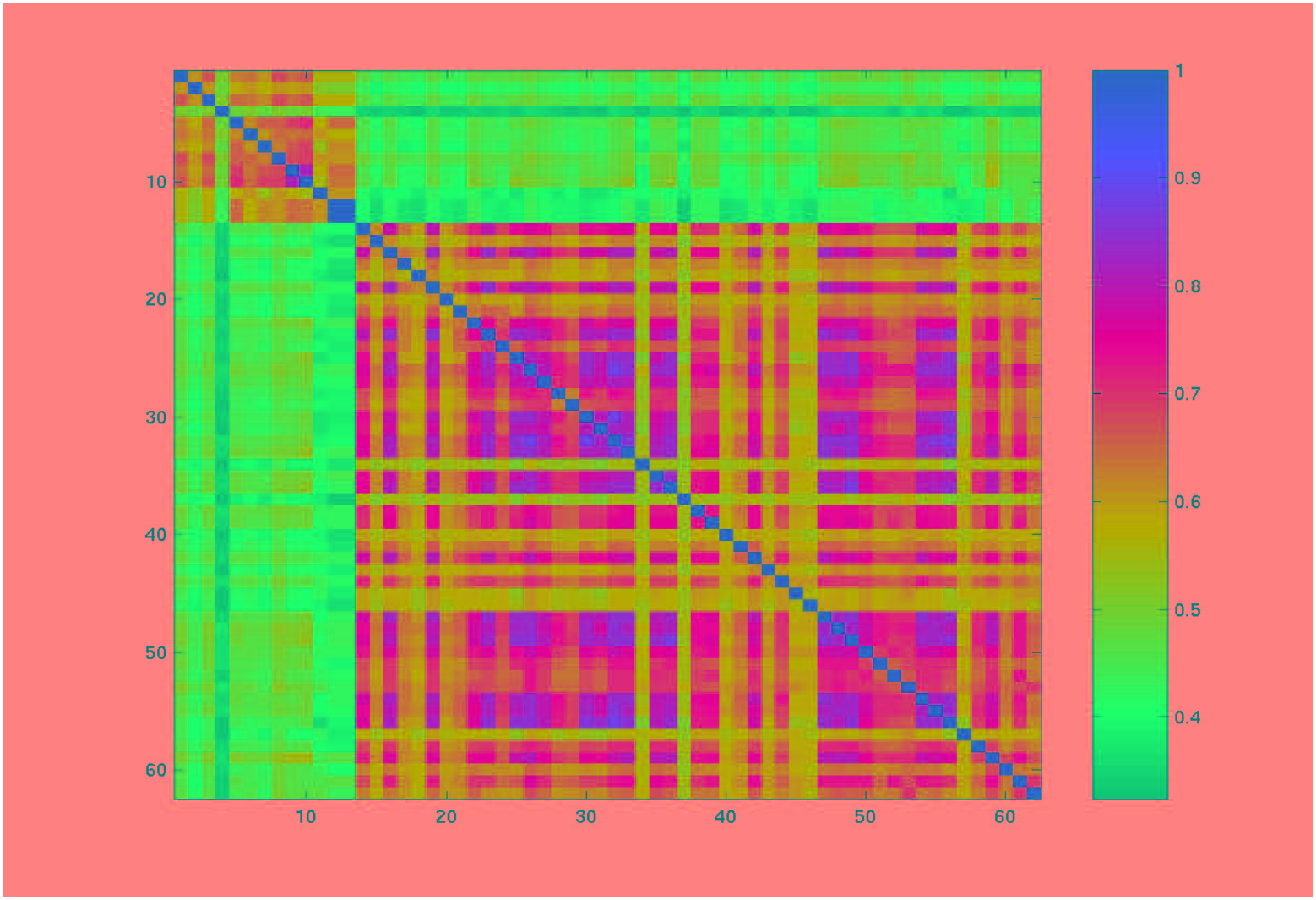
Exploratory data analysis for microarray data

Anja von Heydebreck

Max–Planck–Institute for Molecular Genetics,
Dept. Computational Molecular Biology, Berlin, Germany

heydebre@molgen.mpg.de

Visualization of similarity/distance matrices

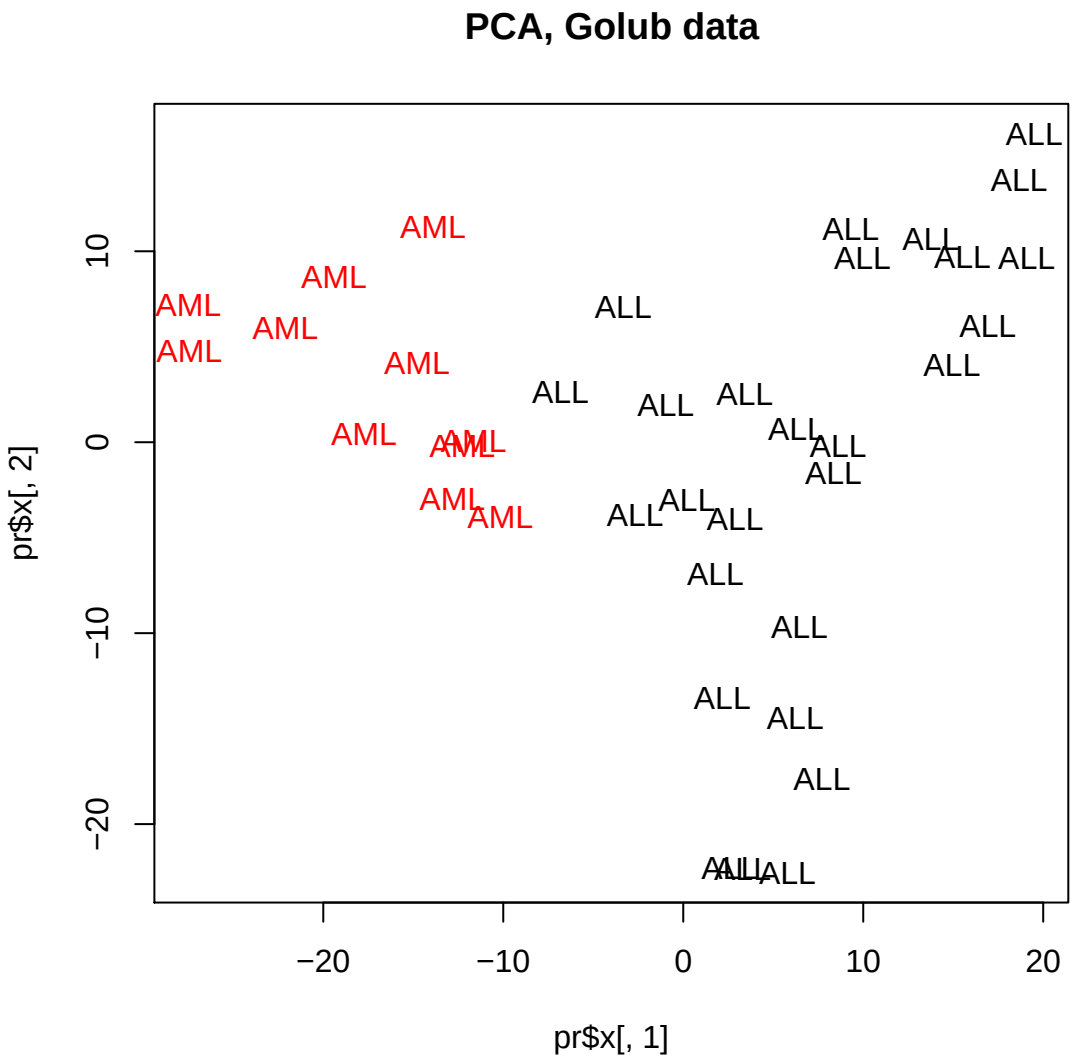


Matrix of correlation coefficients between any two hybridizations, ordered by array batch.

Projection methods

- Map the rows and/or columns of the data matrix to a plane such that similar rows/columns are located close to each other.

- Different methods (principal component analysis, multidimensional scaling, correspondence analysis) use different notions of similarity.



Principal component analysis

❍ Imagine k observations (e.g. tissue samples) as points in n-dimensional space (here: n is the number of genes).

❍ Aim: Dimension reduction while retaining as much of the variation in the data as possible.

❍ Principal component analysis identifies the direction in this space with maximal variance (of the observations projected onto it).

❍ This gives the first principal component (PC). The $i+1$st PC is the direction with maximal variance among those orthogonal to the first i PCs.

❍ The data projected onto the first PCs may then be visualized in scatterplots.

Principal component analysis

❍ PCA can be explained in terms of the eigenvalue decomposition of the covariance/correlation matrix Σ:

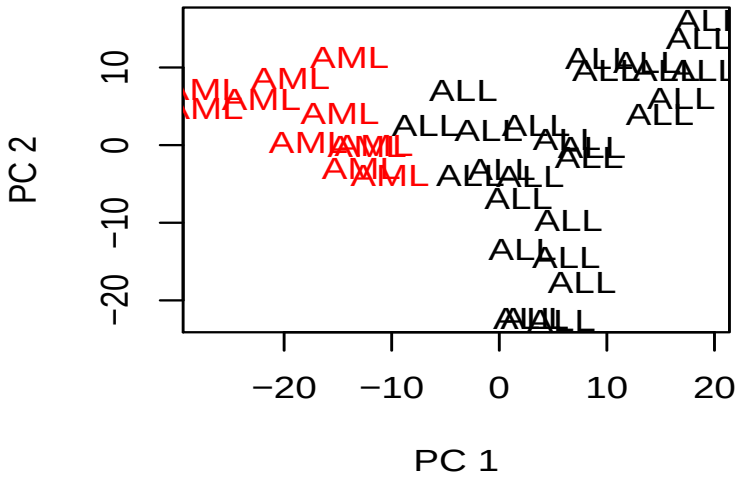
$$\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^{\boldsymbol{t}},$$

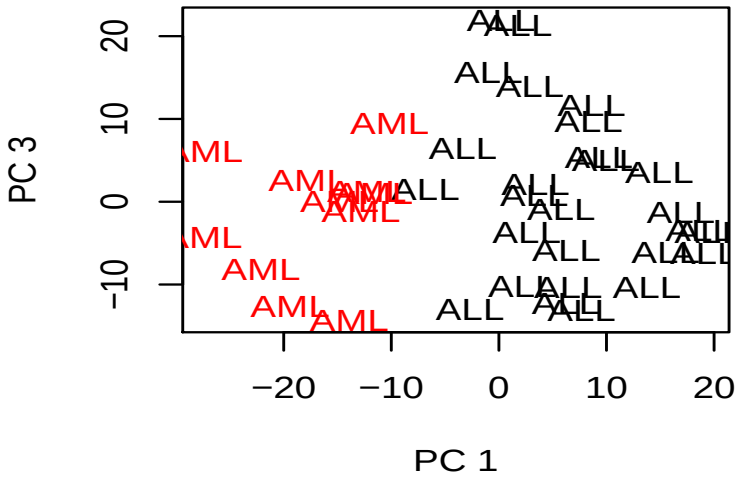
where the columns of $\boldsymbol{S}$ are the eigenvectors of Σ (the principal components), and $\boldsymbol{\Lambda}$ is the diagonal matrix with the eigenvalues (the variances of the principal components).

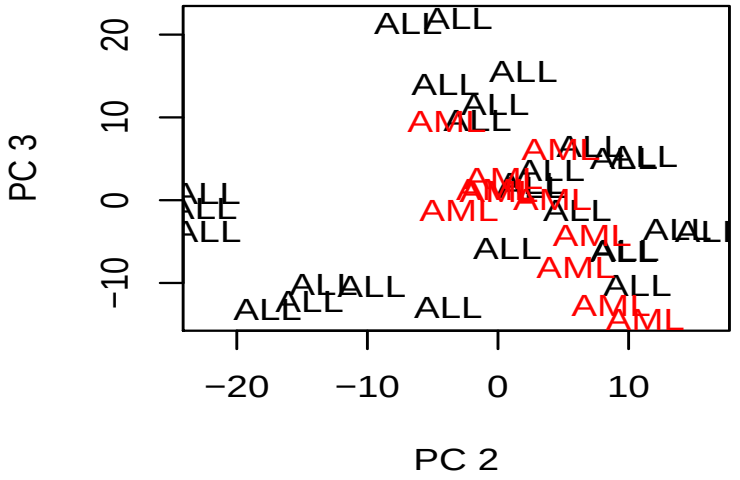
❍ Use of the correlation matrix instead of the covariance matrix amounts to standardizing variables (genes).

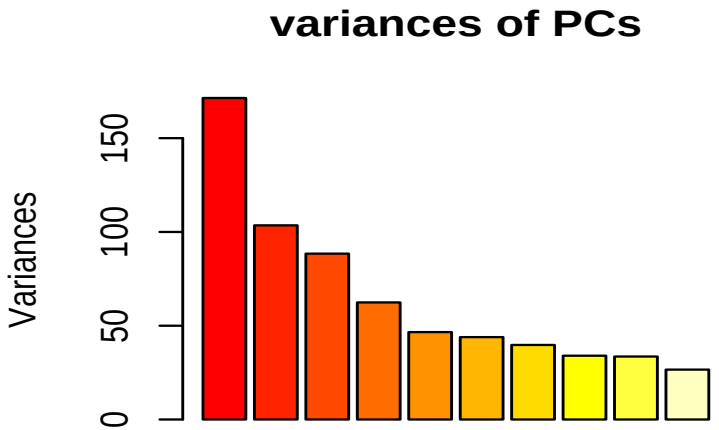
❍ R function `prcomp` in package `mva`.

PCA, Golub data









Multidimensional scaling

❍ Given a *n* x *n* dissimilarity matrix $D = (d_{ij})$ for n objects (e.g. genes or samples), multidimensional scaling (MDS) tries to find n points in Euclidean space (e.g. plane) with a similar distance structure $D' = (d'_{ij})$ - more general than PCA.

❍ The similarity between D and D' is scored by a stress function.

❍ Least-squares scaling: $S(D, D') = (\sum (d_{ij} - d'_{ij})^2)^{1/2}$. Corresponds to PCA if the distances are Euclidean.
In R: `cmdscale` in package `mva`.

❍ Sammon mapping: $S(D, D') = \sum (d_{ij} - d'_{ij})^2 / d_{ij}$. Puts more emphasis on the smaller distances being preserved.
In R: `sammon` in package `MASS`.

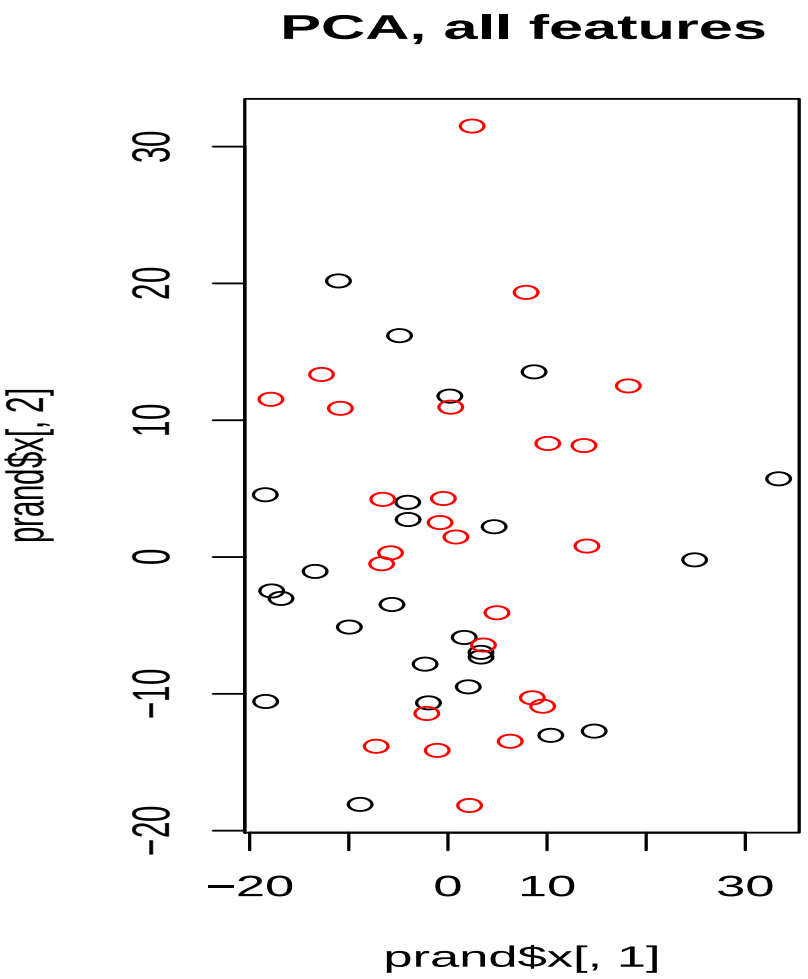
Projection methods: feature selection

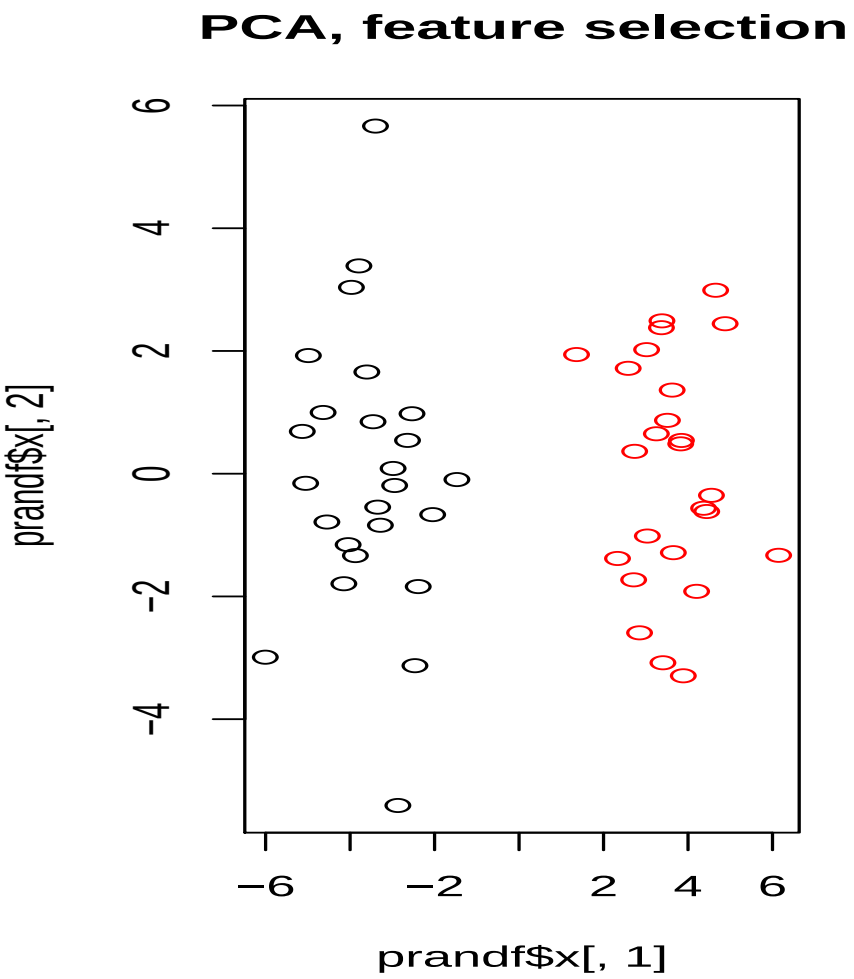
❍ The results of a projection method also depend on the features (genes) selected.

❍ If those genes are selected that discriminate best between two groups, it is no wonder if they appear separated.

❍ This may also happen if there is no real difference between the groups.

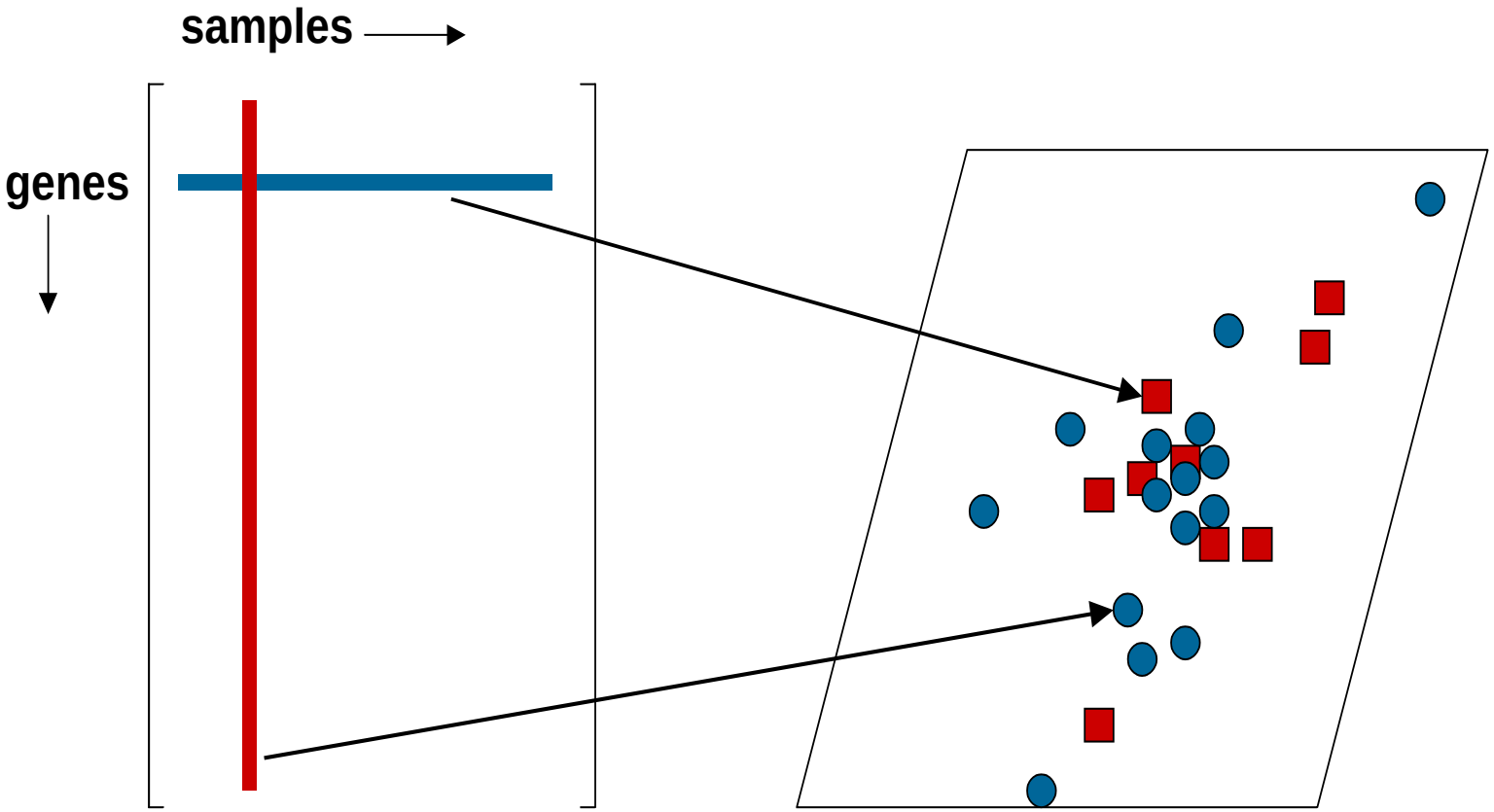
Projection methods: feature selection





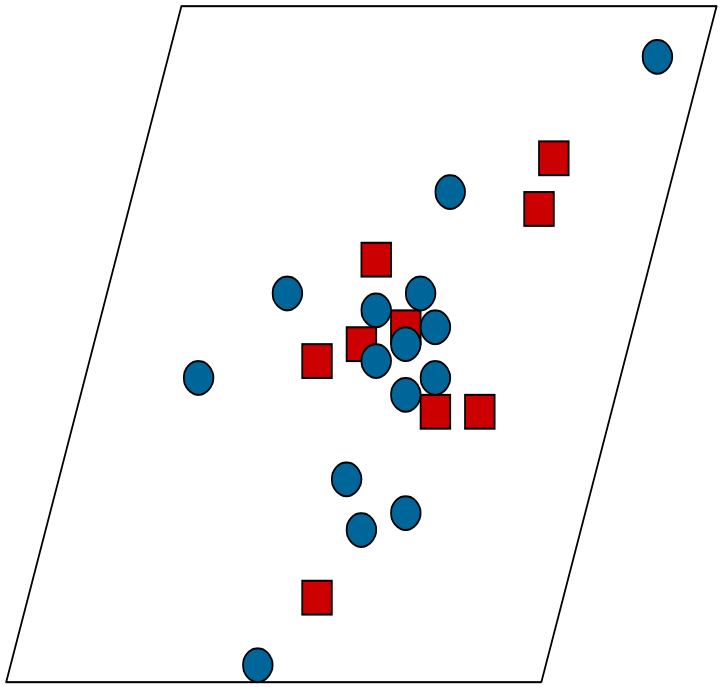
Left: PCA for a 5000 x 50 random data matrix. For the right plot, 90 “genes” with best discrimination between red and black were selected (t-statistic).

Correspondence analysis: Projection onto plane



Correspondence analysis: Properties of projection

- Similar row/column profiles (small χ^2-distance) are projected close to each other.
- A gene with positive/negative association with a sample will lie in the same/opposite direction from the centroid.



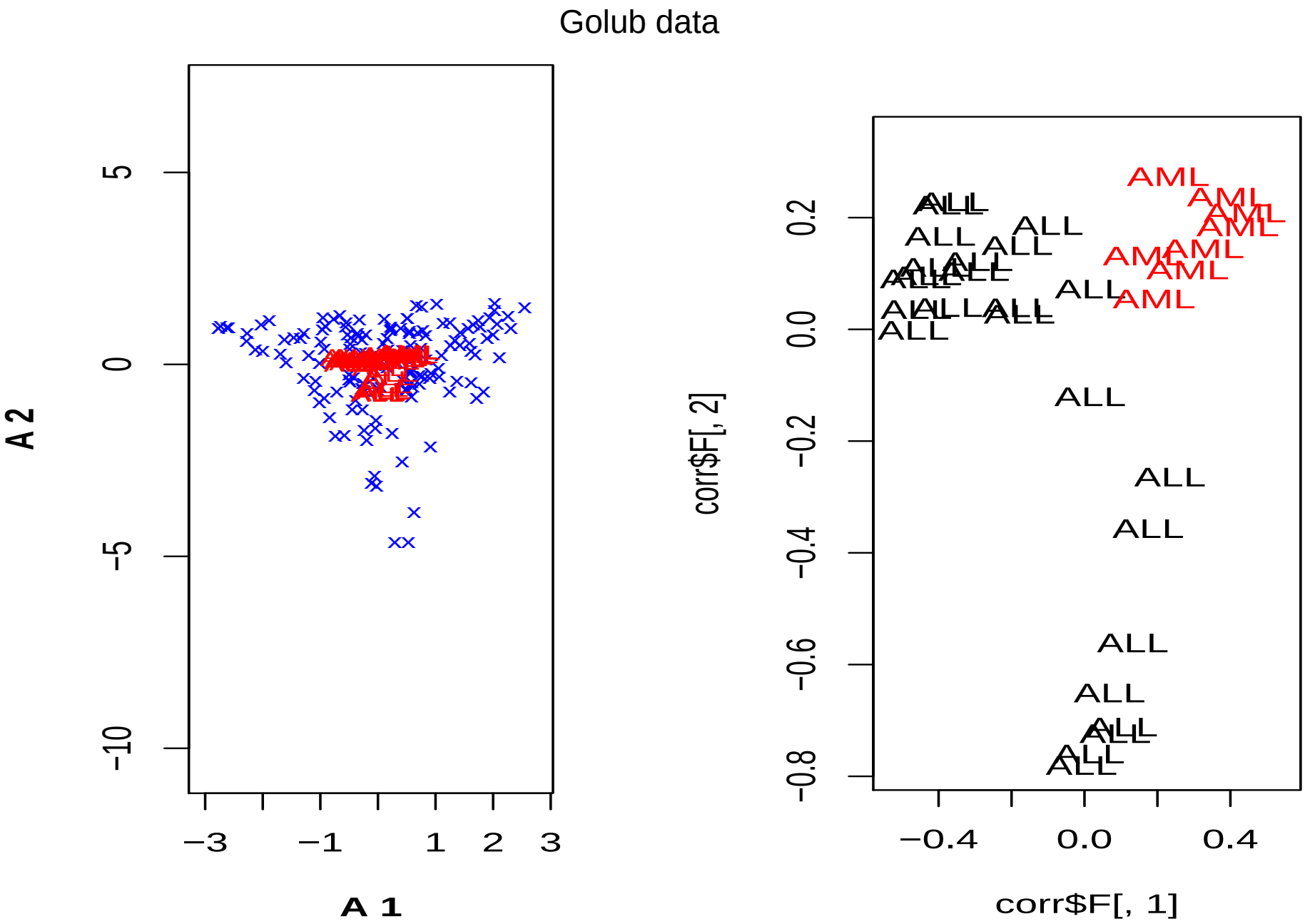
Projection methods: Correspondence analysis

❍ Correspondence analysis is usually applied to tables of frequencies (contingency tables) in order to show associations between particular rows and columns – in the sense of deviations from homogeneity, as measured by the χ^2-statistic.

❍ Data matrix is supposed to contain only positive numbers - may apply global shifting.

❍ R packages `CoCoAn, multiv`.

Correspondence analysis - Example



ISIS - a class discovery method

❍ Aim: detect subtle class distinctions among a set of tissue samples/gene expression profiles (application: search for disease subtypes)

❍ Idea: Such class distinctions may be characterized by differential expression of just a small set of genes, not by global similarity of the gene expression profiles.

❍ The method quantifies this notion and conducts a search for interesting class distinctions in this sense.

❍ R package `ISIS` available at
`http://www.molgen.mpg.de/~heydebre`

References

❍ Duda, Hart and Stork (2000). Pattern Classification. 2nd Edition. Wiley.

❍ Dudoit and Fridlyand (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, Vol. 3(7), research 0036.1-0036.21.

❍ Eisen et al. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, Vol 95, 14863–14868.

❍ Fellenberg et al. (2001): Correspondence analysis applied to microarray data. *PNAS*, Vol. 98, p. 10781–10786.

❍ v. Heydebreck et al. (2001). Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, Suppl. 1, S107–114.

❍ Kerr and Churchill (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *PNAS*, Vol. 98, p. 8961-8965.

❍ Pollard and van der Laan (2002). Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences*, Vol. 176, 99-121.