used to characterize the variability of the empirical eigenvalues. For this case we can therefore compare the values of μ_i with the corresponding λ_i, as shown in Figure 1(a). Figure 1(b) plots the difference between the average (over 1000 samples) of the partial sum of the first i empirical eigenvalues against the same partial sum of the process eigenvalues. These two plots show that for $i = 1$ the average empirical eigenvalue overestimates λ_1, but that for $i > 1$ the converse is true. Figure 1(b) also shows that the empirical partial sum initially overestimates the process partial sum, but that this gradually declines. One of the results of this paper will be bounds on the degree of overestimation for these partial sums in a fully general setting.



(a)



(b)

Fig. 1. (a) A plot of the log eigenvalue against the index of the eigenvalue. The straight line is the theoretical relationship. The centre point (marked with a star) in the error bar is the log of the average value of μ_k. The upper and lower ends of the error bars are the 97.5% and 2.5% centiles of of $\log(\mu_k)$ respectively taken over 1000 repetitions. (b) A plot of the difference between the average of $\sum_{j=1}^{i}\mu_j$ and $\sum_{j=1}^{i}\lambda_j$ against i.

Koltchinskii and Gine [12] discuss a number of results including rates of convergence of the μ-spectrum to the λ-spectrum. The measure they use compares the whole spectrum rather than individual eigenvalues or subsets of eigenvalues. They also do not deal with the estimation problem for PCA residuals.

Johnstone [13] studies the distribution of the largest eigenvalue of the Gram matrix of a set of vectors whose components are independent Gaussians, though his is also an asymptotic analysis as the dimension of the feature space and the number of vectors tends to infinity at a fixed ratio greater than 1.

In an earlier version of this paper, [14] discussed the concentration of spectral properties of Gram matrices and of the residuals of fixed projections. However, these results gave deviation bounds on the sampling variability of μ_i with respect to $\mathbb{E}[\mu_i]$, but did not address the relationship of μ_i to λ_i or the estimation problem of the residual of PCA on new data.

In order to state our main results consider a general probability space $\mathcal{X}$ and a measurable feature mapping $\boldsymbol{\psi}$

$$\boldsymbol{\psi} : \mathbf{x} \in \mathcal{X} \longmapsto \boldsymbol{\psi}(\mathbf{x}) \in F$$

to a real Hilbert space F. We assume a probability measure p on the space $\mathcal{X}$. Note that this implies a distribution on F via the measurable feature map $\boldsymbol{\psi}$. We will assume throughout that the support of this distribution is bounded in a ball of radius R in F. We draw an i.i.d. sample S of m points

$$S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$$

from $\mathcal{X}$ according to p and form the Gram matrix $K(S)$ of their projections into F

$$K(S)_{ij} = \langle \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}_j) \rangle .$$

We refer to the composition of the inner product with the projections as the kernel function κ:

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{z}) \rangle ,$$

and similarly to the matrix $K(S)$ as the kernel matrix. It is often convenient to specify the kernel κ and define the feature space implicitly by this choice. Such a feature space will exist provided the kernel is symmetric and has the property that all finite kernel matrices are positive semi-definite (see [15] for details). We refer to the eigenvalues $\hat{\lambda}_1(S) \geq \hat{\lambda}_2(S) \geq \dots \geq \hat{\lambda}_m(S)$ of $K(S)$ as the empirical eigenvalues dropping the dependency on S if this is clear from the context.

There is a corresponding self-adjoint operator in the inner product space $L^2_p(\mathcal{X})$ defined by

$$\mathcal{K}(f)(\mathbf{x}) = \int_{\mathcal{X}} f(\mathbf{x}')\kappa(\mathbf{x}, \mathbf{x}')dp(\mathbf{x}').$$

We refer to the eigenvalues of this operator as the process eigenvalues and denote them by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_i \geq \cdots$.

Given a sequence of numbers $\nu_1 \geq \nu_2 \geq \dots \geq \nu_m$, where m may be infinity, we use the notations

$$\nu^{>k} = \sum_{i=k+1}^{m} \nu_i \quad \text{and} \quad \nu^{\leq k} = \sum_{i=1}^{k} \nu_i$$

to denote the tail and initial sums respectively.

We must introduce a further definition before quoting the main results of the paper. This is concerned with the procedure known as Principal Components Analysis that projects multi-dimensional data in the feature space F onto the subspace $\hat{V}_k$ spanned by the first k eigenvectors of the correlation matrix

$$C(S) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\psi}(\mathbf{x}_i)\boldsymbol{\psi}(\mathbf{x}_i)'.$$

Note that we do not restrict the space F to be finite dimensional. However, for any finite set of points $\mathbf{x}_1, \dots, \mathbf{x}_m$, the feature vectors $\boldsymbol{\psi}(\mathbf{x}_1), \dots, \boldsymbol{\psi}(\mathbf{x}_m)$ span a finite dimensional subspace of F. Hence, by choosing a basis that spans this subspace and extending to a basis of the whole space, the correlation matrix $C(S)$ becomes effectively finite dimensional.

We denote projection onto a subspace V by $P_V(\boldsymbol{\psi}(\mathbf{x}))$. We denote the projection onto the orthogonal complement of V by $P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}))$. If V is a one dimensional subspace with $\mathbf{v}$ a non-zero element of V, we will also write $P_{\mathbf{v}}$ in place of P_V. The norm of the orthogonal projection is also referred to as the residual since it corresponds to the distance between the original point and its projection.

We can now state the three main results of this paper. The first is concerned with the residual projections and the sum of the last eigenvalues.

Theorem 1: If we perform PCA in the feature space defined by a kernel $\kappa(\mathbf{x}, \mathbf{z})$ then with probability greater than $1 - \delta$ over random m-samples S, for all $1 \leq k \leq m$, if we project new data onto the space $\hat{V}_k$, the expected squared residual is bounded by

$$\begin{aligned} \lambda^{>k} &\leq \mathbb{E}\left[\|P^{\perp}_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\ &\leq \min_{1\leq \ell \leq k}\left[\frac{1}{m}\hat{\lambda}^{>\ell}(S) + \frac{1+\sqrt{\ell}}{\sqrt{m}}\sqrt{\frac{2}{m}\sum_{i=1}^{m}\kappa(\mathbf{x}_i, \mathbf{x}_i)^2}\right] \\ &\qquad + R^2\sqrt{\frac{18}{m}\ln\left(\frac{2m}{\delta}\right)} \end{aligned}$$

where the support of the distribution is in a ball of radius R in the feature space and λ_i and $\hat{\lambda}_i$ are the process and empirical eigenvalues respectively.

The theorem states that when projecting into the empirical eigen-subspace spanned by the first k eigenvectors the expected squared residual of a randomly drawn test point can with high probability be bounded by a minimum over $\ell \leq k$ of the sum of all but the first ℓ empirical eigenvalues plus a complexity term that scales like $\sqrt{\ell/m}$.

The last term on the right hand side represents the usual dependency on the confidence parameter δ. The expression inside the minimisation involves two terms. The first term is the empirical estimate of the squared residual, which decreases as ℓ increases. The second term is the complexity penalty that grows with increasing ℓ. The expression will reach a minimum at a value ℓ_0 approximately where the two expressions have equal values. Hence, the overall bound decreases as k increases up to ℓ_0 and remains constant from that point onwards. In practice we expect that the left hand side will continue to decline slowly beyond this point as further dimensions are included. This effect is indeed evident in the experiments reported in the final section.

For applications of kernel PCA the theorem suggests that good capture of the data can be expected provided the empirical eigenvalues decay before $\sqrt{\ell/m}$ grows too big. Indeed this can be used as a criterion for deciding whether subspace projection is justified based on the available training data.

The second theorem considers the sum of the first k eigenvalues and the projections into the space spanned by the first k eigenvectors.

Theorem 2: If we perform PCA in the feature space defined by a kernel $\kappa(\mathbf{x}, \mathbf{z})$ then with probability greater than $1 - \delta$ over random m-samples S, for all $1 \leq k \leq m$, if we project new data onto the space $\hat{V}_k$, the sum of the largest k process eigenvalues is bounded by

$$\begin{aligned}\lambda^{\leq k} &\geq \mathbb{E}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\ &\geq \max_{1\leq \ell\leq k}\left[\frac{1}{m}\hat{\lambda}^{\leq \ell}(S) - \frac{1+\sqrt{\ell}}{\sqrt{m}}\sqrt{\frac{2}{m}\sum_{i=1}^m \kappa(\mathbf{x}_i,\mathbf{x}_i)^2}\right] \\ &\quad -R^2\sqrt{\frac{19}{m}\ln\left(\frac{2(m+1)}{\delta}\right)}\end{aligned}$$

where the support of the distribution is in a ball of radius R in the feature space and λ_i and $\hat{\lambda}_i$ are the process and empirical eigenvalues respectively.

This result is perhaps more interesting from the perspective of the relation between process and empirical eigenvalues. In particular, it implies a good fit between the partial sums of the largest eigenvalues with indices k for which $\sqrt{k/m}$ is small.

The final result concerns the projections of data into the 1-dimensional subspace determined by a single eigenvector. In this case it is not possible to obtain a relationship with the process eigenvalues, but the 'generalisation' of the empirical projection obeys an even tighter bound than for the larger subspaces.

Theorem 3: If we perform PCA in the feature space defined by a kernel $\kappa(\mathbf{x},\mathbf{z})$ then with probability greater than $1-\delta$ over random m-samples S, for all $1 \leq k \leq m$, if we project new data onto the one dimensional subspace $\hat{U}_k$ spanned by the k-th eigenvector of $C(S)$, the expected value of the projection of new data satisfies

$$\begin{aligned}\mathbb{E}\left[\|P_{\hat{U}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &\geq \frac{1}{m}\hat{\lambda}_k(S) - \frac{2}{\sqrt{m}}\sqrt{\frac{2}{m}\sum_{i=1}^m \kappa(\mathbf{x}_i,\mathbf{x}_i)^2} \\ &\quad -R^2\sqrt{\frac{19}{m}\ln\left(\frac{2(m+1)}{\delta}\right)}\end{aligned}$$

where the support of the distribution is in a ball of radius R in the feature space and $\hat{\lambda}_i$ are the empirical eigenvalues.

The paper is organised as follows. In Section 2 we give the background results and develop the basic techniques that are required to develop the necessary framework in sections 3 and 4. Section 5 then gives the main results of the paper. We provide experimental verification of the theoretical findings in Section 6, before drawing our conclusions.

II. Background and Techniques

We will make use of the following results that can be traced back to the work of Hoeffding [16] and Azuma [17]. We quote versions given by McDiarmid [18]. Results of this type bounding the deviation of a random variable from its expected value are often referred to as concentration inequalities. More advanced results of this type due to Boucheron *et al.* and Talagrand can be found in [19] and [20].

Theorem A: Let $X_1,\ldots,X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$, and that there exist $f_i : A^{n-1} \to \mathbb{R}$ for $1 \leq i \leq n$ satisfying

$$\sup_{x_1,\ldots,x_n} |f(x_1,\ldots,x_n) - f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)| \leq c_i,$$

then for all $\epsilon > 0$,

$$P\{f(X_1,\ldots,X_n) - \mathbb{E}f(X_1,\ldots,X_n) > \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

Theorem B: Let $X_1,\ldots,X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$, for $1 \leq i \leq n$

$$\sup_{x_1,\ldots,x_n,\hat{x}_i} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},\hat{x}_i,x_{i+1},\ldots,x_n)| \leq c_i,$$

then for all $\epsilon > 0$,

$$P\{f(X_1,\ldots,X_n) - \mathbb{E}f(X_1,\ldots,X_n) > \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

We will also make use of the following theorem characterising the eigenvectors of a self-adjoint completely continuous operator in a Hilbert space. This theorem is usually referred to as the Courant Fischer Weyl theorem in its matrix version. We quote it here in the more general form [21].

Theorem C: [Courant-Fischer-Weyl Minimax Theorem] If T is a self-adjoint completely continuous operator on a real Hilbert space, then for $k = 1, 2, \ldots,$

$$\begin{aligned}\lambda_k(T) &= \max_{\dim(V)=k}\ \min_{0\neq\mathbf{v}\in V} \frac{\langle T\mathbf{v},\mathbf{v}\rangle}{\langle \mathbf{v},\mathbf{v}\rangle} \\ &= \min_{\dim(T)=m-k+1}\ \max_{0\neq\mathbf{v}\in T} \frac{\langle T\mathbf{v},\mathbf{v}\rangle}{\langle \mathbf{v},\mathbf{v}\rangle},\end{aligned}$$

with the extrema achieved by the corresponding eigenvector.

The approach we adopt in the first stage of the analysis is to relate the eigenvalues to the sums of squares of residuals. This is well-known particularly in the case of matrices, following from consideration of the singular value decomposition. We sketch the analysis in the more general operator form since we require this for the process eigenvalues mentioned above. The matrix form is a simple consequence of this general result.

Recall the operator of the form

$$\mathcal{K}_q(f)(\mathbf{x}) = \int_{\mathcal{X}} f(\mathbf{x}')\kappa(\mathbf{x},\mathbf{x}')dq(\mathbf{x}'),$$

in the space $L_q^2(\mathcal{X})$, where q is some distribution over $\mathcal{X}$. Furthermore consider the self-adjoint operator

$$\mathcal{C}_q(\cdot) = \int_{\mathcal{X}} \langle \boldsymbol{\psi}(\mathbf{x}), \cdot\rangle \boldsymbol{\psi}(\mathbf{x}) dq(\mathbf{x}).$$

Let $v(\cdot), \lambda$ be an eigenfunction, eigenvalue pair for $\mathcal{K}_q$, that is $\mathcal{K}_q(v)(\mathbf{x}) = \lambda v(\mathbf{x})$. Consider the point

$$\mathbf{u} = f_q(v) = \int_{\mathcal{X}} v(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})dq(\mathbf{x}) \in F.$$

We have

$$\begin{aligned}\mathcal{C}_q(\mathbf{u}) &= \int_{\mathcal{X}}\int_{\mathcal{X}} \kappa(\mathbf{x},\mathbf{z})v(\mathbf{z})dq(\mathbf{z})\boldsymbol{\psi}(\mathbf{x})dq(\mathbf{x}) \\ &= \lambda\int_{\mathcal{X}} v(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})dq(\mathbf{x}) \\ &= \lambda\mathbf{u}.\end{aligned}$$

It follows that $f_q(v), \lambda$ is an eigenvector, eigenvalue pair for $\mathcal{C}_q$. Furthermore, we have

$$\begin{aligned}\|f_q(v)\|^2 &= \int_{\mathcal{X}}\int_{\mathcal{X}} v(\mathbf{x})v(\mathbf{z})\kappa(\mathbf{x},\mathbf{z})dq(\mathbf{x})dq(\mathbf{z}) \\ &= \lambda\int_{\mathcal{X}} v(\mathbf{z})^2 dq(\mathbf{z}) = \lambda\|v\|_q^2,\end{aligned}$$

in the norm determined by the distribution q. Similarly it is easily verified that if $\mathbf{u}, \lambda$ is an eigenvector, eigenvalue pair for $\mathcal{C}_q$ the function

$$g(\mathbf{u})(\cdot) = \langle \boldsymbol{\psi}(\cdot), \mathbf{u}\rangle$$

is an eigenfunction for $\mathcal{K}_q$ with eigenvalue λ and

$$\|g(\mathbf{u})\|_q^2 = \lambda\|\mathbf{u}\|^2.$$

Furthermore, we have that

$$g(f_q(v)) = \mathcal{K}_q(v) \quad \text{and} \quad f_q(g(\mathbf{u})) = \mathcal{C}_q(\mathbf{u}).$$

It follows from this analysis that the two operators have the same non-zero eigenvalues and there is a 1-1 correspondence between the corresponding eigenvectors, eigenfunctions given by the functions f and g.

If we consider the case where q is the empirical distribution, that is the uniform distribution on a fixed m-sample S, we will see that this analysis forms the basis of kernel PCA. If we choose q to be the empirical distribution uniform on a fixed sample S, we will denote the operators $\mathcal{C}_q$ and $\mathcal{K}_q$ by $\mathcal{C}_S$ and $\mathcal{K}_S$ respectively.

If $\mathbf{u}_i, \lambda_i$ are the i-th normalised eigenvector, eigenvalue pair of the operator $\mathcal{C}_S$ in the feature space, this corresponds to the i-th eigenvector of the correlation matrix

$$C(S) = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\psi}(\mathbf{x}_i)\boldsymbol{\psi}(\mathbf{x}_i)'.$$

The PCA projection of an input $\mathbf{x}$ onto $\mathbf{u}_i$ is given by

$$\begin{aligned}\langle \boldsymbol{\psi}(\mathbf{x}), \mathbf{u}_i\rangle &= \lambda_i^{-1/2}\langle \boldsymbol{\psi}(\mathbf{x}), f_q(v_i)\rangle \\ &= \lambda_i^{-1/2} m^{-1}\sum_{j=1}^m v_i(\mathbf{x}_j)\kappa(\mathbf{x}_j,\mathbf{x}),\end{aligned}$$

where $v_i(\cdot), \lambda_i$ are the corresponding eigenfunction, eigenvalue pair of the operator $\mathcal{K}_q$. This equation forms the basis of kernel PCA, since it implies that the projection of a new point into the space spanned by the i-th eigenvector can be computed as

$$P_{\mathbf{u}_i}(\boldsymbol{\psi}(\mathbf{x})) = \left(\hat{\lambda}_i^{-1/2}\sum_{j=1}^m \mathbf{v}_{ij}\kappa(\mathbf{x},\mathbf{x}_j)\right)\mathbf{u}_i,$$

where $(\mathbf{v}_{ij})_{j=1}^m$, $\hat{\lambda}_i$ are the i-th eigenvector and eigenvalue of the kernel matrix $K(S)$.

Now consider the first eigenvalue of the operator $\mathcal{K}_q$ for general distribution q. By Theorem C and the above observations we have

$$\begin{aligned}\lambda_1(\mathcal{K}_q) &= \max_{0\neq\mathbf{v}\in F}\frac{\langle \mathcal{C}_q(\mathbf{v}),\mathbf{v}\rangle}{\langle \mathbf{v},\mathbf{v}\rangle} \\ &= \max_{0\neq\mathbf{v}\in F}\frac{1}{\|\mathbf{v}\|^2}\int_{\mathcal{X}}\langle \boldsymbol{\psi}(\mathbf{x}),\mathbf{v}\rangle^2 dq(\mathbf{x}) \\ &= \max_{0\neq\mathbf{v}\in F}\mathbb{E}_q\left[\|P_{\mathbf{v}}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\ &= \mathbb{E}_q\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \min_{0\neq\mathbf{v}\in F}\mathbb{E}_q\left[\|P_{\mathbf{v}}^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right],\end{aligned}$$

where $\mathbb{E}_q$ denotes expectation with respect to q, since $\|\boldsymbol{\psi}(\mathbf{x})\|^2 = \|P_{\mathbf{v}}(\boldsymbol{\psi}(\mathbf{x}))\|^2 + \|P_{\mathbf{v}}^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2$. It follows that the first eigenvector is characterised as the direction for which the expected square of the residual is minimal.

Applying the same line of reasoning to the first equality of Theorem C, delivers the following equality

$$\lambda_k(\mathcal{K}_q) = \max_{\dim(V)=k, V\subseteq F}\min_{0\neq\mathbf{v}\in V}\mathbb{E}_q\left[\|P_{\mathbf{v}}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]. \qquad (3)$$

Notice that this characterisation implies that if $\mathbf{u}_k$ is the k-th eigenvector of $\mathcal{C}_q$, then

$$\lambda_k(\mathcal{K}_q) = \mathbb{E}_q\left[\|P_{\mathbf{u}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right], \qquad (4)$$

which in turn implies that if V_k is the space spanned by the first k eigenvectors, then

$$\begin{aligned}\sum_{i=1}^k \lambda_i(\mathcal{K}_q) &= \mathbb{E}_q\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\ &= \mathbb{E}_q\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \mathbb{E}_q\left[\|P_{V_k}^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]. \qquad (5)\end{aligned}$$

It readily follows by induction over the dimension of V that we can equally characterise the sum of the first k and last $m-k$ eigenvalues by

$$\begin{aligned}\sum_{i=1}^k \lambda_i(\mathcal{K}_q) &= \max_{\dim(V)=k}\mathbb{E}_q\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\ &= \mathbb{E}_q\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \min_{\dim(V)=k}\mathbb{E}_q\left[\|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right], \qquad (6) \\ \sum_{i=k+1}^{\infty}\lambda_i(\mathcal{K}_q) &= \mathbb{E}_q\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \sum_{i=1}^k \lambda_i(\mathcal{K}_q) \qquad (7) \\ &= \min_{\dim(V)=k}\mathbb{E}_q\left[\|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]. \qquad (8)\end{aligned}$$

Hence, as for the case when $k=1$, the subspace spanned by the first k eigenvalues is characterised as that for which the sum of the squares of the residuals is minimal.

In the case that q is the empirical distribution the results correspond to the matrix form of the residual result, namely that projecting into the eigenspaces corresponding to the largest eigenvalues minimises the average squared residual. If we take q to be the data generating distribution p, the result describes the fact that the eigenvectors of the operator $\mathcal{C}_p$ characterise the subspaces of F capturing the largest expected squared residual:

$$\lambda_k(\mathcal{K}) = \max_{\dim(V)=k}\min_{0\neq\mathbf{v}\in V}\mathbb{E}[\|P_{\mathbf{v}}(\boldsymbol{\psi}(\mathbf{x}))\|^2], \qquad (9)$$

where V is a linear subspace of the feature space F and we use $\mathbb{E}$ to denote expectation with respect to p. Similarly,

$$\begin{aligned}\sum_{i=1}^{k} \lambda_i(\mathcal{K}) &= \max_{\dim(V)=k} \mathbb{E}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\ &= \mathbb{E}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \min_{\dim(V)=k} \mathbb{E}\left[\|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \quad (10) \\ \sum_{i=k+1}^{\infty} \lambda_i(\mathcal{K}) &= \mathbb{E}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \sum_{i=1}^{k} \lambda_i(\mathcal{K}) \\ &= \min_{\dim(V)=k} \mathbb{E}\left[\|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]. \quad (11)\end{aligned}$$

One of the aims of this paper is to elucidate the relationship between these two projections, demonstrating conditions when the quality of the empirical projection matches that of the 'ideal' process projection.

We are now in a position to motivate the main results of the paper. We consider the general case of a kernel defined feature space with input space $\mathcal{X}$ and probability density $p(\mathbf{x})$. We fix a sample size m and a draw of m examples $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ according to p. We fix the feature space determined by the kernel as given by the mapping $\boldsymbol{\psi}$. We can therefore view the eigenvectors of correlation matrices corresponding to finite Gram matrices as lying in this space. Further we fix a feature dimension k. Let $\hat{V}_k$ be the space spanned by the first k eigenvectors of the correlation matrix corresponding to the sample kernel matrix $K(S)$ with corresponding eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k$, while V_k is the space spanned by the first k process eigenvectors with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Similarly, let $\hat{\mathbb{E}}[f(\mathbf{x})]$ denote the expectation with respect to the sample or the empirical mean:

$$\hat{\mathbb{E}}[f(\mathbf{x})] = \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i),$$

while as before $\mathbb{E}[\cdot]$ denotes expectation with respect to p.

We are interested in the relationships between the following quantities:

$$\begin{aligned}\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &= \frac{1}{m}\sum_{j=1}^{m} \|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2 \\ &= \frac{1}{m}\sum_{i=1}^{k} \hat{\lambda}_i \\ \mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &= \sum_{i=1}^{k} \lambda_i \\ \mathbb{E}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \quad &\text{and} \quad \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right].\end{aligned}$$

Bounding the difference between the first and second will relate the process eigenvalues to the sample eigenvalues, while the difference between the first and third will bound the expected performance of the space identified by kernel PCA when used on new data.

Our first two observations follow simply from equation (10),

$$\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \frac{1}{m}\sum_{i=1}^{k} \hat{\lambda}_i \geq \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \quad (12)$$

and

$$\mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \sum_{i=1}^{k} \lambda_i \geq \mathbb{E}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \quad (13)$$

Our strategy will be to show that the right hand side of inequality (12) and the left hand side of inequality (13) are close in value making the two inequalities approximately a chain of inequalities. We then bound the difference between the first and last entries in the chain.

First, however, in the next section we will examine averages over random m samples. We will use the notation $\mathbb{E}_m[\cdot]$ to denote this type of average though we could equivalently write $\mathbb{E}^m[\cdot]$ in the sense that this is simply the expectation with respect to the m-fold product distribution.

III. Averaging over Samples and Population Eigenvalues

The sample correlation matrix is $C(S) = \frac{1}{m}XX'$ with eigenvalues $\mu_1 \geq \mu_2 \ldots \geq \mu_d$. (If $\mathbf{x}$ is a zero-mean random variable then this is also the covariance matrix.) In the notation of the section II $\mu_i = (1/m)\hat{\lambda}_i$. The corresponding population correlation matrix has eigenvalues $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_d$ and eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_d$. Again by the observations above these are the process eigenvalues.

Statisticians have been interested in the sampling distribution of the eigenvalues of $C(S)$ for some time. There are two main approaches to studying this problem, as discussed in section 6 of [22]. In the case that $\mathbf{x}$ has a multivariate normal distribution, the exact sampling distribution of $\mu_1, \ldots, \mu_d$ can be given [23]. Alternatively, the "delta method" can be used, expanding the sample roots about the population roots. For normal populations this has been carried out in [24] (if there are no repeated roots of the population covariance) and [25] (for the general case), and extended in [26] to the non-Gaussian case.

The following proposition describes how $\mathbb{E}_m[\mu_1]$ is related to λ_1 and $\mathbb{E}_m[\mu_d]$ is related to λ_d. It requires no assumption of Gaussianity.

Proposition A: [Anderson, 1963, pp 145-146] $\mathbb{E}_m[\mu_1] \geq \lambda_1$ and $\mathbb{E}_m[\mu_d] \leq \lambda_d$.

Proof: By the results of the previous section we have

$$\begin{aligned}\mu_1 &= \max_{0 \neq \mathbf{c}} \sum_{i=1}^{m} \frac{1}{m}\|P_{\mathbf{c}}(\mathbf{x}_i)\|^2 \\ &\geq \frac{1}{m}\sum_{i=1}^{m} \|P_{\mathbf{u}_1}(\mathbf{x}_i)\|^2 = \hat{\mathbb{E}}\left[\|P_{\mathbf{u}_1}(\mathbf{x})\|^2\right].\end{aligned}$$

We now apply the expectation operator $\mathbb{E}_m$ to both sides. On the RHS we get

$$\mathbb{E}_m\hat{\mathbb{E}}\left[\|P_{\mathbf{u}_1}(\mathbf{x})\|^2\right] = \mathbb{E}\left[\|P_{\mathbf{u}_1}(\mathbf{x})\|^2\right] = \lambda_1$$

by equation (11), which completes the proof. Correspondingly μ_d is characterized by $\mu_d = \min_{0 \neq \mathbf{c}} \hat{\mathbb{E}}\left[\|P_{\mathbf{c}}(\mathbf{x}_i)\|^2\right]$ (minor components analysis). □

Interpreting this result, we see that $\mathbb{E}_m[\mu_1]$ *overestimates* λ_1, while $\mathbb{E}_m[\mu_d]$ *underestimates* λ_d.

Proposition A can be generalized to give the following result where we have also allowed for a kernel defined feature space of dimension $N_F \leq \infty$.

Proposition 4: Using the above notation, for any k, $1 \leq k \leq m$, $\mathbb{E}_m[\sum_{i=1}^k \mu_i] \geq \sum_{i=1}^k \lambda_i$ and $\mathbb{E}_m[\sum_{i=k+1}^m \mu_i] \leq \sum_{i=k+1}^{N_F} \lambda_i$.

Proof: Let V_k be the space spanned by the first k process eigenvectors. Then from the derivations above we have

$$\sum_{i=1}^k \mu_i = \max_{V:\ \dim V = k} \hat{\mathbb{E}}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \geq \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right].$$

Again, applying the expectation operator $\mathbb{E}_m$ to both sides of this equation and taking equation (11) into account, the first inequality follows. To prove the second we turn $\max$ into $\min$, P into $P^\perp$ and reverse the inequality. Again taking expectations of both sides proves the second part. □

Furthermore, [26] (equation 2) gives the asymptotic relationship

$$\mathbb{E}_m[\mu_i] = \lambda_i + \frac{1}{m}\sum_{j=1, j\neq i} \frac{\lambda_i\lambda_j + \kappa_{22}^{ij}}{\lambda_i - \lambda_j} + O(m^{-2}), \qquad (14)$$

where κ_{22}^{ij} is the bivariate cumulant of order 4 of the marginal distribution of ϕ_i and ϕ_j (assumed finite).

Remark 5: Proposition 4 also implies that

$$\mathbb{E}_{N_F}\left[\sum_{i=1}^{N_F}\mu_i\right] = \sum_{i=1}^{N_F}\lambda_i$$

if we sample N_F points.

We can tighten this relation and obtain another relationship from the trace of the matrix when the support of p satisfies $\kappa(\mathbf{x},\mathbf{x}) = C$, a constant. For example if the kernel is stationary, this holds since $\kappa(\mathbf{x},\mathbf{x}) = \kappa(\mathbf{x}-\mathbf{x}) = \kappa(\mathbf{0}) = C$. Thus

$$\mathrm{trace}\left(\frac{1}{m}K\right) = C = \sum_{i=1}^m \mu_i.$$

Also we have for the continuous eigenproblem $\int_{\mathcal{X}} \kappa(\mathbf{x},\mathbf{x})p(\mathbf{x})d\mathbf{x} = C$. Using the feature expansion representation of the kernel $\kappa(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{N_F}\lambda_i\phi_i(\mathbf{x})\phi_i(\mathbf{y})$ and the orthonormality of the eigenfunctions we obtain the following result

$$\sum_{i=1}^m \mu_i = \sum_{i=1}^{N_F}\lambda_i.$$

Applying the results obtained in this section, it follows that $\mathbb{E}_m[\mu_1]$ will overestimate λ_1, and the cumulative sum $\sum_{i=1}^k \mathbb{E}_m[\mu_i]$ will overestimate $\sum_{i=1}^k \lambda_i$. This behaviour is illustrated in Figure 1(b). At the other end, clearly for $N_F \geq k > m$, $\mu_k \equiv 0$ is an underestimate of λ_k.■

IV. Concentration of eigenvalues

Section II outlined the relatively well-known perspective that we now apply to obtain the concentration results for the eigenvalues of positive semi-definite matrices. The key to the results is the characterisation in terms of the sums of residuals given in equations (3) and (8).

Theorem 6: Let $\kappa(\mathbf{x},\mathbf{z})$ be a positive semi-definite kernel function on a space X, and let p be a probability density function on X. Fix natural numbers m and $1 \leq k < m$ and let $S = (\mathbf{x}_1,\ldots,\mathbf{x}_m) \in X^m$ be a sample of m points drawn according to p. Then for all $\epsilon > 0$,

$$P\left\{\left|\frac{1}{m}\hat{\lambda}_k(S) - \mathbb{E}_m\left[\frac{1}{m}\hat{\lambda}_k(S)\right]\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-2\epsilon^2 m}{R^4}\right),$$

where $\hat{\lambda}_k(S)$ is the k-th eigenvalue of the matrix $K(S)$ with entries $K(S)_{ij} = \kappa(\mathbf{x}_i,\mathbf{x}_j)$ and $R^2 = \max_{\mathbf{x}\in X}\kappa(\mathbf{x},\mathbf{x})$.

Proof: The result follows from an application of Theorem A provided

$$\sup_S\left|\frac{1}{m}\hat{\lambda}_k(S) - \frac{1}{m}\hat{\lambda}_k(S\setminus\{\mathbf{x}_i\})\right| \leq R^2/m.$$

Let $\hat{S} = S \setminus \{\mathbf{x}_i\}$ and let V ($\hat{V}$) be the k dimensional subspace spanned by the first k eigenvectors of $\mathcal{C}_S$ ($\mathcal{C}_{\hat{S}}$). Let κ correspond to the feature mapping $\boldsymbol{\psi}$. Using m times equation (3) for the empirical distribution we have

$$\begin{aligned}
\hat{\lambda}_k(S) &\geq \min_{v\in\hat{V}}\sum_{j=1}^m \|P_v(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 \\
&\geq \min_{v\in\hat{V}}\sum_{j\neq i} \|P_v(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 = \hat{\lambda}_k(\hat{S}) \\
\hat{\lambda}_k(\hat{S}) &\geq \min_{v\in V}\sum_{j\neq i} \|P_v(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 \\
&\geq \min_{v\in V}\sum_{j=1}^m \|P_v(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 - R^2 = \hat{\lambda}_k(S) - R^2. \ \square
\end{aligned}$$

Surprisingly a very similar result holds when we consider the sum of the last $m-k$ eigenvalues or the first k eigenvalues.

Theorem 7: Let $\kappa(\mathbf{x},\mathbf{z})$ be a positive semi-definite kernel function on a space X, and let p be a probability density function on X. Fix natural numbers m and $1 \leq k < m$ and let $S = (\mathbf{x}_1,\ldots,\mathbf{x}_m) \in X^m$ be a sample of m points drawn according to p. Then for all $\epsilon > 0$,

$$P\left\{\left|\frac{1}{m}\hat{\lambda}^{>k}(S) - \mathbb{E}_m\left[\frac{1}{m}\hat{\lambda}^{>k}(S)\right]\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-2\epsilon^2 m}{R^4}\right),$$

and

$$P\left\{\left|\frac{1}{m}\hat{\lambda}^{\leq k}(S) - \mathbb{E}_m\left[\frac{1}{m}\hat{\lambda}^{\leq k}(S)\right]\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-2\epsilon^2 m}{R^4}\right),$$

where $\hat{\lambda}^{\leq k}(S)$ ($\hat{\lambda}^{>k}(S)$) is the sum of (all but) the largest k eigenvalues of the matrix $K(S)$ with entries $K(S)_{ij} = \kappa(\mathbf{x}_i,\mathbf{x}_j)$ and $R^2 = \max_{\mathbf{x}\in X}\kappa(\mathbf{x},\mathbf{x})$.

Proof: The result follows from an application of Theorem A provided

$$\sup_S\left|\frac{1}{m}\hat{\lambda}^{>k}(S) - \frac{1}{m}\hat{\lambda}^{>k}(S\setminus\{\mathbf{x}_i\})\right| \leq R^2/m.$$

Let $\hat{S} = S \setminus \{\mathbf{x}_i\}$ and let V $(\hat{V})$ be the k dimensional subspace spanned by the first k eigenvectors of $\mathcal{C}_S$ $(\mathcal{C}_{\hat{S}})$. Let κ correspond to the feature mapping $\boldsymbol{\psi}$. Using m times equation (8) for the empirical distribution we have

$$\begin{aligned}
\hat{\lambda}^{>k}(S) &\leq \sum_{j=1}^{m} \|P_{\hat{V}}^{\perp}(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 \leq \sum_{j\neq i} \|P_{\hat{V}}^{\perp}(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 + R^2 \\
&= \hat{\lambda}^{>k}(\hat{S}) + R^2 \\
\lambda^{>k}(\hat{S}) &\leq \sum_{j\neq i} \|P_{V}^{\perp}(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 \\
&= \sum_{j=1}^{m} \|P_{V}^{\perp}(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 - \|P_{V}^{\perp}(\boldsymbol{\psi}(\mathbf{x}_i))\|^2 \leq \lambda^{>k}(S).
\end{aligned}$$

A similar derivation proves the second inequality. □

Corollary 8: Consider a feature space F defined by a kernel $\kappa(\mathbf{x}, \mathbf{z})$ in a space X with a distribution density $p(\mathbf{x})$. Furthermore let $\hat{\lambda}_i$, $i = 1, \ldots, m$ be the empirical eigenvalues. With probability $1-\delta$ over the selection of a random sample of m points drawn according to $p(\mathbf{x})$

$$\left| \frac{1}{m}\hat{\lambda}^{\leq k}(S) - \mathbb{E}_m\left[\frac{1}{m}\hat{\lambda}^{\leq k}(S)\right] \right| \leq R^2 \sqrt{\frac{1}{m} \ln \frac{2}{\delta}}$$

Our next result concerns the concentration of the residuals with respect to a fixed subspace.

Theorem 9: Let p be a probability density function on X. Fix natural numbers m and a subspace V and let $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in X^m$ be a sample of m points drawn according to a probability density function p. Then for all $\epsilon > 0$,

$$P\{|\hat{\mathbb{E}}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] - \mathbb{E}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]| \geq \epsilon\} \leq 2\exp\left(\frac{-\epsilon^2 m}{2R^4}\right).$$

Proof: Since we have that

$$\mathbb{E}_m\left[\hat{\mathbb{E}}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]\right] = \mathbb{E}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right],$$

the result follows from an application of Theorem B provided

$$\sup_{S,\hat{x}_i} \left| \hat{\mathbb{E}}_S\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] - \hat{\mathbb{E}}_{S\setminus\{\mathbf{x}_i\}\cup\{\hat{\mathbf{x}}_i\}}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \right| \leq R^2/m.$$

Clearly the largest change will occur if one of the points $\boldsymbol{\psi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\hat{\mathbf{x}}_i)$ lies in the subspace V and the other does not. In this case the change will be at most R^2/m. □

We apply the theorem to the subspace V_k spanned by the first k process eigenvalues to obtain the following corollary.

Corollary 10: Consider a feature space F defined by a kernel $\kappa(\mathbf{x}, \mathbf{z})$ in a space X with a distribution density $p(\mathbf{x})$. Furthermore let V_k be the subspace of F spanned by the first k process eigenvectors. With probability $1-\delta$ over the selection of a random sample of m points drawn according to $p(\mathbf{x})$

$$\left| \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] - \mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \right| \leq R^2 \sqrt{\frac{1}{m} \ln \frac{2}{\delta}}$$

The concentration results of this section are very tight. In the notation of the earlier sections they show that with high probability

$$\begin{aligned}
\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &= \frac{1}{m}\sum_{i=1}^{k} \hat{\lambda}_i \approx \mathbb{E}_m\left[\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]\right] \\
&= \mathbb{E}_m\left[\frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i\right] \\
\text{and} \qquad \mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &= \sum_{i=1}^{k} \lambda_i \\
&\approx \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right], \qquad (15)
\end{aligned}$$

where we have used Theorem 7 to obtain the first approximate equality and Theorem 9 with $V = V_k$ to obtain the second approximate equality.

This gives the sought relationship to create an approximate chain of inequalities

$$\begin{aligned}
\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &= \frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i \geq \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\
&\approx \mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \sum_{i=1}^{k}\lambda_i \\
&\geq \mathbb{E}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]. \qquad (16)
\end{aligned}$$

Notice that using Proposition 4 we also obtain the following diagram of approximate relationships

$$\begin{array}{ccc}
\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] & \geq & \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\
\approx & & \approx \\
\mathbb{E}_m\left[\frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i\right] & \geq & \mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right].
\end{array}$$

Hence, the approximate chain could have been obtained in two ways. It remains to bound the difference between the first and last entries in this chain. This together with the concentration results of this section will deliver the required bounds on the differences between empirical and process eigenvalues, as well as providing a performance bound on kernel PCA.

V. Learning a projection matrix

This section will work up to a proof of the three main results given in the introduction. The key observation that enables the analysis bounding the difference between

$$\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i$$

and $\mathbb{E}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]$ is that we can view the projection norm $\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2$ as a linear function of pairs of features from the feature space F.

Proposition 11: Let $\hat{V}$ be the subspace spanned by some fixed subset I of k eigenvectors of the kernel matrix. The projection norm $\|P_{\hat{V}}(\boldsymbol{\psi}(\mathbf{x}))\|^2$ is a linear function $\hat{f}$ in a feature space $\hat{\mathcal{F}}$ for which the kernel function is given by

$$\hat{\kappa}(\mathbf{x}, \mathbf{z}) = \kappa(\mathbf{x}, \mathbf{z})^2.$$

Furthermore the 2-norm of the function $\hat{f}$ is $\sqrt{k}$.

Proof: Let $X = U\Sigma V'$ be the singular value decomposition of the sample matrix X in the feature space. The projection norm is then given by

$$\hat{f}(\mathbf{x}) = \|P_{\hat{V}}(\boldsymbol{\psi}(\mathbf{x}))\|^2 = \boldsymbol{\psi}(\mathbf{x})'U(I)U(I)'\boldsymbol{\psi}(\mathbf{x}),$$

where $U(I)$ is the matrix containing the k columns of U in the set I. Hence we can write

$$\|P_{\hat{V}}(\boldsymbol{\psi}(\mathbf{x}))\|^2 = \sum_{ij=1}^{N_F} w_{ij}\boldsymbol{\psi}(\mathbf{x})_i\boldsymbol{\psi}(\mathbf{x})_j = \sum_{ij=1}^{N_F} w_{ij}\hat{\boldsymbol{\psi}}(\mathbf{x})_{ij},$$

where $\hat{\boldsymbol{\psi}}$ is the projection mapping into the feature space $\hat{\mathcal{F}}$ consisting of all pairs of F features and $w_{ij} = (U(I)U(I)')_{ij}$. The standard polynomial construction gives

$$\begin{aligned} \hat{\kappa}(\mathbf{x},\mathbf{z}) &= \kappa(\mathbf{x},\mathbf{z})^2 = \left(\sum_{i=1}^{N_F}\boldsymbol{\psi}(\mathbf{x})_i\boldsymbol{\psi}(\mathbf{z})_i\right)^2 \\ &= \sum_{i,j=1}^{N_F}\boldsymbol{\psi}(\mathbf{x})_i\boldsymbol{\psi}(\mathbf{z})_i\boldsymbol{\psi}(\mathbf{x})_j\boldsymbol{\psi}(\mathbf{z})_j \\ &= \sum_{i,j=1}^{N_F}(\boldsymbol{\psi}(\mathbf{x})_i\boldsymbol{\psi}(\mathbf{x})_j)(\boldsymbol{\psi}(\mathbf{z})_i\boldsymbol{\psi}(\mathbf{z})_j) \\ &= \left\langle\hat{\boldsymbol{\psi}}(\mathbf{x}),\hat{\boldsymbol{\psi}}(\mathbf{z})\right\rangle_{\hat{\mathcal{F}}}. \end{aligned}$$

It remains to show that the norm of the linear function is k. The norm satisfies (note that $\|\cdot\|_F$ denotes the Frobenius norm and $\mathbf{u}_i$ the columns of U)

$$\begin{aligned} \|\hat{f}\|^2 &= \sum_{i,j=1}^{N_F}\alpha_{ij}^2 = \|U(I)U(I)'\|_F^2 \\ &= \left\langle\sum_{i\in I}\mathbf{u}_i\mathbf{u}_i', \sum_{j\in I}\mathbf{u}_j\mathbf{u}_j'\right\rangle_F = \sum_{i,j\in I}(\mathbf{u}_i'\mathbf{u}_j)^2 = k \end{aligned}$$

as required. □

We are now in a position to apply a learning theory bound where we consider a regression problem for which the target output is the square of the norm of the sample point $\|\boldsymbol{\psi}(\mathbf{x})\|^2$. We restrict the linear function in the space $\hat{\mathcal{F}}$ to have norm $\sqrt{k}$. The loss function is then the shortfall between the output of $\hat{f}$ and the squared norm.

The approach we adopt here makes use of the Rademacher variables and the measure is therefore known as the Rademacher complexity. We refer the reader to Ledoux and Talagrand [27] as a core reference, though we will only be using the results and approach described in [28].

Definition 12: Given a sample $S = \{\mathbf{x}_1,\dots,\mathbf{x}_m\}$ generated by a distribution $\mathcal{D}$ on a set X and a real-valued function class $\mathcal{F}$ with domain X, the empirical Rademacher complexity of $\mathcal{F}$ is the random variable

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{m}\sum_{i=1}^m\sigma_i f(\mathbf{x}_i)\right|\,\middle|\,\mathbf{x}_1,\dots,\mathbf{x}_m\right],$$

where $\sigma = \{\sigma_1,\dots,\sigma_m\}$ are independent uniform $\{-1,+1\}$-valued (Rademacher) random variables. The Rademacher complexity of $\mathcal{F}$ is

$$R_m(\mathcal{F}) = \mathbb{E}_S\left[\hat{R}_m(\mathcal{F})\right] = \mathbb{E}_{S\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{m}\sum_{i=1}^m\sigma_i f(\mathbf{x}_i)\right|\right].$$

Note that we denote the input space with Z in the theorem, so that in the case of supervised learning we would have $Z = Y \times X$. The following theorem follows closely the proof of Theorem 8 in Bartlett and Mendelson [28], the small changes allow us to obtain slightly tighter bounds for our special case. We omit the details just noting that bounding in terms of the empirical Rademacher complexity follows from one further application of Theorem B.

Theorem D: (Bartlett and Mendelson, 2002) Let $\mathcal{F}$ be a class of functions mapping from Z to $[0,1]$ and let $S = (\mathbf{z}_i)_{i=1}^m$ be drawn independently according to a probability distribution $\mathcal{D}$ and fix $\delta \in (0,1)$. Then with probability at least $1-\delta$ over samples of length m every $f \in \mathcal{F}$ satisfies

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}[f(\mathbf{z})] &\le \hat{\mathbb{E}}[f(\mathbf{z})] + R_m(\mathcal{F}) + \sqrt{\frac{2\ln(2/\delta)}{m}} \\ &\le \hat{\mathbb{E}}[f(\mathbf{z})] + \hat{R}_m(\mathcal{F}) + \sqrt{\frac{18\ln(2/\delta)}{m}}. \end{aligned} \quad (17)$$

Given a training set S the class of functions that we will primarily be considering are linear functions with bounded norm

$$\left\{\mathbf{x}\to\sum_{i=1}^m\alpha_i\kappa(\mathbf{x}_i,\mathbf{x}) : \alpha'\mathbf{K}\alpha \le B^2\right\} \subseteq \{\mathbf{x}\to\langle\mathbf{w},\phi(\mathbf{x})\rangle : \|\mathbf{w}\|\le B\} = \mathcal{F}_B,$$

where ϕ is the feature mapping corresponding to the kernel $\kappa(\cdot,\cdot)$. Note that although the choice of functions appears to depend on S, the definition of $\mathcal{F}_B$ does not depend on the particular training set. Bartlett and Mendelson [28] bound the empirical Rademacher complexity of this function class.

Theorem E: (Bartlett and Mendelson, 2002) If $\kappa : X \times X \to \mathbb{R}$ is a kernel, and $S = \{\mathbf{x}_1,\dots,\mathbf{x}_m\}$ is a sample of points from X, then the empirical Rademacher complexity of the class $\mathcal{F}_B$ satisfies

$$\hat{R}_m(\mathcal{F}_B) \le \frac{2B}{m}\sqrt{\sum_{i=1}^m\kappa(\mathbf{x}_i,\mathbf{x}_i)} = \frac{2B}{m}\sqrt{\operatorname{tr}(\mathbf{K})}$$

The final ingredient that will be required to apply the technique are the properties of the Rademacher complexity that allow it to be bounded in terms of large classes. The following standard theorem summarises the properties of the empirical Rademacher complexity that we require.

Theorem F: Let $\mathcal{F}$ and $\mathcal{H}$ be classes of real functions. Then

1) If $\mathcal{F} \subseteq \mathcal{H}$, then $\hat{R}_m(\mathcal{F}) \le \hat{R}_m(\mathcal{H})$.
2) For every $c \in \mathbb{R}$, $\hat{R}_m(c\mathcal{F}) = |c|\hat{R}_m(\mathcal{F})$

The proofs of these results are immediate consequences of the definition of empirical Rademacher complexity. We can now apply these results to the approximation of the norm of the variable by a linear function of bounded norm.

Theorem 13: If we perform PCA on a randomly drawn training set S of size m in the feature space defined by a kernel $\kappa(\mathbf{x},\mathbf{z})$ and project new data onto the space $\hat{V}$ spanned by a subset I of k eigenvectors, with probability greater than $1-\delta$ over the generation of the sample S the expected squared residual is bounded by

$$\begin{aligned}\mathbb{E}\left[\|P_{\hat{V}}^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &\leq \frac{1}{m}\sum_{i\notin I}\hat{\lambda}_i(S)\\ &+\frac{1+\sqrt{k}}{\sqrt{m}}\sqrt{\frac{2}{m}\sum_{i=1}^{m}\kappa(\mathbf{x}_i,\mathbf{x}_i)^2}+R^2\sqrt{\frac{18}{m}\ln\left(\frac{2}{\delta}\right)},\end{aligned}$$

where the support of the distribution is in a ball of radius R in the feature space.

Proof. As indicated in Proposition 11 we consider the function class $\hat{\mathcal{F}}_{\sqrt{k}}$ with respect to the kernel

$$\hat{\kappa}(\mathbf{x},\mathbf{z})=\kappa(\mathbf{x},\mathbf{z})^2,$$

with corresponding feature mapping $\hat{\boldsymbol{\psi}}$. Note that the weight vectors considered satisfy the special condition that they are positive semi-definite, that is that

$$\sum_{ij}w_{ij}\hat{\boldsymbol{\psi}}(\mathbf{x})_{ij}\geq 0,$$

for all $\mathbf{x}$. Furthermore the function corresponds to the norm squared of a projection mapping. We will denote the subset of functions satisfying this condition by $\mathcal{P}$. We augment the corresponding primal weight vectors with one further dimension while augmenting the corresponding input vectors with a feature

$$\begin{aligned}\|\boldsymbol{\psi}(\mathbf{x}))\|^2k^{-0.25} &= \kappa(\mathbf{x},\mathbf{x})k^{-0.25}=k^{-0.25}\sqrt{\hat{\kappa}(\mathbf{x},\mathbf{x})}\\ &= \|\hat{\boldsymbol{\psi}}(\mathbf{x}))\|k^{-0.25}\end{aligned}$$

that is the norm squared in the original feature space divided by the fourth root of k. We now apply Theorem D to the class

$$\begin{aligned}\hat{\mathcal{F}} &= \Big\{f_\ell:(\hat{\boldsymbol{\psi}}(\mathbf{x}),\|\hat{\boldsymbol{\psi}}(\mathbf{x}))\|k^{-0.25})\\ &\qquad\mapsto(\|\hat{\boldsymbol{\psi}}(\mathbf{x}))\|-f(\hat{\boldsymbol{\psi}}(\mathbf{x})))R^{-2}\mid f\in\hat{\mathcal{F}}_{\sqrt{k}}\cap\mathcal{P}\Big\}\\ &\subseteq R^{-2}\hat{\mathcal{F}}'_{\sqrt{k+\sqrt{k}}},\end{aligned}$$

where we have restricted the inputs to images of points in the input space as indicated. The squared norm of the image of the input $\mathbf{x}$ under this feature mapping is $\hat{k}(\mathbf{x},\mathbf{x})(1+k^{-0.5})$. The theorem is applied to the function $\hat{f}_\ell$ where $\hat{f}$ is the projection function of Theorem 11. We must first verify that the range of the function class on the restricted inputs is $[0,1]$. Since we have restricted ourselves to positive semi-definite weight vectors $f(\hat{\boldsymbol{\psi}}(\mathbf{x}))\geq 0$, so that

$$f_\ell(\hat{\boldsymbol{\psi}}(\mathbf{x}))\leq\|\hat{\boldsymbol{\psi}}(\mathbf{x}))\|R^{-2}\leq 1.$$

Furthermore, since we have restricted $\hat{\mathcal{F}}$ to only contain functions that correspond to taking the norm squared of projection mappings in the original feature space we have that

$$f(\hat{\boldsymbol{\psi}}(\mathbf{x}))\leq\|\hat{\boldsymbol{\psi}}(\mathbf{x}))\|,$$

so that $f_\ell(\hat{\boldsymbol{\psi}}(\mathbf{x}))\geq 0$ as required. We can therefore apply Theorem 11. First note that for the function $\hat{f}_\ell$ the left hand side of the expression is equal to

$$\frac{1}{R^2}\mathbb{E}\left[\|P_{\hat{V}}^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right],$$

where $\hat{V}$ is the space spanned by the k eigenvectors in the set I. Hence, to obtain the result it remains to evaluate the two expressions on the right hand side of equation (17). The first is a scaling of the empirical squared residual when projecting into the space $\hat{V}$, that is

$$\frac{1}{R^2}\hat{\mathbb{E}}\left[\|P_{\hat{V}}^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]=\frac{1}{mR^2}\sum_{i\notin I}\hat{\lambda}_i.$$

The second expression is $\hat{R}_m(\hat{\mathcal{F}})$ which by Theorem F parts 1 and 2 can be bounded by $R^{-2}\hat{R}_m\left(\hat{\mathcal{F}}'_{\sqrt{k+\sqrt{k}}}\right)$. Next we apply Theorem E to obtain

$$\begin{aligned}\hat{R}_m\left(\hat{\mathcal{F}}'_{\sqrt{k+\sqrt{k}}}\right) &\leq \frac{\sqrt{k+\sqrt{k}}}{m}\sqrt{\operatorname{tr}(\mathbf{K})}\\ &=\sqrt{\frac{k+\sqrt{k}}{m}}\sqrt{\frac{2(1+k^{-0.5})}{m}\sum_{i=1}^{m}\kappa(\mathbf{x}_i,\mathbf{x}_i)^2}\\ &=\frac{1+\sqrt{k}}{\sqrt{m}}\sqrt{\frac{2}{m}\sum_{i=1}^{m}\kappa(\mathbf{x}_i,\mathbf{x}_i)^2}.\end{aligned}$$

Assembling all the components and multiplying through by R^2 gives the result. ■

We can apply the bound m times to obtain a proof of Theorem 1.

Proof of Theorem 1. We apply Theorem 13 taking $I=\{1,\ldots,k\}$, for $k=1,\ldots,m$, in each case replacing δ by δ/m. This ensures that with probability $1-\delta$ the assertion holds for all m applications. The second inequality of Theorem 1 follows from the observation that for $k\geq\ell$

$$\mathbb{E}\left[\|P_{\hat{V}_k}^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]\leq\mathbb{E}\left[\|P_{\hat{V}_\ell}^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right],$$

while the first inequality follows from the last inequality of equation (16). ■

A similar argument applies for Theorem 2.

Proof of Theorem 2. We apply Theorem 13 taking $I=\{1,\ldots,k\}$, for $k=1,\ldots,m$, in each case replacing δ by $\delta/(m+1)$. This ensures that with probability $1-\delta$ the assertion holds for all m applications together with the assertion that

$$\left|\mathbb{E}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right]-\hat{\mathbb{E}}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right]\right|\leq R^2\sqrt{\frac{1}{m}\ln\frac{2(m+1)}{\delta}}.$$

This final inequality follows from a straightforward application of McDiarmid's inequality. The second inequality of Theorem 2 follows from the observations above together with the fact that

$$\frac{1}{m}\hat{\lambda}^{\leq k}=\hat{\mathbb{E}}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right]-\frac{1}{m}\hat{\lambda}^{>k},$$

while the first inequality again follows from the last inequality of equation (16). ■

Finally we give the proof of Theorem 3.

Proof of Theorem 3. Consider applying Theorem 13 taking $I = \{k\}$, and replacing δ by $\delta/(m+1)$. This ensures that with probability $1-\delta$ the assertion holds for all m applications together with the assertion that

$$\left|\mathbb{E}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \hat{\mathbb{E}}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right]\right| \leq R^2\sqrt{\frac{1}{m}\ln\frac{2(m+1)}{\delta}}.$$

This final inequality follows from a straightforward application of McDiarmid's inequality. The inequality of Theorem 3 follows from the observations above together with the fact that

$$\frac{1}{m}\sum_{i\neq k}\hat{\lambda}_i = \hat{\mathbb{E}}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \frac{1}{m}\hat{\lambda}_k.$$

■

VI. Experiments

To illustrate the results described in this paper experiments were carried out with the Breast cancer data set [29] which contains 683 data points. This dataset is available from the UCI data repository. A normalised cubic polynomial kernel was chosen,

$$\kappa_{NC}\langle\mathbf{x}_i,\mathbf{x}_j\rangle = \frac{\langle\mathbf{x}_i,\mathbf{x}_j\rangle^3}{\sqrt{\langle\mathbf{x}_i,\mathbf{x}_i\rangle^3\langle\mathbf{x}_j,\mathbf{x}_j\rangle^3}} \tag{18}$$

from a range of other kernels, based on the empirical observation that the process eigenspectrum did not decay too fast.

We compare three quantities

(i) $\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \frac{1}{m}\sum_{i=1}^k\hat{\lambda}_i$,

(ii) $\mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \sum_{i=1}^k\lambda_i$,

(iii) $\mathbb{E}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]$.

From inequality (13) we have (ii) $\geq$ (iii) and from Proposition 2 we have (i) $\geq$ (ii) in the expectation $\mathbb{E}_m$ with respect to the product distribution.

We randomly selected 50% of the data as a 'training' set. The process eigenspectrum was obtained by performing an eigenvalue decomposition of the kernel matrix constructed from the entire dataset. Similarly the spectrum $\{\hat{\lambda}_i\}$ was obtained from an eigendecomposition of the appropriate submatrix. The computation of $\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2$ is carried out as explained in [15].

Figure 2(a) shows the projected squared norm plotted against k for these three quantities. Curves (i) and (iii) have been averaged over 20 random choices of the training set. The error bars give one standard deviation. Notice the close agreement between the curves (i) and (iii), indicating that the subspace identified as optimal for the training set is indeed capturing almost the same amount of information for all data points. The very tight error bars show clearly the very tight concentration of the sums of tail of eigenvalues as predicted by Theorem 7. In order to amplify the information depicted in Figure 2(a), Figure 2(b) plots the differences (i)-(ii) and (iii)-(ii). As expected we see that (i)-(ii) ≥ 0 and (iii)-(ii) ≤ 0. For larger projection dimensions the theory predicts that the accuracy will level off and remain constant and this effect can be observed in Figure 2(b).

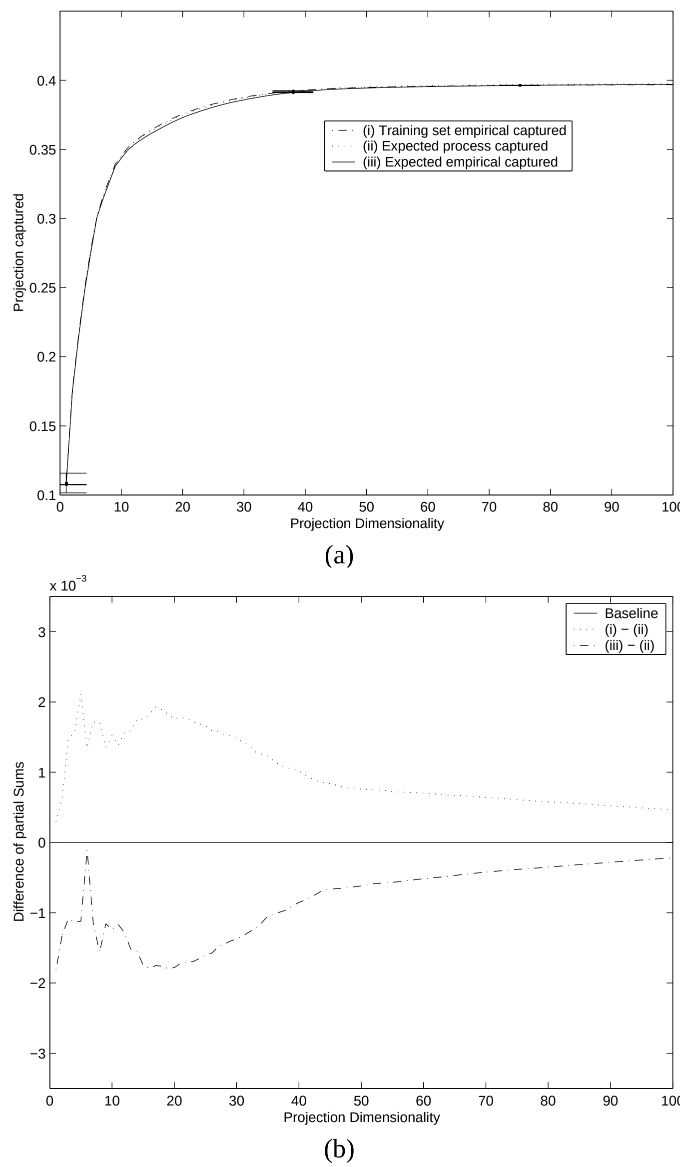


Fig. 2. (a) Plot of the projected squared norm plotted against the projection dimension. The plot shows three curves, (i) expected squared norm for training set when projected into empirical eigenspace averaged over 20 random splits, (ii) expected squared norm for the true process eigenspectrum and (iii) expected squared norm for empirical eigenspace again averaged over 20 random splits. (b) Zooms in on plot (a) by displaying the differences between (i) and (ii) and between (iii) and (ii).

VII. Conclusions

The paper has shown that the eigenvalues of a positive semi-definite matrix generated from a random sample is concentrated. Furthermore the sum of the last $m-k$ eigenvalues is similarly concentrated as is the residual when the data is projected into a fixed subspace.

Furthermore, we have shown that estimating the projection subspace on a random sample can give a good model for future data provided the number of examples is much larger than the dimension of the subspace that captures most of the training data. The results provide a basis for performing PCA or kernel-PCA from a randomly generated sample, as they confirm that the subspace identified by the sample will indeed ‘generalise’ in the sense that it will capture most of the information in a test sample provided that the dimension is small compared to the sample size and that the subspace captures most of the variance in the training data. The result is somewhat counter-intuitive in that the dimension of the feature space does not appear explicitly. The critical quantity is the ratio of the empirical or ‘effective’ dimension of the sample data to the number of examples it comprises.

Experiments are presented that confirm the theoretical predictions on a real world data-set for small projection dimensions. For larger projection dimensions the theory predicts that the accuracy will level off and remain constant. In practice there is a slow attenuation with increasing projection dimension. This is not inconsistent with the theory and accords with intuitive expectations.

ACKNOWLEDGEMENTS

CW thanks Matthias Seeger for comments on an earlier version of the paper. We would like to acknowledge the financial support of EPSRC Grant No. GR/N08575, EU Project KerMIT, No. IST-2000-25341, the Neurocolt working group No. 27150, and the PASCAL Network of Excellence, No. IST-2002-506778.

REFERENCES

[1] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[2] 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*, 1998.

[3] N. Cristianini, H. Lodhi, and J. Shawe-Taylor, “Latent semantic kernels for feature selection,” NeuroCOLT Working Group, http://www.neurocolt.org, Tech. Rep. NC-TR-00-080, 2000.

[4] J. Kleinberg, “Authoritative sources in a hyperlinked environment,” in *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[5] S. Brin and L. Page, “The anatomy of a large-scale hypertextual (web) search engine,” in *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[6] A. Y. Ng, A. X. Zheng, and M. I. Jordan, “Link analysis, eigenvectors and stability,” in *To appear in the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, 2001.

[7] A. Guionnet and O. Zeitouni, “Concentration of the spectral measure for large matrices,” *Electron. Comm. Prob.*, vol. 5, pp. 119–136, 2000.

[8] N. Alon, M. Krivelevich, and V. H. Vu, “On the concentration of eigenvalues of random symmetric matrices,” *Israel Journal of Mathematics*, vol. 131, pp. 259–267, 2002.

[9] C. K. I. Williams and M. Seeger, “The Effect of the Input Density Distribution on Kernel-based Classifiers,” in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, P. Langley, Ed. Morgan Kaufmann, 2000.

[10] C. T. H. Baker, *The numerical treatment of integral equations*. Oxford: Clarendon Press, 1977.

[11] H. Zhu, C. K. I. Williams, R. J. Rohwer, and M. Morciniec, “Gaussian regression and optimal finite dimensional linear models,” in *Neural Networks and Machine Learning*, C. M. Bishop, Ed. Berlin: Springer-Verlag, 1998.

[12] V. Koltchinskii and E. Gine, “Random matrix approximation of spectra of integral operators,” *Bernoulli*, vol. 6(1), pp. 113–167, 2000.

[13] I. Johnstone, “On the distribution of the largest principal component,” Stanford University, http://www-stat.stanford.edu/ imj/Reports/index.html, Tech. Rep., 2000.

[14] J. Shawe-Taylor, N. Cristianini, and J. Kandola, “On the Concentration of Spectral Properties,” in *Advances in Neural Information Processing Systems 14*, T. G. Diettrich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002.

[15] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press, 2004.

[16] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *J. Amer. Stat. Assoc.*, vol. 58, pp. 13–30, 1963.

[17] K. Azuma, “Weighted sums of certain dependent random variables,” *Tohoku Math J.*, vol. 19, pp. 357–367, 1967.

[18] C. McDiarmid, “On the method of bounded differences,” in *Surveys in Combinatorics 1989*. Cambridge University Press, 1989, pp. 148–188.

[19] S. Boucheron, G. Lugosi, and P. Massart, “A sharp concentration inequality with applications,” *Random Structures and Algorithms*, vol. 16, pp. 277–292, 2000.

[20] M. Talagrand, “New concentration inequalities in product spaces,” *Invent. Math.*, vol. 126, pp. 505–563, 1996.

[21] H. Voss, “Variational characterization of eigenvalues of nonlinear eigenproblems,” in *Proceedings of the International Conference on Mathematical and Computer Modelling in Science and Engineering*, 2003, pp. 379–383.

[22] M. L. Eaton and D. E. Tyler, “On Wielandt’s Inequality and Its Application to the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix,” *Annals of Statistics*, vol. 19(1), pp. 260–271, 1991.

[23] A. T. James, “The distribution of the latent roots of the covariance matrix,” *Annals of Math. Stat.*, vol. 31, pp. 151–158, 1960.

[24] D. N. Lawley, “Tests of Significance for the Latent Roots of Covariance and Correlation Matrices,” *Biometrika*, vol. 43(1/2), pp. 128–136, 1956.

[25] T. W. Anderson, “Asymptotic Theory for Principal Component Analysis,” *Annals of Mathematical Statistics*, vol. 34(1), pp. 122–148, 1963.

[26] C. M. Waternaux, “Asymptotic Distribution of the Sample Roots for a Nonnormal Population,” *Biometrika*, vol. 63(3), pp. 639–645, 1976.

[27] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

[28] P. L. Bartlett and S. Mendelson, “Rademacher and gaussian complexities: Risk bounds and structural results,” *Journal of Machine Learning Research*, 2002.

[29] C. Blake and C. Merz, “UCI repository of machine learning databases,” 1998. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html