On the Eigenspectrum of the Gram Matrix and its Relationship to the Process Eigenspectrum

John Shawe-Taylor[1], Chris Williams[2], Nello Cristianini[3], and Jaz Kandola[1]

[1] Department of Computer Science,
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK
[2] Division of Informatics,
University of Edinburgh
[3] Department of Statistics,
University of California at Davies

Abstract. In this paper we analyze the relationships between the eigenvalues of the $m \times m$ Gram matrix K for a kernel $k(\cdot, \cdot)$ 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.

1 Introduction

Over recent years there has been a considerable amount of interest in kernel methods such as Support Vector Machines, 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 $k(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, m$, where $\{\mathbf{x}_i : i = 1, \ldots, m\}$ is a given dataset and $k(\cdot, \cdot)$ 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.

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 ([12]). This representation has also been related to an Information Retrieval algorithm known as latent semantic indexing, again with kernel defined feature spaces ([5]).

Furthermore eigenvectors have been used in the HITS ([8]) and Google's PageRank ([4]) 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. [13] 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.

In this paper we shift the emphasis towards studying the reliability of the estimates gained from a finite sample. 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.

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

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

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

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

As pointed out in [16], the standard numerical method (see, e.g., [3], 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

$$\frac{1}{m} \sum_{k=1}^{m} k(\mathbf{x}_k, \mathbf{x}_j) \phi_i(\mathbf{x}_k) = \lambda_i \phi_i(\mathbf{x}_j).$$

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

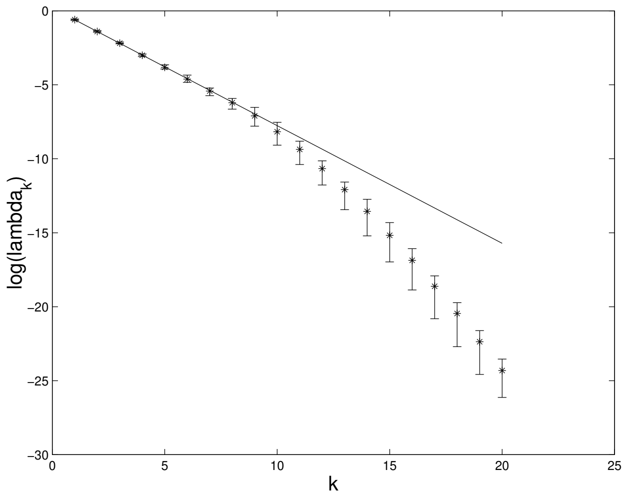


Fig. 1. A plot of the log eigenvalue against the index of the eigenvalue. The straight line is the theoretical relationship. The centre point (marked with a star) in the "error bar" is the log of the average value of μ_k. The upper and lower ends of the error bars are the maximum and minimum values of $\log(\mu_k)$ respectively taken over ten repetitions.

([3], 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 $k(x, y) = \exp -b(x-y)^2$ (the RBF kernel with lengthscale $b^{-1/2}$), there are analytic results for the eigenvalues and eigenfunctions of equation (1) as given in section 4 of [17]. For this example we can therefore compare the values of μ_i with the corresponding λ_i, as shown in Figure 1. Here $m = 500$ points were used, with parameters $b = 3$ and $p(x) \sim N(0, 1/4)$. As the result depends upon the random points chosen for the sample, this process was repeated ten times. We observe good agreement of the process and matrix eigenvalues for small k, but that for larger k the matrix eigenvalues underestimate the process eigenvalues. One of the by-products of this paper will be bounds on the degree of underestimation for this estimation problem in a fully general setting.

[9] 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.

In an earlier paper, [14] discussed the concentration of spectral properties of Gram matrices and of the residuals of fixed projections. However, we note that 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.

The paper is organised as follows. In section 2 we give the background results and develop the basic techniques that are required to derive the main results in section 3. We provide experimental verification of the theoretical findings in section 4, before drawing our conclusions.

2 Background and Techniques

We will make use of the following results due to McDiarmid.

Theorem 1. *([11]) 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 $f_i : A^{n-1} \to \mathbb{R}$ satisfy for $1 \leq i \leq n$*

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

then for all $\epsilon > 0$,

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

Theorem 2. *([11]) 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 $\epsilon > 0$,

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

We will also make use of the following theorem characterising the eigenvectors of a symmetric matrix.

Theorem 3 (Courant-Fischer Minimax Theorem). *If $M \in \mathbb{R}^{m\times m}$ is symmetric, then for $k = 1, \ldots, m$,*

$$\lambda_k(M) = \max_{\dim(T)=k} \min_{0\neq\mathbf{v}\in T} \frac{\mathbf{v}'M\mathbf{v}}{\mathbf{v}'\mathbf{v}} = \min_{\dim(T)=m-k+1} \max_{0\neq\mathbf{v}\in T} \frac{\mathbf{v}'M\mathbf{v}}{\mathbf{v}'\mathbf{v}},$$

with the extrema achieved by the corresponding eigenvector.

The approach adopted in the proofs of the next section is to view the eigenvalues as the sums of squares of residuals. This is applicable when the matrix is positive semi-definite and hence can be written as an inner product matrix $M = X'X$, where X' is the transpose of the matrix X containing the m vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$

as columns. This is the finite dimensional version of Mercer's theorem, and follows immediately if we take $X = V\sqrt{\Lambda}$, where $M = V\Lambda V'$ is the eigenvalue decomposition of M. There may be more succinct ways of representing X, but we will assume for simplicity (but without loss of generality) that X is a square matrix with the same dimensions as M. To set the scene, we now present a short description of the residuals viewpoint.

The starting point is the singular value decomposition of $X = U\Sigma V'$, where U and V are orthonormal matrices and Σ is a diagonal matrix containing the singular values (in descending order). We can now reconstruct the eigenvalue decomposition of $M = X'X = V\Sigma U'U\Sigma V' = V\Lambda V'$, where $\Lambda = \Sigma^2$. But equally we can construct a matrix $N = XX' = U\Sigma V'V\Sigma U' = U\Lambda U'$, with the same eigenvalues as M.

As a simple example consider now the first eigenvalue, which by Theorem 3 and the above observations is given by

$$\begin{aligned}\lambda_1(M) &= \max_{0\neq\mathbf{v}\in\mathbb{R}^m} \frac{\mathbf{v}'N\mathbf{v}}{\mathbf{v}'\mathbf{v}} = \max_{0\neq\mathbf{v}\in\mathbb{R}^m} \frac{\mathbf{v}'XX'\mathbf{v}}{\mathbf{v}'\mathbf{v}} = \max_{0\neq\mathbf{v}\in\mathbb{R}^m} \frac{\|\mathbf{v}'X\|^2}{\mathbf{v}'\mathbf{v}} \\ &= \max_{0\neq\mathbf{v}\in\mathbb{R}^m} \sum_{j=1}^m \|P_{\mathbf{v}}(\mathbf{x}_j)\|^2 = \sum_{j=1}^m \|\mathbf{x}_j\|^2 - \min_{0\neq\mathbf{v}\in\mathbb{R}^m} \sum_{j=1}^m \|P_{\mathbf{v}}^{\perp}(\mathbf{x}_j)\|^2\end{aligned}$$

where $P_{\mathbf{v}}(\mathbf{x})$ ($P_{\mathbf{v}}^{\perp}(\mathbf{x})$) is the projection of $\mathbf{x}$ onto the space spanned by $\mathbf{v}$ (space perpendicular to $\mathbf{v}$), since $\|\mathbf{x}\|^2 = \|P_{\mathbf{v}}(\mathbf{x})\|^2 + \|P_{\mathbf{v}}^{\perp}(\mathbf{x})\|^2$. It follows that the first eigenvector is characterised as the direction for which sum of the squares of the residuals is minimal.

Applying the same line of reasoning to the first equality of Theorem 3, delivers the following equality

$$\lambda_k(M) = \max_{\dim(V)=k} \min_{0\neq\mathbf{v}\in V} \sum_{j=1}^m \|P_{\mathbf{v}}(\mathbf{x}_j)\|^2. \tag{3}$$

Notice that this characterisation implies that if $\mathbf{v}^k$ is the k-th eigenvector of N, then

$$\lambda_k(M) = \sum_{j=1}^m \|P_{\mathbf{v}^k}(\mathbf{x}_j)\|^2, \tag{4}$$

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

$$\sum_{i=1}^k \lambda_i(M) = \sum_{j=1}^m \|P_{V_k}(\mathbf{x}_j)\|^2 = \sum_{j=1}^m \|\mathbf{x}_j\|^2 - \sum_{j=1}^m \|P_{V_k}^{\perp}(\mathbf{x}_j)\|^2, \tag{5}$$

where $P_V(\mathbf{x})$ ($P_V^{\perp}(\mathbf{x})$) is the projection of $\mathbf{x}$ into the space V (space perpendicular to V). 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

$$\sum_{i=1}^{k} \lambda_i(M) = \max_{\dim(V)=k} \sum_{j=1}^{m} \|P_V(\mathbf{x}_j)\|^2 = \sum_{j=1}^{m} \|\mathbf{x}_j\|^2 - \min_{\dim(V)=k} \sum_{j=1}^{m} \|P_V^{\perp}(\mathbf{x}_j)\|^2,$$

$$\sum_{i=k+1}^{m} \lambda_i(M) = \sum_{j=1}^{m} \|\mathbf{x}_j\|^2 - \sum_{i=1}^{k} \lambda_i(M) = \min_{\dim(V)=k} \sum_{j=1}^{m} \|P_V^{\perp}(\mathbf{x}_j)\|^2. \tag{6}$$

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.

Frequently, we consider all of the above as occurring in a kernel defined feature space, so that wherever we have written a vector $\mathbf{x}$ we should have put $\boldsymbol{\psi}(\mathbf{x})$, where $\boldsymbol{\psi}$ is the corresponding feature map

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

to a feature space F. Hence, the matrix M has entries $M_{ij} = \langle \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}_j)\rangle$. The kernel function computes the composition of the inner product with the feature maps,

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

which can in many cases be computed without explicitly evaluating the mapping $\boldsymbol{\psi}$. We would also like to evaluate the projections into eigenspaces without explicitly computing the feature mapping $\boldsymbol{\psi}$.

This can be done as follows. Let $\mathbf{u}_i$ be the i-th singular vector in the feature space, that is the i-th eigenvector of the matrix N, with the corresponding singular value being $\sigma_i = \sqrt{\lambda_i}$ and the corresponding eigenvector of M being $\mathbf{v}_i$. The projection of an input $\mathbf{x}$ onto $\mathbf{u}_i$ is given by

$$\begin{aligned} \boldsymbol{\psi}(\mathbf{x})'\mathbf{u}_i &= (\boldsymbol{\psi}(\mathbf{x})'U)_i \\ &= (\boldsymbol{\psi}(\mathbf{x})'XV)_i\sigma_i^{-1} \\ &= \mathbf{k}'\mathbf{v}_i\sigma_i^{-1}, \end{aligned}$$

where we have used the fact that $X = U\Sigma V'$ and $\mathbf{k}_j = \boldsymbol{\psi}(\mathbf{x})'\boldsymbol{\psi}(\mathbf{x}_j) = k(\mathbf{x},\mathbf{x}_j)$.

Our final background observation concerns the kernel operator and its eigenspaces. The operator in question is

$$K(f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x},\mathbf{z})f(\mathbf{z})p(\mathbf{z})d\mathbf{z},$$

where $p(\mathbf{z})$ is the underlying probability density function that governs the occurrence of different examples. Note that we have moved away from a finite set of examples to a potentially uncountably infinite space $\mathcal{X}$.

Provided the operator is positive semi-definite, by Mercer's theorem we can decompose $k(\mathbf{x},\mathbf{z})$ as a sum of eigenfunctions,

$$k(\mathbf{x},\mathbf{z}) = \sum_{i=1}^{\infty} \lambda_i\phi_i(\mathbf{x})\phi_i(\mathbf{z}) = \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{z})\rangle,$$

where $\lambda_i = \lambda_i(K(f))$ the functions $(\phi_i(\mathbf{x}))_{i=1}^{\infty}$ form a complete orthonormal basis with respect to the inner product $\langle f, g\rangle_p = \int_{\mathcal{X}} f(\mathbf{x})g(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ and $\boldsymbol{\psi}(\mathbf{x})$ is the feature space mapping

$$\boldsymbol{\psi} : \mathbf{x} \longrightarrow (\psi_i(\mathbf{x}))_{i=1}^{\infty} = \left(\sqrt{\lambda_i}\phi_i(\mathbf{x})\right)_{i=1}^{\infty} \in F.$$

Note that $\phi_i(\mathbf{x})$ has norm 1 and satisfies

$$\phi_i(\mathbf{x}) = \lambda_i \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{z})\phi_i(\mathbf{z})p(\mathbf{z})d\mathbf{z}, \quad \text{so that} \int_{\mathcal{X}^2} k(\mathbf{x}, \mathbf{z})\phi_i(\mathbf{x})\phi_i(\mathbf{z})p(\mathbf{z})p(\mathbf{z})d\mathbf{y}d\mathbf{z} = \lambda_i.$$

If we let $\boldsymbol{\phi}(\mathbf{x}) = (\phi_i(\mathbf{x}))_{i=1}^{\infty} \in F$, we can define the unit vector $\mathbf{u}_i \in F$ corresponding to λ_i by

$$\mathbf{u}_i = \int_{\mathcal{X}} \phi_i(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \mathbf{e}_i.$$

For a general function $f(\mathbf{x})$ we can similarly define the vector

$$\mathbf{f} = \int_{\mathcal{X}} f(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})p(\mathbf{x})d\mathbf{x}.$$

Now the expected square of the norm of the projection $P_{\mathbf{f}}(\boldsymbol{\psi}(\mathbf{x}))$ onto the vector $\mathbf{f}$ (assumed to be of norm 1) of an input $\boldsymbol{\psi}(\mathbf{x})$ drawn according to $p(\mathbf{x})$ is given by

$$\begin{aligned}
\mathbb{E}\left[\|P_{\mathbf{f}}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &= \int_{\mathcal{X}} \|P_{\mathbf{f}}(\boldsymbol{\psi}(\mathbf{x}))\|^2 p(\mathbf{x})d\mathbf{x} = \int_{\mathcal{X}} \left(\mathbf{f}'\boldsymbol{\psi}(\mathbf{x})\right)^2 p(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{X}} f(\mathbf{y})\boldsymbol{\phi}(\mathbf{y})'\boldsymbol{\psi}(\mathbf{x})p(\mathbf{y})d\mathbf{y} f(\mathbf{z})\boldsymbol{\phi}(\mathbf{z})'\boldsymbol{\psi}(\mathbf{x})p(\mathbf{z})d\mathbf{z}p(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathcal{X}^3} f(\mathbf{y})f(\mathbf{z}) \sum_{j=1}^{\infty} \sqrt{\lambda_j}\phi_j(\mathbf{y})\phi_j(\mathbf{x})p(\mathbf{y})d\mathbf{y} \sum_{\ell=1}^{\infty} \sqrt{\lambda_\ell}\phi_\ell(\mathbf{z})\phi_\ell(\mathbf{x})p(\mathbf{z})d\mathbf{z}p(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathcal{X}^2} f(\mathbf{y})f(\mathbf{z}) \sum_{j,\ell=1}^{\infty} \sqrt{\lambda_j}\phi_j(\mathbf{y})p(\mathbf{y})d\mathbf{y}\sqrt{\lambda_\ell}\phi_\ell(\mathbf{z})p(\mathbf{z})d\mathbf{z} \int_{\mathcal{X}} \phi_j(\mathbf{x})\phi_\ell(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathcal{X}^2} f(\mathbf{y})f(\mathbf{z}) \sum_{j=1}^{\infty} \lambda_j\phi_j(\mathbf{y})\phi_j(\mathbf{z})p(\mathbf{y})d\mathbf{y}p(\mathbf{z})d\mathbf{z} \\
&= \int_{\mathcal{X}^2} f(\mathbf{y})f(\mathbf{z})k(\mathbf{y}, \mathbf{z})p(\mathbf{y})p(\mathbf{z})d\mathbf{y}d\mathbf{z}.
\end{aligned}$$

Since all vectors $\mathbf{f}$ in the subspace spanned by the image of the input space in F can be expressed in this fashion, it follows that the sum of the finite case characterisation of eigenvalues and eigenvectors is replaced by an expectation

$$\lambda_k(K(f)) = \max_{\dim(V)=k} \min_{0\neq\mathbf{v}\in V} \mathbb{E}[\|P_{\mathbf{v}}(\boldsymbol{\psi}(\mathbf{x}))\|^2], \tag{7}$$

where V is a linear subspace of the feature space F. Similarly,

$$\sum_{i=1}^{k} \lambda_i(K(f)) = \max_{\dim(V)=k} \mathbb{E}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \mathbb{E}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \min_{\dim(V)=k} \mathbb{E}\left[\|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right],$$

$$\sum_{i=k+1}^{\infty} \lambda_i(K(f)) = \mathbb{E}\left[\|\boldsymbol{\psi}(\mathbf{x})\|^2\right] - \sum_{i=1}^{k} \lambda_i(K(f)) = \min_{\dim(V)=k} \mathbb{E}\left[\|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]. \tag{8}$$

where $P_V(\boldsymbol{\psi}(\mathbf{x}))$ ($P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}))$) is the projection of $\boldsymbol{\psi}(\mathbf{x})$ into the subspace V (the projection of $\boldsymbol{\psi}(\mathbf{x})$ into the space orthogonal to V).

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 $\boldsymbol{\psi}$ using the process eigenfunctions. 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 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 expectation with respect to the sample,

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

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

We are interested in the relationships between the following quantities:

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

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 (8),

$$\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i \geq \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right], \tag{9}$$

$$\text{and} \quad \mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \sum_{i=1}^{k}\lambda_i \geq \mathbb{E}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \tag{10}$$

Our strategy will be to show that the right hand side of inequality (9) and the left hand side of inequality (10) 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.

3 Averaging over Samples and Population Eigenvalues

Consider a zero-mean random variable $\mathbf{X} \in \mathbb{R}^p$. Let m samples drawn from $p(\mathbf{X})$ be stored in the $p \times m$ data matrix X. The the sample estimate for the covariance matrix is $S_X = \frac{1}{m}XX'$. Let the eigenvalues of S_X be $\mu_1 \geq \mu_2 \ldots \geq \mu_p$. By the results above these are the same as the eigenvalues of the matrix $\frac{1}{m}X'X$. Note that in the notation of the previous section $\mu_i = (1/m)\hat{\lambda}_i$. The corresponding population covariance be denoted Σ, with eigenvalues $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p$ and eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_p$. Again by the observations above these are the process eigenvalues.

Statisticians have been interested in the sampling distribution of the eigenvalues of S_X for some time. There are two main approaches to studying this problem, as discussed in section 6 of [6]. In the case that $p(\mathbf{X})$ has a multivariate normal distribution, the exact sampling distribution of $\mu_1, \ldots, \mu_p$ can be given [7]. Alternatively, the "delta method" can be used, expanding the sample roots about the population roots. For normal populations this has been carried out by [10] (if there are no repeated roots of the population covariance) and [1] (for the general case), and extended by [15] to the non-Gaussian case.

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

Proposition 1 (Anderson, 1963, pp 145-146). $\mathbb{E}_m[\mu_1] \geq \lambda_1$ *and* $\mathbb{E}_m[\mu_p] \leq \lambda_p$.

Proof: By the results of the previous section we have

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

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

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

Proposition 1 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 2. *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 (8) 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, [10] (section 4) gives the asymptotic relationship

$$\mathbb{E}_m[\mu_i] = \lambda_i + \frac{1}{m} \sum_{j=1, j\neq i}^{p} \frac{\lambda_i \lambda_j}{\lambda_i - \lambda_j} + O(m^{-2}). \tag{11}$$

Note that this is consistent with Proposition 2.

Proposition 2 also implies that

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

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 $k(\mathbf{x}, \mathbf{x}) = C$, a constant. For example if the kernel is stationary, this holds since $k(\mathbf{x}, \mathbf{x}) = k(\mathbf{x} - \mathbf{x}) = k(\mathbf{0}) = C$. Thus

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

Also we have for the continuous eigenproblem $\int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = C$. Using the feature expansion representation of the kernel $k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N_F} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y})$ 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 $\sum_{i=1}^{k} \lambda_i$. At the other end, clearly for $N_F \geq k > m$, $\mu_k \equiv 0$ is an underestimate of λ_k.

4 Concentration of eigenvalues

Section 2 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 (6). Note that these results (Theorems 4 to 6) are reproduced from [14].

Theorem 4. *Let $k(\mathbf{x}, \mathbf{z})$ 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 $K(S)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $R^2 = \max_{\mathbf{x} \in X} k(\mathbf{x}, \mathbf{x})$.

Proof: The result follows from an application of Theorem 1 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 $K(S)$ $(K(\hat{S}))$. Let k correspond to the feature mapping $\boldsymbol{\psi}$. Using equation (3) we have

$$\hat{\lambda}_k(S) \geq \min_{v \in \hat{V}} \sum_{j=1}^{m} \|P_v(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 \geq \min_{v \in \hat{V}} \sum_{j \neq i} \|P_v(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 = \hat{\lambda}_k(\hat{S})$$

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

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

Theorem 5. *Let $k(\mathbf{x}, \mathbf{z})$ 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 $\epsilon > 0$,*

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

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

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 $K(S)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $R^2 = \max_{\mathbf{x} \in X} k(\mathbf{x}, \mathbf{x})$.

Proof: The result follows from an application of Theorem 1 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 $K(S)$ $(K(\hat{S}))$. Let k correspond to the feature mapping $\boldsymbol{\psi}$. Using equation (6) we have

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

$$\lambda^{>k}(\hat{S}) \leq \sum_{j \neq i} \|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 = \sum_{j=1}^m \|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}_j))\|^2 - \|P_V^{\perp}(\boldsymbol{\psi}(\mathbf{x}_i))\|^2 \leq \lambda^{>k}(S).$$

A similar derivation proves the second inequality. □

Our next result concerns the concentration of the residuals with respect to a fixed subspace. For a subspace V and training set S, we introduce the notation

$$\bar{P}_V(S) = \hat{\mathbb{E}}\left[\|P_V(\boldsymbol{\psi}(\mathbf{x}))\|^2\right].$$

Theorem 6. *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 $\epsilon > 0$,*

$$P\{\bar{P}_V(S) - \mathbb{E}_m[\bar{P}_V(S)]| \geq \epsilon\} \leq 2\exp\left(\frac{-\epsilon^2 m}{2R^4}\right).$$

Proof: The result follows from an application of Theorem 2 provided

$$\sup_{S, \hat{x}_i} |\bar{P}_V(S) - \bar{P}(S \setminus \{\mathbf{x}_i\} \cup \{\hat{\mathbf{x}}_i)| \leq R^2/m.$$

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

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

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

$$\text{and} \qquad \mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \sum_{i=1}^{k}\lambda_i \approx \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right], \tag{12}$$

where we have used Theorem 5 to obtain the first approximate equality and Theorem 6 with $V = V_k$ to obtain the second approximate equality.

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

$$\begin{aligned}\hat{\mathbb{E}}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] &= \frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i \geq \hat{\mathbb{E}}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] \\ &\approx \mathbb{E}\left[\|P_{V_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right] = \sum_{i=1}^{k}\lambda_i \geq \mathbb{E}\left[\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2\right]. \end{aligned} \tag{13}$$

Notice that using Proposition 2 we also obtain the following diagram of approximate relationships

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

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.

5 Learning a projection matrix

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 3. *The projection norm $\|P_{\hat{V}_k}(\boldsymbol{\psi}(\mathbf{x}))\|^2$ as a linear function $\hat{f}$ in a feature space $\hat{F}$ for which the kernel function is given by*

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

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

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

where U_k is the matrix containing the first k columns of U. Hence we can write

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

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

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

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)

$$\|\hat{f}\|^2 = \sum_{i,j=1}^{N_F} \alpha_{ij}^2 = \|U_k U_k'\|_F^2 = \left\langle \sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i', \sum_{j=1}^{k} \mathbf{u}_j \mathbf{u}_j' \right\rangle_F = \sum_{i,j=1}^{k} (\mathbf{u}_i' \mathbf{u}_j)^2 = k$$

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 $\hat{F}$ to have norm $\sqrt{k}$. The loss function is then the shortfall between the output of $\hat{f}$ and the squared norm. If we scale this with a factor $1/R^2$ the output is in the range $[0,1]$ and we can apply Theorem 17.7 of [2].

Due to space limitations we will not quote a detailed result of this type here. The expected squared residual of a random test point is with probability $1-\delta$ bounded by

$$O\left(R^2 \epsilon + \frac{1}{m} \sum_{i=k+1}^{m} \hat{\lambda}_i \right),$$

where ϵ is chosen so that

$$\epsilon^2 = \frac{c}{m} \left[\frac{k}{\epsilon^2} \log \frac{1}{\epsilon} \log \frac{m}{\epsilon} + \log \frac{1}{\delta} \right],$$

where c is a constant, $R^2 = \max_i k(\mathbf{x}_i, \mathbf{x}_i)$, k is the projection dimension. The second factor is the average residue after the projection of the training set or in other words the sum of the remaining eigenvalues divided by the sample size. The size of the difference between the expected 2-norm residual and the training set estimate depends on the dimension of the projection space and the number of training examples. For a non-trivial bound the number of examples must be much larger than the projection dimension.

6 Experiments

In order to test the concentration results we performed experiments with the Breast cancer data using a cubic polynomial kernel. The kernel was chosen to ensure that the spectrum did not decay too fast.

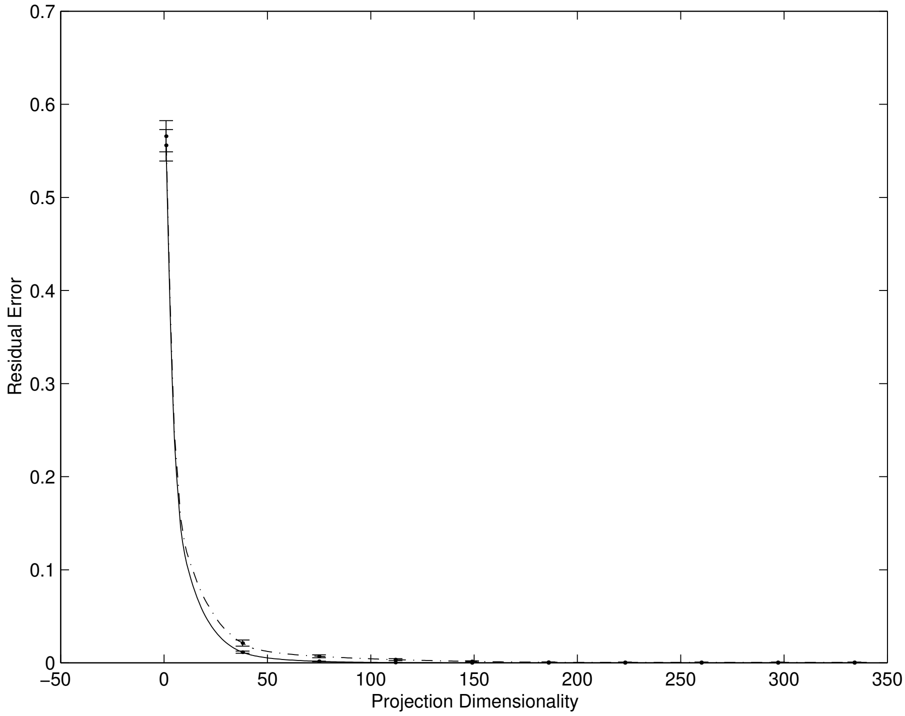
We randomly selected 50% of the data as a 'training' set and kept the remaining 50% as a 'test' set. We centered the whole data set so that the origin of the feature space is placed at the centre of gravity of the training set. We then performed an eigenvalue decomposition of the training set. The sum of the eigenvalues greater than the k-th gives the sum of the residual squared norms of the training points when we project onto the space spanned by the first k eigenvectors. Dividing this by the average of all the eigenvalues (which measures the average square norm of the training points in the transformed space) gives a fraction residual not captured in the k dimensional projection. This quantity was averaged over 5 random splits and plotted against dimension in Figure 2 as the continuous line. The error bars give one standard deviation. The Figure 2a shows the full spectrum, while Figure 2b shows a zoomed in subwindow. The very tight error bars show clearly the very tight concentration of the sums of tail of eigenvalues as predicted by Theorem 5.

In order to test the concentration results for subsets we measured the residuals of the test points when they are projected into the subspace spanned by the first k eigenvectors generated above for the training set. The dashed lines in Figure 2 show the ratio of the average squares of these residuals to the average squared norm of the test points. We see the two curves tracking each other very closely, indicating that the subspace identified as optimal for the training set is indeed capturing almost the same amount of information in the test points.

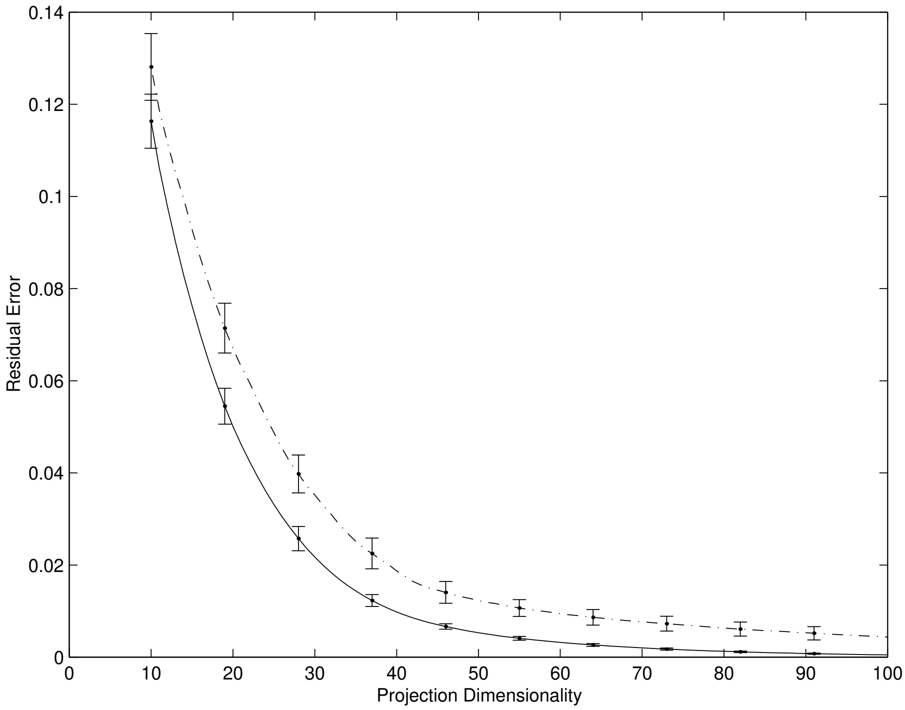
7 Conclusions

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. The results provide



(a)



(b)

Fig. 2. Plots of the fraction of the average squared norm captured in the subspace spanned by the first k eigenvectors for different values of k. Continuous line is fraction for training set, while the dashed line is for the test set. (a) shows the full spectrum, while (b) zooms in on an interesting portion.

a basis for performing PCA or kernel-PCA from a randomly generated sample, as they confirm that the subset identified by the sample will indeed 'generalise' in the sense that it will capture most of the information in a test sample.

Experiments are presented that confirm the theoretical predictions on a real world data-set for small projection dimensions. For larger projection dimensions the theory is unable to make confident predictions, though in practice the residuals became very small in the examples tested. This phenomenon suggests that the theory is unable to accurately model the case when the residuals are very small but the dimension approaches the training set size.

Further research should look at the question of how the space identified by a subsample relates to the eigenspace of the underlying kernel operator since this is not guaranteed by the bounds obtained.

Acknowledgements

I thank Matthias Seeger for comments on an earlier version of the paper.

References

1. T. W. Anderson. Asymptotic Theory for Principal Component Analysis. *Annals of Mathematical Statistics*, 34(1):122–148, 1963.
2. M. Anthony and P. Bartlett. *Learning in Neural Networks: Theoretical Foundations*. Cambridge, England: Cambridge University Press, 1999.
3. C. T. H. Baker. *The numerical treatment of integral equations*. Clarendon Press, Oxford, 1977.
4. 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.
5. Nello Cristianini, Huma Lodhi, and John Shawe-Taylor. Latent semantic kernels for feature selection. Technical Report NC-TR-00-080, NeuroCOLT Working Group, http://www.neurocolt.org, 2000.
6. 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*, 19(1):260–271, 1991.
7. A. T. James. The distribution of the latent roots of the covariance matrix. *Annals of Math. Stat.*, 31:151–158, 1960.
8. J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
9. V. Koltchinskii and E. Gine. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.
10. D. N. Lawley. Tests of Significance for the Latent Roots of Covariance and Correlation Matrices. *Biometrika*, 43(1/2):128–136, 1956.
11. C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, 1989.
12. 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.

13. Andrew Y. Ng, Alice X. Zheng, and Michael I. Jordan. Link analysis, eigenvectors and stability. In *To appear in the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, 2001.
14. J. Shawe-Taylor, N. Cristianini, and J. Kandola. On the Concentration of Spectral Properties. In T. G. Diettrich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.
15. C. M. Waternaux. Asymptotic Distribution of the Sample Roots for a Nonnormal Population. *Biometrika*, 63(3):639–645, 1976.
16. C. K. I. Williams and M. Seeger. The Effect of the Input Density Distribution on Kernel-based Classifiers. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*. Morgan Kaufmann, 2000.
17. H. Zhu, C. K. I. Williams, R. J. Rohwer, and M. Morciniec. Gaussian regression and optimal finite dimensional linear models. In C. M. Bishop, editor, *Neural Networks and Machine Learning*. Springer-Verlag, Berlin, 1998.