Principal Component Analysis and Karhunen-Loeve Transform

Each of the orthogonal transforms discussed previously (Fourier transform, cosine transform, Walsh-Hadamard transform, Haar transform, etc.) is associated with an orthogonal (or unitary) matrix $\mathbf{A}$ that satisfies $\mathbf{A}^{-1} = \mathbf{A}^T$ or $\mathbf{A}^T\mathbf{A} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. This orthogonal matrix can be expressed in terms of its N column vectors:

$$\mathbf{A} = [\mathbf{a}_0, \cdots, \mathbf{a}_{N-1}], \quad \text{or} \quad \mathbf{A}^T = \begin{bmatrix} \mathbf{a}_0^T \\ \cdots \\ \mathbf{a}_{N-1}^T \end{bmatrix}$$

These column vectors are orthogonal and normalized (orthonormal):

$$(\mathbf{a}_i, \mathbf{a}_j) = \mathbf{a}_i^T\mathbf{a}_j = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

and can be used as the basis vectors that span the N-dimensional vector space.

As we have seen before, the transform of any given discrete signal, represented by a vector in the N-dimensional space $\mathbf{x} = [x_0, \cdots, x_{N-1}]^T$, can be carried out as a matrix multiplication:

$$\mathbf{y} = \begin{bmatrix} y_0 \\ \cdots \\ y_{N-1} \end{bmatrix} = \mathbf{A}^T\mathbf{x} = \begin{bmatrix} \mathbf{a}_0^T \\ \cdots \\ \mathbf{a}_{N-1}^T \end{bmatrix} \mathbf{x}$$

where the ith component $y_i = \mathbf{a}_i^T\mathbf{x}$ is the projection of the signal vector $\mathbf{x}$ onto the ith basis vector $\mathbf{a}_i$. Left multiplying $\mathbf{A}$ on both sides of the equation above, we get the inverse transform:

$$\mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{A}^T\mathbf{x} = \mathbf{A}\mathbf{A}^{-1}\mathbf{x} = \mathbf{x}$$

which can be rewritten as:

$$\mathbf{x} = \mathbf{A}\mathbf{y} = [\mathbf{a}_0, \cdots, \mathbf{a}_{N-1}] \begin{bmatrix} y_0 \\ \cdots \\ y_{N-1} \end{bmatrix} = \sum_{i=0}^{N-1} y_i\mathbf{a}_i$$

We see that a given signal vector $\mathbf{x}$ is expressed by the inverse transform as a linear combination of the orthogonal basis vectors $\mathbf{a}_i$, $(i = 0, \cdots, N -$

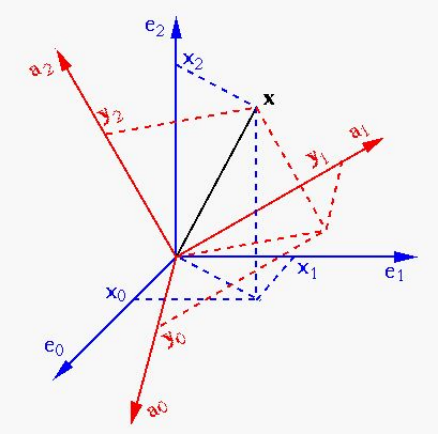


Figure 1: Rotation of coordinate system

1). Geometrically, the transform $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ represents a rotation of the coordinate system of the N-dimensional space. The rotation is specified by the orthogonal transform matrix $\mathbf{A}$. As a special case, the orthogonal transform matrix could be the identity matrix $\mathbf{A} = \mathbf{I} = [\mathbf{e}_0, \cdots, \mathbf{e}_{N-1}]$ where $\mathbf{e}_i = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ is the ith basis vector of the space with the ith component being 1 and all others zeor. The transform associated with this identity matrix is $\mathbf{x} = \mathbf{I}\mathbf{x}$, i.e., the original signal vector:

$$\mathbf{x} = \sum_{i=0}^{N-1} x_i \mathbf{e}_i = \begin{bmatrix} x_0 \\ 0 \\ \cdots \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ \cdots \\ 0 \\ x_{N-1} \end{bmatrix} = \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_{N-1} \end{bmatrix}$$

After seeing quite a few different transform methods and their applications discussed in the previous chapters, we may want to ask some more general questions regarding the common natures of these transforms. Why do we want to carry out such transforms to start with? Do different transforms share some intrinsic properties and essential characteristics in common? If an orthogonal transform is nothing more than a certain rotation in the N-dimensional vector space, what can be achieved by such a rotation? And, finally, is there an optimal rotation among all possible transform rotations? We will address such questions in the following discussion for the Karhunen-Loeve Transform (KLT) and the associated principal component analysis (PCA).

Signal Correlation

First let us quickly review some basic concepts of multivariate random variables. A time signal $x(t)$ can be considered as a random process and its samples x_m $(m = 0, \cdots, N-1)$ form a random vector represented by $\mathbf{x}$:

$$\mathbf{x} = [x_0, \cdots, x_{N-1}]^T$$

with the *mean vector*

$$\mathbf{m}_x \stackrel{\triangle}{=} E(\mathbf{x}) = [E(x_0), \cdots, E(x_{N-1})]^T = [\mu_0, \cdots, \mu_{N-1}]^T$$

and the *covariance matrix*

$$\mathbf{\Sigma}_x \stackrel{\triangle}{=} E[(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^T] = E(\mathbf{x}\mathbf{x}^T) - \mathbf{m}_x \mathbf{m_x}^T = \begin{bmatrix} .. & .. & .. \\ .. & \sigma_{ij}^2 & .. \\ .. & .. & .. \end{bmatrix}$$

where $\mu_i = E(x_i)$ is the expectation or mean of x_i, and $\sigma_{ij}^2 \stackrel{\triangle}{=} E(x_i - \mu_i)(x_j - \mu_j) = E(x_i x_j) - \mu_i \mu_j$ is the covariance of two random variables x_i and x_j. When $i = j$, σ_{ij}^2 becomes the variance of x_i, $\sigma_i^2 \stackrel{\triangle}{=} E(x_i - \mu_i)^2 = E(x_i^2) - \mu_i^2$.

Moreover, the *correlation coefficient* between two variables x_i and x_j is defined as

$$\rho_{ij} = \frac{\sigma_{ij}^2}{\sigma_i \sigma_j}$$

which can be considered as the normalized version of the covariances so that when $x_i = x_j$, $\rho_{ij} = 1$.

After a certain orthogonal transform of a given random vector $\mathbf{x}$, the resulting vector $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ is still random with the following mean vector and covariance matrix:

$$\mathbf{m}_y = E(\mathbf{y}) = E(\mathbf{A}^T\mathbf{x}) = \mathbf{A}^T E(\mathbf{x}) = \mathbf{A}^T \mathbf{m}_x$$

$$\begin{aligned} \mathbf{\Sigma}_y &= E[(\mathbf{y} - \mathbf{m}_y)(\mathbf{y} - \mathbf{m}_y)^T] = E[\mathbf{A}^T(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^T\mathbf{A}] \\ &= \mathbf{A}^T E[(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^T]\mathbf{A} = \mathbf{A}^T \mathbf{\Sigma}_x \mathbf{A} \end{aligned}$$

The mean and the variance of a component x_i of a random vector $\mathbf{x}$ can be estimated by averaging the outcomes of the random experiment concerning the variable repeated K times:

$$\hat{\mu}_i = \frac{1}{K}\sum_{k=1}^{K} x_i^{(k)}, \qquad \hat{\sigma}_i^2 = \frac{1}{K}\sum_{k=1}^{K}(x_i^{(k)} - \hat{\mu}_i)^2 = \frac{1}{K}\sum_{k=1}^{K}(x_i^{(k)})^2 - \hat{\mu}_i^2$$

and the covariance between two variables x_i and x_j can be estimated as

$$\hat{\sigma}_{ij}^2 = \frac{1}{K}\sum_{k=1}^{K}(x_i^{(k)} - \hat{\mu}_i)(x_j^{(k)} - \hat{\mu}_j) = \frac{1}{K}\sum_{k=1}^{K} x_j^{(k)} x_j^{(k)} - \hat{\mu}_i\hat{\mu}_j$$

The meaning of the covariance σ_{ij}^2 between two random variables x_i and x_j can be illustrated by the following examples.

1. Assume an experiment concerning x_i and x_j is repeated $K = 3$ times with the following outcomes:

Experiment	1st	2nd	3rd
$x_i^{(k)}$	1	2	3
$x_j^{(k)}$	1	2	3

   The means and covariances of x_i and x_j can be estimated as

$$\hat{\mu}_i = \frac{1}{K}\sum_{k=1}^{K} x_i^{(k)} = \hat{\mu}_j = \frac{1}{K}\sum_{k=1}^{K} x_j^{(k)} = 2$$

$$\hat{\sigma}_{ij}^2 = \frac{1}{K}\sum_{k=1}^{K} x_i^{(k)} x_j^{(k)} - \hat{\mu}_i\hat{\mu}_j = (1\times 1 + 2\times 2 + 3\times 3)/3 - 2\times 2 = 0.667$$

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}^2}{\hat{\sigma}_i\hat{\sigma}_j} = 1$$

   In this case x_i and x_j are maximally correlated.

2. Assume the outcomes of the 3 experiments are

Experiment	1st	2nd	3rd
$x_i^{(k)}$	1	2	3
$x_j^{(k)}$	3	2	1

   We have $\hat{\mu}_i = \hat{\mu}_j = 2$, $\hat{\sigma}_{ij}^2 = -0.667$, and $\hat{\rho}_{ij} = -1$, indicating that the two variables are negatively or inversely correlated.

3. Assume the outcomes are:

Experiment	1st	2nd	3rd
x_i	1	2	3
x_j	2	2	2

We have $\hat{\mu}_i = \hat{\mu}_j = 2$, $\hat{\sigma}_{ij}^2 = 0$, and $\hat{\rho}_{ij} = 0$, indicating that the two variables are totally uncorrelated.

4. Assume the outcomes are:

Experiment	1st	2nd	3rd
$x_i^{(k)}$	2	2	2
$x_j^{(k)}$	1	2	3

We have $\hat{\mu}_i = \hat{\mu}_j = 2$, $\hat{\sigma}_{ij}^2 = 0$, and $\hat{\rho}_{ij} = 0$, indicating that the two variables are totally uncorrelated.

5. Combine the outcomes of the two previous cases ($K = 5$):

Experiment	1st	2nd	3rd	4th	5th
$x_i^{(k)}$	1	2	2	2	3
$x_j^{(k)}$	2	1	2	3	2

We still have $\hat{\mu}_i = \hat{\mu}_j = 2$, $\hat{\sigma}_{ij}^2 = 0$ and $\hat{\rho}_{ij} = 0$, indicating that the two variables are totally uncorrelated.

From the above examples we see that the covariance σ_{ij}^2 represents how much the two variables x_i and x_j are correlated. If $\sigma_{ij}^2 > 0$, they are positively correlated, $\sigma_{ij}^2 < 0$ they are negatively correlated, and if $\sigma_{ij}^2 = 0$, they are uncorrelated at all.

Karhunen-Loeve Transform (KLT)

Now we consider the *Karhunen-Loeve Transform (KLT)* (also known as *Hotelling Transform* and *Eigenvector Transform*), and the associated *Principal Component Analysis (PCA)*, which is widely used for data analysis in mnay different fields.

Let ϕ_k be the eigenvector corresponding to the kth eigenvalue λ_k of the covariance matrix $\mathbf{\Sigma}_x$, i.e.,

$$\mathbf{\Sigma}_x \phi_k = \lambda_k \phi_k \qquad (k = 0, \cdots, N-1)$$

or in matrix form:

$$\begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \sigma_{ij} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} \\ \phi_k \\ \\ \end{bmatrix} = \lambda_k \begin{bmatrix} \\ \phi_k \\ \\ \end{bmatrix} \qquad (k = 0, \cdots, N-1)$$

As the covariance matrix $\mathbf{\Sigma}_x = \mathbf{\Sigma}_x^T$ is positive definte and symmetric (Hermitian if $\mathbf{x}$ is complex), all of its eigenvalues $\lambda_i > 0$ are positive and its eigenvectors ϕ_i's are orthogonal:

$$(\phi_i, \phi_j) = \phi_i^T, \phi_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

These N orthonormal eigenvectors can be used as the basis vectors of the N-dimensional vector space, and they can be used to construct an $N \times N$ orthogonal matrix $\mathbf{\Phi}$:

$$\mathbf{\Phi} \triangleq [\phi_0, \cdots, \phi_{N-1}]$$

that satisfies

$$\mathbf{\Phi}^T \mathbf{\Phi} = \mathbf{I}, \quad \text{i.e.,} \quad \mathbf{\Phi}^{-1} = \mathbf{\Phi}^T$$

The N eigenequations above can be combined to be expressed as:

$$\begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \sigma_{ij} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} [\phi_0, \cdots, \phi_{N-1}] = [\phi_0, \cdots, \phi_{N-1}] \begin{bmatrix} \lambda_0 & \cdots & \cdots \\ \cdots & \lambda_i & \cdots \\ \cdots & \cdots & \lambda_{N-1} \end{bmatrix}$$

or in matrix form:

$$\mathbf{\Sigma}_x \mathbf{\Phi} = \mathbf{\Phi} \mathbf{\Lambda}$$

where $\mathbf{\Lambda}$ is a diagonal matrix $\mathbf{\Lambda} = diag(\lambda_0, \cdots, \lambda_{N-1})$. Left multiplying $\mathbf{\Phi}^T = \mathbf{\Phi}^{-1}$ on both sides, the covariance matrix $\mathbf{\Sigma}_x$ can be diagonalized:

$$\mathbf{\Phi}^T \mathbf{\Sigma}_x \mathbf{\Phi} = \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{\Lambda} = \mathbf{\Phi}^{-1} \mathbf{\Phi} \mathbf{\Lambda} = \mathbf{\Lambda}$$

We can now define the Karhunen-Loeve Transform of a given signal vector $\mathbf{x}$ as

$$\mathbf{y} = \begin{bmatrix} y_0 \\ \cdots \\ y_{N-1} \end{bmatrix} = \mathbf{\Phi}^T \mathbf{x} = \begin{bmatrix} \phi_0^T \\ \cdots \\ \phi_{N-1}^T \end{bmatrix} \mathbf{x}$$

where the ith component y_i of the transform vector is the projection of $\mathbf{x}$ onto the ith basis vector $\phi_\mathbf{i}$:

$$y_i = (\phi_i, \mathbf{x}) = \phi_i^T \mathbf{x}$$

Left multiplying $\mathbf{\Phi} = (\mathbf{\Phi}^T)^{-1}$ on both sides of the transform equation $\mathbf{y} = \mathbf{\Phi}^T\mathbf{x}$, we get the inverse transform:

$$\mathbf{x} = \mathbf{\Phi}\mathbf{y} = [\phi_0, \cdots, \phi_{N-1}] \begin{bmatrix} y_0 \\ \cdots \\ y_{N-1} \end{bmatrix} = \sum_{i=0}^{N-1} y_i \phi_i$$

We see that by KLT transform, the signal vector $\mathbf{x}$ is expressed in the N-dimensional vector space spanned by the N eigenvectors ϕ_i $(i = 0, \cdots, N-1)$. As shown below, representing a given signal vector $\mathbf{x}$ by this particular set of basis vectors is most advantageous in two specific aspects.

KLT Completely Decorrelates the Signal

Compared with all possible orthogonal transforms, KLT is optimal in terms of the following two properties:

- KLT completely decorrelates the signal
- KLT maximally compacts the energy (information) contained in the signal.

As seen before, most other orthogonal transforms can decorrelate the signal and compact the signal energy into a small number of components to different extents, the KLT transform does these optimally.

The first property is simply due to the definition of the KLT transform, i.e., the covariance matrix is diagonalized after thr transform. The second property is due to the fact that KLT redistributes the energy among the N components in such a way that the energy is maximally compacted into a small number of components of $\mathbf{y} = \mathbf{\Phi}^T\mathbf{x}$.

To see the first property, simply recall that the covariance matrix of the signal after a transform $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ becomes:

$$\mathbf{\Sigma}_y = E[(\mathbf{y} - \mathbf{m}_y)(\mathbf{y} - \mathbf{m}_y)^T] = \mathbf{A}^T E[(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^T]\mathbf{A} = \mathbf{A}^T \mathbf{\Sigma}_x \mathbf{A}$$

In this particular case, $\mathbf{y} = \mathbf{\Phi}^T\mathbf{x}$ and we have

$$\mathbf{\Sigma}_y = \mathbf{\Phi}^T\mathbf{\Sigma}_x\mathbf{\Phi} = \mathbf{\Lambda}$$

or in matrix form:

$$\mathbf{\Sigma}_y = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \sigma_{ij} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} = \mathbf{\Phi}^T\mathbf{\Sigma}_x\mathbf{\Phi} = \mathbf{\Lambda} = \begin{bmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \\ 0 & \cdots & \lambda_{N-1} & \end{bmatrix} = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{N-1}^2 \end{bmatrix}$$

We can make two observations:

- After KLT, the covariance matrix of the signal $\mathbf{y} = \mathbf{\Phi}^T\mathbf{x}$ is diagonalized, i.e., the covariance σ_{ij} between any two different components y_i and y_j is always zero. In other words, the signal is completely decorrelated.
- The variance of y_i is the ith eigenvalue of the covariance matrix of $\mathbf{x}$, i.e., $\sigma_i^2 = \lambda_i$.

KLT Optimally Compacts the Energy

Now we show that KLT redistributes the energy contained in the signal so that it is maximally compacted into a small number of components after the transform. Let $\mathbf{A} = [\mathbf{a}_0, \cdots, \mathbf{a}_{N-1}]$ be an arbitrary orthogonal matrix satisfying $\mathbf{A}^{-1} = \mathbf{A}^T$. An orthogonal transform of a given signal vector $\mathbf{x}$ can be defined as

$$\mathbf{y} = \begin{bmatrix} y_0 \\ \cdots \\ y_{N-1} \end{bmatrix} = \mathbf{A}^T\mathbf{x} = \begin{bmatrix} \mathbf{a}_0^T \\ \cdots \\ \mathbf{a}_{N-1}^T \end{bmatrix} \mathbf{x}$$

where the ith component of $\mathbf{y}$ is $y_i = \mathbf{a}_i^T\,\mathbf{x}$. The inverse transform is:

$$\mathbf{x} = \mathbf{A}\mathbf{y} = [\mathbf{a}_0, \cdots, \mathbf{a}_{N-1}] \begin{bmatrix} y_0 \\ \cdots \\ y_{N-1} \end{bmatrix} = \sum_{i=0}^{N-1} y_i\mathbf{a}_i$$

Consider the variances of the signal components before and after the KLT transform:

$$\sigma_{x_i}^2 = E[(x_i - \mu_{x_i})^2] \triangleq E(e_{x_i}), \quad \text{and} \quad \sigma_{y_i}^2 = E[(y_i - \mu_{y_i})^2] \triangleq E(e_{y_i})$$

where $e_{x_i} \triangleq (x_i - \mu_{x_i})^2$ can be considered as the dynamic energy or information contained in the ith component of the signal, and the trace of the covarance matrix $tr\mathbf{\Sigma}_x$ represents the total energy or information contained in the signal:

$$tr\mathbf{\Sigma}_x = \sum_{i=0}^{N-1} \sigma_{x_i}^2 = \sum_{i=0}^{N-1} E[(x_i - \mu_{x_i})^2] = \sum_{i=0}^{N-1} E(e_{x_i})$$

Due to the commutativity of trace: $tr(\mathbf{AB}) = tr(\mathbf{BA})$, we have:

$$tr\mathbf{\Sigma}_\mathbf{y} = tr(\mathbf{A}^T\mathbf{\Sigma}_x\mathbf{A}) = tr(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}_x)) = tr\mathbf{\Sigma}_x$$

We see that the total energy or information of the signal is conserved after the KLT transform. However, as shown below, the energy redistribution among the N signal components is drasticallychanged in such a way that the energy is optimally compacted into a small number of components.

Define the energy contained in the first $M < N$ components after the transform $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ as

$$S_M(\mathbf{A}) \triangleq \sum_{i=0}^{M-1} E[(y_i - \mu_{y_i})^2] = \sum_{i=0}^{M-1} \sigma_{y_i}^2 = \sum_{i=0}^{M-1} E(e_{y_i})$$

Since the total energy $S_N(\mathbf{A}) = \sum_{i=0}^{N-1} E(e_{x_i})$ is conserved, $S_M(\mathbf{A})$ represents the percentage of energy contained in the first M components. We now show that $S_M(\mathbf{A})$ is maximized if and only if the transform matrix $\mathbf{A}$ is $\mathbf{\Phi}$, the transform matrix of the KLT:

$$S_M(\mathbf{\Phi}) \geq S_M(\mathbf{A})$$

i.e., the KLT optimally compacts energy into a small number of components of the signal. Consider

$$\begin{aligned} S_M(\mathbf{A}) &= \sum_{i=0}^{M-1} E(y_i - \mu_{y_i})^2 = \sum_{i=0}^{M-1} E[\mathbf{a}_i^T(\mathbf{x} - \mathbf{m}_{x_i})\ \mathbf{a}_i^T(\mathbf{x} - \mathbf{m}_{x_i})] \\ &= \sum_{i=0}^{M-1} \mathbf{a}_i^T E[(\mathbf{x} - \mathbf{m}_{x_i})(\mathbf{x} - \mathbf{m}_{x_i})^T]\mathbf{a}_i^T = \sum_{i=0}^{M-1} \mathbf{a}_i^T\mathbf{\Sigma}_x\mathbf{a}_i \end{aligned}$$

The optimal transform matrix $\mathbf{A}$ should therefore satisfy

$$\begin{cases} S_M(\mathbf{A}) = \sum_{i=0}^{M-1} \mathbf{a}_i^T\mathbf{\Sigma}_x\mathbf{a}_i \rightarrow max \\ \text{subject to:} \quad \mathbf{a}_j^T\mathbf{a}_j = 1 \quad (j = 0, \cdots, M-1) \end{cases}$$

The constraint $\mathbf{a}_j^T\mathbf{a}_j = 1$ is to guarantee that the column vectors of $\mathbf{A}$ are orthogonal and normalized. This constrained optimization problem can be solved using Lagrange multiplier method by letting the following partial derivative be zero:

$$\frac{\partial}{\partial \mathbf{a}_i}[S_M(\mathbf{A}) - \sum_{j=0}^{M-1}\lambda_j(\mathbf{a}_j^T\mathbf{a}_j - 1)] = \frac{\partial}{\partial \mathbf{a}_i}[\sum_{j=0}^{M-1}(\mathbf{a}_j^T\mathbf{\Sigma}_x\mathbf{a}_j - \lambda_j\mathbf{a}_j^T\mathbf{a}_j + \lambda_j)]$$

$$= \frac{\partial}{\partial \mathbf{a}_i}[\mathbf{a}_i^T\mathbf{\Sigma}_x\mathbf{a}_i - \lambda_i\mathbf{a}_i^T\mathbf{a}_i] \stackrel{*}{=} 2\mathbf{\Sigma}_x\mathbf{a}_i - 2\lambda_i\mathbf{a}_i = 0$$

(The equal sign with a * is due to the derivative of a scalar function with respect to its vector argument, see appendix A). We see that the column vectors of $\mathbf{A}$ must be the eigenvectors of $\mathbf{\Sigma}_x$:

$$\mathbf{\Sigma}_x\mathbf{a}_i = \lambda_i\mathbf{a}_i \qquad (i = 0, \cdots, M-1)$$

i.e., the optimal transform matrix is

$$\mathbf{A} = [\mathbf{a}_0, \cdots, \mathbf{a}_{N-1}] = \mathbf{\Phi} = [\phi_0, \cdots, \phi_{N-1}]$$

where ϕ_i's are the orthogonal eigenvectors of $\mathbf{\Sigma}_x$ corresponding to eigenvalues λ_i $(i = 0, \cdots, N-1)$:

$$\mathbf{\Sigma}_x\phi_i = \lambda_i\phi_i, \qquad \text{i.e.} \qquad \phi^T\mathbf{\Sigma}_x\phi_i = \lambda_i\phi_i^T\phi_i = \lambda_i$$

Thus we have proved that the optimal transform is indeed the KLT transform, and

$$S_M(\mathbf{\Phi}) = \sum_{i=0}^{M-1}\phi_i^T\mathbf{\Sigma}_x\phi_i = \sum_{i=0}^{M-1}\lambda_i$$

where the ith eigenvalue λ_i of $\mathbf{\Sigma}_x$ is also the average energy contained in the ith component of the signal. If we choose the M $\phi_i's$ that correspond to the M largest eigenvalues of $\mathbf{\Sigma}_x$: $\lambda_0 \geq \cdots \lambda_{M-1} \cdots \geq \lambda_{N-1}$, then $S_M(\mathbf{\Phi})$ is maximized.

Due to its properties of signal decorrelation and energy compaction, KLT can be used to reduce the dimensionality of the data set for data compression. The signal components after the KLT are called the *principal components*, and the data analysis method based on the KLT transform is called *principal component analysis (PCA)*, which is widely used in a large variety of fields.

In summary, the PCA can be carried out in the following steps:

1. Estimate the mean vector $\mathbf{m}_x$ and the covariance matrix $\mathbf{\Sigma}_x$ of the signal vectors $\mathbf{x}$.

2. Find $\mathbf{\Sigma}_x$'s eigenvalues λ_i and associated eigenvector ϕ_i ($i = 0, \cdots, N-1$). Sort the eigenvalues in descending order, together with their corresponding eigenvectors.

3. Choose a lowered dimensionality $M < N$ so that the percentage of energy contained $\sum_{i=0}^{M-1} \lambda_i / \sum_{i=0}^{N-1} \lambda_i$ is no less than a given threshold (e.g., 95%).

4. Construct an N by M transform matrix composed of M eigenvectors corresponding to the M largest eigenvalues of $\mathbf{\Sigma}_x$:

$$\mathbf{\Phi}_M = [\phi_0, \cdots, \phi_{M-1}]_{N \times M}$$

and carry out KLT based on $\mathbf{\Phi}_M$:

$$\mathbf{y} = \mathbf{\Phi}_M^T \mathbf{x}$$

or

$$\begin{bmatrix} y_0 \\ \cdots \\ y_{M-1} \end{bmatrix}_{M \times 1} = \begin{bmatrix} \phi_0^T \\ \cdots \\ \phi_{M-1}^T \end{bmatrix}_{M \times N} \begin{bmatrix} x_0 \\ \cdots \\ x_{N-1} \end{bmatrix}_{N \times 1}$$

As the dimensionality M of $\mathbf{y}$ is less than the dimensionality N of $\mathbf{x}$, data compression is achieved for storage and/or transmission. This is a lossy compression with the error representing the percentage of information lost: $\sum_{i=M}^{N-1} \lambda_i / \sum_{i=0}^{N-1} \lambda_i$. But as these λ_i's are the smallest eigenvalues, the error is minimum (e.g., 5%).

5. Carry out inverse KLT for reconstruction:

$$\mathbf{x} = \mathbf{\Phi}_M \mathbf{y}$$

or

$$\begin{bmatrix} x_0 \\ \cdots \\ x_{N-1} \end{bmatrix}_{N \times 1} = \begin{bmatrix} \phi_0 \cdots \phi_{M-1} \end{bmatrix}_{N \times M} \begin{bmatrix} y_0 \\ \cdots \\ y_{M-1} \end{bmatrix}_{M \times 1}$$

Although KLT is optimal, other transforms are still widely used for two reasons. First, the KLT transform depends on the specific data set being processed, as the transform matrix is composed of the eiganvectors of the covariance matrix $\mathbf{\Sigma}_x$ of the signal vector $\mathbf{x}$, which can be estimated only if sufficient amount of data is available. Second, the computational cost for KLT is much higher than other transforms. This is because there does not exist any fast algorithm for the KLT transform, and the computation complexity of the KLT transform is $O(N^2)$, instead of $O(N \log_2 N)$ for most of other transforms discussed before. Moreover, in order to obtain the KLT transform matrix **Phi**, we also need to estimate the covariance matrix $\mathbf{\Sigma}_x$ from the available data, and to solve its eigenvalue problem. These tasks will further increase the computational complexity significantly. For these reasons, in many applications, the DCT transform (or some other transform) is the more preferable method.

Geometric Interpretation of KLT

Assume the N random variables in a signal vector $\mathbf{x} = [x_0, \cdots, x_{N-1}]^T$ have a normal joint probability density function:

$$p(x_0, \cdots, x_{N-1}) = p(\mathbf{x}) = N(\mathbf{x}, \mathbf{m}_x, \mathbf{\Sigma}_x) = \frac{1}{(2\pi)^{N/2} \left|\mathbf{\Sigma}_x\right|^{1/2}} exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_x)^T \mathbf{\Sigma}_x^{-1}(\mathbf{x}-\mathbf{m}_x)]$$

In particular, when $N = 1$, $\mathbf{\Sigma}_x$ and $\mathbf{m}_x$ become σ_x and μ_x, respectively, and the density function becomes the familiar single variable normal distribution:

$$p(x) = N(x, \mu_x, \sigma_x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} exp[-\frac{(x-\mu_x)^2}{2\sigma_x^2}]$$

The shape of this normal distribution in the N-dimensional space can be found by considering the iso-value hyper-surface in the space determined by equation

$$N(\mathbf{x}, \mathbf{m}_x, \mathbf{\Sigma}_x) = c_0$$

where c_0 is a constant. Or, equivalently, this equation can be written as

$$(\mathbf{x} - \mathbf{m}_x)^T \mathbf{\Sigma}_x^{-1} (\mathbf{x} - \mathbf{m}_x) = c_1$$

where c_1 is another constant related to c_0, $\mathbf{m}_x$ and $\mathbf{\Sigma}_x$. In particular, with $N = 2$ variables x_0 and x_1, we have

$$\begin{aligned}(\mathbf{x}-\mathbf{m}_x)^T\mathbf{\Sigma}_x^{-1}(\mathbf{x}-\mathbf{m}_x) &= [x_0-\mu_{x_0}, x_1-\mu_{x_1}]\begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix}\begin{bmatrix} x_0-\mu_{x_0} \\ x_1-\mu_{x_1} \end{bmatrix} \\ &= a(x_0-\mu_{x_0})^2 + b(x_0-\mu_{x_0})(x_1-\mu_{x_1}) + c(x_1-\mu_{x_1})^2 = c_1\end{aligned}$$

Here we have assumed

$$\mathbf{\Sigma}_x^{-1} = \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix}$$

As $\mathbf{\Sigma}_x^{-1}$ and $\mathbf{\Sigma}_x$ are positive definite, i.e.,

$$\left|\mathbf{\Sigma}_x^{-1}\right| = ac - b^2/4 > 0$$

the above quadratic equation represents an ellipse (instead of other quadratic curves such as a hyperbola or a parabola) centered at $\mathbf{m}_x = [\mu_0, \mu_1]^T$, When $N > 2$, the equation $N(\mathbf{x}, \mathbf{m}_x, \mathbf{\Sigma}_x) = c_0$ represents a hyper ellipsoid in the N-dimensional space. The center and spatial distribution of this ellipsoid are determined by $\mathbf{m}_x$ and $\mathbf{\Sigma}_x$, respectively. When $\mathbf{x} = [x_0, \cdots, x_{N-1}]^T$ is completely decorrelated by KLT:

$$\mathbf{y} = \mathbf{\Phi}^T\mathbf{x}$$

the covariance matrix becomes diagonalized:

$$\mathbf{\Sigma}_y = \mathbf{\Lambda} = \begin{bmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_{N-1} \end{bmatrix} = \begin{bmatrix} \sigma_{y_0}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{y_1}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{y_{N-1}}^2 \end{bmatrix}$$

and the quadratic equation becomes:

$$(\mathbf{y}-\mathbf{m}_y)^T\mathbf{\Sigma}_y^{-1}(\mathbf{y}-\mathbf{m}_y) = \sum_{i=0}^{N-1}\frac{(y_i-\mu_{y_i})^2}{\lambda_i} = \sum_{i=0}^{N-1}\frac{(y_i-\mu_{y_i})^2}{\sigma_{y_i}^2} = c_1$$

This equation represents a standard ellipsoid in the N-dimensional space. In other words,the KLT transform $\mathbf{y} = \mathbf{\Phi}^T\mathbf{x}$ rotates the coordinate system in such a way that the semi-principal axes of the ellipsoid associated with the normal distribution of $\mathbf{x}$ are in parallel with ϕ_i $(i = 0, \cdots, N-1)$, the axes of

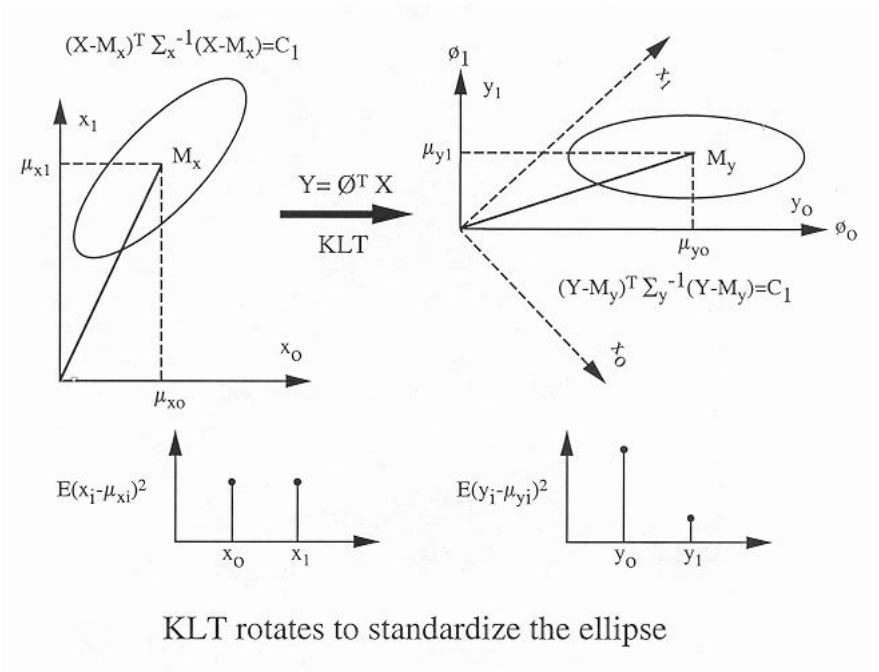


Figure 2: Geometric interpretation of KLT

the new coordinate system. Moreover, the length of the semi-principal axis parallel to the basis vector ϕ_i is equal to the square root of the corresponding eigenvalue $\sqrt{\lambda_i} = \sigma_{y_i}$.

The standardization of the ellipsoid is the essential reason why the rotation of KLT can achieve two highly desirable outcomes: (a) the complete decorrelation of the signal components, and (b) optimal distribution and compaction of the energy or information contained in the signal, as illustrated in the figure below.

Comparison with Other Orthogonal Transforms

To illustrate the optimality of the KLT transform in terms of the two desirable properties discussed above, we compare KLT with other orthogonal transforms such as identity transform (no transform), Walsh-Hadamard transform, discrete cosine transform and discrete Fourier transform DFT in the following examples.

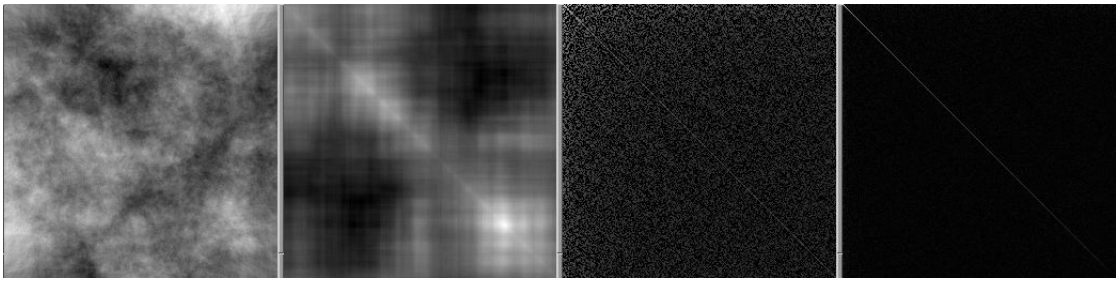


Figure 3: Image of clouds and covariance matrics after various transforms

Example 1

Each row of a 256×256 image of clouds (left panel in the figure below) can be treated as one observation of a 1D random vector $\mathbf{x}$ (with 256 components). Different orthogonal transforms $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ are carried out and the corresponding covariance matrices $\mathbf{\Sigma}_y$ are obtained and compared to see how well each transform decorrelates the signal and compacts its energy. The figure below shows the original image (left panel) and three covariance matrices corresponding to no transform (identify transform matrix), DCT, and KLT. As the behaviors of DFT and WHT are very similar to that of DCT, they are not discussed here. The pixel intensities of the images for covariance matrices are rescaled by a mapping $y = x^{0.3}$ so that those low values can still be visible.

In the secnd panel showing the covariance matrix of the original signal without any transform, there exist quite a lot bright areas off the main diagonal, indicating that many signal components are highly correlated ($\sigma_{ij}^2 > 0$). In the third panel showing the covariance matrix after a DCT, the values of the off-diagonal elements are much reduced, indicating that the signal components are significantly decorrelated. Finally, in the last panel showing the covariance matrix after a KLT, the off-diagonal elements are zero, i.e., the signal components are completely decorrelated.

The effect of energy compaction can also be seen in the figure, as the brightness of the elements along the main diagonal is gradually reduced from top-left to bottom-right. This effect is more clearly shown in the figure below, where the energy distribution among the N elements is plotted. The flat curve is the original energy distribution (no transform), while the curve of steepest descent (high on the left and low on the right) represents the energy distribution after KLT. The intermediate ones are by DCT and WHT

Figure 4: Signal energy distribution after various transforms

with similar effects.

The effect of energy compaction is also illustrated by the table below showing the number of components needed in order to keep certain percentage of the total dynamic energy (information) in the signal. For example, if one wants to keep 99% of the total energy contained in the original signal, 250 out of the total 256 components are needed without transform, 97 out of 256 are needed after DCT, and only 55 are needed after KLT.

Percentage:	90	95	99	100
no transform:	209	230	250	256 (all)
DCT:	10	22	97	256 (all)
KLT:	7	13	55	256 (all)

Based on the example above, some observations could be made. First, all orthogonal transforms have the tendency of decorrelating the given signals, and KLT does it optimally. Specifically, given the value of a time sample of a signal as a function of time (e.g., the temperature as a function of time), the value of the next sample can be predicted with reasonable confidence to be close to the current one, i.e., two consecutive time samples are highly correlated. On the other hand, after an orthogonal transform, the magnitude (or the energy proportional to the magnitude squared) of a certain frequency component bares little information in terms of the magnitude (or energy) of the next frequency component, i.e., the two frequency components are much less correlated than the time samples before the transform.

Second, at the same time, an orthogonal transform tends to compact the energy contained in the signal into a small number of signal components. For example, after a DFT or DCT, most of the energy is concentrated in a small

number of low frequency components as well as the DC component. Most of the high frequency components carry little energy.

Third, although the KLT is optimal in terms of signal decorrelation and energy compaction, the performance of other transforms are not too different from that of the KLT. In the example above, the performance of DCT is reasonably close to that of the KLT, indicating that the DCT could be used as a suboptimal transform to achieve significant signal decorrelation and energy compaction, although not optimally, but with much reduced computational complexity.

Example 2

The example above clearly demonstrates that after an orthogonal transform the signal is less correlated and its energy more compacted. However, is this always true? The answer is, it depends on the nature of the specific signal at hand. The general claim that orthogonal transforms tend to reduce signal correlation is based on an implicit assumption that signals in reality are mostly continuous and smooth due to the nature of underlying physics. Given the value of the current sample of a time signal, one could estimate the value of the next sample to be within a certain neighborhood of the current one (i.e., the two signal components are highly correlated), as any major discontinuity in a time signal corresponds to an energe surge in the physical process, which is in general not very likely.

However, when the assumption of smooth signal is not necessarily valid, orthogonal transforms such as DCT may not perform well in terms of signal decorrelation and energy compaction, sometimes the signal correlation may even *increase* after the transform. Also the energy is not necessarily always compacted by the transform. This is illustrated in the following example.

The left panel of the figure below is an image showing the texture of sand, where the pixels are not correlated as in the image of clouds, since the color of a grain of sand is not related to that of the neighboring grains. The second panel shows the covariance matrix of the row vectors of the image, where all off-diagonal elements have very low values, indicating the pixels are hardly correlated. In comparison, the third panel shows the covariance matrix after the DCT, with most of the off-diagonal elements having higher values than those before the transform, indicating the signal correlation is significantly increased. Finally, the last panel is the covariance matrix after the KLT, showing that the signal is completely decorrelated.

The energy distribution plots shown below indicate that DCT does not make any improvement in term of energy compaction, compared to the orig-

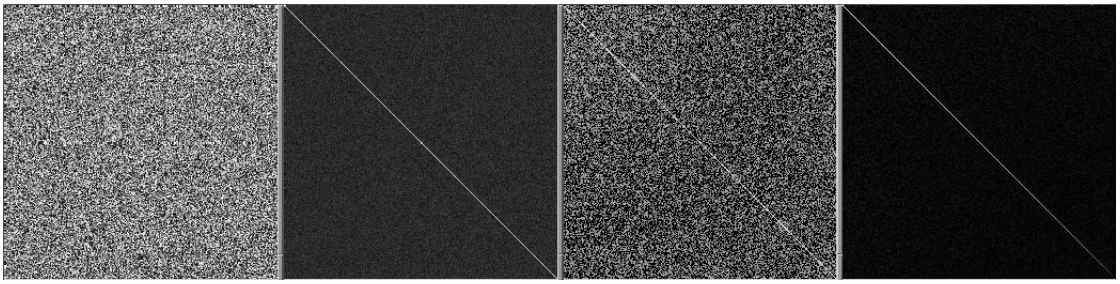


Figure 5: Image of sands and covariance matrics after various transforms

Figure 6: Signal energy distribution after various transforms

inal signal (the two very similar flat plots), but KLT can still compact the energy (the other plot high on the left low on the right), although this compaction by KLT is much less effective than in the previous example.

From the two examples above, one can see that whether an orthogonal transform can decorrelate the signal or not depends on the nature of the signal. If it is initially highly correlated, as is true for most of the physical signals, orthogonal transform will significantly decorrelate the signal, as well as compacting its energy. This is essentially the reason why orthogonal transforms are widely used in data processing. However, in the not too likely case where the signal is not correlated to start with, an orthogonal transform may not reduce the signal correlation, sometimes it may even increase it, as shown in the second example above. Only the KLT can always guarantee that the signal is complete decorrelated, and its energy optimally compacted.

Example 3

While the KLT is the optimal transform in terms of energy compaction, all other orthogonal transforms share the same property to a less extent,

as discussed before. Here we take another look at this property through a specific example, referred to as *Fourier descriptor* in image processing literatures.

A two-dimensional shape in an image can be described by all the pixels along its boundary, in terms of there coordinates $(x[m], y[m])$, $(m = 0, \cdots, N-1)$, where N is the total number of pixels along the boundary. The coordinates $x[m]$ and $y[m]$ can be treated, respectively, as the real and imaginary components of a complex number $z[m] = x[m] + j\ y[m]$, and the Fourier transform can be carried out to obtain the Fourier coefficients (Fourier descriptors) of the shape:

$$Z[n] = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} z[m] e^{-j2\pi mn/N}, \qquad n = 0, \cdots, N-1$$

Based on these coefficients $Z[n]$, the original shape can be reconstructed by inverse Fourier transform:

$$z[m] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} Z[n] e^{j2\pi mn/N}, \qquad m = 0, \cdots, N-1$$

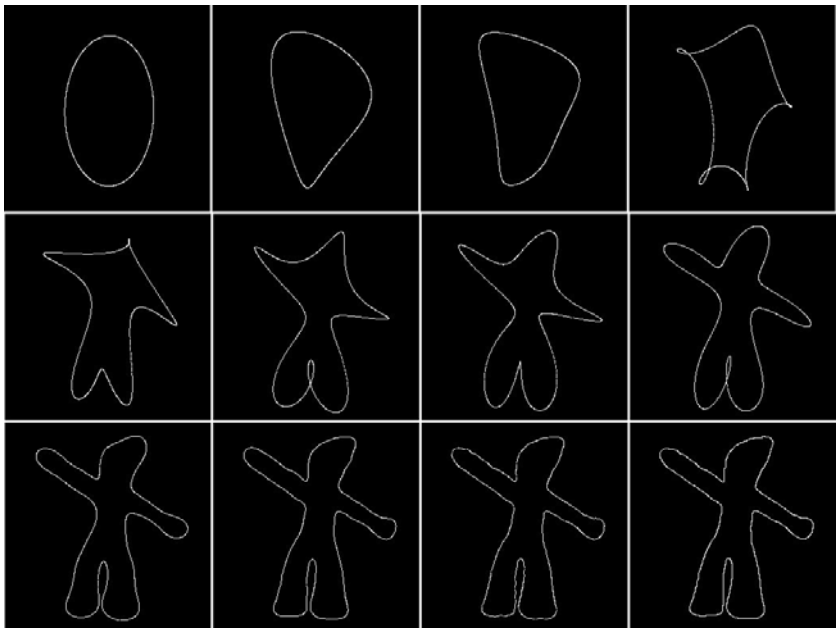
The inverse Fourier transform using all N coefficients will perfectly reconstruct the original one. While this result is not surprising at all, it is interesting to observe the reconstructed shape using only the first $M < N$ low frequency components. Note that since the Fourier transform is a complex transform with both negative frequencies as well as positive ones in the frequency spectrum, the inverse transform with M components needs to contain both positive and negative terms symmetric to the DC component in the middle:

$$\hat{z}[m] = \sum_{k=-M/2}^{M/2} Z[k] e^{j2\pi mk/N} \quad (m = 0, \cdots, N-1)$$

As an example, the shape of Gumby is used to illustrate this idea, as shown below. The coordinates of the $N = 1,157$ boundary pixels are first transformed to frequency domain, and then the shape is reconstructed by inverse transform using different number of the total N frequency components as shown in the figure. It can be seen that the reconstructed shape using less than 5% of the total components ($M = 50$ out of $N = 1,157$, 2nd to the right in the bottom row of the figure) is virtually the same as the original shape,

which can also be perfectedly reconstructed using all $1,157$ components (last shape bottom row). This example clearly illustrates the fact that most of the information (energy) representing the 2D shape is contained in a small number of the low frequency components, while all remaining high frequency components carry little information and can therefore be neglected.

(a) Gumby



(b) Reconstructed Gumby shapes

Figure 7: Reconstructed shapes using 1,2,3,4,5,6,7,8,20,50,100, and all 1,157 components

Applications

As the optimal orthogonal transform, the KLT transform finds many applications in a wide variety of fields. Here we will just discuss some of such applications.

Image compression

Assume a set of N images of size $K = rows \times columes$ are to be stored or transmitted. The pixels of the smae position in all these images are used to form a N-dimensional vector and there are in total K such vectors. Treating these vectors as random vectors, we can find their mean vector $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$, and the KLT can be carried out to transform these vectors into a lower dimensional space of $M << N$ dimensions.

Example:

A set of twenty face images are KLT transformed to obtaine the eigenimages in the transform domain as shown in the figures:

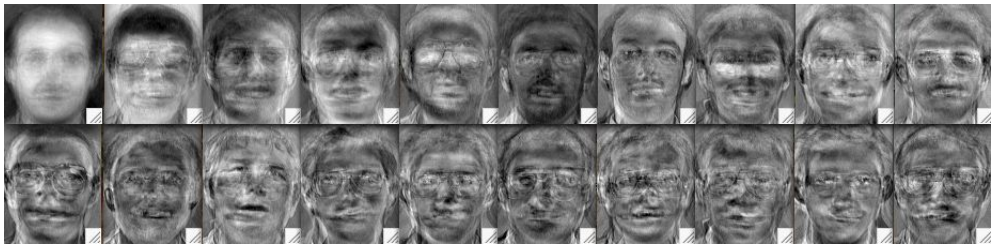
It can be seen that the first few eigenfaces capture the most essential features shared by all of the faces. For example, the first eigenface represents a most generic face in the dark background, and the second eigenface represents the dark hair. The rest of the eigenfaces represent some other features with progressively less importance.

The table below shows the percentage of energy contained in each component:

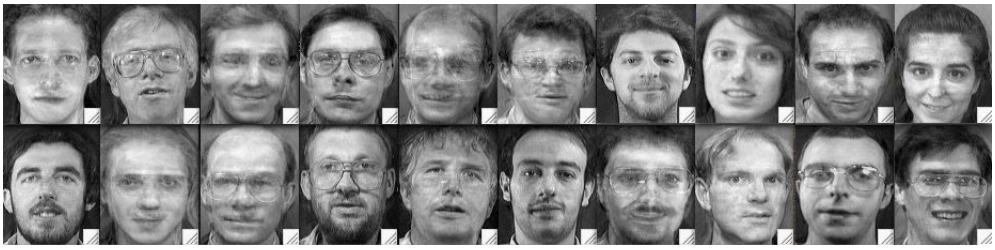
components	1	2	3	4	5	6	7	8	9	10
percentage energy	48.5	11.6	6.1	4.6	3.8	3.7	2.6	2.5	1.9	1.9
accumulative energy	48.5	60.1	66.2	70.8	74.6	78.3	81.0	83.5	85.4	87.3
	11	12	13	14	15	16	17	18	19	20
	1.8	1.6	1.5	1.4	1.3	1.2	1.1	1.1	0.9	0.8
	89.	90.7	92.2	93.6	94.9	96.1	97.2	98.2	99.2	100.0

Reconstructed faces using 95% of the total information are also shown in the figure. The method of eigenfaces has also been used in facial recognition.

(a) The original face images



(b) The eigenfaces



(c) Reconstructed faces

Figure 8: KLT transform of face images

Remote sensing

In remote sensing, images of the surface of either the Earth or other planets such as Mars are taken by satellites, for various studies (e.g., geology, geography, etc.). The camera system on the satellite has a set of N sensors each sensitive to a different wavelength band in the visible and infrared range of the electromagnetic spectrum. Depending on the number of sensors N, the data are referred to as either multi or hyper-spectral images. At each pixel in the image, a set of N values each produced by one of the N sensors form a spectral profile that characterizes the surface material.

As different types of materials on the ground surface usually have different spectral profiles, one typical application of the multi- or hyper-spectral data is to classify the pixels in the image into different classes each corresponding to a certain surface material. When N is large, KLT can be used to reduce the dimensionality without loss of essential information. Specifically, the N values associated with each pixel are considered as a vector in the N-dimensional vector space, whose dimensionality will then be reduced from N to $M << N$ by the KLT transform. All classification can then be subsequently carried out in this low dimensional space, thereby significantly reducing the computional complexity.

Feature extraction for pattern recognition

In many applications, various objects, called patterns in the field of machine learning, in the images (e.g., hand-written characters, human faces, etc.) need to be classified. As the first step of this process, a set of features pertaining to the patterns of interest need to be extracted. KLT can be used for this purpose. Assum a set of images are taken, each containing one of the ten numbers from 0 to 9 (or the face of one individual). Each image is treated as a vector by concatenating all of its rows one after another. Next the mean vector and covarance matrix of these vectors are obtained. Based on the covariance matrix, the KLT is carried out to reduce the dimensionality of the vectors from N to $M << N$. Alternatively, to better extract the information pertaining to the difference between different classes of patterns, the KLT can be based on a different matrix called between-class scatter matrix, which represnts the separability of the classes. Specifically, we use the eigenvectors corresponding to the M largest eiganvalues of the between-class scatter matrix to form an $M \times N$ transform matrix. After the transform

by this matrix, the classification is carried out in this M-dimensional space with much reduced computational complexity.

Data visualization

In various data analysis applications, it is sometimes desirable to visualize the data, for example, to find out how the data points are distributed in the feature space. However, visualization is obviously impossible if the dimensionality of the data is higher than three. In such cases the data points can be projected from the original N-dimensional space to a $M = 2$ dimensional space by the KLT transform based on the covariance matrix of the data points, so that most of the information characterizing the spatial distribution of the data points is conserved. In some cases $M = 3$ dimensions can be used if a 3D rotation can be simulated to show multiple 2D projections of the 3D space from different vantage points.