Credits for some of the material in these slides go to:

Prof. Pier Luca Lanzi lectures on DATA MINING

Prof. Shireen Elhabian and Prof. Aly Farag A TUTORIAL ON DATA REDUCTION

Principal Component Analysis

Matteo Matteucci, PhD (matteo.matteucci@polimi.it)

Artificial Intelligence and Robotics Laboratory
Politecnico di Milano

Dimensionality Reduction

We have already encountered this idea before to remove unnecessary features dealing with the bias-variance trade-off in regression …

Dimensionality reduction can be used with several aims:

- May help to eliminate irrelevant features or reduce noise
- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized

Principal Component Analysis is just one of the possible techniques to perform dimensionality reduction … it is linear and easy to understand!

Variance and Spread

Variance is a measure of the spread of the data along dimension X_i having mean $\overline{X_i}$ (claimed to be the original measure of data variability)

$$\sigma_{ii} = \sigma_i^2 = \frac{\sum_{n=1}^{N}\left(X_{ni} - \overline{X_i}\right)^2}{N-1} = \frac{\sum_{n=1}^{N}\left(X_{ni} - \overline{X_i}\right)\left(X_{ni} - \overline{X_i}\right)}{N-1}$$

Covariance is a measure of how much *each of the dimensions* varies from the mean *with respect to each other*.

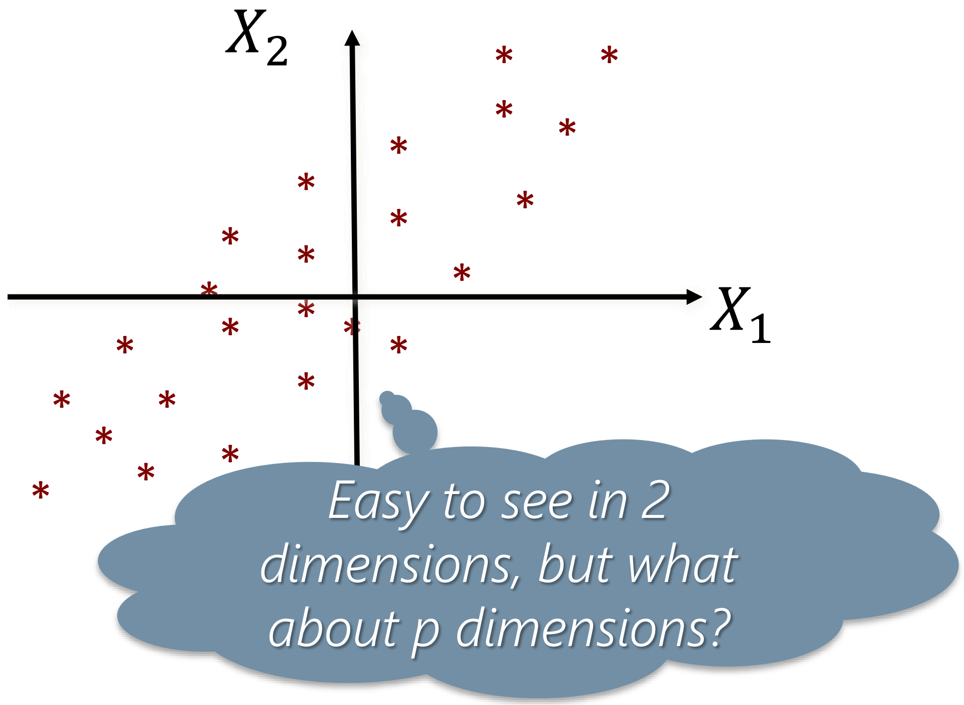
$$\sigma_{ij} = \frac{\sum_{n=1}^{N}\left(X_{ni} - \overline{X_i}\right)\left(X_{nj} - \overline{X_j}\right)}{N-1}$$

Covariance is measured between 2 dimensions to see if there is a relationship between the spread in the 2 dimensions …

Covariance Interpretation

Say you have a 2-dimensional data set

- X_1: number of hours studied for a subject
- X_2: marks obtained in that subject



Assume the covariance is: 104.53

What does this value mean?

- Exact value is not as important as its sign
- A positive value indicates that both dimensions increase or decrease together
- A negative value indicates while one increases the other decreases
- If covariance is zero the two dimensions are independent of each other

Covariance Matrix (1/2)

Covariance Matrix represents covariance, i.e., dependency/redundancy, among data dimensions

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$$

Properties:

- Diagonal σ_{ii} represents variance of X_i variable
- σ_{ij} represents covariance between X_i and X_j variables
- $\sigma_{ij} = \sigma_{ji}$, hence matrix is symmetrical about the diagonal
- p-dimensional data will result in a $p \times p$ covariance matrix

Covariance Matrix (2/2)

Let's consider zero mean data in the form of $N \times p$ data matrix X

- Columns of X correspond to all observed measurements of an attribute X_j
- Rows of X correspond to the measurements from each data point X_i

We can write the $p \times p$ covariance matrix Σ_X of attributes from the data

$$\Sigma_X = \frac{1}{N-1}(X - \overline{X})^T (X - \overline{X}) = \frac{1}{N-1} X^T X$$

- The diagonal terms of Σ_X are the variances of the attributes
- The off-diagonal terms of Σ_X are the covariances between the attributes

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad N=3 \qquad P=1$$

$$\frac{1}{N-1}\left[X - \bar{X}\right]^T \left[X - \bar{X}\right] =$$

$$= \frac{1}{N-1}\begin{bmatrix} x_1-\bar{x} & x_2-\bar{x} & x_3-\bar{x} \end{bmatrix}\begin{bmatrix} x_1-\bar{x} \\ x_2-\bar{x} \\ x_3-\bar{x} \end{bmatrix} =$$

$$= \frac{1}{N-1}\left((x_1-\bar{x})(x_1-\bar{x}) + (x_2-\bar{x})(x_2-\bar{x}) + (x_3-\bar{x})(x_3-\bar{x})\right)$$

$$= \frac{1}{N-1}\left((x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2\right) =$$

$$= \frac{1}{N-1}\sum_n (x_n-\bar{x})^2 \quad \leftarrow \text{VARIANCE}$$

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \end{bmatrix} \qquad N = 3 \quad P = 2$$

$$\frac{1}{N-1}\left[X - \bar{X}\right]^T \left[X - \bar{X}\right] =$$

$$= \frac{1}{N-1}\begin{bmatrix} x_1 - \bar{x} & x_2 - \bar{x} & x_3 - \bar{x} \\ y_1 - \bar{y} & y_2 - \bar{y} & y_3 - \bar{y} \end{bmatrix} \begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ x_2 - \bar{x} & y_2 - \bar{y} \\ x_3 - \bar{x} & y_3 - \bar{y} \end{bmatrix} =$$

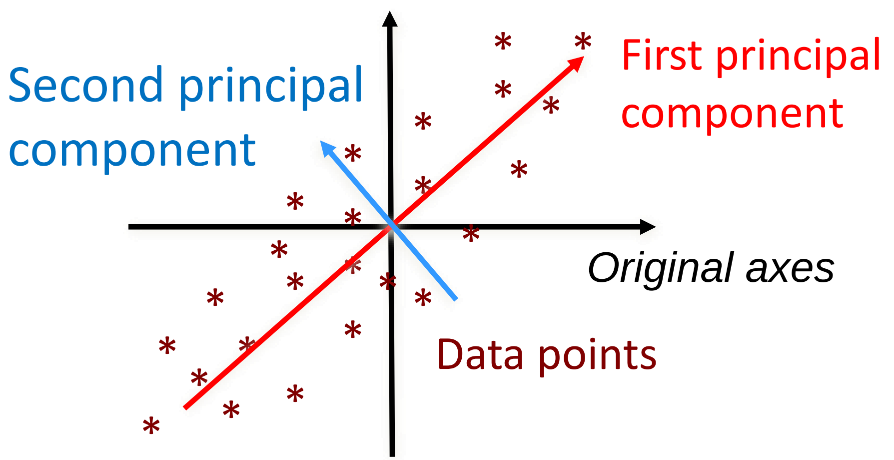
$$= \frac{1}{N-1}\begin{bmatrix} (x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 & (x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + (x_3-\bar{x})(y_3-\bar{y}) \\ (y_1-\bar{y})(x_1-\bar{x}) + (y_2-\bar{y})(x_2-\bar{x}) + (y_3-\bar{y})(x_3-\bar{x}) & (y_1-\bar{y})^2 + (y_2-\bar{y})^2 + (y_3-\bar{y})^2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{N-1}\sum_n (x_n - \bar{x})^2 & \frac{1}{N-1}\sum_n (x_n-\bar{x})(y_n-\bar{y}) \\ \frac{1}{N-1}\sum_n (y_n-\bar{y})(x_n-\bar{x}) & \frac{1}{N-1}\sum_n (y_n-\bar{y})^2 \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}$$

Principal Component Analysis (PCA)

Given N data vectors $X \in \Re^p$ find $k \leq p$ orthogonal vectors, i.e., the principal components, that can be best used to represent data

- The first principal component is the *normalized linear combination* of the features that has *maximal variance* (captures the highest variability in data)
- The second principal components is the linear combination that has *maximal variance* among all combinations *uncorrelated* to the first one



$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p; \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

$$Z_2 \perp Z_1$$

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \cdots + \phi_{p2}X_p; \sum_{j=1}^{p} \phi_{j2}^2 = 1$$

First Principal Component

The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features with the largest variance:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Some important notes about this definition

- Elements $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ are called *loadings* of the first principal component
- Loadings make the $\phi_1 = [\phi_{11}\ \phi_{21}\ \ldots\ \phi_{p1}]^T$ the *principal component vector*
- Normalized means $\sum_{j=1}^{p} \phi_{j1}^2 = \phi_1^T \phi_1 = 1$, otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance

Computing the First Principal Component (1/2)

Suppose we have a $N \times p$ data set X in the form of a matrix with rows representing our data. Each point ***principal score*** is defined as:

$$Z_{n1} = \phi_{11}X_{n1} + \phi_{21}X_{n2} + \cdots + \phi_{p1}X_{np}; \quad with \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

If we force each of the features X_j to have zero mean, so does Z_1 (for any values of loadings ϕ_{j1}), the sample variance of Z_{n1} can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\left(Z_{n1} - \overline{Z_1}\right)^2 = \frac{1}{N}\sum_{n=1}^{N} Z_{n1}^2 = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{j=1}^{p}\phi_{j1}X_{nj}\right)^2; \quad with \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

Computing the First Principal Component (2/2)

To find the first principal component

$$Z_{n1} = \phi_{11}X_{n1} + \phi_{21}X_{n2} + \cdots + \phi_{p1}X_{np}; \quad with \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

we have to find

$$argmax_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{j=1}^{p}\phi_{j1}X_{nj}\right)^2; \quad subject\ to \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

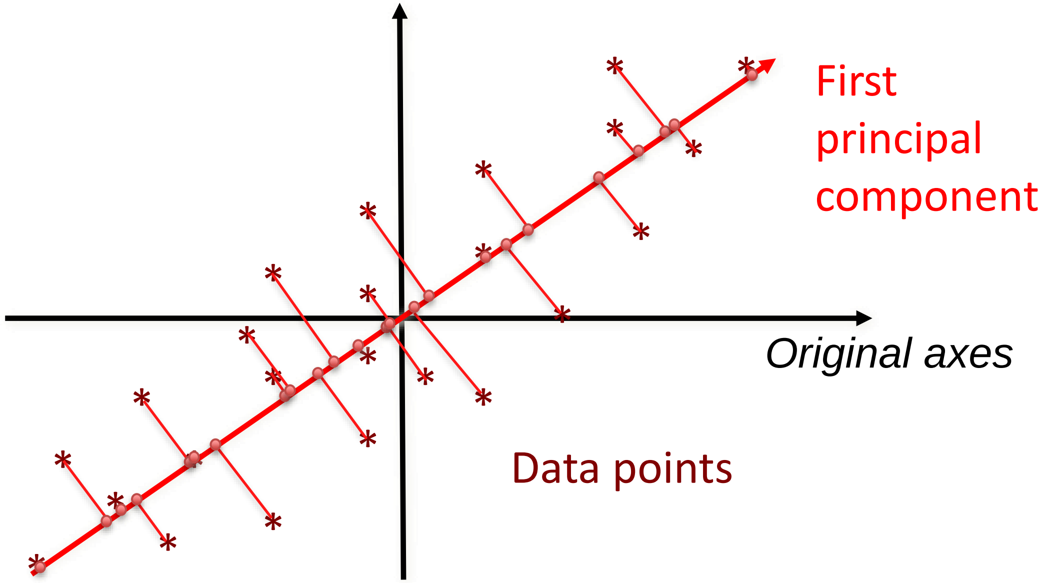
This can be solved via Singular Value Decomposition (SVD) of matrix X

More on this later …

Geometric Interpretation of the First Principal Component

The loading vector $\phi_1 = [\phi_{11}\ \phi_{21}\ \dots\ \phi_{p1}]$ defines the direction in feature space along which the data vary the most

If we project the N data points $X \in \Re^p$ onto this direction, the projected values are the principal component scores $Z_{11}, \dots, Z_{n1}$ themselves.



Best linear projection on a one-dimensional subspace of the original dataset, i.e., it preserves most of the variance/spread in the data

Further Principal Components

Second principal component Z_2 is the linear combination of $X_1, X_2, \ldots, X_p$ with maximal variance among all combinations uncorrelated with Z_1

$$Z_{n2} = \phi_{12}X_{n1} + \phi_{22}X_{n2} + \cdots + \phi_{p2}X_{np}; \quad \sum_{j=1}^{p} \phi_{j2}^2 = 1$$

with second principal component scores $Z_{12}, \ldots, Z_{n2}$, and second principal component loading vector $\phi_2 = [\phi_{12}\ \phi_{22}\ \ldots\ \phi_{p2}]$

More on this later ...

There are at most $\min(N-1, p)$ principal components, sometimes less

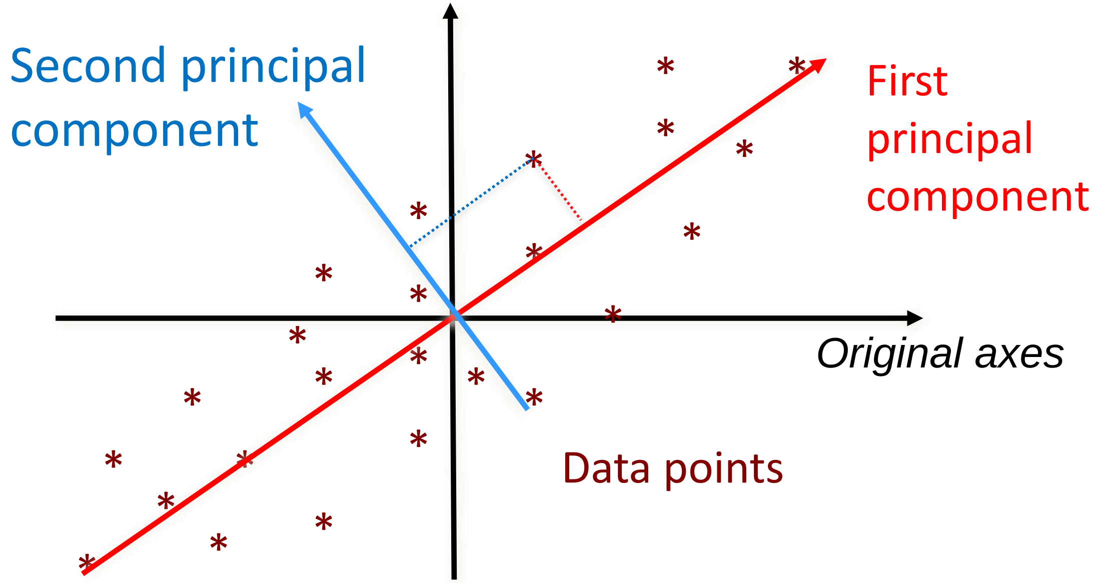
The principal component directions $\phi_1, \phi_2, \ldots, \phi_{\min(N-1,p)}$ are the right singular vectors of the data matrix X and the component variances are $1/N$ times the squares of the singular values

More on this later ...

Geometric Interpretation of PCA (continued)

The loading vector $\phi_1 = [\phi_{11}\ \phi_{21}\ \dots\ \phi_{p1}]$ defines a direction in feature space along which the data vary the most, the loading vector $\phi_2 = [\phi_{12}\ \phi_{22}\ \dots\ \phi_{p2}]$ defines an orthogonal direction.

The principal component scores $Z_{11}, \dots, Z_{n1}$ and $Z_{12}, \dots, Z_{n2}$ are the points coordinates in the new reference system (subspace) defined by the principal components.



The relationship between the subspaces is a rotation with a stretch, you have also a projection if $k < p$

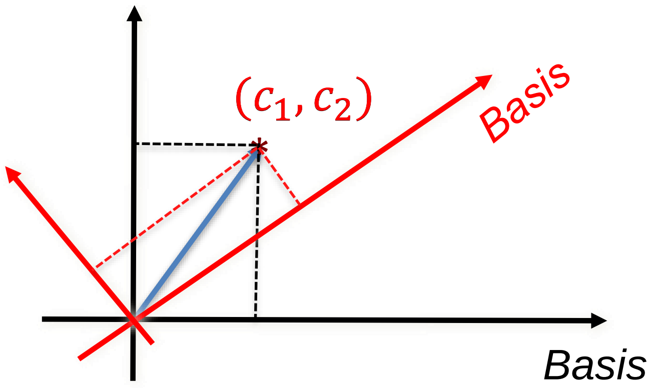
Change of Basis

A span of a set of vectors $X_1, X_2, \ldots, X_p$ is the set of vectors that can be written as a linear combination of $X_1, X_2, \ldots, X_p$

$$span(X_1, X_2, \ldots, X_p) = \{c_1 X_1 + c_2 X_2 + \cdots + c_p X_p | c_1, c_2, \ldots, c_p \in \Re\}$$

A basis for $\Re^p$ is a set of vectors which

- Spans $\Re^p$, i.e., any vector in the p-dimensional space can be written as linear combination of these vectors.
- Are linearly independent, i.e., orthogonal

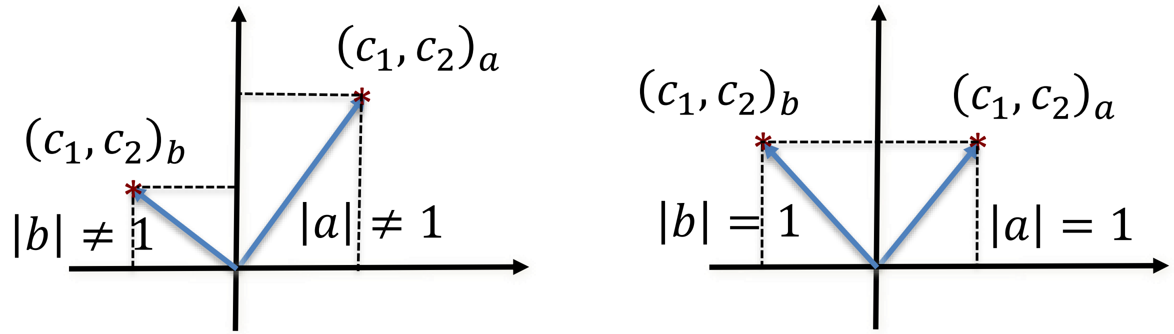


Any set of p-linearly independent vectors, i.e., orthogonal vectors, form a basis vectors for $\Re^p$

Orthogonal/Orthonormal Basis

Two vectors are *orthogonal* if their inner product is zero.

$$a^T b = \sum_{i=1}^{p} a_i b_i = 0$$



An *orthonormal* basis of a vector space V, with inner product, is a set of basis vectors whose elements are orthogonal and of magnitude 1

- To change the vectors of an orthogonal basis into an orthonormal basis just multiply by the inverse of their norm
- The standard basis of the p-dimensional Euclidean space $\Re^p$, i.e.,$(1,0)(0,1)$, is an example of orthonormal (and ordered) basis

PCA as a Change of Basis

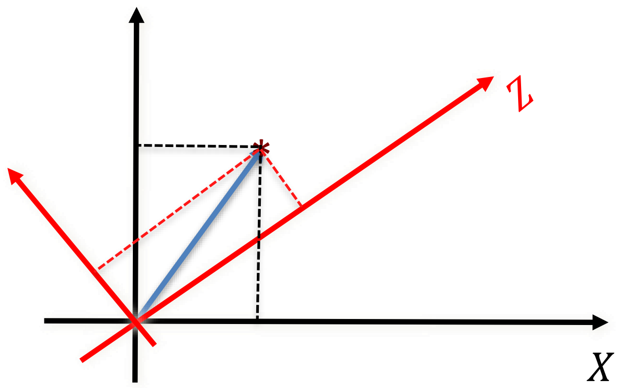
Let X and Z be two $N \times p$ matrices related by a linear transformation Φ, being X the original recoded dataset and Z a re-representation of it

$$Z = X\Phi$$

being $\phi_1, \phi_2, \dots, \phi_p$ the columns of Φ, $\mathrm{X_n}$ the rows of X, Z_n the rows of Z

What we have here is that:

- Φ is a matrix that transforms X into Z
- Geometrically, Φ is a *rotation* and a *stretch* (scaling)
- The columns of Φ, $[\phi_1^T \ \phi_2^T \ \dots \phi_p^T]$ are a set of new basis vectors for expressing the rows of X



PCA as a Change of Basis

Let X and Z be two $N \times p$ matrices related by a linear transformation Φ, being X the original recoded dataset and Z a re-representation of it

$$Z = X\Phi$$

$$Z = [X_1 \quad \cdots \quad X_p][\phi_1 \quad \cdots \quad \phi_p]$$

Projection of X_i point on the p components, i.e., on the new $\phi_1, \dots, \phi_p$ basis

$$Z = \begin{bmatrix} \phi_1 X_{11} & \cdots & \phi_p X_{1p} \\ \vdots & \ddots & \vdots \\ \phi_1 X_{N1} & \cdots & \phi_p X_{Np} \end{bmatrix}$$

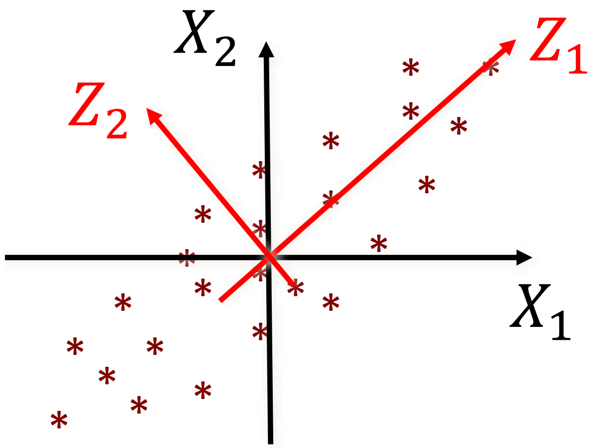
How do we select the new basis?

It does not change the data, just the representation. If $\phi_1, \dots, \phi_p$ are orthonormal we have a pure rotation, otherwise we have also a stretch

How to select the new basis?

PCA extracts relevant information from the given data, i.e., removes redundant information, while retaining the maximum information.

Uncorrelated signals have no redundancy, while correlated signals introduce redundancy



Information can be represented by the spread of the data, or as signal-to-noise ratio (SNR)

$$SNR = \sigma^2_{signal} / \sigma^2_{\text{noise}}$$

Principal components have high SNR, i.e., **high variance**, and they are orthogonal, i.e., have **low redundancy**

Data Covariance Matrix and Change of Basis

Suppose we can manipulate Σ_X via the change of basis

$$Z = X\Phi$$

Our goals are to find the Φ so that covariance matrix Σ_Z

1. Shows minimal redundancy as measured by off-diagonal elements, i.e. we would like each variable to co-vary as little as possible with other variables, so to minimize data redundancy
2. Maximizes the signal measured by variance terms on the diagonal, so to maximize signal-to-noise ratio

The optimized covariance matrix Σ_Z should be a diagonal matrix

PCA and Diagonalization (1/2)

To compute X principal components Φ we want Σ_Z to become diagonal

$$\Sigma_Z = \frac{1}{N-1} Z^T Z = \frac{1}{N-1} (X\Phi)^T (X\Phi)$$

$$= \frac{1}{N-1} \Phi^T X^T X \Phi$$

I suppose you know about eigenvectors and eigenvalues ☹

Let ϕ_j composed by the $X^T X$ eigenvectors

We know that $X^T X$ is *symmetric*, and it can be diagonalized by the orthogonal matrix V formed with its $r \leq p$ eigenvectors arranged by columns, where r is the rank of $X^T X$

$$X^T X = V \mathrm{D} V^T$$

Matrix D is the diagonal matrix of $X^T X$ eigenvalues

PCA and Diagonalization (2/2)

By choosing ϕ_j as the set of $X^T X$ eigenvectors, we get $\Phi = V$

$$\Sigma_Z = \frac{1}{N-1}(X\Phi)^T(X\Phi) = \frac{1}{N-1}\Phi^T X^T X\Phi$$

$$= \frac{1}{N-1}\Phi^T V D V^T \Phi = \frac{1}{N-1}\Phi^T(\Phi \mathrm{D} \Phi^T)\Phi$$

$$= \frac{1}{N-1}\left(\Phi^{\mathrm{T}}\Phi\right)\mathrm{D}\left(\Phi^{\mathrm{T}}\Phi\right)$$

Recall the constrain
$\sum_{j=1}^{p} \phi_{j1}^2 = 1$

In an *orthonormal* basis we have for the transpose $\Phi^{\mathrm{T}}\Phi = \Phi^{-1}\Phi = I$

$$\Sigma_Z = \frac{1}{N-1}\left(\Phi^{\mathrm{T}}\Phi\right)D\left(\Phi^{\mathrm{T}}\Phi\right) = \frac{1}{N-1}\mathrm{D}$$

Selecting ϕ_j to be $X^T X$ eigenvectors works!

PCA and Singular Values Decomposition (1/5)

The Singular Values Decomposition (SVD) of a $N \times p$ matrix A is:

$$A = U\Lambda \mathrm{V}^{\mathrm{T}}$$

Where we have:

- U is the $\mathrm{N} \times r$ orthonormal matrix, i.e., $U^T U = I$, of AA^T eigenvectors
- V is the $\mathrm{r} \times p$ orthonormal matrix, i.e, $V^T V = I$, of $A^T A$ eigenvectors
- Λ is the $r \times r$ diagonal matrix of the squared root eigenvalues of A arranged in *non increasing order*
- r is the rank of A, i.e., the number of linearly independent columns

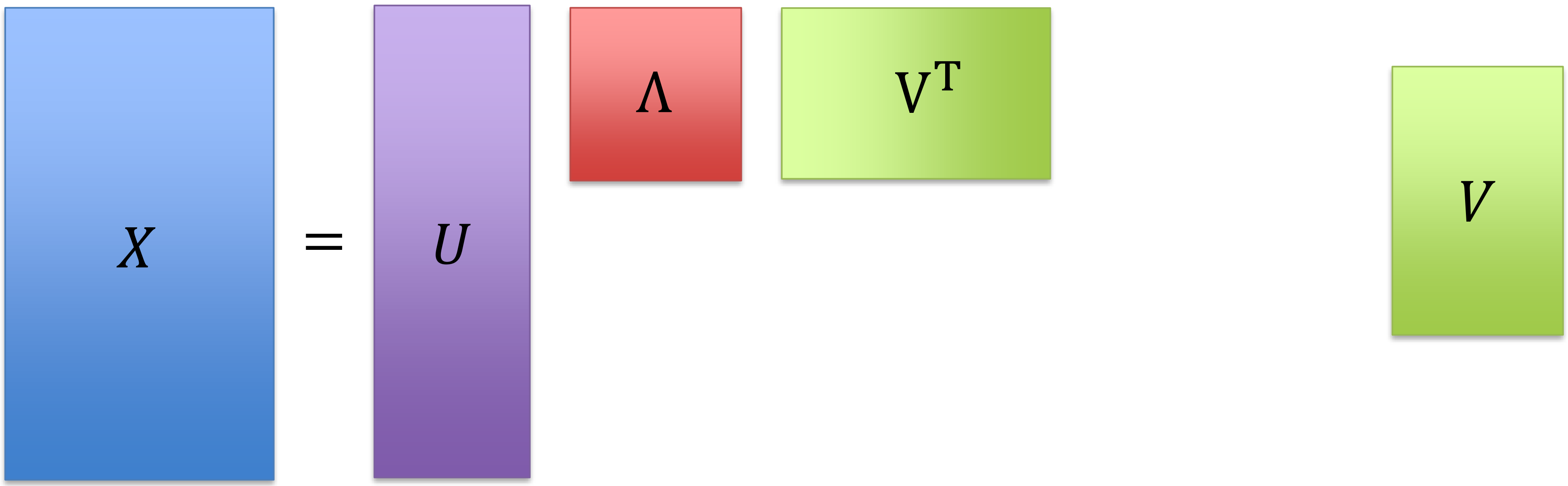
Note that from $X = U\Lambda V^T$ we get the previous result

$$X^T X = (U\Lambda V^T)^T (U\Lambda V^T) = V\Lambda U^T U\Lambda V^T = V\Lambda\Lambda V^T = VDV^T$$

PCA and Singular Values Decomposition (2/5)

If we compute the SVD of a X then we can get $\Phi = V$:

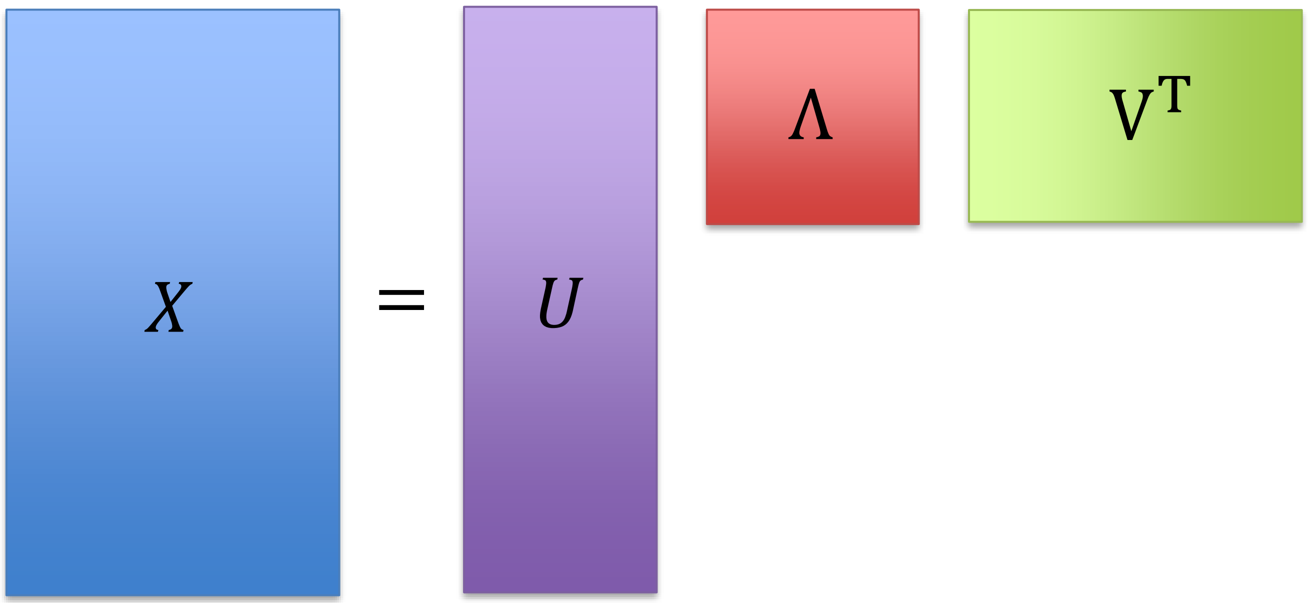
$$X = U\Lambda V^{T} \qquad\qquad \Phi = V$$



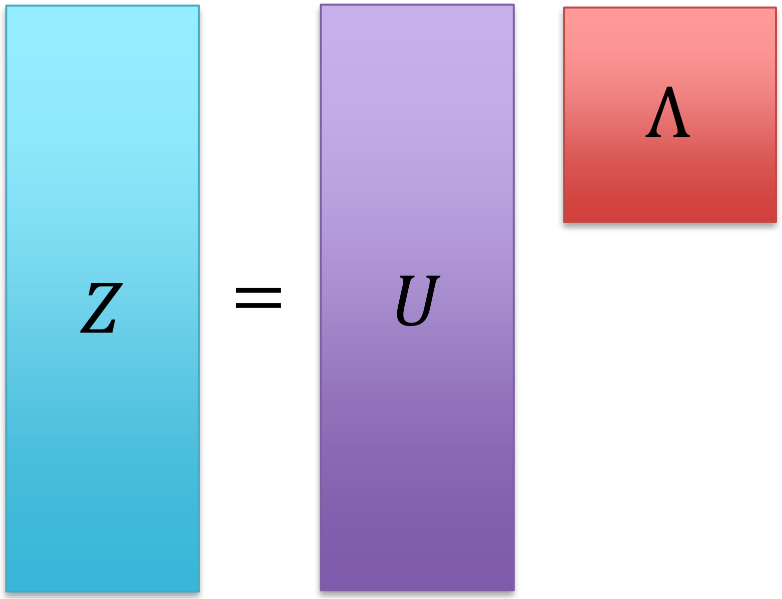
PCA and Singular Values Decomposition (2/5)

If we compute the SVD of a X then we can get $\Phi = V$:

$$X = U\Lambda V^{T}$$



$$Z = X\Phi = U\Lambda V^{T}V = U\Lambda$$

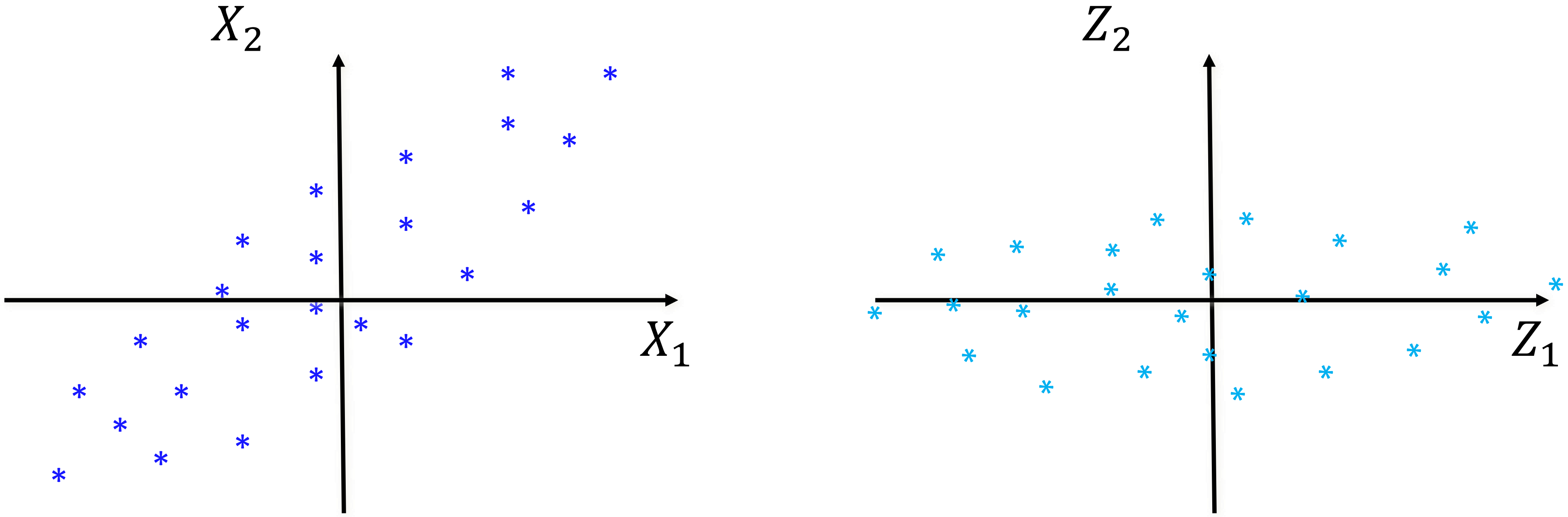


PCA and Singular Values Decomposition (3/5)

If we compute the SVD of a X then we can get $\boldsymbol{\Phi} = V$:

$$X = U\Lambda V^{T}$$

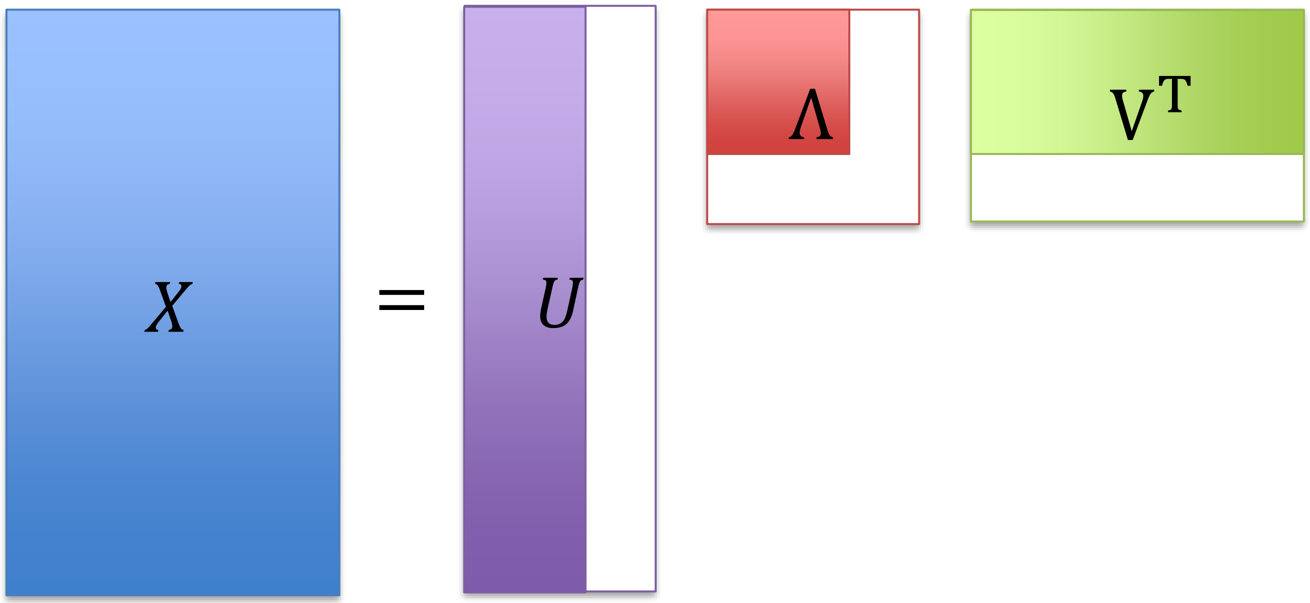
$$Z = X\Phi = U\Lambda V^{T}V = U\Lambda$$



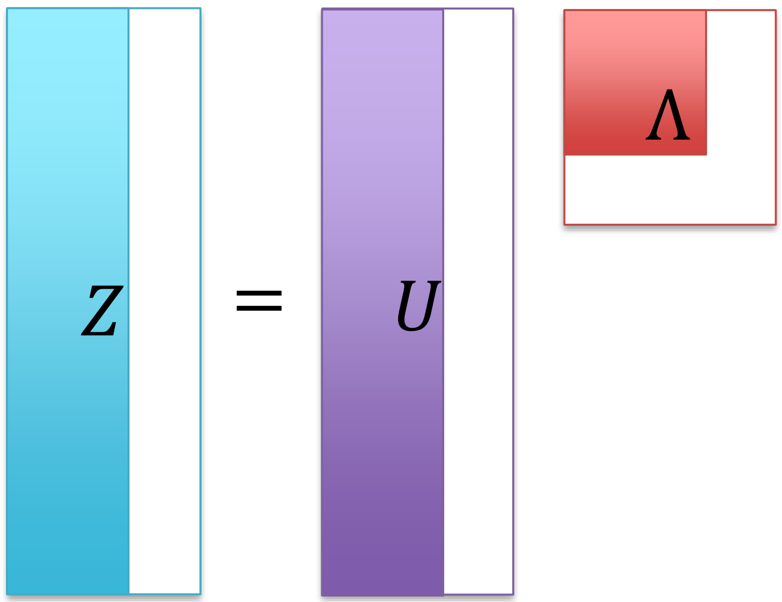
PCA and Singular Values Decomposition (4/5)

We can project X in a lower space selecting $k < r \leq p$ components

$$X = U\Lambda V^{T}$$



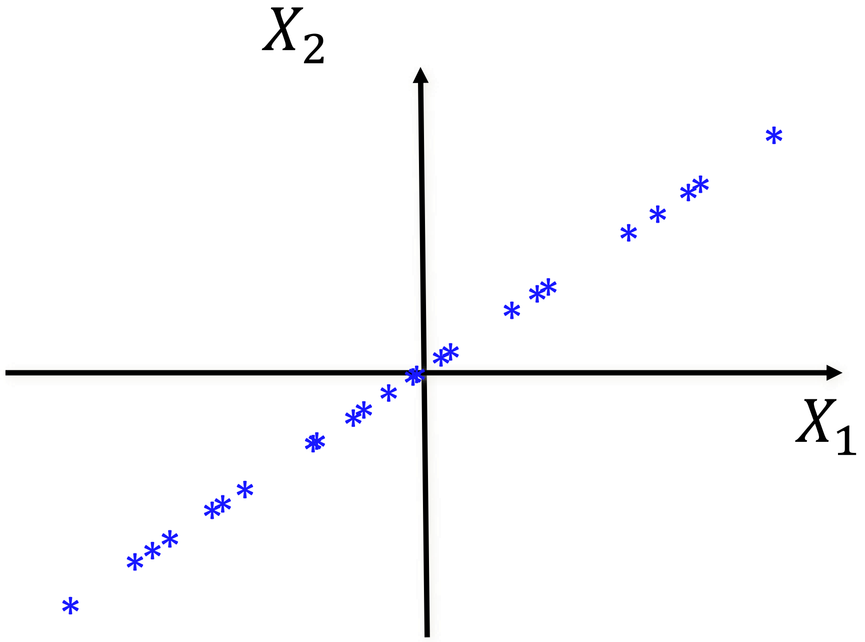
$$Z = X\Phi = U\Lambda V^{T}V = U\Lambda$$



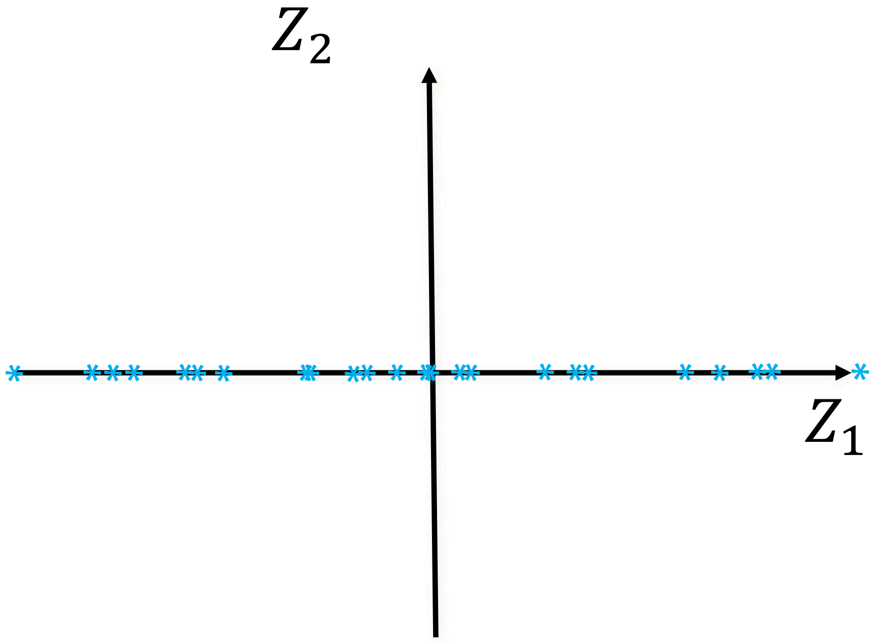
PCA and Singular Values Decomposition (5/5)

We can project X in a lower space selecting $k < r \leq p$ components

$$X = U\Lambda V^{T}$$



$$Z = X\Phi = U\Lambda V^{T}V = U\Lambda$$



Proportion of Variance Explained (1/2)

The total variance present in a data set (assuming that the variables have been centered to have mean zero) is defined as

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{N} \sum_{n=1}^{N} X_{nj}^2$$

the variance explained by the k^{th} principal component is:

$$Var(Z_k) = \frac{1}{N} \sum_{n=1}^{N} Z_{nk}^2$$

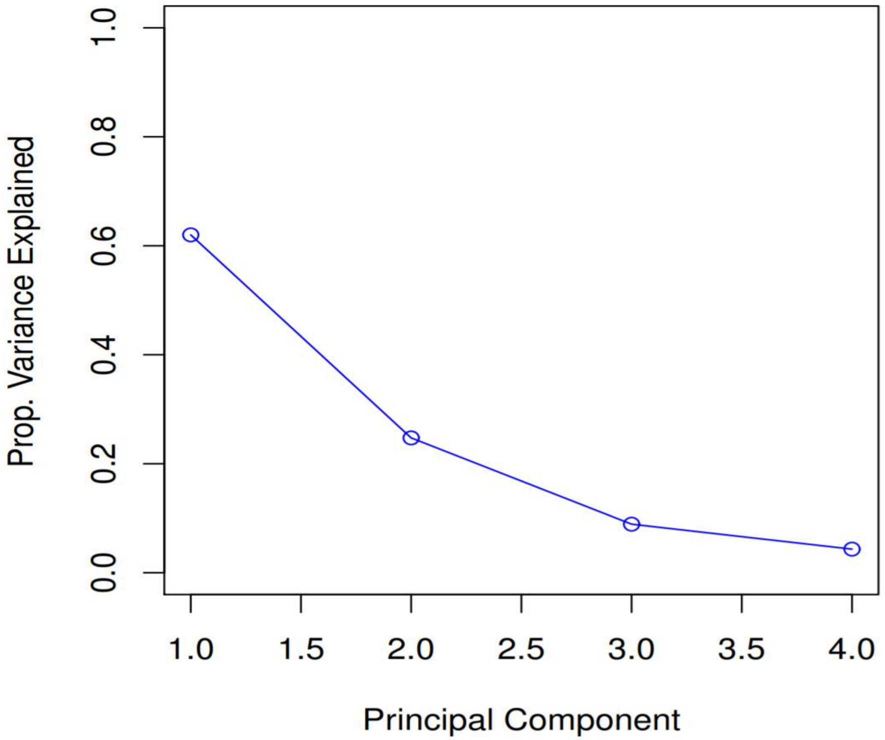
It can be shown that

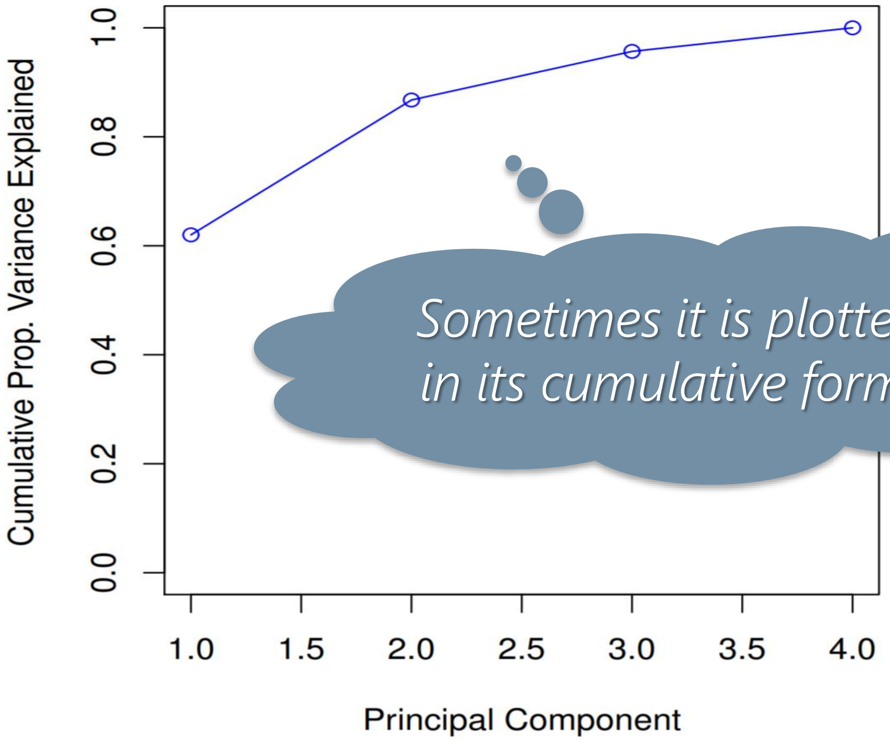
$$\sum_{j=1}^{p} Var(X_j) = \sum_{k=1}^{M} Var(Z_k), \ \ with \quad M = \min(N-1, p)$$

Proportion of Variance Explained (2/2)

Proportion of Variance Explained (PVE) of the k^{th} principal component

$$PVE_k = \frac{Var(Z_k)}{\sum_{j=1}^{p} Var(X_j)} = \frac{\sum_{n=1}^{N} Z_{nk}^2}{\sum_{j=1}^{p} \sum_{n=1}^{N} X_{nj}^2}$$

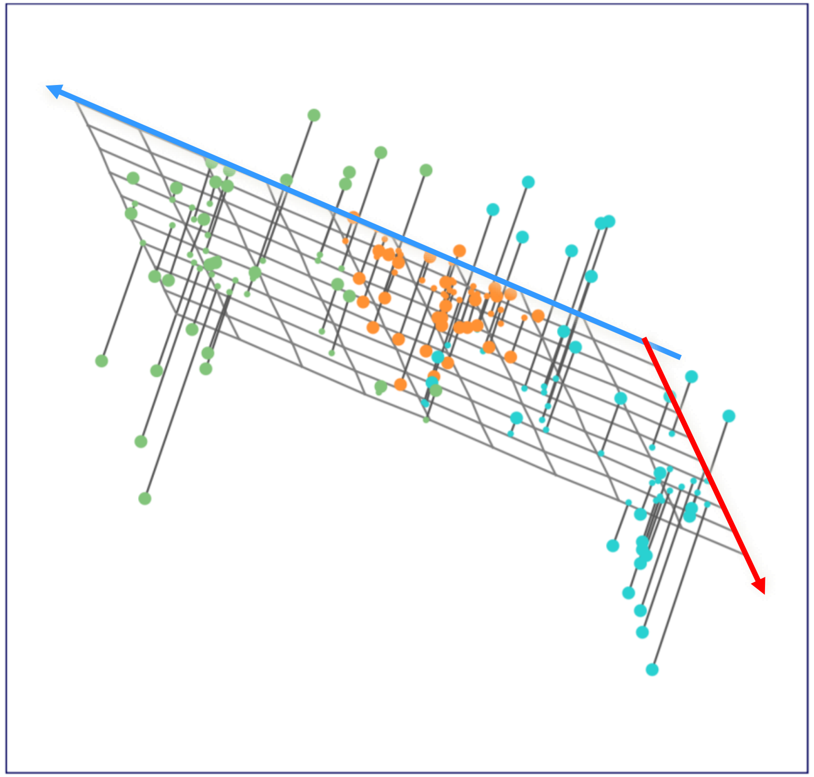


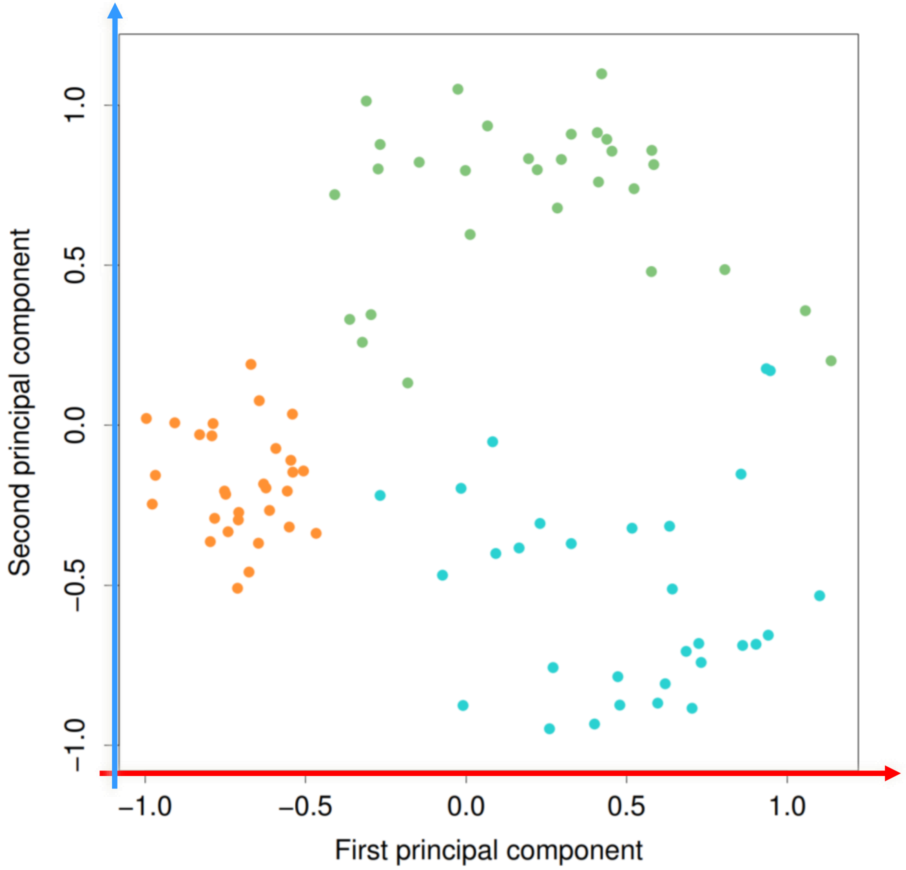


Sometimes it is plotted in its cumulative form.

PCA and Hyperplanes

The first k^{th} principal components $\phi_1 \ldots \phi_k$ define the k-dimensional hyperplane which is closest, in the Euclidean sense, to the N observations





Dimensionality Reduction Uses

Dimensionality reduction can be used with several aims:

- Eliminate irrelevant features or reduce noise
- Remove features which are highly correlated
- Allow data to be more easily visualized
- Avoid curse of dimensionality by projecting a in low dimensionality subspace

Uses which you might think about immediately

- Feature projection before regression -> Principal Component Regression
- Feature projection for 2D/3D visualization -> Clusters preview
- Feature projection before KNN classification
- Feature projection before k-means clustering

Other uses are related to its geometrical meaning ...