Applied Statistical Multivariate Analysis

Principal Component Analysis

Instructor: Yaser Zerehsaz

موسسه آموزش عالی آزاد توسعه

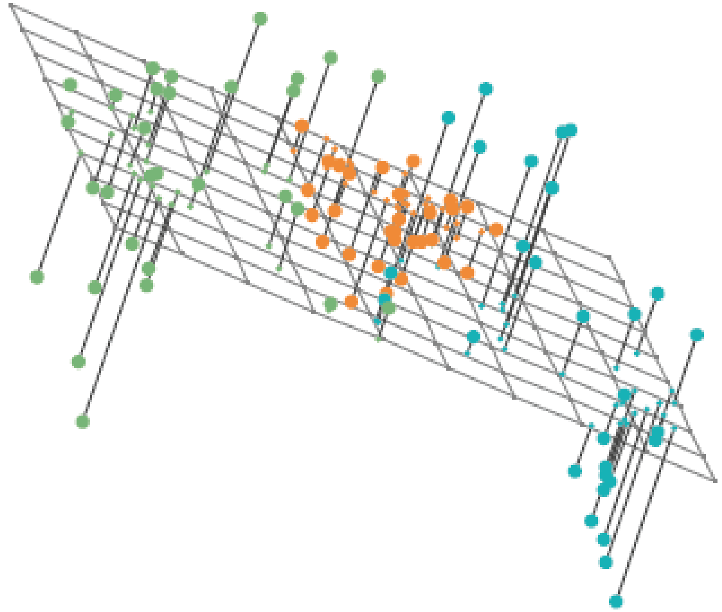
برگزار کننده دوره‌های تخصصی علم داده

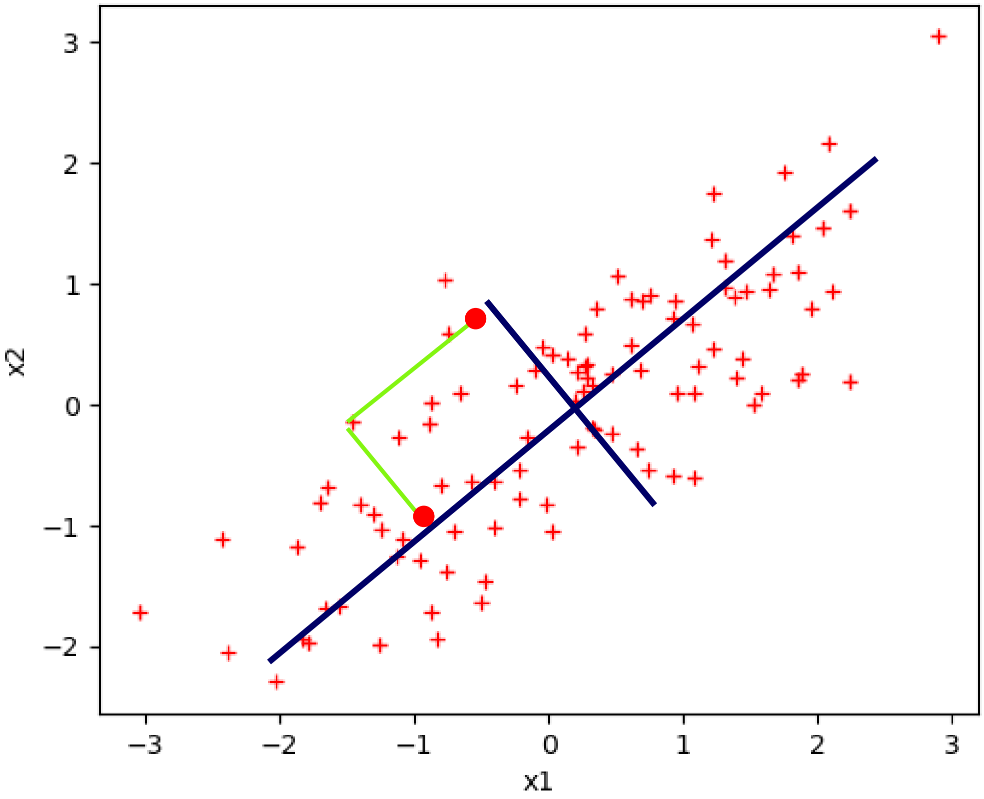
Tihe.ac.ir

Motivation

- Dimension reduction of a dataset
- Extract the relevant features for monitoring, fault diagnosis and prediction
- PCA is a linear dimension reduction method, it constructs a linear combination of the original variables. The new space will be q-dimensional where $q < p$.
- It works based on a projection method
- It can be used for visualization of data
- It can be employed for noise reduction

Geometric representation





What is an appropriate one-dimensional projection?

Mathematical formulation

Suppose that the columns of matrix $\mathbf{X}$ are centered i.e. $E(\mathbf{X}) = \mathbf{0}$ or the mean of each column is zero, and let's denote each column of $\mathbf{X}$ by $\mathbf{x}_i \in \mathbb{R}^n; i = 1,2,\dots,p$. Then:

PCA problem is to find $q \le p$ new variables $\mathbf{y}_j \in \mathbb{R}^n; j = 1,2,\dots,q$, such that:

$$\mathbf{y}_j = \sum_{i=1}^{p} w_i \mathbf{x}_i \text{ maximizes } \operatorname{var}\left(\mathbf{y}_j\right) = \mathbf{w}_j{}^T \mathbf{\Sigma} \mathbf{w}_j$$
$$\text{with } \mathbf{w}_j^T \mathbf{w}_j = 1,\ \mathbf{w}_j^T \mathbf{w}_k = 0 \text{ for } k \ne j = 1,2,\dots,q$$

In matrix form:

$\mathbf{Y} = \mathbf{X}\mathbf{W}$ where $\mathbf{W}$ solves

$$\text{Maximize } \mathbf{W}^T \mathbf{\Sigma} \mathbf{W}$$
$$\text{Subject to: } \mathbf{W}^T \mathbf{W} = \mathbf{I}$$

$\mathbf{y}_j{}'s$ are called PC scores or features

Solution to the PCA problem

Derive the first PC:

$$\text{Maximize } f(\mathbf{w}_1) = \mathbf{w}_1{}^T \boldsymbol{\Sigma} \mathbf{w}_1$$

$$\text{S.t. } \mathbf{w}_1{}^T \mathbf{w}_1 = 1$$

Incorporate the constraint in the objective function using Lagrange multipliers and take the derivative with respect to $\mathbf{w}_1$:

$$\boldsymbol{\Sigma} \mathbf{w}_1 = \lambda_1 \mathbf{w}_1$$

λ_1 must be the eigenvalue of $\boldsymbol{\Sigma}$ and $\mathbf{w}_1$ is the corresponding eigenvector.

Solution to the PCA problem

- Recall the eigendecomposition problem

$$\boldsymbol{\Sigma} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^T = \sum_{i=1}^{p} \lambda_i \mathbf{w}_i {\mathbf{w}_i}^T$$

$\mathbf{W}$ is the matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues.

- To obtain $\mathbf{w}_1$, we should choose one eigenvalue and its associated eigenvector, but which one?
- Choose λ_i maximizing $\sum_{i=1}^{p} \lambda_i \mathbf{w}_i {\mathbf{w}_i}^T$
- How can we get the second direction?
- They should be perpendicular to the first direction $\mathbf{w}_1$ i.e. $\mathbf{w}_1^T \mathbf{w}_2 = 0$
- Recall that in eigendecomposition, different eigenvectors are orthogonal. So, the solution to PCA problem is given by the eigendecomposition of $\boldsymbol{\Sigma}$ and $\mathbf{Y} = \mathbf{X}\mathbf{W}$

Properties of PCs

- $E(\mathbf{y}_i) = 0$
- $Cov(\mathbf{y}) = \Lambda$; as a result, PCs are uncorrelated
- $Var(\mathbf{y}_i) = \lambda_i$
- if $\mathbf{X}$ is a multivariate normal, $\mathbf{Y}$ will be multivariate normal and PCs are independent
- $Cov(\mathbf{X}, \mathbf{Y}) = \mathbf{W\Lambda}$.

Example- Athletic performance data

- Records for 55 countries in the following men's track events: 100, 200, 400, 800, 1500, 5000, 10000 meters and marathon
- The data are in seconds for the first three variables and in minutes for the rest

Example- Components

	PC1	PC2	PC3
100	0.019865	0.210690	0.029042
200	0.041554	0.358926	0.018390
400	0.110632	0.827863	0.377669
800	0.005488	0.023175	-0.005342
1500	0.014387	0.044653	-0.050004
5000	0.079308	0.129961	-0.336449
10,000	0.181099	0.298854	-0.848723
Marathon	0.972787	-0.180807	0.141872

First component:

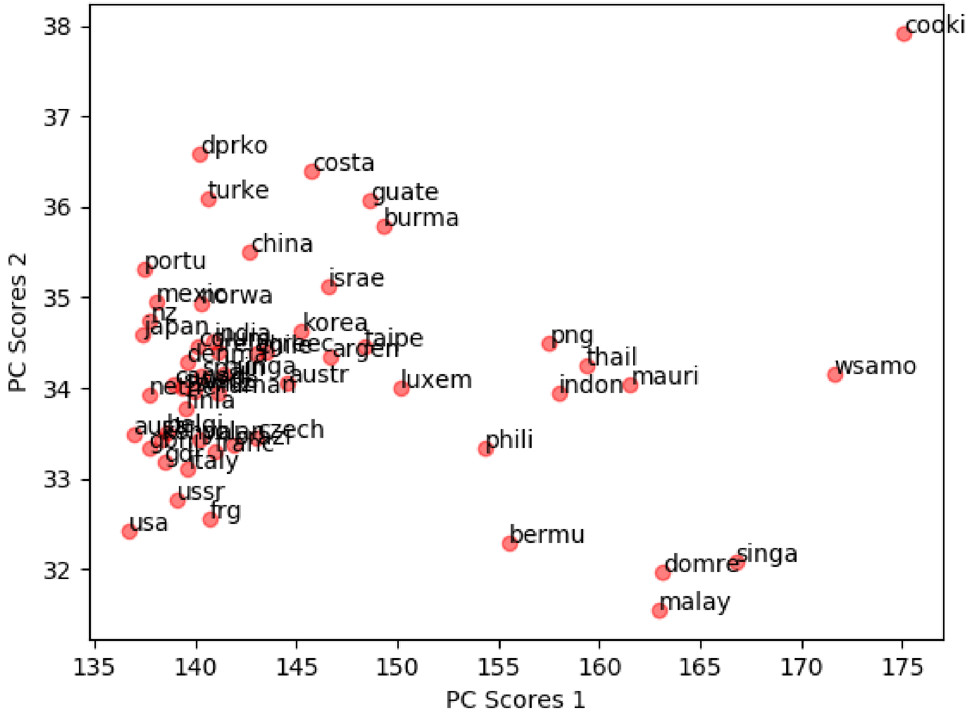
- Overall mean with much more emphasis on Marathon
- Higher variability

Second component:

- A contrast between Marathon and the rest of variables

Example- First two PCs

	PC1	PC2
100	0.019865	0.210690
200	0.041554	0.358926
400	0.110632	0.827863
800	0.005488	0.023175
1500	0.014387	0.044653
5000	0.079308	0.129961
10,000	0.181099	0.298854
Marathon	0.972787	-0.180807



Number of PCs to keep

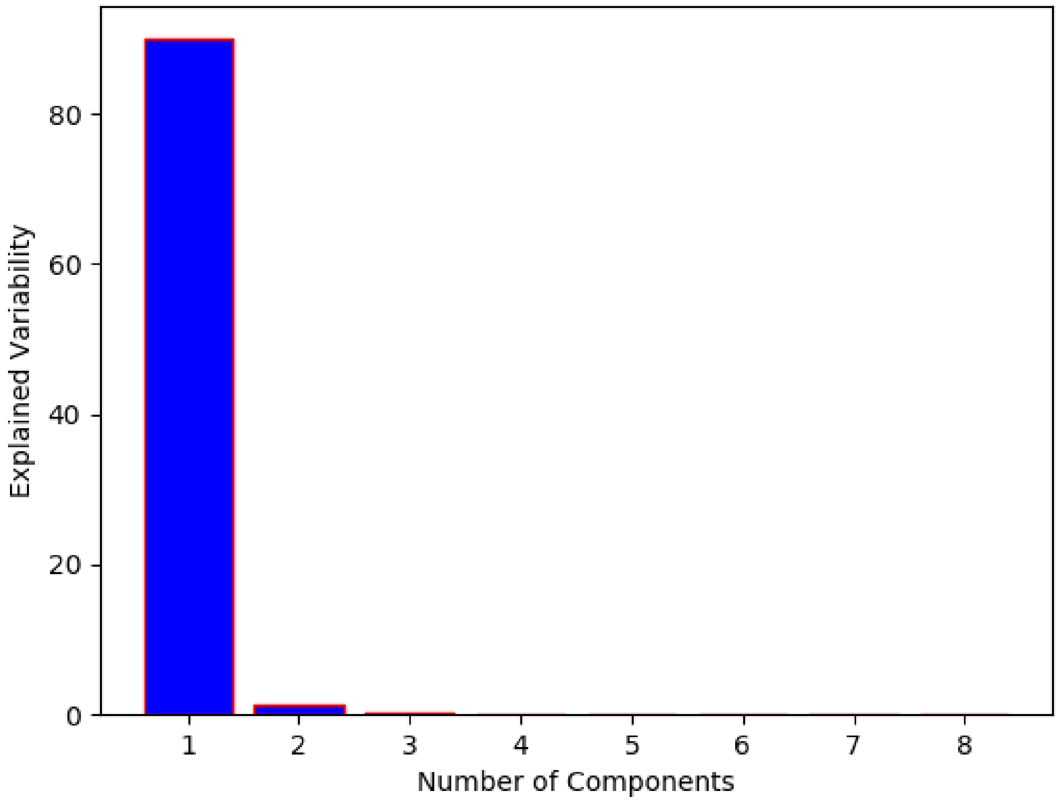
- The criterion is to keep enough PCs to appropriately represent the data
- Scree plot: plot λ_i against i and for an elbow
- Explained variability of each component i is $\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$
- Percentage of explained variability: pick the first q components such that

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geq 1 - \alpha$$

where α is a pre-specified small value.

- There is no universal rule

Scree plot



Explained Variability
0
20
40
60
80
1
2
3
4
5
6
7
8
Number of Components

Using SVD in PCA

- If the original matrix $\mathbf{X}$ is centered, the covariance matrix $\mathbf{\Sigma}$ can be estimated by $\widehat{\mathbf{\Sigma}} = \frac{\mathbf{X}^T\mathbf{X}}{n-1}$
- When p is large, computing $\mathbf{X}^T\mathbf{X}$ is cumbersome
- Based on SVD, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Then we can write

$$(n-1)\widehat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$$

- $\mathbf{V}$ is the matrix of eigenvectors of $\widehat{\mathbf{\Sigma}}$ and $\mathbf{D}$ contains the square roots of the eigenvalues of $(n-1)\widehat{\mathbf{\Sigma}}$
- Hence, we can compute the PC scores as $\mathbf{Y} = \mathbf{X}\mathbf{V}$

Using SVD in PCA- Power Method

- The solution for SVD goes through eigendecomposition
- It can be showed that for a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ the eigenvalues can be obtained by solving

$$c_n \lambda^n + c_{n-1} \lambda^n + c_0 = 0$$

- For large values of n, polynomial equations like this one are difficult and time-consuming to solve

Solution

Use the power method algorithm based upon the following theorem

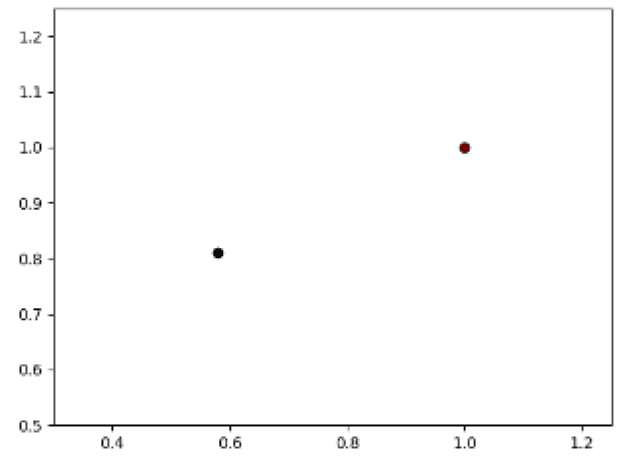
Theorem: if $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a diagonalizable matrix with a dominant eigenvalue, then there exist a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ such that the sequence of

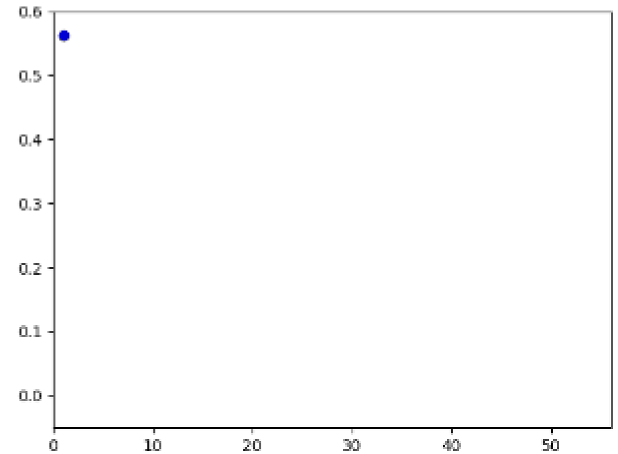
$$\mathbf{Ax}, \mathbf{A}^2\mathbf{x}, \dots, \mathbf{A}^k\mathbf{x}$$

Converges to a multiple of the dominant eigenvector of $\mathbf{A}$

Using SVD in PCA- Power method

$$\mathbf{A} = \begin{pmatrix} -4 & 10 \\ 7 & 5 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$





Problems with PCA

- Since we do not know $\mathbf{\Sigma}$, we replace it with $\widehat{\mathbf{\Sigma}}$
- When $p \gg n$, the estimated eigenvalues and eigenvectors might not be very accurate
- Also, in some cases it is not easy to choose the appropriate number of PCs

Bootstrap as a possible solution

Use bootstrap method (Efron 1979):

- To make an inference regarding some of the estimators, like the explained variance

Fundaments:

- Apply bootstrap when the there is no theory to compute standard errors, confidence intervals, etc.
- It may not always work, but it works generally
- Basic idea is to pretend that the observed sample is the population
- For this reason, we can resample several samples
- Compute the estimators and evaluate the variability

Bootstrap- Algorithm

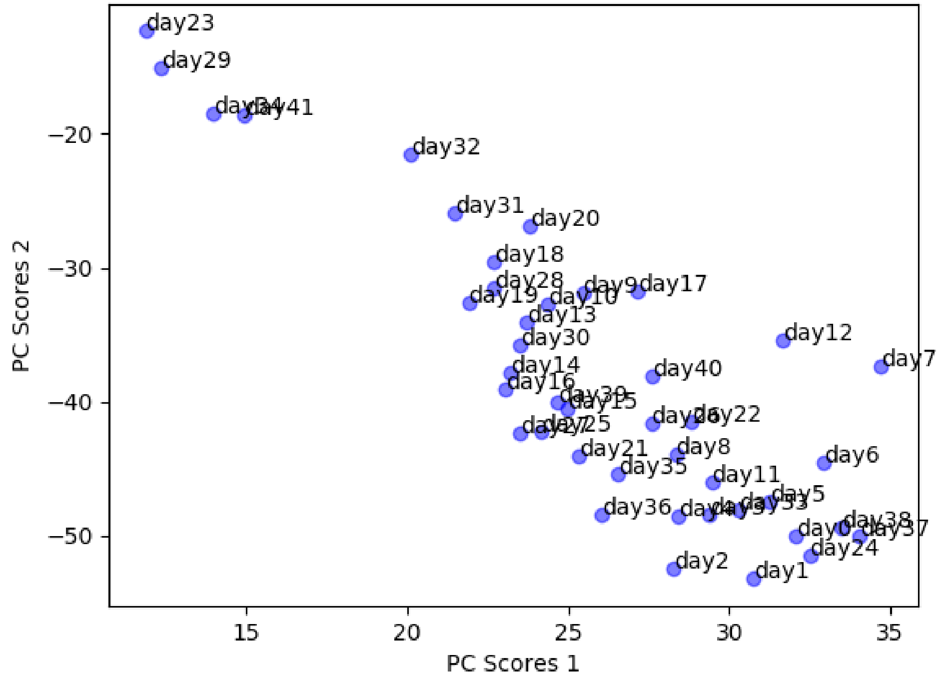
- Sample with replacement from data $x_1, x_2, \ldots, x_n$ to generate $x_1^*, x_2^*, \ldots, x_n^*$
- Compute the parameter of interest $\hat{\theta}^*$e.g. $\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$ and repeat the process for N times
- You can now plot a histogram of all $\hat{\theta}^*$'s or compute a $(1-p)\%$ confidence interval by taking the $\frac{p}{2}$ th or $(1-\frac{p}{2})$th percentiles of $\hat{\theta}^*$

Example- LA air pollution data

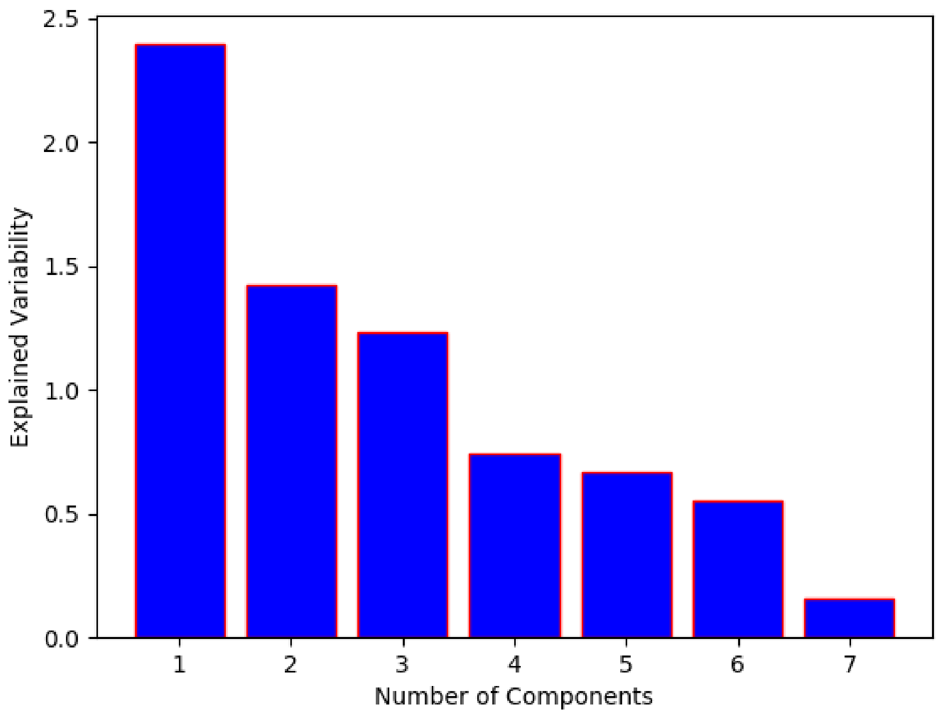
	PC1	PC2	PC3
Wind	-0.236821	0.278445	-0.643474
Solar radiation	0.205567	-0.526614	-0.224469
carbon monoxide	0.551084	-0.006820	0.113609
nitric oxide	0.377615	0.434674	0.407098
nitrogen dioxide	0.498016	0.199767	-0.196557
ozone	0.324551	-0.566974	-0.159847
hydrocarbon content	0.319403	0.307883	-0.541048

First two PCs

	PC1	PC2
Wind	-0.236821	0.278445
Solar radiation	0.205567	-0.526614
carbon monoxide	0.551084	-0.006820
nitric oxide	0.377615	0.434674
nitrogen dioxide	0.498016	0.199767
ozone	0.324551	-0.566974
hydrocarbon content	0.319403	0.307883

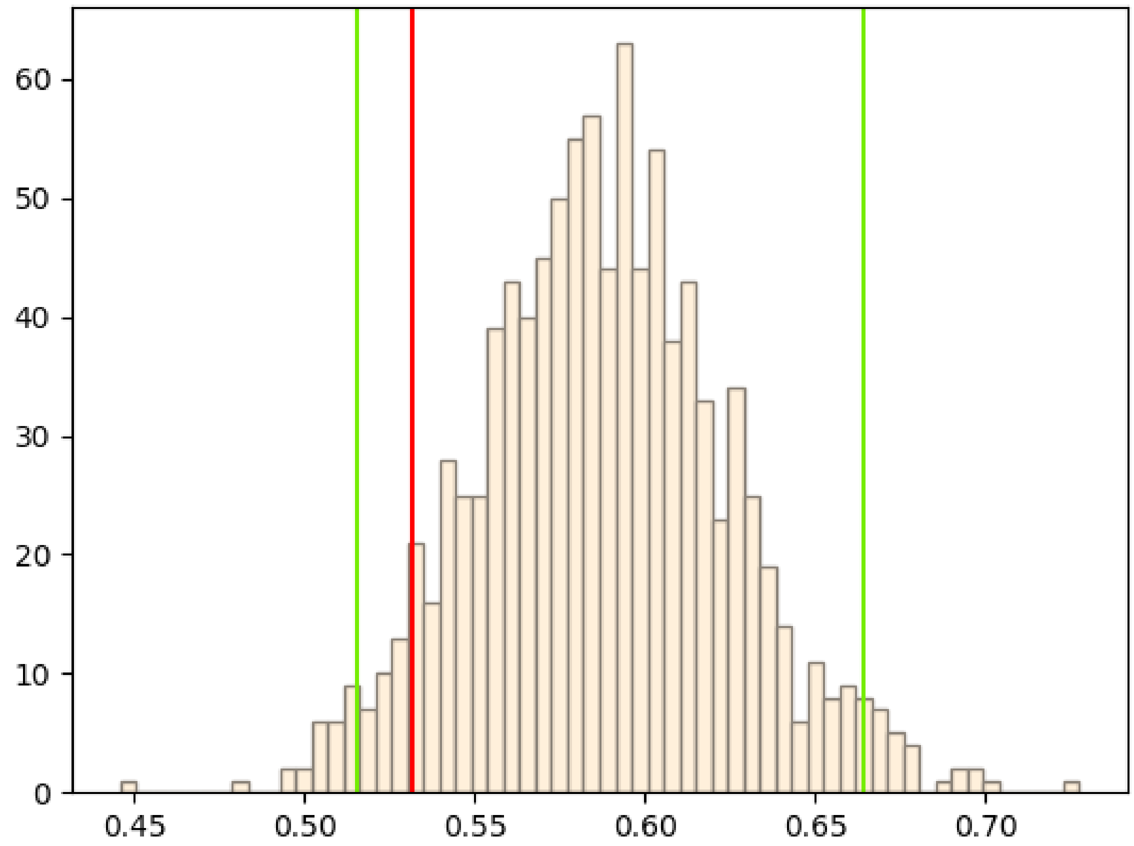


Variance explained



Can we build an empirical confidence interval for the explained variance by the 1st two PCs?

Bootstrap histogram for explained variance

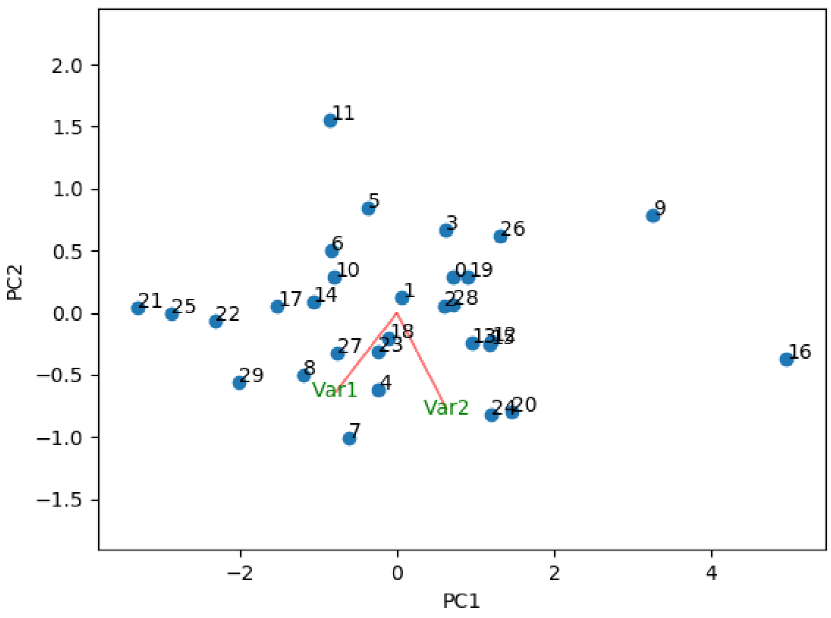


Biplots

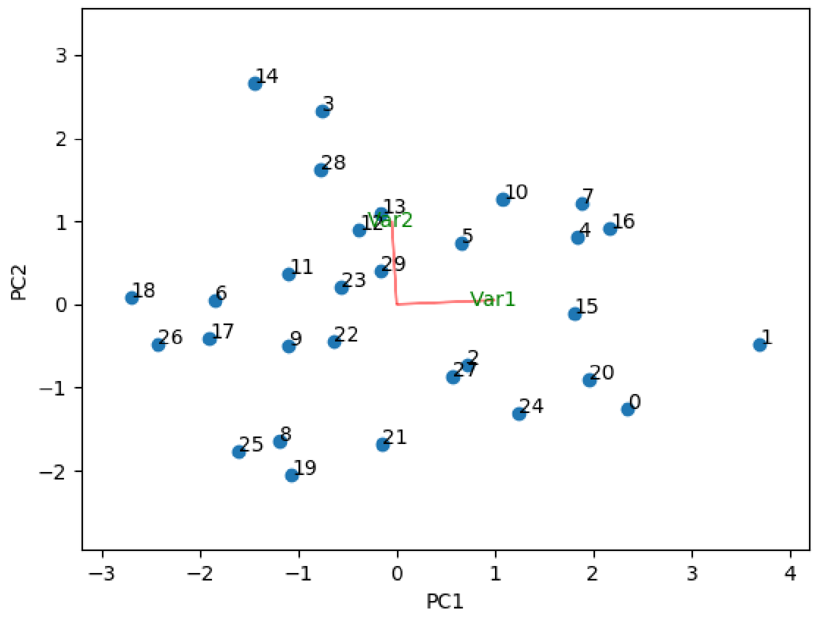
- Demonstrating both observations and variables on the same plot
- The variables are represented by directions
- (Cosines of) angles between directions are proportional to correlations between variables
- If the first two components do not explain most of variance, be cautious in interpretation of the biplot

Biplot-Simple example

$\rho = 0.9$

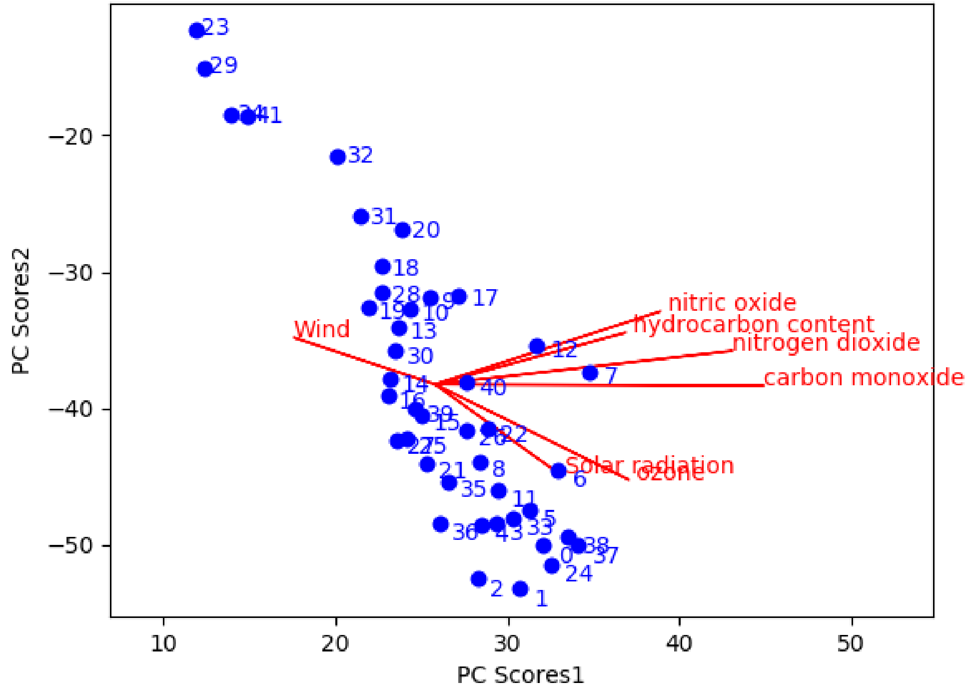


$\rho = 0$



Biplot for LA Air Pollution Data

	PC1	PC2
Wind	-0.236821	0.278445
Solar radiation	0.205567	-0.526614
carbon monoxide	0.551084	-0.006820
nitric oxide	0.377615	0.434674
nitrogen dioxide	0.498016	0.199767
ozone	0.324551	-0.566974
hydrocarbon content	0.319403	0.307883



Useful functions in Python

from sklearn.decomposition import PCA

P = PCA(n_components)

p.fit(X): fit a PCA model to a dataset X

p.components_: matrix of loadings

p.explained_variance_: The amount of variance explained by each of the selected components.

p.fit_transform(X): Compute the PC scores