INTRODUCTORY APPLIED MACHINE LEARNING

Yan-Fu Kuo

Dept. of Bio-industrial Mechatronics Engineering

National Taiwan University

Today:

- Principal component analysis
- Principal component regression
- Partial least squares regression

Outline

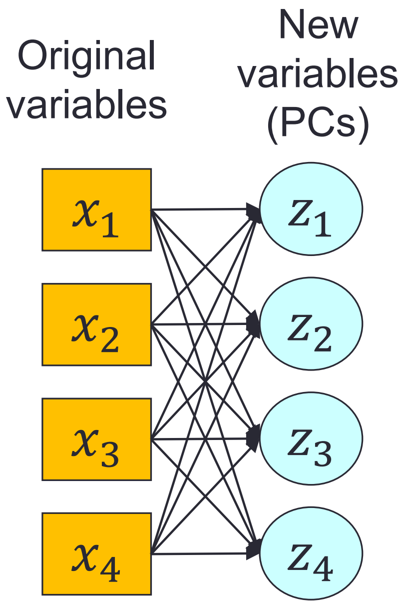
- Goals
- Variance and covariance
- Representing data set by a point, a line, or multiple lines
- Algebraic interpretation
- Properties of principal component analysis (PCA)
- PCA clustering
- Principal component regression (PCR)
- Biased PCR
- Partial least squares regression (PLSR)

Goals

- After this, you should be able to:
  - Calculate principal components (PC) for a set of data
  - Recognize conditions under which principal component analysis (PCA) or partial least squares discriminant analysis (PLSDA) may be useful
  - Perform principal component regression (PCR) and partial least squares regression (PLSR)
  - Select appropriate principal components for your regression model

What Is Principal Component Analysis?

- Principal component analysis (PCA) is a mathematical procedure that converts a set of possibly correlated variables x_i into a set of uncorrelated variables
- The new variables are called principal components (PC) z_i
- The numbers of the original variables and PCs are the same
- The PCs may help in data analysis, model development, and etc.

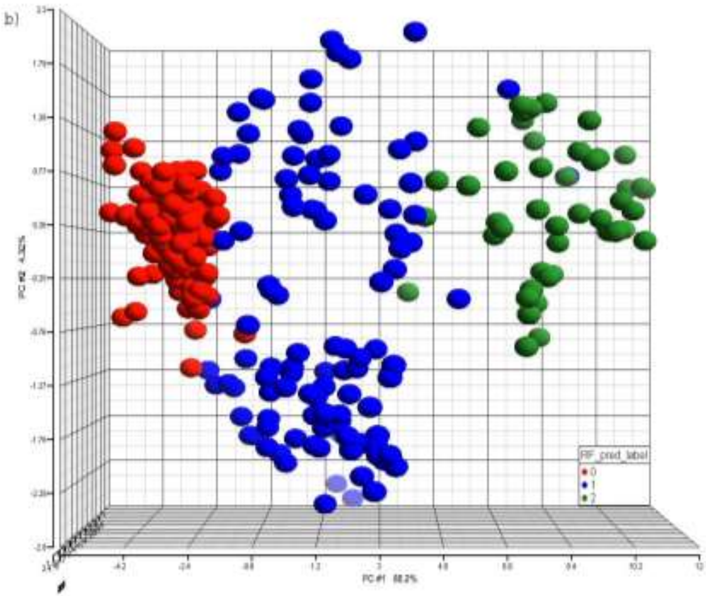


Why PCA? 1. Multicollinearity

- In multivariate analysis, the variables can be correlated to each other
- For example, the data of height, weight, sex, age, and race from a population

- The correlated data can be redundant and can reduce the accuracy of machine learning model
- Instead, use the uncorrelated PCs as the attributes of machine learning models

Why PCA? 2. Dimension Reduction



PC scatter plot

- Not all the PCs are important
- Feature extraction – only keeping a few key feature PCs
- The first PCs from PCA are associated with the largest variance and are usually regarded as key features
- PCA can be used as a clustering tool

History of Principal Component Analysis

- Invented by Pearson (1901) and Hotelling (1933)

- Since 1970 actually used (high performance computer)
- Also named as discrete Karhunen–Loève transform (KLT), proper orthogonal decomposition (POD), and Hotelling transform
- Applications: compression, pattern recognition, spectral image data analysis

Statistical Background

- Variance – a measure of how far a set of numbers are spread out from each other
- Let $x_i \in \Re$, $i = 1 \ldots N$, denote a variable
- The variance of x_i is: $var(x_i) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- Variance operates on 1 dimension, independently of the other dimensions
- Example:

Data set 1 = [0, 8, 12, 20], Mean = 10, Variance = 69.33

Data set 2 = [8, 9, 11, 12], Mean = 10, Variance = 3.33

Statistical Background (Cont'd)

- Covariance – a measure of how much two variables change together
- Suppose there exists two random variables $x_{1i}, x_{2i} \in \Re$, $i = 1 \dots N$, the covariance between x_{1i} and x_{2i} is:

$$cov(x_{1i}, x_{2i}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \in \Re$$

- Positive/negative covariance – higher than average values of one variable tend to be paired with higher/lower than average values of the other variable
- Zero covariance – the two random variables are independent

Statistical Background (Cont'd)

- Covariance matrix – a matrix whose element in the i, j position is the covariance between the ith and jth elements of a <u>random vector</u> $\boldsymbol{x}_i = [x_{1i} \ \dots x_{Mi}]^{\mathrm{T}} \in \Re^M$
- The covariance matrix of the sample matrix $\boldsymbol{X}$ $(i = 1 \dots N)$ of the random vector is:

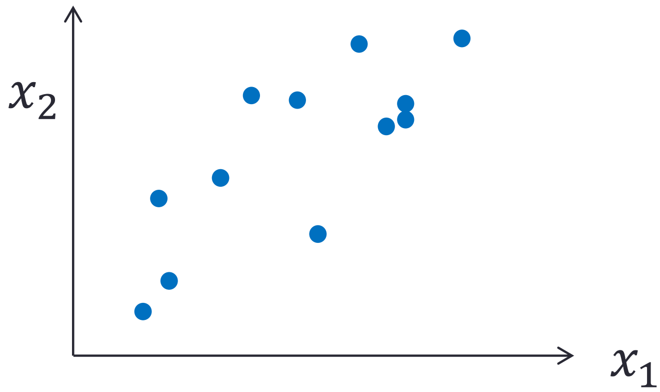
$$cov(\boldsymbol{X}) = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}} = E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^{\mathrm{T}}]$$

$$= \begin{bmatrix} cov(x_{1i}, x_{1i}) & \dots & cov(x_{1i}, x_{Mi}) \\ \vdots & \ddots & \vdots \\ cov(x_{Mi}, x_{1i}) & \dots & cov(x_{Mi}, x_{Mi}) \end{bmatrix} \in \Re^{M \times M}$$

- The matrix $cov(\boldsymbol{X})$ is positive-semidefinite and symmetric

The Goal of PCA – Feature Reduction

- We wish to the underlying variance-covariance structure of a large set of variables through a few linear combinations of these variables
- Suppose we have $\boldsymbol{x}_i \in \Re^M$, $i = 1 \dots N$, sample points in a M-dimensional space
- How does one represent the data set by a point, a line, or multiple lines?



Representing A Data Set with A Point

- If one would like to represent these data set by one point $\boldsymbol{x}_0$, what should it be?

$$\text{The mean } \bar{\boldsymbol{x}} \equiv \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \quad ?$$

- How does one prove it mathematically?
- Problem statement: find $\boldsymbol{x}_0$ such that the cost function

$$s(\boldsymbol{x}_0) \equiv \sum_{i=1}^{N}\|\boldsymbol{x}_i - \boldsymbol{x}_0\|^2$$

is minimized

Solution

$$s(\boldsymbol{x}_0) = \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{x}_0\|^2 = \sum_{i=1}^{N} \|(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) + (\overline{\boldsymbol{x}} - \boldsymbol{x}_0)\|^2$$

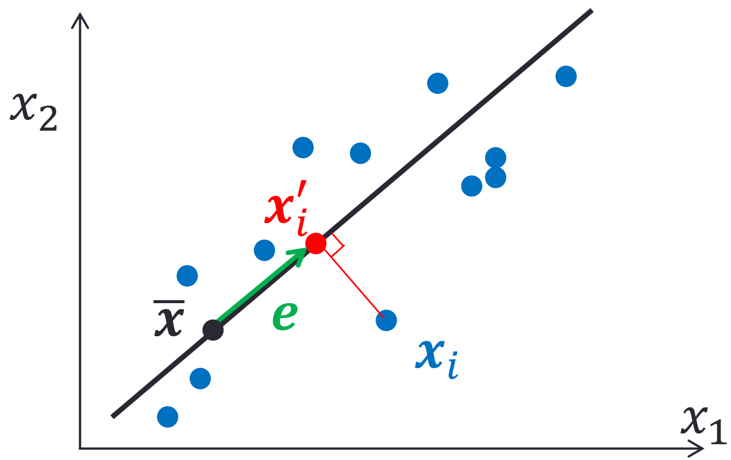
$$= \sum_{i=1}^{N} \|(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\|^2 + 2\sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}} (\overline{\boldsymbol{x}} - \boldsymbol{x}_0) + \sum_{i=1}^{N} \|(\overline{\boldsymbol{x}} - \boldsymbol{x}_0)\|^2$$

$$= \sum_{i=1}^{N} \|(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\|^2 + 2\left(\sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right)^{\mathrm{T}} (\overline{\boldsymbol{x}} - \boldsymbol{x}_0) + N\|(\overline{\boldsymbol{x}} - \boldsymbol{x}_0)\|^2$$

$$= \sum_{i=1}^{N} \|(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\|^2 + N\|(\overline{\boldsymbol{x}} - \boldsymbol{x}_0)\|^2$$

Representing A Dataset with A Line

- The line passes through $\bar{\boldsymbol{x}}$
- Suppose the unit vector of the line is $\boldsymbol{e} \in \Re^M$, $||\boldsymbol{e}||^2 = 1$
- For each point $\boldsymbol{x}_i$, there is a point $\boldsymbol{x}_i'$ on the line that is the projection of $\boldsymbol{x}_i$ to the line
- Every point on the line can be represented as $\boldsymbol{x}_i' = \bar{\boldsymbol{x}} + \alpha_i \boldsymbol{e}$, where $\alpha_i \in \Re$



Problem Statement

Find $\boldsymbol{e}$ such that the cost function

$$s(\alpha_1, \ldots, \alpha_N, \boldsymbol{e}) \equiv \sum_{i=1}^{N} \|\boldsymbol{x}_i' - \boldsymbol{x}_i\|^2 = \sum_{i=1}^{N} \|\overline{\boldsymbol{x}} + \alpha_i \boldsymbol{e} - \boldsymbol{x}_i\|^2$$

is minimized

- We want to prove that $\boldsymbol{e}$ is the eigenvector of $cov(\boldsymbol{X})$
- Recall that $cov(\boldsymbol{X})$ is the covariance matrix

Solving the Optimization Problem

$$s(\alpha_1, \dots, \alpha_N, \boldsymbol{e}) = \sum_{i=1}^{N} \|\overline{\boldsymbol{x}} + \alpha_i \boldsymbol{e} - \boldsymbol{x}_i\|^2 = \sum_{i=1}^{N} \|\alpha_i \boldsymbol{e} - (\boldsymbol{x}_i - \overline{\boldsymbol{x}})\|^2$$

$$= \sum_{i=1}^{N} {\alpha_i}^2 \|\boldsymbol{e}\|^2 - 2\sum_{i=1}^{N} \alpha_i \boldsymbol{e}^{\mathrm{T}} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) + \sum_{i=1}^{N} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}\|^2$$

- Differentiate s against α_i

$$\frac{\partial s(\alpha_1, \dots, \alpha_N, \boldsymbol{e})}{\partial \alpha_i} = 2\alpha_i - 2\boldsymbol{e}^{\mathrm{T}} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})$$

- The minimal of s shows up when $\alpha_i = \boldsymbol{e}^{\mathrm{T}} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})$

Problem Reformulation

$$s(\alpha_1, \dots, \alpha_N, \boldsymbol{e}) = -\sum_{i=1}^{N} \boldsymbol{e}^{\mathrm{T}}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{e} + \sum_{i=1}^{N} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}\|^2$$

$$= -\boldsymbol{e}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{e} + \sum_{i=1}^{N} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}\|^2$$

where $\boldsymbol{S} = \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}}$

- Note that the scatter matrix looks very similar to the covariance matrix:

$$cov(\boldsymbol{X}) = \frac{1}{N-1}\sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}} = (N-1)\boldsymbol{S}$$

Solving the Problem with Lagrangian

- Minimizing $s(\alpha_1, \ldots, \alpha_N, \boldsymbol{e})$ is equivalent to maximizing $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{e}$ subject to the constraint $||\boldsymbol{e}||^2 - 1 = 0$
- Using Lagrange multiplier to solve the problem:

Let $\lambda \in \Re$

$$L = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{e} - \lambda(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} - 1)$$

- Differentiate L against $\boldsymbol{e}$:

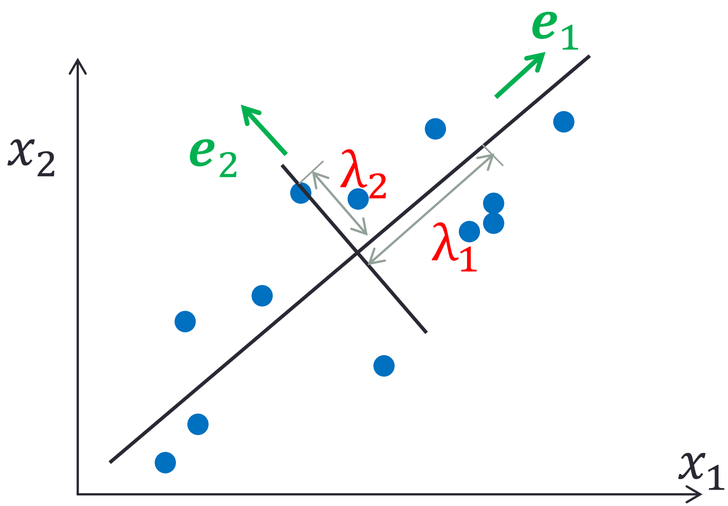
$$\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{S}\boldsymbol{e} - 2\lambda\boldsymbol{e} = 0$$

$$\boldsymbol{S}\boldsymbol{e} - \lambda\boldsymbol{e} = 0$$

- $\boldsymbol{e}$ is the eigenvector of $\boldsymbol{S}$ corresponding to the <u>largest eigenvalue</u>

Algebraic Interpretation

- PC space is a rotated orthogonal coordinate system
- The origin is the mean of the data points
- Eigenvectors show the direction of axes
- Eigenvalues are positive and show the significance of the corresponding axis

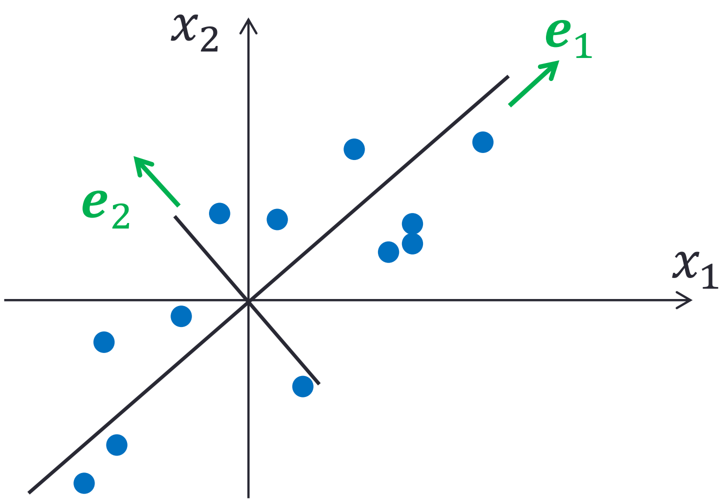


What Are Principal Components?

- A principal component (PC) $\boldsymbol{z}_i$ are linear transformation of the original variables, i.e.,

$$\boldsymbol{z}_i = (z_{i1}, z_{i2}) = (\boldsymbol{e}_1{}^{\mathrm{T}}\boldsymbol{x}_i, \boldsymbol{e}_2{}^{\mathrm{T}}\boldsymbol{x}_i)$$

where the coefficients w_{ij} are computed as projection of the principal component $\boldsymbol{z}_i$on to the basis vectors $\boldsymbol{x}_j$



$$\boldsymbol{z}_i = \begin{bmatrix} - \boldsymbol{e}_1 - \\ \vdots \\ - \boldsymbol{e}_2 - \end{bmatrix} \boldsymbol{x}_i$$

Example

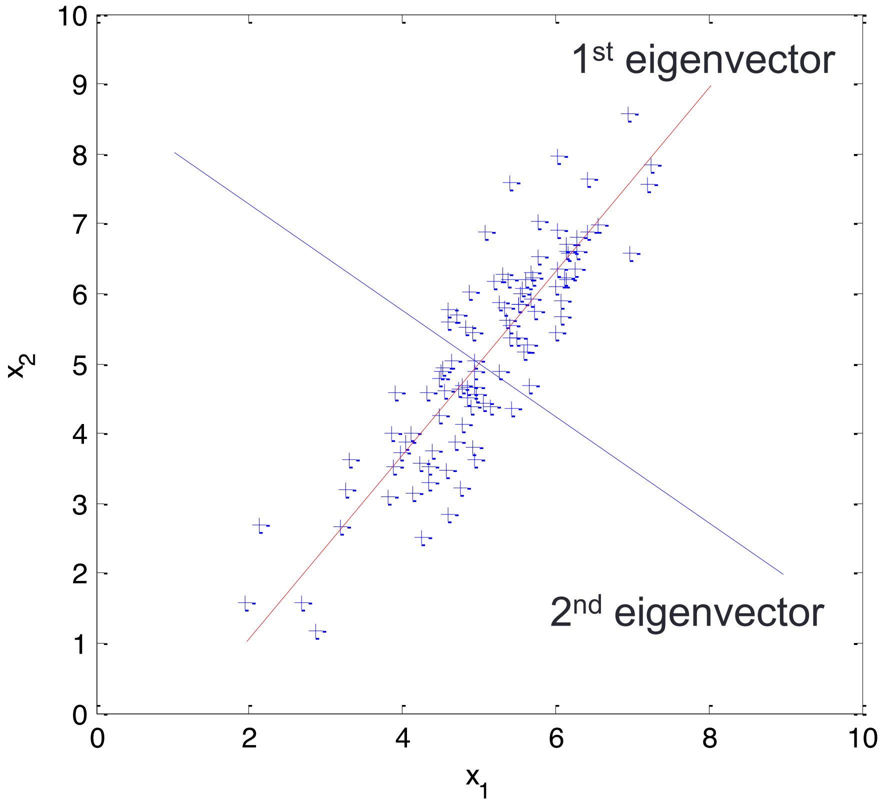
```
% generate data
Data = mvnrnd([5, 5],[1 1.2; 1.2 2], 100);
figure(1); plot(Data(:,1), Data(:,2), '+');
xlim([0 10]); ylim([0 10]);
xlabel('x_1'); ylabel('x_2');

% center the data
for i = 1:size(Data,1)
    Data(i, :) = Data(i, :) - mean(Data);
end

DataCov = cov(Data); %covariance matrix
[PC, variances, explained] = pcacov(DataCov); %eigen

% plot principal components
set( gcf, 'Color', [1, 1, 1]);
hold on; plot(PC(1,1)*[-5 5]+5, PC(2,1)*[-5 5]+5, '-r');
```

Example Figure



Representing A Dataset with Multiple Lines

- Project the data set on a d-dimensional plane of the form:

$$\boldsymbol{x}_i = \overline{\boldsymbol{x}} + \alpha_{i1}\boldsymbol{e}_1 + \cdots + \alpha_{id}\boldsymbol{e}_d$$

  where $\Re \ni d \ll M$

- Cost function to be minimized:

$$s(\alpha_1, \dots, \alpha_N, \boldsymbol{e}_1, \dots, \boldsymbol{e}_d) \equiv \sum_{i=1}^{N} \left\| \overline{\boldsymbol{x}} + \sum_{j=1}^{\boldsymbol{d}} \alpha_{ij}\boldsymbol{e}_j - \boldsymbol{x}_i \right\|^2 ,$$

  where the vectors $\boldsymbol{e}_1, \dots, \boldsymbol{e}_d$ are d eigenvectors corresponding to d largest eigenvalues of the scatter matrix

PCA through Singular Value Decomposition

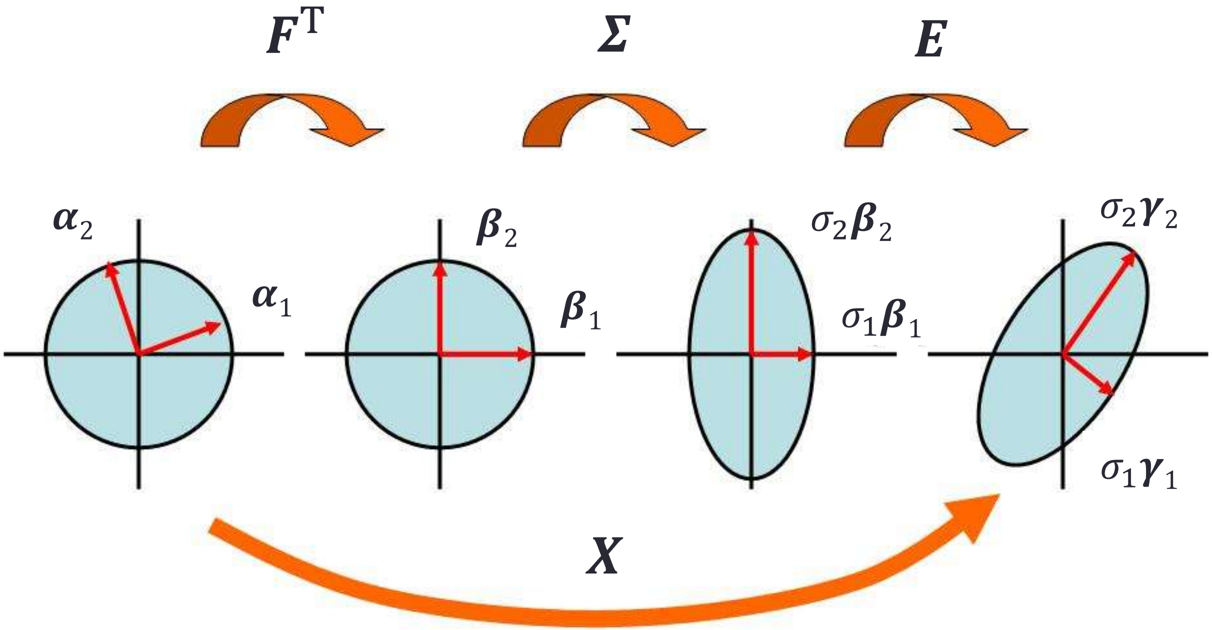
- In practice PCA is conducted via singular value decomposition (SVD)
- SVD: it is always possible to decompose any matrix $\boldsymbol{X}$ into

$$\boldsymbol{X} = \boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}$$

$$\underset{M \times N}{\begin{bmatrix} \\ \boldsymbol{X} \\ \\ \end{bmatrix}} = \underset{M \times M}{\begin{bmatrix} | & & | \\ \boldsymbol{e}_1 & \cdots & \boldsymbol{e}_M \\ | & & | \end{bmatrix}} \underset{M \times N}{\begin{bmatrix} \sigma_1 & & \varnothing \\ & \ddots & \\ \varnothing & & \sigma_M \\ \hline \varnothing & & \varnothing \end{bmatrix}} \underset{N \times N}{\begin{bmatrix} — & \boldsymbol{f}_1 & — \\ — & \boldsymbol{f}_N & — \end{bmatrix}}$$

where $\boldsymbol{E}$ and $\boldsymbol{F}$ are orthogonal matrices, i.e., $\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E} = \boldsymbol{I}$ and $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} = \boldsymbol{I}$

Geometric Explanation of SVD



PCA through SVD

- The SVD of $\boldsymbol{X}$: $\boldsymbol{X} = \boldsymbol{E\Sigma F}^{\mathrm{T}}$
- Suppose $\boldsymbol{X}$ is centered, the covariance matrix $cov(\boldsymbol{X})$ can be written as:

$$cov(\boldsymbol{X}) = \boldsymbol{XX}^{\mathrm{T}} = \left(\boldsymbol{E\Sigma F}^{\mathrm{T}}\right)\left(\boldsymbol{F\Sigma E}^{\mathrm{T}}\right) = \boldsymbol{E\Sigma}^2\boldsymbol{E}^{\mathrm{T}}$$

- Note that $cov(\boldsymbol{X})$ is a real symmetric matrix, which can be factored into

$$cov(\boldsymbol{X}) = \boldsymbol{Q\Lambda Q}^{\mathrm{T}}$$

with orthonormal eigenvectors in $\boldsymbol{Q}$ and eigenvalues in $\boldsymbol{\Lambda}$

- This gives that $\boldsymbol{E\Sigma}^2\boldsymbol{E}^{\mathrm{T}} = cov(\boldsymbol{X}) = \boldsymbol{Q\Lambda Q}^{\mathrm{T}}$

PCA through SVD (Cont'd)

- The eigenvectors of $cov(\boldsymbol{X})$ are the columns of $\boldsymbol{E}$
- The eigenvalues of $cov(\boldsymbol{X})$ are the diagonal elements of $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^2$

- The PC matrix $\boldsymbol{Z}$ is defined as

$$\boldsymbol{Z} = \boldsymbol{E}^{\mathrm{T}}\boldsymbol{X} = \begin{bmatrix} -\boldsymbol{e}_1- \\ \vdots \\ -\boldsymbol{e}_M- \end{bmatrix} \begin{bmatrix} & \boldsymbol{X} & \end{bmatrix}$$

Correlation between Principal Components

- The PCs are pairwise uncorrelated
- Recall that the PC matrix is $\boldsymbol{Z} = \boldsymbol{E}^{\mathrm{T}}\boldsymbol{X}$
- Vectors in the PC matrix $\boldsymbol{Z}$ is orthogonal, i.e., covariance matrix of $\boldsymbol{Z}$ is diagonal
- Proof:

$$cov(\boldsymbol{Z}) = \frac{1}{N-1}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} = \frac{1}{N-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{E}$$

$$= \frac{1}{N-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{\Sigma}^2\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E} = \frac{1}{N-1}\boldsymbol{\Sigma}^2$$

Sorting Variance of PCs

- The PCs have a variance equal to their corresponding eigenvalue

$$cov(\boldsymbol{z}_i) = \lambda_i = \sigma_i^2, \text{where } i = 1 \dots M$$

- Small $\lambda_i \Leftrightarrow$ small variance $\Leftrightarrow$

  data change little along the eigenvector $\boldsymbol{e}_i$

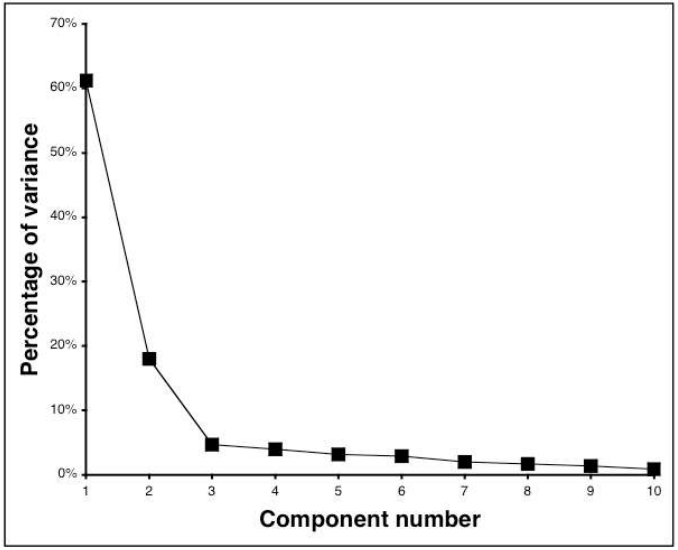
- The eigenvalues are usually sorted in an ascending order, i.e., $\lambda_1 \geq \cdots \geq \lambda_M \geq 0$
- The percentage variance explained by each PC is given by

$$\frac{\lambda_i}{\sum \lambda_i} = \frac{\sigma_i^2}{\sum \sigma_i^2}$$

PC to Be Included

- Typically the first m eigenvectors corresponding to the m largest eigenvalues are retained
- Enough PCs to have a cumulative variance explained by the PCs that is larger than 60-90%
- Scree plot:

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100.0%



Dimensionality Reduction through PCA

- Assume information is additive

$$\boldsymbol{X} = \boldsymbol{X}_S + \boldsymbol{X}_N = \boldsymbol{E}_S\boldsymbol{\Sigma}_S(\boldsymbol{F}_S)^{\mathrm{T}} + \boldsymbol{E}_N\boldsymbol{\Sigma}_N(\boldsymbol{F}_N)^{\mathrm{T}}$$

$$\begin{bmatrix} & & \\ & \boldsymbol{X} & \\ & & \end{bmatrix} = \begin{bmatrix} | & & | \\ \boldsymbol{e}_1 & \cdots & 0 \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \varnothing \\ & \ddots & \\ \varnothing & & 0 \\ \hdashline \varnothing & & \varnothing \end{bmatrix} \begin{bmatrix} — & \boldsymbol{f}_1 & — \\ — & 0 & — \end{bmatrix}$$

$$+ \begin{bmatrix} | & & | \\ 0 & \cdots & \boldsymbol{e}_M \\ | & & | \end{bmatrix} \begin{bmatrix} 0 & & \varnothing \\ & \ddots & \\ \varnothing & & \sigma_M \\ \hdashline \varnothing & & \varnothing \end{bmatrix} \begin{bmatrix} — & 0 & — \\ — & \boldsymbol{f}_N & — \end{bmatrix}$$

- Reserve PC with high variance and disregard the rest
- Common application area: compression, noise reduction

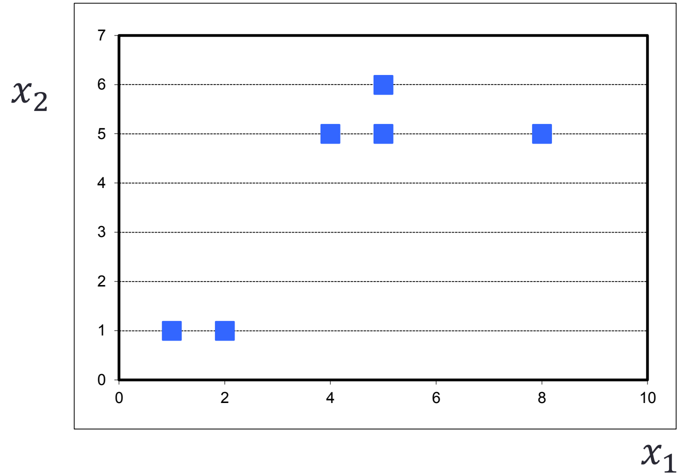
Using PCA in Divisive Clustering

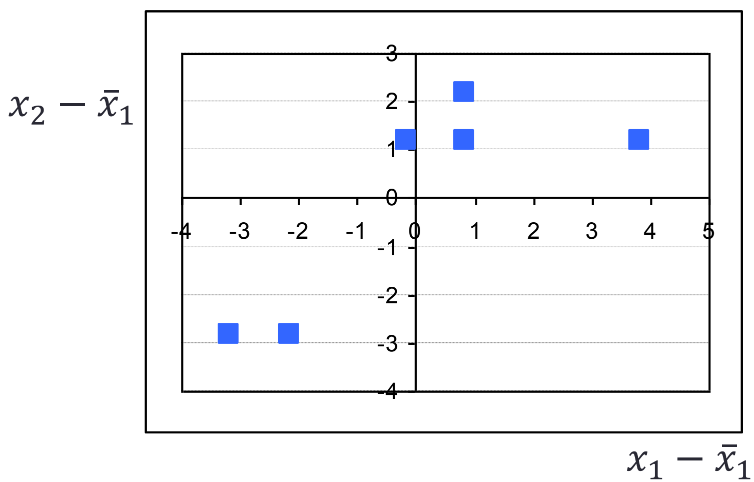
1. Calculate the eigenvalues and eigenvectors
2. Choose the eigenvector with the highest eigenvalue of the covariance matrix
3. Try each points along the principal axis as the dividing point and select the one with the largest margin
4. Determine the separating

PCA Clustering Example

- Data

Point	x_{1i}	x_{2i}	$x_{1i}-\bar{x}_1$	$x_{1i}-\bar{x}_1$
1	1	1	-3.17	-2.83
2	2	1	-2.17	-2.83
3	4	5	-0.17	1.17
4	5	5	0.83	1.17
5	5	6	0.83	2.17
6	8	5	3.83	1.17





PCA Clustering Example (Cont'd)

- Covariance matrix: $cov(\boldsymbol{X}) = \begin{pmatrix} 5.14 & 3.69 \\ 3.69 & 4.14 \end{pmatrix}$
- Eigenvalues λ: $cov(\boldsymbol{X}) - \lambda \cdot I_2 = \begin{pmatrix} 5.13 - \lambda & 3.69 \\ 3.69 & 4.13 - \lambda \end{pmatrix}$

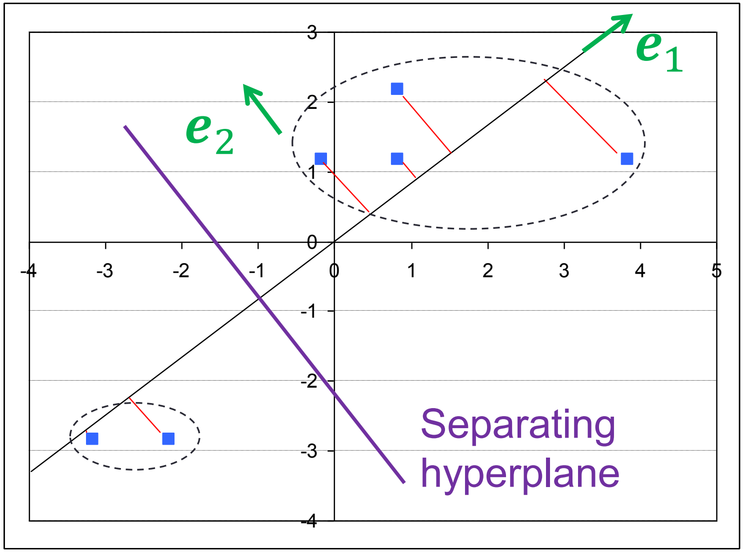
  $\lambda_1 = 8.367$ and $\lambda_2 = 0.911$
- Eigenvectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$:

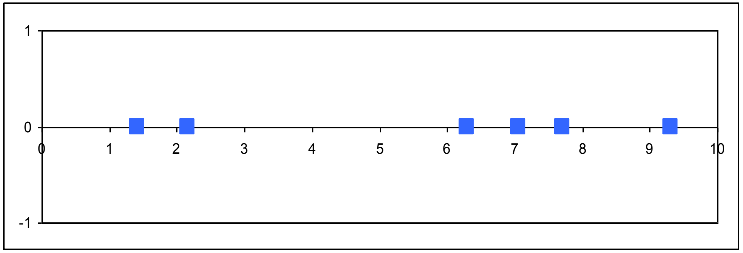
  $\begin{pmatrix} 5.14 & 3.69 \\ 3.69 & 4.14 \end{pmatrix} \boldsymbol{e}_1 = 8.36\boldsymbol{e}_1$ and $\begin{pmatrix} 5.14 & 3.69 \\ 3.69 & 4.14 \end{pmatrix} \boldsymbol{e}_2 = 0.91\boldsymbol{e}_2$

  $\boldsymbol{e}_1 = (-0.75, -0.66)^{\mathrm{T}}$ and $\boldsymbol{e}_2 = (0.66, -0.75)^{\mathrm{T}}$

PCA Clustering Example (Cont'd)

- Projections on the selected eigenvector
- Try the median of each pair of adjacent points as the dividing point
- Choose the one gives the largest margin as the separating hyperplane



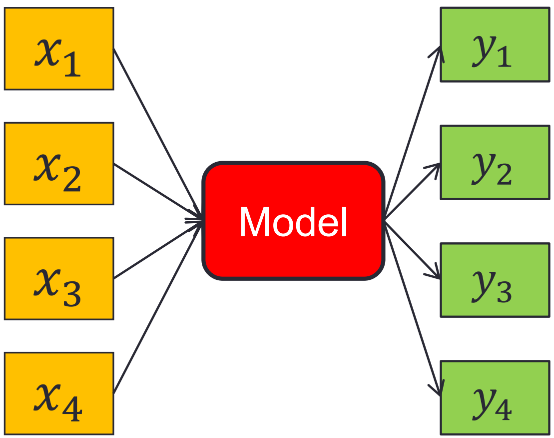


Summary of PCA

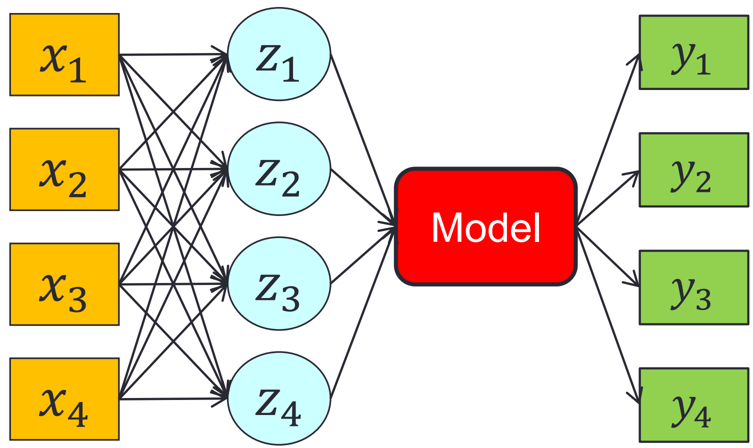
- A method for identifying the important “directions” in the data
- Maps data into a (reduced) coordinate system that is given by those “directions”
- The new variables are called PCs
- Practically implemented by using SVD
- One of the most common feature reduction techniques
- Unsupervised in a sense that it does not consider the output class/value of an instance
- What about using PCs as explanatory variables for regression?

Multiple Regression vs. Principal Component Regression

- Multiple regression



- Principal component regression (PCR)



Review of Multiple Regression Model

- Multiple regression model:

$$y = f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M = \boldsymbol{\beta x}$$

where $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_M]$ and $\boldsymbol{x} = [1, x_1, \ldots, x_M]^{\mathrm{T}}$

- Suppose N observation is made, i.e.,

$$\boldsymbol{X} = \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1M} & \cdots & x_{NM} \end{bmatrix} \qquad \boldsymbol{y} = [y_1 \quad \cdots \quad y_N]$$

- The regression model coefficients: $\boldsymbol{\beta} = \boldsymbol{y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}$

Problem with Multiple Regression – Multicollinearity

- The situation where the explanatory variables are highly inter-correlated is referred to as multicollinearity
- When the explanatory variables are highly correlated, it becomes difficult to disentangle the separate effects of each of the explanatory variables on the response variable
- In the other words, the inverse of $(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}$ may be changed dramatically once there is noise in the data, and hence the regression model coefficients are marginally unstable

Principal Component Regression

- Regress the response variable $\boldsymbol{y}$ using $\boldsymbol{Z}$ rather than $\boldsymbol{X}$
- Reformulate the model: $\boldsymbol{y} = \boldsymbol{\beta}'\boldsymbol{Z}$
- Because the PC matrix $\boldsymbol{Z}$ is orthogonal, $\boldsymbol{\beta}'$ can be directly calculated following the least-squares, i.e.,

$$\boldsymbol{\beta}' = \boldsymbol{y}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}$$

- Matching the models

$$\boldsymbol{y} = \boldsymbol{\beta}'\boldsymbol{Z} = \boldsymbol{\beta}'\boldsymbol{E}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{\beta}\boldsymbol{X} = \boldsymbol{y},$$

the regression model coefficients are $\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{E}^{\mathrm{T}}$

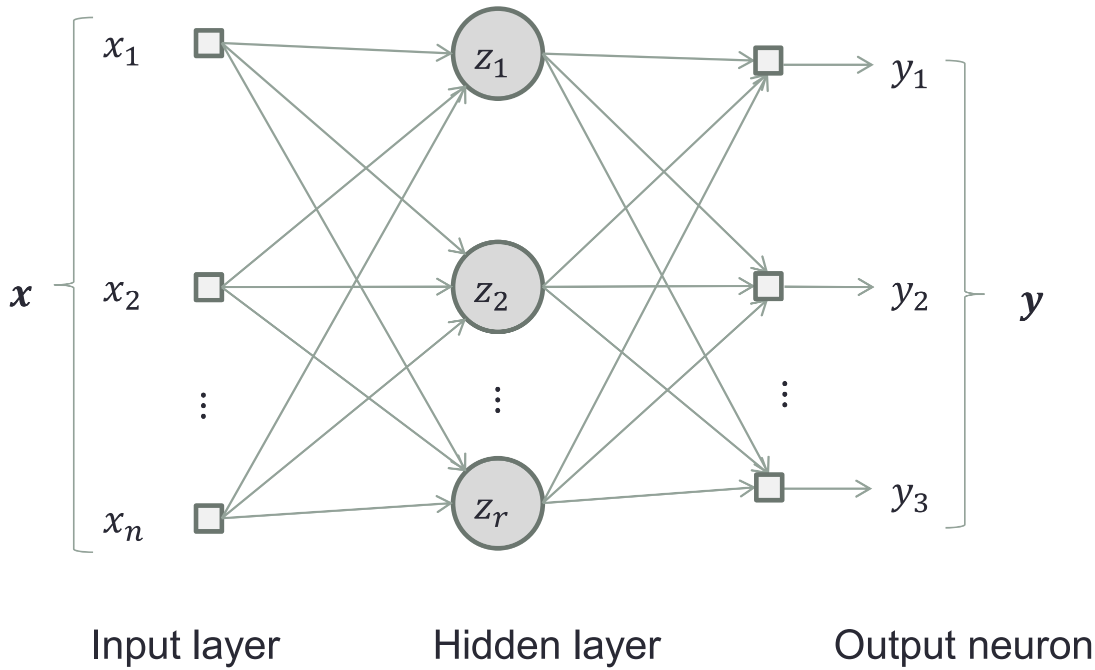
Biased Principal Component Regression

- Assume the signal and the noise are additive
- Decompose the explanatory variable matrix into a signal matrix and a noise matrix:

$$\boldsymbol{X} = \boldsymbol{X}_S + \boldsymbol{X}_N = \boldsymbol{E}_S \boldsymbol{\Sigma}_S (\boldsymbol{F}_S)^{\mathrm{T}} + \boldsymbol{E}_N \boldsymbol{\Sigma}_N (\boldsymbol{F}_N)^{\mathrm{T}}$$

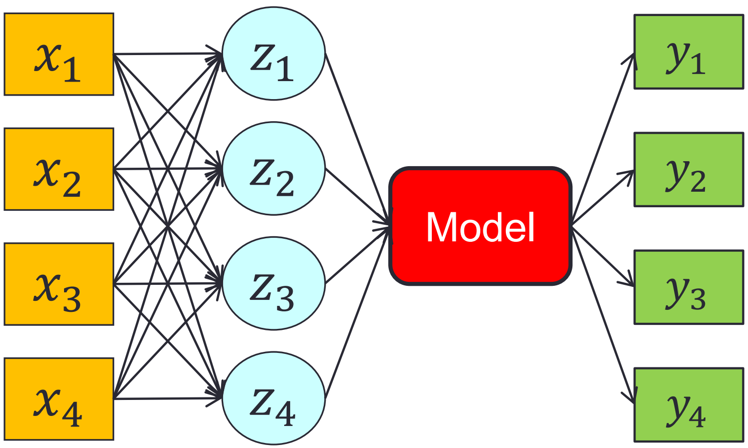
- Regress the response variable $\boldsymbol{y}$ using the signal matrix $\boldsymbol{X}_S$ only
- How to select PCs to be retained in the signal matrix $\boldsymbol{X}_S$?

Architecture of Principal Component Regression



Problem of PCR

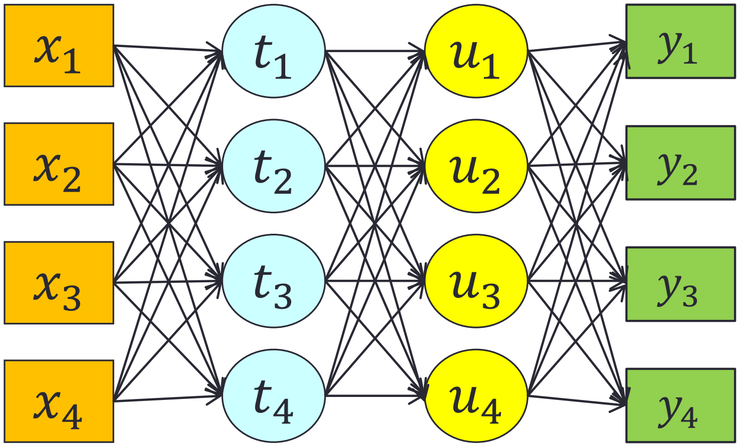
- PCs are created from the covariance matrix of the explanatory variables (hence unsupervised)



- What if we consider the relationship between x_i and y_i while creating the new latent variables?

Partial Least Squares Regression (PLSR)

- Partial least squares regression (PLSR) is similar to PCR except in how the component scores are computed
- In PLSR, the weights reflect the covariance structure between the explanatory and response variables
- Denote t and u as the PLSR latent variables



PLSR Formulation

- Let $\boldsymbol{T} \in \Re^{N\times H}$ denote the latent variable (also called scores) matrix of $\boldsymbol{X}$, and $\boldsymbol{U} \in \Re^{N\times H}$ denote the latent variable matrix of $\boldsymbol{Y}$
- The $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be expressed as

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} \qquad \text{with} \quad \boldsymbol{T}^{\mathrm{T}}\boldsymbol{T} = \boldsymbol{I}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{F} = \boldsymbol{T}\boldsymbol{B}\boldsymbol{C}^{T} + \boldsymbol{F}$$

where $\boldsymbol{P} \in \Re^{H\times M}$ and $\boldsymbol{C} \in \Re^{L\times H}$ are the loading matrices, and $\boldsymbol{E} \in \Re^{N\times M}$ and $\boldsymbol{F} \in \Re^{N\times L}$ are the random errors of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively

PLSR Formulation (Cont'd)

- The matrices $\boldsymbol{T}$ and $\boldsymbol{U}$ are identified column by column
- Let vector $\boldsymbol{t}$ be a column in $\boldsymbol{T}$, and vector $\boldsymbol{u}$ be the corresponding column in $\boldsymbol{U}$
- Objective: find a pair of $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$ and $\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{c}$ such that the covariance between $\boldsymbol{t}$ and $\boldsymbol{u}$ is maximized, i.e.,

$$\max \boldsymbol{t}^{\mathrm{T}}\boldsymbol{u}$$

  with the constraints that $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = 1$ and $\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t} = 1$
- SIMPLS is one of the most common algorithm to solve this optimization problem

SIMPLS Algorithm

- Initialize $\boldsymbol{u}$ with random values, and let $\boldsymbol{E} = \boldsymbol{X}$, $\boldsymbol{F} = \boldsymbol{Y}$
- Run these four steps until $\boldsymbol{t}$ is converged
  1. $\boldsymbol{w} = \boldsymbol{E}^{\mathrm{T}}\boldsymbol{u}/\left|\boldsymbol{E}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}\right|$ (estimate $\boldsymbol{X}$ weights)
  2. $\boldsymbol{t} = \boldsymbol{E}^{\mathrm{T}}\boldsymbol{w}/\left|\boldsymbol{E}^{\mathrm{T}}\boldsymbol{w}\right|$ (estimate $\boldsymbol{X}$ factor scores)
  3. $\boldsymbol{c} = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{t}/\left|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{t}\right|$ (estimate $\boldsymbol{Y}$ weights)
  4. $\boldsymbol{u} = \boldsymbol{F}\boldsymbol{c}$ (estimate $\boldsymbol{Y}$ scores)
- Record $\boldsymbol{w}$, $\boldsymbol{t}$, $\boldsymbol{c}$ and $\boldsymbol{u}$
- Let $\boldsymbol{p} = \boldsymbol{E}^{\mathrm{T}}\boldsymbol{t}$, $b = \boldsymbol{t}^{\mathrm{T}}\boldsymbol{u}$, $\boldsymbol{E} = \boldsymbol{E} - \boldsymbol{t}\boldsymbol{p}^{\mathrm{T}}$, and $\boldsymbol{F} = \boldsymbol{F} - b(\boldsymbol{t}^{\mathrm{T}}\boldsymbol{c})$
- The $\boldsymbol{t}$, $\boldsymbol{u}$, $\boldsymbol{w}$, $\boldsymbol{c}$, and $\boldsymbol{p}$ are stored in the corresponding matrices, and the b is stored as a diagonal element of $\boldsymbol{B}$
- Calculate the next pair of $\boldsymbol{t}$ and $\boldsymbol{u}$

PLSR Model

- The response variables are predicted using the formula

$$\widehat{\boldsymbol{Y}} = \boldsymbol{TBC}^{\mathrm{T}} = \boldsymbol{X}\left(\boldsymbol{P}^{\mathrm{T}+}\right)\boldsymbol{BC}^{\mathrm{T}}$$

  where $\boldsymbol{P}^{\mathrm{T}+}$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{P}^{\mathrm{T}}$

- If all the latent variables of $\boldsymbol{X}$ are used, this regression is equivalent to PCR

Why PLS?

- Similar to PCR, PLSR can handle strong collinear data and under-determined problems
- The regression is performed based on the relationship between the explanatory and response variables (hence supervised)

Partial Least Squares Discriminant Analysis

- In partial least squares discriminant analysis (PLSDA), the entries of the response variable matrix $\boldsymbol{Y}$ are binary
- Comparison to PCA
  - PLSDA is a supervised classification technique while PCA is an unsupervised clustering technique
  - PLSDA enhances the separation between groups of observations by rotating newly defined components (e.g., scores) such that a maximum separation among classes is obtained

Reference

- T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*
- Hervé Abdi, “Partial least square regression”