Distances and expression measures

Sandrine Dudoit and Robert Gentleman

Bioconductor short course

Summer 2002

Outline

- Distances
- Standardization
- Absolute versus relative expression measures

The importance of distance

Any clustering or classification of samples and/or genes involves combining or identifying objects that are *close* or *similar* to each other.

Distances or similarities are mathematical representations of what we mean by close or similar.

The choice of distance is extremely important and should not be taken lightly. In some cases, a Euclidean metric will be sensible while in others a Manhattan metric will be a better choice.

Generally, some experience or subject matter knowledge is very helpful in selecting an appropriate distance for a given project.

Metrics and distances

A **metric**, d, satisfies the following five properties

(i) non–negativity $d(a, b) \geq 0$;

(ii) symmetry $d(a, b) = d(b, a)$;

(iii) identification mark $d(a, a) = 0$;

(iv) definiteness $d(a, b) = 0$ if and only if $a = b$;

(v) triangle inequality $d(a, b) + d(b, c) \geq d(a, c)$.

We can also consider pairwise **distances**, which are functions that are required to 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.

Similarity functions

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

(i) non–negativity $S(a,b) \geq 0$;

(ii) symmetry $S(a,b) = S(b,a)$;

(iii) The more *similar* the objects a and b, the greater $S(a,b)$.

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.

Distances between clusters

For many clustering algorithms, distances between groups (clusters) of observations 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.

Gene expression data

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

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

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

An array of conormalized arrays.

Table 1: *Metrics and 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	$d_C(\mathbf{x}_i, \mathbf{x}_j) = \sum_g \frac{\lvert x_{gi} - x_{gj} \rvert}{(x_{gi} + x_{gj})}$
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).

Distances

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 values 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 level data.

Weighted distances may also be used for the purpose of feature selection in classification (see lecture on classification).

Standardization

- Standardization of the features is an important issue when considering distances between objects.
- Samples or genes are assigned to classes on the basis of their *distance* from other objects.
- The distance or similarity function that is used generally has a large effect on the performance of the classification or clustering procedure.
- 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

A common type of data transformation for continuous measurements is **standardization**.

For microarray data both genes and/or observations (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

$$x_{gi} \leftarrow (x_{gi} - \bar{x}_{g.})/s_{g.},$$

so that each gene has mean zero and unit variance across arrays.

Standardizing arrays

$$x_{gi} \leftarrow (x_{gi} - \bar{x}_{.i})/s_{.i},$$

so that each array has mean zero and unit variance across genes.

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

Distance between observations	Standardize variables	Standardize observations
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 cDNA 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 cDNA 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 cDNA arrays.

Affymetrix versus cDNA arrays

Hence, there is a belief that the expression measures of different genes can be compared directly for cDNA 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 cDNA 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 patient A is our reference sample. Then the **relative** expression measures y_{gi} represent the standard data from a cDNA 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 patient 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 if 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 measurements and relative expression measurements.

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

Absolute versus relative expression measures

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

Absolute versus relative expression measures

	Distance between	
	samples	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.