3

The Karhunen-Loève (KL) Expansion

3.1 INTRODUCTION

The goal of this chapter is to address how dimensionality reducing mappings are optimized over the class of orthogonal transformations. Orthogonal transformations are linear, thus the optimization problem amounts to finding a best basis for the change of coordinates in the sense of Equation (2.1). We may visualize this transformation as a rotation of the coordinates which reveals the subspace in which the data resides. Given the nature of the theory in the linear framework it will be seen that optimal, in a sense to be made explicit, orthogonal transformations are natural and easy to use.

The main result is the well-known Karhunen-Loève (KL) expansion [28, 35]. Its importance in pattern analysis is substantiated by the number of aliases under which the technique is known, which include the Proper Orthogonal Decomposition (POD) [35], Principal Component Analysis (PCA) [24, 26], and Empirical Orthogonal Functions (EOFs) [36]. It is also closely related to the well-known singular value decomposition (SVD)[22]. It is rather astonishing, given its rich history and widespread application, that the KL procedure has received so little attention in standard courses in applied mathematics, linear algebra and engineering.

Our presentation begins with with the framework for defining *optimal bases* in Section 3.2. This is followed by a special case, i.e., computing the best line for a collection of points in the plane. This reduces the KL procedure to its simplest setting. The formulae derived here are indeed a special case of the general equations in higher dimensions. We then derive the general KL procedure in Section 3.4. The resulting approach will be referred to as the *direct method*, following [43, 44, 45]. Section 3.5 presents the most important and widely used properties of the KL expansion. The mathematical framework of the linear theory of optimal transformations is discussed in detail and the optimality criteria which lead to the derivation of the KL expansion are fully characterized.

The direct method for implementing the KL transformation cannot be applied to elements of high-dimensional vector spaces, e.g., dimensions above 1000, unless an alternative approach is used, which we will refer to as the *snapshot method* given its natural application to digital images. In Section 3.6, we present this technique and apply it to the Rogue's Gallery problem, i.e., the characterization of high resolution digital images of human faces, a problem initiated in [46, 29].

Section 3.7 re-examines the KL transformation from the perspective of the singular value decomposition described in Section 2.9. The SVD permits a deeper understanding of the relationship between the Direct and Snapshot methods for computing the eigenvectors.

In Section 3.8 we present an extension of the KL procedure for gappy data proposed in [13]. This is followed by a discussion of the application of the KL procedure in the presence of noise in Section 3.9.

3.2 WHAT IS AN OPTIMAL BASIS?

Consider an N-dimensional inner product space V equipped with an ordered o.n. basis $\mathcal{B} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$. Every point in V may be expressed without error in terms of the basis vectors as

$$\mathbf{x}^{(\mu)} = a_1^{(\mu)}\mathbf{v}^{(1)} + \cdots + a_N^{(\mu)}\mathbf{v}^{(N)} \tag{3.1}$$

where $a_i^{(\mu)} = (\mathbf{x}^{(\mu)}, \mathbf{v}^{(i)})$. It is clear that the coefficients $a_i^{(\mu)}$ in the expansion depend on the basis $\mathcal{B}$.

The purpose of this chapter is to mathematically characterize the notion of an *optimal basis*. Given a data set, how should a basis $\mathcal{B}$ be constructed such that the truncation of the full N-term expansion in equation (3.1) to a D-term term expansion

$$\mathbf{x}_D^{(\mu)} = a_1^{(\mu)}\mathbf{v}^{(1)} + a_2^{(\mu)}\mathbf{v}^{(2)} \cdots + a_D^{(\mu)}\mathbf{v}^{(D)} \tag{3.2}$$

will produce a minimum error? Given $\mathbf{x}_D^{(\mu)}$ is an approximation to $\mathbf{x}^{(\mu)}$ we write

$$\mathbf{x}^{(\mu)} \approx \mathbf{x}_D^{(\mu)}$$

and the accuracy of this expression will be at the center of our discussion.

Now we assume we have a collection, or ensemble, of P patterns $\{\mathbf{x}^{(\mu)}\}$, with each $\mathbf{x}^{(\mu)} \in V \subset \mathbb{R}^N$. In practice, an optimal basis for V will extract, or package, the salient features and information in the data. Ideally, this setting will enhance our ability to study the data in terms of a significantly reduced number of expansion coefficients.

The basis will be optimal because it minimizes and error over the set of all o.n. bases. The error vector for each pattern $\varepsilon_D^{(\mu)}$ is the difference between the exact point $\mathbf{x}^{(\mu)}$ and the truncated expansion $\mathbf{x}_D^{(\mu)}$, i.e.,

$$\varepsilon_D^{(\mu)} = \mathbf{x}^{(\mu)} - \mathbf{x}_D^{(\mu)}.$$

A scalar measure of the error is then simply

$$\epsilon = \|\varepsilon_D^{(\mu)}\|$$

As shown below, a closed form formula for the best basis may be obtained for the square error

$$\epsilon_{se} = \|\varepsilon_D^{(\mu)}\|^2.$$

Actually, we will be interested in representing a whole family of patterns with minimum error and require the basis to characterize all of the patterns equally well on average. To quantify this, define the ensemble average of a set of vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(P)}$ as

$$\langle \mathbf{x} \rangle = \frac{1}{P}\sum_{\mu=1}^{P} \mathbf{x}^{(\mu)} \tag{3.3}$$

It should be noted that the above addition is applied component-wise; it is standard practice to omit the pattern index μ for terms within the angled brackets when writing an ensemble average.

It is customary to mean subtract each pattern in the ensemble. This is geometrically equivalent to moving the center of the coordinate system of the patterns to the ensemble average (or *centroid*) of all the data set. Thus we define a new ensemble

$$\tilde{\mathbf{x}}^{(\mu)} = \mathbf{x}^{(\mu)} - \langle \mathbf{x} \rangle.$$

Definition 3.1. *The quantity $\tilde{\mathbf{x}}^{(\mu)}$ is called the fluctuating field, or charicature, of the pattern $\mathbf{x}^{(\mu)}$.*

In what follows we will assume, unless otherwise stated, that all the pattern vectors have been mean subtracted and we drop the tilda notation for convenience.

Definition 3.2. *The mean square error ϵ_{mse} of a $D-$term approximation to an ensemble of vectors is defined as*

$$\epsilon_{mse} = \langle \|\mathbf{x} - \mathbf{x}_D\|^2 \rangle \tag{3.4}$$

$$= \langle \|\varepsilon_D^{(\mu)}\|^2 \rangle \tag{3.5}$$

Unless explicitly stated otherwise, we shall assume that the norm above is induced by the usual Euclidean inner product.

Let's re-examine our previous remarks in terms of subspaces. We may decompose $\mathbf{x}^{(\mu)}$ in two pieces as

$$\mathbf{x}^{(\mu)} = \sum_{i=1}^{D} a_i^{(\mu)} \mathbf{v}^{(i)} + \sum_{i=D+1}^{N} a_i^{(\mu)} \mathbf{v}^{(i)} \tag{3.6}$$

$$= \mathbf{x}_D^{(\mu)} + \varepsilon_D^{(\mu)} \tag{3.7}$$

The basis for this vector expansion may be used to define the subspaces $W_D = span\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(D)}\}$ and $W_D^{\perp} = span\{\mathbf{v}^{(D+1)}, \ldots, \mathbf{v}^{(N)}\}$. These subspaces split V into two pieces

$$V = W_D \oplus W_D^{\perp}$$

where the truncated representations $\mathbf{x}_D^{(\mu)}$ lie in W_D and the error vectors $\varepsilon_D^{(\mu)}$ lie in $W_D^{\perp}$.

While the orthogonal projection theorem tells the orthogonal expansion provides a "best approximation", it says nothing about how to find W_D such that we obtain the best possible best approximation. Thus our task is to determine a single basis which provides the *optimal subspace* W_D for any level of truncation $1 \leq D < N$. Again, optimal here means that a well-defined error should be a minimized over all possible D-dimensional subspaces.

We now continue our discussion of optimal bases in terms of a familiar data analysis problem.

3.3 ON LINES OF BEST FIT

In 1901 Karl Pearson published a paper entitled "On Lines and Planes of Closest Fit to a System of Points" [39]. This paper presented many of the important ideas that eventually developed into a general approach for dimensionality reduction. In this section we consider a simple example examined by Pearson, i.e., the problem of finding the *closest* line through a collection of points. Our treatment will be from a somewhat different perspective, specifically, we wish to determine the best one-dimensional subspace W_1 such that the decomposition

$$\mathbb{R}^2 = W_1 \oplus W_2$$

is optimal. This discussion permits us to understand the KL procedure in a simple setting and to compare it with other classical approaches.

One of the first approaches one encounters in the modeling of data is the well-known method of least squares [48]. For instance, this technique can be used to determine the best linear model for a collection of points lying in the plane. We begin by assuming that an ensemble of 2-tuples $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ are connected by the relation

$$y = ax + b \tag{3.8}$$

Here, it is assumed that the $\{x^{(i)}\}$ are known exactly and the $\{y^{(i)}\}$ should be fit as well as possible. Equation (3.8) is the best line that relates y to x and is referred to as the the regression line of y on x. Now we must determine the *best* choices for the parameters a and b such that this linear model produces a minimum error. Indeed, how we choose to compute a and b is what distinguishes standard least squares fitting of data with the method proposed by Pearson.

Thus, it is of interest to investigate different possibilities for computing the error of the best line approximation to a collection of points. This discussion will reveal the connection between least squares and a special two-dimensional case of the Karhunen-Loève expansion. It is noteworthy that this is one of the relatively few examples which permit an explicit calculation of what we will refer to later as the optimal eigenvectors.

The straight line approximation provided by least squares is determined by computing the values of a and b which minimize the average error

$$\langle \epsilon_y^2 \rangle = (y^{(1)} - ax^{(1)} - b)^2 + \cdots + (y^{(N)} - ax^{(N)} - b)^2. \tag{3.9}$$

The quantity $(y^{(1)} - ax^{(1)} - b)$ is the difference between the actual value of the data point $y^{(1)}$ and its *modeled* value $\tilde{y}^{(1)} = ax^{(1)} + b$. It is a straight forward application (see problem 3.1) of linear algebra to show that the best values of a and b are given by

$$a = \frac{\sum_{i=1}^N x^{(i)} y^{(i)}}{\sum_{i=1}^N (x^{(i)})^2}, \qquad b = 0 \tag{3.10}$$

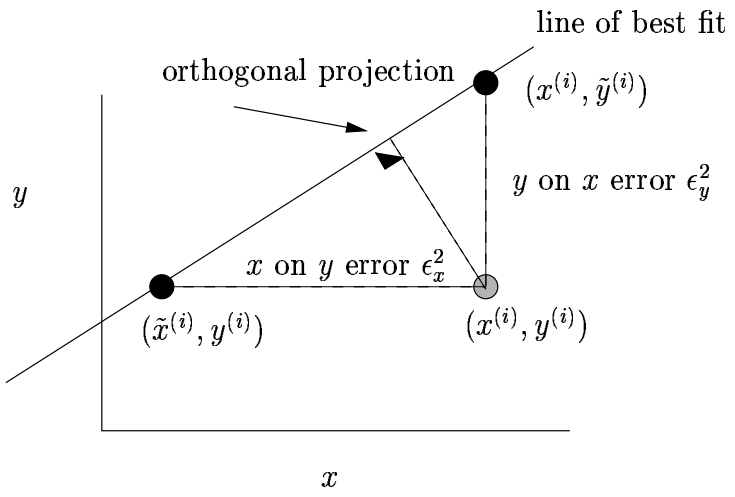

Fig. 3.1 The line of best fit. In standard least squares the line is computed by minizing one of the errors $\langle \epsilon_y^2 \rangle$ or $\langle \epsilon_x^2 \rangle$. The best KL fit corresponds to minimizing the magnitude of the component orthogonal to the best line.

where we have assumed for simplicity that the mean of the data is zero, i.e., $\sum_{i=1}^{N} x^{(i)} = \sum_{i=1}^{N} y^{(i)} = 0$.

As Pearson pointed out, it is inherent in this model that y is the dependent variable and that x is the independent variable. In practice, however, it is often not a simple matter to designate one variable as the dependent variable and the other as the independent. For instance, consider $x^{(i)}$ to be the length of someone's leg and $y^{(i)}$ to be the length of their arm. It might be conjectured that there is a linear relation between these but it seems rather arbitrary to identify one as the dependent variable.

It is simple to develop an equivalent to equation (3.9) where we view x as the dependent variable and y as the dependent variable, namely

$$\langle \epsilon_x^2 \rangle = (x^{(1)} - cy^{(1)} - d)^2 + \cdots + (x^{(N)} - cy^{(N)} - d)^2. \tag{3.11}$$

This model is based on the equation $x = cy + d$ and it suggests that given the value $y^{(i)}$ minimize the predicted value $x^{(i)}$ and again the error is minimized over all the the points in the ensemble.

Based on these considerations then, it is natural to consider a measure of the error which does not require that the variables be treated differently. This is accomplished by finding the line which is closest on average to every point in the data set. See Figure 3.3 for a geometrical comparison of the errors which may be minimized in the construction of the line of best fit.

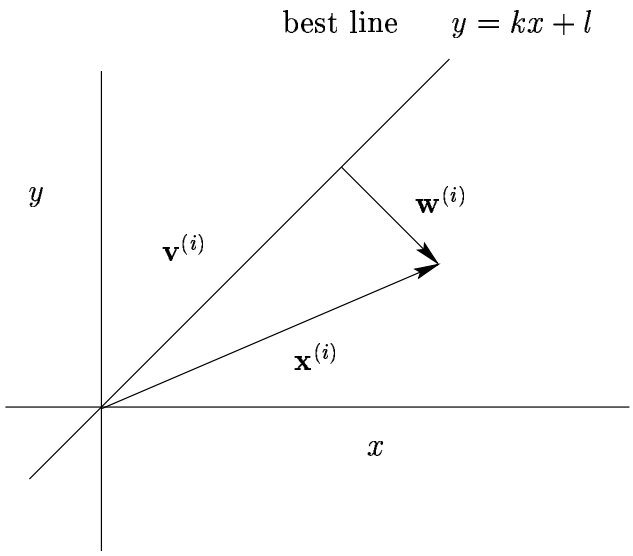

Fig. 3.2 The line of best fit is found by minimizing the magnitude of $\|\mathbf{w}^{(i)}\|$ on average. Equivalently, the average magnitude of the orthogonal projection $\mathbf{v}^{(i)}$ should be maximized.

We begin by writing the model equation as $y = kx + l$. Note that, similar to the example above, if data has been mean subtracted then one can show that the intercept $l = 0$, see Problem 3.2. First write a single point as a vector $\mathbf{u}^{(i)} = (x^{(i)}, y^{(i)}) \in \mathbb{R}^2$. Also, write the vector pointing in the direction of the model line as $\mathbf{m} = (1, k)$. Thus the line $y = kx$ may be written as the vector $\mathbf{v} = \alpha\mathbf{m}$ where $\alpha \in \mathbb{R}$ as shown in Figure 3.3.

To examine this question further, write $\mathbf{w}^{(i)} = \mathbf{x}^{(i)} - \mathbf{v}^{(i)}$ where $\mathbf{v}^{(i)} = \alpha_i\mathbf{m}$. There is a unique value of α_i which corresponds to the orthogonal projection of $\mathbf{x}^{(i)}$ onto $\mathbf{v}^{(i)}$. It is this value of α_i which provides the vector along $\mathbf{m}$ best approximating $\mathbf{x}^{(i)}$. Thus α_i may be found geometrically by requiring that $\mathbf{v}^{(i)}$ and $\mathbf{w}^{(i)}$ be orthogonal, i.e.,

$$(\mathbf{w}^{(i)}, \mathbf{v}^{(i)}) = 0.$$

Substituting $\mathbf{w}^{(i)} = \mathbf{x}^{(i)} - \mathbf{v}^{(i)}$ we have the following series of calculations which provide α_i:

$$(\mathbf{x}^{(i)} - \mathbf{v}^{(i)}, \mathbf{v}^{(i)}) = 0$$
$$(x^{(i)} - \alpha_i)\alpha_i + (y^{(i)} - \alpha_i k)\alpha_i k = 0$$

and assuming $\alpha_i \neq 0$

$$x^{(i)} - \alpha_i + y^{(i)}k - \alpha_i k^2 = 0$$
$$\alpha_i = \frac{x^{(i)} + ky^{(i)}}{1 + k^2}$$

Recall that for each data point it is the vector $\mathbf{w}^{(i)}$ which gives the error. This is the fact we will use to determine an expression for the ensemble average error as a function of k.

We now evaluate the expression for the magnitude of the error $\|\mathbf{w}^{(i)}\|^2$ as a function of the line slope k:

$$\begin{aligned}\|\mathbf{w}^{(i)}\|^2 &= (\mathbf{w}^{(i)}, \mathbf{w}^{(i)}) \\ &= (\mathbf{x}^{(i)} - \alpha_i \mathbf{m}, \mathbf{w}^{(i)}) \\ &= (\mathbf{x}^{(i)}, \mathbf{w}^{(i)}) - \alpha_i (\mathbf{m}, \mathbf{w}^{(i)}) \\ &= (\mathbf{x}^{(i)}, \mathbf{w}^{(i)})\end{aligned}$$

where we have used the fact that $(\mathbf{m}, \mathbf{w}^{(i)}) = 0$. Now

$$\begin{aligned}\|\mathbf{w}^{(i)}\|^2 &= (\mathbf{x}^{(i)}, \mathbf{w}^{(i)}) \\ &= (\mathbf{x}^{(i)}, \mathbf{x}^{(i)} - \alpha_i \mathbf{m}) \\ &= \|\mathbf{x}^{(i)}\|^2 - \alpha_i (\mathbf{x}^{(i)}, \mathbf{m}) \\ &= (x^{(i)})^2 + (y^{(i)})^2 - \frac{(x^{(i)} + ky^{(i)})^2}{1 + k^2}\end{aligned}$$

Since we are to minimize the error over all the data points in the ensemble we must find the quantity k^* such that

$$\langle \|\mathbf{w}(k^*)\|^2 \rangle = \min_k \langle \|\mathbf{w}(k)\|^2 \rangle. \tag{3.12}$$

We first write the error explicitly as a function of k

$$\langle \|\mathbf{w}(k)\|^2 \rangle = \langle x^2 + y^2 - \frac{(x + ky)^2}{1 + k^2} \rangle \tag{3.13}$$

where we have dropped the subscripts of the data points as per convention. We now have an expression for the ensemble average error which we may view as a function of the slope of the line k. To determine the k^* which minimizes this error we require

$$\frac{\partial \langle \|\mathbf{w}(k)\|^2 \rangle}{\partial k} = 0 \tag{3.14}$$

Differentiating equation (3.13) gives

$$\frac{\partial \langle \|\mathbf{w}(k)\|^2 \rangle}{\partial k} = -\langle \frac{2(x + ky)y(1 + k^2) - 2k(x + ky)^2}{(1 + k^2)^2} \rangle \tag{3.15}$$

and upon setting this equal to zero and simplifying we obtain

$$\langle (x + k^* y) y (1 + k^{*2}) - k^* (x + k^* y)^2 \rangle = 0. \tag{3.16}$$

Factoring further simplifies this expression

$$\langle (x + k^* y)(y - k^* x) \rangle = 0$$

from which it follows

$$\langle (1 - k^{*2}) xy + k^* (y^2 - x^2) \rangle = 0$$

and consequently

$$(1 - k^{*2}) \langle xy \rangle + k^* \langle y^2 - x^2 \rangle = 0.$$

If we define

$$\beta = \frac{\langle x^2 \rangle - \langle y^2 \rangle}{\langle xy \rangle}$$

then we have for our optimal value of k^* the solutions to

$$k^{*2} + \beta k^* - 1 = 0 \tag{3.17}$$

which has the solutions

$$k^*_{\pm} = \frac{-\beta \pm \sqrt{\beta^2 + 4}}{2} \tag{3.18}$$

Hence, after some simplification

$$k^*_{\pm} = \frac{\langle y^2 \rangle - \langle x^2 \rangle}{2\langle xy \rangle} \pm \frac{\sqrt{\langle x^2 - y^2 \rangle^2 + 4\langle xy \rangle^2}}{2\langle xy \rangle} \tag{3.19}$$

Observe that there are two values of k^*, i.e., k^*_- and k^*_+, which satisfy the constraint equation (3.14). The line which is orthogonal to the value of k^* which minimizes the error would be the line which would correspond to the worst value of k.

We note that the procedure described in this section is sometimes referred to as *total least squares* [17].

3.4 CONSTRUCTION OF THE OPTIMAL BASIS.

We now consider two related derivations of the Karhunen-Loève expansion. The first approach is based on minimizing the error term resulting from an orthogonal projection onto a D-dimensional subspace. We refer to the derivation as the *simulataneous approach* given that the subspaces are defined by a single optimization problem. The second approach proceeds by maximizing the mean-square projection of the data onto sequential one-dimensional subspaces. The two approaches produce equivalent results while providing different interpretations of the procedure.

3.4.1 The Simultaneous Approach

Now we consider the actual construction of the optimal basis which, for any $1 \leq D \leq N$, produces a pair of orthogonal subspaces

$$V = W_D \dot{\oplus} W_D^{\perp}$$

that minimizes the mean square truncation error ϵ_{mse}. Given an ensemble of vectors $\{\mathbf{x}^{(\mu)}\}_{\mu=1}^{P}$, with each $\mathbf{x}^{(\mu)} \in V$ and $\dim V = N$, we seek a set of basis vectors $\{\boldsymbol{\phi}^{(j)}\}_{j=1}^{N}$ such that the error of the truncated expansion is minimized in the mean-square sense. Recall that any pattern vector $\mathbf{x}^{(\mu)}$ may be written without error as

$$\mathbf{x}^{(\mu)} = \sum_{j=1}^{N} a_j^{(\mu)} \boldsymbol{\phi}^{(j)}.$$

The expansion error vector may be expressed in terms of the basis since

$$\begin{aligned} \varepsilon_D^{(\mu)} &= \mathbf{x}^{(\mu)} - \mathbf{x}_D^{(\mu)} \\ &= \sum_{j=D+1}^{N} a_j^{(\mu)} \boldsymbol{\phi}^{(j)}. \end{aligned}$$

On average we have

$$\begin{aligned} \epsilon_{mse} &= \langle \|\varepsilon_D^{(\mu)}\|^2 \rangle \\ &= \langle (\varepsilon_D^{(\mu)}, \varepsilon_D^{(\mu)}) \rangle \\ &= \langle (\sum_{j=D+1}^{N} a_j \boldsymbol{\phi}^{(j)}, \sum_{k=D+1}^{N} a_k \boldsymbol{\phi}^{(k)}) \rangle \\ &= \langle \sum_{j,k=D+1}^{N} a_j a_k (\boldsymbol{\phi}^{(j)}, \boldsymbol{\phi}^{(k)}) \rangle \end{aligned}$$

which upon invoking the orthonormality relation gives

$$\epsilon_{mse} = \langle \sum_{j=D+1}^{N} a_j^2 \rangle$$

and hence

$$\epsilon_{mse} = \langle \sum_{j=D+1}^{N} (\mathbf{x}, \boldsymbol{\phi}^{(j)})^2 \rangle.$$

Noting that $(\mathbf{x}, \boldsymbol{\phi})^2 = (\boldsymbol{\phi}, \mathbf{x}\mathbf{x}^T\boldsymbol{\phi})$ and defining $\mathbf{C} = \langle \mathbf{x}\mathbf{x}^T \rangle$ we can write

$$\begin{aligned} \epsilon_{mse} &= \langle \sum_{j=D+1}^{N} (\boldsymbol{\phi}^{(j)}, \mathbf{x}\mathbf{x}^T\boldsymbol{\phi}^{(j)}) \rangle \\ &= \sum_{j=D+1}^{N} (\boldsymbol{\phi}^{(j)}, \langle \mathbf{x}\mathbf{x}^T \rangle \boldsymbol{\phi}^{(j)}) \end{aligned}$$

Thus the total mean-square-error due to truncating the expansion is

$$\epsilon_{mse} = \sum_{j=D+1}^{N} (\boldsymbol{\phi}^{(j)}, \mathbf{C}\boldsymbol{\phi}^{(j)}). \tag{3.20}$$

We compute the eigenvectors $\boldsymbol{\phi}^{(i)}$ which make ϵ_{mse} a minimum using the technique of Lagrange multipliers [1] subject to the constraints $(\boldsymbol{\phi}^{(j)}, \boldsymbol{\phi}^{(j)}) = 1$. These constraints ensure nontrivial solutions for the extrema. As usual, define the functional

$$g(\boldsymbol{\phi}^{(D+1)}, \ldots, \boldsymbol{\phi}^{(N)}) = \sum_{j=D+1}^{N} (\boldsymbol{\phi}^{(j)}, \mathbf{C}\boldsymbol{\phi}^{(j)}) - \sum_{j=D+1}^{N} \lambda_j ((\boldsymbol{\phi}^{(j)}, \boldsymbol{\phi}^{(j)}) - 1)$$

In what follows we employ the notation

$$\nabla_{\mathbf{v}}(\cdot) = (\frac{\partial(\cdot)}{\partial v_1}, \ldots, \frac{\partial(\cdot)}{\partial v_N})^T.$$

It is a simple exercise to show that

$$\nabla_{\mathbf{v}}(\mathbf{v}, \mathbf{v}) = 2\mathbf{v}$$

and that given $\mathbf{C}$ is a symmetric matrix

$$\nabla_{\mathbf{v}}(\mathbf{v}, \mathbf{C}\mathbf{v}) = 2\mathbf{C}\mathbf{v}.$$

To obtain the extrema we must simultaneously require that

$$\nabla_{\boldsymbol{\phi}^{(j)}} g(\boldsymbol{\phi}^{(D+1)}, \ldots, \boldsymbol{\phi}^{(N)}) = 0$$

for $j = D+1 \ldots N$.

The best basis vectors are then provided by solving

$$\nabla_{\boldsymbol{\phi}^{(j)}} g = 2\mathbf{C}\boldsymbol{\phi}^{(j)} - \lambda_j 2\boldsymbol{\phi}^{(j)} = 0,$$

i.e., the eigenvector problem

$$\mathbf{C}\boldsymbol{\phi}^{(j)} = \lambda_j \boldsymbol{\phi}^{(j)}. \tag{3.21}$$

We refer to the basis $\{\phi^{(i)}\}$ as the KL eigenvectors, or KL basis.

The astute reader may object to the omission of the constraint that the eigenvectors be required to be orthogonal in the formulation of the Lagrange multiplier problem. This did not disrupt the derivation given that the lagrange multipliers of the omitted constraints are in fact zero. This is established in the next derivation. Note also that the resulting symmetric eigenvector problem always will produce a set of o.n. basis vectors.

Depending on the size of the ambient dimension N and the number of patterns P, we employ either the *direct* method ($N \geq P$) or the *snapshot* method. These techniques are discussed in full later and given a unified treatment via the singular value decomposition.

3.4.2 The Sequential Approach

In order to develop a better intuition concerning the properties of the KL eigenvectors we consider another derivation of the equations for the best basis. It proceeds sequentially as follows:

- Find the best one-dimensional subspace W_1.
- Find the best one-dimensional subspace W_2 with the restriction that it be orthogonal to W_1.
- Find the best one-dimensional subspace W_i with the restriction that $W_i \perp W_j$ for all $j < i$.

This process will result in the same eigenvector problem as in the simultaneous derivation of the previous section. In fact, the same equation is produced after the first step, an indication of the tightly knit structure of the linear theory. For simplicity we will assume that the eigenvalues corresponding to the best basis vectors are distinct.

Now we define the best first eigenvector $\boldsymbol{\phi}^{(1)}$ to be the one which maximizes the mean-square projection of all patterns in the ensemble onto itself. Namely, find

$$\max_{\boldsymbol{\phi}^{(1)}} \langle (\boldsymbol{\phi}^{(1)}, \mathbf{x})^2 \rangle$$

$$\text{subject to } \; (\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(1)}) = 1.$$

Again, this problem may be solved via the technique of Lagrange multipliers. Write as before

$$\begin{aligned} g_1(\boldsymbol{\phi}^{(1)}) &= \langle (\boldsymbol{\phi}^{(1)}, \mathbf{x})^2 \rangle - \lambda_1 [(\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(1)}) - 1] \\ &= (\boldsymbol{\phi}^{(1)}, \mathbf{C}\boldsymbol{\phi}^{(1)}) - \lambda_1 [(\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(1)}) - 1] \end{aligned}$$

Differentiating w.r.t. $\phi^{(i)}$ we obtain

$$\nabla_{\boldsymbol{\phi}^{(1)}} D(\boldsymbol{\phi}^{(1)}) = 2\mathbf{C}\boldsymbol{\phi}^{(1)} - \lambda_1 2\boldsymbol{\phi}^{(1)} = 0$$

or

$$\mathbf{C}\boldsymbol{\phi}^{(1)} = \lambda_1 \boldsymbol{\phi}^{(1)}.$$

Note that this eigenvector problem is a necessary condition. It is associated with an extremum, i.e., either a maxima or a minima. Hence both the best and worst directions will satisfy this equation. The next best basis direction should satisfy the above requirements of maximum projection, but with the added restriction that it be orthogonal to the best direction $\boldsymbol{\phi}^{(1)}$. Thus, the second eigenvector $\boldsymbol{\phi}^{(2)}$ is found requiring

$$\max_{\boldsymbol{\phi}^{(2)}} \langle (\boldsymbol{\phi}^{(2)}, \mathbf{x})^2 \rangle$$

$$\text{subject to } (\boldsymbol{\phi}^{(2)}, \boldsymbol{\phi}^{(2)}) = 1 \text{ and } (\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}) = 0$$

where now $\phi^{(1)}$ is assumed to be the (now fixed) o.n. vector found above.

Again, the method of Lagrange multipliers requires us to find the extrema of

$$g_2(\boldsymbol{\phi}^{(2)}) = (\boldsymbol{\phi}^{(2)}, \mathbf{C}\boldsymbol{\phi}^{(2)}) - \lambda_2[(\boldsymbol{\phi}^{(2)}, \boldsymbol{\phi}^{(2)}) - 1] - \mu(\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}).$$

Now differentiating w.r.t. $\boldsymbol{\phi}^{(2)}$ we obtain

$$\nabla_{\boldsymbol{\phi}^{(2)}} D(\boldsymbol{\phi}^{(2)}) = 2\mathbf{C}\boldsymbol{\phi}^{(2)} - 2\lambda_2 \boldsymbol{\phi}^{(2)} - 2\mu\boldsymbol{\phi}^{(1)} = 0.$$

Hence, taking the inner product with $\boldsymbol{\phi}^{(1)}$ we can show that μ must be zero because of the orthogonality condition. Namely,

$$(\boldsymbol{\phi}^{(1)}, \mathbf{C}\boldsymbol{\phi}^{(2)}) - \lambda_2(\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}) - \mu(\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(1)}) = 0.$$

But

$$(\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}) = 0$$

and

$$\begin{aligned}(\boldsymbol{\phi}^{(1)}, \mathbf{C}\boldsymbol{\phi}^{(2)}) &= (\mathbf{C}\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}) \\ &= (\lambda_1 \boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}) \\ &= 0.\end{aligned}$$

Using these facts we have

$$\mu(\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(1)}) = 0$$

from which we conlude $\mu = 0$ since $\|\boldsymbol{\phi}^{(1)}\| = 1$.

The process for determining the i'th best eigenvector given the first $i-1$ eigenvectors is analogous and is investigated in Exercise 3.17.

Ordering of the Optimal Basis

A natural ordering for the optimal basis $\{\boldsymbol{\phi}^{(j)}\}$ is provided by the spectrum (i.e., the discrete set of eigenvalues) of $\mathbf{C}$. Recall that the first eigenvector was found by requiring that

$$\langle(\boldsymbol{\phi}^{(1)}, \mathbf{x})^2\rangle = \text{maximum}$$

or, equivalently, that $\langle a_1^2\rangle = \lambda_1 = \text{maximum}$. Proceeding in this fashion, the second eigenvalue is defined so that $\lambda_2 = \text{maximum}$, subject to the constraint that the associated coordinate direction $\boldsymbol{\phi}^{(2)}$ be orthogonal to $\boldsymbol{\phi}^{(1)}$. Hence

$$\lambda_1 \geq \lambda_2.$$

The remaining eigenvalues are defined iteratively such that each λ_i is a maximum subject to the requirement that the associated coordinate direction $\boldsymbol{\phi}^{(i)}$ be orthogonal to $\{\boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(i-1)}\}$. Hence at each step $\lambda_i \geq \lambda_{i+1}$, so we conclude

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0.$$

Therefore the eigenvectors can be ordered naturally according to the amount of variance contained in their respective directions. The pattern $\mathbf{x}^{(\mu)}$ is then approximated by the basis vectors $\boldsymbol{\phi}$ corresponding to the largest eigenvalues of $\mathbf{C}$.

Special Cases

It is interesting to consider two special cases that may occur for the eigenvalues and to interpret their significance geometrically.

$\lambda_i = \lambda_j$ with $i \neq j$: In this first case, the eigenvectors associated with these eigenvalues are not unique, although the subspace they span is. We will see below that this may also be interpretted as each direction storing exactly the same amount of information. If there exist a set of k equal eigenvectors then they define a unique k-dimensional subspace although the eigenvectors themselves are not unique.

$\lambda_i = 0$ for some i: In this second case, there is no variance in the coordinate along this direction. Hence we may conclude that no information is present and this coordinate may be safely truncated.

3.5 GENERAL PROPERTIES OF THE KL EXPANSION

In this section we outline the useful properties of the KL decomposition. The framework of the optimal orthogonal transformation is especially rich, and the optimality condition takes on many (equivalent) aspects.

Property 3.5.1. *The $N \times N$ matrix $\mathbf{C}$ is referred to as the ensemble averaged covariance matrix. It is symmetric and determines an ordered set of N orthogonal eigenvectors with associated real eigenvalues.*

The following property is actually true for any data set which has zero mean.

Property 3.5.2. *For an ensemble of fluctuating vectors, the coordinate values a_j are zero on average. To see this write*

$$\begin{aligned}\langle a_j \rangle &= \langle (\mathbf{x}, \boldsymbol{\phi}^{(j)}) \rangle \\ &= (\langle \mathbf{x} \rangle, \boldsymbol{\phi}^{(j)}) \\ &= (0, \boldsymbol{\phi}^{(j)}) \\ &= 0\end{aligned}$$

Property 3.5.3. *The KL expansion coefficients are uncorrelated on average, i.e.,*

$$\langle a_j a_k \rangle = 0$$

when $j \neq k$.

$$\begin{aligned}\langle a_j a_k \rangle &= \langle (\mathbf{x}, \boldsymbol{\phi}^{(j)})(\mathbf{x}, \boldsymbol{\phi}^{(k)}) \rangle \\ &= \langle (\boldsymbol{\phi}^{(j)}, \mathbf{x}\mathbf{x}^T \boldsymbol{\phi}^{(k)}) \rangle \\ &= (\boldsymbol{\phi}^{(j)}, \langle \mathbf{x}\mathbf{x}^T \rangle \boldsymbol{\phi}^{(k)}) \\ &= (\boldsymbol{\phi}^{(j)}, \mathbf{C}\boldsymbol{\phi}^{(k)}) \\ &= (\boldsymbol{\phi}^{(j)}, \lambda_k \boldsymbol{\phi}^{(k)}) \\ &= \lambda_k \delta_{jk}\end{aligned}$$

In particular, $\langle a_j a_k \rangle = 0$ when $j \neq k$. Although the data is uncorrelated *on average* in the KL coordinate system, it is possible that the data is correlated on subsets of the total ensemble. E.g., in a time-series setting it is possible for the data to have short time correlations in the KL basis coordinates. If $\tilde{X} \subset X$ then we may write $\langle a_j a_k \rangle_{\tilde{X}} \neq 0$.

Property 3.5.4. *The eigenvalues of $\mathbf{C}$ are non-negative*

$$\lambda_j = \langle a_j^2 \rangle \geq 0,$$

for $j = 1 \ldots N$.

This follows directly from the previous property for the case $j = k$. Note that the number of non-zero eigenvalues is equal to the rank of the matrix C which in turn equals the dimension of the space spanned by the data set. See Section 3.7 for more details.

It is also possible to view our derivation as maximizing the variance along each coordinate direction, subject to orthogonality constraints.

Property 3.5.5. *For mean subtracted data, the statistical variance of the jth coordinate direction is proportional to the jth eigenvalue of $\mathbf{C}$.*

We write the statistical variance of the jth coordinate direction over the ensemble of patterns as $\text{var}(a_j)$ where

$$\begin{aligned}\text{var}(a_j) &= \frac{1}{P-1}\sum_{\mu=1}^{P}(a_j^{(\mu)} - \langle a_j\rangle)^2 \\ &= \frac{P}{P-1}\frac{1}{P}\sum_{\mu=1}^{P}(a_j^{(\mu)})^2 \\ &= \frac{P}{P-1}\langle a_j^2\rangle = \frac{P}{P-1}\lambda_j\end{aligned}$$

That is,

$$\text{var}(a_j) \propto \lambda_j$$

where we used the fact $\langle a_j\rangle = 0$.

Property 3.5.6. *The eigenvalues of C give a measure of the truncation error.*

$$\epsilon_{mse} = \sum_{j=D+1}^{N}\lambda_j \tag{3.22}$$

Substituting the eigenvector equation $\mathbf{C}\boldsymbol{\phi}^{(j)} = \lambda_j\boldsymbol{\phi}^{(j)}$, into

$$\epsilon_{mse} = \sum_{j=D+1}^{N}(\boldsymbol{\phi}^{(j)}, \mathbf{C}\boldsymbol{\phi}^{(j)})$$

then,

$$\epsilon_{mse} = \sum_{j=D+1}^{N}(\boldsymbol{\phi}^{(j)}, \lambda_j\boldsymbol{\phi}^{(j)}) \tag{3.23}$$

$$= \sum_{j=D+1}^{N}\lambda_j \tag{3.24}$$

Property 3.5.7. *The KL basis captures more statistical variance than any other basis. Let $\{\boldsymbol{\psi}^{(i)}\}_{i=1}^{N}$ be any other basis for the inner product space V and write the D-term expansion for an element of V as*

$$\mathbf{x}_D^{(\mu)} = \sum_{j=1}^{D} b_j^{(\mu)}\boldsymbol{\psi}^{(i)}.$$

Define

$$\rho_j = \langle b_j^2 \rangle.$$

$$\sum_{j=1}^{D} \rho_j \leq \sum_{j=1}^{D} \lambda_j \tag{3.25}$$

with equality when $\{\psi^{(i)}\}$ *is the KL basis.*

Definition 3.3. *A data set is said to be translationally invariant if* $\mathbf{x} \in X$ *implies that any cyclic permutation of the vector* $\mathbf{x}$ *is also in* X.

Property 3.5.8. *If* X *is a translationally invariant data set, then the optimal eigenvectors are the Fourier vectors, i.e., sinusoids.*

We will prove this in Section 3.7.1. Thus, for translationally invariant data, the discrete Fourier transform provides an analaytical form for the best basis.

Shannon's Entropy

A standard measure of information is provided by Shannon's entropy which is defined as

$$H = -\sum_{i=1}^{N} P_i \ln P_i$$

where $\sum_{i=1}^{N} P_i = 1$. If we interpret the normed eigenvalues of the covariance matrix

$$\tilde{\lambda}^{(i)} = \frac{\lambda^{(i)}}{\sum_{j=1}^{N} \lambda^{(j)}}$$

as the probabilities P_i, then it is possible to show that the KL eigenvectors are optimal in an information theoretic sense, i.e., they minimize H [51].

The significance of Shannon's entropy H in this context is that it provides a measure of the distribution of the magnitude of the eigenvalues, or energy, across the coordinates of the basis. In particular, if the probabilities are all constant with

$$P_i = \frac{1}{N}$$

for all $i = 1 \ldots N$, then

$$H = \ln N$$

Also, if

$$\begin{cases} P_i = 1 & \text{if } i = 1, \\ P_i = 0 & \text{if } 1 < i \leq N \end{cases}$$

then

$$H = 0.$$

See [27] for a proof of these facts.

In these two extreme cases we can see that if all the eigenvalues are equal then there is no compression of information, i.e., there is no preferred coordinate direction and H is a maximum. On the other hand, if there is only one non-zero eigenvalue, then all the information is contained along one coordinate and H is a minimum.

Property 3.5.9. *The Karhunen-Loève basis minimizes Shannon's entropy.*

Proof. Following [11], consider an arbitrary o.n. basis $\mathcal{B}$ for the data set X consisting of ordered vectors $\{\boldsymbol{\psi}^{(j)}\}_{j=1}^{N}$ and the expansion for $\mathbf{x} \in X$, i.e.,

$$\mathbf{x}^{(\mu)} = \sum_{j=1}^{N} b_j^{(\mu)} \boldsymbol{\psi}^{(j)}$$

With respect to this basis, the total variance of the j'th coordinate is given by

$$\rho^{(j)} = \langle b_j^2 \rangle$$

and the normalized variance by

$$\tilde{\rho}^{(j)} = \frac{\rho^{(j)}}{\sum_{i=1}^{N} \rho^{(i)}}$$

Furthermore, we assume that the basis $\{\boldsymbol{\psi}^{(j)}\}_{j=1}^{N}$ is ordered according to the variance

$$\rho^{(1)} \geq \rho^{(2)} \geq \cdots \geq \rho^{(N)} \geq 0.$$

The total entropy may be written as a function of the basis

$$H(\mathcal{B}) = -\sum_{j=1}^{N} \tilde{\rho}^{(j)} \ln \tilde{\rho}^{(j)}.$$

Because the KL system maximizes the variance we have

$$\sum_{k=1}^{j} \tilde{\rho}^{(k)} \leq \sum_{k=1}^{j} \tilde{\lambda}^{(k)}$$

Now define this left term as $\alpha_j = \sum_{k=1}^{j} \tilde{\rho}^{(k)}$. This quantity corresponds to the fraction of variance represented by the first j coordinates. Now the entropy may be rewritten as

$$H = -\sum_{j=1}^{N} (\alpha_j - \alpha_{j-1}) \ln(\alpha_j - \alpha_{j-1})$$

To obtain the extrema we differentiate

$$\frac{\partial H}{\partial \alpha_m} = -\sum_{j=1}^{N}(\frac{\partial \alpha_j}{\partial \alpha_m} - \frac{\partial \alpha_{j-1}}{\partial \alpha_m})(1 + \ln(\alpha_j - \alpha_{j-1}))$$
$$= \ln\left(\frac{\alpha_{j+1} - \alpha_j}{\alpha_j - \alpha_{j-1}}\right)$$

It is an exercise to verify that

$$\frac{\alpha_{j+1} - \alpha_j}{\alpha_j - \alpha_{j-1}} \leq 1 \tag{3.26}$$

from which we conclude that

$$\frac{\partial H}{\partial \alpha_m} \leq 0$$

so H is a decreasing function of α_m. Thus, H is minimized when α_m is maximized, but this is exactly the property of the KL basis. □

Truncation Criteria

It is common in some applications to refer to the statistical variance as *energy* given it is a measure of amplitudes squared. Using this terminology, the total energy in the data is denoted

$$E_N = \sum_{i=1}^{N} \lambda_i.$$

The energy captured by a D term expansion is given by

$$E_D = \sum_{i=1}^{D} \lambda_i.$$

Typically, for purposes of comparison, we will be interested in the normalized energy

$$\tilde{E}_D = \frac{E_D}{E_N}$$

Now one can also interpret the quantity

$$\tilde{\lambda}_i = \frac{\lambda_i}{E_N}$$

as the probablity that a pattern is contained in the subspace spanned by the eigenvector $\phi^{(i)}$. Note that the normalized mean square error is readily available as

$$\epsilon_{nmse} = \sum_{i=D+1}^{N} \frac{\lambda_i}{E_N} = \tilde{E}_N - \tilde{E}_D.$$

We will refer to a plot of the eigenvalues versus the eigenvalue index as a KL-spectrum plot. Often we will plot $\log(\lambda_i)$ versus i to enhance visualization of sharp decreases in the eigenvalues. Also, it is often useful to plot $\tilde{E}_D$ as a function of the number of terms D in the expansion. These plots are used to estimate the so called KL dimension of the data. This dimension is generally taken as the number of terms required to ensure that some minimum quantity of energy is captured by the data.

Several *ad-hoc* criteria have been proposed for determining the number of terms D to retain in the expansion

$$\mathbf{x} = \sum_{j=1}^{D} a_j \boldsymbol{\phi}^{(j)}.$$

A simple but widely used energy based criterion is to retain the number of terms necessary to capture a specified fraction of the total energy [16]. Specifically, we have the normalized energy criterion

$$\tilde{E}_D > \gamma \tag{3.27}$$

or equivalently, that the normalized mean square error

$$\epsilon_{nmse} < 1 - \gamma \tag{3.28}$$

should be less than some constant γ, typically taken to be $\gamma \in [0.90, 0.99]$. The equivalent constraints specified by Equations (3.27) and (3.28) can be shown to be connected to the Frobenius norm of the data matrix, see Problem 3.20.

In addition it is often useful to add the restriction that

$$\frac{\lambda_{D+1}}{\lambda_1} < \delta \tag{3.29}$$

where $\delta = 0.01$, for example. This is a restriction on the 2-norm of the data matrix. We summarize these remarks with the following definitions:

Definition 3.4. *The KL energy dimension, written dim(KLE_γ), is defined to be the minimum number of terms required in the orthogonal expansion to ensure that $\tilde{E}_D > \gamma$.*

Definition 3.5. *The KL magnification dimension, written dim(KLM_δ), is defined to be the minimum number D required to ensure that $\lambda_{D+1}/\lambda_1 < \delta$.*

In addition, it is useful to combine these definitions into a total KL dimension, written $dim(KLD_{\gamma,\delta})$ which may be defined as the maximum of KLE_γ and KLM_δ. Note that the utility of these global definitions of dimension is limited by the requirement of making *ad-hoc* choices for γ and δ. In Section 4.6, we will present a scaling argument which eliminates the need for these parameters in the estimation of local dimension.

While this definition of KL-dimension seems to be connected to other measures of dimensionality, there are also other criteria which have been proposed for determining the number of terms to retain in a best basis expansion. It has been observed that the KL-spectrum often can be viewed as two lines. The point where these lines intersect determines the value of D. In fact, there is considerable evidence that in many cases the data along the second line corresponds to noise. Although this is a risky assumption as we shall see. Finally, it should be emphasized that the utility of these measures is greatly enhanced if they can be implemented in a problem dependent fashion. It is clear that one may require far fewer terms for a classification problem than a reconstruction problem where more details in the pattern are required.

Matrix Notation

In this section we would like to reinterpret the expansions as linear transformations and emphasize the dimensionality reducing properties of these transformations. We begin by constructing a matrix Φ made up of the eigenvectors of C, i.e.,

$$\Phi = [\boldsymbol{\phi}^{(1)}|\boldsymbol{\phi}^{(2)}|\ldots|\boldsymbol{\phi}^{(N)}]$$

Thus the coefficients of the pattern vector w.r.t. the KL basis are now

$$\mathbf{a}^{(\mu)} = \boldsymbol{\Phi}^T \mathbf{x}^{(\mu)}$$

where $\mathbf{a}^{(\mu)} = (a_1^{(\mu)}, \ldots, a_N^{(\mu)})^T$. These relations may be combined to give

$$A = \Phi^T X \tag{3.30}$$

If we have determined a number of terms D to retain in our expansion, clearly it is not required to compute all the terms in the expansion. Hence, it is useful to define a dimensionality reducing transformation based on Φ_D where

$$\Phi_D = [\boldsymbol{\phi}^{(1)}|\boldsymbol{\phi}^{(2)}|\ldots|\boldsymbol{\phi}^{(D)}]$$

is a matrix with D columns, namely the first D eigenvectors. Now, the D coefficients are given by

$$\hat{\mathbf{a}}^{(\mu)} = \Phi_D^T \mathbf{x}^{(\mu)}$$

where $\hat{\mathbf{a}}^{(\mu)} = (a_1^{(\mu)}, \ldots, a_D^{(\mu)})^T$. Or, in matrix notation,

$$\hat{A} = \Phi_D^T X$$

where $\hat{A}$ is a $D \times P$ matrix. It is identical to the first D rows of A in Equation (3.30).

Property 3.5.10. *The KL basis diagonalizes the ensemble averaged covariance matrix C*

$$\begin{aligned}\langle \mathbf{a}\mathbf{a}^T \rangle &= \langle (\Phi^T \mathbf{x})(\mathbf{x}^T \Phi) \rangle \\ &= \Phi^T C \Phi \\ &= \Lambda\end{aligned}$$

where $\Lambda_{ii} = \lambda_i$ and all the off diagonal elements are zero.

Property 3.5.11. *We now have the spectral decomposition of the covariance matrix as $C = \Phi \Lambda \Phi^T$, i.e.,*

$$C = \lambda_1 \boldsymbol{\phi}^{(1)} {\boldsymbol{\phi}^{(1)}}^T + \lambda_2 \boldsymbol{\phi}^{(2)} {\boldsymbol{\phi}^{(2)}}^T + \cdots + \lambda_N \boldsymbol{\phi}^{(N)} {\boldsymbol{\phi}^{(N)}}^T$$

This allows us to decompose the covariance matrix in an optimal way. Note that if $P < N$ then we do not expect more than a basis of P vectors. The remaining $N - P$ vectors belong to the null space of C.

3.6 THE SNAPSHOT METHOD

The construction of a *data-dependent* basis as outlined above requires solving the eigenvector problem

$$\mathbf{C}\boldsymbol{\phi}^{(j)} = \lambda_j \boldsymbol{\phi}^{(j)}$$

where $\mathbf{C}$ is an $N \times N$ matrix consisting of the average of P rank one covariance matrices. If N is large, say of order $O(10^4)$, then it is generally not possible to solve this problem directly. There are a variety of techniques for computing the largest eigenvalues and eigenvectors but if the matrix $\mathbf{C}$ is a singular matrix, then the problem may be reduced without approximation to an eigenvector problem of size $P \times P$. The technique is referred to as the *Snapshot Method* because of its applicability to data sets consisting of high-resolution digital snapshots [46, 43, 29].

The fact that the basis is data dependent may be made explicit.

Proposition 3.1. *If $\lambda_j > 0$, then*

$$\boldsymbol{\phi}^{(j)} = \sum_{\nu=1}^{P} \alpha_\nu^{(j)} \mathbf{x}^{(\nu)} \tag{3.31}$$

The data spans the same space as the eigenvectors corresponding to eigenvalues with non-zero variance.

For simplicity write $\boldsymbol{\phi}^{(1)} = \boldsymbol{\phi}$ and $\alpha_\nu^{(1)} = \alpha_\nu$ We can reformulate the variational problem for the first eigenvector as

$$h(\boldsymbol{\phi}) = \langle (\boldsymbol{\phi}, \mathbf{x})^2 \rangle - \lambda((\boldsymbol{\phi}, \boldsymbol{\phi}) - 1).$$

Then

$$h(\boldsymbol{\alpha}) = \langle (\sum_{\nu=1}^{P} \alpha_\nu \mathbf{x}^{(\nu)}, \mathbf{x})^2 \rangle - \lambda((\sum_{\nu} \alpha_\nu \mathbf{x}^{(\nu)}, \sum_{\xi} \alpha_\xi \mathbf{x}^{(\xi)}) - 1)$$
$$= \langle [\sum_{\nu=1}^{P} \alpha_\nu (\mathbf{x}^{(\nu)}, \mathbf{x})]^2 \rangle - \lambda(\sum_{\nu,\xi} \alpha_\nu \alpha_\xi (\mathbf{x}^{(\nu)}, \mathbf{x}^{(\xi)}) - 1)$$

Then

$$\frac{\partial h(\boldsymbol{\alpha})}{\partial \alpha_\tau} = 0$$

for $\tau = 1, \ldots, P$. Now the basis may be optimized w.r.t. the coefficient vectors $\boldsymbol{\alpha}^{(j)}$ in place of the eigenvectors, the advantage being that there are only P of them, rather than N original basis vectors.

Differentiating $h(\boldsymbol{\alpha})$ gives

$$\frac{\partial h(\boldsymbol{\alpha})}{\partial \alpha_\tau} = \langle 2 \sum_{\nu} \alpha_\nu (\mathbf{x}^{(\nu)}, \mathbf{x})(\mathbf{x}^{(\tau)}, \mathbf{x}) \rangle - 2\lambda \sum_{\xi} \alpha_\xi (\mathbf{x}^{(\xi)}, \mathbf{x}^{(\tau)}) = 0 \qquad (3.32)$$

Expanding the ensemble average gives

$$\frac{1}{P} \sum_{\nu,\xi} \alpha_\nu (\mathbf{x}^{(\nu)}, \mathbf{x}^{(\xi)})(\mathbf{x}^{(\tau)}, \mathbf{x}^{(\xi)}) - \lambda \sum_{\xi} \alpha_\xi (\mathbf{x}^{(\xi)}, \mathbf{x}^{(\tau)}) = 0$$

Note that a common term may be factored as

$$\sum_{\xi} (\mathbf{x}^{(\tau)}, \mathbf{x}^{(\xi)}) [\sum_{\nu} \alpha_\nu (\mathbf{x}^{(\nu)}, \mathbf{x}^{(\xi)}) - P\lambda\alpha_\xi] = 0$$

If we write

$$\beta_\xi = \sum_{\nu} \alpha_\nu (\mathbf{x}^{(\nu)}, \mathbf{x}^{(\xi)}) - P\lambda\alpha_\xi \qquad (3.33)$$

then for each $\tau = 1 \ldots P$ it follows

$$\sum_{\xi} (\mathbf{x}^{(\tau)}, \mathbf{x}^{(\xi)}) \beta_\xi = 0. \qquad (3.34)$$

Defining

$$L_{\nu\mu} = (\mathbf{x}^{(\nu)}, \mathbf{x}^{(\mu)})$$

Equation (3.34) may be written in matrix notation as

$$L\boldsymbol{\beta} = 0 \qquad (3.35)$$

and Equation (3.33) as

$$L\boldsymbol{\alpha} - \lambda\boldsymbol{\alpha} = \boldsymbol{\beta}$$

The trivial solution $\boldsymbol{\beta} = 0$ of Equation (3.35) leads to the eigenvalue problem

$$L\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \tag{3.36}$$

i.e., a $P \times P$ eigenvector problem.

Remark 3.6.1. *Again, the proof here is for the first eigenvector, but the same approach can be used to show that the equation above gives all the solutions.*

This result is very useful if the number of patterns P is manageable, typically $P < 1000$. The dimension N now enters in only in storage space and add/multiplies.

3.6.1 The Rogues Gallery Problem

The Rogue's Gallery problem, introduced in [46, 29], is an application of the KL expansion for the low-dimensional characterization of human faces. This problem is an excellent example of a situation where the direct computation of the eigenvectors (i.e., solutions of equation (3.21)) which form the optimal basis is impossible. Instead we must employ the snapshot method and solve equation (3.36).

For this application an ensemble of 200 digital photographs of human faces was collected. The population was restricted to be homogeneous, i.e., neither eyeglasses nor beards were permitted. Each digital image is given as an $M \times N$ array of pixels and in this particular experiment the raw data is captured with $M = N = 256$. In addition all of the photographs were aligned to match a template and adjusted for depth. Furthermore, the data was normalized so that each image would have uniform lighting on average.

The average face is required to perform the analysis, this is displayed in Figure 3.3. After the faces where mean subtracted the eigenpictures where calculated. The first four of these are shown in Figure 3.4. A sampling of eigenpictures, i.e., numbers 9, 18, 116, 196, are shown in Figure 3.5 to give an idea of the change in the images as a function of the associated eigenvalue.

To demonstrate the efficacy of the basis for representing the digital images partial reconstructions retaining 10, 20, 30, 40, 50 and 60 terms are shown in Figure 3.6. The face being reconstructed is *not* part of the original data set of 200 faces.

This example is extended to exploit symmetry properties in Section 4.3. In addition, a constrast of the linear reconstruction with a nonlinear reconstruction is discussed in Section 9.3.1.

3.7 THE SINGULAR VALUE DECOMPOSITION AND KL

In this section we re-examine the direct method and the snapshot method for implementing the Karhunen-Loève decomposition via the singular value

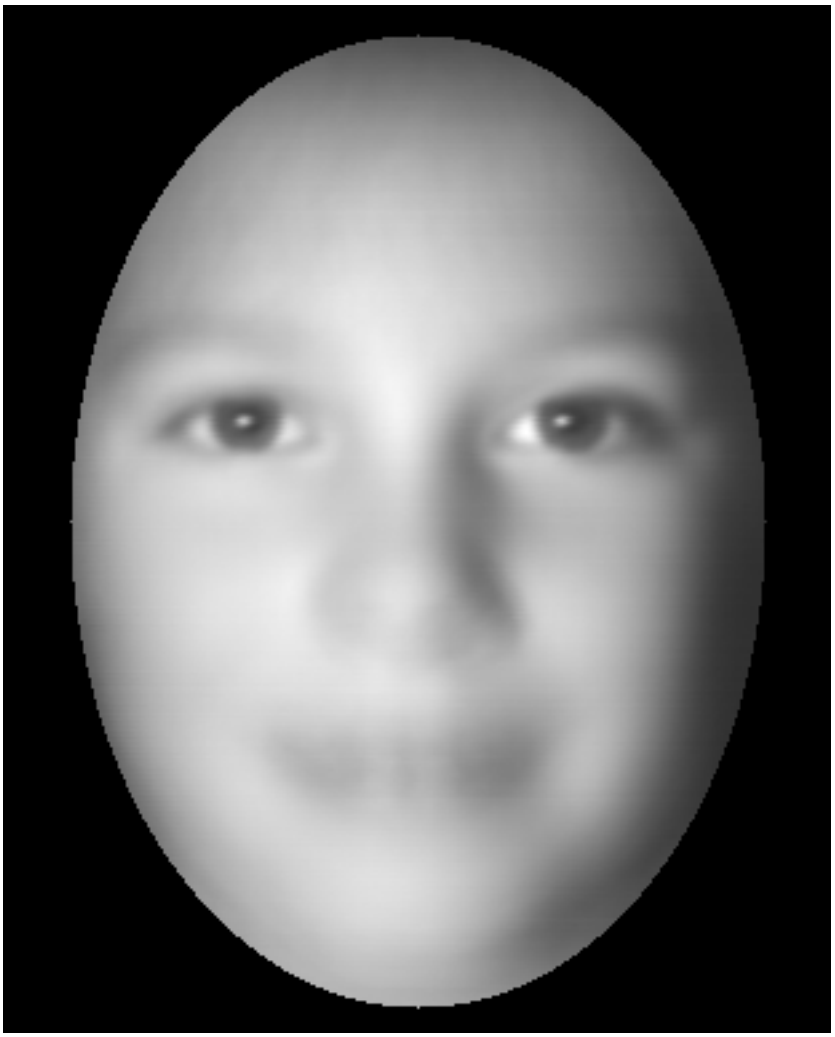

Fig. 3.3 The average face.

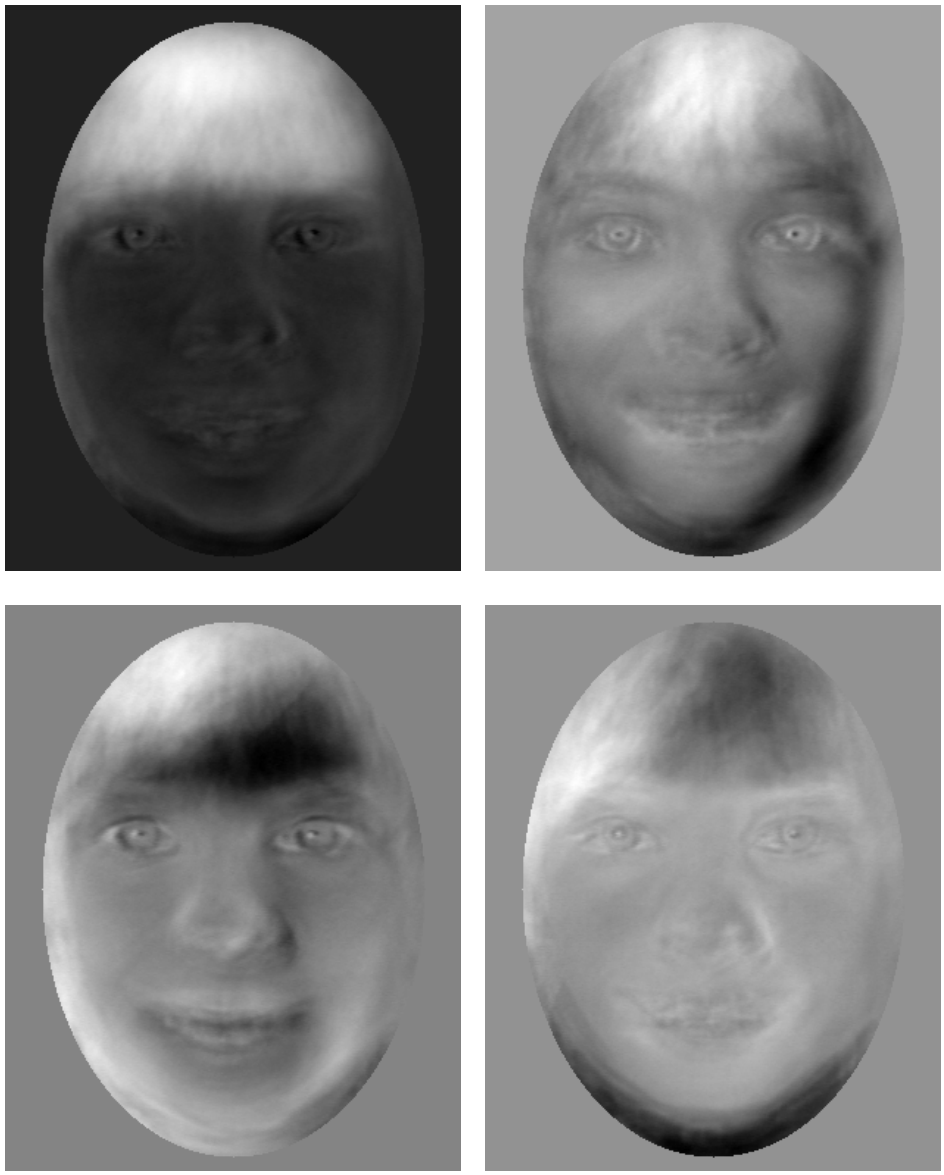

Fig. 3.4 The first 4 eigenfaces.

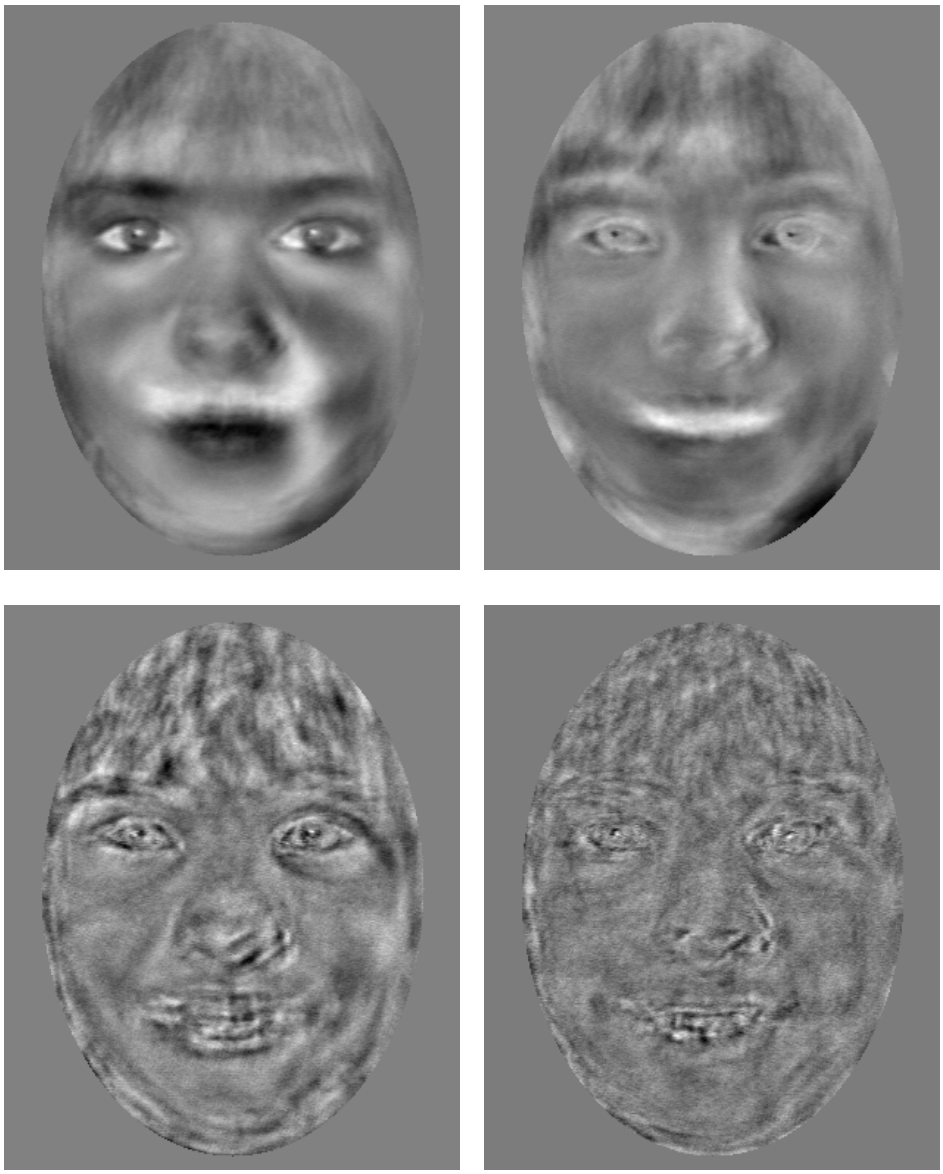

Fig. 3.5 Top left: eigeface number 9; top right: eigenface 18; bottom left: eigenface 116; bottom right: eigenface 196.

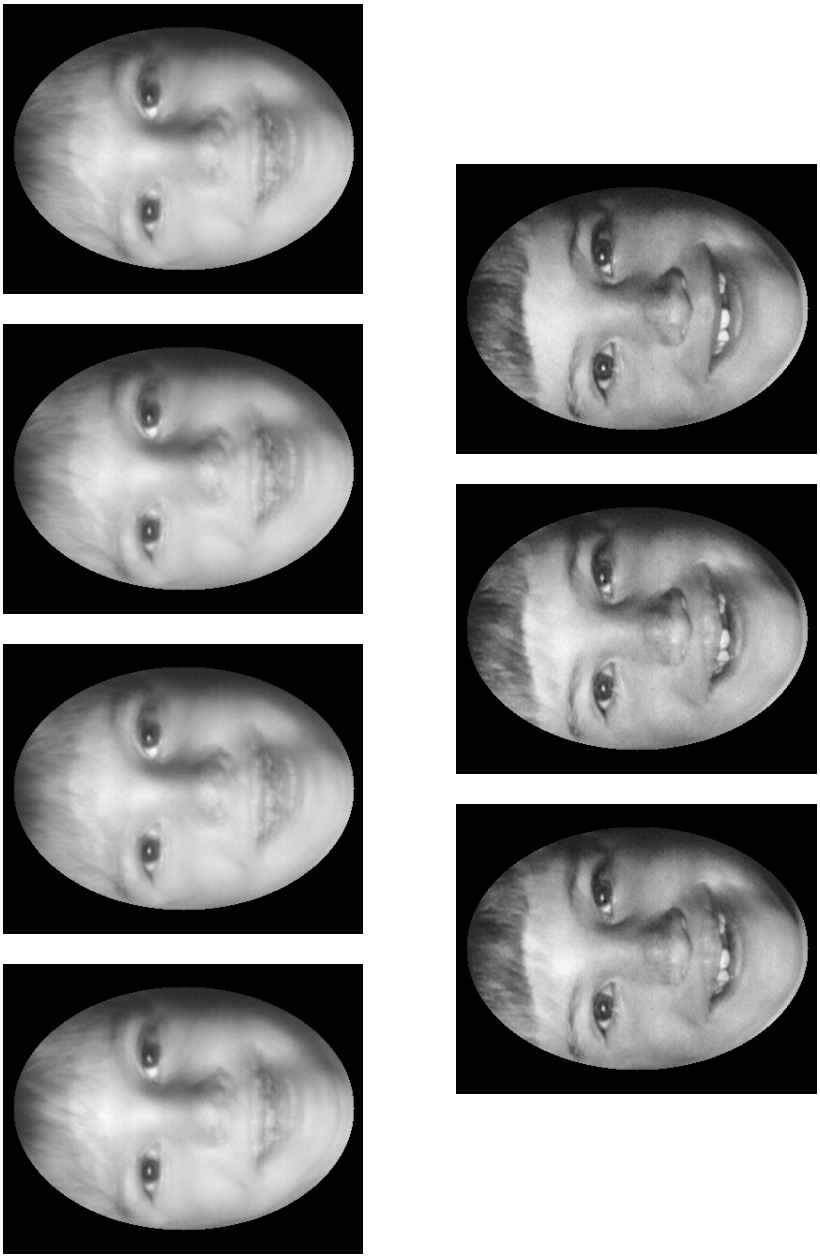

Fig. 3.6 Eigenface reconstruction. The reconstructions consist of 10, 20, 30, 40, 50 and 60 terms from left to right, top to bottom. The picture in the bottom right corresponds to the original image.

decomposition (SVD). The SVD, as described in Section 2.9, is a classical and powerful tool in numerical linear algebra. Detailed textbook discussions are available, see e.g., [17, 50]; additional references and theory may be found in [23]. Here our purpose is to demonstrate that the eigenvector problems stated in equations (3.21) and (3.36) associated with the direct method snapshot method, respectively, fit neatly into the mathematical framework of the SVD. In particular, we shall see that the left singular vectors are the eigenvectors computed in the direct method, while the right singular vectors are the eigenvectors computed in the snapshot method. The SVD may also be seen to be equivalent to what has been referred to as the bi-orthogonal decomposition, see, e.g., [2].

We begin by constructing a $N \times P$ data matrix X out of our ensemble $\{\mathbf{x}^{(\mu)}\}_{\mu=1}^{P}$ of pattern vectors in $\mathbb{R}^N$ where the columns of X are the pattern vectors $X = [\mathbf{x}^{(1)}|\ldots|\mathbf{x}^{(P)}]$.

To assist in the interpretation of our results, we will assume that the ensemble consists of time-dependent vectors. For simplicity we assume that the spatial variable is 1-dimensional. It should be emphasized that these assumptions are for convenience and are not a requirement of the theory. For these time dependent observations, $(X)_{ij} = x_i^{(j)}$, the column index is the time index $j = 1, \ldots, P$, and the row index, $i = 1, \ldots, N$ is the spatial index. Thus our *spatio-temporal* data matrix

$$X = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(P)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(P)} \\ \vdots & \vdots & \ddots & \vdots \\ x_N^{(1)} & x_N^{(2)} & \cdots & x_N^{(P)} \end{bmatrix}$$

is indexed left-to-right by time and top-to-bottom by space. Note that the size of this matrix may be enormous in practice and its actual formation may not be possible due to computer memory limitations. However, this does not prevent us from applying the mathematics of the SVD.

The transpose of X will be written X^T and is given by

$$X^T = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_N^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_N^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(P)} & x_2^{(P)} & \cdots & x_N^{(P)} \end{bmatrix}$$

In what follows, it will be useful to write the matrix X^T in terms of its column vectors. Therefore, we introduce the notation

$$X^T = [\mathbf{y}^{(1)}|\ldots|\mathbf{y}^{(N)}]$$

where $y_j^{(i)} = x_i^{(j)}$. One might view X^T as a new data matrix where the roles of time and space have been interchanged.

The motivation for collecting the data in matrix form comes from the observation that we may rewrite the ensemble averaged correlation matrix in terms of XX^T. To see this, write out the jk'th element of this matrix

$$(XX^T)_{jk} = \sum_{\mu=1}^{P} x_j^{(\mu)} x_k^{(\mu)} \tag{3.37}$$

$$= P\langle x_j x_k \rangle \tag{3.38}$$

where $j, k = 1, \ldots, N$. In other words,

$$\frac{1}{P} XX^T = \langle \mathbf{x}\mathbf{x}^T \rangle.$$

From the above expression, we see that the patterns are being correlated over the spatial variable and averaged over time. Thus, we define the ensemble averaged *spatial correlation matrix* C_x of the observations as

$$C_x = \frac{1}{P} XX^T. \tag{3.39}$$

In an analogous manner we may form the ensemble averaged *temporal correlation matrix*

$$C_t = \frac{1}{N} X^T X \tag{3.40}$$

As before, let's write out the jk'th element of this matrix

$$(X^T X)_{jk} = \sum_{l=1}^{N} x_l^{(j)} x_l^{(k)} \tag{3.41}$$

$$= N\langle y_j y_k \rangle \tag{3.42}$$

where $j, k = 1, \ldots, P$. In other words,

$$\frac{1}{N} X^T X = \langle \mathbf{y}\mathbf{y}^T \rangle$$

so we see that C_t is in fact a temporal correlation matrix. Note that the definition of the ensemble, and hence ensemble average, has changed from that given above. When determining C_x, the spatial correlations are averaged over time while in determining C_t the temporal correlations are averaged over the spatial domain. Hence we will refer to the eigenvectors of C_x as the *spatial eigenvectors* and to the eigenvectors of C_t as the *temporal eigenvectors.*

We note that C_x is an $N \times N$ matrix and the spatial eigenvectors are solutions of

$$XX^T U = U\Lambda_x \tag{3.43}$$

where the columns of U correspond to the eigenvectors of C_x, i.e., $U = [\mathbf{u}^{(1)}|\ldots|\mathbf{u}^{(N)}]$. Also, Λ_x is an $N \times N$ matrix

$$\Lambda_x = \begin{bmatrix} \lambda^{(1)} & & & \\ & \lambda^{(2)} & & \\ & & \ddots & \\ & & & \lambda^{(N)} \end{bmatrix}.$$

Similarly, C_t is an $P \times P$ matrix and the temporal eigenvectors are solutions of

$$X^T X V = V \Lambda_t \tag{3.44}$$

where the columns of V correspond to the eigenvectors of C_t, i.e., $V = [\mathbf{v}^{(1)}|\ldots|\mathbf{v}^{(N)}]$. Also, and Λ_t is a $P \times P$ matrix

$$\Lambda_t = \begin{bmatrix} \lambda^{(1)} & & & \\ & \lambda^{(2)} & & \\ & & \ddots & \\ & & & \lambda^{(P)} \end{bmatrix}.$$

We have used the same notation for the eigenvalues of XX^T and $X^T X$ in view of the following proposition which is a consequence of the fact $\det(XX^T) = \det(X^T X)$:

Proposition 3.2. *The non-zero entries of Λ_x and Λ_t are equal.*

Hence, the KL spectrum can be determined from computing the eigenvalues of either C_x or C_t. Note that the eigenvalues of C_x are actually given by the matrix $\frac{1}{P}\Lambda_x$ and that the eigenvalues of C_t are given by the matrix $\frac{1}{N}\Lambda_t$.

Now we recast these results in terms of the SVD. We recognize that the spatial eigenvectors U of the spatial correlation matrix XX^T are exactly the left-singular vectors of the data matrix X. Also, the temporal eigenvectors V of the temporal correlation matrix $X^T X$ are exactly the right-singular vectors of the data matrix X. Thus, by the singular value decomposition theorem 2.4 we may decompose the data matrix

$$X = U \mathbf{\Sigma} V^T \tag{3.45}$$

where $\mathbf{\Sigma}$ is the $N \times P$ diagonal matrix given by

$$\mathbf{\Sigma} = \text{diag}(\sigma^{(1)}, \ldots, \sigma^{(r)}, 0, \ldots, 0)$$

where $\sigma_i = \sqrt{\lambda_i}$, i.e., the singular values are the square roots of the eigenvalues of XX^T.

Given the columns of U for a basis for the data, we have the *orthogonal expansion*

$$\mathbf{x}^{(\mu)} = \sum_{j=1}^{r} a_j^{(\mu)} \mathbf{u}^{(j)}$$

Recall that this may be rewritten in terms of matrices as

$$X = UA$$

where A is an $N \times P$ matrix of expansion coefficients $A = [\mathbf{a}^{(1)}|\mathbf{a}^{(2)}|\ldots|\mathbf{a}^{(P)}]$. Furthermore, since $U^T = U^{-1}$. the expansion coefficients are given by

$$A = U^T X.$$

The following propositions provide very useful relationships between the expansion coefficients and the eigenvectors.

Proposition 3.3. *If $\mathbf{\Sigma}$ is the matrix of singular values of X and V the matrix of associated temporal eigenvectors (right-singular vectors) then the matrix of expansion coefficients A, i.e., the projections of the data onto the optimal spatial eigenvectors, is given by*

$$A = \mathbf{\Sigma} V^T$$

Proof. By the SVD

$$X = U\mathbf{\Sigma}V^T.$$

Multiplying both sides of the relationship by U^T and using the fact $U^T U = \mathbf{I}$ gives

$$U^T X = \mathbf{\Sigma} V^T.$$

Recognizing $A = U^T X$ completes the result which has an extremely useful interpretation. Namely that the expansion coefficients A are contained in the temporal eigenvectors, i.e., the right-singular vectors. □

Thus, the time dependent coefficients given by the matrix A, may be computed using two different methods. Firstly, we can compute the spatial eigenvectors and the associated projections in the usual fashion. Alternatively, we can compute these coefficients directly from the temporal correlation matrix. Usually one approach is significantly more efficient than the other.

The next proposition states that the spatial eigenvectors may be written as the superposition of data where the appropriate expansion coefficients are provided by temporal eigenvectors, i.e., the right-singular vectors. Compare this result with equation (3.31).

Proposition 3.4.

$$\mathbf{u}^{(j)} = \frac{1}{\sigma_j} \sum_{k=1}^{P} v_k^{(j)} \mathbf{x}^{(k)}$$

where $j = 1, \ldots, rank(X)$.

Proof. This result also follows directly from the SVD

$$X = U\mathbf{\Sigma}V^T$$
$$XV = U\mathbf{\Sigma}$$

where the fact that V is an orthogonal matrix was used. Defining $\mathbf{\Sigma}^+$, $P \times N$ matrix, as the pseudoinverse of $\mathbf{\Sigma}$ we have

$$U = XV\mathbf{\Sigma}^+$$

from which the proposition follows. □

The next proposition presents a relation which is in a sense symmetrical to the previous one. It states that the temporal eigenvectors may be written as the superposition of data where the appropriate expansion coefficients are provided by spatial eigenvectors.

Proposition 3.5.

$$\mathbf{v}^{(j)} = \frac{1}{\sigma_j} \sum_{k=1}^{N} u_k^{(j)} \mathbf{y}^{(k)}$$

where $j = 1, \ldots, r$, $r = rank(X)$ *and* $y_i^{(k)} = x_k^{(i)}$ *as defined above.*

Proof. Again, this result follows directly from the SVD

$$X = U\mathbf{\Sigma}V^T$$
$$U^T X = \mathbf{\Sigma}V^T$$
$$V^T = \mathbf{\Sigma}^+ U^T X$$

where the fact that U is an orthogonal matrix was used. Hence,

$$V = X^T U\mathbf{\Sigma}^+ \tag{3.46}$$

from which the proposition follows. □

Example 3.1. Recall Example 2.16 where the left and right singular vectors were computed for the data matrix

$$X = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad X^T = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

The columns of X^T are $\mathbf{y}^{(1)} = (1,1)^T$, $\mathbf{y}^{(2)} = (0,1)^T$ and $\mathbf{y}^{(3)} = (1,0)^T$.

We now confirm proposition (3.4). To this end, we compute

$$\mathbf{u}^{(j)} = \frac{1}{\sigma_j} \sum_{k=1}^{P} v_k^{(j)} \mathbf{x}^{(k)}$$

for $\mathbf{u}^{(1)}$. Evaluating this formula gives

$$\mathbf{u}^{(1)} = \frac{1}{\sqrt{3}}(\mathbf{v}_1^{(1)}\mathbf{x}^{(1)} + \mathbf{v}_2^{(1)}\mathbf{x}^{(2)})$$

Recalling $\mathbf{v} = \frac{1}{\sqrt{2}}(1,1)^T$ we obtain

$$\mathbf{u}^{(1)} = \frac{1}{\sqrt{3}}(\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix})$$

Therefore,

$$\mathbf{u}^{(1)} = \frac{1}{\sqrt{6}}(2,1,1)^T$$

which checks.

We now confirm proposition (3.5) by computing

$$\mathbf{v}^{(j)} = \frac{1}{\sigma_j}\sum_{k=1}^{N} u_k^{(j)}\mathbf{y}^{(k)}$$

for $\mathbf{v}^{(1)}$. Recalling $\mathbf{u}^{(1)} = \frac{1}{\sqrt{6}}(2,1,1)^T$ we obtain

$$\mathbf{v}^{(1)} = \frac{1}{\sqrt{3}}(\frac{2}{\sqrt{6}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{\sqrt{6}}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 0 \end{pmatrix}) = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

which checks. It is also reassuring to confirm the formula $A = \mathbf{\Sigma}V^T$ which provides the spatial expansion coefficients in terms of the temporal eigenvectors. By direct computation

$$\begin{aligned} A &= U^T X \\ &= \begin{pmatrix} \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{3}{\sqrt{6}} & \frac{3}{\sqrt{6}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix} \end{aligned}$$

According to the proposition, the $N \times P$ matrix $A = \mathbf{\Sigma}V^T$ is also found as

$$\frac{1}{\sqrt{2}}\begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{3}{2}} & \sqrt{\frac{3}{2}} \\ -\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ 0 & 0 \end{pmatrix}$$

which agrees with the previous result.

3.7.1 Translationally Invariant Data

We consider a data set consisting of points $\{x_i\}$ to be translationally invariant if the spatial domain is periodic and a cyclic permutation of the components of a given data point generates another element of the data set. In matrix notation we can consider the (rectangular) matrix X whose rows are the x_i^T. There is a group of cyclic permutations of the columns of X, that is the components of x_i^T such that which we represent by the circulant matrix $\{C_i\}$. Corresponding to each of the C_i there is a permutation of the rows of X, denoted by P_i, which rearranges the rows of X into their original order

$$P_i X C_i = X$$

Note that $P_i^{-1} = P_i^T$. The right singular vectors of X may be determined by forming $X^T X$, i.e.,

$$X^T X = C_i^T X^T X C_i$$

hence

$$X^T X C_i = C_i^T X^T X$$

So the group of circulant matrices commutes with $X^T X$. Given the C_i and $X^T X$ are *simple*, it follows that they share the same eigenvectors [33]. Since the eigenvectors of the circulant matrices are sinusoids we conclude that the right singular vectors are also sinusoids. The fact that the optimal basis for translationally invariant data is sinusoids is well known. The argument present here follows [6].

3.8 IMPLEMENTATION WITH MISSING DATA

Now we turn to the problem of using the KL procedure on data sets which have gaps, or missing components. The algorithm presented here is due to [13]. The development follows [13], although here we simplify the setting of the presentation using only discrete vector spaces, rather than function spaces. We distinguish this extension of the KL procedure for *gappy data* due to [13] from the case of *noisy data* which is developed in the next section.

3.8.1 Estimating Missing Data

Let $\mathbf{x} \in \mathbb{R}^N$ an a vector which has a reduced expansion in terms of the KL basis as

$$\mathbf{x} \approx \mathbf{x}_D = \sum_{n=1}^{D} a_n \mathbf{u}^{(n)}$$

It follows that only D points of information are required to reproduce the original vector. Consider now an incomplete, or gappy, copy $\tilde{\mathbf{x}}$ of the original