Applied Machine Learning in Health Sciences 2023
-
Principal component analysis

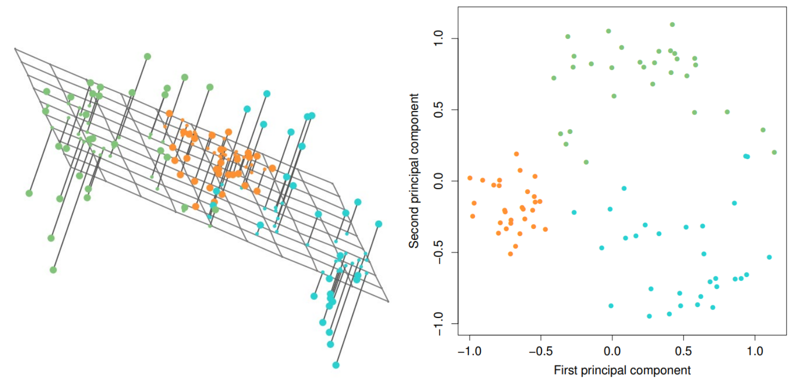
Peter Mondrup Rasmussen

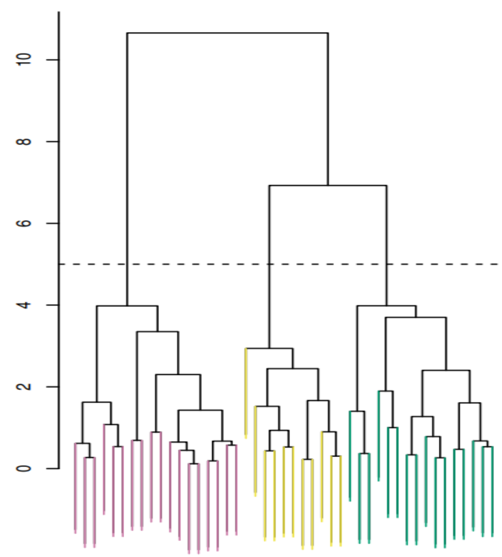
CFIN

pmr@cfin.au.dk

Unsupervised learning

- In *supervised learning* we have an input X with p features $X_1, X_2, \dots, X_p$ and a corresponding response Y. Our goal often is to predict Y based on the input X.
- In *unsupervised learning* we only consider the features $X_1, X_2, \dots, X_p$ and the goal is often to discover new *structure* in the data or to learn new *representations* of the data.
- Unsupervised learning is often used for *exploratory data analysis* where the analysis tend to be more subjective compared to supervised learning, where there often is a well-defined goal (predicting the response as good as possible).





Principal component analysis

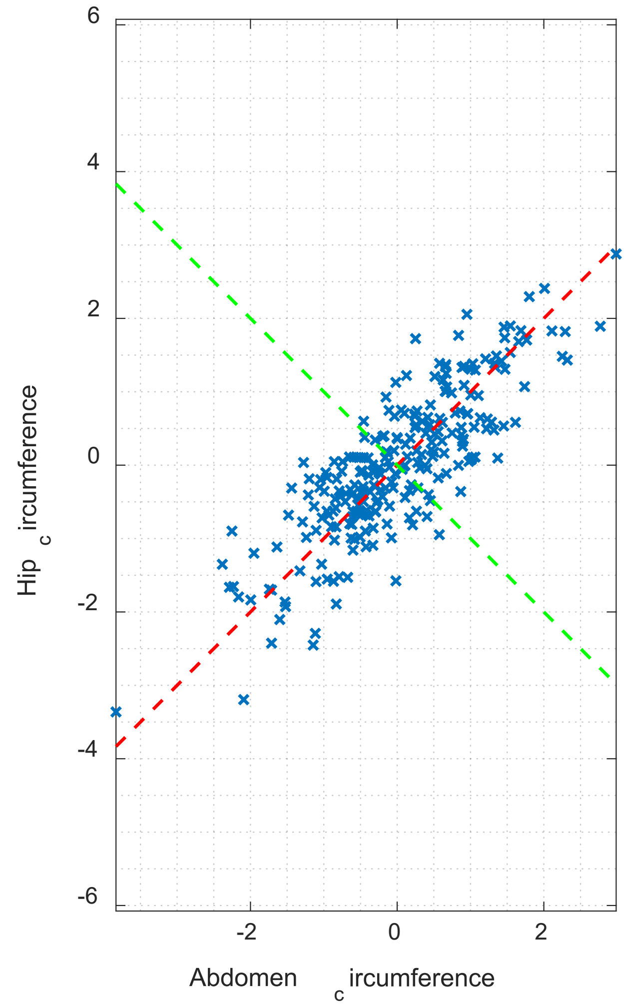
Principal component analysis

PCA provides a tool for finding a *low-dimensional representation* of the data that captures as much of the information as possible. PCA finds new *interesting* dimensions along which the observations vary the most.

These dimensions or *principal components* are found as linear combinations of the original input features, e.g. the first principal component

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

The individual ϕ_{jm}'s above are elements of the *loading vector* ϕ_1 (which is normalized to unit length).



Principal component analysis

- Consider a $(n \times p)$ data set $\boldsymbol{X}$ that has been centered (column mean is zero). For each of the observations x_i we can compute the *scores* of the first principal component

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip} = \sum_{j}^{p} \phi_{j1}x_{ij} \,.$$

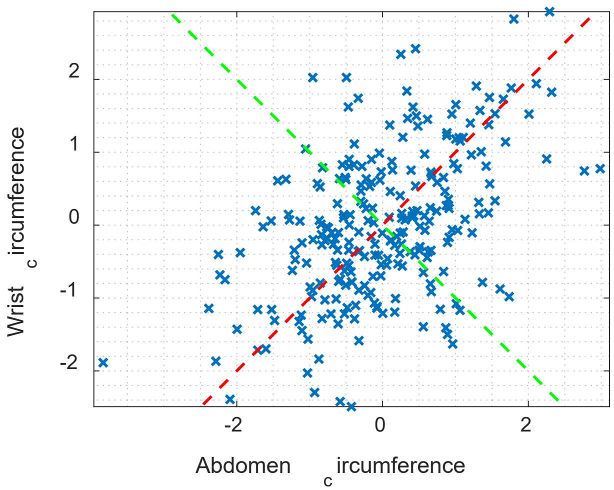
- The goal is to find ϕ's so that these projections has the largest variance

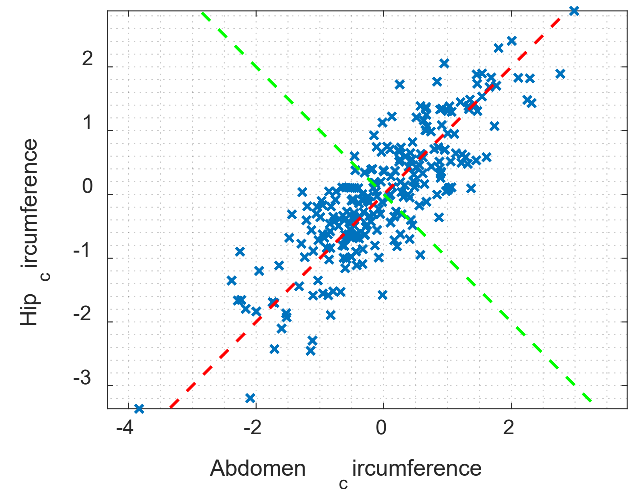
$$\underset{\phi}{\text{maximise}} \left\{ \frac{1}{n} \sum_{i=1}^{n} z_{i1}^2 \right\} \text{ subject to } \left|\left|\phi_1\right|\right| = 1.$$

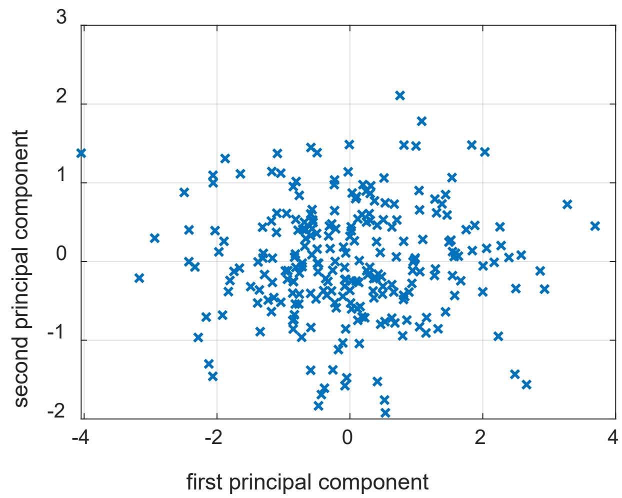
- The loading vector $\phi_1$defines the direction in feature space along which the projected data or scores vary the most.
- The second principal component Z_2 is then found in a similar way as a linear combination of $X_1, X_2, \dots, X_p$ but with the additional requirements that Z_2must be *uncorrelated* with Z_1. This constraint is equivalent to constraining ϕ_2 to be *orthogonal* to ϕ_1.
- The scores of the second principal component are

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip}.$$

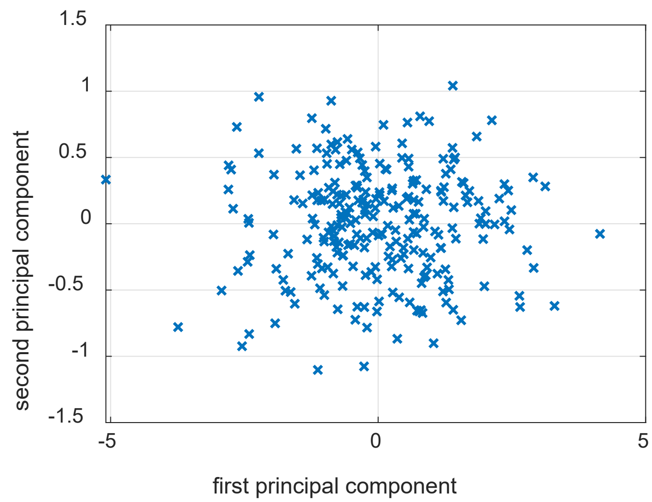
Principal component analysis





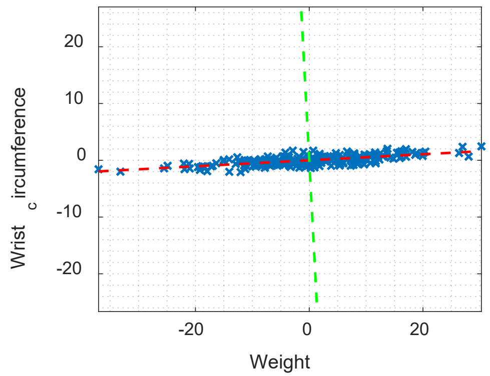


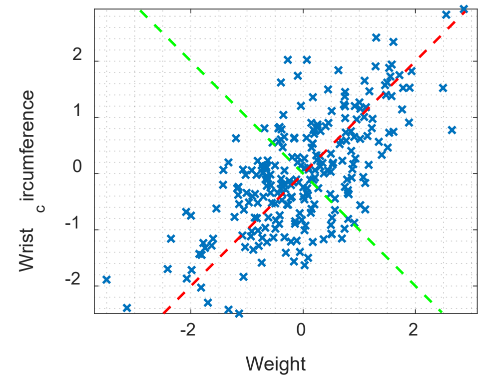
var exp.: 0.74 0.26

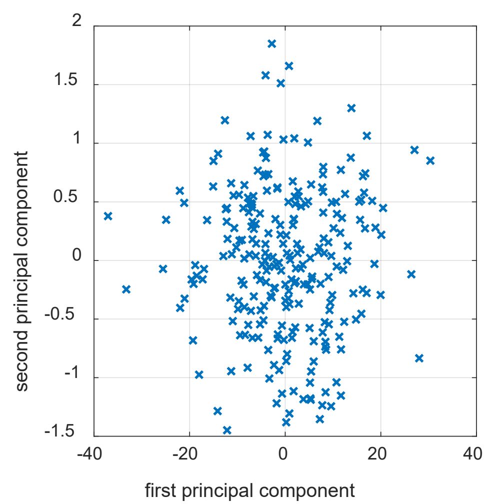


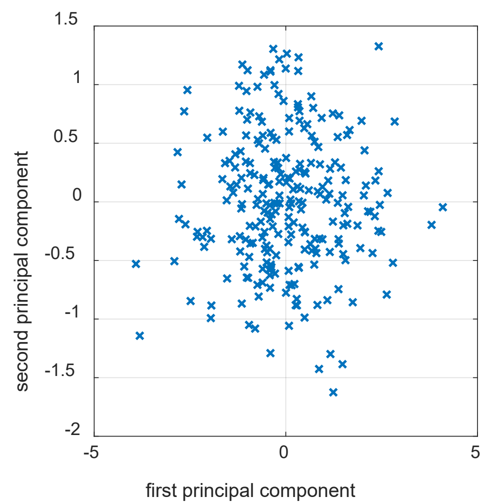
var exp.: 0.92 0.08

Principal component analysis









original feature scale
var exp.: 0.997 0.003

standardized features
var exp.: 0.83 0.17

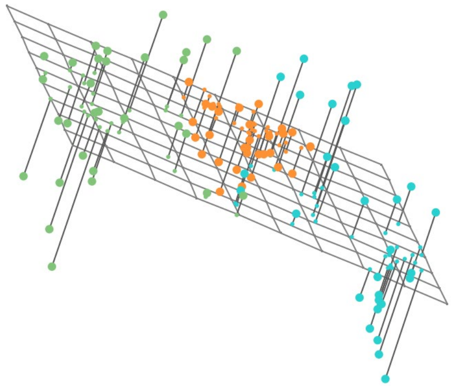
Principal component analysis – reconstruction

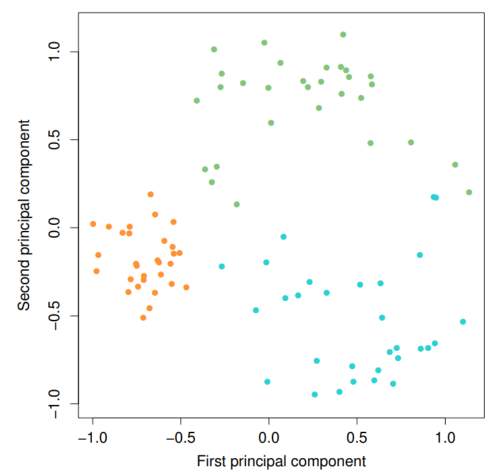
- Given a set of the first M principal component score vectors and the corresponding loading vectors, we can compute reconstructed points in the original feature space $\hat{x}_i$ as approximations of the original observations with coordinates

$$\hat{x}_{ij} = \sum_{m=1}^{M} z_{im}\phi_{jm}$$

- Principal component analysis can be interpreted as a method that minimize the squared Euclidean distance between the original observations and the reconstructions

$$\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - \hat{x}_{ij}\right)^2$$

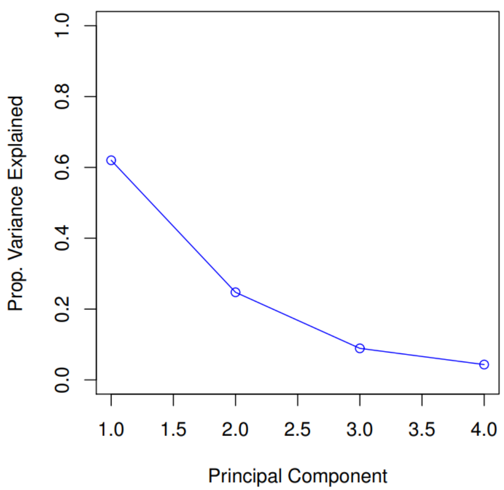


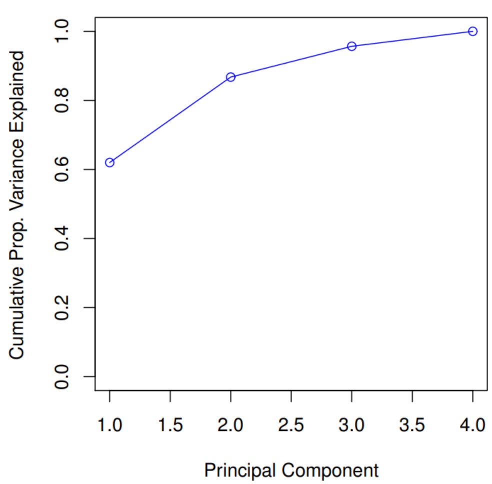


Principal component analysis – variance explained

- Often, we are interested in assessing how much of the total variance that the first M principal components accounts for.
- Proportion of variance explained (PVE) describe how much variance individual PCs account for

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



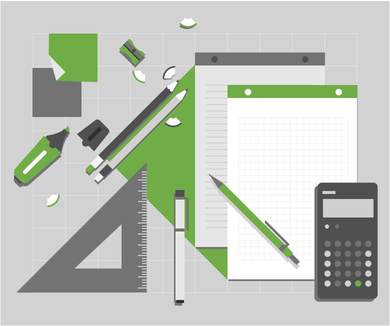


Principal component analysis

Practical issues

Principal component analysis – practical issues

- **Scaling the variables:** The relative scale of input features has an importance. Generally advisable to scale/standardize data, especially if different measurement units are used. If the same measurement unit is used (i.e. features are comparable) the it could be useful not to scale the variables before PCA.
- **Uniqueness of the PCs:** Principal component loading vectors are only unique up to a sign flip.
- **Deciding how many PCs to use:** Look at scree plot for an elbow. Look for interesting structure starting from first components and continue until no further interesting structure is found.
- **Noise reduction:** PCA could be used for noise reduction, and the resulting new data representation could be used as inputs to e.g. a supervised learning methods.
- **Linear method:** PCA is a linear method and can only find new linear representations. Other techniques exist for finding non-linear representations, e.g. auto-encoders, kernel PCA, local linear embedding, multidimensional scaling.



In-class exercises

9 Principal component analysis

References

Figures from James et al. *An Introduction to Statistical Learning*, second edition, https://www.statlearning.com/resources-second-edition