Principal Component Analysis

Data Mining

Prof. Sujee Lee

Department of Systems Management Engineering

Sungkyunkwan University

Dimensionality Reduction

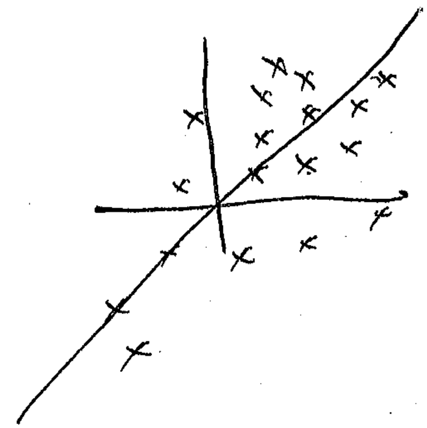
- **Dimensionality Reduction**
  - Find a new way to represent this data that summarizes the essential characteristics with fewer features.
  - **Motivation**: Algorithms like k-means are more computationally intensive in higher dimension and/or might take longer to converge
  - **Benefits**:
    - Computational Savings: compress data -> saving in time/space efficiency
    - Statistical Benefits: fewer dimension -> better generalization from fewer observations
    - Visualization: look at data in 1, 2, or 3 dimension (for identifying outliers, etc.)
  - **Methods** :
    - (Projection) Principal Component Analysis (PCA)
    - (Manifold Learning) t-distributed Stochastic Neighbor Embedding (t-SNE)

Principal Component Analysis

- **Principal Component Analysis**
  - Given a (training) dataset $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ such that $\boldsymbol{x}_i = \left(x_{i,1}, x_{i,2}, \ldots, x_{i,p}\right)$ is the i-th input vector of d features
  - **Goal** : Find "best" linear transformation $f: \mathbb{R}^p \rightarrow \mathbb{R}^q$ where $q < p$, which maps data into a q-dimensional space.
    - "best" means that the transformations maximally captures variation in the data.

Principal Component Analysis: simple case

- **Consider a case $\boldsymbol{p} = \mathbf{2}$, $\boldsymbol{q} = \mathbf{1}$** (1-dimensional projection)
  - Just like LDA, we want to find a direction $u \in \mathbb{R}^d$ (unit vector) that maximiz variance of data, so we consider the projected points $\{\boldsymbol{u}^T \boldsymbol{x}_i\}_{i=1}^n$, and we want **maximize sample variance**:



$$\frac{1}{n}\sum_i^n(\boldsymbol{u}^T\boldsymbol{x}_i - \boldsymbol{u}^T\overline{\boldsymbol{x}})^2 \text{ , where } \overline{\boldsymbol{x}} = \frac{1}{n}\sum_i^n \boldsymbol{x}_i$$

  - We can rewrite the **objective function** as

$$\frac{1}{n}\sum_i^n(\boldsymbol{u}^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}}))^2 = \frac{1}{n}\sum_i^n(\boldsymbol{u}^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\boldsymbol{T}}\boldsymbol{u})$$

$$= \boldsymbol{u}^T\left(\frac{1}{n}\sum_i^n(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\boldsymbol{T}}\right)\boldsymbol{u} = \boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{S}\boldsymbol{u}$$

for $\boldsymbol{S}$ defined from data

Principal Component Analysis: simple case

- So PCA amounts to solving:

$$\max_{||\boldsymbol{u}||_2=1} \boldsymbol{u}^T \boldsymbol{S}\boldsymbol{u}, \qquad \text{where } \boldsymbol{S} = \frac{1}{n}\sum_i^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\boldsymbol{T}}$$

- The solution is the top eigenvector of S, i.e., the eigenvector corresponding to the max eigenvalues

- **PCA algorithm for $\boldsymbol{q} = \mathbf{1}$**:

1. Construct the matrix $\boldsymbol{S} = \frac{1}{n}\sum_i^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\boldsymbol{T}}$
2. Compute the top eigenvector $\boldsymbol{u}$
3. Project all the data to $\{\boldsymbol{u}^T \boldsymbol{x}_i\}_{i=1}^n$

Principal Component Analysis: simple case

- It is convenient to write $\boldsymbol{S}$ in terms of data matrix $\boldsymbol{X}$:

  - Recall: If $\boldsymbol{X} = \begin{pmatrix} -\boldsymbol{x}_1^T - \\ \cdots \\ -\boldsymbol{x}_n^T - \end{pmatrix}$, then $\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} = \frac{1}{n}\sum_i^n \boldsymbol{x}_i\boldsymbol{x}_i^T$

- To construct $\boldsymbol{S}$, we can just construct a matrix $\widetilde{\boldsymbol{X}}$ by subtracting column means form row of $\boldsymbol{X}$:

$$\widetilde{\boldsymbol{X}} = \begin{pmatrix} -(\boldsymbol{x}_1 - \overline{\boldsymbol{x}})^T - \\ \cdots \\ -(\boldsymbol{x}_n - \overline{\boldsymbol{x}})^T - \end{pmatrix}, \text{ then } \boldsymbol{S} = \frac{\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}}{\boldsymbol{n}}$$

Singular Value Decomposition

- **Singular Value Decomposition (SVD)**

  - Why is $\max_{||\boldsymbol{u}||_2=\mathbf{1}} \boldsymbol{u}^T\boldsymbol{S}\boldsymbol{u}$ solved with max eigenvector of S?

  - Fact1: Any real, symmetric matrix has an orthonormal basis of eigenvectors

  - Face2: If $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ (real, not necessarily symmetric), we can always construct a singular value decomposition (SVD) of **A**:

$$\boldsymbol{A} = \boldsymbol{U}\begin{pmatrix} \boldsymbol{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\boldsymbol{V}^T$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$, $\boldsymbol{V} \in \mathbb{R}^{n\times n}$, the columns of $\boldsymbol{U}$ are orthonormal vectors, columns of $\boldsymbol{V}$ ar orthonormal, and $\boldsymbol{D}$ is diagonal with positive entries.

  - Columns of $\boldsymbol{U}$ are left singular vectors of $\boldsymbol{A}$
  - Columns of $\boldsymbol{V}$ are right singular vectors of $\boldsymbol{A}$
  - Values in $\boldsymbol{D}$ are singular values of $\boldsymbol{A}$

  - In particular, if $\boldsymbol{A}$ is symmetric, then $\boldsymbol{U} = \boldsymbol{V}$ and columns correspond to eigenvectors of $\boldsymbol{A}$.

Principal Component Analysis

- For any vector $\boldsymbol{u} \in \mathbb{R}^p$, we can write it as a linear combination of orthonormal eigenvectors $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_n\}$ so that

$$\boldsymbol{u} = c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 + \cdots + c_p\boldsymbol{u}_p$$

where $c_1^2 + c_2^2 + \cdots + c_p^2 = 1$ if $\boldsymbol{u}$ is a unit vector

- Now, we can write

$$\boldsymbol{u}^T\boldsymbol{S}\boldsymbol{u} = \left(c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 + \cdots + c_p\boldsymbol{u}_p\right)^T S\left(c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 + \cdots + c_p\boldsymbol{u}_p\right)$$

$$= \left(c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 + \cdots + c_p\boldsymbol{u}_p\right)^T\left(c_1\lambda_1\boldsymbol{u}_1 + c_2\lambda_2\boldsymbol{u}_2 + \cdots + c_p\lambda_p\boldsymbol{u}_p\right)$$

$$= c_1^2\lambda_1 + c_2^2\lambda_2 + \cdots + c_p^2\lambda_p$$

this is a weighted sum of $\lambda_1, \cdots, \lambda_p$, and maximized if $c_1 = 1$, and $c_2 = c_3 = \cdots = c_p = 0$.

i.e., $\boldsymbol{u} = \boldsymbol{u_1}$ is maximizer.

Principal Component Analysis

- Returning to PCA with q>1, we want to project, the data $\{\boldsymbol{x}_i\}_{i=1}^{n}$ onto a subspace spanned by orthogonal unit vectors $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_q\}$
- The coordinates with respect to the new basis are

$$\widetilde{\boldsymbol{x}}_{\boldsymbol{i}} \longmapsto \left(\boldsymbol{u}_{\boldsymbol{1}}^T\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}, \boldsymbol{u}_{\boldsymbol{2}}^T\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}, \cdots, \boldsymbol{u}_{\boldsymbol{q}}^T\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}\right) \in \mathbb{R}^q$$

We want to maximize the sum of squared lengths of projected vectors:

$$\max_{\{\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_q\}} \frac{1}{n}\sum_{i}^{n} \left\|\left(\boldsymbol{u}_{\boldsymbol{1}}^T\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}, \boldsymbol{u}_{\boldsymbol{2}}^T\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}, \cdots, \boldsymbol{u}_{\boldsymbol{q}}^T\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}\right)\right\|_2^2$$

Principal Component Analysis

- Using the matrix notation $\boldsymbol{U} = (\boldsymbol{u}_1\ \boldsymbol{u}_2 \dots \boldsymbol{u}_q) \in \mathbb{R}^{p\times q}$,

PCA is :

$$\max_{\boldsymbol{U}\in\mathbb{R}^{p\times q}:\boldsymbol{U}^T\boldsymbol{U}=I} \frac{1}{n}\sum_{i}^{n} \left|\left|\boldsymbol{U}^{\boldsymbol{T}}\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}\right|\right|_2^2$$

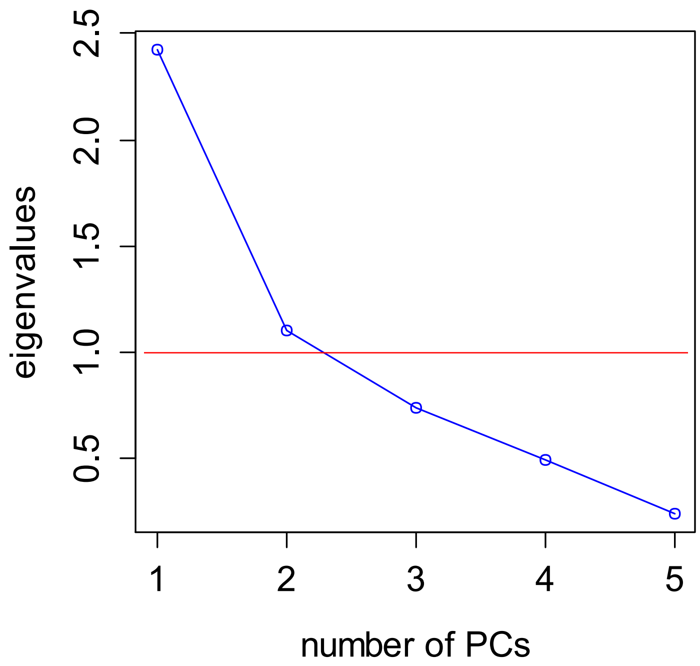
Maximizer is when $\boldsymbol{U}$ consists of top q eigenvectors of $\boldsymbol{S} = \frac{\widetilde{\boldsymbol{X}}^{\boldsymbol{T}}\widetilde{\boldsymbol{X}}}{\boldsymbol{n}}$

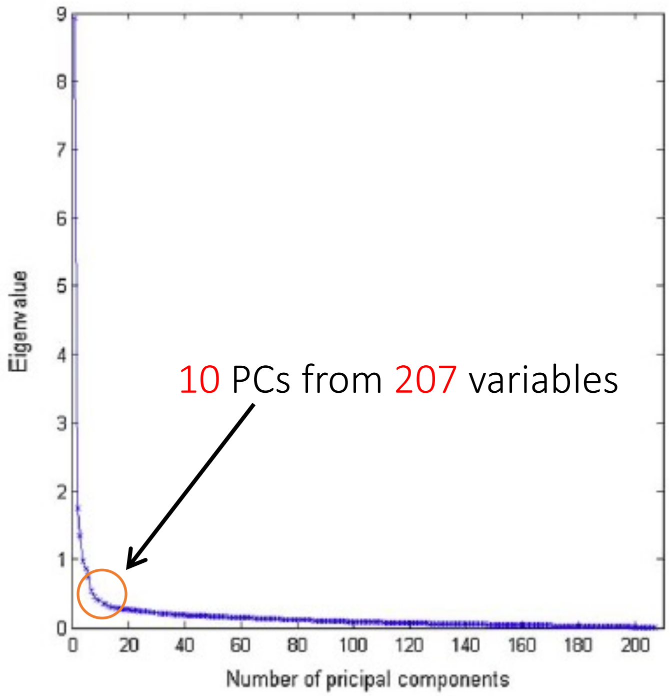
- PCA output is to map each $\boldsymbol{x}_{\boldsymbol{i}}$ to $\boldsymbol{U}^{\boldsymbol{T}}\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}$
- Note: another interpretation of the objective function is to minimize sum of squared distances to

space spanned by $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_q\}$

Number of Principal Components

- **How many PCs? – PCA as Dimension Reduction Tech.**
  - 1) visualization: q=1,2, or 3
  - 2) computational consideration: q is at most something
  - 3) Elbow plot: plot eigenvalues of $\boldsymbol{S}$, find elbow





Number of Principal Components

- 4) compute % of variation explained by principal components:

Consider eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$,

Compute ratio $\frac{\lambda_1+\lambda_2+\cdots+\lambda_q}{\sum_i^p \lambda_i}$, which is a fraction of explained variance.

Determine a proper q considering it

Principal Component Analysis

- **PCA recap.**
  - PCA maximizes variation in p-dimensional data when projected onto a q-dimensional subspace

q<p

1. Recenter the data: $\boldsymbol{x_i} \rightarrow \widetilde{\boldsymbol{x}}_{\boldsymbol{i}}$, where $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}} = \boldsymbol{x_i} - \overline{\boldsymbol{x}}$

   this gives a data matrix instead of $\widetilde{\boldsymbol{X}}$ instead of $\boldsymbol{X}$

2. Construct the matrix $\boldsymbol{S} = \frac{1}{n}\widetilde{\boldsymbol{X}}^{\boldsymbol{T}}\widetilde{\boldsymbol{X}} = \frac{1}{n}\sum_i^n \widetilde{\boldsymbol{x}}_{\boldsymbol{i}}\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}^T \in \mathbb{R}^{p \times p}$

3. Compute the q eigenvectors corresponding to the largest q eigenvalues of $\boldsymbol{S}$. Store them in matrix $\boldsymbol{U} = (\boldsymbol{u}_1\ \boldsymbol{u}_2 \ldots \boldsymbol{u}_q) \in \mathbb{R}^{p \times q}$

4. Map data to $\{\boldsymbol{U}^T \widetilde{\boldsymbol{x}}_{\boldsymbol{i}}\}_{i=1}^n$