Introduction to Statistical learning

Omid Safarzadeh

January 19, 2022

Table of contents

Unsupervised learning

- Supervised learning: $Trainingdata = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Unsupervised learning: $Trainingdata = \{\mathbf{x}_i\}_{i=1}^{n}$
- Can be used on its own, but also as a pre-processing step before supervised learning!

Principal component analysis (PCA)

- $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{ip}]^T$
- Can we find $\mathbf{x}_i = [z_{i1}, z_{i2}, ..., z_{ik}]^T, k << p$ such that $\mathbf{x}_i \approx \mathbf{z}_i$ (in some sense)?
- Find a k-dimensional subspace in which the data approximately lies.
- I.e., the subspace captures almost all variation in the data.

Centralize and standardize

- $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$, $j = 1, ..., p$
- $\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$, $j = 1, ..., p$
- $x_{ij} \leftarrow \frac{x_{ij} - \bar{x}_j}{\sigma_j}$, $j = 1, ..., p$, $i = 1, ..., n$.
- Dividing by σ_j is not necessary if attributes are on the same scale. Ex: Each correspond to value in USD

Computing the major axis of variation

- Major axis of variation = first principal component
- Projection of $\mathbf{x} = [x_1, ..., x_p]$ to $\mathbf{u} = [u_1, ..., u_p]^T$ is $\mathbf{z} = \mathbf{x}^t\mathbf{u}\mathbf{u}$

Variance of the projections:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{u})^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{u})^T(\mathbf{x}_i^T\mathbf{u})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{u}$$

$$= \mathbf{u}^T(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T)\mathbf{u}$$

Optimization problem

- $\Sigma = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ is the sample covariance matrix of the data

$$\text{maximize } \mathbf{u}^T \Sigma \mathbf{u} \text{ subject to } ||\mathbf{u}|| = 1$$

Method of Lagrange multipliers:

$$\Rightarrow \text{maximize } \mathbf{u}^T \Sigma \mathbf{u} \text{ subject to } \mathbf{u}^T \mathbf{u} = 1$$

$$\Rightarrow \text{maximize } \mathscr{L}(\mathbf{u}, \lambda) = \mathbf{u}^T \Sigma \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1),$$

$$\Rightarrow \Sigma \mathbf{u} = \lambda \mathbf{u}$$

Result

- Principal eigenvector of Σ is the principal component
- k eigenvectors with the k largest eigenvalues form all k principal components

PCA algorithm

1. Given $\mathbf{X} = [\mathbf{x}_1^T, ..., \mathbf{x}_n^T]^T$, compute $\Sigma = \frac{1}{n}\mathbf{X}^T\mathbf{X}$
2. Compute all eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ of Σ, and the corresponding eigenvectors $\mathbf{u}_1, ..., \mathbf{u}_p$
3. Pick $k \leq p$ eigenvectors with the largest eigenvalues, i.e., $\mathbf{u}_1, ..., \mathbf{u}_k$
4. For all $i = 1, ..., n$, let $\mathbf{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{ik} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^T\mathbf{u}_1 \\ \mathbf{x}_i^T\mathbf{u}_2 \\ \vdots \\ \mathbf{x}_i^T\mathbf{u}_k \end{bmatrix}$
5. $\{\mathbf{z}_i\}_{i=1}^n$ is the k-dimensional approximation to $\{\mathbf{x}_i\}_{i=1}^n$

Can we recover $\mathbf{x}_i$ exactly from $\mathbf{z}_i$?

$$\hat{\mathbf{x}}_i = z_{i1}\mathbf{u}_1 + z_{i2}\mathbf{u}_2 + \cdots + z_{ik}\mathbf{u}_k$$

PCA example - Arrest dataset

- Number of arrests per 100; 000 residents for 50 states in the US ($n = 50$)
- $X_1 = Assault, X_2 = Murder, X_3 = Rape, X_4 = UrbanPop$ (percent living in urban areas), $p = 4$
- Centralize and standardize the data, and then, apply PCA ($k = 2$)
- Principal component loading vectors:

$$\mathbf{u}_1 = \begin{bmatrix} 0.54 \\ 0.58 \\ 0.54 \\ 0.28 \end{bmatrix} \qquad \mathbf{u}_2 = \begin{bmatrix} -0.42 \\ -0.19 \\ 0.17 \\ 0.87 \end{bmatrix}$$

- For each $\mathbf{x}_i = [x_{i1}, x_{i2}, x_{i3}, x_{i4}]^T$ compute principal component scores $\mathbf{z}_i = [z_{i1}, z_{i2}]^T$, where

$$z_{i1} = \mathbf{x}_i^T \mathbf{u}_1 = 0.54x_{i1} + 0.58x_{i2} + 0.54x_{i3} + 0.28x_{i4}$$

$$z_{i2} = \mathbf{x}_i^T \mathbf{u}_2 = -0.42x_{i1} - 0.19x_{i2} + 0.17x_{i3} + 0.87x_{i4}$$

Data visualization using PCA

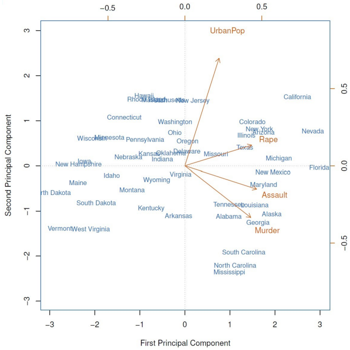

Figure: James et al., 2013

$$z_{i1} = \mathbf{x}_i^T \mathbf{u}_1 = 0.54x_{i1} + 0.58x_{i2} + 0.54x_{i3} + 0.28x_{i4}$$
$$z_{i2} = \mathbf{x}_i^T \mathbf{u}_2 = -0.42x_{i1} - 0.19x_{i2} + 0.17x_{i3} + 0.87x_{i4}$$

Interpretation of the results

- Crime-related variables are correlated with each other
- UrbonPop is less correlated with crime-related variables
- States with high value in first component = states with high crime rates
- States with high value in the second component = states with high level of urbanization

Importance of standardization in PCA

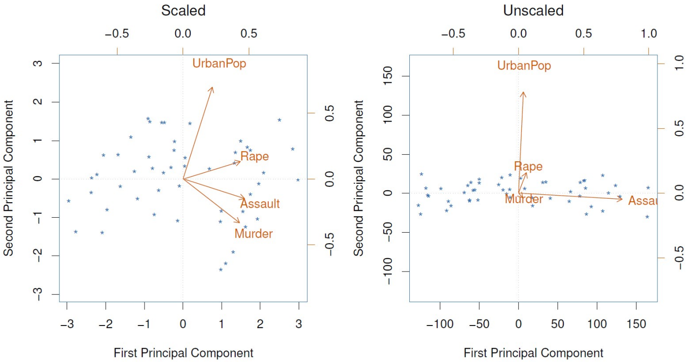

Figure: James et al., 2013

- Left: unit variance, Right: no scaling

How to choose k?

- Proportion of variance explained (PVE)
- Total variance:

$$\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

- Variance explained by the mth principal component

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{u}_m)^2 = \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} x_{ij} u_{mj})^2$$

- $PVE(m) = \frac{\sum_{i=1}^{n} (\sum_{j=1}^{p} x_{ij} u_{mj})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$
- $PVE(\text{first } k) = \sum_{m=1}^{k} PVE(m)$

How to choose k?

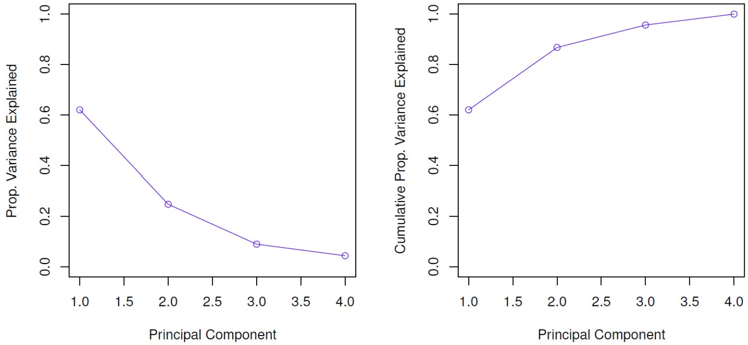

Figure: James et al., 2013

- Left: $PVE(m)$. Right:$PVE(\text{first } k)$

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. Springer New York. https://books.google.fr/books?id=qcI%5C_AAAAQBAJ