

# Lecture 10: { Karhunen-Loève Transform Principal Component Analysis

Note Title

Consider a stochastic process in  $\mathbb{C}^n$  s.t.

$$\mathbb{X} \in \mathbb{C}^n \sim f_{\mathbb{X}}(x_1, \dots, x_n) \text{ pdf.}$$

$\mathbb{X}$  random vector, i.e., each coordinate is a r.v. (or random signal)

Now consider its **covariance**:

$$\Gamma_{\mathbb{X}}[k, l] := E[(X_k - EX_k)(X_l - EX_l)]$$

$\xrightarrow{\text{expectation}}$   $k, l = 1:n$

$$\begin{aligned} \text{i.e., } \Gamma_{\mathbb{X}} &:= E(\mathbb{X} - E\mathbb{X})(\mathbb{X} - E\mathbb{X})^* \\ &= E\mathbb{X}\mathbb{X}^* - (E\mathbb{X})(E\mathbb{X})^* \in \mathbb{C}^{n \times n} \end{aligned}$$

Let  $\{\mathbb{x}_1, \dots, \mathbb{x}_N\}$  be  $N$  realizations of  $\mathbb{X}$ .  
Then, the **sample estimate** of  $\Gamma_{\mathbb{X}}$  is:

$$\hat{\Gamma}_{\mathbb{X}} := \frac{1}{N} \sum_{j=1}^N \mathbb{x}_j \mathbb{x}_j^* - \bar{\mathbb{X}} \bar{\mathbb{X}}^*$$

$$\text{where } \bar{\mathbb{X}} := \frac{1}{N} \sum_{j=1}^N \mathbb{x}_j$$

If we define the **data matrix**

$$X := [\mathbb{x}_1 | \mathbb{x}_2 | \dots | \mathbb{x}_N] \in \mathbb{C}^{n \times N},$$

$$\text{then } \hat{\Gamma}_{\mathbb{X}} = \frac{1}{N} X X^* - \bar{\mathbb{X}} \bar{\mathbb{X}}^*.$$

Suppose we want a **data-adaptive ONB** of  $\mathbb{C}^n$  (unlike fixed ONBs such as DFT, DCT, DST) s.t. the realizations of this stochastic process as a whole (i.e., **on average**) can be **best** approximated by  $m$  coordinates with  $m \ll n$  in the **mean-squared ( $L^2$ ) error sense**.

R.V.:  
upper-  
case

Realizations  
(obs.'s)  
 $\Rightarrow$  lower-  
case

Let  $W \in U(n) :=$  a set of all **unitary** matrices in  $\mathbb{C}^n$  and let  $\underline{Y} = W^* \underline{X}$ .

$\Rightarrow$  viewing  $\underline{X}$  relative to  $W$  (or the ONB consisting of col's of  $W$ )  
 $\underline{X}$  is viewed relative to  $\underline{I}_n$ .

$$\begin{aligned} \Rightarrow \underline{X} &= W \underline{Y} = [\omega_1 | \dots | \omega_n] \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\ &= [e_1 | \dots | e_n] \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \\ &= Y_1 \omega_1 + \dots + Y_n \omega_n. \end{aligned}$$

Suppose we retain  $Y_1 \sim Y_m$  and replace  $Y_{m+1} \sim Y_n$  by  $\alpha_{m+1} \sim \alpha_n$  (predetermined val's).  
 Then the approx. of  $\underline{X}$  is given by

$$\underline{X}^{(m)} := \sum_{j=1}^m Y_j \omega_j + \sum_{j=m+1}^n \alpha_j \omega_j$$

The error is  $\Delta \underline{X} := \underline{X} - \underline{X}^{(m)} = \sum_{j=m+1}^n (Y_j - \alpha_j) \omega_j$ ,  
 and the mean-squared error is

$$\mathcal{E}^{(m)} := E \|\Delta \underline{X}\|^2 = E (\Delta \underline{X})^* (\Delta \underline{X})$$

$$\begin{aligned} &= E \left[ \sum_{j=m+1}^n \sum_{k=m+1}^n \overline{(Y_j - \alpha_j)} (Y_k - \alpha_k) \underbrace{\omega_j^* \omega_k}_{= \delta_{jk}} \right] \\ &= E \left[ \sum_{j=m+1}^n |Y_j - \alpha_j|^2 \right] \end{aligned}$$

Step 1.  $\frac{\partial \mathcal{E}^{(m)}}{\partial \alpha_j} = -2 E(Y_j - \alpha_j) = 0, j = m+1 : n.$

$\Rightarrow \alpha_j = E Y_j = E [w_j^* X] = w_j^* E[X].$

Then, 
$$\begin{aligned} \mathcal{E}^{(m)} &= E \left[ \sum_{j=m+1}^n (Y_j - \alpha_j) \overline{(Y_j - \alpha_j)} \right] \\ &= \sum_{j=m+1}^n E w_j^* (X - EX) (w_j^* (X - EX))^* \\ &= \sum_{j=m+1}^n E w_j^* (X - EX) (X - EX)^* w_j \\ &= \sum_{j=m+1}^n w_j^* E (X - EX) (X - EX)^* w_j \\ &= \sum_{j=m+1}^n w_j^* \Gamma_X w_j \end{aligned}$$

Step 2. What kind of  $\{w_j\}$  minimizes the above quantity subject to  $w_j^* w_j = 1$ ?

$\Rightarrow$  Use the **Lagrange multiplier**:

$$\begin{aligned} \tilde{\mathcal{E}}^{(m)} &:= \mathcal{E}^{(m)} - \sum_{j=m+1}^n \lambda_j (w_j^* w_j - 1) \\ &= \sum_{j=m+1}^n [w_j^* \Gamma_X w_j - \lambda_j (w_j^* w_j - 1)] \end{aligned}$$

$$\frac{\partial \tilde{\mathcal{E}}^{(m)}}{\partial w_j} = 2 \Gamma_X w_j - 2 \lambda_j w_j = 0$$

$$\Rightarrow \Gamma_X w_j = \lambda_j w_j \quad \text{Eigenvalue problem!}$$

$1 \leq m \leq n$  was arbitrary. So, for any  $1 \leq m \leq n$ , we need to solve  $\Gamma_X w_j = \lambda_j w_j, j = 1 : m$ .  
 $\Gamma_X$  : hermitian  $\rightarrow$  unitarily diagonalizable &  $\lambda_j \in \mathbb{R}$

## Remarks

(1) Analyzing (or transforming) the input  $\mathbb{X}$  via the eigenvectors of  $\Gamma_{\mathbb{X}}$  is called **PCA** (or **KLT**).

(2) In practice,  $\hat{\Gamma}_{\mathbb{X}}$  is used for  $\Gamma_{\mathbb{X}}$ .

The quality of  $\hat{\Gamma}_{\mathbb{X}}$  depends on  $n > N$  or  $n < N$ .

classical setting:  $n \ll N$  (e.g., census)  $\Rightarrow \hat{\Gamma}_{\mathbb{X}}$ : good  
neo classical setting:  $n \gg N$  (e.g., images)  $\Rightarrow \hat{\Gamma}_{\mathbb{X}}$ : poor

(3) "Only" optimal in terms of the mean-squared (or the entropy) criterion.

Satosi Watanabe (1965): The Entropy Minimization Criterion.

Define  $\mathcal{D}_{\mathbb{X}} := \text{diag}(\Gamma_{\mathbb{X}}) / \|\text{diag}(\Gamma_{\mathbb{X}})\|_1 \in \mathbb{R}^n$

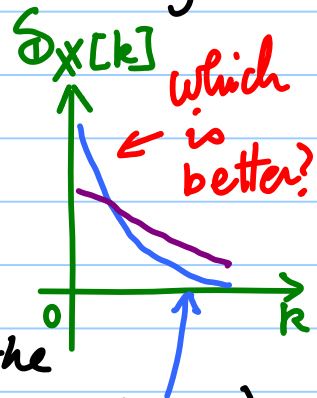
Consider all possible  $W \in U(n)$  and the coordinate transf's  $W^* \mathbb{X}$ .

Define  $H(p) := - \sum_{i=1}^n p_i \log p_i$ ,  $p_i \geq 0$ ,  $\sum_i p_i = 1$ .

$\mathcal{D}_{\mathbb{X}}$  above qualifies as  $p$  thanks to the normalization.  
Then Watanabe showed that

$$H(\mathcal{D}_{W_{KL}^* \mathbb{X}}) = \min_{W \in U(n)} H(\mathcal{D}_{W^* \mathbb{X}})$$

i.e.,  $W_{KL} = \arg \min_{W \in U(n)} H(\mathcal{D}_{W^* \mathbb{X}})$ .



That is, the KLT (or PCA) provides the **minimum entropy** coordinates (**sharper distribution!**)

$\Rightarrow$  Packing more energy (or variance) into the first few coordinates!

$$(4) \quad \mathbb{Y} = W_{KL}^* \mathbb{X}, \quad W_{KL} = [w_1 | \dots | w_n]$$

eigenvectors of  $\Gamma_{\mathbb{X}}$ .

Then, what about  $\Gamma_{\mathbb{Y}}$ ?

$$\begin{aligned} \Gamma_{\mathbb{Y}} &= E(\mathbb{Y} - E\mathbb{Y})(\mathbb{Y} - E\mathbb{Y})^* \\ &= E W_{KL}^* (\mathbb{X} - E\mathbb{X})(\mathbb{X} - E\mathbb{X})^* W_{KL} \\ &= W_{KL}^* E(\mathbb{X} - E\mathbb{X})(\mathbb{X} - E\mathbb{X})^* W_{KL} \\ &= W_{KL}^* \Gamma_{\mathbb{X}} W_{KL} = \text{diag}(\lambda_1, \dots, \lambda_n) \\ &\quad \Gamma_{\mathbb{X}} w_j = \lambda_j w_j, \quad \lambda_j > 0. \end{aligned}$$

$\Rightarrow$  The components of  $\mathbb{Y}$  are **decorrelated!**

$$E(\mathbb{Y}_i - E\mathbb{Y}_i)(\mathbb{Y}_j - E\mathbb{Y}_j) = \lambda_i \delta_{ij}$$

## ★ Relationship between KLT/PCA and SVD

For simplicity, let's consider the **centered** data matrix  $\tilde{X}$  of  $X \in \mathbb{C}^{n \times N}$

$$\tilde{X} := X - \frac{1}{N} X \mathbb{1} \mathbb{1}^T = X \left( \mathbb{I} - \frac{1}{N} \mathbb{1} \mathbb{1}^T \right)$$

where  $\mathbb{1} := (1, 1, \dots, 1)^T \in \mathbb{R}^N$  **centering matrix**

Note  $\bar{X} = \frac{1}{N} X \mathbb{1}$ , so multiplying the centering matrix from right subtracts  $\bar{X}$  from each col. vector  $X_j$  of  $X$ ,  $j=1, \dots, N$ .

$\Rightarrow$  The mean of the col. vectors of  $\tilde{X} = \mathbf{0}$ .

Let the **Singular Value Decomposition (SVD)** of  $\tilde{X}$  be  $\tilde{X} = U \Sigma V^*$  (full SVD)

where  $U \in U(n) \subset \mathbb{C}^{n \times n}$

$V \in U(N) \subset \mathbb{C}^{N \times N}$

$\Sigma \in \mathbb{R}^{n \times N}$  : diagonal  $\sigma_1, \dots, \sigma_{\min(n,N)}$

$$\Sigma = \begin{matrix} n & \left[ \begin{array}{c|c} \sigma_1 & \\ \vdots & \\ \sigma_r & \\ \vdots & \\ 0 & \end{array} \right] & \text{or } N & \left[ \begin{array}{c|c} \sigma_1 & \\ \vdots & \\ \sigma_r & \\ \vdots & \\ 0 & \\ \hline & \mathbf{0} \end{array} \right] \\ \underbrace{\hspace{1.5cm}}_n & & \underbrace{\hspace{1.5cm}}_{N-n} & & \underbrace{\hspace{1.5cm}}_{n-N} & & \underbrace{\hspace{1.5cm}}_N \end{matrix}$$

$n < N$   
classical

$n > N$   
neoclassical

$$r = \text{rank}(\tilde{X}) \leq \min(n, N-1).$$

Note that if  $\{x_1, \dots, x_N\}$  are linearly independent, then  $\text{rank}(X) \leq \min(n, N)$ . However, in the case of its centered version  $\tilde{X} = [\tilde{x}_1 | \dots | \tilde{x}_N]$ ,  $\tilde{x}_j = x_j - \bar{x}$ ,  $j=1:N$ , its column vectors are not linearly indep. because  $\tilde{x}_1 + \dots + \tilde{x}_N = \mathbf{0}$ .

$$\Rightarrow \text{rank}(\tilde{X}) \leq \min(n, N-1).$$

(1) If  $n < N$ , then

$$\begin{aligned}\hat{\Gamma}_{\tilde{X}} &= \frac{1}{N} \tilde{X} \tilde{X}^* = \frac{1}{N} U \Sigma V^* V \Sigma^* U^* \\ &= \frac{1}{N} U \Sigma \Sigma^* U^* \\ &= U \begin{bmatrix} \sigma_1^2/N & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2/N \end{bmatrix} U^* \Rightarrow \text{The eigenvalue} \\ & \text{decomposition} \\ & \text{of } \hat{\Gamma}_{\tilde{X}} \text{ !!}\end{aligned}$$

Hence, in this case

$$U = \underline{W}_{KL}!$$

(2) If  $n > N$ , then we should not compute all  $n$  KLB vectors because  $\text{rank}(\tilde{X}) \leq N-1$ , so computing more than  $N-1$  KLB vectors is useless. Moreover  $n$  could be huge.

$\Rightarrow$  How to compute the top  $N-1$  KLB vec's?

The first  $N-1$  column vectors of  $\tilde{X}V \in \mathbb{C}^{n \times N}$   
= the top  $N-1$  KLB vectors!

why?  $\hat{\Gamma}_{\tilde{X}} = \frac{1}{N} \tilde{X} \tilde{X}^*$

$$\begin{aligned}\text{so, } \hat{\Gamma}_{\tilde{X}} \tilde{X}V &= \frac{1}{N} \tilde{X} \tilde{X}^* \tilde{X}V \\ &= \frac{1}{N} \tilde{X} (V \Sigma^* U^* U \Sigma V^*) V\end{aligned}$$

$$= \frac{1}{N} \tilde{X} V \underbrace{\Sigma^* \Sigma}_{N \times N \text{ diagonal!}} = \tilde{X} V \begin{bmatrix} \sigma_{1/N}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{N-1/N}^2 \\ & & & 0 \end{bmatrix}$$

⇒ Each column vector of  $\tilde{X} V$  is a linear combination of the column vectors of  $\tilde{X}$ , i.e., belongs to  $\text{span}\{\tilde{X}_1, \dots, \tilde{X}_N\}$  !!

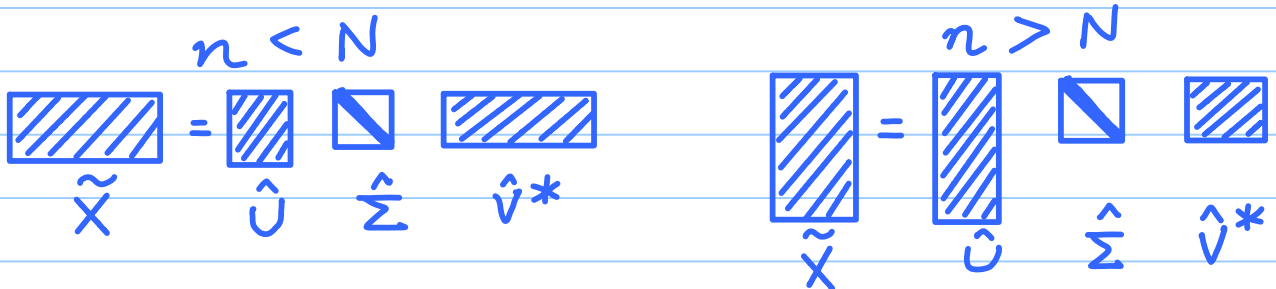
Remark: In either case ( $n < N$  or  $n > N$ ), you do not need to compute  $\hat{\Gamma}_{\tilde{X}}$ . Moreover, you should use the **reduced SVD** instead of the full SVD for the purpose of KLB/PCA computation.

The reduced SVD of  $\tilde{X} \in \mathbb{C}^{n \times N}$

$$\tilde{X} = \hat{U} \hat{\Sigma} \hat{V}^*$$

Let  $p := \min(n, N-1) \geq r = \text{rank}(\tilde{X})$ .

Then  $\hat{U} \in \mathbb{C}^{n \times p}$ ,  $\hat{\Sigma} \in \mathbb{R}^{p \times p}$ ,  $\hat{V} \in \mathbb{C}^{N \times p}$



In Julia, this is done by  
`julia>  $\hat{U}, \hat{\Sigma}, \hat{V} = \text{svd}(\tilde{X}; \text{full} = \text{false});$`   
 while in MATLAB, it's  
`>> [ $\hat{U}, \hat{\Sigma}, \hat{V}$ ] = \text{svd}(\tilde{X}, 'econ');`



Note also  $\tilde{X} \hat{V} = \hat{U} \hat{\Sigma} \hat{V}^* \hat{V} = \hat{U} \hat{\Sigma}$

$$= [\sigma_1 u_1, \dots, \sigma_p u_p]$$

$$= [\tilde{X} v_1, \dots, \tilde{X} v_p]$$

So,  $u_j = \frac{1}{\sigma_j} \tilde{X} v_j$ ,  $j=1, \dots, p = \min(n, N-1)$   
 $= N-1$  if  $n \gg N$ .

In other words, each principal axis  $u_j$  is just a linear combination of the (centered) input vectors  $\tilde{X}_1, \dots, \tilde{X}_N$ !

### Example: The Rogues' Gallery Dataset

- Through the courtesy of Prof. Larry Sirovich
- A set of digitized photos of 143 faces each of which has  $128 \times 128$  pixels i.e.,  $n = 128^2 = 16384$ .
- These faces were of a specific group of people, i.e., Caucasian male students (and some faculty) at Brown Univ., without glasses, mustache, beard.
- Horizontal dilation has been applied s.t. the pupils are placed on two fixed common points.
- 143 faces were randomly divided into 72 and 71 faces for the training and test datasets, i.e.,  $N = 72$ .
- KLB/PCA were computed on the training data set.

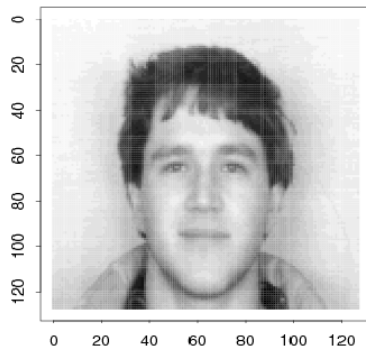
# 72 Faces !

$x_1$   $x_2$  ...

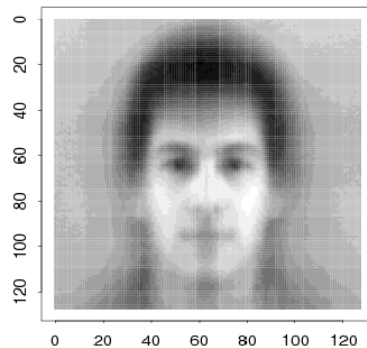


$x_{71}$   $x_{72}$

Original



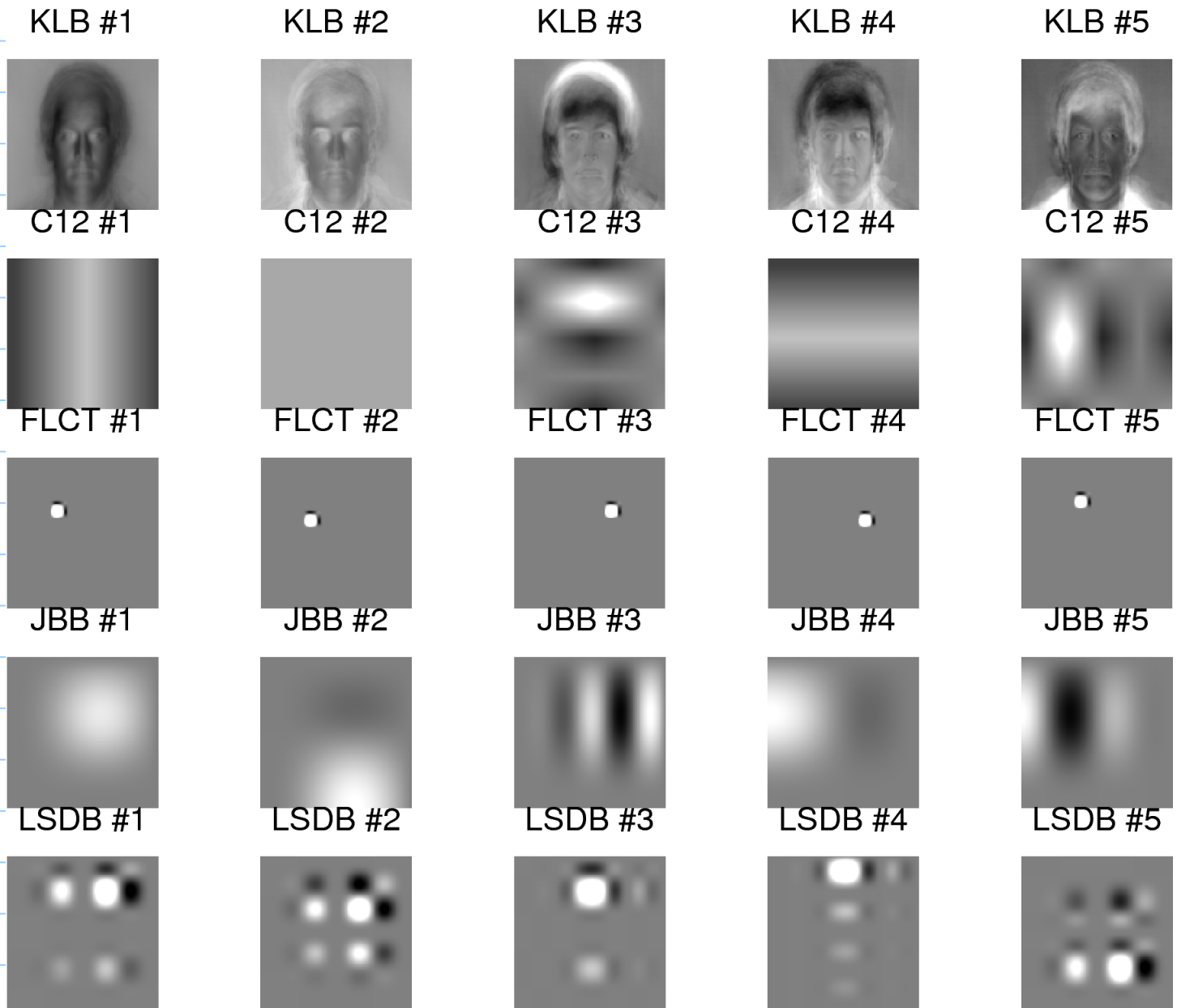
Average face



one of the  
test image

$\bar{x}$

# Comparison of Bases



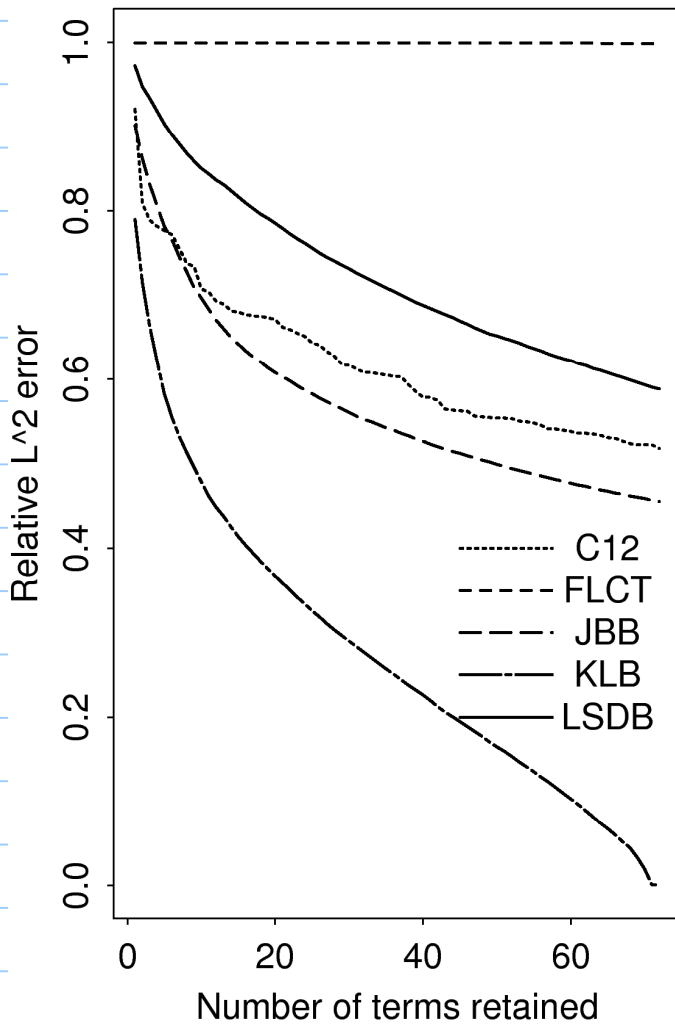
## Remarks:

(1) I am only displaying the top 5 KLB vectors. There are totally 71 KLB vectors plus the mean vector  $\bar{x}$ .

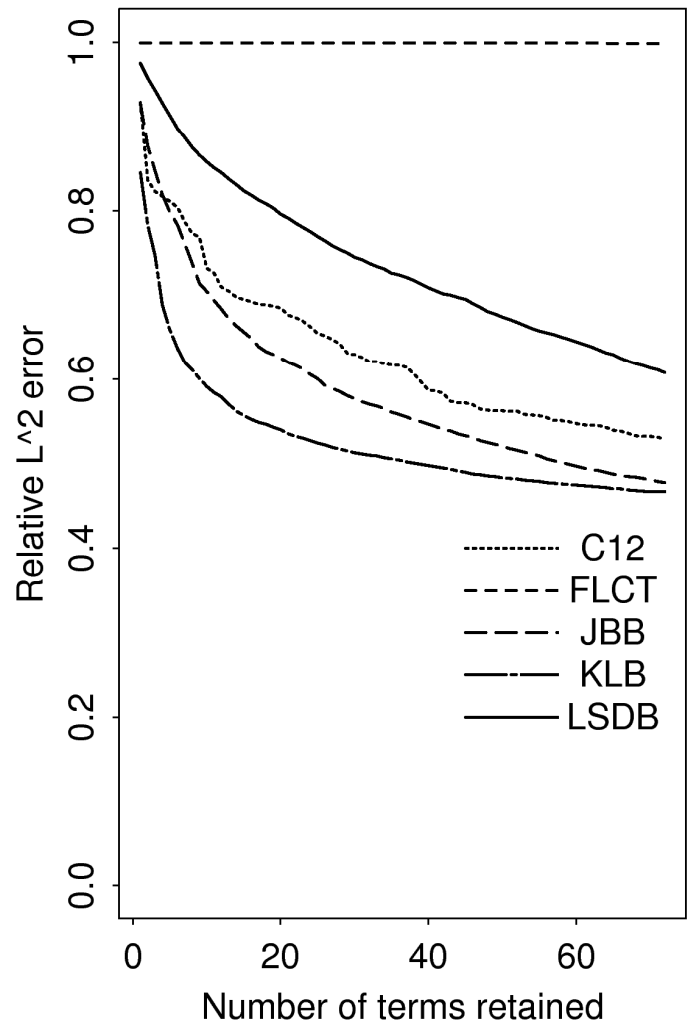
(2) As you can see these KLB vectors = lin. combi's of the 72 faces.

# Approximation Errors

(a) Training Dataset Errors

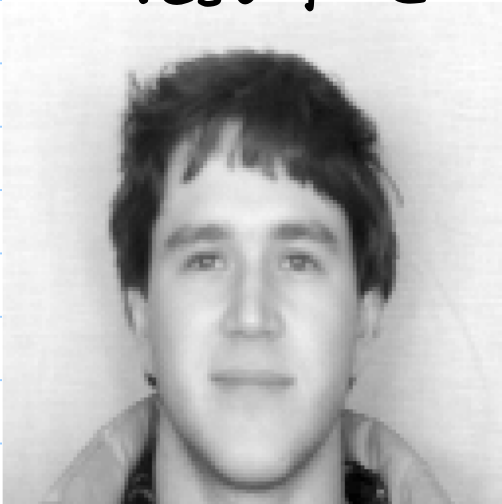


(b) Test Dataset Errors



test face

71 term approx. + ave.



## Remarks:

Note that for each training face, one can reconstruct it within machine precision using the full 71 KLB vectors and the average face  $\bar{x}$ .

However, for the test faces that were not used to generate the KLB, the reconstructions are not perfect.

Why? Basically, viewing these faces as the realizations of a stochastic process using only up to the 2nd order statistics (i.e., mean + covariance) and only using 72 realizations lead to serious limitation of the KLB applicability.

## ★ Relationship between KLT and DCT

For simplicity, let's assume  $E X_j = 0, j=1:n$ .  
Suppose the underlying stochastic process is stationary in wide sense, i.e.,

$$\Gamma_{\times}[k, l] = \Gamma_{\times}[k+m, l+m], \quad \forall m \in \mathbb{Z}$$

(appropriate mod  $N$  needs here)

Furthermore, assume

$$\Gamma_{\times}[k, l] = \rho^{|k-l|} \quad 0 < \rho < 1.$$

This is called the **1st order Markov model**, i.e.,  $X_{k+1}$  depends only on  $X_k$  (plus independent noise). This can be explained as follows.

Suppose  $X_{k+1} = \rho X_k + Z_k$   
where  $\rho \in (0, 1)$ ,  $Z_k$ : iid.  $E Z_k = 0$   
 $\text{var } Z_k = 1$ .

Then

$$E X_k X_{k+1}^* = \rho E |X_k|^2 + E X_k Z_k^* = E X_k E Z_k^*$$

$$E X_k X_{k+2}^* = E X_k (\rho X_{k+1}^* + Z_{k+1}^*)$$
$$= \rho E X_k X_{k+1}^* + E X_k Z_{k+1}^*$$

$$\vdots = \rho^2 E |X_k|^2$$

$$E X_k X_{k+l}^* = \rho^l E |X_k|^2$$

Since we assumed the wide sense stationarity,  $E|X_k|^2$  does not depend on  $k$ , i.e., we can assume

$$E|X_k|^2 = \sigma^2 = \text{const.}, \text{ and}$$

$$\Gamma_{XX} = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \dots & \rho & 1 \end{bmatrix} \Rightarrow \text{Toeplitz matrix!}$$

For further simplicity, let's assume  $\sigma^2 = 1$ .

One can show that

The eigenvalues & eigenvectors of this Toeplitz matrix are:

$$\begin{cases} \lambda_k = \frac{1 - \rho^2}{1 - 2\rho \cos \omega_k + \rho^2}, & k = 0, 1, \dots, n-1. \\ \psi_k[l] = \sqrt{\frac{2}{n + \lambda_k}} \sin\left[\omega_k \left(l - \left(\frac{n-1}{2}\right)\right) + (k+1)\frac{\pi}{2}\right] \end{cases}$$

$l = 0, 1, \dots, n-1.$

where  $\omega_k$  is a solution of the following secular equation:

$$\tan(n\omega_k) = -\frac{(1 - \rho^2) \sin \omega_k}{(1 + \rho^2) \cos \omega_k - 2\rho}$$

Now consider the case when  $\rho \uparrow 1$ .

Then  $\tan n\omega_k = 0 \Leftrightarrow \omega_k = \frac{k\pi}{n}$ ,  $k \neq 0$

For  $k=0$ , we use the small angle perturbation:  $\tan \alpha \approx \alpha$ ,  $\sin \alpha \approx \alpha$ ,  $\cos \alpha \approx 1 - \frac{\alpha^2}{2}$   
using these, we get

$$n \omega_0 = - \frac{(1-\rho^2) \omega_0}{(1+\rho^2)(1-\omega_0^2/2) - 2\rho}$$

$$\Leftrightarrow \omega_0^2 = \frac{2}{1+\rho^2} \cdot \frac{1-\rho^2}{n} + \frac{2(1-\rho)^2}{1+\rho^2}$$

$$\approx \frac{1}{n} (1-\rho^2) \text{ as } \rho \uparrow 1. \quad (*)$$

So, one can say  $\omega_0 \rightarrow 0$  as  $\rho \uparrow 1$ .  
 after all,  $\omega_k = k\pi/n$ ,  $k=0, 1, \dots, n-1$ .

$$\text{Now, } \omega_k[l] = \sqrt{\frac{2}{n+\lambda_k}} \sin \left[ \frac{k\pi}{n} \left( l - \left( \frac{n-1}{2} \right) \right) + (k+1) \frac{\pi}{2} \right]$$

$$= \sqrt{\frac{2}{n+\lambda_k}} \sin \left[ \frac{\pi k}{n} \left( l + \frac{1}{2} \right) + \frac{\pi}{2} \right]$$

$$= \sqrt{\frac{2}{n+\lambda_k}} \cos \left[ \frac{\pi k}{n} \left( l + \frac{1}{2} \right) \right]$$

How about  $\lambda_k$  as  $\rho \uparrow 1$ ?

$$\text{For } k \neq 0, \quad \lambda_k = \frac{1-\rho^2}{1-2\rho \cos \omega_k + \rho^2} \rightarrow 0 \text{ as } \rho \uparrow 1$$

$$\text{For } k=0, \quad \lambda_0 \approx \frac{1-\rho^2}{1-2\rho(1-\omega_0^2/2) + \rho^2}$$

$$(*) \rightarrow \frac{1-\rho^2}{1+\rho}$$

$$= \frac{1-\rho^2}{1-\rho + \rho(1+\rho)/n}$$

$$\rightarrow n \text{ as } \rho \uparrow 1.$$

$$\text{So, } \begin{cases} \omega_0[l] = \frac{1}{\sqrt{n}} \\ \omega_k[l] = \sqrt{\frac{2}{n}} \cos \left[ \frac{\pi k \left( l + \frac{1}{2} \right)}{n} \right] \end{cases} \quad \text{DCT-II !!}$$