Distances

Katie Pollard

BMI 206

In this unit we will learn …

- How commonly used distance measurements encode different notions of “close”
- How to measure similarity of high-dimensional vectors

Distances

Multivariate statistical methods require a notion of pairwise distance between objects.

- Dissimilarity
  - Non-negative: $d(x,y) \geq 0$
  - Symmetric: $d(x,y)=d(y,x)$
  - Monotone: $d(x,y)>d(x,z)$ if z more similar to x
- Metric (additional conditions)
  - Definite: $d(x,y)=0$ iff $x=y$
  - Triangle inequality: $d(x,y)+d(y,z) \geq d(x,z)$

Distance Metrics

- Manhattan distance (Hamming for binary data)

$$d(x,y) = \sum_i |x_i - y_i| \in (0,\infty)$$

- Euclidean distance

Examples of the Minkowski metric

Correlation Distances

$$d(x,y) = 1 - r(x,y) \in (0,2)$$

Sample correlation measures r(x,y):

- Pearson
- Uncentered (cosine-angle distance)
- Spearman
- Kendall's Tau
- Maximal Information Coefficient

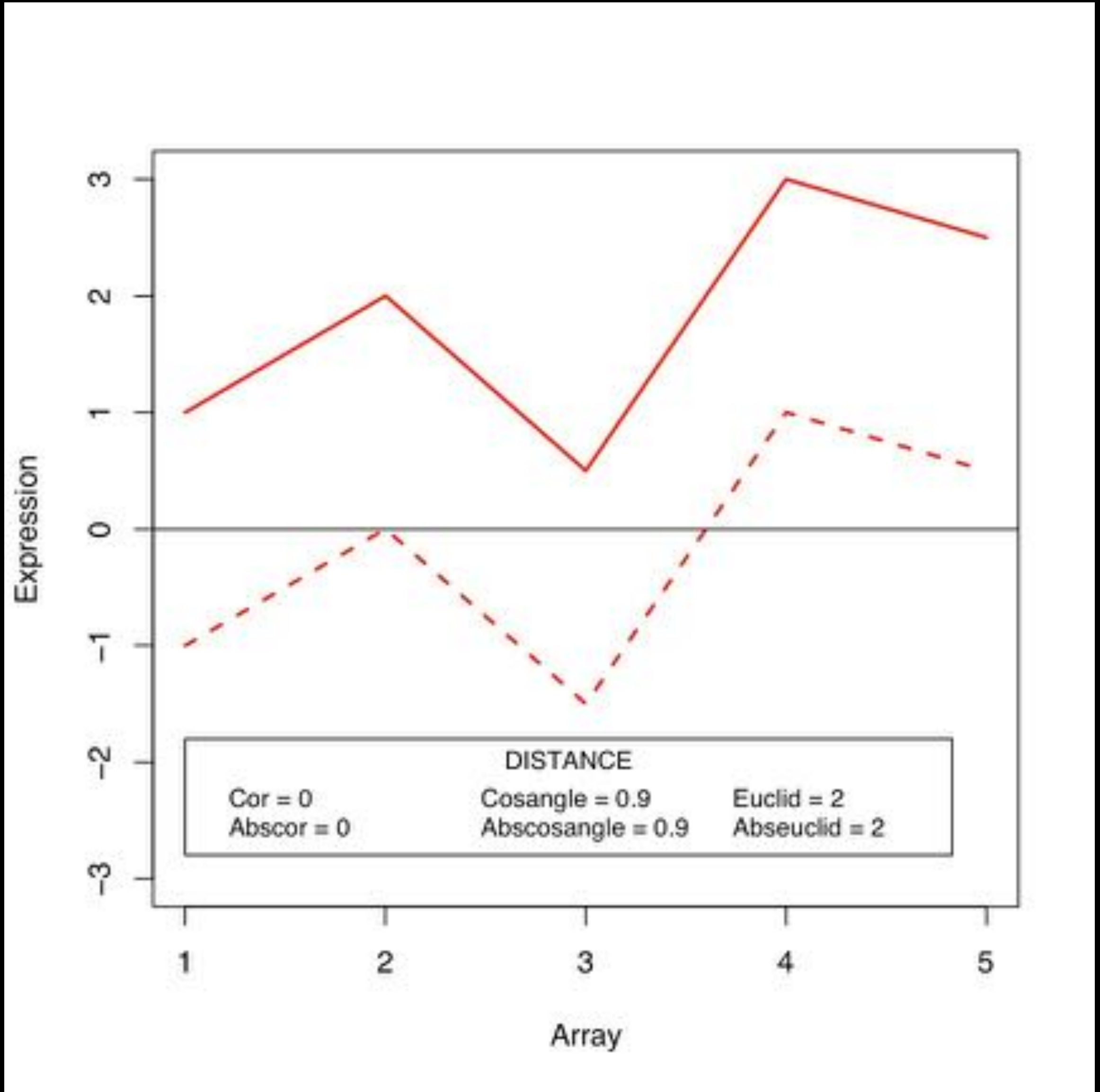
More on Distances

- Minkowski metrics: magnitude
- Correlation distances: pattern (or both)
- The absolute value of any distance can also be used, e.g.

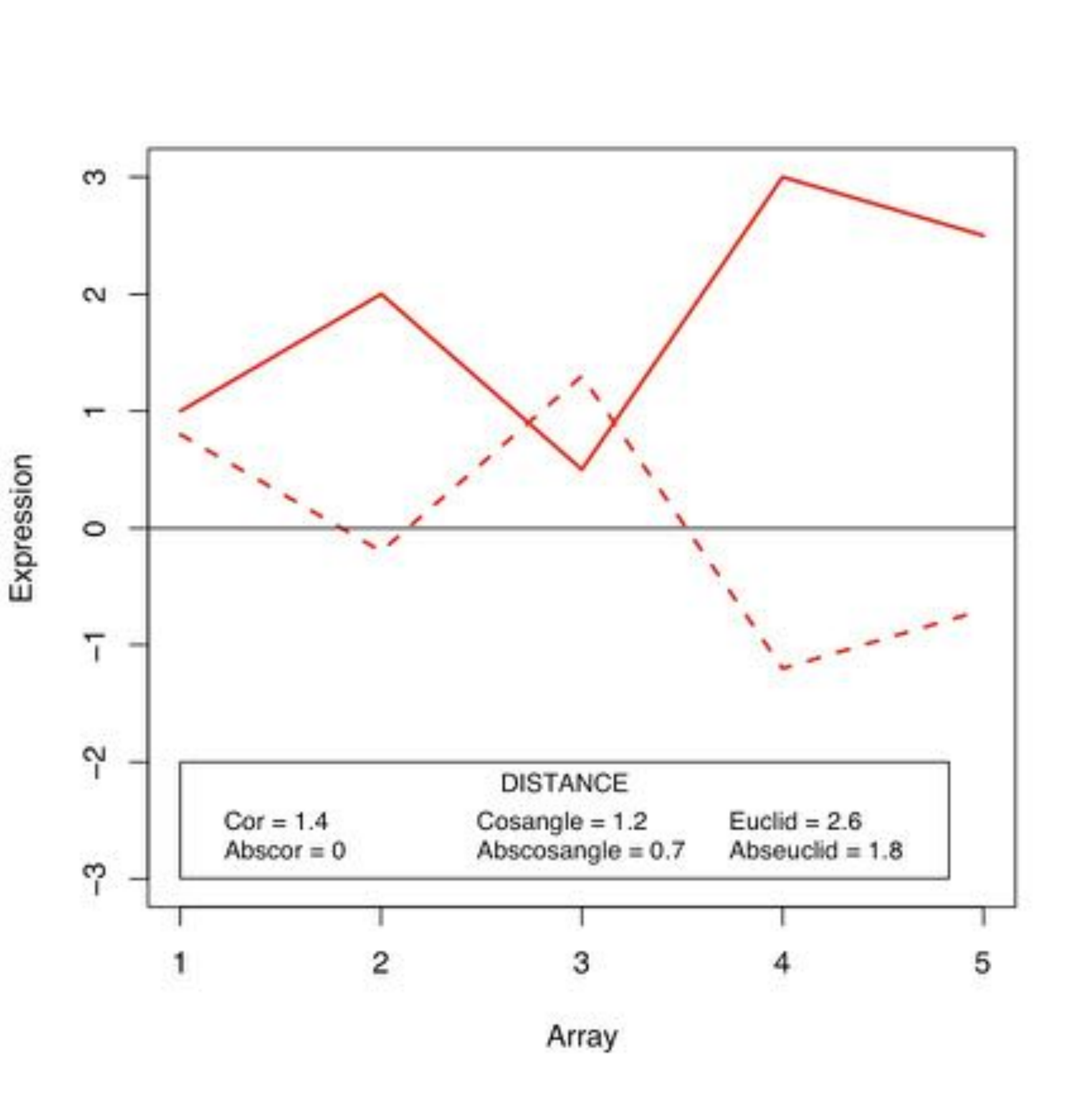
- Distances between distributions are a different concept, e.g., Kullback-Leibler

$$D_{KL}(p(X)||q(X)) = -\sum_x p(x) \log\{q(x)/p(x)\}$$
$$= \sum_x p(x) \log\{p(x)/q(x)\}$$

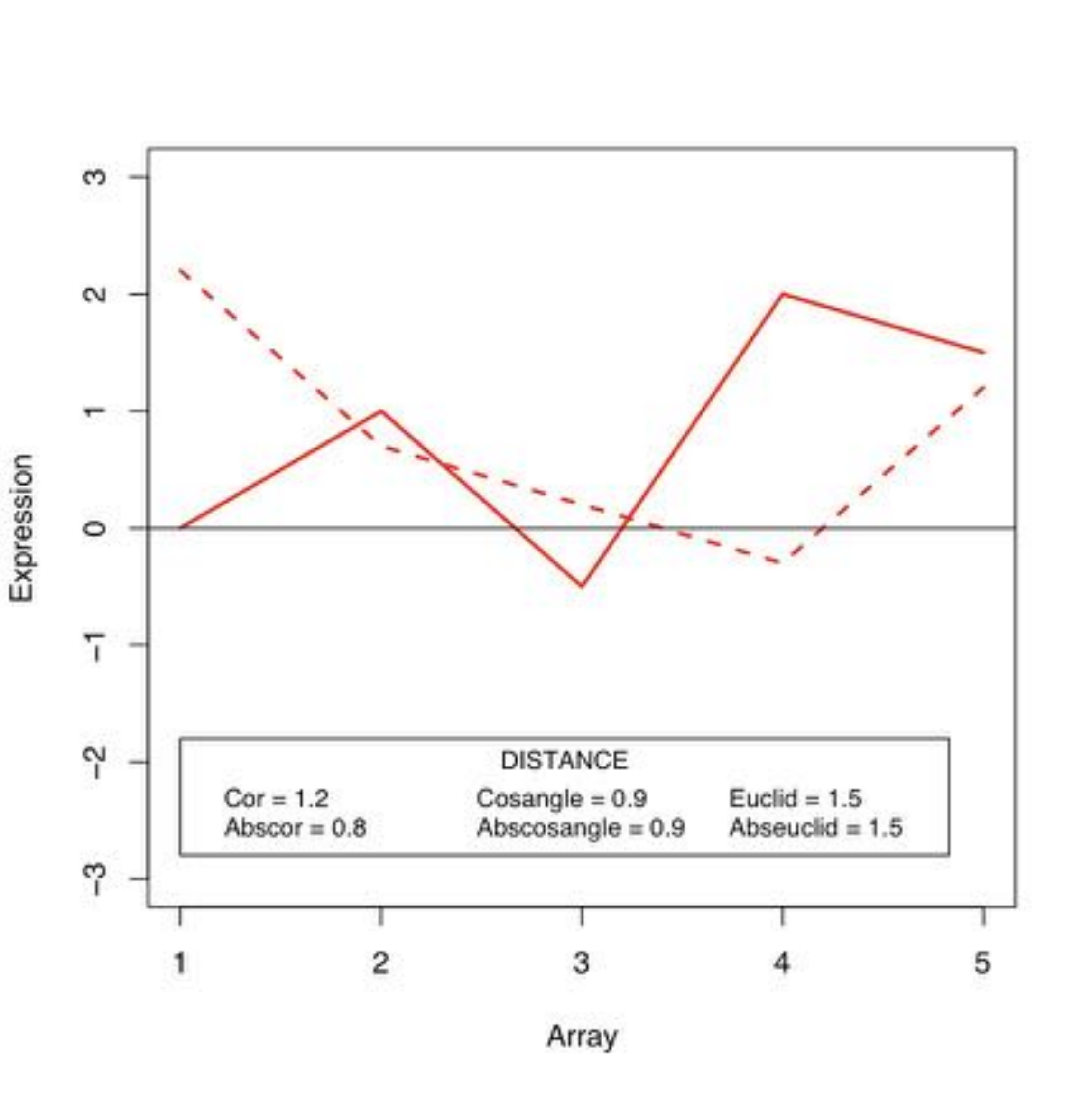
Perfectly Correlated



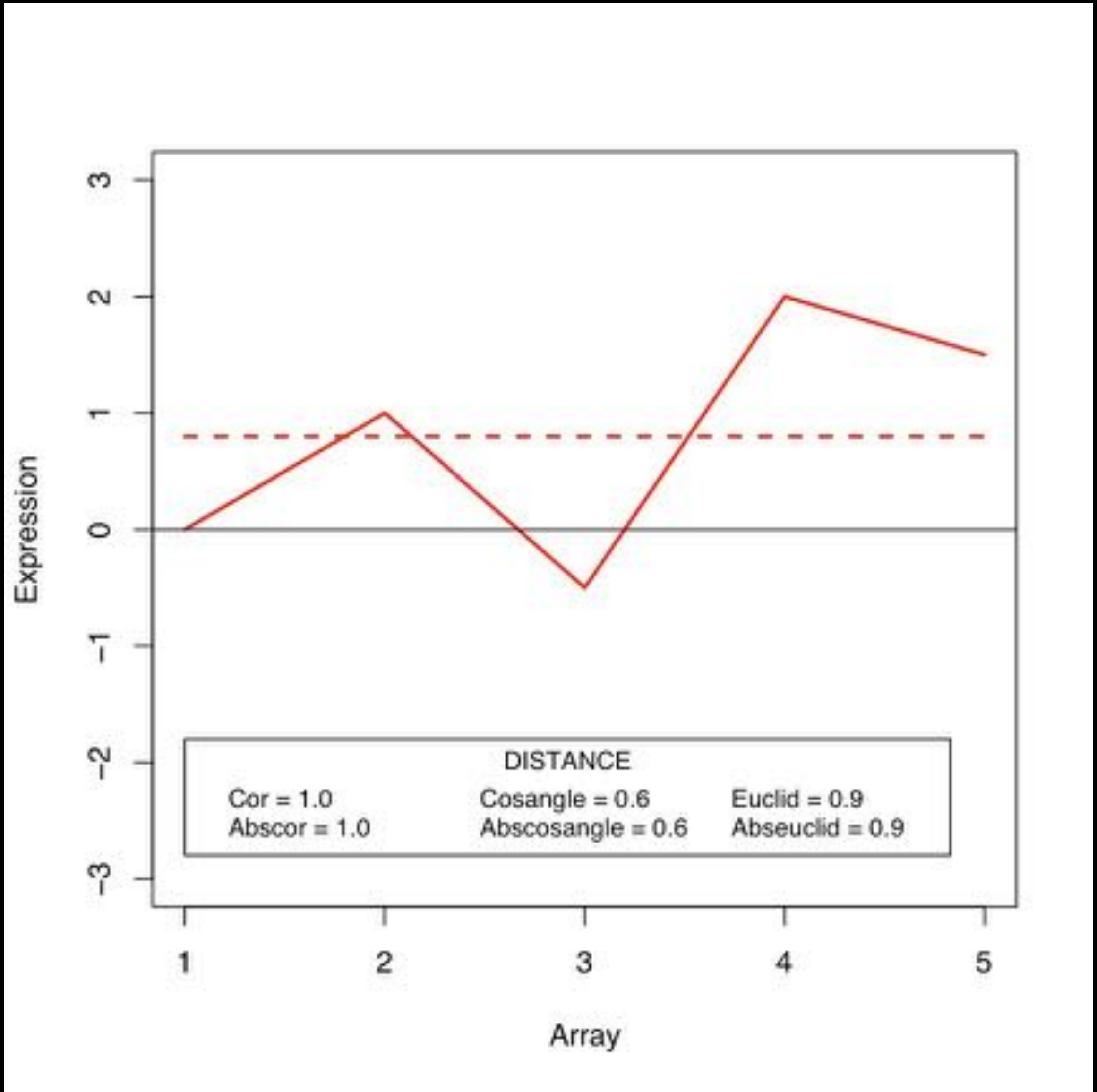
Anti-Correlated



Same Mean, Uncorrelated



Same Mean, No Variation



Distances in R

Function	Package	Distances
`dist`	`stats`	Euclidean, Manhattan, Canberra, max, binary
`daisy`	`cluster` `bioDist`	Euclidean, Manhattan
`distancematrix` `distancevector`	`hopach`	Euclidean, cor, cosine-angle (abs versions)