# Principal Component Analysis

Dr. H.G.J. van Mil 

# Contents

# Introduction

Principal component analysis (PCA) is a dimension reduction method, or *unsupervised* machine learning method in modern statistical language. PCA is an exploratory method and needs to be contrasted with confirmatory methods (significance tests), and prediction methods (*supervised* machine learning).

Exploratory methods are used when there is no clear enough understanding of our system of interest to generate a meaningful hypothesis; a hypothesis for confirmatory statistics. One example is the measurement of many morphological or genetic features of different species within a genus. The dataset is then explored to identify those features that that can be used to classify or relate species to each other.

Now imagine a spreadsheet with individual subjects as rows and features or properties as columns. An example could be a dataset where each row is an ID representing an individual, and each of 250 columns are the features measured on that individual. In the context of dimension reduction (PCA) this dataset contains 250 dimensions. It's clear that we cannot visualize 250 dimensions effectively in order to explore the dataset, and that is the problem PCA will tackle.

As a dimension reduction method, PCA reduces the dimensions of the problem (250 in our example), to a two dimensional space that can be represent in a $(x, y)$ plot that we can investigate on paper (and also the flat computer screen). In doing so it tries to extract the maximal amount of information from the higher dimensional space and concentrate it in two dimensions. Information used to accomplish this takes the following statistical measures:

- **Variance**: If a variable (dimensions) varies a lot in you data, it means that its dimension have discriminating properties and caries information. The opposite, no variation, would mean that the property is shared by all individuals and no groups of individuals can be singled out, therefore this dimension would not carry any information.

- **Correlation**: If variables (dimensions) vary, then they can correlate. If features are correlated, then we might expect that there is some morphological or functional relations and that feature therefore caries information.

We shall work out these ideas in the following screencasts. The fundamentals of PCA are however, based on linear algebra which is unfortunately outside the scoop of this course. Therefore we will take a more intuitive approach in the screencasts, using the concepts introduced in the previous two weeks, the aforementioned variance and correlations.

---

PCA is also covered in Elements of Biostatistics, chapter 4.

---

This online course is an excerpt of General Research Skills.

# Chapter 1

# Motivation Behind PCA

This screencast explains the basic motivation behind dimensions reduction PCA. First the explanation behind the dimension reduction of a 3D object into 2D. It also addresses the meaning of the modern classification of PCA as an unsupervised machine learning methodology by comparing it with supervised machine learning.

As an simple example the Iris data is used. This data set will we analyzed in more detail in another screencast.

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed,
```

Questions:

- What is the difference between supervised and unsupervised machine learning methodologies.

# Chapter 2

# PCA Fundamentals

PCA algorithm is solidly based on linear algebra. Without going in to detail we will use a concept and two measures from linear algebra:

- **Orthogonality**: Dimensions (variables) are independent of each other if they are Orthogonal; geometrically we would visualize that as perpendicular dimensions (axes).
- **Eigenvectors**: The directions of these **orthogonal** axes, and thus the principal components, are given by the eigenvectors.
- **Eigenvalues**: The length, or weights, of the eigenvectors are given by the eigenvalues. It turns out that the eigenvalues are directly related to the variance of the principle component.

PCA captures the most variance in the first new axes, the first principal component, by capturing most of the information (variance or eigenvalues). This is repeated for the next orthogonal principal component by capturing the most information (variance or eigenvalue) that is left after the previous step. These steps are repeated until we have the same number of principle components as dimensions.

- How we need to interpret the loading of a variables in a principal component?

# Chapter 3

# PCA in classification

In this simple example the Iris data, introduced earlier, are analyzed using PCA. Note that normally we do not have the information on which species the individuals belong to, unsupervised. The species are only used here for a didactic purpose.

Questions:

- Which of the sepal or petal widths and lengths were important in the first, and which in the second principal component?
- Which feature(s) (variable(s)) is response for the separation between Versicolor and Virginica?

# Chapter 4

# PCA and Gradients

The Heptathlon dataset give the scores, for every athlete, for the different disciplines of the woman's Olympic Heptathlon of 1988 in Seoul. The row are the individual athletes and the columns the scores per discipline. This dataset might be interesting for the coaches to investigate if one of disciplines might contribute more to a success than another. Here we do not look in particular to clusters of groups of athletes but if the distribution of the individual athlete is related to the final score when visualizing the principal components. The relation with linear regression and supervised learning is also discussed.

Question

- Explain in your own words which disciplines contributed to the success of Jackie Joyner-Kersee.