PS5841

Data Science in Finance & Insurance

PCA

Yubo Wang

Autumn 2021

COLUMBIA UNIVERSITY

Unsupervised Learning

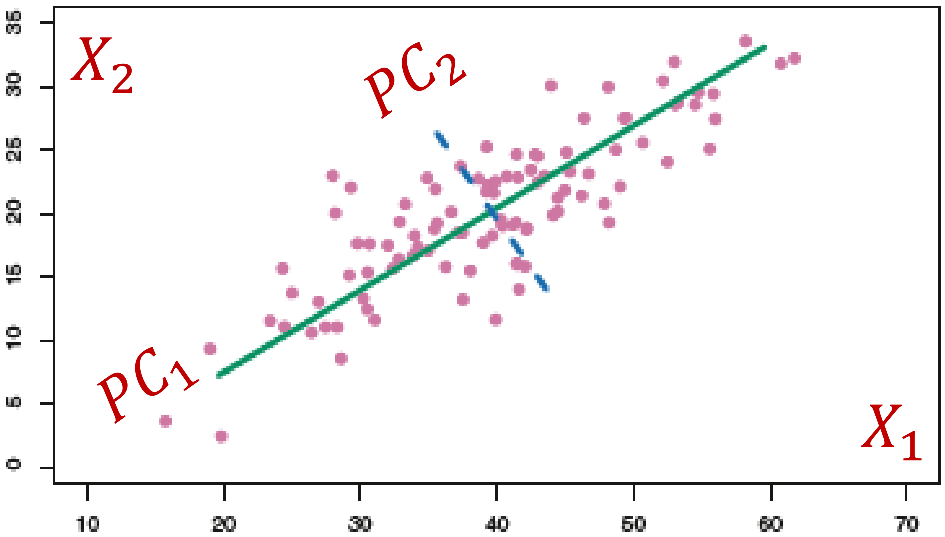
- No responses to supervise learning
  - Difficult to access the quality of unsupervised learning
- Feature visualization
- Data pre-processing
- Discover unknown subgroups in data

PCA

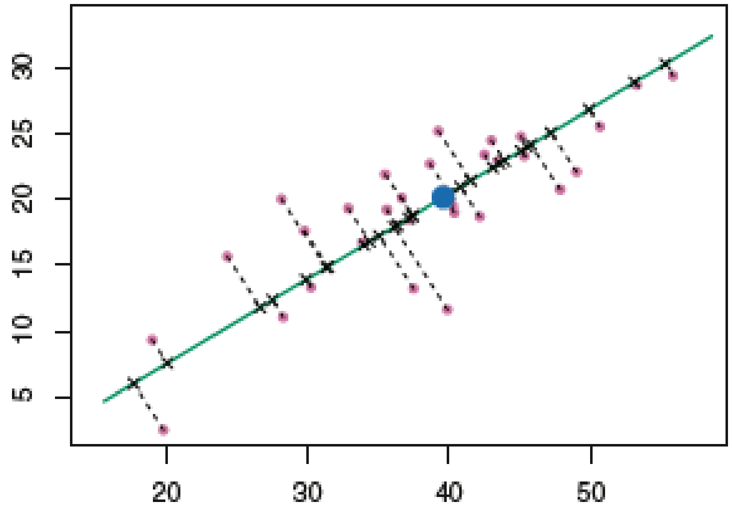
- Low(er)-dimensional representation of feature space capturing variation as much as possible
  - Loading vectors: orthogonal unit vectors in feature space with the most variations
  - Score vectors: projections along loading vectors
- Q-dimensional hyperplane that is closest (in terms of Euclidean distance) to the observations.
- At most $\min(n-1, p)$ principal components.

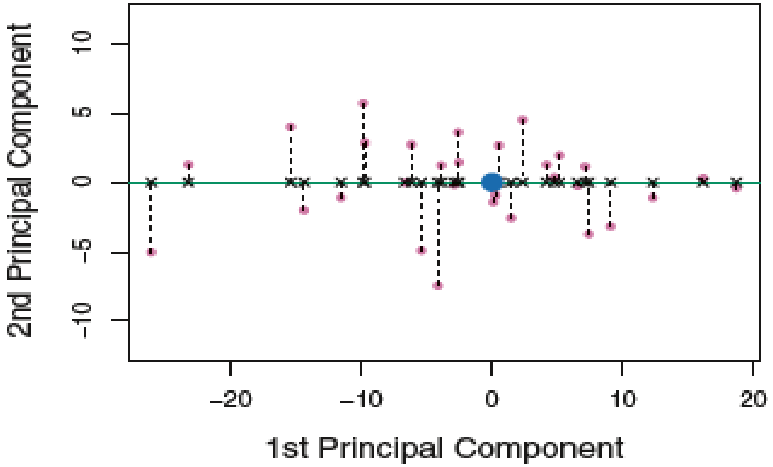
COLUMBIA UNIVERSITY

Example: PC in 2D Data

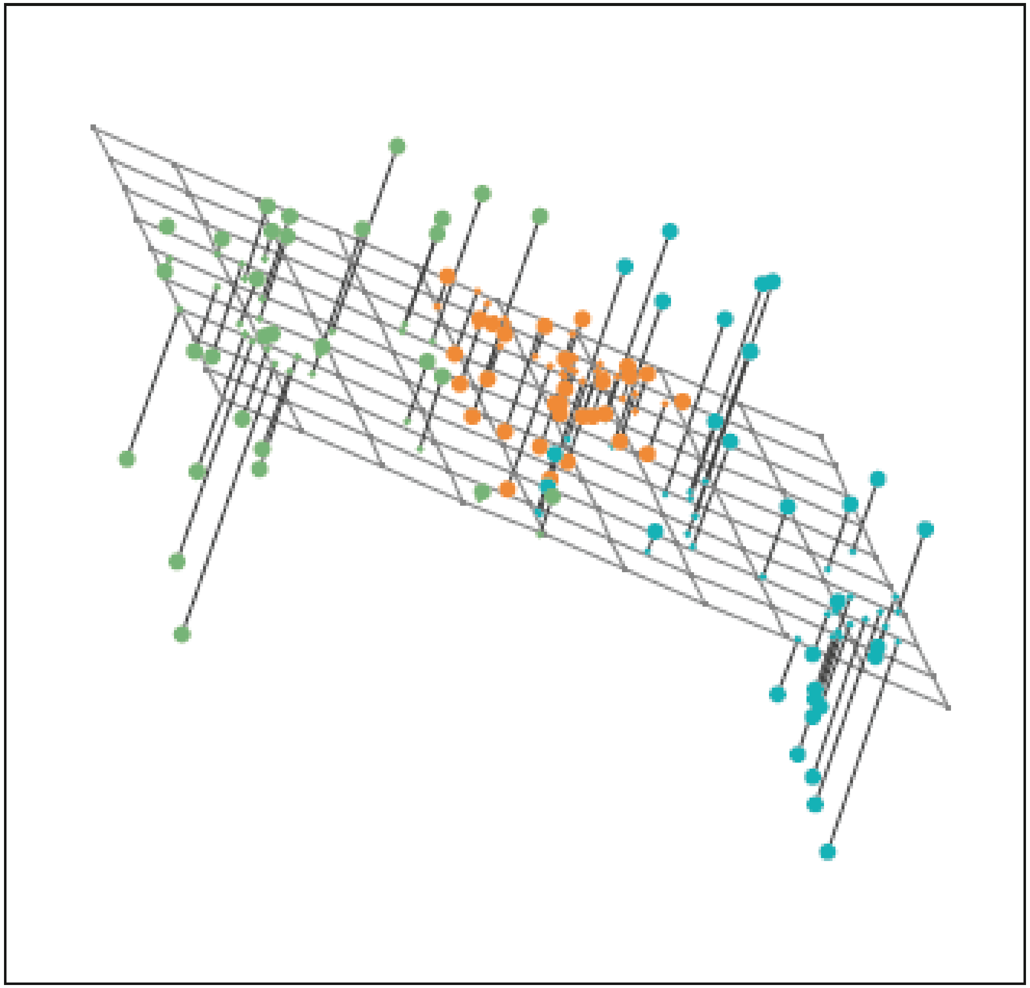


- Centered "rotation"
- PC_1 minimizes SS distance from data to projection onto PC_1 $\Leftrightarrow$
- PC_1 maximizes SS distance from projection onto PC_1 to the center





Example: PC in 3D Data



Standardizing Data

- De-mean (zero mean) to make variance calculation more tractable
- Unit variance unless measured in the same units

COLUMBIA UNIVERSITY

PC notation

$$\underset{n\times q}{\boldsymbol{Z}} = \underset{n\times p}{\boldsymbol{X}}\underset{p\times q}{\boldsymbol{\Phi}}$$

$$\left(\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_q\right)^T = \boldsymbol{X}\left(\boldsymbol{\Phi}_1,\dots,\boldsymbol{\Phi}_q\right)^T$$

$$\begin{bmatrix} z_{11} & \cdots & z_{1q} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nq} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \phi_{11} & \cdots & \phi_{1q} \\ \vdots & \ddots & \vdots \\ \phi_{p1} & \cdots & \phi_{pq} \end{bmatrix}$$

$$\boldsymbol{Z}_k = \sum_{j=1}^{p} \phi_{jk}\boldsymbol{X}_j = \phi_{1k}\boldsymbol{X}_1 + \cdots + \phi_{pk}\boldsymbol{X}_p$$

$$z_{ik} = \sum_{j=1}^{p} \phi_{jk}x_{ij} = \phi_{1k}x_{i1} + \cdots + \phi_{pk}x_{ip}$$

k-th score vector $\boldsymbol{Z}_k$, loading vector $\boldsymbol{\Phi}_k$ for PC_1

Columbia University

PC_1

- Perform PCA on standardized data (zero mean, plus unit variance unless measured in the same units)
- $\boldsymbol{\Phi}_1$ maximizes

$$\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2$$

subject to $\sum_{j=1}^{p}\phi_{j1}^2 = 1$

where

$$z_{ik} = \sum_{j=1}^{p}\phi_{jk}x_{ij} = \phi_{1k}x_{i1} + \cdots + \phi_{pk}x_{ip}$$

COLUMBIA UNIVERSITY

PC_k

- Perform PCA on standardized data (zero mean, plus unit variance unless measured in the same units)
- $\boldsymbol{\Phi}_k$ maximizes

$$\frac{1}{n}\sum_{i=1}^{n} z_{ik}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jk}x_{ij}\right)^2$$

subject to

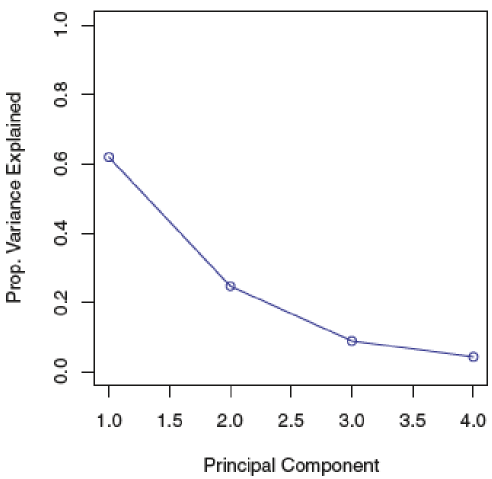
$$\sum_{j=1}^{p}\phi_{jk}^2 = 1$$

$$\boldsymbol{\Phi}_2\ orthogonal\ to\ \boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_{k-1}$$

COLUMBIA UNIVERSITY

PVE

- Proportion of variance (total variation) explained by PC_k



$$\frac{\frac{1}{n}\sum_{i=1}^{n} z_{ik}^2}{\sum_{j=1}^{p} Var(\boldsymbol{X}_j)}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} \phi_{jk} x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2 s}$$

Columbia University

Representing Data

- Standardized data (zero mean, plus unit variance unless measured in the same units) can be represented via score and loading vectors

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}\,, M < q \leq p$$

- When $\mathrm{M} = \min(n-1, p)$

$$x_{ij} = \sum_{m=1}^{M} z_{im}\phi_{jm}$$

Columbia University

Uniqueness

- Loading vectors are unique up to a sign flip
  - sign flip does not alter the coordinate system
- Score vectors are unique up to a sign flip
  - Variance of Z and –Z are the same
- Flipping signs on loading and score vectors simultaneously has no effect on

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}, M < q \leq p$$

- But the right sign may improve interpretability

COLUMBIA UNIVERSITY

Principal Components Regression

- Dimension Reduction
  - A small number of PCs may be able to explain most of the variability in data, as well as the relationship with the response (no guarantee)
  - Assumption: the directions in which the predictors show the most variation are the direction that are associated with the response
  - Not a variable selection method
    - Each PC is a linear combination of all original predictors

Columbia University

PCR Recipe

- Standardize all predictors
- Construct the first M principal components
  - Can determine M by cross validation
- Apply LS fit using M principal components
- Recover β_j to make predictions

Columbia University

Dimension Reduction

- From p predictors to $M < p$ predictors
  - Standardizing the predictors necessary to have predictors on the same scale

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i$$

$$= \theta_0 + \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} x_{ij} \phi_{jm} + \varepsilon_i$$

$$= \theta_0 + \sum_{m=1}^{M} \sum_{j=1}^{p} x_{ij} \phi_{jm} \theta_m + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j + \varepsilon_i$$

$$\beta_j = \sum_{M=1}^{M} \phi_{jm} \theta_m$$

$$\beta_0 = \theta_0$$

Columbia University

That was

Columbia University