CHAPTER 6

FEATURE GENERATION I: LINEAR TRANSFORMS

6.1 INTRODUCTION

The goal of this chapter is the generation of features via linear transforms of the input (measurement) samples. A number of transforms will be presented and reviewed. The basic concept is to transform a given set of measurements to a new set of features. If the transform is suitably chosen, transform domain features can exhibit high "*information packing*" properties compared with the original input samples. This means that most of the classification-related information is "squeezed" in a relatively small number of features, leading to a reduction of the necessary feature space dimension.

The basic reasoning behind transform-based features is that an appropriately chosen transform can exploit and remove information redundancies, which usually exist in the set of samples obtained by the measuring devices. Let us take for example an image resulting from a measuring device, for example, X-rays or a camera. The pixels (i.e., the input samples) at the various positions in the image have a large degree of correlation, due to the internal morphological consistencies of real-world images that distinguish them from noise. Thus, if one uses the pixels as features, there will be a large degree of redundant information. Alternatively, if one obtains the Fourier transform, for example, of a typical real-world image, it turns out that most of the energy lies in the low-frequency components, due to the high correlation between the pixels. Hence, using the Fourier coefficients as features seems a reasonable choice, because the low-energy, high-frequency coefficients can be neglected, with little loss of information. In this chapter we will see that the Fourier transform is just one of the tools from a palette of possible transforms.

6.2 BASIS VECTORS AND IMAGES

Let $x(0), x(1), \ldots, x(N-1)$ be a set of input samples and $\boldsymbol{x}$ be the $N \times 1$ corresponding vector,

$$\boldsymbol{x}^T = [x(0), \ldots, x(N-1)]$$

Given a unitary $N \times N$ matrix A[1] we define the transformed vector $\boldsymbol{y}$ of $\boldsymbol{x}$ as

$$\boldsymbol{y} = A^H \boldsymbol{x} \equiv \begin{bmatrix} \boldsymbol{a}_0^H \\ \vdots \\ \boldsymbol{a}_{N-1}^H \end{bmatrix} \boldsymbol{x} \tag{6.1}$$

where "H" denotes the Hermitian operation, that is, complex conjugation and transposition. From (6.1) and the definition of unitary matrices we have

$$\boldsymbol{x} = A\boldsymbol{y} = \sum_{i=0}^{N-1} y(i)\,\boldsymbol{a}_i \tag{6.2}$$

The columns of A, $\boldsymbol{a}_i, i = 0, 1, \ldots, N-1$, are called the *basis vectors* of the transform. The elements $y(i)$ of $\boldsymbol{y}$ are nothing but the projections of $\boldsymbol{x}$ onto these basis vectors. Indeed, taking the inner product of $\boldsymbol{x}$ with $\boldsymbol{a}_j$ we have

$$< \boldsymbol{x}, \boldsymbol{a}_j > \equiv \boldsymbol{x}^H \boldsymbol{a}_j = \sum_{i=0}^{N-1} y(i) < \boldsymbol{a}_i, \boldsymbol{a}_j > = \sum_{i=0}^{N-1} y(i)\delta_{ij} = y(j) \tag{6.3}$$

This is due to the unitary property of A, that is, $A^H A = I$ or $< \boldsymbol{a}_i, \boldsymbol{a}_j > = \boldsymbol{a}_i^H \boldsymbol{a}_j = \delta_{ij}$.

In many problems, such as in image analysis, the input set of samples is a two-dimensional sequence $X(i, j), i, j = 0, 1, \ldots, N-1$, defining an $N \times N$ matrix X instead of a vector. In such cases, one can define an equivalent N^2 vector $\boldsymbol{x}$, for example, by ordering the rows of the matrix one after the other (*lexicographic ordering*)

$$\boldsymbol{x}^T = [X(0,0), \ldots, X(0, N-1), \ldots, X(N-1, 0), \ldots, X(N-1, N-1)]$$

and then transform this equivalent vector. However, this is not the most efficient way to work. The number of operations required to multiply an $N^2 \times N^2$ square matrix (A) with an $N^2 \times 1$ vector $\boldsymbol{x}$ is of the order of $O(N^4)$, which is prohibitive for many applications. An alternative possibility is to transform matrix X via a set of *basis matrices* or *basis images*. Let U and V be unitary $N \times N$ matrices.

[1] A complex matrix is called unitary if $A^{-1} = A^H$. Real matrices are equivalently called *orthogonal* if $A^{-1} = A^T$.

Define the transformed matrix Y of X as

$$Y = U^H X V \tag{6.4}$$

or

$$X = U Y V^H \tag{6.5}$$

The number of operations is now reduced to $O(N^3)$. Equation (6.5) can alternatively be written (Problem 6.1) as

$$X = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} Y(i, j) \boldsymbol{u}_i \boldsymbol{v}_j^H \tag{6.6}$$

where $\boldsymbol{u}_i$ are the column vectors of U and $\boldsymbol{v}_j$ the column vectors of V. Each of the outer products $\boldsymbol{u}_i \boldsymbol{v}_j^H$ is an $N \times N$ matrix

$$\boldsymbol{u}_i \boldsymbol{v}_j^H = \begin{bmatrix} u_{i0} v_{j0}^* & \cdots & u_{i0} v_{jN-1}^* \\ \vdots & \ddots & \vdots \\ u_{iN-1} v_{j0}^* & \cdots & u_{iN-1} v_{jN-1}^* \end{bmatrix} \equiv \mathcal{A}_{ij}$$

and (6.6) is an expansion of matrix X in terms of these N^2 basis images (matrices). Furthermore, if Y turns out to be diagonal, then (6.6) becomes

$$X = \sum_{i=0}^{N-1} Y(i, i) \boldsymbol{u}_i \boldsymbol{v}_i^H$$

and the number of basis images is reduced to N. An interpretation similar to (6.3) is also possible. To this end, let us define the inner product between two arrays as

$$< A, B > \equiv \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} A(m, n) B^*(m, n) \tag{6.7}$$

Then it is not difficult to show that (Problem 6.1)

$$Y(i, j) = < X, \mathcal{A}_{ij} > \tag{6.8}$$

In words, the (i, j) element of the transformed matrix results from multiplying each element of X by the conjugate of the corresponding element of $\mathcal{A}_{ij}$ and summing up all products.

Transformations of the type (6.4) are also known as *separable* (Problem 6.2). The reason is that one can look at them as a succession of one-dimensional transforms, first applied on column vectors and then on row vectors. For example, the intermediate result in (6.4), $Z = U^H X$, is equivalent to N transforms applied to the column vectors of X, and $(U^H X)V = (V^H Z^H)^H$ is equivalent to a second sequence of N transforms acting upon the rows of Z. All the two-dimensional transforms that we will deal with in this chapter are separable ones.

Example 6.1. Given the image X and the orthogonal transform matrix U

$$X = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}, \qquad U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

the transformed image $Y = U^T X U$ is

$$Y = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 4 & -1 \\ -1 & 0 \end{bmatrix}$$

The corresponding basis images are

$$\mathcal{A}_{00} = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} [1, 1] = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathcal{A}_{11} = \frac{1}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} [1, -1]$$
$$= \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

and similarly

$$\mathcal{A}_{01} = \mathcal{A}_{10}^T = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$$

Now verify that the elements of Y are obtained via the matrix inner products $< X, \mathcal{A}_{ij} >$.

6.3 THE KARHUNEN–LOÈVE TRANSFORM

Let $\boldsymbol{x}$ be the vector of input samples. In the case of an image array, $\boldsymbol{x}$ may be formed by lexicographic ordering of the array elements. We have already mentioned in a number of places in this book that a desirable property of the generated features is to be mutually uncorrelated in order to avoid information redundancies. The goal of this section is to generate features that are optimally uncorrelated, that is, $E[y(i)y(j)] = 0,\ i \neq j$. Let[2]

$$\boldsymbol{y} = A^T \boldsymbol{x} \tag{6.9}$$

From the definition of the correlation matrix we have

$$R_y \equiv E[\boldsymbol{y}\boldsymbol{y}^T] = E[A^T \boldsymbol{x}\boldsymbol{x}^T A] = A^T R_x A \tag{6.10}$$

However, R_x is a symmetric matrix, and hence its eigenvectors are mutually orthogonal (Appendix B). Thus, if matrix A is chosen so that its columns are the orthonormal eigenvectors $\boldsymbol{a}_i,\ i = 0, 1, \ldots, N-1$, of R_x, then R_y is diagonal

[2] We deal with real data. The complex case is a straightforward extension.

(Append

where Λ
eigenvalu
itive defi
is known
of genera
significar
applicatio

Mean sq

Let us nov

where only
projection
involved in
resulting m

Our goal no
(6.14) and t
have

Combining

(Appendix B)

$$R_y = A^T R_x A = \Lambda \tag{6.11}$$

where Λ is the diagonal matrix having as elements on its diagonal the respective eigenvalues λ_i, $i = 0, 1, \ldots, N-1$, of R_x. Furthermore, assuming R_x to be positive definite (Appendix B) the eigenvalues are positive. The resulting transform is known as the *Karhunen–Loève (KL)* transform, and it achieves our original goal of generating mutually uncorrelated features. The KL transform is of fundamental significance in pattern recognition and in a number of signal and image processing applications. Let us look at some of its important properties.

Mean square error approximation. From Eqs. (6.2) and (6.3) we have

$$\boldsymbol{x} = \sum_{i=0}^{N-1} y(i)\,\boldsymbol{a}_i \quad \text{and} \quad y(i) = \boldsymbol{a}_i^T \boldsymbol{x} \tag{6.12}$$

Let us now define a new vector in the m-dimensional subspace

$$\hat{\boldsymbol{x}} = \sum_{i=0}^{m-1} y(i)\,\boldsymbol{a}_i \tag{6.13}$$

where only m of the basis vectors are involved. Obviously, this is nothing but the projection of $\boldsymbol{x}$ onto the subspace spanned by the m (orthonormal) eigenvectors involved in the summation. If we try to approximate $\boldsymbol{x}$ by its projection $\hat{\boldsymbol{x}}$, the resulting mean square error is given by

$$E\big[||\boldsymbol{x} - \hat{\boldsymbol{x}}||^2\big] = E\left[\,||\sum_{i=m}^{N-1} y(i)\,\boldsymbol{a}_i||^2\right] \tag{6.14}$$

Our goal now is to choose the eigenvectors that result in the minimum MSE. From (6.14) and taking into account the orthonormality property of the eigenvectors, we have

$$E\left[\,||\sum_{i=m}^{N-1} y(i)\,\boldsymbol{a}_i||^2\right] = E\left[\sum_i \sum_j (y(i)\,\boldsymbol{a}_i^T)(y(j)\,\boldsymbol{a}_j)\right] \tag{6.15}$$

$$= \sum_{i=m}^{N-1} E[y^2(i)] = \sum_{i=m}^{N-1} \boldsymbol{a}_i^T E[\boldsymbol{x}\boldsymbol{x}^T]\,\boldsymbol{a}_i \tag{6.16}$$

Combining this with (6.14) and the eigenvector definition, we finally get

$$E\big[||\boldsymbol{x} - \hat{\boldsymbol{x}}||^2\big] = \sum_{i=m}^{N-1} \boldsymbol{a}_i^T \lambda_i \boldsymbol{a}_i = \sum_{i=m}^{N-1} \lambda_i \tag{6.17}$$

Thus, if we choose in (6.13) the eigenvectors corresponding to the *m largest* eigenvalues of the *correlation matrix*, then the error in (6.17) is *minimized, being the sum of the $N - m$ smallest eigenvalues. Furthermore, it can be shown (Problem 6.3) that this is also the minimum MSE, compared with any other approximation of* $\boldsymbol{x}$ *by an m-dimensional vector.*

A different form of the KL transform results if we compute A in terms of the eigenvectors of the covariance matrix. This transform diagonalizes the covariance matrix Σ_y,

$$\Sigma_y = A^T \Sigma_x A = \Lambda \tag{6.18}$$

In general, the two are different and coincide for zero mean random vectors. In practice this is usually the case, because if it is not true one can replace each vector by $\boldsymbol{x} - E[\boldsymbol{x}]$. Despite that, it is still interesting to point out a difference between the two variants of the KL transform. It can be shown (Problem 6.4) that in this case, the resulting orthonormal basis (Σ_x eigenvectors $\hat{\boldsymbol{a}}_i$) guarantees that the mean square error between $\boldsymbol{x}$ and its approximation given by

$$\hat{\boldsymbol{x}} = \sum_{i=0}^{m-1} y(i)\,\hat{\boldsymbol{a}}_i + \sum_{i=m}^{N-1} E[y(i)]\,\hat{\boldsymbol{a}}_i, \quad y(i) \equiv \hat{\boldsymbol{a}}_i^T \boldsymbol{x} \tag{6.19}$$

is minimum. In words, the last $N - m$ components are not random but are frozen to their respective mean values.

The optimality of the KL transform, with respect to the MSE approximation, leads to excellent information packing properties and offers us a tool to select the m dominant features out of N measurement samples. However, although this may be a good criterion, in many cases it does not necessarily lead to maximum class separability in the lower dimensional subspace. This is reasonable, since the dimensionality reduction is not optimized with respect to class separability, as was, for example, the case with the scattering matrix criteria of the previous chapter. This is demonstrated via the example of Figure 6.1. The feature vectors in the two classes follow the Gaussian distribution with the same covariance matrix. The ellipses show the curves of constant pdf values. We have computed the eigenvectors of the overall correlation matrix, and the resulting eigenvectors are shown in the figure. Eigenvector $\boldsymbol{a}_1$ is the one that corresponds to the largest eigenvalue. It does not take time for someone to realize that projection on $\boldsymbol{v}_1$ makes the two classes almost coincide. However, projecting on $\boldsymbol{a}_2$ keeps the two class separable.

Total variance. Let $E[\boldsymbol{x}]$ be zero. If this is not the case, the mean can always be subtracted. Let $\boldsymbol{y}$ be the KL transformed vector of $\boldsymbol{x}$. From the respective definitions we have that $\sigma_{y(i)}^2 \equiv E[y^2(i)] = \lambda_i$. That is, *the eigenvalues of the input*

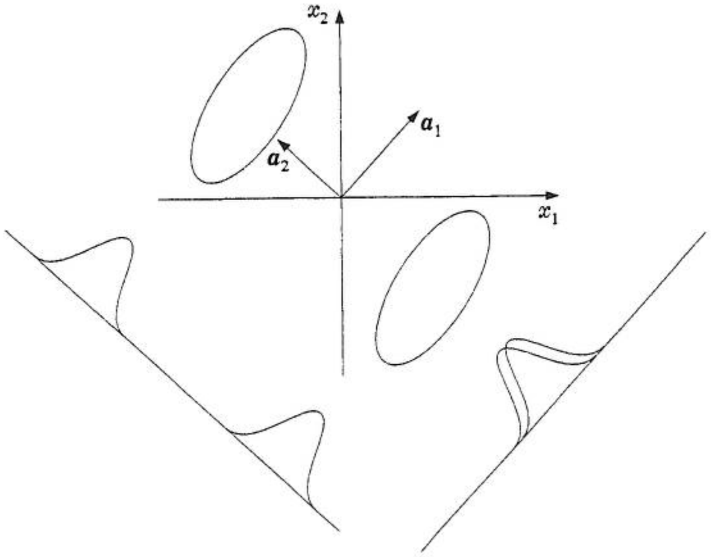

FIGURE 6.1: The KL transform is not always best for pattern recognition. In this example, projection on the eigenvector with the larger eigenvalue makes the two classes coincide. On the other hand, projection on the other eigenvector keeps the classes separated.

correlation matrix are equal to the variances of the transformed features. Thus, selecting those features, $y(i) \equiv \boldsymbol{a}_i^T \boldsymbol{x}$, corresponding to the m largest eigenvalues makes their sum variance $\sum_i \lambda_i$ maximum. In other words, the selected m features retain most of the total variance associated with the original random variables $x(i)$. Indeed, the latter is equal to the trace of R_x, which we know from linear algebra to be equal to the sum of the eigenvalues $\sum_{i=0}^{N-1} \lambda_i$ [Stra 80]. It can be shown that this is a more general property. That is, from all possible sets of m features, obtained via any orthogonal linear transformation on $\boldsymbol{x}$, the ones resulting from the KL transform have the largest sum variance (Problem 6.3).

Entropy. We know from Chapter 2 that the entropy of a process is defined as

$$h_y = -E[\ln p_y(\boldsymbol{y})]$$

and it is a measure of the randomness of the process. For a zero mean Gaussian multivariable m-dimensional process the entropy becomes

$$h_y = \frac{1}{2} E[\boldsymbol{y}^T R_y^{-1} \boldsymbol{y}] + \frac{1}{2} \ln |R_y| + \frac{m}{2} \ln(2\pi) \tag{6.20}$$

However,

$$E[y^T R_y^{-1} y] = E[\text{trace}\{y^T R_y^{-1} y\}] = E[\text{trace}\{R_y^{-1} y y^T\}] = \text{trace}(I) = m$$

and using the known property from linear algebra the determinant is

$$\ln|R_y| = \ln(\lambda_0 \lambda_1 \ldots \lambda_{m-1})$$

In words, selection of the m features that correspond to the m largest eigenvalues maximizes the entropy of the process. This is expected, because variance and randomness are directly related.

Remarks

- The concept of principal eigenvectors subspace has also been exploited as a classifier. First, the sample mean of the whole training set is subtracted from the feature vectors. For each class, ω_i, the correlation matrix R_i is estimated and the principal m eigenvectors (corresponding to the m largest eigenvalues) are computed. A matrix A_i is then formed using the respective eigenvectors as columns. An unknown feature vector $\boldsymbol{x}$ is then classified in the class ω_j for which

$$||A_j^T \boldsymbol{x}|| > ||A_i^T \boldsymbol{x}||, \quad \forall\, i \neq j \tag{6.21}$$

  that is, the class corresponding to the maximum norm subspace projection of $\boldsymbol{x}$ [Wata 73]. From Pythagoras' theorem this is equivalent to classifying a vector in its *nearest class subspace*. The decision surfaces are hyperplanes if all the subspaces have the same dimension or quadric surfaces in the more general case. *Subspace classification integrates the stages of feature generation/selection and classifier design.*

  If this approach results in a relatively high classification error, the performance may be improved by suitable modifications known as *learning subspace methods*. For example, one can iteratively rotate the subspaces to adjust the lengths of the projections of the training vectors. The basic idea is to increase the length of a projection in the subspace of the correct class and decrease it for the rest. Such techniques have been applied successfully in a number of applications, such as speech recognition, texture classification, and character recognition. The interested reader may consult, for example, [Oja 83, Koho 89, Prak 97].
- For the computation of the correlation matrix eigenvectors, a number of iterative schemes have been developed. The computation is performed working directly with the vectors, without having to estimate the corresponding correlation matrix, using neural network concepts [Oja 83, Diam 96].

Example

Compute t

The eig

we can alv

The KL tra

where $y(0)$,

6.4 THE

Given a ma

U and V s

X

where $\Lambda^{\frac{1}{2}}$ is

eigenvalues

other words

Y is diagon

where $\boldsymbol{u}_i$, $\boldsymbol{v}_i$

Example 6.2. The correlation matrix of a vector $\boldsymbol{x}$ is given by

$$R_x = \begin{bmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{bmatrix}$$

Compute the KL transform of the input vector.

The eigenvalues of R_x are $\lambda_0 = 0.1$, $\lambda_1 = \lambda_2 = 0.4$. Since the matrix R_x is symmetric, we can always construct orthonormal eigenvectors. For this case we have

$$\boldsymbol{a}_0 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \quad \boldsymbol{a}_1 = \frac{1}{\sqrt{6}}\begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{a}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

The KL transform is then given by

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \end{bmatrix} = \begin{bmatrix} 2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{3} & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \end{bmatrix}$$

where $y(0)$, $y(1)$ correspond to the two largest eigenvalues.

6.4 THE SINGULAR VALUE DECOMPOSITION

Given a matrix X of rank r, we will show that there exist $N \times N$ unitary matrices U and V so that

$$X = U \begin{bmatrix} \Lambda^{\frac{1}{2}} & O \\ O & 0 \end{bmatrix} V^H \quad \text{or} \quad Y \equiv \begin{bmatrix} \Lambda^{\frac{1}{2}} & O \\ O & 0 \end{bmatrix} = U^H X V \tag{6.22}$$

where $\Lambda^{\frac{1}{2}}$ is the $r \times r$ diagonal matrix with elements $\sqrt{\lambda_i}$, and λ_i are the r nonzero eigenvalues of the associated matrix $X^H X$. O denotes a zero element matrix. *In other words, there exist unitary matrices U and V so that the transformed matrix Y is diagonal.* From (6.22) it is readily shown that

$$X = \sum_{i=0}^{r-1} \sqrt{\lambda_i}\, \boldsymbol{u}_i \boldsymbol{v}_i^H \tag{6.23}$$

where $\boldsymbol{u}_i$, $\boldsymbol{v}_i$ are the first r columns of U and V, respectively. More precisely, APPENDIX **B**

LINEAR ALGEBRA BASICS

POSITIVE DEFINITE AND SYMMETRIC MATRICES

- An $l \times l$ real matrix A is called *positive definite* if for *every* nonzero vector x the following is true:

$$x^T Ax > 0 \tag{B.1}$$

  If equality with zero is allowed, A is called *nonnegative or positive semidefinite*.

- It is easy to show that all eigenvalues of such a matrix are positive. Indeed, let λ_i be one eigenvalue and v_i the corresponding unit norm eigenvector $(v_i^T v_i = 1)$. Then by the respective definitions

$$Av_i = \lambda_i v_i \quad \text{or} \tag{B.2}$$

$$0 < v_i^T A v_i = \lambda_i \tag{B.3}$$

  Since the determinant of a matrix is equal to the product of its eigenvalues, we conclude that the determinant of a positive definite matrix is also positive.

- Let A be an $l \times l$ symmetric matrix, $A^T = A$. Then the eigenvectors corresponding to distinct eigenvalues are orthogonal. Indeed, let $\lambda_i \neq \lambda_j$ be two such eigenvalues. From the definitions we have

$$Av_i = \lambda_i v_i \tag{B.4}$$

$$Av_j = \lambda_j v_j \tag{B.5}$$

  Multiplying (B.4) on the left by v_j^T and the transpose of (B.5) on the right by v_i, we obtain

$$v_j^T A v_i - v_j^T A v_i = 0 = (\lambda_i - \lambda_j) v_j^T v_i \tag{B.6}$$

  Thus, $v_j^T v_i = 0$. Furthermore, it can be shown that even if the eigenvalues are not distinct, we can still find a set of orthogonal eigenvectors. The same is true for Hermitian matrices, in case we deal with more general complex-valued matrices.

- Based on this, it is now straightforward to show that a symmetric matrix A can be diagonalized by the similarity transformation

$$\Phi^T A \Phi = \Lambda \tag{B.7}$$

where matrix Φ has as its columns the unit eigenvectors ($v_i^T v_i = 1$) of A, that is,

$$\Phi = [v_1, v_2, \ldots, v_l] \tag{B.8}$$

and Λ is the diagonal matrix with elements the corresponding eigenvalues of A. From the orthonormality of the eigenvectors it is obvious that $\Phi^T \Phi = I$, that is, Φ is a unitary matrix, $\Phi^T = \Phi^{-1}$. The proof is similar for Hermitian complex matrices as well.

CORRELATION MATRIX DIAGONALIZATION

Let x be a random vector in the l-dimensional space. Its correlation matrix is defined as $R = E[xx^T]$. Matrix R is readily seen to be positive semidefinite. For our purposes we will assume that it is positive definite, thus invertible. Moreover, it is symmetric, and hence it can always be diagonalized

$$\Phi^T R \Phi = \Lambda \tag{B.9}$$

where Φ is the matrix consisting of the (orthogonal) eigenvectors and Λ the diagonal matrix with the corresponding eigenvalues on its diagonal. Thus, we can always transform x into another random vector whose elements are uncorrelated. Indeed

$$x_1 \equiv \Phi^T x \tag{B.10}$$

Then the new correlation matrix is $R_1 = \Phi^T R \Phi = \Lambda$. Furthermore, if $\Lambda^{1/2}$ is the diagonal matrix whose elements are the square roots of the eigenvalues of R ($\Lambda^{1/2}\Lambda^{1/2} = \Lambda$), then it is readily shown that the transformed random vector

$$x_1 \equiv \Lambda^{-1/2}\Phi^T x \tag{B.11}$$

has uncorrelated elements with unit variance. $\Lambda^{-1/2}$ denotes the inverse of $\Lambda^{1/2}$. It is now easy to see that if the correlation matrix of a random vector is the identity matrix I, then this is invariant under any unitary transformation $A^T x$, $A^T A = I$.

That is, the
byproduct (

Lemma. L
R_x, R_y, res
matrices si

Proof.
mation

generates t
tively. Nov
vectors gen

have correl
elements tl
vectors by

diagonalize
All these ar

That is, the transformed variables are also uncorrelated with unit variance. A useful byproduct of this is the following lemma.

Lemma. Let $\boldsymbol{x}$, $\boldsymbol{y}$ be two zero mean random vectors with correlation matrices R_x, R_y, respectively. Then there is a linear transformation that diagonalizes both matrices simultaneously.

Proof. Let Φ be the eigenvector matrix diagonalizing R_x. Then the transformation

$$\boldsymbol{x}_1 \equiv \Lambda^{-1/2}\Phi^T \boldsymbol{x} \tag{B.12}$$

$$\boldsymbol{y}_1 \equiv \Lambda^{-1/2}\Phi^T \boldsymbol{y} \tag{B.13}$$

generates two new random vectors with correlation matrices $R_x^1 = I$, R_y^1, respectively. Now let Ψ be the eigenvector matrix diagonalizing R_y^1. Then the random vectors generated by the unitary transformation ($\Psi^T\Psi = I$)

$$\boldsymbol{x}_2 \equiv \Psi^T \boldsymbol{x}_1 \tag{B.14}$$

$$\boldsymbol{y}_2 \equiv \Psi^T \boldsymbol{y}_1 \tag{B.15}$$

have correlation matrices $R_x^2 = I$, $R_y^2 = D$, where D is the diagonal matrix with elements the eigenvalues of R_y^1. Thus, the linear transformation of the original vectors by the matrix

$$A^T = \Psi^T \Lambda^{-1/2}\Phi^T \tag{B.16}$$

diagonalizes both correlation matrices simultaneously (one to an identity matrix). All these are obviously valid for covariance matrices as well.