Lecture 2:

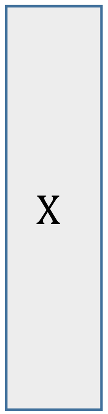
Linear Methods of Machine Learning:

1) Principal Component Analysis (PCA)

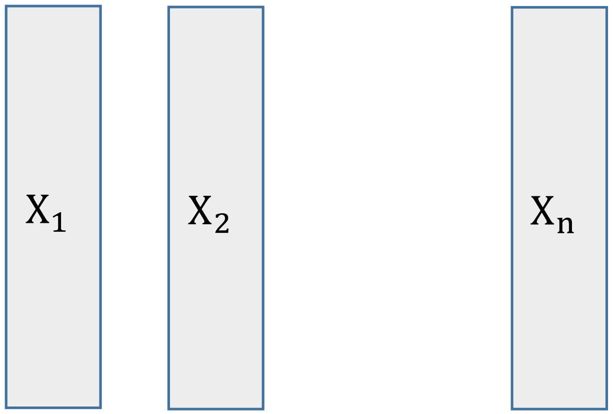
1. **Principal Component Analysis (PCA) - most popular linear data analysis technique**
2. **PCA as the best linear approximation**
3. **PCA as the best solution to linear dimensionality reduction problem**
4. **PCA from Singular Value Decomposition technique**
5. **PCA as the best solution to Metric Multi Dimensionality Scaling**
6. **PCA as maximum variance preserving technique**
7. **PCA: how many components should be left**
8. **PCA: example (human faces)**

Notations

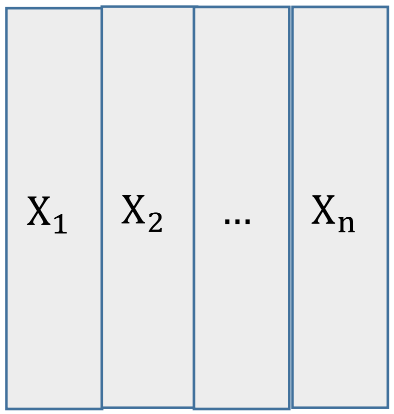
$X = \begin{pmatrix} x_1 \\ \cdots \\ x_p \end{pmatrix} \in R^p$ - p-dimensional vector with components $x_1, x_2, \ldots, x_p$



$\{X_1, X_2, \ldots, X_n\}$ – dataset, $X_i = \begin{pmatrix} x_{i1} \\ \cdots \\ x_{ip} \end{pmatrix}$, $i = 1, 2, \ldots, n$



described by $p \times n$ data matrix $\mathbf{X} = (X_1 \ldots X_n) = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ \cdots & \cdots & \cdots \\ x_{1p} & \cdots & x_{np} \end{pmatrix}$



Mean vector $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Centering dataset $\bar{X}_1 = X_1 - \bar{X},\ \bar{X}_2 = X_2 - \bar{X},\ \dots,\ \bar{X}_n = X_n - \bar{X}$

described by $p \times n$ centering data matrix $\bar{\mathbf{X}} = (\bar{X}_1\ \bar{X}_2\ \dots\ \bar{X}_n)$

$\bar{X}_1$	$\bar{X}_2$	...	$\bar{X}_n$

$\mathbf{H} = I_n - \frac{1}{n}\mathbf{1}\times\mathbf{1}^T$ - $n \times n$ centering matrix, $\mathbf{1}$ – n-dimensional vector of all 1's:

$$\mathbf{H} \times \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} = \begin{pmatrix} a_1 - \bar{a} \\ a_2 - \bar{a} \\ \dots \\ a_n - \bar{a} \end{pmatrix}, \qquad \bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$\bar{\mathbf{X}} = \mathbf{H}\times\mathbf{X}$ and $\{X_1, X_2, \dots, X_n\} \leftrightarrow (\bar{X}, \bar{\mathbf{X}})$

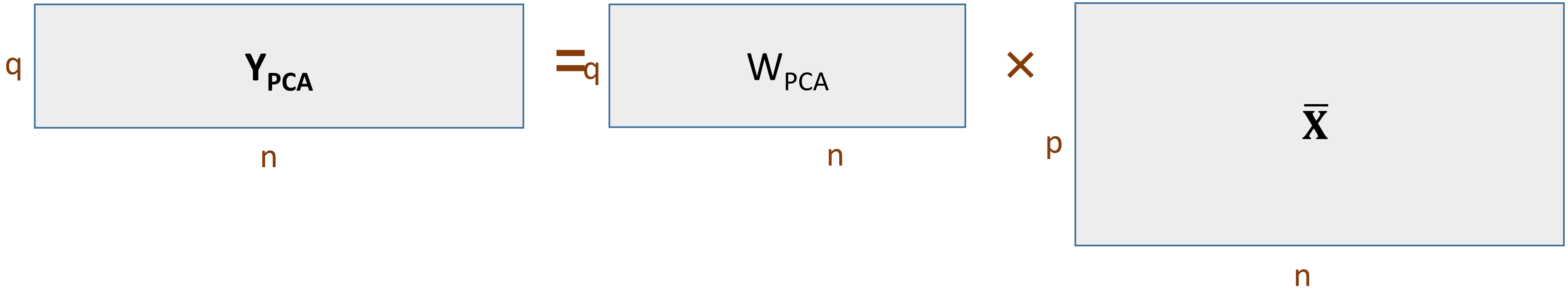
Principal Component Analysis (PCA):

Linear data analysis technique that transform a vector consisting of **possibly correlated variables** (original features) into a vector consisting of **low-dimensional uncorrelated features** called **principal components** with certain desired properties

$X \in R^p$ - original features $\rightarrow$ $Y_{PCA} = W_{PCA} \times (X - \bar{X}) \in R^q$ - reduced features

W_{PCA} - $q \times p$ - PCA-matrix

data $\{X_1, X_2, \ldots, X_n\}$ $\rightarrow$ reduced data matrix $\mathbf{Y}_{PCA} = (y_1\ y_2\ \ldots\ y_n)$



Most popular and best-known, one of the oldest (K. Pearson, 1901) data analysis method widely used **as preprocessing step** in multivariate analysis, machine learning, pattern recognition, data mining, image processing, visualization, etc.

Underlies the many **other linear and nonlinear** data analysis methods

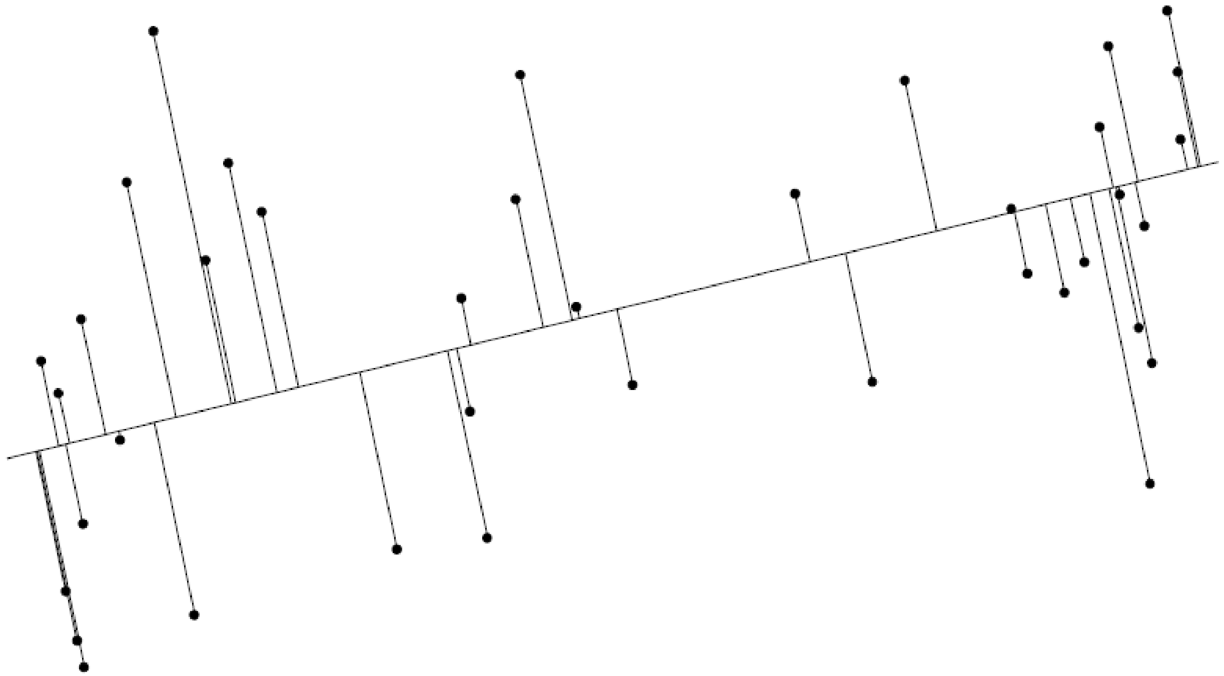
Various names:

- **discrete Karhunen-Lòeve transform** in stochastic analysis, image/video compression, and signal processing,
- **Hotelling transform** in multivariate quality control;
- **proper orthogonal decomposition** (POD) in mechanical engineering,
- **singular value decomposition** (SVD) or **eigenvalue decomposition** (EVD) in linear algebra, etc.

PCA can be justified in several ways

(from various points of view, a few various criteria leading to PCA)

1) The best linear approximation: to find low-dimensional linear affine subspace that best approximates given high-dimensional dataset

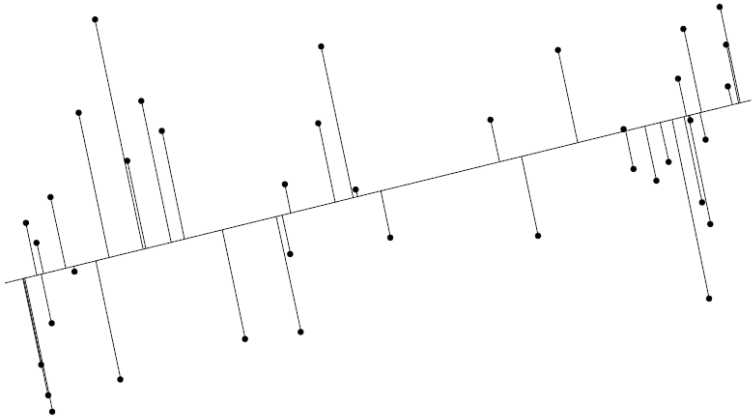


- $L(q)$ - desired q-dimensional linear affine subspace in R^p,

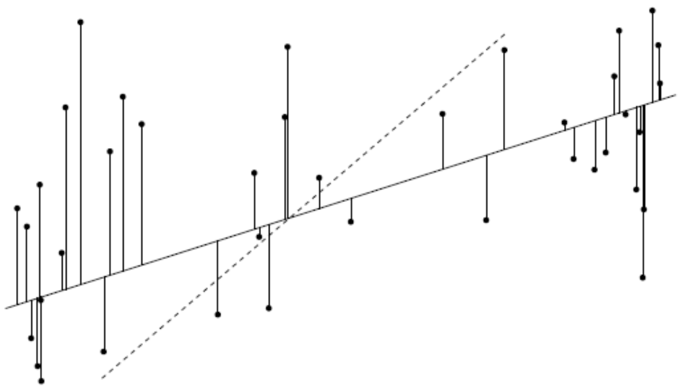
  for the time being, the dimension q is assumed to be chosen (given)

- $Pr_{L(q)}(X)$ - **orthogonal** projection of p-dimensional vector X into $L(q)$

Best subspace minimizes the objective function $J(L(q)) = \frac{1}{n}\sum_{i=1}^{n}\left\|X_i - Pr_{L(q)}(X_i)\right\|^2$



Not regression line



$$L(q) = L(q, X_0, E) = \{X_0 + \sum_{k=1}^{q} t_k \times e_k \in R^p : t = (t_1, t_2, \dots, t_q)^T \in R^q\} -$$

q-dimensional linear affine plane

- passing through a point $X_0 \in R^p$, and
- spanned by **orthonormal** vectors $\{e_1, e_2, \dots, e_q\} \subset R^p$
- $E = (e_1 \dots e_q)$ - orthogonal $p \times q$ matrix:

$E^T \times E = I_q$ - unit $q \times q$ matrix

$Pr_{L(q)}(X) = X_0 + \sum_{k=1}^{q} (X - X_0, e_k) \times e_k \in L(q, X_0, E)$ - orthogonal projection of vector X into $L(q)$

$$J(L(q, X_0, E)) = \frac{1}{n}\sum_{i=1}^{n}\left\|X_i - Pr_{L(q)}(X_i)\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|(X_i - X_0) - \sum_{k=1}^{q}(X_i - X_0, e_k) \times e_k\right\|^2$$

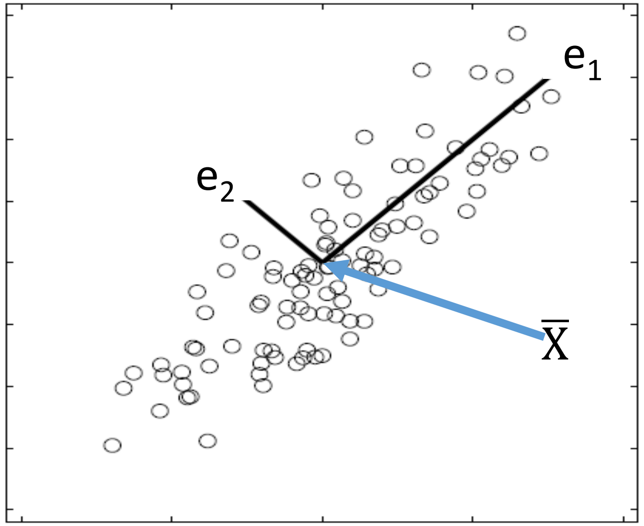
Let $\bar{X}$ - mean vector and $L^*(q) = L(q, \bar{X}, E)$ - affine subspace passing through the mean vector.

For any orthogonal $p \times q$ matrix E:

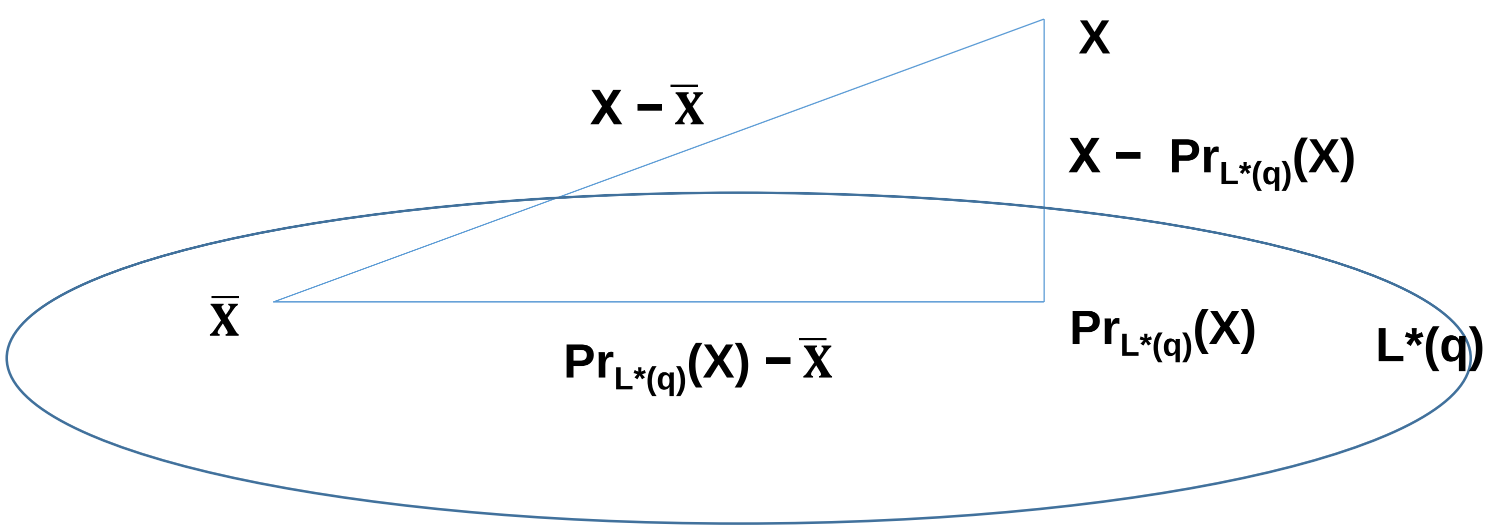
$$J(L(q, X_0, E)) = J(L(q, \bar{X}, E)) + \left\| X_0 - \Pr_{L^*(q)}(X_0) \right\|^2$$

Hence:

- X_0 must belong to the affine space $L(q, \bar{X}, E)$
- mean vector $\bar{X}$ can be taken as X_0



$L_{PCA}(q)$



$$X - Pr_{L^*(q)}(X) = (X - \overline{X}) - (Pr_{L^*(q)}(X) - \overline{X}) \text{ and } X - Pr_{L^*(q)}(X) \perp Pr_{L(q)}(X) - \overline{X})$$

$$\left\|X - Pr_{L^*(q)}(X)\right\|^2 = \|X - \overline{X}\|^2 - \left\|\overline{X} - Pr_{L^*(q)}(X)\right\|^2$$

$$J(L(q, \overline{X}, E)) = \frac{1}{n}\sum_{i=1}^{n}\left\|X_i - Pr_{L^*(q)}(X_i)\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|(X_i - \overline{X}) - \sum_{k=1}^{q}(X_i - \overline{X}, e_k) \times e_k\right\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2 - \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{k=1}^{q}(X_i - \overline{X}, e_k) \times e_k\right\|^2$$

Orthogonal $p \times q$ matrix $E = (e_1 \dots e_q)$ must maximize quadratic form

$$\Phi(E) = \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{k=1}^{q}(X_i - \overline{X}, e_k) \times e_k\right\|^2$$

using an orthonormality of $\{e_1, e_2, \dots, e_q\}$:

$$\Phi(E) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}|(X_i - \overline{X}, e_k)|^2$$

using a representation $(u, v) = u^T \times v = v^T \times u$:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\left(e_k^T \times (X_i - \overline{X})\right) \times \left((X_i - \overline{X})^T \times e_k\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q} e_k^T \times [(X_i - \overline{X}) \times (X_i - \overline{X})^T] \times e_k$$

changing the order of summation:

$$\Phi(E) = \frac{1}{n}\sum_{k=1}^{q} e_k^T \times \left(\sum_{i=1}^{n}(X_i - \overline{X}) \times (X_i - \overline{X})^T\right) \times e_k$$

$$= \mathrm{Tr}\left(E^T \times \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \times (X_i - \overline{X})^T\right) \times E\right)$$

- X - p-dimensional random vector
- $Cov(X) = \mathbf{M}(X - \mathbf{M}X)\times(X - \mathbf{M}X)^T$ – $p\times p$ covariance matrix
- $\{X_1, X_2, \ldots, X_n\}$ – i.i.d. (sample)
- $\overline{X}$ - sample mean (estimator of $\mathbf{M}X$)
- $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \times (X_i - \overline{X})^T$ – sample covariance matrix - estimator of $Cov(X)$

(1) Optimizing problem: maximize the quadratic form

$$\Phi(e_1, e_2, \dots , e_q) = \sum_{k=1}^{q} e_k^T \times \Sigma \times e_k$$

over vectors $e_1, e_2, \dots , e_q \in R^p$ under ‘orthonormality’ constraints

$$(e_i, e_j) = \delta_{ij}, \quad 1 \le i \le j \le n$$ (δ_{ij} - Kronecker symbol)

(2) – **Eigenvector problem**: maximize the quadratic form

$$\Phi(E) = Tr(E^T \times \Sigma \times E)$$

over $p \times q$ matrix E under constraint $E^T \times E = I_q$

Denote $St(p, q)$ – a set consisting of orthogonal $p \times q$ matrices – **Stiefel manifold**

(3) Optimizing problem: to maximize the quadratic form $\Phi(E) = Tr(E^T \times \Sigma \times E)$ over $E \in St(p, q)$

Solution to the Eigenvector problem: the columns $e_1, e_2, \dots, e_q \in R^p$ of the desired matrix E

- are p-dimensional eigenvectors of $p \times p$ matrix Σ:

$$\Sigma e_k = \lambda_k \times e_k, \quad k = 1, 2, \dots, q$$

- corresponding to q largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$ of the matrix Σ, respectively.

The best approximating affine subspace: $L_{PCA}(q) = L(q, \overline{X}, E_{PCA})$

$$Pr_{PCA}(X) = \overline{X} + \sum_{k=1}^{q} y_k \times e_k \in L_{PCA}(q), \qquad Y_{PCA} = (y_1, y_2, \dots, y_q)^T = (E_{PCA})^T \times (X - \overline{X}) \in R^q$$

- orthogonal projection $Pr_{PCA}(X)$ of p-dimensional vector X into $L_{PCA}(q)$

$W_{PCA} = (E_{PCA})^T$: defines

- best approximating affine subspace: $L_{PCA}(q) = L(q, \overline{X}, (W_{PCA})^T)$
- PCA-features $Y_{PCA} = W_{PCA} \times (X - \overline{X})$

‘Direct’ solution to the optimizing problem

(1) maximize the quadratic form

$$\Phi_n(e_1, e_2, \ldots, e_q) = \sum_{k=1}^{q} e_k^T \times \Sigma \times e_k$$

over vectors $\{e_1, e_2, \ldots, e_q\}$ under ‘orthonormality’ constraints $(e_i, e_j) = \delta_{ij}$, $1 \leq i \leq j \leq n$

Step 1. Maximizing the first summand $(e_1)^T\times\Sigma\times e_1$ over e_1 subject to $(e_1)^T\times e_1 = 1$

Lagrange multipliers: maximize

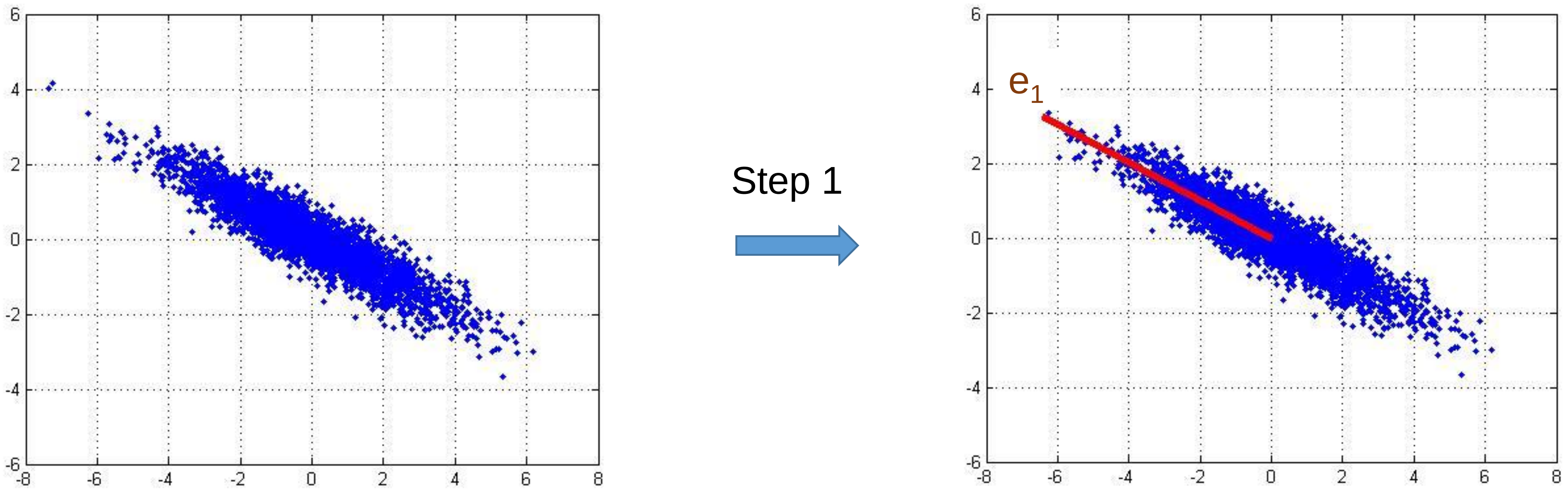
$$(e_1)^T\times\Sigma\times e_1 + \lambda\times(1 - (e_1)^T\times e_1)$$

Setting the gradient $\nabla = 2\Sigma\times e_1 - 2\lambda\times e_1$ w.r.t. e_1 to zero, we find that

$\Sigma\times e_1 = \lambda\times e_1$ $\rightarrow$ e_1 – eigenvector, λ - eigenvalue

$$(e_1)^T\times\Sigma\times e_1 = (e_1)^T\times\lambda\times e_1 = \lambda\times(e_1\times (e_1)^T) = \lambda \in \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$$

$$\rightarrow \quad \max((e_1)^T\times\Sigma\times e_1) = \lambda_{max} = \lambda_1$$



1^{st} PCA axis e_1

Step 2. Maximizing second summand $(e_2)^T \times \Sigma \times e_2$ over e_2 subject to $(e_2)^T \times e_2 = 1$ and $(e_2)^T \times e_1 = 0$

Lagrange multipliers: maximize

$$(e_2)^T \times \Sigma \times e_2 + \lambda \times (1 - (e_2)^T \times e_{12}) + b \times (e_2)^T \times e_1$$

the gradient w.r.t. $e_2 = 0$: $\rightarrow$ $\Sigma \times e_2 - \lambda \times e_2 + b \times e_1 = 0$ (*)

$(e_2)^T \times e_1 = 0$: $\rightarrow$ $(e_2)^T \times \Sigma \times e_1 = (e_2)^T \times (\lambda \times e_1) = \lambda_1 \times (e_2)^T \times e_1 = 0$

$\rightarrow$ $(e_1)^T \times \Sigma \times e_2 = (e_2)^T \times \Sigma \times e_1 = 0$

a multiplication of equation (*) on the left by $(e_1)^T$ gives

$$(e_1)^T \times (\Sigma \times e_2 - \lambda \times e_2 + b \times e_1) = (e_1)^T \times \Sigma \times e_2 – \lambda \times ((e_1)^T \times e_2) + b \times ((e_1)^T \times e_1) = 0,$$

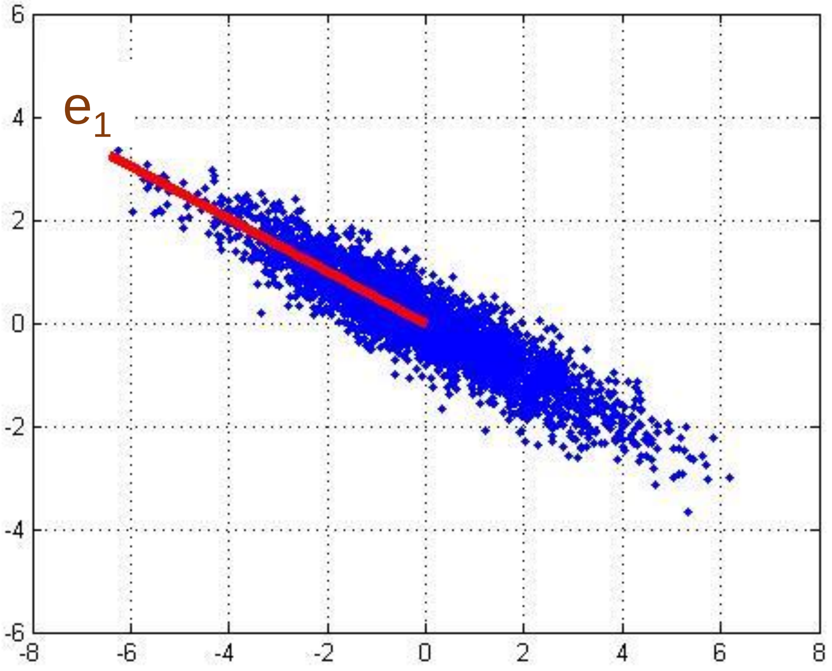
hence, $b = 0$ and (*) yields

$\Sigma \times e_2 = \lambda \times e_2$ (e_2 – eigenvector, λ - eigenvalue) and $(e_2)^T \times \Sigma \times e_2 = \lambda$

Let $u_1 = e_1, u_2, \ldots, u_p \in R^p$ be all the eigenvectors of the $p \times p$ matrix Σ corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ of the matrix Σ, respectively.

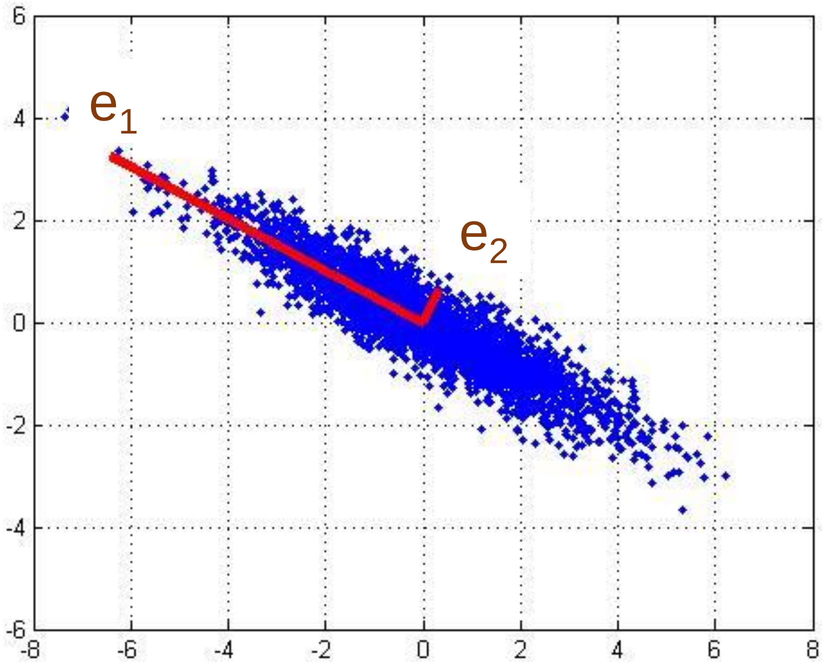
Thus, we must choose e_2 among the vectors $\{u_2, u_3, \ldots, u_p\}$ and

$$\max_{2 \leq k \leq p}((u_k)^T \times \Sigma \times u_k) = \max_{2 \leq k \leq p} \lambda_k = \lambda_2$$



1^{st} PCA axis e_1

Step 2



2^{nd} PCA axis e_2

… and in like manner vectors $e_3, e_4, \ldots, e_q$ are sought

Solution to the Eigenvector problem:

- vectors $e_1, e_2, \ldots, e_q \in R^p$ which are the p-dimensional eigenvectors of $p \times p$ matrix Σ:

$$\Sigma e_k = \lambda_k \times e_k, \quad k = 1, 2, \ldots, q$$

- corresponding to q largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$ of the matrix Σ, respectively,

maximize the objective function $\Phi(e_1, e_2, \ldots, e_q) = \sum_{k=1}^{q} e_k^T \times \Sigma \times e_k$ and

$$\max\left(\sum_{k=1}^{q} e_k^T \times \Sigma \times e_k\right) = \sum_{k=1}^{q} \lambda_k$$

2) The best solution of linear dimensionality reduction problem:

to find low-dimensional features which provide minimal recovering error

Dimensionality reduction

Reduced q-dimensional features $\{y_1, y_2, \ldots, y_n\}$ are the results of applying of **Embedding mapping**

$$h: X \in R^p \to y = h(X) \in R^q$$

to original p-dimensional features $\{X_1, X_2, \ldots, X_n\}$

Low-dimensional reduced features must **preserve as much as possible available information** contained in high-dimensional original features: the possibility for **accurate recovery** of the original vectors from their low-dimensional features

Preserving information - the possibility for **recovery** of original vectors from reduced low-dimensional features **with small recovering error**

- an existence of a recovering mapping

$$g: y = h(X) \in R^q \rightarrow g(y) \in R^p$$

from q-dimensional Reduced feature space to p-dimensional Original feature space

- a recovered value

$$y = h(X) \in R^q \rightarrow \widehat{X} = g(y) = g(h(X))$$

such that a recovering error $\Delta(X) = \left\|\widehat{X} - X\right\|$ is small

Minimum average recovering error: $\Delta(W, V) = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{X}_i - X_i\right|^2\right)^{1/2}$ $\rightarrow$ min over h, g

Linear Dimensionality reduction

Linear embedding mapping h is defined by **orthogonal** $q \times p$ matrix W $W \times W^T = I_q$:

$$y_i = W \times (X_i - \overline{X}) \in R^q,\ i = 1, 2, \dots, n$$

Linear recovering mapping is defined by $p \times q$ matrix V and gives recovered values

$$\widehat{X}_i = \overline{X} + V \times y_i,\ i = 1, 2, \dots, n$$

Minimum average recovering error: $\Delta(W, V) = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{X}_i - X_i\right|^2\right)^{1/2}$ $\rightarrow$ min over W, V

Under chosen $q \times p$ matrix W, the best recovering mapping $V = V(W)$ minimizes the recovering error

$$\|\widehat{X} - X\| = \|(X - \overline{X}) - V \times W \times (X - \overline{X})\|$$

$V(W) = \arg\min_V \|\widehat{X} - X\| = W^+$ - $p \times q$ (left) pseudoinverse Moore-Penrose matrix to $q \times p$ matrix W

Due orthogonality of W

$$W^+ = (W^T \times W)^{-1} \times W^T = W^T$$

Recovered value: $\widehat{X}(W) = \overline{X} + W^T \times W \times (X - \overline{X})$

Squared averaged recovering error:

$$\Delta^2(W) = \frac{1}{n} \sum_{i=1}^{n} \left| \widehat{X}_i(W) - X_i \right|^2 \to \min$$

Squared recovering error:

$$\Delta^2(W) = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{X}_i(W) - X_i\right|^2 = \frac{1}{n}\sum_{i=1}^{n}\left|\overline{X} + W^T \times W \times (X_i - \overline{X}) - X_i\right|^2$$

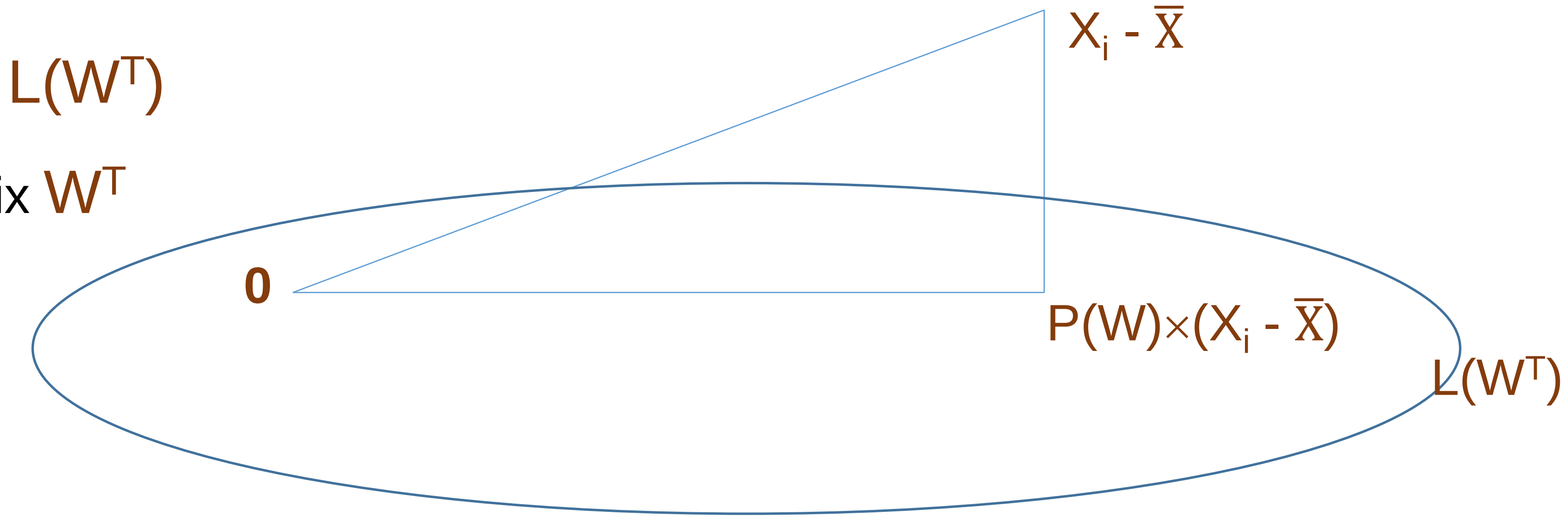
Denote $P(W) = W^T \times W$ - $p \times p$ matrix

$$\Delta^2(W) = \frac{1}{n}\sum_{i=1}^{n}\left|(X_i - \overline{X}) - P(W) \times (X_i - \overline{X})\right|^2$$

$L(W^T)$ - q-dimensional linear space $L(W^T)$

spanned by q columns of $p\times q$ matrix W^T

$P(W)$ - projection matrix into $L(W^T)$



$$\Delta^2(W) = \frac{1}{n}\sum_{i=1}^{n}|(X_i - \bar{X}) - P(W)\times(X_i - \bar{X})|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \bar{X}\|^2 - \frac{1}{n}\sum_{i=1}^{n}|P(W)\times(X_i - \bar{X})|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \bar{X}\|^2 - \frac{1}{n}\sum_{i=1}^{n}\{(X_i - \bar{X})^T\times[P^T(W)\times P(W)]\times(X_i - \bar{X})\}$$

$$\boxed{P^T(W) = P^2(W) = P(W)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \bar{X}\|^2 - \frac{1}{n}\sum_{i=1}^{n}\{(X_i - \bar{X})^T\times[W^T\times W]\times(X_i - \bar{X})\}$$

number

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2 - \frac{1}{n}\sum_{i=1}^{n}\{(X_i - \overline{X})^T \times [W^T \times W] \times (X_i - \overline{X})\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}\left((X_i - \overline{X})^T \times \left(W^T \times W\right) \times (X_i - \overline{X})\right)$$

Tr(A×A^T) = Tr(A^T×A)

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}\left(W \times (X_i - \overline{X}) \times (X_i - \overline{X})^T \times W^T\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2 - \mathrm{Tr}\left(W \times \frac{1}{n}\sum_{i=1}^{n}\left((X_i - \overline{X}) \times (X_i - \overline{X})^T\right) \times W^T\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2 - \mathrm{Tr}(W \times \Sigma \times W^T)$$

$$\Delta^2(W) = \frac{1}{n}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2 - Tr(W \times \Sigma \times W^T) \rightarrow \min$$

$$Tr(W \times \Sigma \times W^T) \rightarrow \max$$

$E = W^T$: $\quad Tr(E^T \times \Sigma \times E) \rightarrow \max$, E – orthogonal matrix

$$E_{PCA} = \arg\max_E Tr(E^T \times \Sigma \times E)$$

PCA-matrix $W_{PCA} = E_{PCA}^T$ minimizes the recovering error $\Delta(W)$

PCA as Dimensionality reduction

- Linear embedding mapping defined by $q \times p$ matrix W_{PCA}:

$$y_i = h(X_i) = W_{PCA} \times (X_i - \overline{X}),\ i = 1, 2, \ldots, n$$

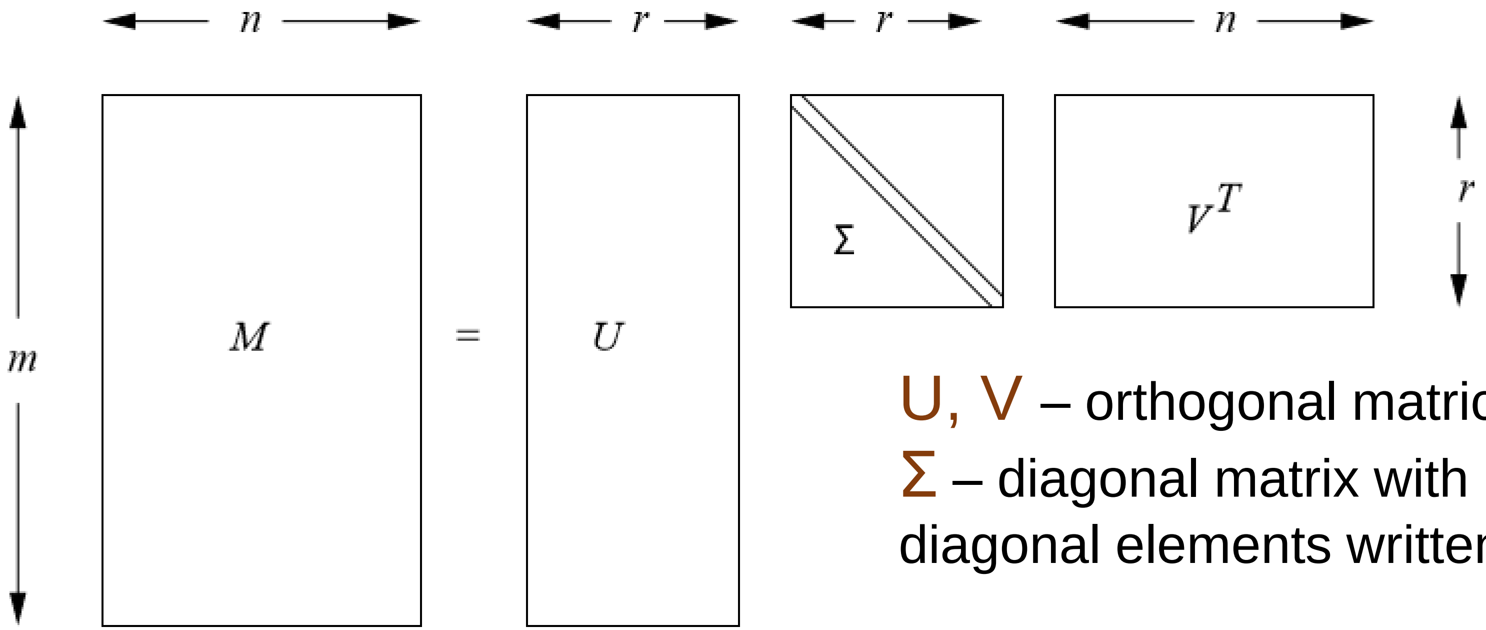
- Linear recovering mapping defined by $p \times q$ matrix $(W_{PCA})^T$: gives recovered values

$$\widehat{X}_{PCA,i} = \overline{X} + W_{PCA}^T \times y_i,\ i = 1, 2, \ldots, n$$

- PCA averaged recovering error:

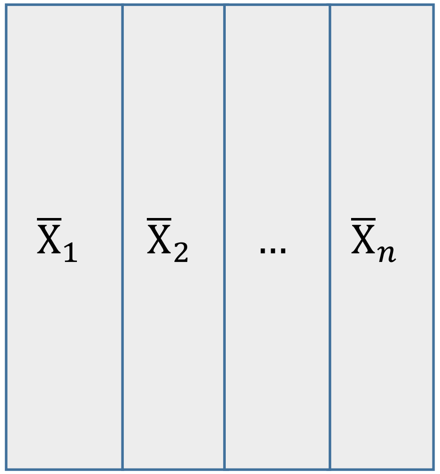
$$\Delta_{PCA}^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{X}_{PCA,i} - X_i\right|^2\right)^{1/2} = \sum_{k=1}^{p} \lambda_k - \sum_{k=1}^{q} \lambda_k = \sum_{k=q+1}^{p} \lambda_k$$

3) PCA from Singular Value Decomposition (SVD)



U, V – orthogonal matrices

Σ – diagonal matrix with non-negative diagonal elements written in descending order



$\bar{\mathbf{X}} = (\bar{X}_1\ \bar{X}_2\ \ldots\ \bar{X}_n)$ - p×n centering data matrix

Singular Value Decomposition: $\overline{\mathbf{X}} = \mathbf{U}_p \times \boldsymbol{\Lambda}_p \times \mathbf{V}_p^T$

$$\underset{p \times n}{\overline{\mathbf{X}}} = \begin{bmatrix} \overline{X}_1 & \overline{X}_2 & \cdots & \overline{X}_n \end{bmatrix} = \underset{p \times p}{\mathbf{U}_p} = \begin{bmatrix} e_1 & e_2 & \cdots & e_p \end{bmatrix} \times \boldsymbol{\Lambda}_p = \begin{pmatrix} \lambda_1 & \cdots & 0 & & & \\ \cdots & \cdots & \cdots & & \mathbf{0} & \\ 0 & \cdots & \lambda_q & & & \\ & & & \lambda_{q+1} & \cdots & 0 \\ & \mathbf{0} & & \cdots & \cdots & \cdots \\ & & & 0 & \cdots & \lambda_p \end{pmatrix} \times \underset{p \times n}{\mathbf{V}_p^T} = \begin{bmatrix} V_1^{(p)} & V_2^{(p)} & \cdots & V_1^{(p)} \end{bmatrix}$$

- $\mathbf{U}_p$ - $p \times p$ orthogonal matrix with p orthonormal eigenvectors $e_1, e_2, \ldots, e_p \in R^p$ of $p \times p$ matrix

$$\overline{\mathbf{X}} \times \overline{\mathbf{X}} = \sum_{i=1}^{n} \left((X_i - \overline{X}) \times (X_i - \overline{X})^T \right) = n \times \Sigma$$

corresponding to eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ of the matrix Σ, respectively, as columns

- $\boldsymbol{\Lambda}_p$ - $p \times p$ diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_p$
- $\mathbf{V}_p$ - $n \times p$ orthogonal matrix: $\mathbf{V}_p^T \times \mathbf{V}_p = I_p$

$\mathbf{U}_p$ - orthogonal matrix $\qquad \bar{\mathbf{X}} = \mathbf{U}_p \times \boldsymbol{\Lambda}_p \times \mathbf{V}_p^T \quad \rightarrow \quad \mathbf{U}_p^T \times \bar{\mathbf{X}} = \boldsymbol{\Lambda}_p \times \mathbf{V}_p^T$

Write in ‘column form’ for all columns

$$Z_{p,i} = \begin{pmatrix} z_{i1} \\ \cdots \\ z_{ip} \end{pmatrix} = \boldsymbol{\Lambda}_p \times \mathbf{V}_i^{(p)} = \mathbf{U}_p^T \times (X_i - \bar{X}) \in R^p,\ i = 1, 2, \dots, n$$

new ‘transformed features’

New p-dimensional features $\{Z_{p,1}, Z_{p,2}, \dots, Z_{p,n}\}$ are ‘the same’ centered features $\{\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n\}$ but written in other coordinate system in R^p defined by orthonormal vectors $e_1, e_2, \dots, e_p \in R^p$

$$Z_{p,i} = \begin{pmatrix} z_{i1} \\ \cdots \\ z_{ip} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{i1} \\ \mathbf{Z}_{i2} \end{pmatrix} \in R^p, \qquad \mathbf{Z}_{i1} = \begin{pmatrix} z_{i1} \\ \cdots \\ z_{iq} \end{pmatrix} \in R^q, \ \mathbf{Z}_{i2} = \begin{pmatrix} z_{i,q+1} \\ \cdots \\ z_{ip} \end{pmatrix} \in R^{p-q}$$

$$\Lambda_p = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_p \end{pmatrix} = \begin{pmatrix} \Lambda_q & \mathbf{0} \\ \mathbf{0} & \Lambda_{p-q} \end{pmatrix}$$

$$\Lambda_q = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_q \end{pmatrix}$$ - $q \times q$ diagonal matrix

$$\Lambda_{p-q} = \begin{pmatrix} \lambda_{q+1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_p \end{pmatrix}$$ - $(p-q) \times (p-q)$ diagonal matrix

$$\mathbf{Z}_{i1} = (\Lambda_q \quad \mathbf{0}) \times \mathbf{V}_i^{(p)} \qquad \mathbf{Z}_{i2} = (\mathbf{0} \quad \Lambda_{p-q}) \times \mathbf{V}_i^{(p)}$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

Let last $(p - q)$ eigenvalues $\lambda_{q+1}, \lambda_{q+2}, \ldots, \lambda_p \approx 0$ - are **small**

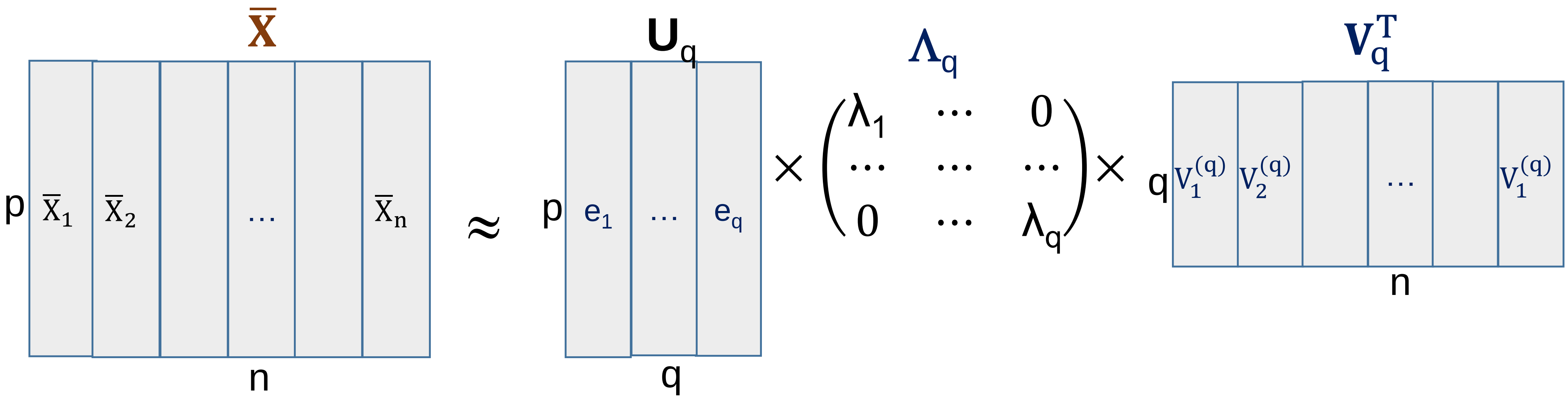
$$\Lambda_{p-q} = \begin{pmatrix} \lambda_{q+1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_p \end{pmatrix} \approx \mathbf{0}$$

$$\mathbf{Z}_{i2} = (\mathbf{0} \quad \Lambda_{p-q}) \times \mathbf{V}_i^{(p)} \approx \mathbf{0}$$

Transformed features: $Z_{p,i} = \begin{pmatrix} \mathbf{Z}_{i1} \\ \mathbf{Z}_{i2} \end{pmatrix} \approx \begin{pmatrix} \mathbf{Z}_{i1} \\ \mathbf{0} \end{pmatrix}, \quad i = 1, 2, \ldots, n$

Thus, the features $\left\{\mathbf{Z}_{i2} = \begin{pmatrix} z_{i,q+1} \\ \cdots \\ z_{ip} \end{pmatrix} \in R^{p-q}\right\}$ are **irrelevant** and **can be removed**

$$\bar{\mathbf{X}} \approx \mathbf{U}_q \times \boldsymbol{\Lambda}_q \times \mathbf{V}_q^T$$



- $\mathbf{U}_q$ - $p \times q$ orthogonal matrix spanned by eigenvectors $e_1, e_2, \ldots, e_q \in R^p$ – first q columns of $\mathbf{U}_p$,
- Λ_q - $q \times q$ diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_q$
- $\mathbf{V}_q$ - $n \times q$ orthogonal matrix: $\mathbf{V}_q^T \times \mathbf{V}_q = I_q$ which determines q-dimensional reduced features:

$$Z_{q,i} = \mathbf{Z}_{i,1} = \begin{pmatrix} z_{i1} \\ \cdots \\ z_{iq} \end{pmatrix} = \Lambda_q \times \mathbf{V}_i^{(q)} = \mathbf{U}_q^T \times (X_i - \bar{X}) \in R^p, \ i = 1, 2, \ldots, n$$

Orthonormal vectors $e_1, e_2, \dots, e_q$ - p-dimensional eigenvectors of $p \times p$ matrix Σ corresponding to q largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$ of the matrix Σ

$Z_{q,i}$ – reduced q-dimensional 'decorrelated' PCA-features

$\widehat{X}_i = \overline{X} + Z_{q,i} = \widehat{X}_{PCA,i}$ - 'PCA-reconstructed' original features

4) best solution to Metric Multi Dimensionality Scaling (Metric MDS):

to find low-dimensional features which preserve the Euclidean distances

Multi Dimensionality Scaling

- $\{X_1, X_2, \ldots, X_n\}$ – p-dimensional original features
- $\{y_1, y_2, \ldots, y_n\}$ – q-dimensional reduced features
- $\delta_p(X, X')$ and $\delta_q(y, y')$ – chosen distance (or dissimilarity) functions in original feature space R^p and reduced feature space R^q, respectively

MDS: to find reduced features $\{y_1, y_2, \ldots, y_n\}$ which preserve chosen distances - to minimize

$$\sum_{i,j=1}^{n} r_{ij} \left(\delta_p(X_i, X_j) - \delta_q(y_i, y_j) \right)^2 , \qquad \{r_{ij}\} \text{ – some chosen weights}$$

Example: images $X_1, X_2, \ldots, X_n$, $\delta_p(X, X')$ – chosen specific 'dissimilarity' function - 'distance'

Visualization: transform the images to 2 (or 3)-dimensional points $\{y_1, y_2, \ldots, y_n\}$

Euclidean distances $\{D_q(y_i\ y_j) = |y_i - y_j|\}$ between the 'visualized' images should preserve as much as possible distances $\{\delta_q(X_i\ X_j)\}$ between the original images

Metric Multi Dimensionality Scaling (1)

- Euclidean distances

- linear feature transform (embedding) defined by **orthogonal** $q \times p$ matrix W

$$y_i = W \times (X_i - \bar{X}) \in R^q, \quad i = 1, 2, \ldots, n$$

- a preserving the **Averaged Pairwise** Euclidean **Distances (APD)**:

$$\Delta_{APD}(W) = \sum_{i,j=1}^{n} \left(\left\| X_i - X_j \right\|^2 - \left\| y_i - y_j \right\|^2 \right)^2 \to \min$$

$\mathbf{D}_X$ - $n \times n$ Euclidean distance matrix with elements $\|X_i - X_j\|^2$, i, j = 1, 2, … , n

$\mathbf{S}_X$ - $n \times n$ matrix with 'inner product' elements $S_{ij} = (X_i - \bar{X}, X_j - \bar{X})$, i, j = 1, 2, … , n

$$\|X_i - X_j\|^2 = S_{ii} + S_{jj} - 2S_{ij} \qquad \rightarrow \qquad S_{ij} = -\frac{1}{2}\|X_i - X_j\|^2 - S_{ii} - S_{jj}$$

$\mathbf{H} = I_n - \frac{1}{n}\mathbf{1}\times\mathbf{1}^T$ - $n \times n$ centering matrix, $\mathbf{1}$ – n-dimensional vector of all 1's:

$$\mathbf{H} \times \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} = \begin{pmatrix} a_1 - \bar{a} \\ a_2 - \bar{a} \\ \dots \\ a_n - \bar{a} \end{pmatrix}, \qquad \bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$\mathbf{H}\times\mathbf{D}_X\times\mathbf{H}$:

- subtracts from each entry of $\mathbf{D}_X$ the means of the corresponding row and column
- and adds back the mean of all entries of $\mathbf{D}_X$

$$\boxed{\mathbf{S}_X = -\frac{1}{2}\mathbf{H}\times\mathbf{D}_X\times\mathbf{H}}$$

$\mathbf{S}_Y$ - $n \times n$ matrix with elements (y_i, y_j), $i, j = 1, 2, \dots, n$

$\mathbf{D}_Y$ - $n \times n$ Euclidean distance matrix with elements $\|y_i - y_j\|^2$, $i, j = 1, 2, \dots, n$

$$\boxed{\mathbf{S}_Y = -\frac{1}{2}\mathbf{H} \times \mathbf{D}_y \times \mathbf{H}}$$

$$\Delta_{APD} = \sum_{i,j=1}^{n} \left(\|X_i - X_j\|^2 - \|y_i - y_j\|^2\right)^2 = \|\mathbf{D}_X - \mathbf{D}_Y\|_F^2$$

$\mathbf{A} = (a_{ij})$: $\|\mathbf{A}\|_F^2 = \sum_{i,j} a_{ij}^2$ - Frobenius matrix norm

$$\mathbf{S}_X - \mathbf{S}_Y = -\frac{1}{2}\mathbf{H} \times (\mathbf{D}_X - \mathbf{D}_Y) \times \mathbf{H}$$

$$\boxed{\mathbf{S}_X - \mathbf{S}_Y = -\frac{1}{2}\mathbf{H} \times (\mathbf{D}_X - \mathbf{D}_Y) \times \mathbf{H}}$$

Metric Multi Dimensionality Scaling (2)

Minimizing the averaged pairwise distances

$$\Delta_{APD} = \sum_{i,j=1}^{n}\left(\left\|X_i - X_j\right\|^2 - \left\|y_i - y_j\right\|^2\right)^2 = \|\mathbf{D}_X - \mathbf{D}_Y\|_F^2$$

is reduced to Metric Multi Dimensionality Scaling Problem: to minimize

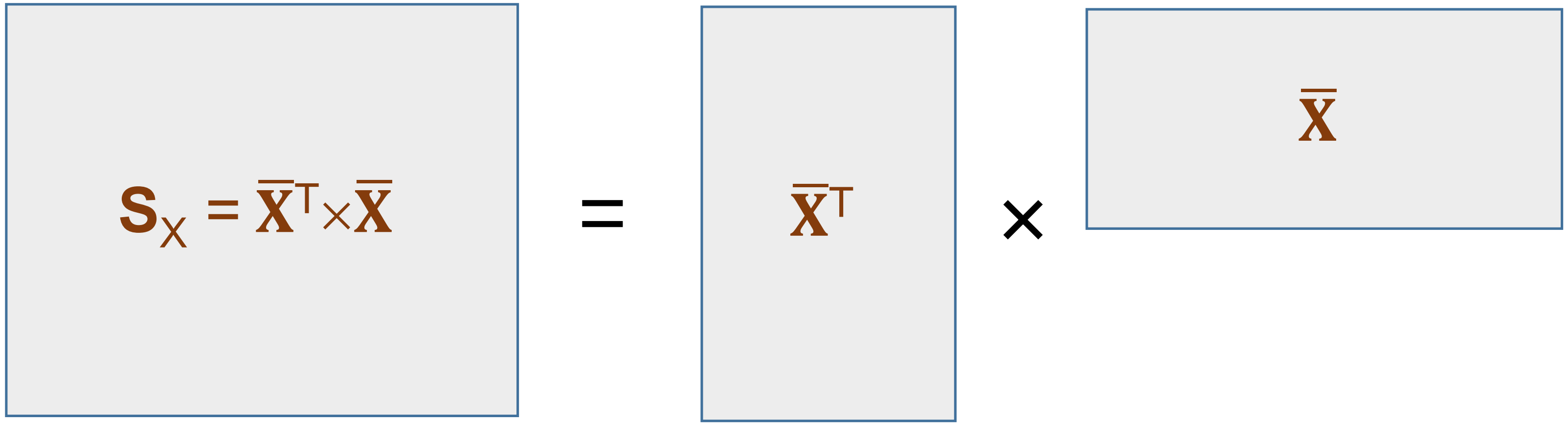
$$\Delta_{MDS} = \|\mathbf{S}_X - \mathbf{S}_Y\|_F^2 = \sum_{i,j=1}^{n}((X_i - \bar{X}, X_j - \bar{X}) - (y_i, y_j))^2$$

Metric Multi Dimensionality Scaling: the best preserving the inner products of data (centering original features and reduced features) in Original and Reduced feature spaces - minimizing

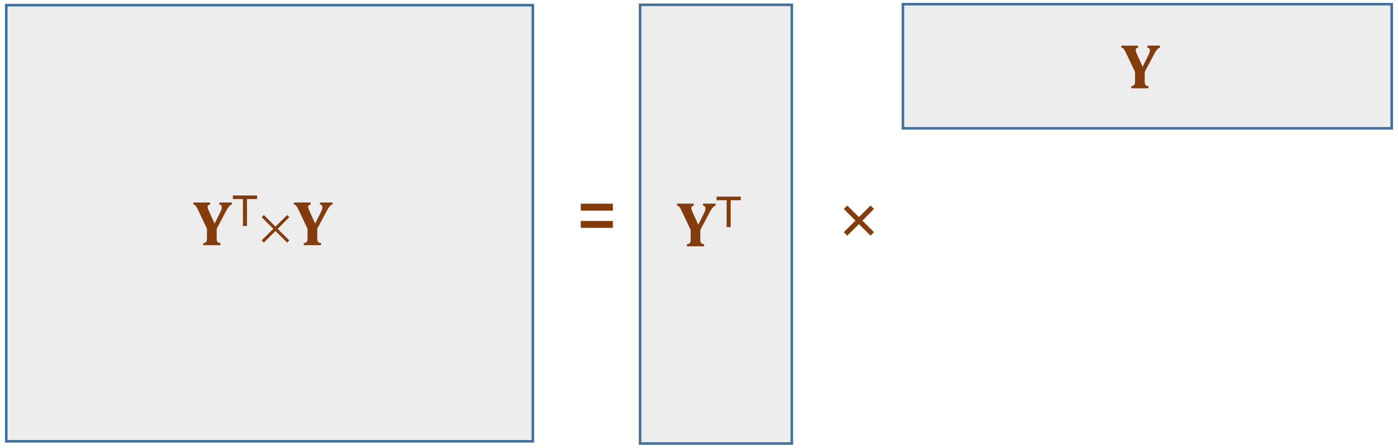
$$\Delta(W) = \|\mathbf{S}_X - \mathbf{S}_Y\|_F^2 = \sum_{i,j=1}^{n}((X_i - \bar{X}, X_j - \bar{X}) - (y_i, y_j))^2$$

$\overline{\mathbf{X}}$: p×n matrix with columns $X_1 - \overline{X}, X_2 - \overline{X}, \ldots , X_n - \overline{X}$

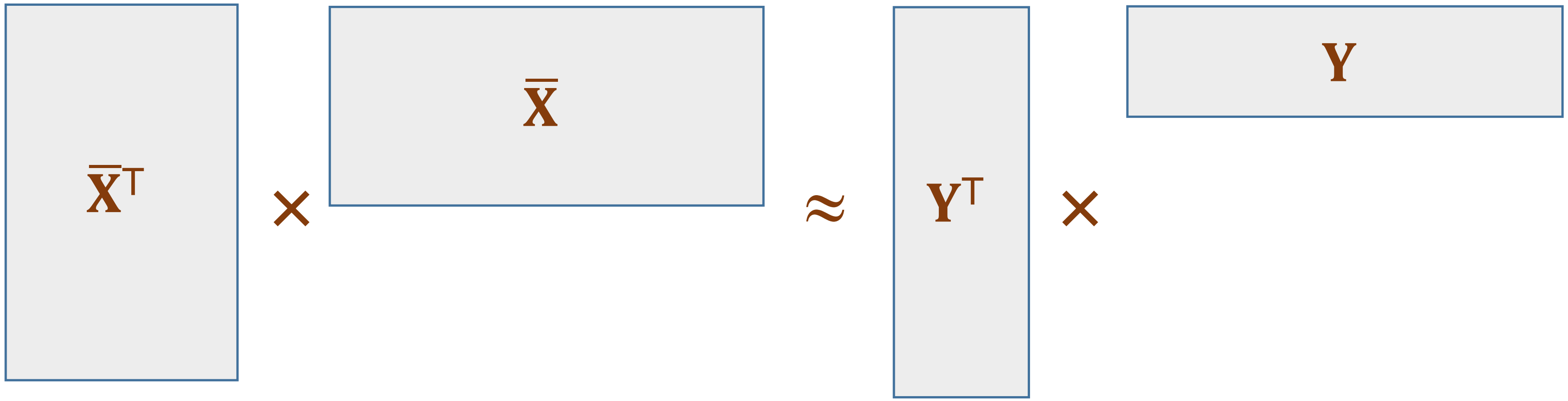
$$\rightarrow \qquad \mathbf{S}_X = \overline{\mathbf{X}}^T \times \overline{\mathbf{X}}$$



$$\mathbf{S}_Y = \mathbf{Y}^T \times \mathbf{Y}$$



$$\|\mathbf{S}_{\mathrm{X}} - \mathbf{S}_{\mathrm{Y}}\|_{\mathrm{F}}^2 = \|\bar{\mathbf{X}}^{\mathrm{T}} \times \bar{\mathbf{X}} - \mathbf{Y}^{\mathrm{T}} \times \mathbf{Y}\|_{\mathrm{F}}^2 \rightarrow \text{min over } \mathbf{Y} = \mathrm{W} \times \bar{\mathbf{X}}$$



$\bar{\mathbf{X}} = \mathbf{U}_p \times \Lambda_p \times \mathbf{V}_p^T$ - SVD

$$\bar{\mathbf{X}}^T \times \bar{\mathbf{X}} = (\mathbf{V}_p \times \Lambda_p \times \mathbf{U}_p^T) \times (\mathbf{U}_p \times \Lambda_p \times \mathbf{V}_p^T) = (\mathbf{V}_p \times \Lambda_p) \times (\mathbf{U}_p^T \times \mathbf{U}_p) \times (\Lambda_p \times \mathbf{V}_p^T)$$

$\mathbf{U}_p$ - orthogonal matrix: $\mathbf{U}_p^T \times \mathbf{U} = \mathrm{I}_p$

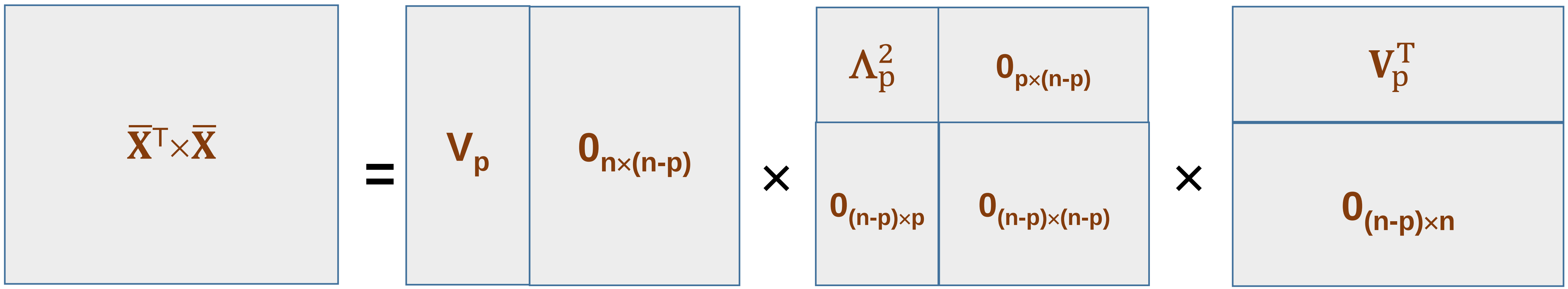
$$\bar{\mathbf{X}}^T \times \bar{\mathbf{X}} = \mathbf{V}_p \times \Lambda_p^2 \times \mathbf{V}_p^T$$

SVD for $\bar{\mathbf{X}}^T \times \bar{\mathbf{X}}$: $\mathrm{Rank}(\bar{\mathbf{X}}^T \times \bar{\mathbf{X}}) \le p$

$$\bar{\mathbf{X}}^T \times \bar{\mathbf{X}} = (\mathbf{V}_p \quad \mathbf{0}) \times \begin{pmatrix} \Lambda_p^2 & 0 \\ 0 & 0 \end{pmatrix} \times \begin{pmatrix} \mathbf{V}_p^T \\ 0 \end{pmatrix}$$

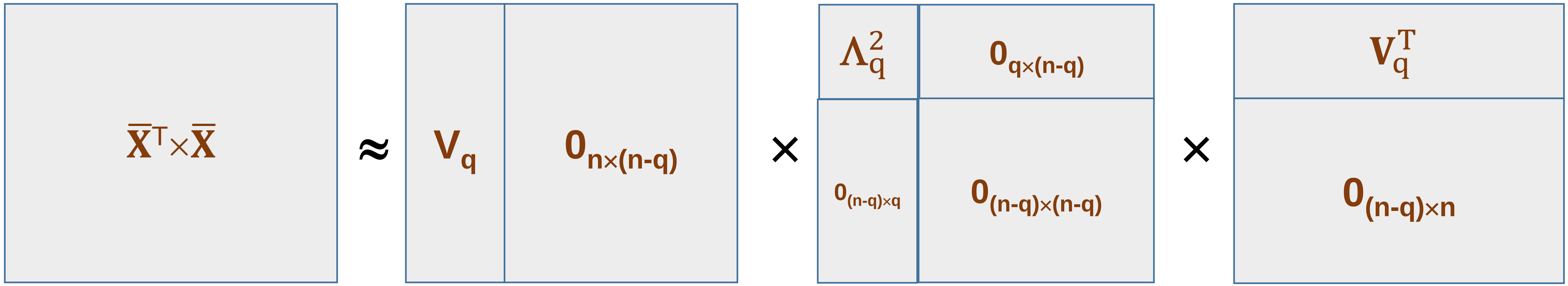
$$\bar{\mathbf{X}}^T \times \bar{\mathbf{X}} = \begin{pmatrix} \mathbf{V}_p & \mathbf{0}_{n\times(n-p)} \end{pmatrix} \times \begin{pmatrix} \Lambda_p^2 & \mathbf{0}_{p\times(n-p)} \\ \mathbf{0}_{(n-p)\times p} & \mathbf{0}_{(n-p)\times(n-p)} \end{pmatrix} \times \begin{pmatrix} \mathbf{V}_p^T \\ \mathbf{0}_{(n-p)\times n} \end{pmatrix}$$

$$\bar{\mathbf{X}}^{T}\times\bar{\mathbf{X}} = (\mathbf{V}_p \quad \mathbf{0}) \times \begin{pmatrix} \Lambda_p^2 & 0 \\ 0 & 0 \end{pmatrix} \times \begin{pmatrix} \mathbf{V}_p^T \\ 0 \end{pmatrix}$$



The best $n\times q$ approximation:

$$\bar{\mathbf{X}}^{T}\times\bar{\mathbf{X}} \approx (\mathbf{V}_q \quad \mathbf{0}) \times \begin{pmatrix} \Lambda_q^2 & 0 \\ 0 & 0 \end{pmatrix} \times \begin{pmatrix} \mathbf{V}_q^T \\ 0 \end{pmatrix} = \mathbf{V}_q \times \Lambda_q^2 \times \mathbf{V}_q^T$$



The best $n \times q$ approximation: $$\bar{\mathbf{X}}^{T} \times \bar{\mathbf{X}} \approx \mathbf{V}_q \times \Lambda_q^2 \times \mathbf{V}_q^T$$

$\mathbf{V}_q$ - $n \times q$ orthogonal matrix: $\mathbf{V}_q^T \times \mathbf{V}_q = I_q$

Λ_q - $q \times q$ diagonal matrix

$\mathbf{V}_q^T$ - $q \times n$ matrix

$$\mathbf{Y}_{PCA} = \Lambda_q \times \mathbf{V}_q^T = W_{PCA} \times \bar{\mathbf{X}} \quad \rightarrow \quad \mathbf{V}_q \times \Lambda_q^2 \times \mathbf{V}_q^T = (\Lambda_q \times \mathbf{V}_q^T)^T \times (\Lambda_q \times \mathbf{V}_q^T) = \mathbf{Y}_{PCA}^T \times \mathbf{Y}_{PCA},$$

$$\boxed{\bar{\mathbf{X}}^{T} \times \bar{\mathbf{X}}} \approx \mathbf{Y}_{PCA}^T \times \mathbf{Y}_{PCA} = \boxed{\mathbf{Y}_{PCA}^T} \times \boxed{\mathbf{Y}_{PCA}}$$

$$\bar{\mathbf{X}}^{\mathrm{T}} \times \bar{\mathbf{X}} \approx \mathbf{Y}_{\mathrm{PCA}}^{\mathrm{T}} \times \mathbf{Y}_{\mathrm{PCA}} = \mathbf{Y}_{\mathrm{PCA}}^{\mathrm{T}} \times \mathbf{Y}_{\mathrm{PCA}}$$

The best $n \times q$ approximation

$$\mathbf{Y}_{\mathrm{MDS}} = \arg\min_{\mathbf{Y}} \|\bar{\mathbf{X}}^{\mathrm{T}} \times \bar{\mathbf{X}} - \mathbf{Y}^{\mathrm{T}} \times \mathbf{Y}\|_{\mathrm{F}}^{2} = \mathrm{W}_{\mathrm{PCA}} \times \bar{\mathbf{X}} \quad \rightarrow \quad \mathbf{Y}_{\mathrm{MDS}} = \mathrm{W}_{\mathrm{MDS}} \times \bar{\mathbf{X}} = \mathrm{W}_{\mathrm{PCA}} \times \bar{\mathbf{X}}$$

$$\mathrm{W}_{\mathrm{MDS}} = \mathrm{W}_{\mathrm{PCA}}$$

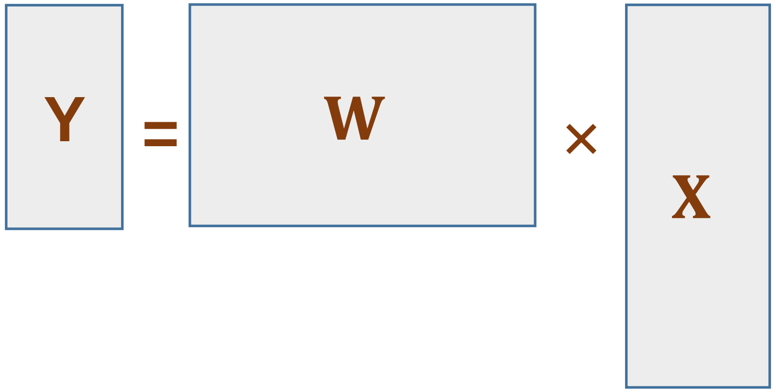
5) Maximum variance preserving:

to find low-dimensional features which:

- uncorrelated
- maximizing mutual information (Gaussian data)
- maximally preserving the dispersion (variance) in data

Probabilistic information model

- X - p-dimensional Gaussian random vector with zero mean
- $\Sigma_X = \mathrm{Cov}(X) = \mathbf{M}(X \times X^T)$ – $p \times p$ covariance matrix



W - orthogonal $q \times p$ matrix

$Y = W \times X$ - q-dimensional projected vector into space $\mathrm{Span}(W^T)$ spanned by columns of W^T

Y – Gaussian, $\mathbf{M}Y = 0$, $\Sigma_Y = \mathrm{Cov}(Y) = \mathbf{M}(Y \times Y^T) = W \times \Sigma_X \times W^T$ – $q \times q$ covariance matrix

$I(X, Y) = H(Y) - H(Y|X)$ - mutual information between X and Y

- $H(Y)$ – entropy (measure of uncertainty) of Y,
- $H(Y|X)$ - conditional entropy of Y given X

$$\boxed{I(X, Y) = I_W(X, Y): \qquad I_W(X, Y) \to \max \quad \text{over } W}$$

$$\boxed{I(X, Y) = H(Y) - H(Y|X) = I_W(X, Y): \qquad I_W(X, Y) \rightarrow \max \quad \text{over } W}$$

W is deterministic $\rightarrow$ $H(Y|X) = 0$ $\rightarrow$ $\boxed{I(X, Y) = H(Y)}$

$H(y) = - \int p(y) \times \log_2 p(y) dy$ – entropy of random variable Y with density $p(y)$

$p(y) = ((2\pi)^q \times Det(\Sigma_Y))^{1/2} \times \exp\left\{-\frac{1}{2} \exp(y^T \times \Sigma_Y^{-1} \times y)\right\}$ – density of $Y \sim N(0, \Sigma_Y)$

$$H(y) = \frac{1}{2} \log_2 (e(2\pi)^q) + \frac{1}{2} \times \log_2(Det(\Sigma_Y)) = \frac{1}{2} \log_2 (e(2\pi)^q) + \frac{1}{2} \times \log_2(Det(W \times \Sigma_X \times W^T))$$

$$Det(W \times \Sigma_X \times W^T) \rightarrow \max \quad \text{over } W$$

$$\mathrm{Det}(W \times \Sigma_X \times W^T) \rightarrow \max \quad \text{over } W$$

The solution: $W^* = U \times E^T$

- U - an arbitrary $q \times q$ orthogonal matrix, does not affect the entropy
- $p \times q$ orthogonal matrix $E = (e_1\ e_2\ \ldots\ e_q)$ consists of q p-dimensional **orthogonal** eigenvectors of Σ corresponding to q largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$

$q = 1$:

- e_1 - a direction in which random variable $Y_e = (X, e)$ has maximal variance
- $Y_1 = (X, e_1)$ and $Var(Y_1) = \lambda_1$

The larger the variance has a random variable, the more information we get after obtaining its value

$q > 1$: orthogonal directions $e_1, \dots , e_{k-1}$ are chosen $\rightarrow Y_1, \dots , Y_{k-1}$

Remaining information in X – information in random vector $X_k = X - E(X \mid Y_1, \dots , Y_{k-1})$

We are looking for the best direction e which maximally preserves the remaining information:

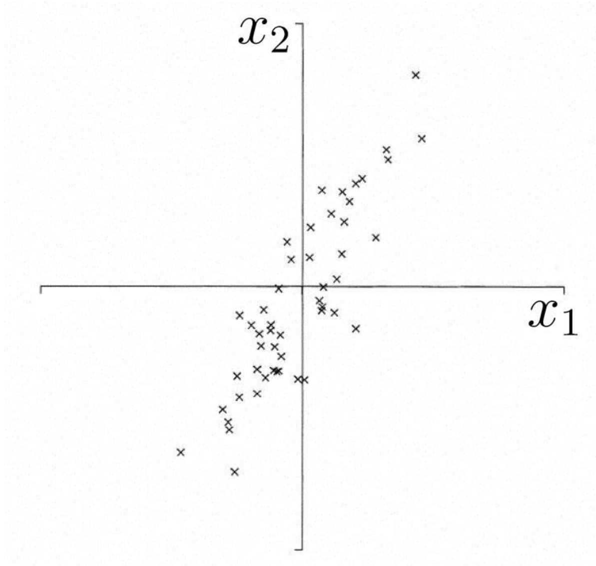
- $e \in Span(e_1, \dots , e_{k-1})^{\perp}$: $e \perp e_1, \dots , e_{k-1}$
- e_k – the best remaining direction

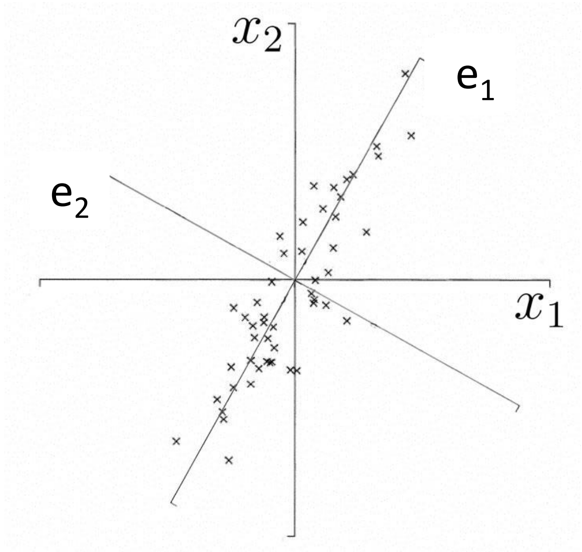
Statistical analysis

- X - p-dimensional random vector
- $\mathrm{Cov}(X) = \mathbf{M}(X - \mathbf{M}X)\times(X - \mathbf{M}X)^T$ – $p\times p$ covariance matrix
- $\{X_1, X_2, \ldots , X_n\}$ – i.i.d. (sample)
- $\overline{X}$ - sample mean (estimator of $\mathbf{M}X$)
- $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \times (X_i - \overline{X})^T$ – sample covariance matrix - estimator of $\mathrm{Cov}(X)$

An applying the above technique to the sample covariance matrix results:

- decorrelation and ordering by variance





$$X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

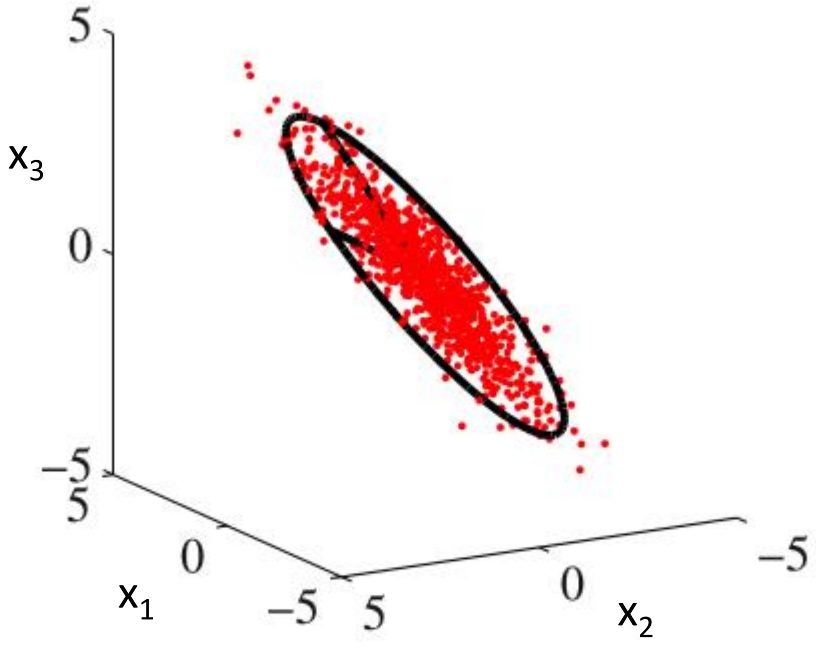
consists of correlated components

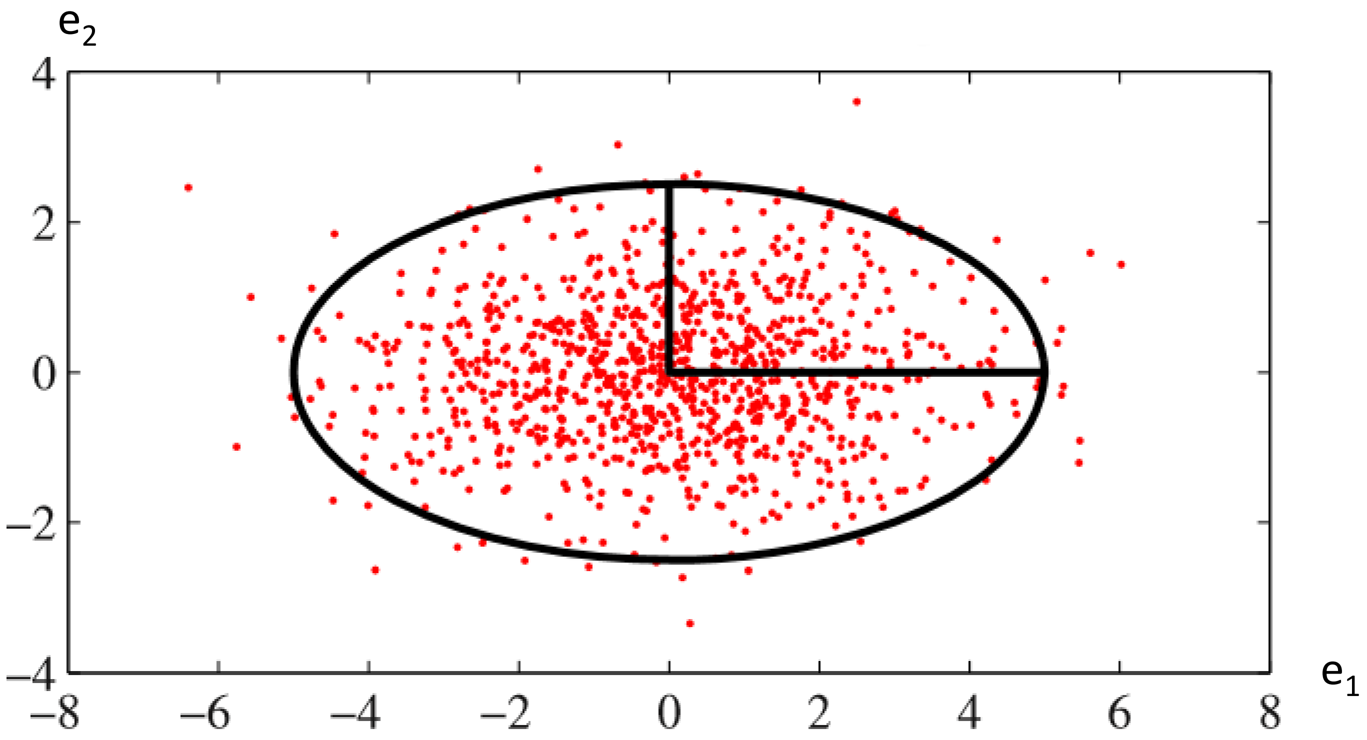
$$Y = \begin{pmatrix} y_1 = (X, e_1) \\ y_2 = (X, e_2) \end{pmatrix}$$

consists of uncorrelated components

$$Var(Y_1) = \lambda_1 > Var(Y_2) = \lambda_2$$

- Maximum variance preserving with decorrelation





Variance in original vectors:

$$\frac{1}{n}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2 = \sum_{k=1}^{p}\lambda_k$$

Variance in reduced vectors:

$$\frac{1}{n}\sum_{i=1}^{n}\|Y_i\|^2 = \sum_{k=1}^{q}\lambda_k$$

Machine Learning

- X - p-dimensional original feature vector, $\{X_1, X_2, \dots, X_n\}$ – training dataset

$E_p = (e_1 \dots e_p)$ - $q \times p$ matrix, $Y_p = E_p^T \times (X - \bar{X}) = \begin{pmatrix} y_1 \\ \cdots \\ y_p \end{pmatrix} \in R^p$- transformed feature vector

consists of ‘the same’ original features but written in other coordinate system with ‘variance ordering’

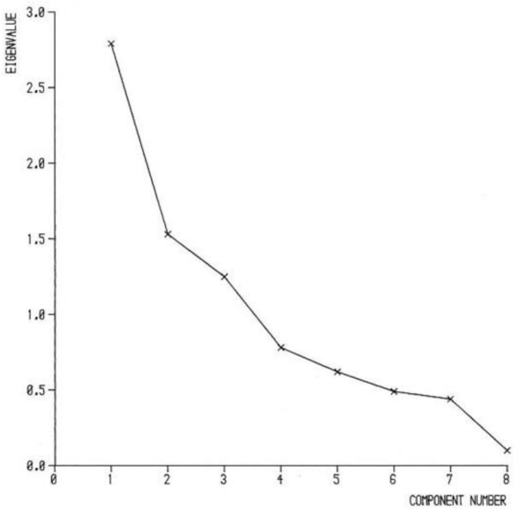
Let dispersion in p-th transformed components $\frac{1}{n}\sum_{i=1}^{n}\left|y_{p,i}\right|^2 = \lambda_p \approx 0$ is small:

- all $\{y_{p,i}\}$ have ‘the same values’ (zeros)
- all examples $\{X_i\}$ are obtained under the same value of p-th transformed feature
- impossible to discover ‘an influence’ of the p-th transformed feature on any Variable of interest
- the p-th transformed feature is useless, ‘non-informative’ and can be removed

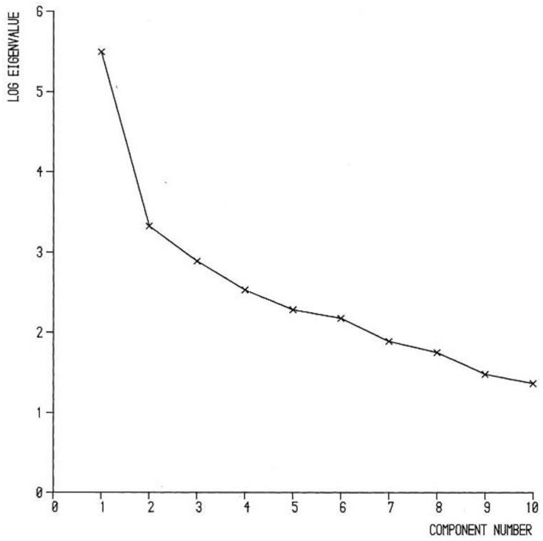
PCA transforms the original possibly correlated features to uncorrelated ‘variance ordered’ transformed features and removes the non-informative transformed features (with small variance)

How many transformed features should be left

Typical examples



Blood chemistry data



Gas chromatography data

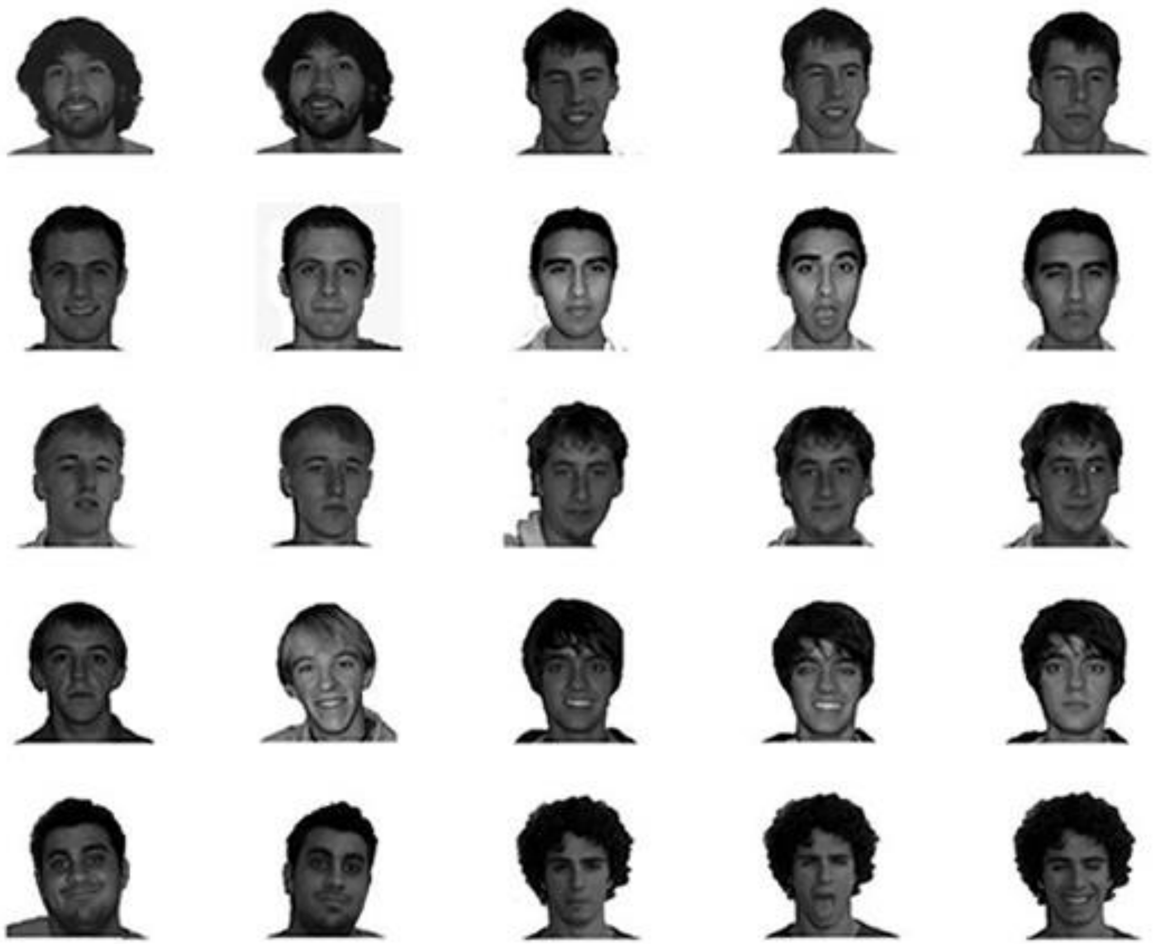
Typical 'rules of thumb':

- to retain enough eigenvalues ($q(P)$, say) to explain at least the fraction of P of dispersion in the data

$$q(P) = \text{minimal } q: \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \geq P \qquad P \sim 0{,}9;\ 0;95$$

- based on quantities $\left\{\frac{\lambda_q}{\sum_{k=q+1}^{p} \lambda_k}\right\}$, $\{\lambda_q - \lambda_{q+1}\}$, etc.

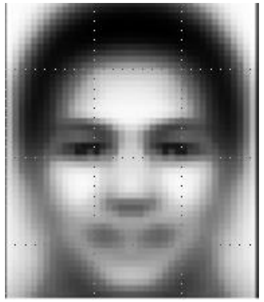
Training Face dataset



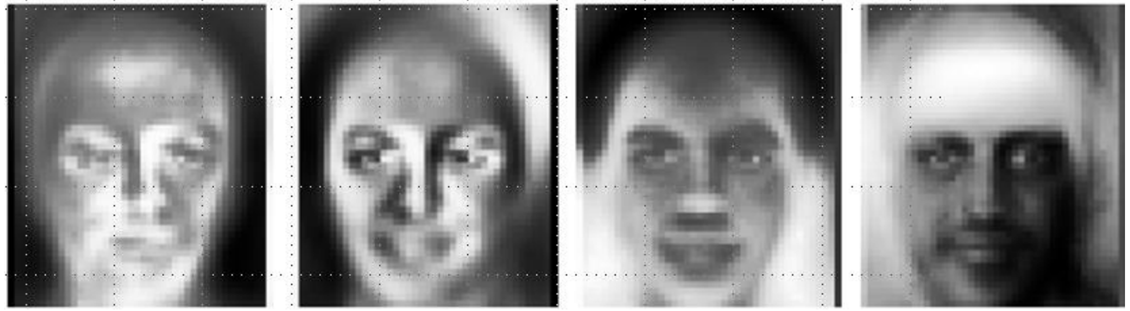
Each face has dimension p = 2061

PCA applying to the Training face dataset

‘Mean’ face

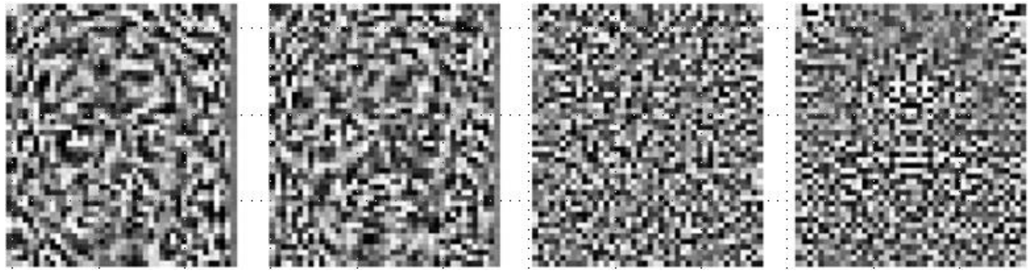


“Principal faces”
(eigenfaces)



first 4 ‘principal faces’

‘principal faces’
15, 100, 200, 250, 300



‘principal faces’
400, 450, 1000, 2000

Original face

Keep only **the first 8 principal components** (eigenvectors)

Recovered face: X* = mean face + $z_1 \times \mathbf{e}_1 + z_2 \times \mathbf{e}_2 + \ldots + z_8 \times \mathbf{e}_8$

- linear combination of 8 first principal faces $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_8\}$

- projection onto 8-dimensional space in R^{2061} spanned by $\{\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_8\}$

$(z_1, z_2, \ldots, z_8)$ - 8-dimensional reduced feature vector

$y_1\times\mathbf{e}_1$ $y_2\times\mathbf{e}_2$ $y_3\times\mathbf{e}_3$ $y_4\times\mathbf{e}_4$ $y_5\times\mathbf{e}_6$ $y_6\times\mathbf{e}_7$ $y_7\times\mathbf{e}_7$ $y_8\times\mathbf{e}_8$