Distances and expression measures

Sandrine Dudoit and Robert Gentleman

Bioconductor Short Course

Winter 2002

Gene expression data

Gene expression data on G genes (features) for n mRNA samples (observations)

$$X_{G\times n} = \begin{matrix} \text{mRNA samples} \\ \end{matrix} \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{G1} & x_{G2} & \dots & x_{Gn} \end{bmatrix} \begin{matrix} \\ \\ \\ \text{Genes} \end{matrix}$$

x_{gi} = expression measure for gene g in mRNA sample i.

An array of conormalized arrays.

Role of distances

Microarray data analysis often involves

- clustering genes or samples;
- classifying genes or samples.

Inherent in every machine learning approach, unsupervised and supervised, is a notion of *distance* or *similarity* between the objects to be clustered or classified.

Role of distances

- Many clustering procedures operate directly on a matrix of pairwise distances between the objects to be clustered: e.g., partitioning around medoid (PAM) and hierarchical clustering methods.
- In supervised learning, new observations are typically assigned to classes on the basis of their distances from objects with known class labels.
  - *k-nearest neighbor*: based on an explicit choice of distance function.
  - *Linear discriminant analysis*: based on the Mahalanobis distance of observations from class means.
  - *Support vector machines*: based on the Euclidean distance between individual observations and a separating hyperplane (margin).

Role of distances

The choice of distance is important and can have a large impact on the results of supervised and unsupervised learning analyses.

In some cases, the Euclidean metric will be sensible, while in others, a distance based on correlations will be a better choice.

Subject matter knowledge is very helpful in selecting an appropriate distance for a given project.

Outline

- Basic notions
- Distances and standardization
- Absolute versus relative expression measures
- Experiment specific distances between genes
- Multidimensional scaling

Definitions: metrics and distances

A **metric**, d, is a function that satisfies the following five properties

(i) non-negativity $d(\mathbf{x}, \mathbf{y}) \geq 0$;

(ii) symmetry $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$;

(iii) identification mark $d(\mathbf{x}, \mathbf{x}) = 0$;

(iv) definiteness $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$;

(v) triangle inequality $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

We can also consider **distance functions** that satisfy the first three properties only.

We will refer to *distances* which include *metrics* and only mention metrics when the behavior of interest is specific to them.

Definitions: similarity and dissimilarity functions

A **similarity function**, s, is more loosely defined and satisfies the three following properties

(i) non-negativity $s(\mathbf{x}, \mathbf{y}) \geq 0$;

(ii) symmetry $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$;

(iii) The more *similar* the objects $\mathbf{x}$ and $\mathbf{y}$, the greater $s(\mathbf{x}, \mathbf{y})$.

A **dissimilarity function**, d, satisfies (i) and (ii), but for (iii), $d(\mathbf{x}, \mathbf{y})$ is larger the more dissimilar the objects.

Distances

There is a great deal of choice (and hence literature) on selecting a distance function.

Some books that pay particular attention to distances in the context of classification and clustering include

- Section 4.7 of Duda, Hart, & Stork (2000);
- Chapter 2 of Gordon (1999);
- Chapter 1 of Kaufman and Rousseeuw (1990);
- Chapter 13 of Mardia, Kent, & Bibby (1979).

When some variables are continuous and others categorical, there are more choices and the implications of the different choices should be weighed carefully.

Examples of distances

- Euclidean metric (possibly standardized);
- Mahalanobis metric;
- Manhattan metric;
- Minkowski metric (special cases are Euclidean and Manhattan metrics);
- Canberra metric;
- One minus correlation;
- etc.

Examples of distances

Name	Formula
Euclidean metric	$d_E(\mathbf{x}_i, \mathbf{x}_j) = \{\sum_g w_g(x_{gi} - x_{gj})^2\}^{1/2}$
Unstandardized	$w_g = 1$
Standardized by s.d.	$w_g = 1/s_{g.}^2$
(Karl Pearson distance)	
Standardized by range	$w_g = 1/R_{g.}^2$
Mahalanobis metric	$d_{Ml}(\mathbf{x}_i, \mathbf{x}_j) = \{(\mathbf{x_i} - \mathbf{x_j})S^{-1}(\mathbf{x_i} - \mathbf{x_j})'\}^{1/2}$
	$= \{\sum_g \sum_{g'} s^{-1}_{gg'}(x_{gi} - x_{gj})(x_{g'i} - x_{g'j})\}^{1/2}$
	where $S = (s_{gg'})$ is any $G \times G$ positive definite matrix, usually the sample covariance matrix of the variables. When the matrix is the identity, this reduces to the unstandardized Euclidean distance.
Manhattan metric	$d_{Mn}(\mathbf{x}_i, \mathbf{x}_j) = \sum_g w_g \lvert x_{gi} - x_{gj}\rvert$
Minkowski metric	$d_{Mk}(\mathbf{x}_i, \mathbf{x}_j) = \{\sum_g w_g \lvert x_{gi} - x_{gj}\rvert^{\lambda}\}^{1/\lambda}$, $\lambda \geq 1$.
	$\lambda = 1$: Manhattan distance
	$\lambda = 2$: Euclidean distance
Canberra metric	$d_C(\mathbf{x}_i, \mathbf{x}_j) = \sum_g \frac{\lvert x_{gi} - x_{gj}\rvert}{\lvert x_{gi} + x_{gj}\rvert}$
One minus Pearson correlation	$d_{corr}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\sum_g (x_{gi} - \bar{x}_{.i})(x_{gj} - \bar{x}_{.j})}{\{\sum_g (x_{gi} - \bar{x}_{.i})^2\}^{1/2}\{\sum_g (x_{gj} - \bar{x}_{.j})^2\}^{1/2}}$

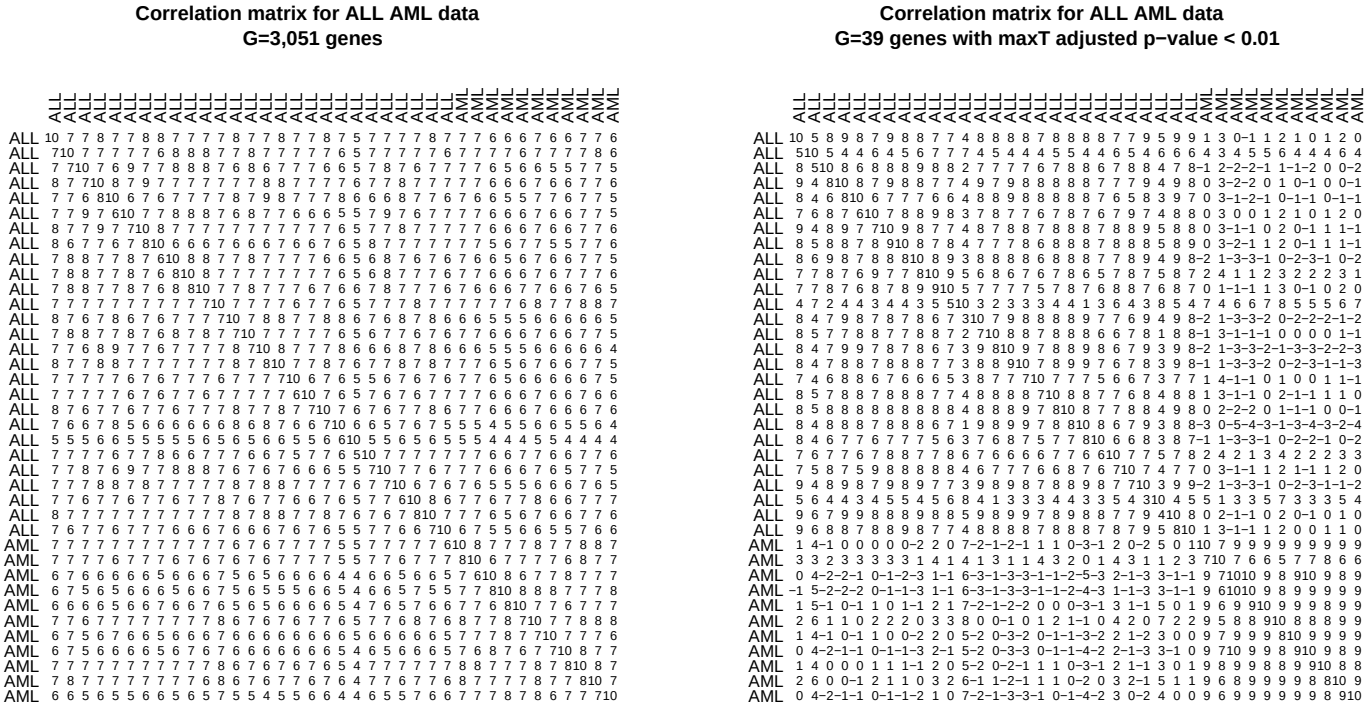
The formulae refer to distances between observations (arrays).

R distance functions

R has a number of functions for computing and displaying distance and similarity matrices.

- Distance functions
  - `dist` (`mva`): Euclidean, Manhattan, Canberra, binary;
  - `daisy` (`cluster`): Euclidean, Manhattan.
- Correlation functions
  - `cor`, `cor.wt`.
- Plotting functions
  - `image`;
  - `plotcorr` (`ellipse`);
  - `levelplot` (`lattice`);
  - `plot.cor`, `plot.mat` (`sma`).

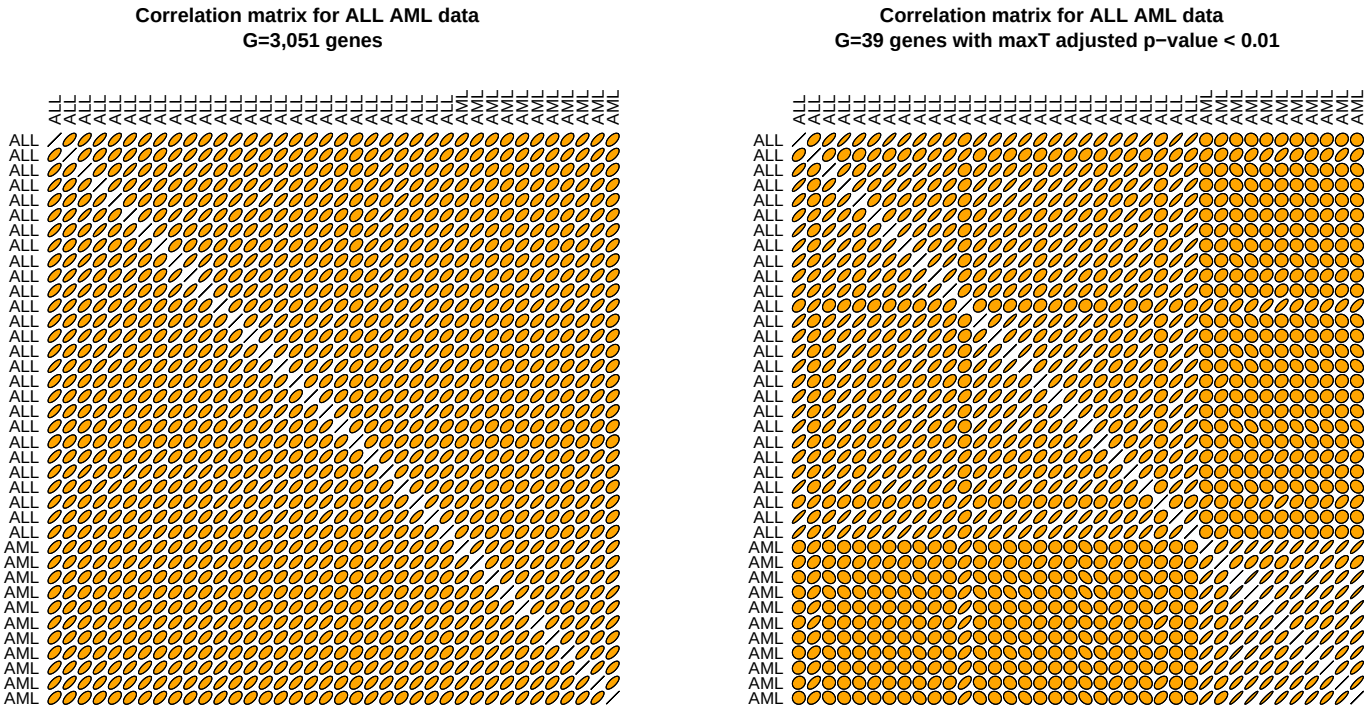
Correlation matrices: `plotcorr`



(a) (b)

Figure 1: Correlation matrix for Golub et al. (1999) ALL AML data: (a) G=3,051 genes and (b) G=39 genes with maxT adjusted p-value < 0.01.

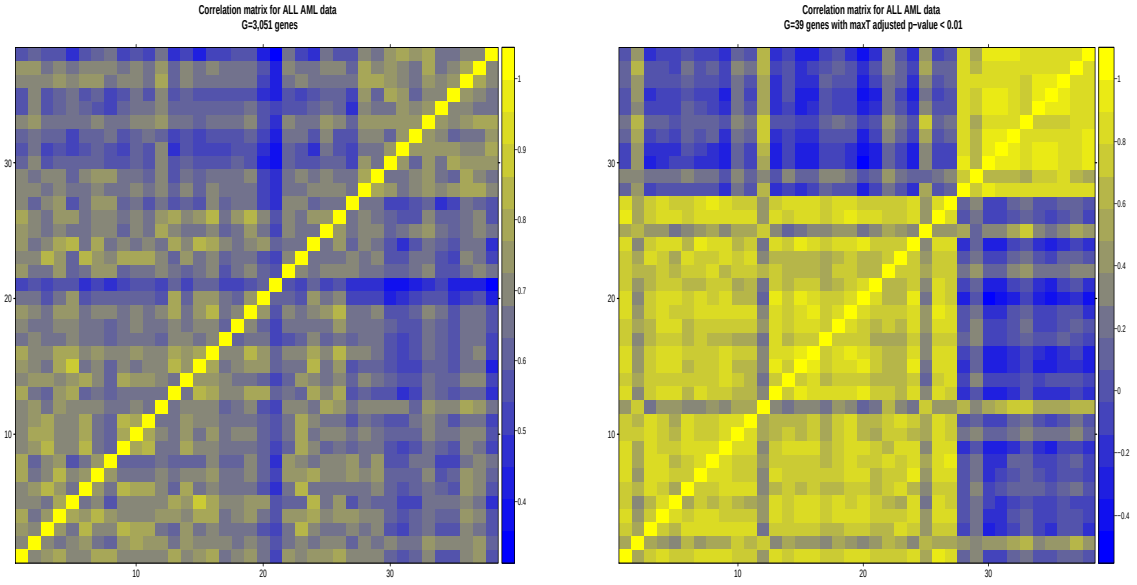
Correlation matrices: `plotcorr`



(a) (b)

Figure 2: Correlation matrix for Golub et al. (1999) ALL AML data: (a) G=3,051 genes and (b) G=39 genes with maxT adjusted p-value < 0.01.

Correlation matrices: `levelplot`



(a) (b)

Figure 3: Correlation matrix for Golub et al. (1999) ALL AML data: (a) G=3,051 genes and (b) G=39 genes with maxT adjusted p-value < 0.01.

Distances between clusters

For many clustering algorithms, *distances between clusters* will be necessary. There are a number of different ways of defining a distance between groups, or between one observation and a group of observations.

Single linkage The distance between two clusters is the *minimum* distance between any two objects, one from each cluster.

Average linkage The distance between two clusters is the *average* of all pairwise distances between the members of both clusters.

Complete linkage The distance between two clusters is the *maximum* distance between two objects, one from each cluster.

Centroid distance The distance between two clusters is the distance between their *centroids*. The definition of centroid may depend on the clustering algorithm being used.

Distances between clusters

The choice of distance measure between clusters has a large effect on the shape of the resulting clusters.

For instance, single linkage leads to long thin clusters, while average linkage leads to round clusters.

Standardization

- Standardization of variables is an important issue when considering distances between objects.
- The distance function and its behavior are intimately related to the *scale* on which measurements are made.
- There are no objective methods for dealing with this problem. The solution is generally problem specific.

Standardization

For microarray data both genes and/or arrays can be standardized. Which of the two should be carried out is dependent upon whether samples or genes are being clustered or classified.

Standardizing genes

$$x_{gi} \leftarrow (x_{gi} - center_g)/scale_g,$$

where $center_g$ is a measure of the center of the distribution of the set of values $\{x_{gi} : i = 1, \ldots, n\}$, such as the mean or median, and $scale_g$ is a measure of scale, such as the standard deviation, IQR, or MAD.

Standardizing arrays

$$x_{gi} \leftarrow (x_{gi} - center_i)/scale_i.$$

Standardizing genes

- Gene standardization in some sense puts all genes on an equal footing and weighs them equally in the classification or clustering. Common standardization procedures are
- $x_{gi} \leftarrow \frac{x_{gi} - \bar{x}_{g.}}{s_{g.}}$,
  where $\bar{x}_{g.}$ and $s_{g.}$ denote respectively the average and standard deviation of gene g's expression levels across the n arrays.
- $x_{gi} \leftarrow \frac{x_{gi} - m_{g.}}{mad_{g.}}$,
  where $m_{g.}$ and $mad_{g.}$ denote respectively the median and median absolute deviation (MAD) of gene g's expression levels across the n arrays. These are robust estimates of location and scale.
- $x_{gi} \leftarrow \frac{x_{gi} - x_{g(1)}}{x_{g(n)} - x_{g(1)}}$,
  where $x_{g(j)}$ denote the ordered expression levels for gene g,
  $x_{g(1)} \leq x_{g(2)} \leq \ldots \leq x_{g(n)}$.

Standardizing arrays

Standardization of arrays can be viewed as part of the **normalization** step.

It is consistent with the common practice of using the correlation between the gene expression profiles of two mRNA samples to measure their similarity.

In practice, we recommend more general adaptive and robust normalization methods which correct for intensity, spatial, and other types of bias using robust local regression (see lecture on pre-processing).

Table 1: *Impact of standardization of observations and variables on the distance function.*

Distance between observations	Standardize variables (genes)	Standardize observations (arrays)
Euclidean, $w_g = 1$	Changed	Changed
Euclidean, $w_g = 1/s_{g.}^2$	Unchanged	Changed
Mahalanobis	Changed, unless S diagonal	Changed
One minus Pearson correlation	Changed	Unchanged

Standardization

Note the relationship between the Euclidean distance $d_E(\cdot, \cdot)$ between standardized vectors and the distance defined as one minus the Pearson correlation:

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{2m(1 - r_{xy})},$$

where r_{xy} denotes the Pearson correlation between the m-vectors $\mathbf{x}$ and $\mathbf{y}$.

Affymetrix versus spotted arrays

A main difference between these two technologies is that Affymetrix arrays are typically used to measure the *overall* abundance of a probe sequence in a target sample, while spotted arrays typically measure the *relative* abundance of a probe sequence in two target samples (one of the two samples is often a reference sample used in multiple experiments).

The expression measures for Affymetrix arrays are typically *absolute* (log) intensities, while they are (log) *ratios* of intensities for spotted arrays.

Affymetrix versus spotted arrays

There is a belief that the expression measures of different genes can be compared directly for spotted arrays but not for Affymetrix arrays.

The distinction is somewhat artificial, since one could always take ratios of expression measures from an Affymetrix experiment with some reference sample and hence have data that are the equivalent of spotted array data.

Whether there is any real difference between the use of absolute and relative expression measures depends on the distance that is being considered.

Absolute versus relative expression measures

Consider the standard situation where we have x_{gi} represent the *absolute* log expression measure for gene g on patient sample/array i.

Let $y_{gi} = x_{gi} - x_{gA}$, where array A is our reference sample. Then the *relative* expression measures y_{gi} represent the standard data from a spotted array experiment with a common reference sample.

Use of relative expression measures amounts to a location transformation for each gene, cf. *gene centering*.

Absolute versus relative expression measures

For m-vectors $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_m)$, consider distance functions of the form

$$d(\mathbf{x}, \mathbf{y}) = F\big(d_1(x_1, y_1), \ldots, d_m(x_m, y_m)\big),$$

where d_k are themselves distance functions.

E.g. the Minkowski metric : $d_k(x_k, y_k) = |x_k - y_k|$ and $F(z_1, \ldots, z_m) = (\sum_{k=1}^{m} z_k^{\lambda})^{1/\lambda}$.

The representation is quite general. There is, in particular, no need for the d_k to all be the same.

Absolute versus relative expression measures

First, suppose that we want to measure the **distance between samples** i and j. Then

$$\begin{aligned} d(\mathbf{y}_{.i}, \mathbf{y}_{.j}) &= F\big(d_1(y_{1i}, y_{1j}), \ldots, d_G(y_{Gi}, y_{Gj})\big) \\ &= F\big(d_1(x_{1i} - x_{1A}, x_{1j} - x_{1A}), \ldots, d_G(x_{Gi} - x_{GA}, x_{Gj} - x_{GA})\big). \end{aligned}$$

If all of the $d_k(a_k, b_k)$ are simply functions of $a_k - b_k$, then $d(\mathbf{y}_{.i}, \mathbf{y}_{.j}) = d(\mathbf{x}_{.i}, \mathbf{x}_{.j})$ and it does not matter whether we look at relative (the $\mathbf{y}$'s) or absolute (the $\mathbf{x}$'s) expression measures.

Examples include the Minkowski metric.

Absolute versus relative expression measures

Suppose now that we are interested in the **distance between genes** and not samples. If

$$d(\mathbf{y}_{g.}, \mathbf{y}_{j.} + a) = d(\mathbf{y}_{g.}, \mathbf{y}_{j.})$$

for any vectors $\mathbf{y}_{g.}$ and $\mathbf{y}_{j.}$ and for any scalar a, then the distance will be the same for both absolute expression measures and relative expression measures.

One minus the Pearson correlation is a distance with this property.

Absolute versus relative expression measures

Thus, for the Minkowski metric (e.g., Euclidean), the distance between samples is the same for relative (spotted array) and absolute (Affymetrix) expression measures. This does not hold for the distance between genes.

For the one minus Pearson correlation distance, the distance between genes is the same for relative (spotted array) and absolute (Affymetrix) expression measures. This does not hold for the distance between samples.

Absolute versus relative expression measures

	Distance between samples	Distance between genes
Minkowski	Unchanged	Changed
One minus correlation	Changed	Unchanged

Changed (unchanged) means that absolute and relative expression measures yield different (the same) results.

Absolute versus relative expression measures

One can argue in favor of both of these properties, i.e., invariance of (i) gene distances or (ii) sample distances, for absolute and relative expression measures.

In general, the correct way in which to analyze the data will depend on the biological question of interest and the relative merits of the two types of expression measures.

Distances and expression measures

Distances may need to be extended in various ways to deal with different types of problems.

Weights may be incorporated in any of the distances above to deal with different types of variables. For example, mixing patient level covariates with gene expression measures may be best dealt with by weighting.

In other cases, one might want to consider mixed versions of the distances. Again, if mixing patient level covariates (e.g., categorical variables) together with gene expression measures, then the Euclidean distance might be appropriate for the gene expression data, but not for the patient covariates.

Experiment specific distances between genes

The gene distance functions considered thus far do not take into account the *structure* or *design* of the microarray experiment, i.e., they treat the columns of the genes-by-arrays matrix of expression measures interchangeably.

However, microarray experiments can be highly-structured, e.g., as in timecourse and multifactorial experiments.

Below are general approaches to *supervise* the distances so that they reflect the design of the experiment under consideration.

Experiment specific distances between genes

One can exploit covariate information (e.g., treatment, cell type, dose, time) to derive suitable transformations of the genes-by-arrays data matrix, e.g., using linear modeling.

Instead of computing distances directly on the genes-by-arrays data matrix, distances can then be computed on the new genes-by-estimated-effects matrix.

Experiment specific distances between genes

In timecourse experiments, it makes sense to consider distances that are not time exchangeable and use the time index in an essential way.

For a large enough number of timepoints, one may

- penalize for non-smoothness as in Sobolev metrics;
- use one of the standard wavelet decompositions to transform expression profiles into potentially interpretable quantities corresponding to local frequency components.

Distances can then be computed for the new profiles and genes clustered based on these distances.

Experiment specific distances between genes

The transformed expression profiles can be matched to a library of profiles of interest for the particular experiment.

For instance, in factorial experiments across time, interesting reference profiles for main effects and interactions might include: cyclical, early, or late effects, or the effects over time for a known gene.

One may also compare genes based on biological metadata, e.g., co-citation in PubMed abstracts.

Multidimensional scaling

Given any $n \times n$ distance matrix D, *multidimensional scaling* (MDS) is concerned with identifying n points in Euclidean space with a *similar* distance structure D'.

The purpose is to provide a lower dimensional representation of the distances which conveys information on the relationships among the n objects, such as the existence of clusters or one-dimensional structure in the data (e.g., seriation).

Multidimensional scaling

There are different approaches for reducing dimensionality, depending on how we define similarity between the old and new distance matrices for the n objects, i.e., depending on the objective or stress function S that we seek to minimize.

- Least-squares scaling: $S(D, D') = (\sum_{i,j}(d_{ij} - d'_{ij})^2)^{1/2}$.
- Sammon mapping: $S(D, D') = \sum_{i,j}(d_{ij} - d'_{ij})^2/d_{ij}$.
  Places more emphasis on smaller distances.
- Shepard-Kruskal non-metric scaling: based on ranks, i.e., the order of the distances is more important than their actual values.

MDS and PCA

When the distance matrix D is the Euclidean distance matrix between the rows of a $n \times m$ matrix X, there is a duality between principal component analysis (PCA) and MDS.

The k-dimensional classical solution to the MDS problem is given by the centered scores of the n objects on the first k principal components.

The classical solution of MDS in k-dimensional space minimizes the sum of squared differences between the entries of the new and old distance matrices, i.e., is optimal for least-squares scaling.

Multidimensional scaling

As with PCA, the quality of the representation will depend on the magnitude of the first k eigenvalues.

The data analyst should choose a value for k that is small enough for ease representation but also corresponds to a substantial "proportion of the distance matrix explained".

Multidimensional scaling

N.B. The MDS solution reflects not only the choice of a distance function, but also the features selected.

If features were selected to separate the data into two groups (e.g., on the basis of two-sample t-statistics), it should come as no surprise that an MDS plot has two groups!

R MDS software

- `cmdscale`: Classical solution to MDS, in package `mva`.
- `sammon`: Sammon mapping, in package `MASS`.
- `isoMDS`: Kruskal's non-metric MDS, in package `MASS`.

Multidimensional scaling

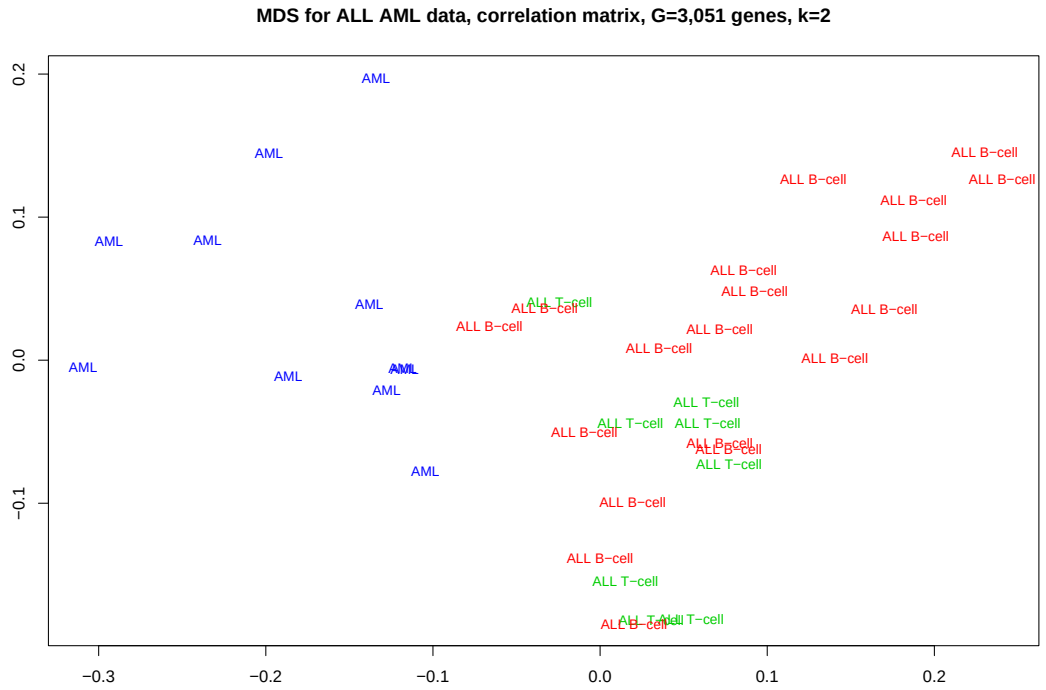


Figure 4: Classical MDS: Golub et al. (1999) ALL AML data, correlation matrix, $G = 3,051$ genes, $k = 2$, $\frac{|\lambda_1|+|\lambda_2|}{\sum_l |\lambda_l|} = 0.43$.

Multidimensional scaling

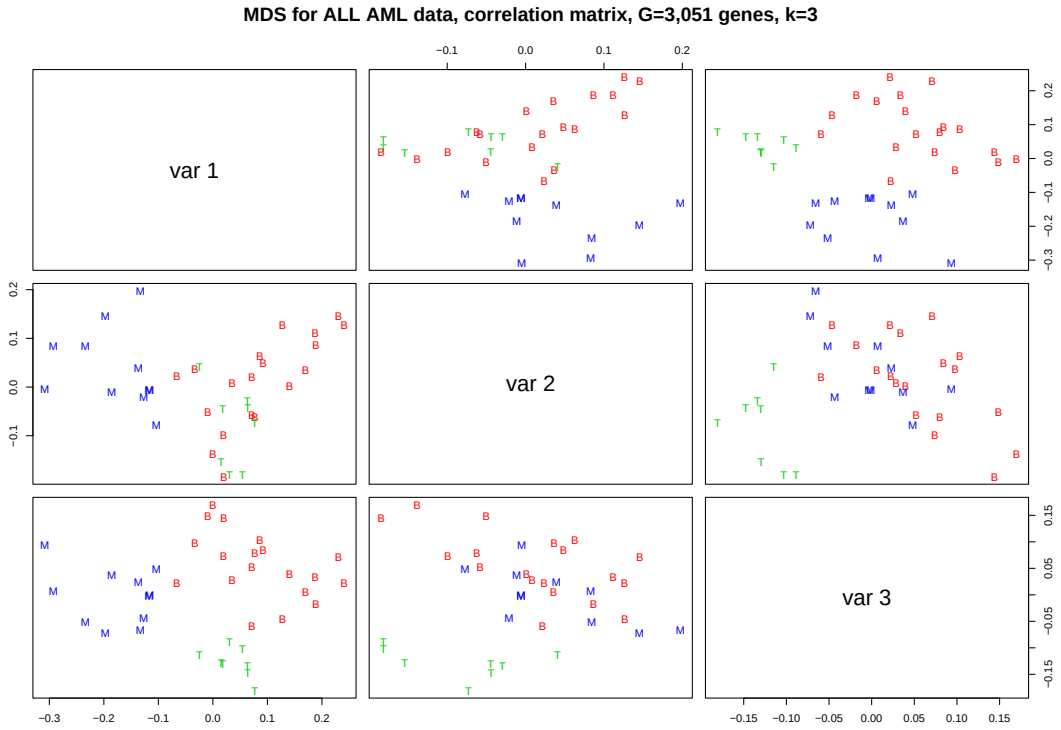


Figure 5: Classical MDS: Golub et al. (1999) ALL AML data, correlation matrix, $G = 3,051$ genes, $k = 3$, $\frac{|\lambda_1|+|\lambda_2|}{\sum_l |\lambda_l|} = 0.55$.