Lecture 4: Finish PCA

Reading: 10.3, 10.5

STATS 202: Data mining and analysis

Jonathan Taylor, 10/1
Slide credits: Sergio Bacallado

PCA: Summary of last lecture

- The first principal component ϕ_1 is a unit vector of length p, which maximizes the variance of the projections or *scores* $z_{j,1} = x_j \cdot \phi_1$ for $j = 1, \ldots, n$.

PCA: Summary of last lecture

- The first principal component ϕ_1 is a unit vector of length p, which maximizes the variance of the projections or *scores* $z_{j,1} = x_j \cdot \phi_1$ for $j = 1, \ldots, n$.
- The second principal component ϕ_2 is a unit vector, orthogonal to ϕ_1, which maximizes the variance of the scores $z_{j,2}$, $j = 1, \ldots, n$.

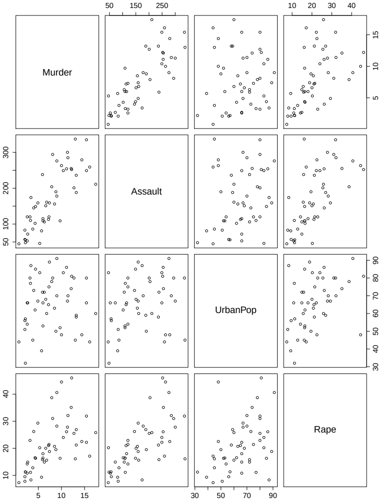
PCA: Summary of last lecture

- The first principal component ϕ_1 is a unit vector of length p, which maximizes the variance of the projections or *scores* $z_{j,1} = x_j \cdot \phi_1$ for $j = 1, \ldots, n$.
- The second principal component ϕ_2 is a unit vector, orthogonal to ϕ_1, which maximizes the variance of the scores $z_{j,2}$, $j = 1, \ldots, n$.
- The third principal component ϕ_3 is orthogonal to ϕ_1 and ϕ_2, and so on...

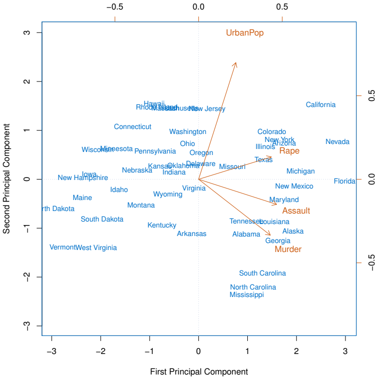
PCA: Summary of last lecture

- The first principal component ϕ_1 is a unit vector of length p, which maximizes the variance of the projections or *scores* $z_{j,1} = x_j \cdot \phi_1$ for $j = 1, \ldots, n$.
- The second principal component ϕ_2 is a unit vector, orthogonal to ϕ_1, which maximizes the variance of the scores $z_{j,2}$, $j = 1, \ldots, n$.
- The third principal component ϕ_3 is orthogonal to ϕ_1 and ϕ_2, and so on...
- If $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T$ is the *singular value decomposition* of $\mathbf{X}$, the principal components are the columns of Φ.

How many principal components are enough?



How many principal components are enough?



We said 2 principal components capture most of the relevant information. But how can we tell?

The proportion of variance explained

We can think of the top **principal components** as directions in space in which the data vary the most.

The proportion of variance explained

We can think of the top **principal components** as directions in space in which the data vary the most.

The ith **score vector** $(z_{1i}, \ldots, z_{ni})$ can be interpreted as a *new* variable. The variance of this variable decreases as we take i from 1 to p.

The proportion of variance explained

We can think of the top **principal components** as directions in space in which the data vary the most.

The ith **score vector** $(z_{1i}, \ldots, z_{ni})$ can be interpreted as a *new* variable. The variance of this variable decreases as we take i from 1 to p. However, the total variance of the score vectors is the same as the total variance of the original variables:

$$\sum_{i=1}^{p} \frac{1}{n} \sum_{j=1}^{n} z_{ji}^2 = \sum_{i=1}^{p} \mathsf{Var}(x_{:,i}).$$

The proportion of variance explained

We can think of the top **principal components** as directions in space in which the data vary the most.

The ith **score vector** $(z_{1i}, \ldots, z_{ni})$ can be interpreted as a *new* variable. The variance of this variable decreases as we take i from 1 to p. However, the total variance of the score vectors is the same as the total variance of the original variables:

$$\sum_{i=1}^{p} \frac{1}{n} \sum_{j=1}^{n} z_{ji}^2 = \sum_{i=1}^{p} \mathsf{Var}(x_{:,i}).$$

We can quantify how much of the variance is captured by the first m principal components/score variables.

The proportion of variance explained

The variance of the mth score variable is:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2$$

The proportion of variance explained

The variance of the mth score variable is:

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

The proportion of variance explained

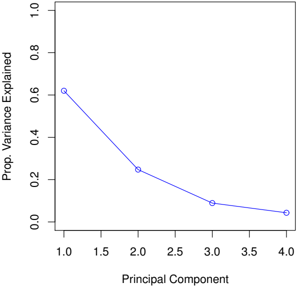
The variance of the mth score variable is:

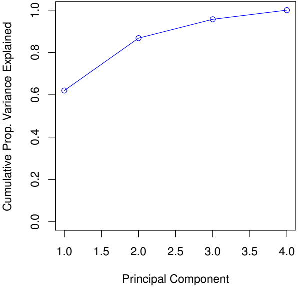
$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2 = \frac{1}{n}\mathbf{\Sigma}_{mm}^2.$$

The proportion of variance explained

The variance of the mth score variable is:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2 = \frac{1}{n}\Sigma_{mm}^2.$$

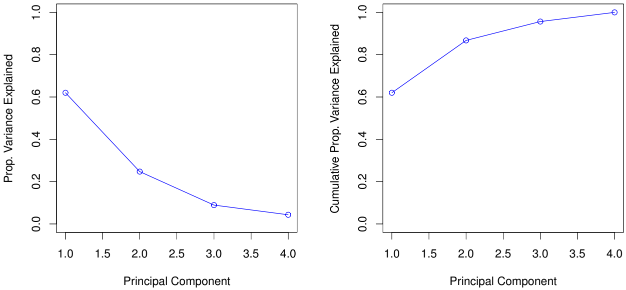




The proportion of variance explained

The variance of the mth score variable is:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2 = \frac{1}{n}\mathbf{\Sigma}_{mm}^2.$$



Scree plot