Principal Component Transform

Multivariate Random Signals

A real time signal $x(t)$ can be considered as a random process and its samples $x_m \ (m = 0, \cdots, N-1)$ a random vector:

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

The *mean vector* of X is

$$M_X \triangleq E(X) = [E(x_0), \cdots, E(x_{N-1})]^T = [\mu_{x_0}, \cdots, \mu_{x_{N-1}}]^T$$

The *covariance matrix* of X is

$$\Sigma_X \triangleq E[(X - M_X)(X - M_X)^T] = E(XX^T) - M_X M_X^T = \begin{bmatrix} .. & .. & .. \\ .. & \sigma_{ij}^2 & .. \\ .. & .. & .. \end{bmatrix}$$

where $\sigma_{ij}^2 \triangleq E(x_i x_j) - \mu_{x_i}\mu_{x_j}$ is the covariance of two random variables x_i and x_j. When $i = j$, σ_{ij}^2 becomes the variance of x_i, $\sigma_i^2 \triangleq E(x_i^2) - \mu_{x_i}^2$.

The *correlation matrix* of X is

$$R_X \triangleq E(XX^T) = \begin{bmatrix} .. & .. & .. \\ .. & r_{ij} & .. \\ .. & .. & .. \end{bmatrix}$$

where $r_{ij} = \sigma_{ij}^2 + \mu_{x_i}\mu_{x_j}$. Both Σ_X and R_X are symmetric matrices (Hermitian if X is complex).

A signal vector X can always be easily converted into a zero-mean vector $X' = X - M_X$ with all of its information (or dynamic energy) conserved. In the following, without loss of generality, we will assume $M_X = 0$ and therefore $\Sigma_X = R_X$.

Before reading on, it is highly recommended to review the basics of

Covariance and Correlation

The mean and the covariance of two variables x and y can be estimated by averaging the outcomes of the random experiment concerning the variables repeated K times:

$$\hat{\mu}_x = \frac{1}{K}\sum_{i=1}^{K} x_i$$

and

$$\hat{\sigma}_{xy} = \frac{1}{K}\sum_{i=1}^{K} x_i y_i$$

The *correlateion coefficient* between two variables x and y are defined as the normalized version of the estimated covariance between the two variables:

$$r_{xy} = \frac{\sigma_{xy}^2}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2}} = \frac{\sigma_{xy}^2}{\sigma_x \cdot \sigma_y}$$

Due to Cauchy-Schwarz inequality,

$$|X \cdot Y| = \parallel X \parallel \cdot \parallel Y \parallel |cos\ \phi| = \left|\sum_{i=1}^{K} x_i y_i\right| \leq \sqrt{\sum_{i=1}^{K} x_i^2}\sqrt{\sum_{i=1}^{K} y_i^2} = \parallel X \parallel \cdot \parallel Y \parallel$$

where ϕ is the angle between the two K-dimensional vectors $X = [x_1, \cdots, x_K]^T$ and $Y = [y_1, \cdots, y_K]^T$. we see that

$$|r_{xy}| = \frac{\left|\sum_{i=1}^{K} x_i y_i\right|}{\sqrt{\sum_{i=1}^{K} x_i^2}\sqrt{\sum_{i=1}^{K} y_i^2}} \leq 1$$

Examples

Consider the following examples showing the correlation coefficient r_{xy} between two random variables x and y under various situations.

- Example 1: assume the experiment concerning x and y is repeated $K = 3$ times with the outcomes:

	1st	2nd	3rd
x	1	2	3
y	1	2	3

The means and covariance of x and y can be estimated as

$$\hat{\mu}_x = \frac{1}{K}\sum_{i=1}^{K} x_i = \hat{\mu}_y = \frac{1}{K}\sum_{i=1}^{K} y_i = 2$$

$$\hat{\sigma}_{xy}^2 = \frac{1}{K}\sum_{i=1}^{K}(x_i - \mu_x)(y_i - \mu_y) = (1\times 1+2\times 2+3\times 3)/3 - 2\times 2 = 0.667$$

$$\hat{\sigma}_x^2 = \hat{\sigma}_y^2 = 0.667$$

And the correlation coefficient is:

$$r_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \cdot \sigma_y} = \frac{0.667}{\sqrt{0.667}\sqrt{0.667}} = 1$$

We see that in this case x and y are highly (maximally) correlated $r_{xy} = 1$.

- Example 2: assume the outcomes of the 3 experiments are

	1st	2nd	3rd
x	1	2	3
y	3	2	1

Similarly, we get

$$\hat{\mu}_x = \frac{1}{K}\sum_{i=1}^{K} x_i = \hat{\mu}_y = \frac{1}{K}\sum_{i=1}^{K} y_i = 2$$

$$\hat{\sigma}_{xy}^2 = \frac{1}{K}\sum_{i=1}^{K}(x_i - \mu_x)(y_i - \mu_y) = -0.667$$

$$\hat{\sigma}_x^2 = \hat{\sigma}_y^2 = 0.667$$

And the correlation coefficient is:

$$r_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \cdot \sigma_y} = \frac{-0.667}{\sqrt{0.667}\sqrt{0.667}} = -1$$

indicating that the two variables are highly negatively correlated.

- Example 3: assume the outcomes are:

	1st	2nd	3rd
x	1	2	3
y	2	2	2

We have

$$\hat{\mu}_x = \hat{\mu}_y = 2$$
$$\hat{\sigma}^2_{xy} = r_{xy} = 0$$

indicating that the two variables are totally uncorrelated (unrelated).

- Example 4: assume the outcomes are:

	1st	2nd	3rd
x	2	2	2
y	1	2	3

We have

$$\hat{\mu}_x = \hat{\mu}_y = 2$$
$$\hat{\sigma}^2_{xy} = r_{xy} = 0$$

indicating that the two variables are totally uncorrelated (unrelated).

- Example 5: assume the experiment is carried out $K = 5$ times (combination of examples 4 and 5) with the outcomes:

	1st	2nd	3rd	4th	5th
x	1	2	2	2	3
y	2	2	2	3	2

We still have

$$\hat{\mu}_x = \hat{\mu}_y = 2$$
$$\hat{\sigma}^2_{xy} = r_{xy} = 0$$

indicating that the two variables are totally uncorrelated (unrelated).

The Principal Component Transform

The *Principal Component Transform* is also called *Karhunen-Loeve Transform (KLT)*, *Hotelling Transform*, or *Eigenvector Transform.*

Let ϕ_i and λ_i be the ith eigenvector and eigenvalue of the covariance matrix Σ_X:

$$\Sigma_X \phi_i = \lambda_i \phi_i \qquad (i = 0, \cdots, N-1)$$

We can construct an $N \times N$ matrix Φ

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

Since the columns of Φ are the eigenvectors of a symmetric (Hermitian if X is complex) matrix R_X, Φ is orthogonal (unitary):

$$\Phi^T \Phi = I$$

i.e.,

$$\Phi^{-1} = \Phi^T$$

and we have

$$\Sigma_X \Phi = \Phi \Lambda$$

where $\Lambda = diag(\lambda_0, \cdots, \lambda_{N-1})$. Or, we have

$$\Phi^{-1} \Sigma_X \Phi = \Phi^T \Sigma_X \Phi = \Lambda$$

We can now define the orthogonal (unitary if X is complex) *Principal Component Transform* of X by

$$\begin{cases} Y = \Phi^T X \\ X = \Phi Y \end{cases}$$

The ith component of the forward transform $Y = \Phi^T X$ is the projection of X on ϕ_i:

$$y_i = (\phi_i, X) = \phi_i^T X$$

and the inverse transform $X = \Phi Y$ represents X in the N-dimensional space spanned by ϕ_i $(i = 0, 1, \cdots, N-1)$:

$$X = \sum_{i=0}^{N-1} y_i \phi_i$$

KLT Completely Decorrelates the Signal

KLT is the optimal orthogonal transform in the following sense:

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

The first property is simply due to the definition of KLT, and the second property is due to the fact that KLT redistributes the energy among the N components in such a way that most of the energy is contained in a small number of components of $Y = \Phi^T X$.

To see the first property, consider the mean vector M_Y and covariance matrix Σ_Y of Y:

$$M_Y = E(Y) = E(\Phi^T X) = \Phi^T E(X) = \Phi^T M_X$$

$$\begin{aligned}
\Sigma_Y &= E(YY^T) - M_Y M_Y^T = E[\Phi^T X(\Phi^T X)^T] - \Phi^T M_X(\Phi^T M_X)^T \\
&= E[\Phi^T (XX^T)\Phi] - \Phi^T M_X M_X^T \Phi \\
&= \Phi^T [E(XX^T) - M_X M_X^T]\Phi \\
&= \Phi^T \Sigma_X \Phi = \Lambda
\end{aligned}$$

We see that after KLT, the correlation matrix of the signal is diagonalized, i.e., the correlation $\sigma_{ij} = 0$ between any two components x_i and x_j is always zero. In other words, the signal is completely decorrelated.

Also, we have

$$R_Y = \Sigma_Y + M_Y M_Y^T = \Lambda + M_Y M_Y^T$$

KLT Optimally Compacts the Energy

Consider a general orthogonal transform pair defined as

$$\begin{cases} Y = A^T X \\ X = AY \end{cases}$$

where X and Y are N by 1 vectors and A is an arbitrary N by N orthogonal matrix $A^{-1} = A^T$.

We represent A by its column vectors $A_i, \quad (i = 0, \cdots, N-1)$ as

$$A = [A_0, \cdots, A_{N-1}]$$

or

$$A^T = \begin{bmatrix} A_0^T \\ \cdot \\ \cdot \\ A_{N-1}^T \end{bmatrix}$$

Now the ith component of Y can be written as

$$y_i = A_i^T \, X$$

As we assume the mean vector of X is zero $M_X = 0$ (and obviously we also have $M_Y = A^T M_x = 0$), we have $\Sigma_X = R_X$, and the variance of the ith element in both X and Y are

$$\sigma_{x_i}^2 = E(x_i^2) \triangleq E(e_{x_i})$$

and

$$\sigma_{y_i}^2 = E(y_i^2) \triangleq E(e_{y_i})$$

where $e_{x_i} \triangleq x_i^2$ and $e_{y_i} \triangleq y_i^2$ represent the energy contained in the ith component of X and Y, respectively. In order words, the trace of Σ_X (the sum of all the diagonal elements of the matrix) represents the expectation of the total amount of energy contained in the signal X

$$\text{Total energy contained in } X = tr\Sigma_X = \sum_{i=0}^{N-1} \sigma_{x_i}^2 = \sum_{i=0}^{N-1} E(x_i^2) = E(\sum_{i=0}^{N-1} e_{x_i})$$

Since an orthogonal transform A does not change the length of a vector X, i.e., $\| Y \|=\| AX \|=\| X \|$, where

$$\| X \| \triangleq \sqrt{\sum_{i=0}^{N-1} x_i^2} = \sqrt{\sum_{i=0}^{N-1} e_{x_i}}$$

the total energy contained in the signal vector X is conserved after the orthogonal transform. (This conclusion can also be obtained from the fact that orthogonal transforms do not change the trace of a matrix.)

We next define

$$S_m(A) \triangleq \sum_{i=0}^{m-1} E(y_i^2) = \sum_{i=0}^{m-1} \sigma_{y_i}^2 = \sum_{i=0}^{m-1} E(e_{y_i})$$

where $m \leq N$. $S_m(A)$ is a function of the transform matrix A and represents the amount of energy contained in the first m components of $Y = A^T X$. Since the total energy is conserved, $S_m(A)$ also represents the percentage of energy contained in the first m components. In the following we will show that $S_m(A)$ is maximized if and only if the transform A is the KLT:

$$S_m(\Phi) \geq S_m(A)$$

i.e., KLT optimally compacts energy into a few components of the signal.

Consider

$$\begin{aligned} S_m(A) &\triangleq \sum_{i=0}^{m-1} E(y_i^2) = \sum_{i=0}^{m-1} E[A_i^T X (A_i^T X)^T] \\ &= \sum_{i=0}^{m-1} E[A_i^T X (X^T A_i)] = \sum_{i=0}^{m-1} A_i^T E(X X^T) A_i \\ &= \sum_{i=0}^{m-1} A_i^T R_X A_i \end{aligned}$$

Now we need to find a transform matrix A so that

$$\begin{cases} S_m(A) \rightarrow max \\ \text{subject to } A_j^T A_j = 1 \quad (j = 0, \cdots, m-1) \end{cases}$$

The constraint $A_j^T A_j = 1$ is to guarantee that the column vectors in A are normalized. This constrained optimization problem can be solved by Lagrange multiplier method as shown below.

We let

$$
\begin{aligned}
& \frac{\partial}{\partial A_i}[S_m(A) - \sum_{j=0}^{m-1} \lambda_j (A_j^T A_j - 1)] = 0 \\
= \quad & \frac{\partial}{\partial A_i}[\sum_{j=0}^{m-1} (A_j^T R_X A_j - \lambda_j A_j^T A_j + \lambda_j)] \\
= \quad & \frac{\partial}{\partial A_i}[A_i^T R_X A_i - \lambda_i A_i^T A_i] \\
\stackrel{*}{=} \quad & 2R_x A_i - 2\lambda_i A_i = 0
\end{aligned}
$$

(* the last equal sign is due to explanation in the handout of review of linear algebra.) We see that the column vectors of A must be the eigenvectors of R_X:

$$R_X A_i = \lambda_i A_i \qquad (i = 0, \cdots, m-1)$$

i.e., the transform matrix must be

$$A = [A_0, \cdots, A_{N-1}] = \Phi = [\phi_0, \cdots, \phi_{N-1}]$$

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

$$S_m(\Phi) = \sum_{i=0}^{m-1} \phi_i^T R_X \phi_i = \sum_{i=0}^{m-1} \lambda_i$$

where the ith eigenvalue λ_i of R_X is also the average (expectation) energy contained in the ith component of the signal. If we choose those $\phi_i^{\prime} s$ that correspond to the m largest eigenvalues of R_X: $\lambda_0 \geq \lambda_1 \geq \cdots \lambda_m \cdots \geq \lambda_{N-1}$, then $S_m(\Phi)$ will achieve maximum.

Geometric Meaning of KLT

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

$$p(x_0, \cdots, x_{N-1}) = N(X, M_X, \Sigma_X) = \frac{1}{(2\pi)^{N/2}\,|\Sigma_X|^{1/2}} exp[-\frac{1}{2}(X-M_X)^T \Sigma_X^{-1}(X-M_X)]$$

where M_X and Σ_X are the mean vector and covariance matrix of X, respectively. When $N = 1$, Σ_X and M_X become σ_x and μ_x, respectively, and the density function becomes single variable normal distribution.

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

$$N(X, M_X, \Sigma_X) = c_0$$

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

$$(X - M_X)^T \Sigma_X^{-1}(X - M_X) = c_1$$

where c_1 is another constant related to c_0, M_X and Σ_X. In particular, with $N = 2$ variables x_0 and x_1, we have

$$\begin{aligned}(X - M_X)^T \Sigma_X^{-1}(X - M_X) &= [x_0 - \mu_{x_0}, x_1 - \mu_{x_1}] \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} x_0 - \mu_{x_0} \\ x_1 - \mu_{x_1} \end{bmatrix} \\ &= a(x_0 - \mu_{x_0})^2 + b(x_0 - \mu_{x_0})(x_1 - \mu_{x_1}) + c(x_1 - \mu_{x_1})^2 = c_1\end{aligned}$$

Here we have assumed

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

The above quadratic equation represents an ellipse (instead of any other quadratic curve) centered at $M_X = [\mu_{x_0}, \mu_{x_1}]^T$, because Σ_X^{-1}, as well as Σ_X, is positive definite:

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

When $N > 2$, the equation $N(X, M_X, \Sigma_X) = c_0$ represents a hyper ellipsoid in the N-dimensional space. The center and spatial distribution of this ellipsoid are determined by M_X and Σ_X, respectively.

When $X = [x_0, \cdots, x_{N-1}]^T$ is completely decorrelated by KLT:

$$Y = \Phi^T X$$

the covariance matrix becomes diagonalized:

$$\Sigma_Y = \Phi^T \Sigma_X \Phi = \Lambda = \begin{bmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_{N-1} \end{bmatrix}$$

and equation $N(X, M_X, \Sigma_X) = c_0$ becomes $N(Y, M_Y, \Sigma_Y) = c_0$, or

$$(Y - M_Y)^T \Sigma_Y^{-1} (Y - M_Y) = \sum_{i=0}^{N-1} \frac{(y_i - \mu_{y_i})^2}{\sqrt{\lambda_i}^2} = c_1$$

This equation represents a standard ellipsoid in the N-dimensional space. In other words, KLT $Y = \Phi^T X$ rotates the coordinate system so that the ellipsoid associated with the normal distribution of X becomes a standardized ellipsoid associated with the normal distribution of Y, whose axes are parallel to ϕ_i $(i = 0, \cdots, N-1)$, the axes of the new coordinate system, with the corresponding semi axes equal to $\sqrt{\lambda_i}$.