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

Note Title

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

R.V.: upper-case

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

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

Now consider its **covariance**:

$$\Gamma_{\mathbf{X}}[k,\ell] := \underset{\sim}{E}\left[(X_k - \underset{\sim}{E}X_k)\overline{(X_\ell - \underset{\sim}{E}X_\ell)}\right]$$

$\underset{\sim}{E}$ ↳ expectation

$k, \ell = 1:n$

i.e., $\Gamma_{\mathbf{X}} := E(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^*$

$$= E\mathbf{X}\mathbf{X}^* - (E\mathbf{X})(E\mathbf{X})^* \in \mathbb{C}^{n\times n}$$

Realizations (obs.'s) ⇒ lower-case

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be $N$ realizations of $\mathbf{X}$.

Then, the **sample estimate** of $\Gamma_{\mathbf{X}}$ is:

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

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

If we define the **data matrix**

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

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

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.**

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

$\Rightarrow$ viewing $\mathbb{X}$ relative to $W$ (or the ONB consisting of col's of $W$)

$\mathbb{X}$ is viewed relative to $I_n$.

$$\Rightarrow \mathbb{X} = W\mathbb{Y} = [w_1 | \cdots | w_n] \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

$$= [e_1 | \cdots | e_n] \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

$$= Y_1 w_1 + \cdots + Y_n w_n .$$

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 $\mathbb{X}$ is given by

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

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

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

$$= E \left[ \sum_{j=m+1}^{n} \sum_{k=m+1}^{n} \overline{(Y_j - \alpha_j)} (Y_k - \alpha_k) \underbrace{w_j^* w_k}_{= \delta_{jk}} \right]$$

$$= E \left[ \sum_{j=m+1}^{n} |Y_j - \alpha_j|^2 \right]$$

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

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

Then, $\mathcal{E}(m) = E\left[\sum_{j=m+1}^{n} (Y_j - \alpha_j)\overline{(Y_j - \alpha_j)}\right]$

$$= \sum_{j=m+1}^{n} E\, w_j^* (X - EX)\left(w_j^* (X - EX)\right)^*$$

$$= \sum_{j=m+1}^{n} E\, w_j^* (X - EX)(X - EX)^* w_j$$

$$= \sum_{j=m+1}^{n} w_j^* E(X - EX)(X - EX)^* w_j$$

$$= \sum_{j=m+1}^{n} w_j^* \Gamma_X w_j$$

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

$\Rightarrow$ Use the Lagrange multiplier:

$$\tilde{\mathcal{E}}(m) := \mathcal{E}(m) - \sum_{j=m+1}^{n} \lambda_j (w_j^* w_j - 1)$$

$$= \sum_{j=m+1}^{n} \left[w_j^* \Gamma_X w_j - \lambda_j (w_j^* w_j - 1)\right]$$

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

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

$1 \le m \le n$ was arbitrary. So, for any $1 \le m \le n$, 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

{ neoclassical 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 $\delta_{\mathbb{X}} := \mathrm{diag}(\Gamma_{\mathbb{X}}) / \| \mathrm{diag}(\Gamma_{\mathbb{X}}) \|_1 \in \mathbb{R}^n$

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

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

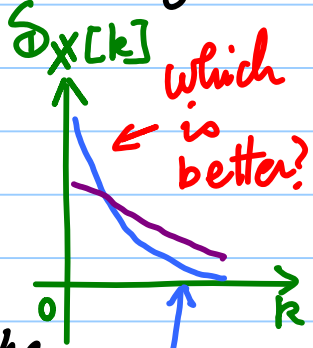
the Shannon entropy

$\delta_{\mathbb{X}}$ above qualifies as $\mathbb{P}$ thanks to the normalization.

Then Watanabe showed that

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

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




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) $\mathbb{Y} = W_{KL}^* \mathbb{X}$, $W_{KL} = [w_1 | \cdots | w_n]$ eigenvectors of $\Gamma_{\mathbb{X}}$.

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

$$\Gamma_{\mathbb{Y}} = E(\mathbb{Y} - E\mathbb{Y})(\mathbb{Y} - E\mathbb{Y})^*$$

$$= E\, W_{KL}^* (\mathbb{X} - E\mathbb{X})(\mathbb{X} - E\mathbb{X})^* W_{KL}$$

$$= W_{KL}^* E(\mathbb{X} - E\mathbb{X})(\mathbb{X} - E\mathbb{X})^* W_{KL}$$

$$= W_{KL}^* \Gamma_{\mathbb{X}} W_{KL} = \underbrace{\text{diag}(\lambda_1, \cdots, \lambda_n)}$$

$$\Gamma_{\mathbb{X}} w_j = \lambda_j w_j, \quad \lambda_j > 0.$$

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

$$E(Y_i - EY_i)\overline{(Y_j - EY_j)} = \lambda_i \delta_{ij}$$

## ☆ Relationship between KLT/PCA and SVD

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

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

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

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

$\Rightarrow$ The mean of the col. vectors of $\tilde{X} = \mathbb{O}$.

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

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

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

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

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \\ \hline & & & \mathbb{O} & & \end{bmatrix}$$

($n$ rows; $n$ columns, $N-n$ columns) or ($N$ rows, $n-N$ rows; $N$ columns)

$n < N$ classical

$n > N$ neoclassical

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

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

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

(1) If $n < N$, then

$$\hat{\Gamma}_{\tilde{X}} = \frac{1}{N}\tilde{X}\tilde{X}^* = \frac{1}{N}U\Sigma V^* V\Sigma^* U^*$$

$$= \frac{1}{N}U\Sigma\Sigma^* U^*$$

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

Hence, in this case

$$U = W_{KL} \text{ !}$$

(2) If $n > N$, then we should not compute all $n$ KLB vectors because $\text{rank}(\tilde{X}) \le N-1$, 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? $\hat{\Gamma}_{\tilde{X}} = \frac{1}{N}\tilde{X}\tilde{X}^*$

so, $\hat{\Gamma}_{\tilde{X}}\tilde{X}V = \frac{1}{N}\tilde{X}\tilde{X}^*\tilde{X}V$

$$= \frac{1}{N}\tilde{X}(V\Sigma^* U^* U\Sigma V^*)V$$

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

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

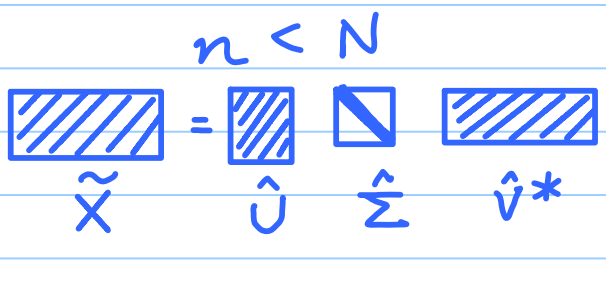
**Remark**: In either case ($n < N$ or $n > N$), you do not need to compute $\hat{\Gamma}_{\tilde{x}}$. Moreover, you should use the reduced SVD instead of the full SVD for the purpose of KLB/PCA computation.

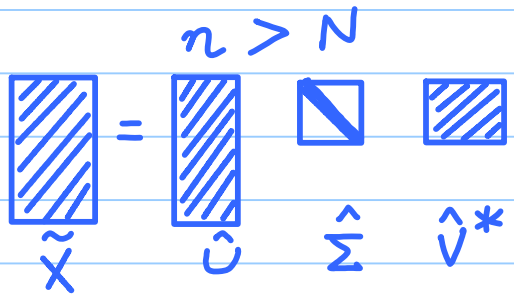
The reduced SVD of $\tilde{X} \in \mathbb{C}^{n \times N}$

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

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 MATLAB, this is done by

```
>> [Û, Σ̂, V̂] = svd(X̃, 'econ');
```

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

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

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

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

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

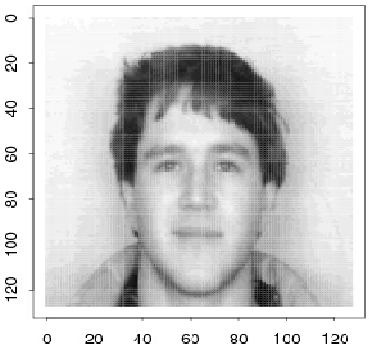
Example: The Rogues' Gallery Dataset

- 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.

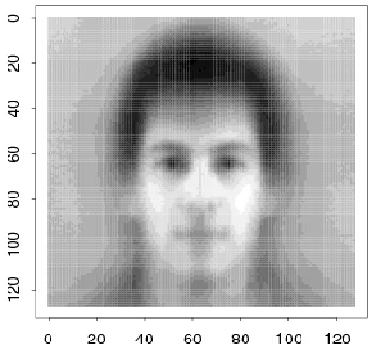
# 72 Faces !

$x_1$ $x_2$ - - -

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

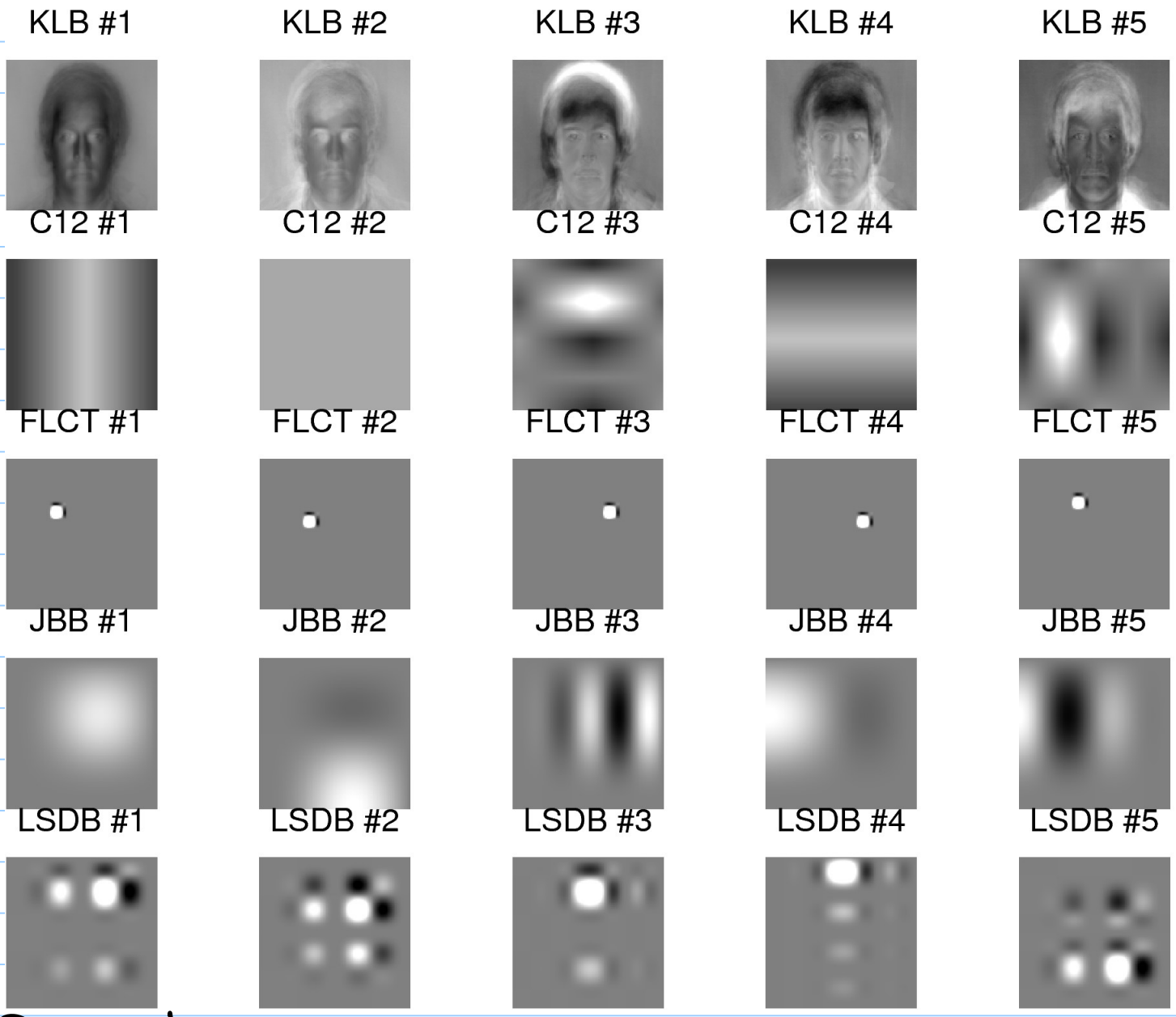



one of the test image




$\bar{x}$

# Comparison of Bases

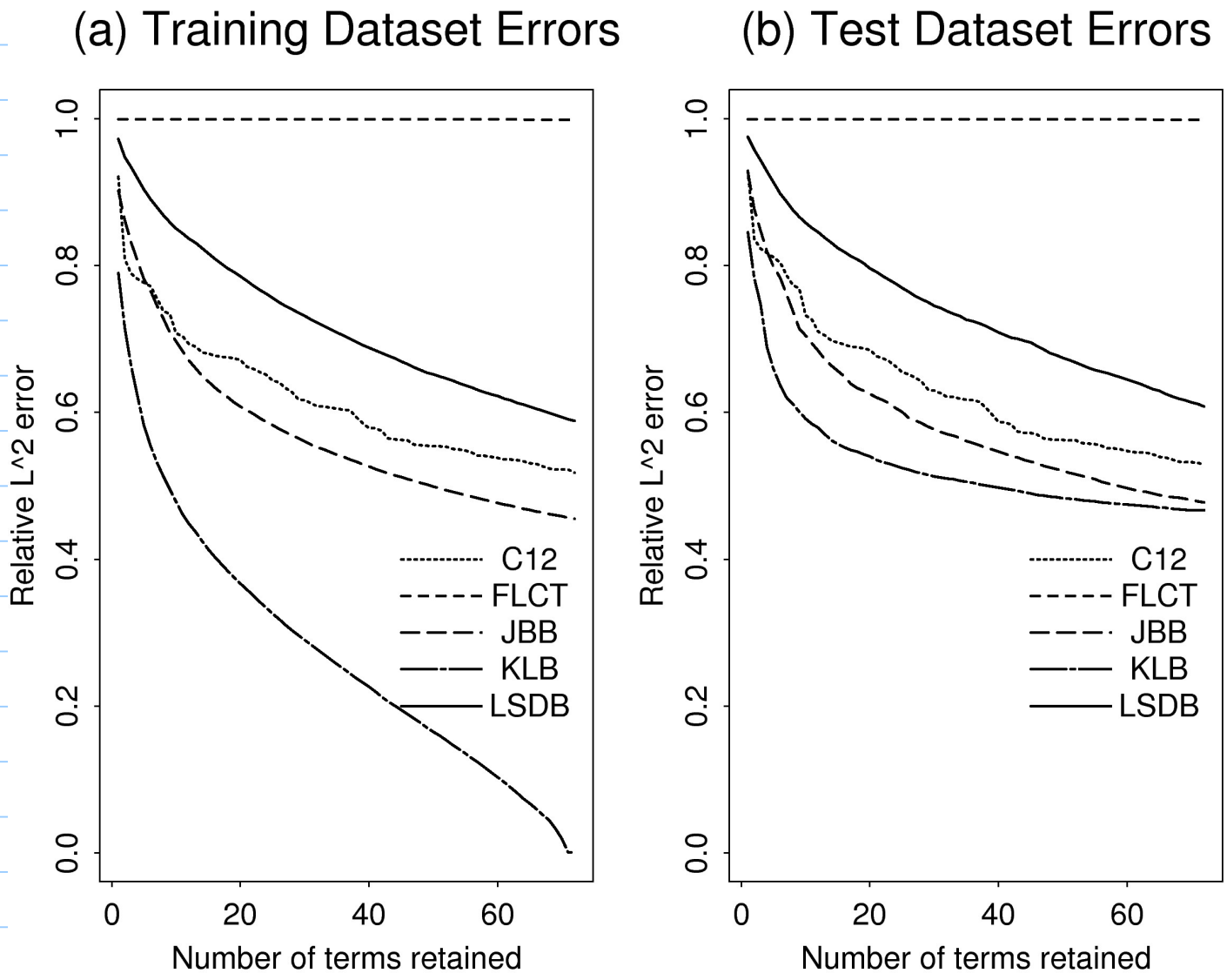



Remarks:

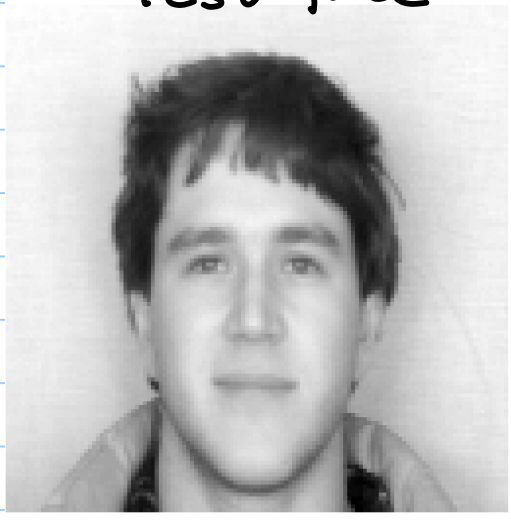
(1) I am only displaying the top 5 KLB vectors. There are totally 71 KLB vectors plus the mean vector $\bar{x}$.

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

# Approximation Errors




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.,

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

(appropriate mod N needs here)

Furthermore, assume

$$\Gamma_X[k,l] = \rho^{|k-l|} \qquad 0 < \rho < 1.$$

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. $E Z_k = 0$
$\operatorname{var} Z_k = 1$.

Then

$$E X_k X_{k+1}^* = \rho E|X_k|^2 + E X_k Z_k^* \quad (= E X_k E Z_k^*, \ E Z_k^* = 0)$$

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

$$= \rho E X_k X_{k+1}^* + E X_k Z_{k+1}^* \quad (E X_k Z_{k+1}^* = 0)$$

$$\vdots$$

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

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

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

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

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

For further simplicity, let's assume $\sigma^2 = 1$.
One can show that
The eigenvalues & eigenvectors of this Toeplitz matrix are:

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

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

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

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

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

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

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

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

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

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

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

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

$$\overset{(*)}{=} \frac{1+\rho}{1-\rho+\rho(1+\rho)/n}$$

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

So,
$$\begin{cases} w_0[l] = \frac{1}{\sqrt{n}} \\ w_k[l] = \sqrt{\frac{2}{n}}\cos\left[\frac{\pi k(l+\frac{1}{2})}{n}\right] \end{cases}$$

DCT-II !!