Data visualization and dimensional reduction, Principal Component Analysis (PCA)

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MED3007
Statistical Principles in Genomics: an Introduction with Rstudio
16.01.2024

1. Unsupervised Learning
    - what it means? key concepts
    - focus of today

2. Principal Component Analysis (PCA)
    - Some theoretical concepts
    - Visualization
    - One example

What is unsupervised learning?

Unsupervised methods

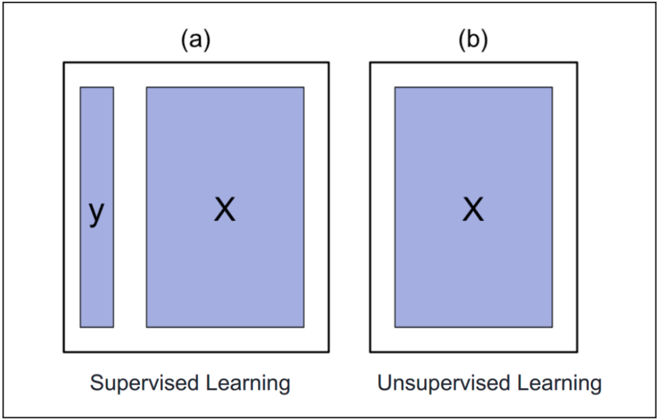
Methods that do not make use of external information: groups / class assignments, clinical outcomes, covariates.
Aim: finding hidden structure in the data

How can we find and/or visualise structure?

- Reducing dimensionality by removing noise and “uninteresting” stuff
  → TODAY – Lecture 2
- Ordering and grouping together via clustering
  → TOMORROW – Lecture 3

Supervised vs Unsupervised learning

- Supervised Learning: both **X** and **y** is known (panel (a) below)
- Unsupervised Learning: only **X** is known (panel (b) below)



Supervised vs Unsupervised learning

- **Different purpose:**
  - In supervised learning, we are interested in using **X** to predict an associated response variable **y**
  - In unsupervised learning we have no **y**, and thus this is often the first data exploration that one can perform to understand the data
- **Different evaluation / validation:**
  - In supervised learning, accuracy of results can be evaluated by comparing predictions with the true **y**
  - In unsupervised learning **y** is unknown, thus making it hard to judge/validate the results
- **Different interpretation:**
  - Supervised learning is more objective, as predicting **y** gives a clear goal
  - As there is no goal for the analysis, unsupervised learning is more subjective

Unsupervised learning

Unsupervised learning: recap

Goal is to discover hidden structure in the observed data $\mathbf{X}$.
Therefore, **visualisation tools** can be very useful

- informative ways to visualize the data
  - Principal Component Analysis → **Lecture 2**
  - t-distributed stochastic neighbor embedding (t-SNE), . . .
    → not for this course!
- discover subgroups to enhance visualization
  - Clustering (hierarchical, k-means) → **Lecture 3**
  - UMAP, Spectral Clustering, . . .
    → not for this course!

Dimension reduction methods

- also called feature extraction methods
- **Aim:** project the high-dimensional data to smaller dimensions
- **Side-product:** easier visualization
- **Assumption:** Small number of hidden factors determine most of the variability in the high-dimensional data

Examples:

- Principal component analysis (PCA):
  Identify the directions of largest variance in the data
- t-distributed stochastic neighbor embedding (t-SNE):
  Find non-linear transformations to represent original data in 2 or 3 dimensions such that similar points are nearby with high probability
- Multi-dimensional scaling (MDS):
  Classical (similar to PCA), aims to preserve the pairwise distances
- Non-negative matrix factorisation (NMF):
  Factorization method applicable when all data are non-negative

1. Unsupervised Learning
   - what it means? key concepts
   - focus of today

2. Principal Component Analysis (PCA)
   - Some theoretical concepts
   - Visualization
   - One example

Principal component analysis (PCA)

Principal Component Analysis (PCA)

PCA is a dimension reduction method that seeks linear combinations of the original variables that

- capture maximal variance
- are mutually uncorrelated

Principal component analysis (PCA)

- The first PC is a linear combination of the original variables

$$v_1 = u_{11}x_1 + u_{21}x_2 + \ldots + u_{p1}x_p \tag{1}$$

  that explains the maximum amount of variation in the data

- The second PC is the linear combination of the original variables

$$v_2 = u_{12}x_1 + u_{22}x_2 + \ldots + u_{p2}x_p$$

  that describes the maximum amount of remaining variation in direction orthogonal to the first PC

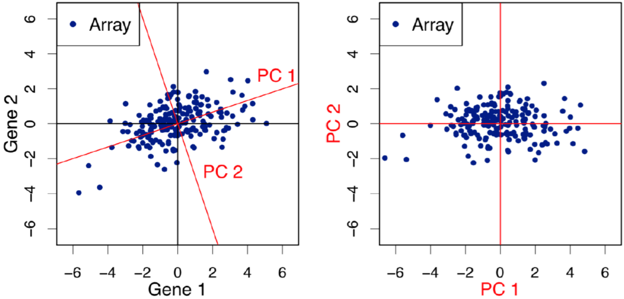
- This can be iterated: The k-th PC is the linear combination of the original variables that describes the maximum amount of remaining variation in direction orthogonal to all the first $(k-1)$ PCs

Principal component analysis (PCA)

Two elements characterize a PCA:

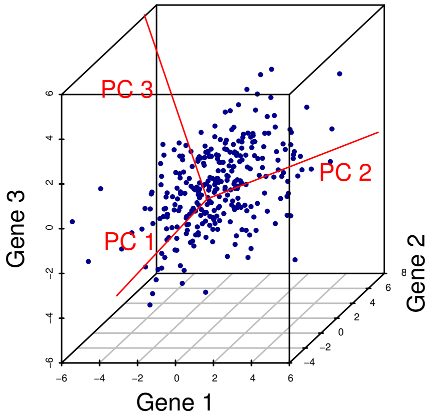
- Loadings: contributions of the original variables to a PC (the loadings of PC 1 are the numbers $u_{11}, \ldots, u_{p1}$ in eq. (1))
- Scores: coordinates of the observations onto the direction of the PC (the scores of PC 1 are the v_{i1}'s that one obtains from (1) by using the observed data $x_{i1}, \ldots, x_{ip}$ of each element of the sample i)

Principal component analysis (PCA)

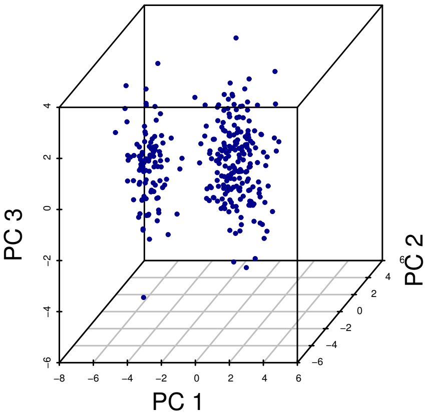


1. the loadings vector
   - defines the direction in the original *p*-dimensional space in which the data vary the most
   - can be “looked at” to interpret the principal components
2. the scores vector
   - they are obtained by projection in these new directions
   - visually, they are the coordinates of the points in the “new space” created by the components (blue points in the right panel above)

Principal component analysis - 3D visualization

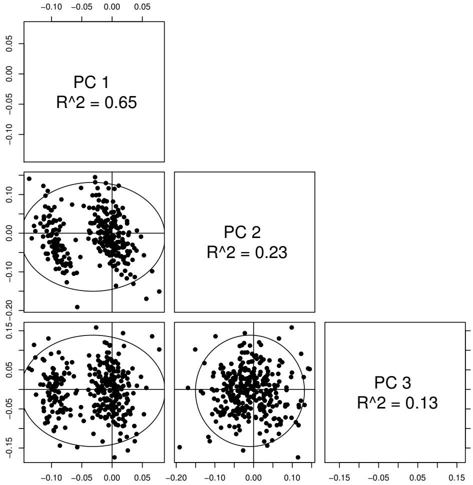


Raw data

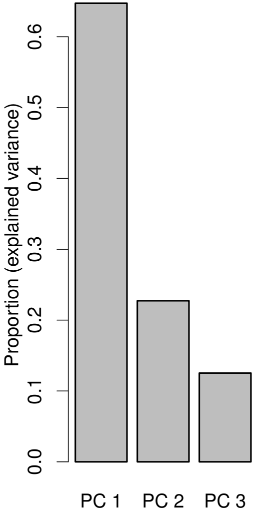


Rotation into PCA coordinate system

Principal component analysis - 2D visualization



2D scatter plots after rotation



barplot of PEV

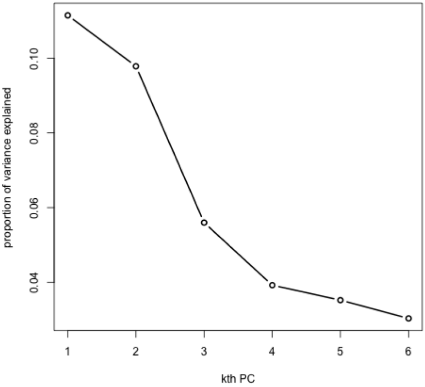
Principal component analysis - 2D visualization, scree plot

Last slide, on the right, a barplot of the proportion of variance explained by each component. This is the same as a screeplot.

In the screeplot, we plot

$$\frac{d_j^2}{\sum_{j=1}^{p} d_j^2}$$

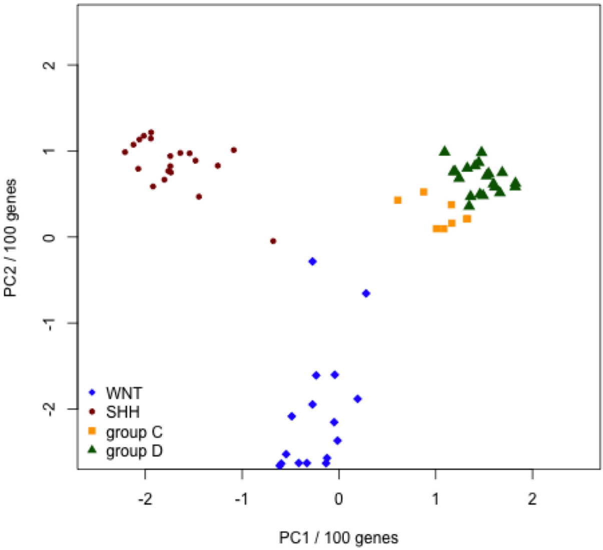
for each component j



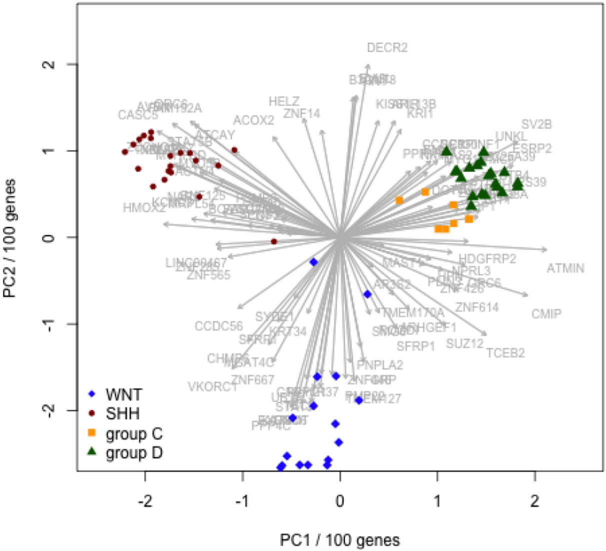
Example PCA: medulloblastoma gene expression

- High-dimensional gene expression data set comprising 18406 genes measured using the 4x44K Agilent Whole Genome Oligo-microarray
- The samples comprise 4 sub-types of medulloblastoma: 8 group C, 20 group D, 20 SHH and 16 WNT tumors
- PCA may fail in high-dimensional low sample size settings (HDLSS), i.e. can not consistently estimate the true underlying direction of maximal variance
- Typically, an unsupervised variable selection is performed, e.g. select the first 100 variables with highest SD or MAD

Example PCA: Scatterplot of the scores



Example PCA: Biplot of scores and loadings



Take-home messages

- PCA is sensitive to which scales the data are on:
  - When variables are on different scales, the data are usually scaled to have the same variance
- How to choose the number of relevant components?
  - Retain the number of PCs required to explain some percentage of the total variation (e.g. 90%), or look for an “elbow” in the scree plot
  - Compare eigenvalues with eigenvalues derived from resampled data
- When very high-dimensional data?
  The interpretation of the PCs and loadings is difficult, and PCA may fail to estimate the true underlying direction of maximal variance
  - Perform unsupervised variable selection, i.e. filtering by SD
  - Use sparse PCA methods