Classnote - Modeling complex phenotypes with latent variables and principal components

Dimensionality reduction

When working with high dimensional data, which happens often with biological data, we often want to represent it using less dimensions. This means that we approximate the original data in less dimensional space. To achieve this goal techniques such as **Principal Components Analysis** and **Singular Values Decomposition** are used.

SVD

Singular values decomposition is a method that allows for a factorisation of a matrix (for simplicity I'll assume it's real valued, however, it holds also for complex valued matrices), into product of two orthogonal matrices and one diagonal, it can be written as: $X = U\Sigma V^{\mathsf{T}}$. Where: X is a real valued matrix to be factorised, U and V are orthogonal matrices and Σ is a diagonal matrix with singular values of X.

Intuitively speaking SVD allows for generalisation of diagonalisation of symmetric square matrix for non square and non symmetric matrices. This allows to reduce dimensionality by substituting small singular values with 0s. This leads to loss of some information, but when conducted with proper care the loss can be relatively small. Such operation guarantees that the resulting matrix A will be the closest $n - k$ - rank (where n is the rank of X and k is the number of singular values substituted with zeros) matrix to the X matrix in the sense of least squares distance, i. e. $||X - A||_F$ is minimal.

PCA

Principal components analysis is similar to **SVD**. To compute the **PCA** one shall compute eigendecomposition of data covariance matrix, which is square and symmetric so it's easily diagonalisable. Primary components are then the eigenvalues of the covariance matrix and can be used for dimensionality reduction in similar way to singular values.

Intuition (taken from Wikipedia article on PCA)

PCA can be thought of as fitting a *p*-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipsoid is small, then the variance along that axis is also small.

To find the axes of the ellipsoid, we must first center the values of each variable in the dataset on 0 by subtracting the mean of the variable's observed values from each of those values. These transformed values are used instead of the original observed values for each of the variables. Then, we compute the covariance matrix of the data and calculate the eigenvalues and corresponding eigenvectors of this covariance matrix. Then we must normalise each of the orthogonal eigenvectors to turn them into unit vectors. Once this is done, each of the mutually-orthogonal unit eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data. This choice of basis will transform the covariance matrix into a diagonalised form, in which the diagonal elements represent the variance of each axis. The proportion of the variance that each eigenvector represents can be calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all eigenvalues.

Biplots and scree plots (degree of explained variance) are used to interpret findings of the PCA.

Latent variable models

What is latent variable?

Latent variables are the variables that we assume occur in the process that are not observed, but influence and are influenced by the observed variables.

For example let's assume that we have data fro car insurance company. In this data we can find information about: car model, driver's age, whether the driver has had an accident in the past year. Then this three are observable variable, and for example the year at which the driver's has got theirs drivers licence is a latent variable. It wasn't observed, but there is relationship between observed variables and this one, e.g. driver couldn't have got theirs drivers' licence before they were 18 years old, and 'new' drivers are more prone to having had an accident in the last year.

What are latent variable models and how do they use latent variables?

Latent variable models relate latent and observable variables. One example of such a model might be a Gaussian Mixture Model in which we want to cluster observations according to their classes (which are latent variables).

Phenotypes

Phenotype is a set of observable features of a given organism.

Bibliography

Matematyka obliczeniowa - slajdy 197 - 202
Wikipedia article on PCA
Wall et al. 2014
Statistical Data Analysis 2 slides from the academic year 2022/23 (author: Ewa Szczurek)