Principal Component Analysis

Linear Algebra

This work was initially funded by an Institutional OER Grant from the Colorado Department of Higher Education (CDHE). For similar OER materials in other courses funded by this project in the Department of Mathematical and Statistical Sciences at the University of Colorado Denver, visit
`https://github.com/CU-Denver-MathStats-OER`

Data Matrices

Suppose that we've conducted a study and measured various characteristics. We put our data into a matrix D, where the columns are each feature and the rows are each member of the study (observations).

Eg, people, and measuring height, weight, and age.

$$\begin{array}{ccccc} & \text{height} & \text{weight} & \text{age} & \cdots \\ \text{Irwin} & 72 \text{ inches} & 187 \text{ pounds} & 32 \text{ years} & \cdots \\ \text{Quinn} & 61 \text{ inches} & 132 \text{ pounds} & 55 \text{ years} & \cdots \\ \vdots & & & & \end{array}$$

If we have p features and N observations, D is an $N \times p$ matrix.

Mean and Covariance

The mean μ or average of a feature $\mathbf{x}$ can be computed by $\frac{1}{N}\sum x_i$.

The (sample) variance of a distribution is the average[1] square deviation from the mean.

$$\text{var} = \frac{1}{N-1}\sum (x_i - \mu)^2.$$

[1] almost

Mean and Covariance

The mean μ or average of a feature $\mathbf{x}$ can be computed by $\frac{1}{N}\sum x_i$.

The (sample) variance of a distribution is the average[1] square deviation from the mean.

$$\text{var} = \frac{1}{N-1}\sum(x_i - \mu)^2.$$

The covariance of two distributions $\mathbf{x}$ and $\mathbf{y}$ is given by

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{N-1}\sum(x_i - \mu_x)(y_i - \mu_y).$$

The covariance is 0 when the two distributions are uncorrelated.

[1] almost

Mean and Covariance

The mean μ or average of a feature $\mathbf{x}$ can be computed by $\frac{1}{N}\sum x_i$.

The (sample) variance of a distribution is the average[1] square deviation from the mean.

$$\mathrm{var} = \frac{1}{N-1}\sum (x_i - \mu)^2.$$

The covariance of two distributions $\mathbf{x}$ and $\mathbf{y}$ is given by

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{N-1}\sum (x_i - \mu_x)(y_i - \mu_y).$$

The covariance is 0 when the two distributions are uncorrelated.

From our data matrix D, let $A = [a_{ij}]$, where $a_{ij} = d_{ij} - \mu_j$ is the deviation from the mean.

Then the matrix of covariances is $S = \frac{1}{N-1}A^T A$.

[1] almost

Principal Component Analysis

We would like the features to be uncorrelated so that we can see the individual effects.

Principal Component Analysis

We would like the features to be uncorrelated so that we can see the individual effects.

Find an orthogonal matrix P such that $AP^T = B$, ie, for each observation, produce a new list of features and measurements that are linear combinations of the original features.

We want the columns $\mathbf{b}_i$ to be uncorrelated,

Principal Component Analysis

We would like the features to be uncorrelated so that we can see the individual effects.

Find an orthogonal matrix P such that $AP^T = B$, ie, for each observation, produce a new list of features and measurements that are linear combinations of the original features.

We want the columns $\mathbf{b}_i$ to be uncorrelated, ie, covariance matrix $\frac{1}{N-1}B^TB$ to be diagonal.

The cov matrix of B is $\frac{1}{N-1}PA^TAP^T$. Since $\frac{1}{N-1}A^TA$ is symmetric, we can orthogonally diagonalize it.

The columns of P (which are unit eigenvectors of A^TA) are called the principal components of the data. These are also right singular vectors of A.

Principal Component Analysis

We would like the features to be uncorrelated so that we can see the individual effects.

Find an orthogonal matrix P such that $AP^T = B$, ie, for each observation, produce a new list of features and measurements that are linear combinations of the original features.

We want the columns $\mathbf{b}_i$ to be uncorrelated, ie, covariance matrix $\frac{1}{N-1}B^TB$ to be diagonal.

The cov matrix of B is $\frac{1}{N-1}PA^TAP^T$. Since $\frac{1}{N-1}A^TA$ is symmetric, we can orthogonally diagonalize it.

The columns of P (which are unit eigenvectors of A^TA) are called the principal components of the data. These are also right singular vectors of A.

The sum $\sum \lambda_i$ of the eigenvalues of $\frac{1}{N-1}A^TA$ is a measure of the total variance of the data.

The scaled values $\frac{\lambda_i}{\sum \lambda_i}$ indicate how important each principal component is. Less important principal components can be dropped for dimension reduction.