Principal Component Analysis (PCA)

Prof. Gustavo Nonato

NYU / CUSP - GX 5006

January 31, 2017

Eigenvectors and Eigenvalues

Eigenvectors and Eigenvalues

Given a $d \times d$ matrix $\mathbf{A}$, a pair $(\lambda, \mathbf{u})$ that satisfies

$$\mathbf{Au} = \lambda\mathbf{u}$$

is called eigenvalue (λ) and corresponding eigenvector ($\mathbf{u}$) of $\mathbf{A}$.

Eigenvectors and Eigenvalues

Given a $d \times d$ matrix $\mathbf{A}$, a pair $(\lambda, \mathbf{u})$ that satisfies

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$$

is called eigenvalue (λ) and corresponding eigenvector ($\mathbf{u}$) of $\mathbf{A}$.

$\mathbf{A}$ is a linear transformation $\mathbf{A} : \mathbb{R}^d \to \mathbb{R}^d$, so, for any $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{A}\mathbf{x}$ corresponds to scale, rotate, and/or reflect $\mathbf{x}$ in $\mathbb{R}^d$.

Eigenvectors and Eigenvalues

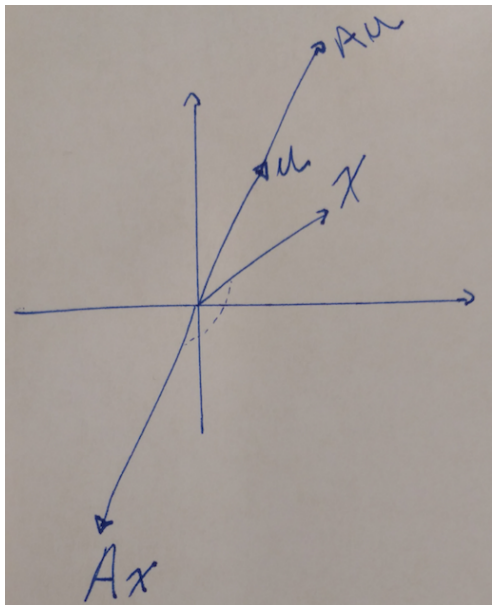
Given a $d \times d$ matrix $\mathbf{A}$, a pair $(\lambda, \mathbf{u})$ that satisfies

$$\mathbf{Au} = \lambda\mathbf{u}$$

is called eigenvalue (λ) and corresponding eigenvector ($\mathbf{u}$) of $\mathbf{A}$.

$\mathbf{A}$ is a linear transformation $\mathbf{A} : \mathbb{R}^d \to \mathbb{R}^d$, so, for any $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{Ax}$ corresponds to scale, rotate, and/or reflect $\mathbf{x}$ in $\mathbb{R}^d$.

By definition, an eigenvector of $\mathbf{A}$ is never rotated by $\mathbf{A}$.



Symmetric Matrices

Symmetric Matrices

When $\mathbf{A}$ is a symmetric matrix, that is, $\mathbf{A}^\top = \mathbf{A}$, its eigenvectors $\mathbf{u}$ and eigenvalues λ bear particular properties:

Symmetric Matrices

When $\mathbf{A}$ is a symmetric matrix, that is, $\mathbf{A}^\top = \mathbf{A}$, its eigenvectors $\mathbf{u}$ and eigenvalues λ bear particular properties:

- $\lambda \in \mathbb{R}$ and $\mathbf{u} \in \mathbb{R}^d$ (no complex numbers involved).

Symmetric Matrices

When $\mathbf{A}$ is a symmetric matrix, that is, $\mathbf{A}^\top = \mathbf{A}$, its eigenvectors $\mathbf{u}$ and eigenvalues λ bear particular properties:

- $\lambda \in \mathbb{R}$ and $\mathbf{u} \in \mathbb{R}^d$ (no complex numbers involved).
- The eigenvectors are orthogonal, that is, given eigenvectors $\mathbf{u}_i$ and $\mathbf{u}_j$ from $\mathbf{A}$

$$\mathbf{u}_i^\top \mathbf{u}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{otherwise} \end{cases}$$

(assuming $\|\mathbf{u}_i\| = 1$)

Symmetric Matrices

Suppose $\mathbf{A}$ symmetric with d distinct eigenvalues λ_i.
The equations $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ can be written in matrix form as:

Symmetric Matrices

Suppose $\mathbf{A}$ symmetric with d distinct eigenvalues λ_i.
The equations $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ can be written in matrix form as:

$$\underbrace{\begin{bmatrix} a_{11} & & a_{1d} \\ \vdots & \cdots & \vdots \\ a_{d1} & & a_{dd} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}}_{\mathbf{U}} = \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}}_{\mathbf{U}}$$

Symmetric Matrices

Suppose $\mathbf{A}$ symmetric with d distinct eigenvalues λ_i.
The equations $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ can be written in matrix form as:

$$\underbrace{\begin{bmatrix} a_{11} & & a_{1d} \\ \vdots & \cdots & \vdots \\ a_{d1} & & a_{dd} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}}_{\mathbf{U}} = \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}}_{\mathbf{U}}$$

In matrix notation

$$\mathbf{AU} = \mathbf{UD}$$

Symmetric Matrices

Suppose $\mathbf{A}$ symmetric with d distinct eigenvalues λ_i.
The equations $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ can be written in matrix form as:

$$\underbrace{\begin{bmatrix} a_{11} & & a_{1d} \\ \vdots & \cdots & \vdots \\ a_{d1} & & a_{dd} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}}_{\mathbf{U}} = \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}}_{\mathbf{U}}$$

In matrix notation

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{D}$$

Since $\mathbf{U}$ is an orthogonal matrix, $\mathbf{U}^\top = \mathbf{U}^{-1}$, thus

Symmetric Matrices

Suppose $\mathbf{A}$ symmetric with d distinct eigenvalues λ_i.
The equations $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ can be written in matrix form as:

$$\underbrace{\begin{bmatrix} a_{11} & & a_{1d} \\ \vdots & \cdots & \vdots \\ a_{d1} & & a_{dd} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}}_{\mathbf{U}} = \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}}_{\mathbf{U}}$$

In matrix notation

$$\mathbf{AU} = \mathbf{UD}$$

Since $\mathbf{U}$ is an orthogonal matrix, $\mathbf{U}^\top = \mathbf{U}^{-1}$, thus

Spectral Decomposition of a Symmetric Matrix

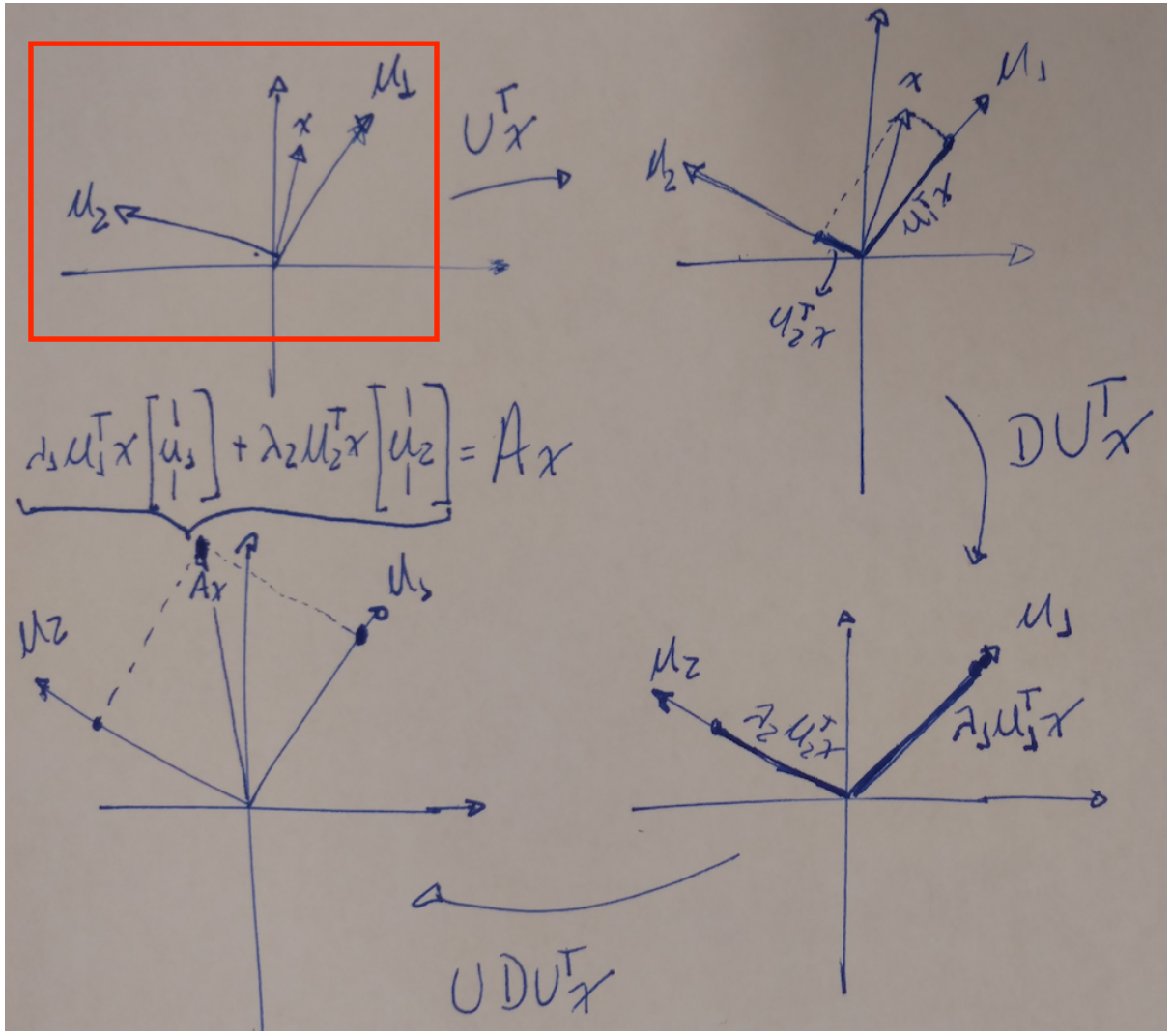
$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

Symmetric Matrices

$$\mathbf{Ax} = \mathbf{UDU}^\top \mathbf{x}$$

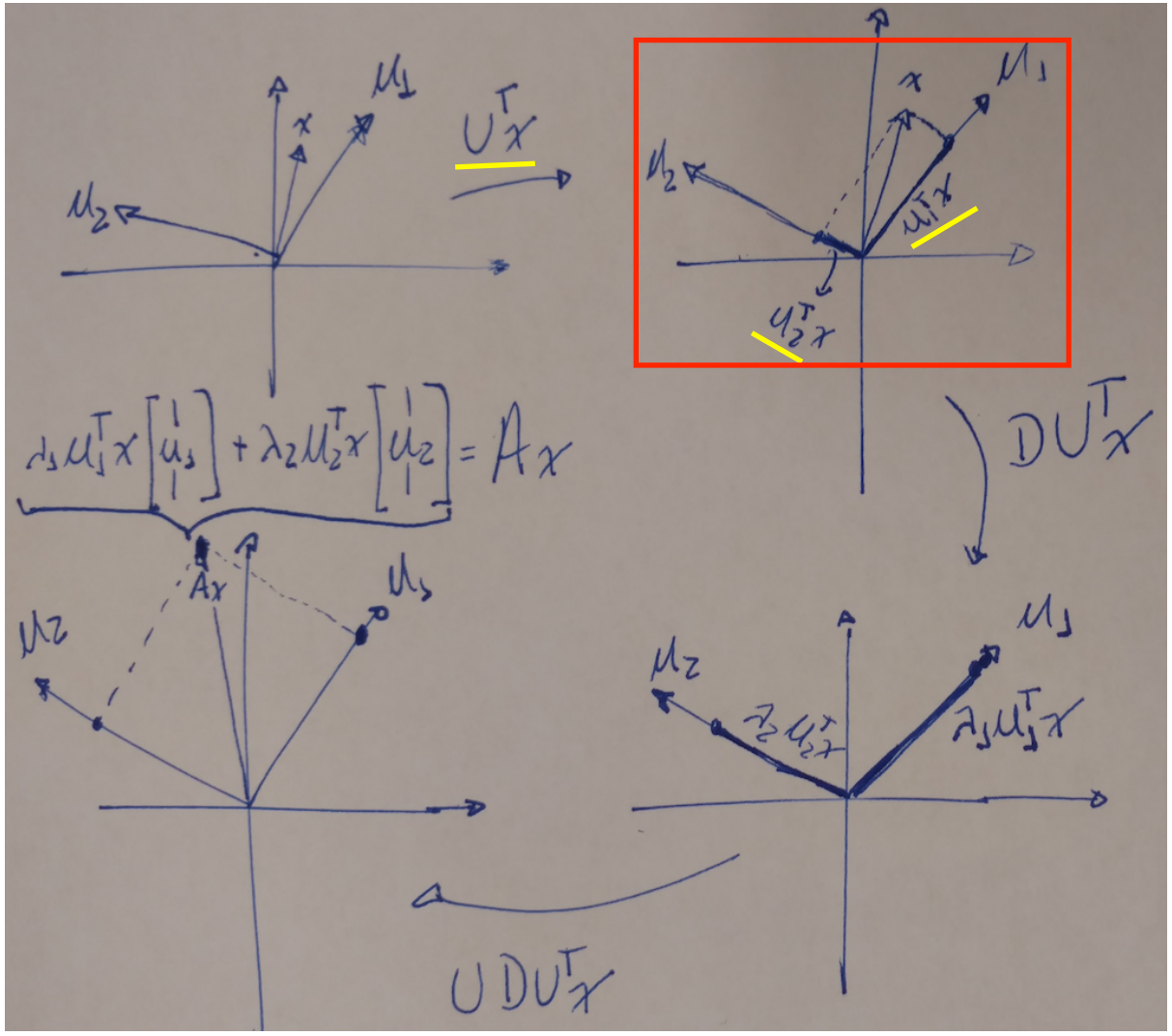
Symmetric Matrices

$$\mathbf{Ax} = \mathbf{UDU}^\top \mathbf{x}$$



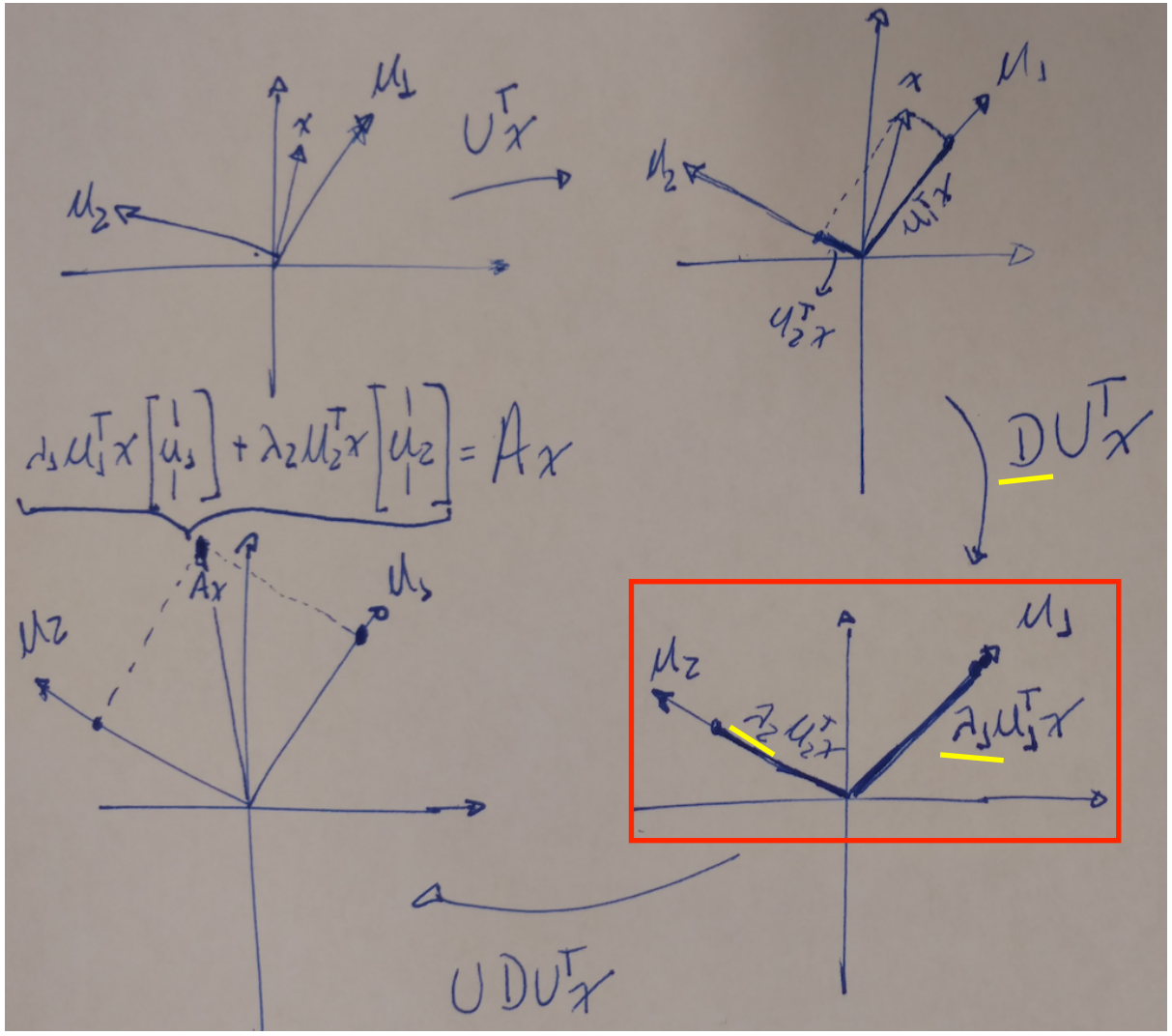
Symmetric Matrices

$$\mathbf{Ax} = \mathbf{UDU}^\top \mathbf{x}$$



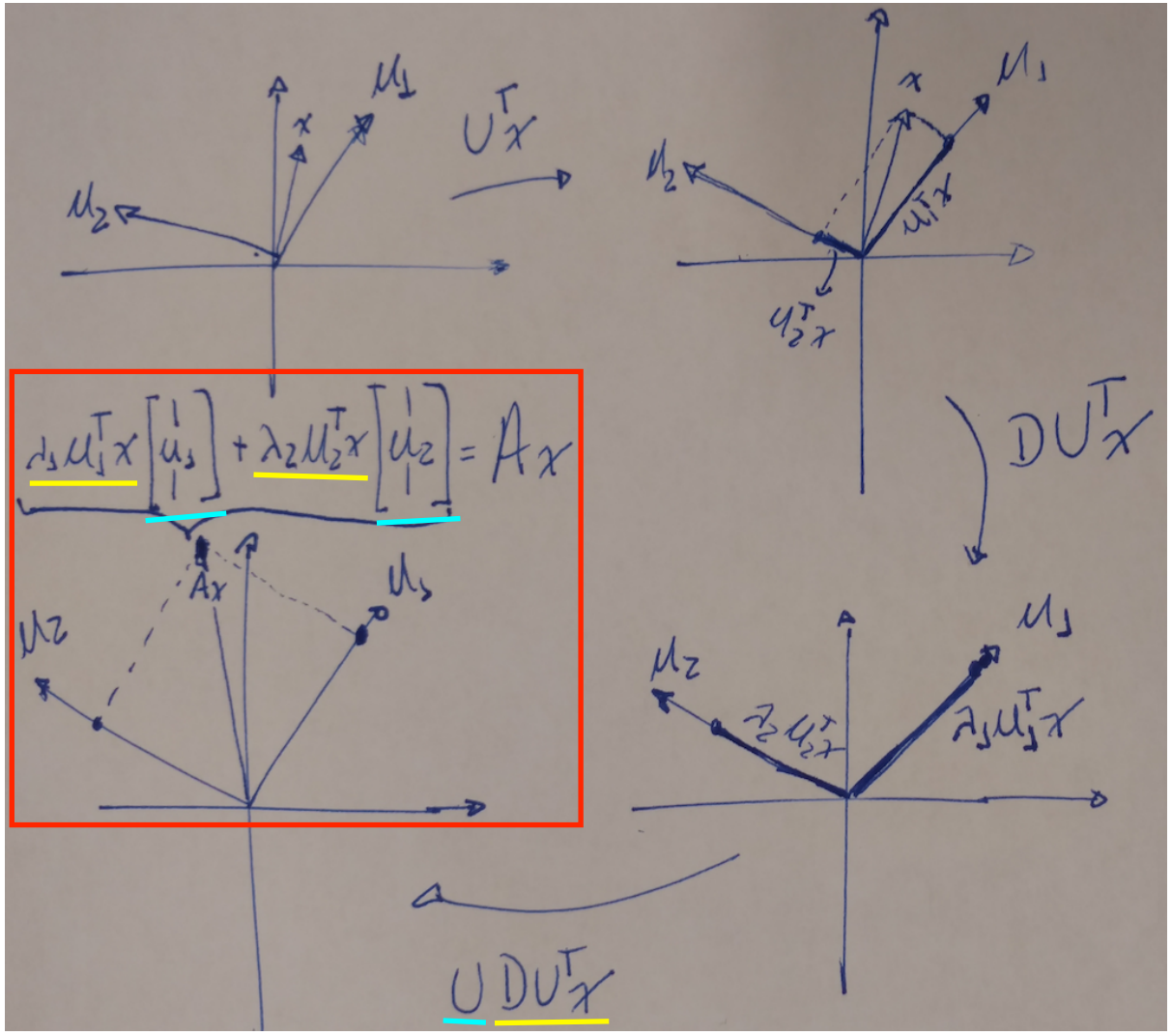
Symmetric Matrices

$$\mathbf{Ax} = \mathbf{UDU}^\top \mathbf{x}$$



Symmetric Matrices

$$\mathbf{Ax} = \mathbf{UDU}^\top \mathbf{x}$$



Quadratic Form

Let $\mathbf{A}$ be a symmetric matrix, then

Quadratic Form

Let $\mathbf{A}$ be a symmetric matrix, then

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$$

is a quadratic form.

Quadratic Form

Let $\mathbf{A}$ be a symmetric matrix, then

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$$

is a quadratic form.

$$f(\mathbf{x}) = f(x_1, x_2) = [x_1\, x_2] \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + x_2^2$$

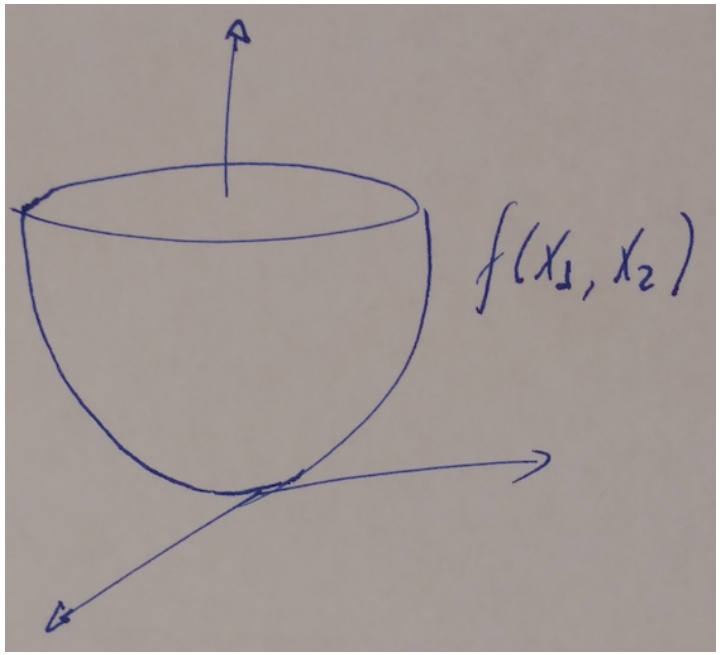
Quadratic Form

Let $\mathbf{A}$ be a symmetric matrix, then

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$$

is a quadratic form.

$$f(\mathbf{x}) = f(x_1, x_2) = [x_1\, x_2] \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + x_2^2$$



Quadratic Form

A symmetric with eigenvalues $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_d$ and corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$.

Quadratic Form

A symmetric with eigenvalues $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_d$ and corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$.

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$$

Quadratic Form

A symmetric with eigenvalues $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_d$ and corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$.

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$$

$$\max_{\|\mathbf{x}\|=1} \{f(\mathbf{x})\} = \mathbf{u}_1^\top \mathbf{A}\mathbf{u}_1 = \lambda_1$$

$$\min_{\|\mathbf{x}\|=1} \{f(\mathbf{x})\} = \mathbf{u}_d^\top \mathbf{A}\mathbf{u}_d = \lambda_d$$

Positive (Semi)Definite Matrices

A is symmetric.

Positive (Semi)Definite Matrices

$\mathbf{A}$ is symmetric.

If for all $\mathbf{x}$

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} \begin{cases} \geq 0, \mathbf{A} \text{ is semipositive definite} \\ > 0, \mathbf{A} \text{ is positive definite} \end{cases}$$

Positive (Semi)Definite Matrices

A is symmetric.

If for all **x**

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} \begin{cases} \geq 0, \mathbf{A} \text{ is semipositive definite} \\ > 0, \mathbf{A} \text{ is positive definite} \end{cases}$$

$$f(x) \geq 0 \longrightarrow \lambda_1 \geq \lambda_2 \ldots \geq \lambda_d \geq 0$$

$$f(x) > 0 \longrightarrow \lambda_1 \geq \lambda_2 \ldots \geq \lambda_d > 0$$

Covariance Matrix

Let $\mathbf{x}_i = [x_{1i}, \ldots, x_{di}]^\top$, $\mathbf{x}_j = [x_{1j}, \ldots, x_{dj}]^\top$

Covariance Matrix

Let $\mathbf{x}_i = [x_{1i}, \ldots, x_{di}]^\top$, $\mathbf{x}_j = [x_{1j}, \ldots, x_{dj}]^\top$

The covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by

$$cov(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{d-1} \sum_{s=1}^{d} (x_{si} - \bar{x}_i)(x_{sj} - \bar{x}_j)$$

where $\bar{x}_i = \frac{1}{d} \sum_s x_{si}$ and $\bar{x}_j = \frac{1}{d} \sum_s x_{sj}$

Covariance Matrix

Let $\mathbf{x}_i = [x_{1i}, \ldots, x_{di}]^\top$, $\mathbf{x}_j = [x_{1j}, \ldots, x_{dj}]^\top$

The covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by

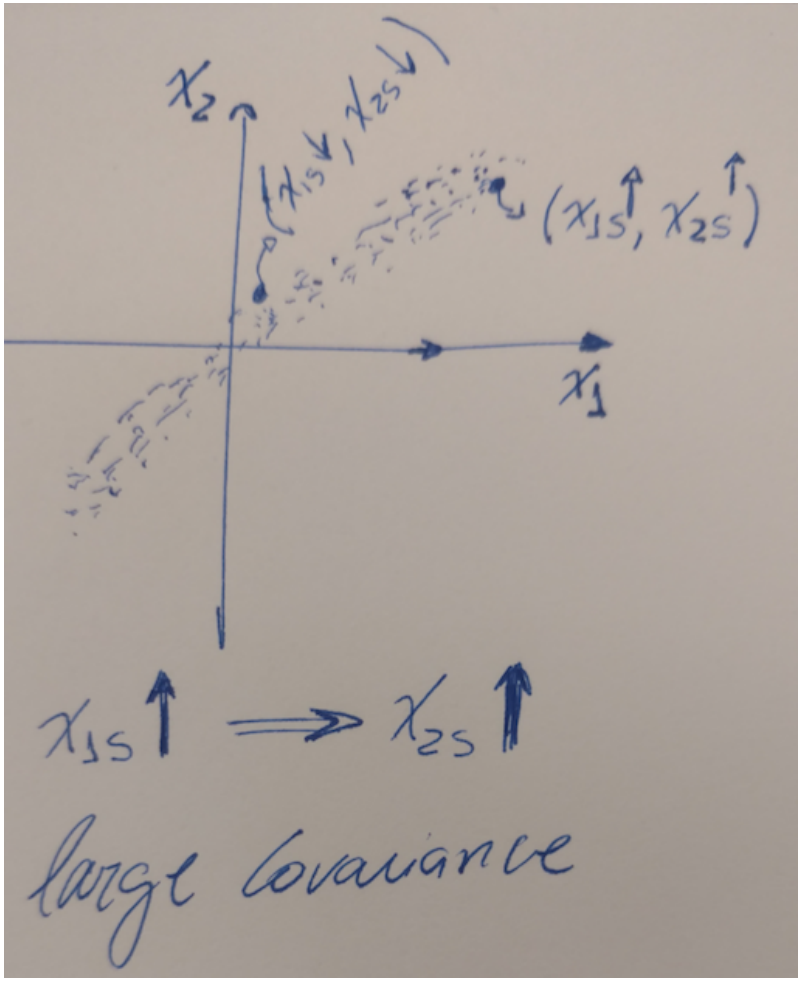
$$cov(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{d-1} \sum_{s=1}^{d} (x_{si} - \bar{x}_i)(x_{sj} - \bar{x}_j)$$

where $\bar{x}_i = \frac{1}{d} \sum_s x_{si}$ and $\bar{x}_j = \frac{1}{d} \sum_s x_{sj}$

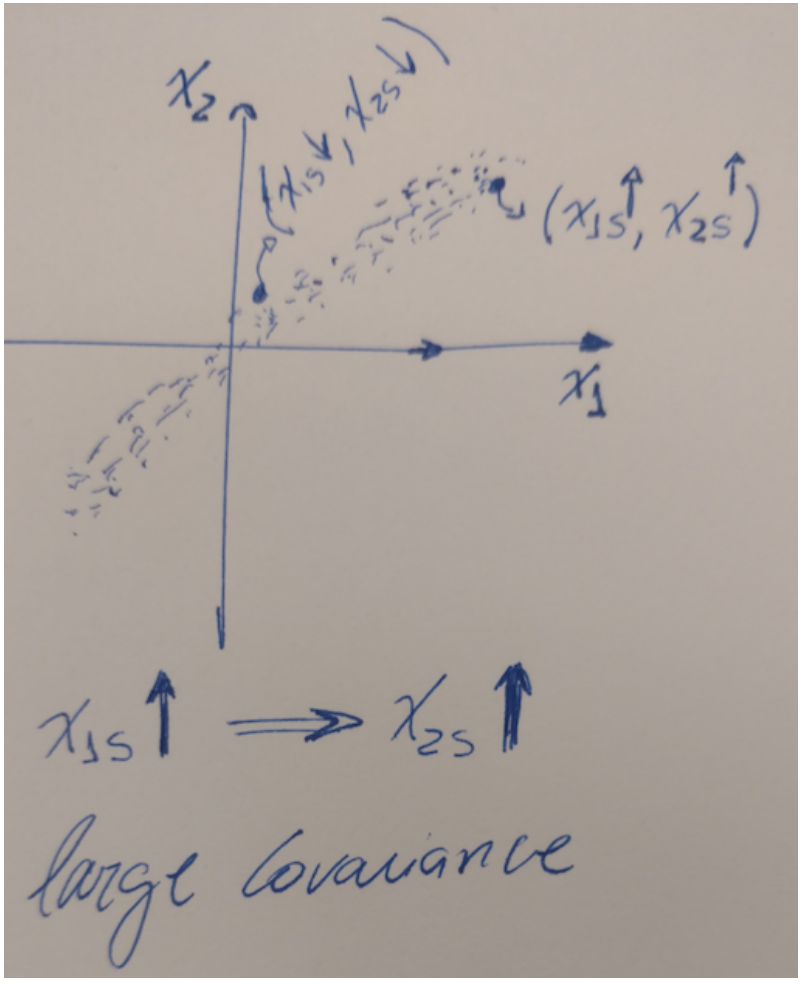
If we assume $\mathbf{x}_i$ and $\mathbf{x}_j$ centered, that is, $\bar{x}_i = 0$ and $\bar{x}_j = 0$

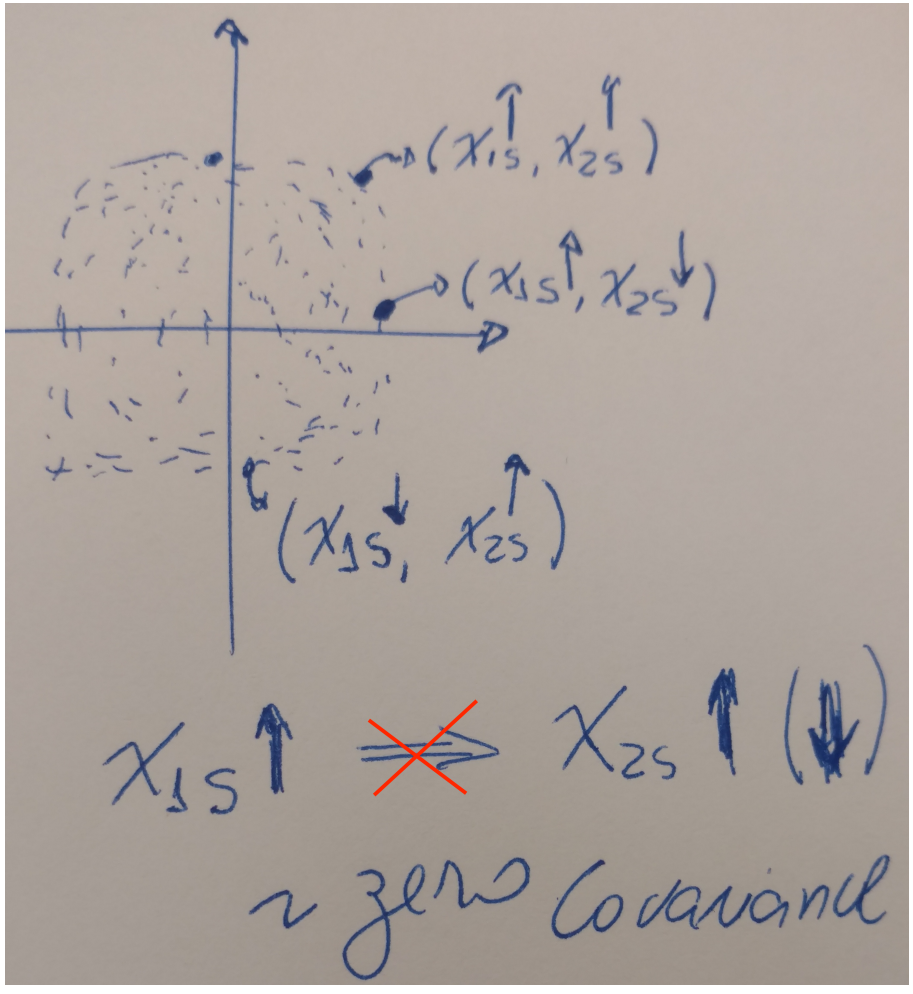
$$cov(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{d-1} \sum_s x_{si} x_{sj}$$

Covariance Matrix



Covariance Matrix





Covariance Matrix

Let $\mathbf{x}_i$, $i = 1, \ldots, n$ a set of data instances (points in $\mathbb{R}^d$) arranged as columns in a data matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \ldots & x_{dn} \end{bmatrix} \tag{1}$$

Covariance Matrix

Let $\mathbf{x}_i$, $i = 1, \ldots, n$ a set of data instances (points in $\mathbb{R}^d$) arranged as columns in a data matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \ldots & x_{dn} \end{bmatrix} \quad (1)$$

Columns correspond to data instances

Rows correspond to a particular attribute.

Covariance Matrix

Lets assume $\mathbf{X}$ is centered, that is, the average of each rows is zero.

Covariance Matrix

Lets assume $\mathbf{X}$ is centered, that is, the average of each rows is zero.

The covariance matrix of $\mathbf{X}$ is given by:

$$\frac{1}{n-1}\mathbf{X}\mathbf{X}^\top = \begin{bmatrix} cov(x_{1:}, x_{1:}) & cov(x_{1:}, x_{2:}) & \dots & cov(x_{1:}, x_{d:}) \\ cov(x_{2:}, x_{1:}) & cov(x_{2:}, x_{2:}) & \dots & cov(x_{2:}, x_{d:}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_{d:}, x_{1:}) & cov(x_{d:}, x_{2:}) & \dots & cov(x_{d:}, x_{d:}) \end{bmatrix}$$

Variances are in the main diagonal

Covariance Matrix

Some important properties:

Covariance Matrix

Some important properties:

- $\mathbf{X}\mathbf{X}^\top$ is symmetric

Covariance Matrix

Some important properties:

- $\mathbf{X}\mathbf{X}^\top$ is symmetric
- $\mathbf{X}\mathbf{X}^\top$ is positive semidefinite

Covariance Matrix

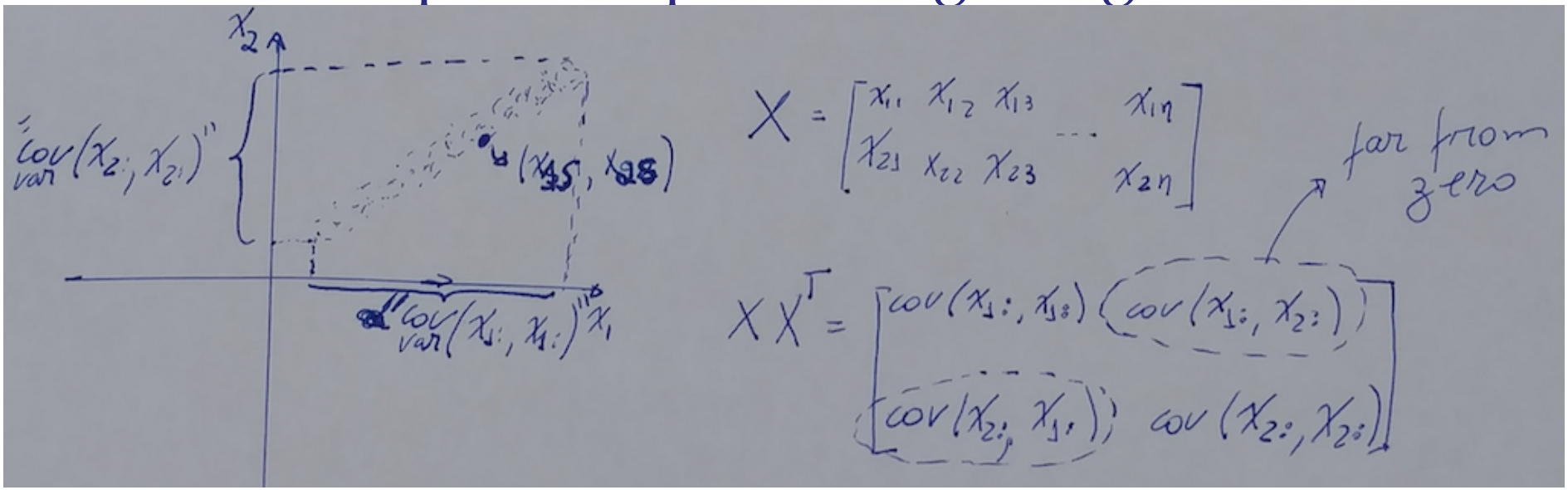
Some important properties:

- $\mathbf{X}\mathbf{X}^\top$ is symmetric
- $\mathbf{X}\mathbf{X}^\top$ is positive semidefinite

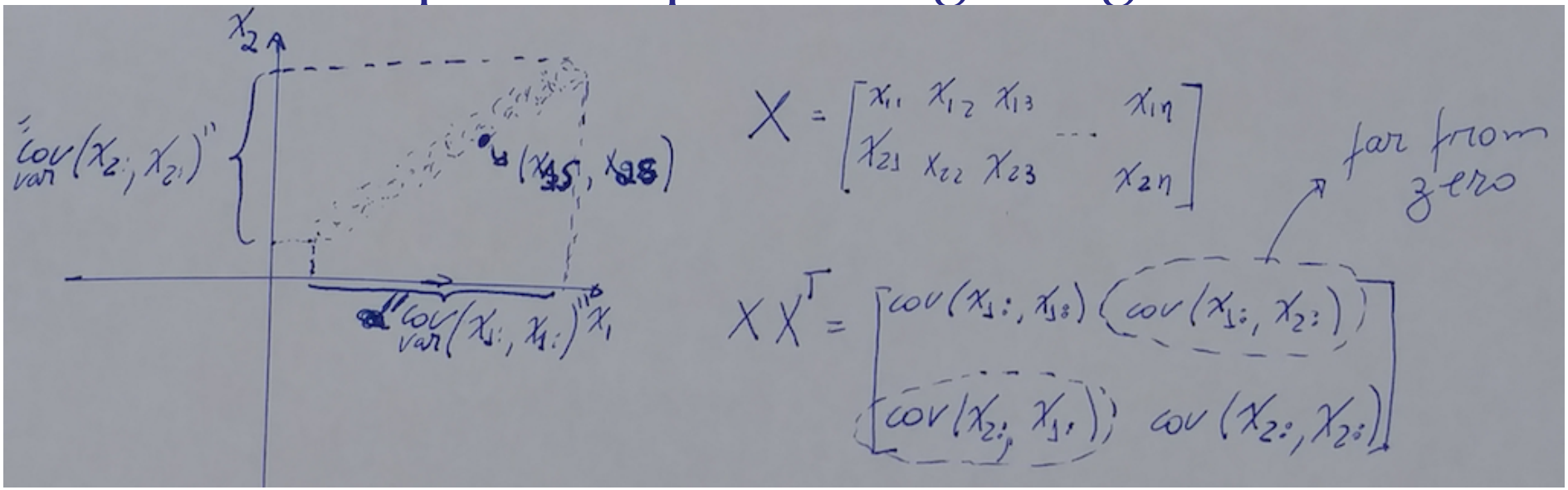
Therefore, all the properties we saw before as to symmetric matrices and positive definite quadratic forms hold for the covariance matrix.

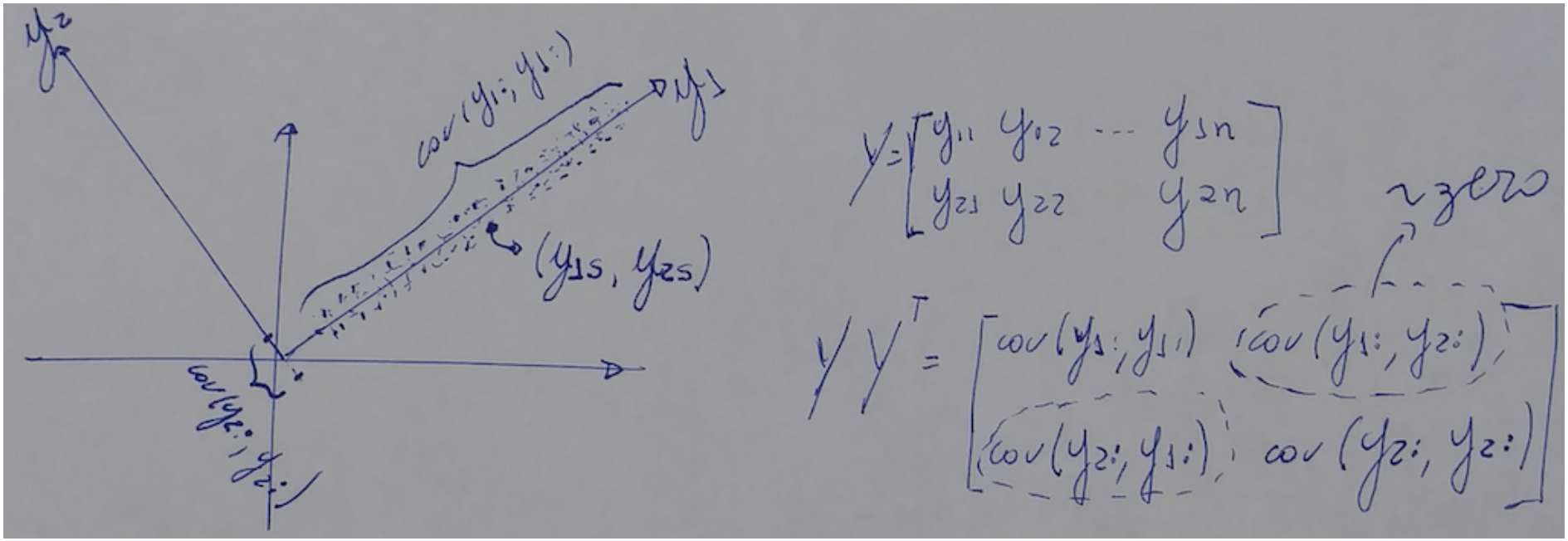
Principal Components: getting some intuition

Principal Components: getting some intuition



Principal Components: getting some intuition





Principal Components

The idea of PCA is to find a new basis to write the data so as to vanish the covariance between distinct attributes.

Principal Components

The idea of PCA is to find a new basis to write the data so as to vanish the covariance between distinct attributes.

Mathematically, we are looking for a change of basis matrix $\mathbf{P}$ such that

$$\mathbf{Y} = \mathbf{PX} \Longrightarrow \mathbf{YY}^\top = \mathbf{D}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements corresponding to the variance of each coordinate/attribute.

Principal Components

The idea of PCA is to find a new basis to write the data so as to vanish the covariance between distinct attributes.

Mathematically, we are looking for a change of basis matrix $\mathbf{P}$ such that

$$\mathbf{Y} = \mathbf{PX} \Longrightarrow \mathbf{YY}^\top = \mathbf{D}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements corresponding to the variance of each coordinate/attribute.

By fiding $\mathbf{P}$:

Principal Components

The idea of PCA is to find a new basis to write the data so as to vanish the covariance between distinct attributes.

Mathematically, we are looking for a change of basis matrix $\mathbf{P}$ such that

$$\mathbf{Y} = \mathbf{PX} \Longrightarrow \mathbf{YY}^\top = \mathbf{D}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements corresponding to the variance of each coordinate/attribute.

By fiding $\mathbf{P}$:

- the new attributes/coordinates will be decorrelated (redundancy removed)

Principal Components

The idea of PCA is to find a new basis to write the data so as to vanish the covariance between distinct attributes.

Mathematically, we are looking for a change of basis matrix $\mathbf{P}$ such that

$$\mathbf{Y} = \mathbf{PX} \Longrightarrow \mathbf{YY}^\top = \mathbf{D}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements corresponding to the variance of each coordinate/attribute.

By fiding $\mathbf{P}$:

- the new attributes/coordinates will be decorrelated (redundancy removed)
- some coordinates will tend to be of low variance (noise related coordinates)

Principal Components

The idea of PCA is to find a new basis to write the data so as to vanish the covariance between distinct attributes.

Mathematically, we are looking for a change of basis matrix $\mathbf{P}$ such that

$$\mathbf{Y} = \mathbf{PX} \Longrightarrow \mathbf{YY}^\top = \mathbf{D}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements corresponding to the variance of each coordinate/attribute.

By fiding $\mathbf{P}$:

- the new attributes/coordinates will be decorrelated (redundancy removed)
- some coordinates will tend to be of low variance (noise related coordinates)
- we will be able to reduce the dimension of the data without loosing relevant information.

Principal Components

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

Principal Components

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

$$\mathbf{Y}\mathbf{Y}^\top = (\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^\top = \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top$$

Principal Components

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

$$\mathbf{Y}\mathbf{Y}^\top = (\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^\top = \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top$$

Reminder

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$
$$\downarrow$$
$$\mathbf{U}^\top\mathbf{A}\mathbf{U} = \mathbf{D}$$

Principal Components

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

$$\mathbf{Y}\mathbf{Y}^\top = (\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^\top = \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top$$

Eigenvectors of $\mathbf{X}\mathbf{X}^\top \rightarrow \mathbf{U}$

Reminder

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$
$$\downarrow$$
$$\mathbf{U}^\top\mathbf{A}\mathbf{U} = \mathbf{D}$$

Principal Components

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

$$\mathbf{Y}\mathbf{Y}^\top = (\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^\top = \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top$$

Eigenvectors of $\mathbf{X}\mathbf{X}^\top \rightarrow \mathbf{U}$

$$\mathbf{P} = \mathbf{U}^\top$$

Reminder

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$
$$\downarrow$$
$$\mathbf{U}^\top\mathbf{A}\mathbf{U} = \mathbf{D}$$

Principal Components

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

$$\mathbf{Y}\mathbf{Y}^\top = (\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^\top = \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top$$

Reminder

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$
$$\downarrow$$
$$\mathbf{U}^\top\mathbf{A}\mathbf{U} = \mathbf{D}$$

Eigenvectors of $\mathbf{X}\mathbf{X}^\top \rightarrow \mathbf{U}$

$$\mathbf{P} = \mathbf{U}^\top$$

$$\mathbf{Y}\mathbf{Y}^\top = \mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}$$

Principal Components

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

$$\mathbf{Y}\mathbf{Y}^\top = (\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^\top = \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top$$

Reminder

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$
$$\downarrow$$
$$\mathbf{U}^\top\mathbf{A}\mathbf{U} = \mathbf{D}$$

Eigenvectors of $\mathbf{X}\mathbf{X}^\top \rightarrow \mathbf{U}$

$$\mathbf{P} = \mathbf{U}^\top$$

$$\mathbf{Y}\mathbf{Y}^\top = \mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}$$

$$\mathbf{Y}\mathbf{Y}^\top = \mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U} = \mathbf{D}$$

Principal Components

The coordinates of the data in the new basis is given by:

$$\mathbf{Y} = \mathbf{U}^\top \mathbf{X}$$

Principal Components

The coordinates of the data in the new basis is given by:

$$\mathbf{Y} = \mathbf{U}^\top \mathbf{X}$$

The diagonal matrix $\mathbf{D}$ in the decomposition $\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ contains the variances of each new coordinate.

Principal Components

The coordinates of the data in the new basis is given by:

$$\mathbf{Y} = \mathbf{U}^\top \mathbf{X}$$

The diagonal matrix $\mathbf{D}$ in the decomposition $\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ contains the variances of each new coordinate.

Moreover,

$$\mathbf{u}_1^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}_1 = \lambda_1 \text{ (maximum of the quadratic form)}$$

$$\mathbf{u}_d^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}_d = \lambda_d \text{ (minimum of the quadratic form)}$$

Principal Components

The coordinates of the data in the new basis is given by:

$$\mathbf{Y} = \mathbf{U}^\top \mathbf{X}$$

The diagonal matrix $\mathbf{D}$ in the decomposition $\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ contains the variances of each new coordinate.

Moreover,

$$\mathbf{u}_1^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}_1 = \lambda_1 \text{ (maximum of the quadratic form)}$$

$$\mathbf{u}_d^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}_d = \lambda_d \text{ (minimum of the quadratic form)}$$

$\mathbf{u}_1$ is the direction that maximizes the variance and $\mathbf{u}_d$ the direction that minimizes the variance.

Principal Components

We can filter out low variance directions (corresponding to small λ_i), since they typically correspond to noise.

Principal Components

We can filter out low variance directions (corresponding to small λ_i), since they typically correspond to noise.

We can reconstruct "noise-free" data by $\hat{\mathbf{X}} = \mathbf{U}\hat{\mathbf{Y}}$

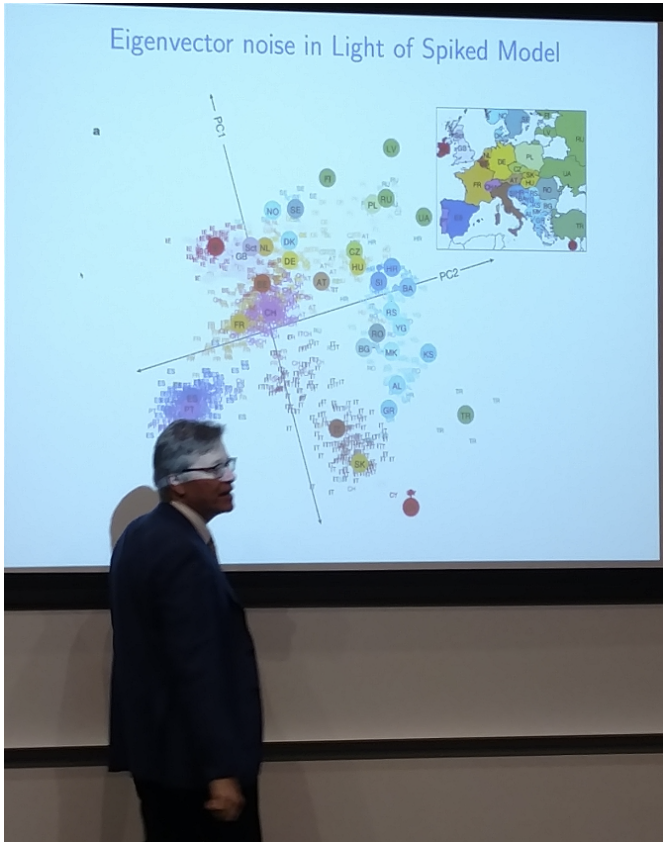
$$\hat{\mathbf{Y}} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ & & \vdots & \\ y_{k1} & y_{k2} & \cdots & y_{kn} \\ 0 & 0 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

Principal Component Analysis

How to choose the ideal variance (singular value) cut off?

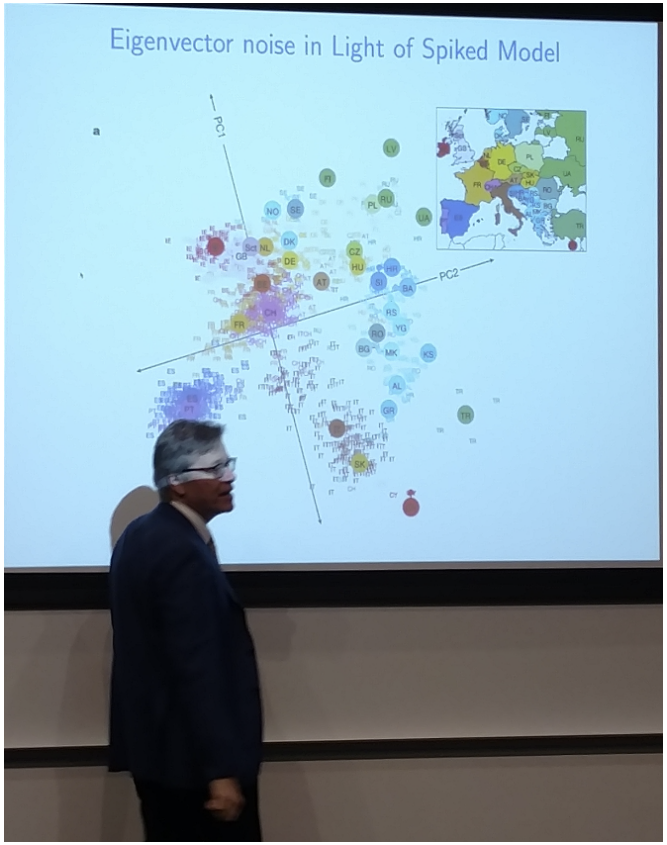
Principal Component Analysis

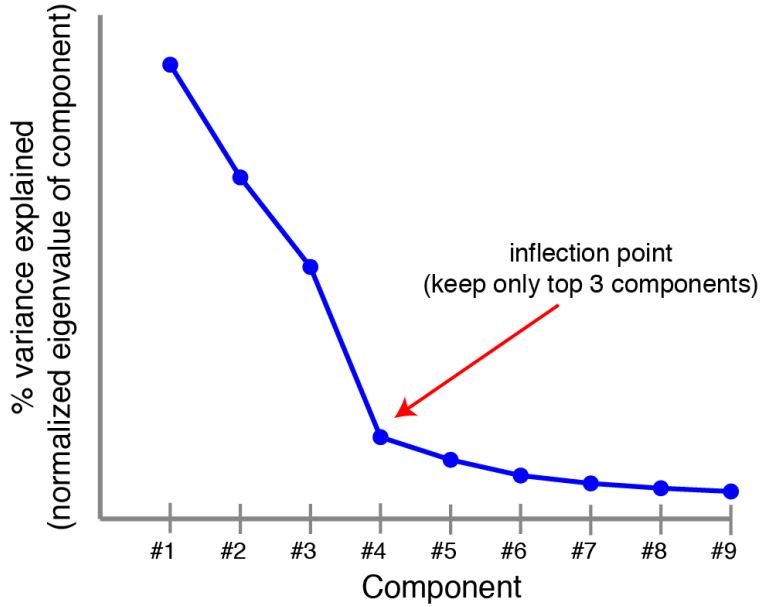
How to choose the ideal variance (singular value) cut off?



Principal Component Analysis

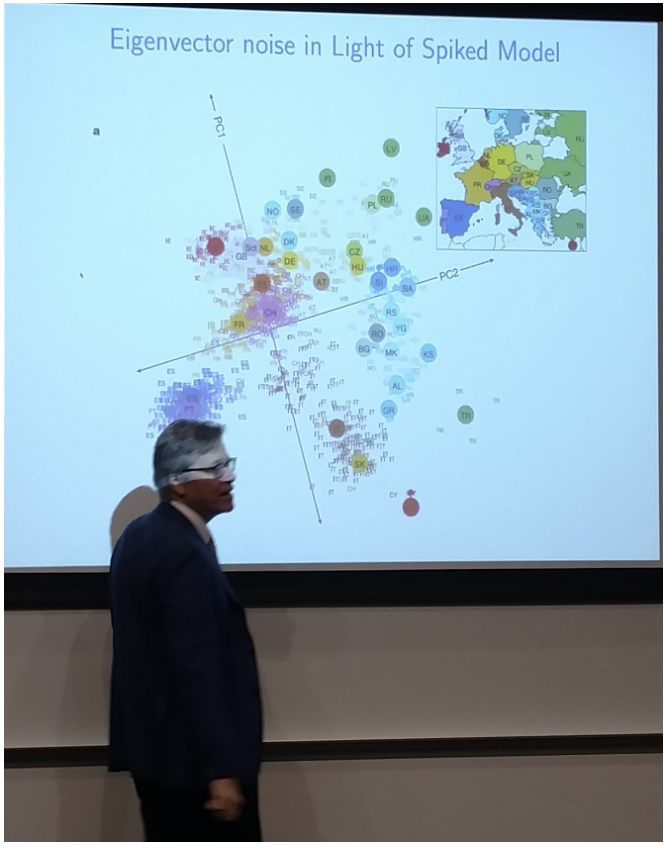
How to choose the ideal variance (singular value) cut off?

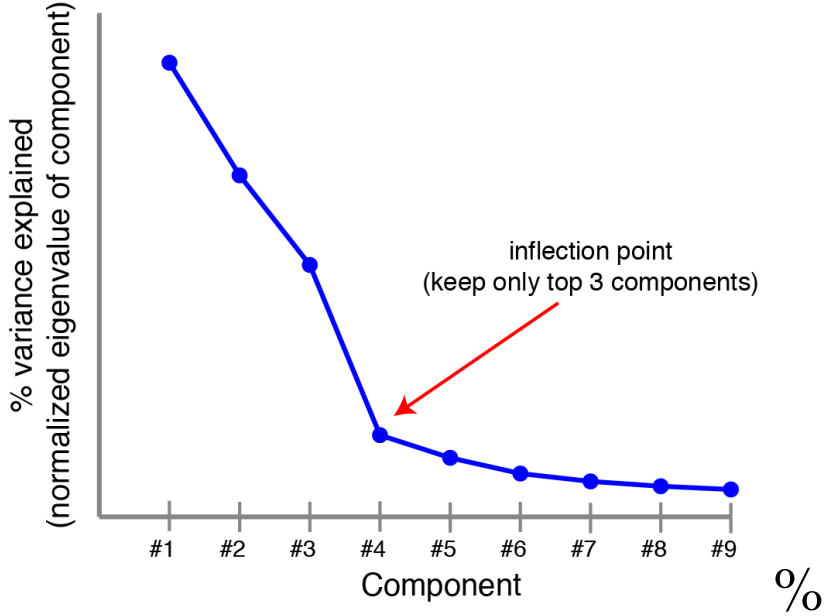




Principal Component Analysis

How to choose the ideal variance (singular value) cut off?





Variance Explained

$$T = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

Some Properties/Facts:

Some Properties/Facts:

Good Properties

Some Properties/Facts:

Good Properties

- PCA is invariant to rotation

Some Properties/Facts:

Good Properties

- PCA is invariant to rotation
- PCA minimizes $\sum \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|$, $\hat{\mathbf{x}}_i = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top \mathbf{x}_i$ (sum of residuals)

Some Properties/Facts:

Good Properties

- PCA is invariant to rotation
- PCA minimizes $\sum \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|$, $\hat{\mathbf{x}}_i = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top \mathbf{x}_i$ (sum of residuals)
- PCA minimizes $\sum \|\mathbf{x}_i - \mathbf{x}_j\| - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|)$ (distance preserving)

Some Properties/Facts:

Good Properties

- PCA is invariant to rotation
- PCA minimizes $\sum \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|$, $\hat{\mathbf{x}}_i = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top \mathbf{x}_i$ (sum of residuals)
- PCA minimizes $\sum \|\mathbf{x}_i - \mathbf{x}_j\| - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|)$ (distance preserving)

Properties to Concern About

Some Properties/Facts:

Good Properties

- PCA is invariant to rotation
- PCA minimizes $\sum \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|$, $\hat{\mathbf{x}}_i = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top \mathbf{x}_i$ (sum of residuals)
- PCA minimizes $\sum \|\mathbf{x}_i - \mathbf{x}_j\| - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|)$ (distance preserving)

Properties to Concern About

- PCA is not scale invariant (scale + PCA $\neq$ PCA + scale)

Some Properties/Facts:

Good Properties

- PCA is invariant to rotation
- PCA minimizes $\sum \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|$, $\hat{\mathbf{x}}_i = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top \mathbf{x}_i$ (sum of residuals)
- PCA minimizes $\sum \|\mathbf{x}_i - \mathbf{x}_j\| - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|)$ (distance preserving)

Properties to Concern About

- PCA is not scale invariant (scale + PCA $\neq$ PCA + scale)
- The new coordinates have no semantic meaning

Some Properties/Facts:

Good Properties

- PCA is invariant to rotation
- PCA minimizes $\sum \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|$, $\hat{\mathbf{x}}_i = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top \mathbf{x}_i$ (sum of residuals)
- PCA minimizes $\sum \|\mathbf{x}_i - \mathbf{x}_j\| - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|)$ (distance preserving)

Properties to Concern About

- PCA is not scale invariant (scale + PCA $\neq$ PCA + scale)
- The new coordinates have no semantic meaning
- PCA is sensitive to outliers

PCA via SVD

PCA via SVD

Given $\mathbf{X}$, a $d \times n$ (centered) data matrix, consider the two symmetric matrices $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$.

PCA via SVD

Given $\mathbf{X}$, a $d \times n$ (centered) data matrix, consider the two symmetric matrices $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$.

It can be shown that $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$ has the same (non-zero) eigenvalues $\lambda_1, \ldots, \lambda_d$ (assuming $d < n$).

PCA via SVD

Given $\mathbf{X}$, a $d \times n$ (centered) data matrix, consider the two symmetric matrices $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$.

It can be shown that $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$ has the same (non-zero) eigenvalues $\lambda_1, \ldots, \lambda_d$ (assuming $d < n$).

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{D}\mathbf{U}^\top \text{ and } \mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$$

PCA via SVD

Given $\mathbf{X}$, a $d \times n$ (centered) data matrix, consider the two symmetric matrices $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$.

It can be shown that $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$ has the same (non-zero) eigenvalues $\lambda_1, \ldots, \lambda_d$ (assuming $d < n$).

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{D}\mathbf{U}^\top \text{ and } \mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$$

Singular Value Decomposition (SVD)

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

where $\mathbf{\Sigma}$ is a diagonal matrix with non-zero entries $\sqrt{\lambda_i}$.

PCA via SVD

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

PCA via SVD

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

$$\mathbf{U}^\top\mathbf{X} = \mathbf{\Sigma}\mathbf{V}^\top$$

PCA via SVD

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

$$\mathbf{U}^\top\mathbf{X} = \mathbf{\Sigma}\mathbf{V}^\top$$

Therefore,

$\mathbf{U}$: the principal directions

$\mathbf{\Sigma}\mathbf{V}^\top$: the coordinates of the data in the PCA space

Centering the Data Matrix

All the development so far assumes that $\mathbf{X}$ is centered (average of each row is zero).

Centering the Data Matrix

All the development so far assumes that $\mathbf{X}$ is centered (average of each row is zero).

When $\mathbf{X}$ is not centered, we can center it doing:

Centering the Data Matrix

All the development so far assumes that $\mathbf{X}$ is centered (average of each row is zero).

When $\mathbf{X}$ is not centered, we can center it doing:

$$\overline{\mathbf{X}} = \begin{bmatrix} \overline{x}_{1:} & \dots & \overline{x}_{1:} \\ \vdots & \ddots & \vdots \\ \overline{x}_{d:} & \dots & \overline{x}_{d:} \end{bmatrix} = \begin{bmatrix} \overline{x}_{1:} & & & \\ & \overline{x}_{2:} & & \\ & & \ddots & \\ & & & \overline{x}_{d:} \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$$

where $\overline{x}_{i:}$ is the average of the ith row of $\mathbf{X}$.

Centering the Data Matrix

All the development so far assumes that $\mathbf{X}$ is centered (average of each row is zero).

When $\mathbf{X}$ is not centered, we can center it doing:

$$\overline{\mathbf{X}} = \begin{bmatrix} \overline{x}_{1:} & \dots & \overline{x}_{1:} \\ \vdots & \ddots & \vdots \\ \overline{x}_{d:} & \dots & \overline{x}_{d:} \end{bmatrix} = \begin{bmatrix} \overline{x}_{1:} & & & \\ & \overline{x}_{2:} & & \\ & & \ddots & \\ & & & \overline{x}_{d:} \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$$

where $\overline{x}_{i:}$ is the average of the ith row of $\mathbf{X}$.

A centered version $\tilde{\mathbf{X}}$ of $\mathbf{X}$ is given by:

$$\tilde{\mathbf{X}} = \mathbf{X} - \overline{\mathbf{X}}$$

Recap

Summary of the Lecture:

Recap

Summary of the Lecture:

- The principal directions are the eigenvectors of $\mathbf{X}\mathbf{X}^\top$

Recap

Summary of the Lecture:

- The principal directions are the eigenvectors of $\mathbf{X}\mathbf{X}^\top$
- The first principal directions (corresponding to the largest eigenvalues) concentrate most of the variance/information

Recap

Summary of the Lecture:

- The principal directions are the eigenvectors of $\mathbf{X}\mathbf{X}^\top$
- The first principal directions (corresponding to the largest eigenvalues) concentrate most of the variance/information
- Attributes are decorrelated in the PCA space (feature space)

Recap

Summary of the Lecture:

- The principal directions are the eigenvectors of $\mathbf{X}\mathbf{X}^\top$
- The first principal directions (corresponding to the largest eigenvalues) concentrate most of the variance/information
- Attributes are decorrelated in the PCA space (feature space)
- Noise can be removed by cutting off principal directions with low variance

Recap

Summary of the Lecture:

- The principal directions are the eigenvectors of $\mathbf{X}\mathbf{X}^\top$
- The first principal directions (corresponding to the largest eigenvalues) concentrate most of the variance/information
- Attributes are decorrelated in the PCA space (feature space)
- Noise can be removed by cutting off principal directions with low variance
- SVD decomposition provides the PCA analysis