Introduction to Microarray Data Analysis and Gene Networks Lecture 4

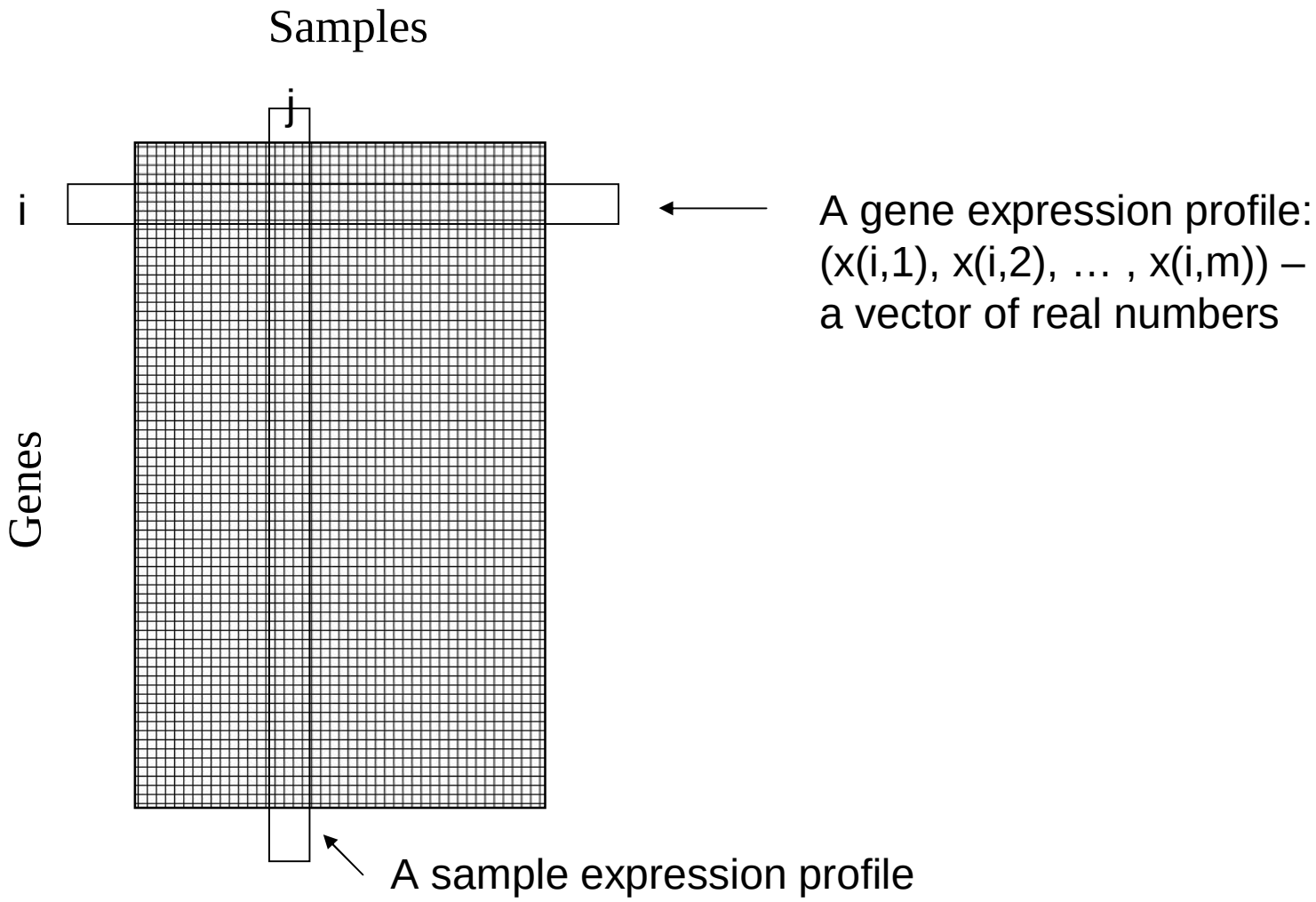
Alvis Brazma

European Bioinformatics Institute

Lecture 4

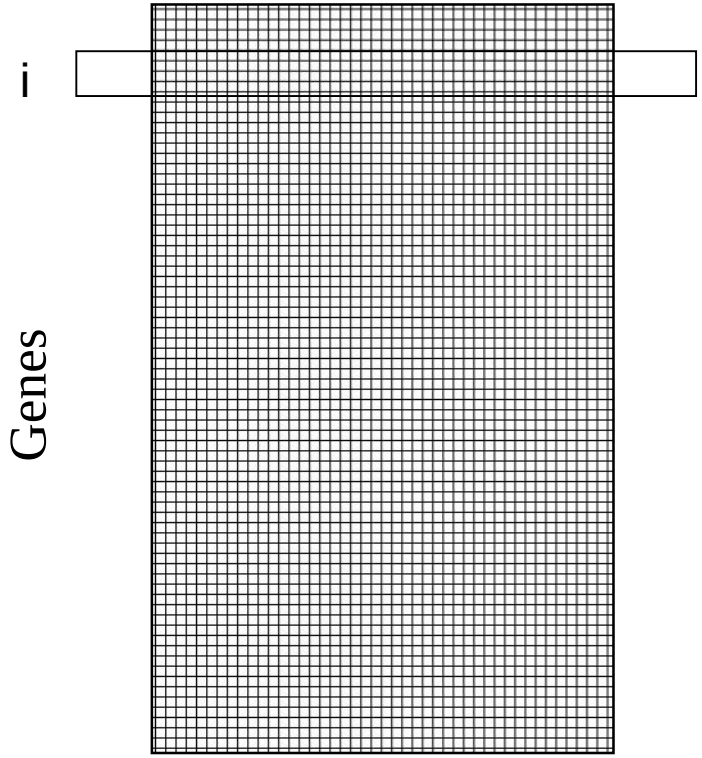
- Distance measures used to estimate similarity between gene and sample expression profiles
- Missing data-point imputation
- Practical

Gene Expression Profile

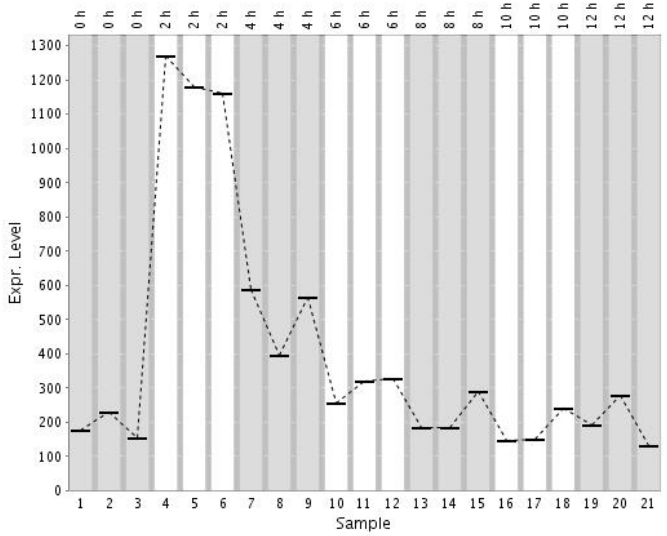


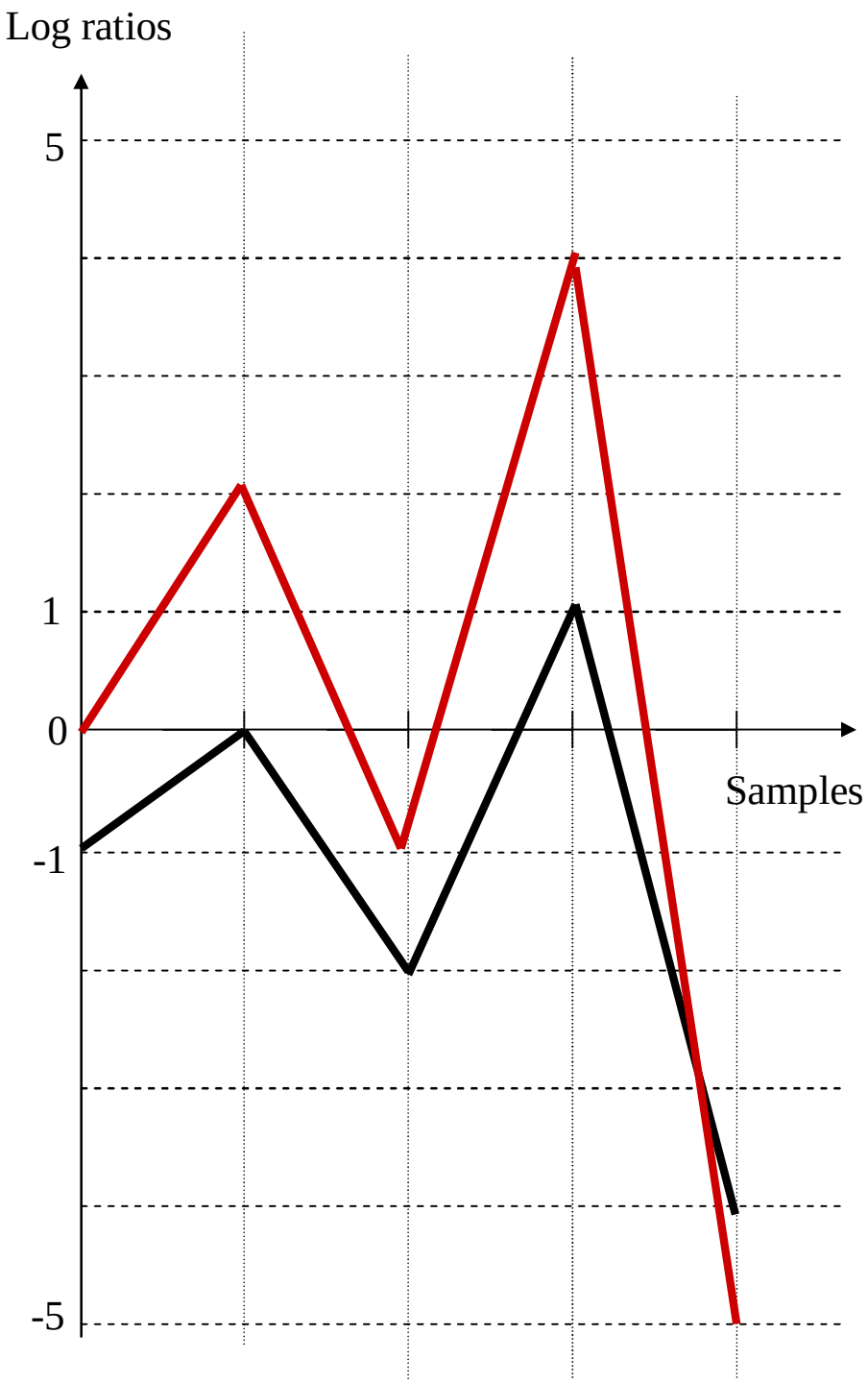
Gene Expression Profile

Samples



A gene expression profile:
(x(i,1), x(i,2), … , x(i,m)) –
a vector of real numbers

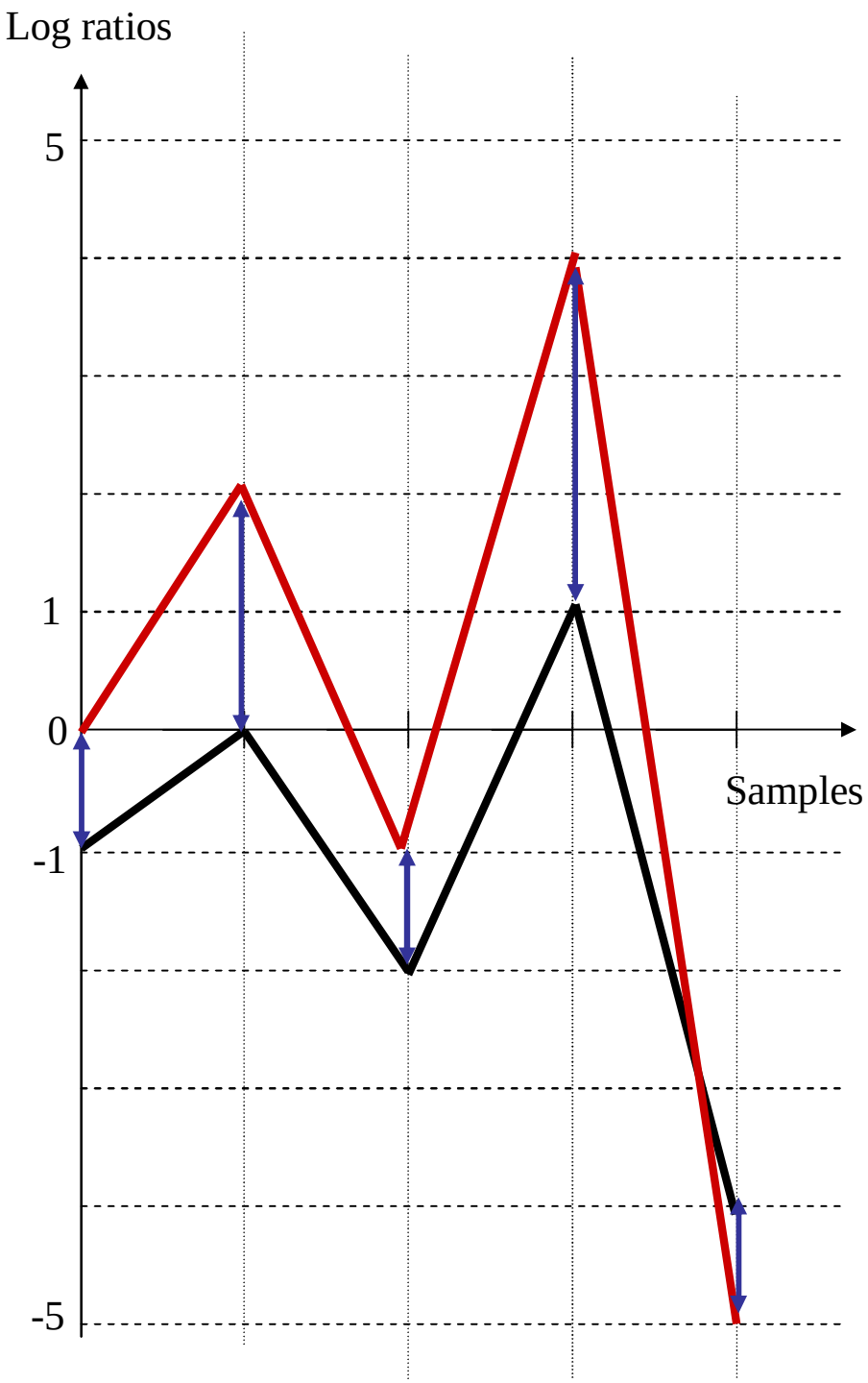




How to measure distance between two gene (or sample) expression profiles?

A = (-1, 0,-2,1,-4)

B = (0,2,-1,4,-5)



A = (-1, 0,-2,1,-4)

B = (0,2,-1,4,-5)

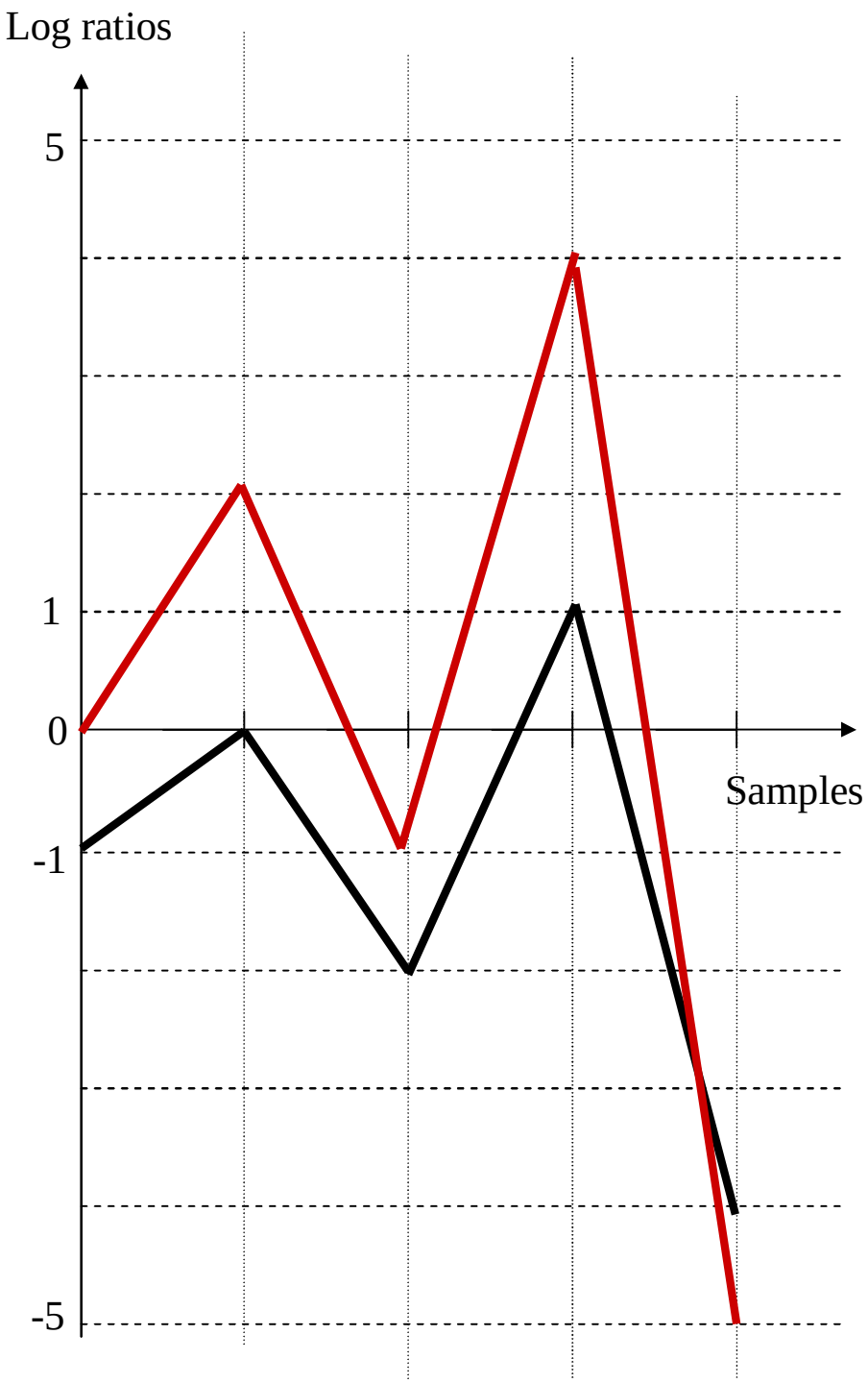
Euclidean distance =

(add up the squares of all arrows and take a square root)

$= (1+4+1+9+1)^{1/2} = 4$

Euclidean distance

$$D_{Eucl}(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

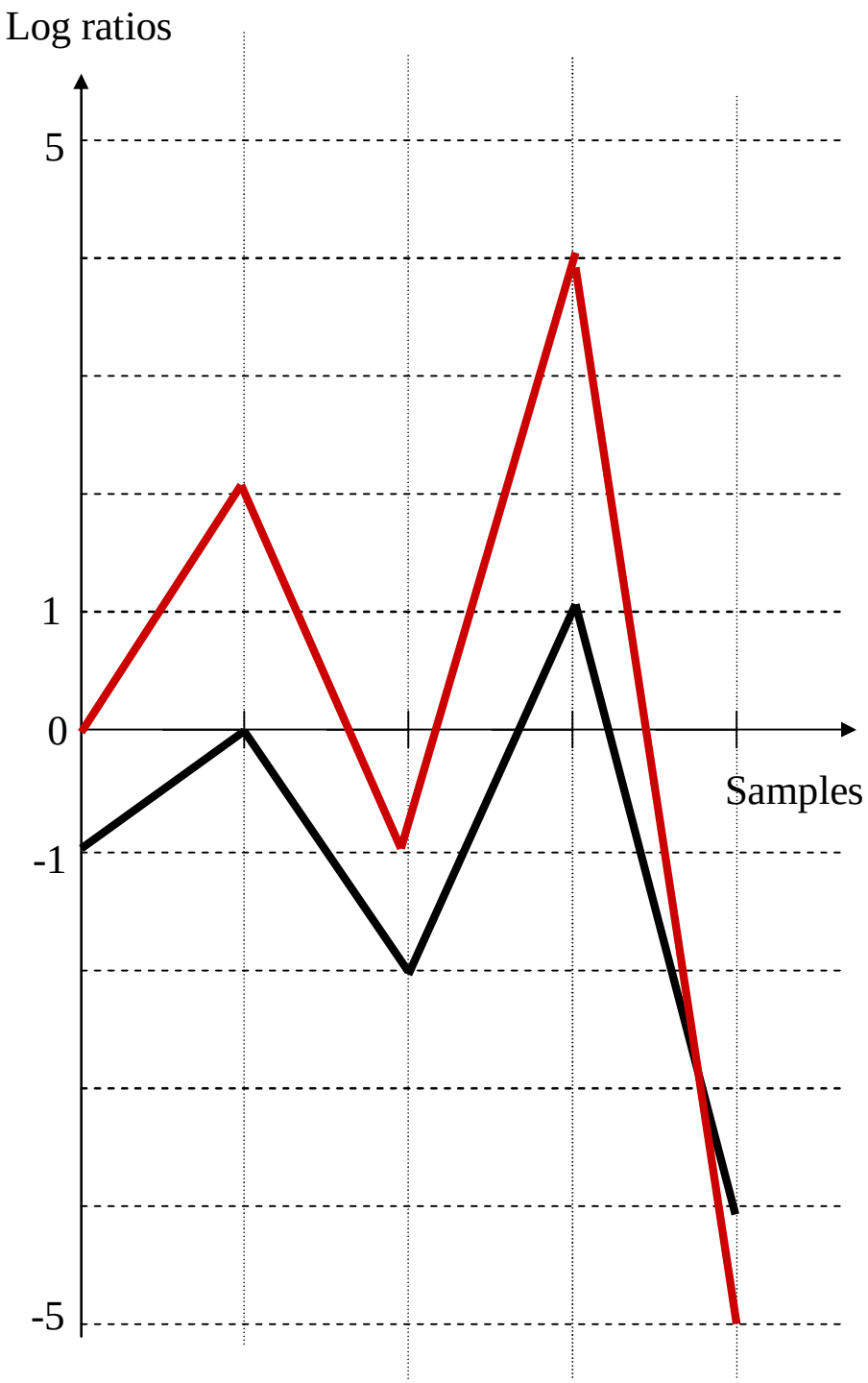


A = (-1, 0,-2,1,-4)

B = (0,2,-1,4,-5)

The absolute values are not very meaningful (remember that *sequence effects* are large) - the Euclidean distance may not be the best

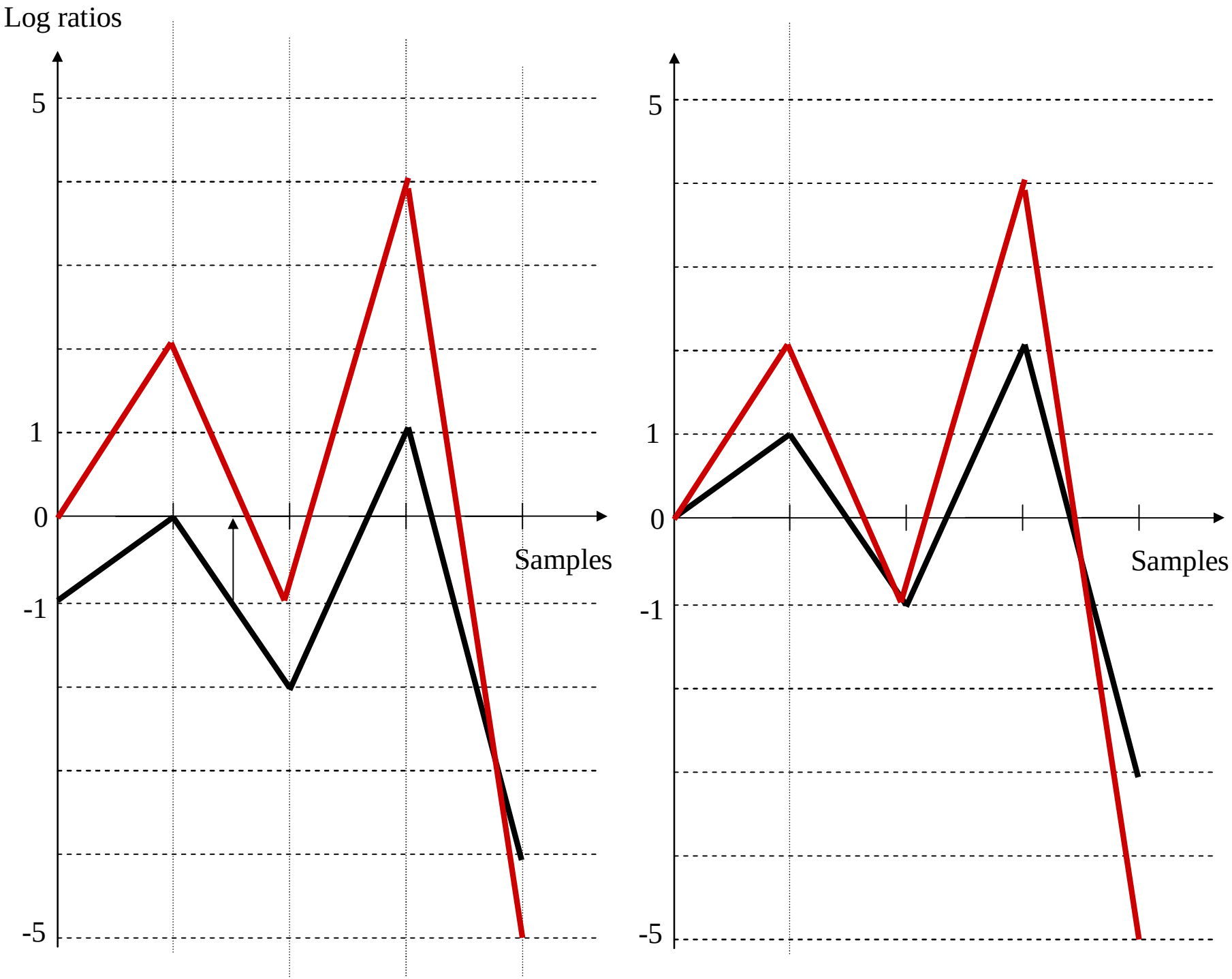
How to measure similarities in *trends*?

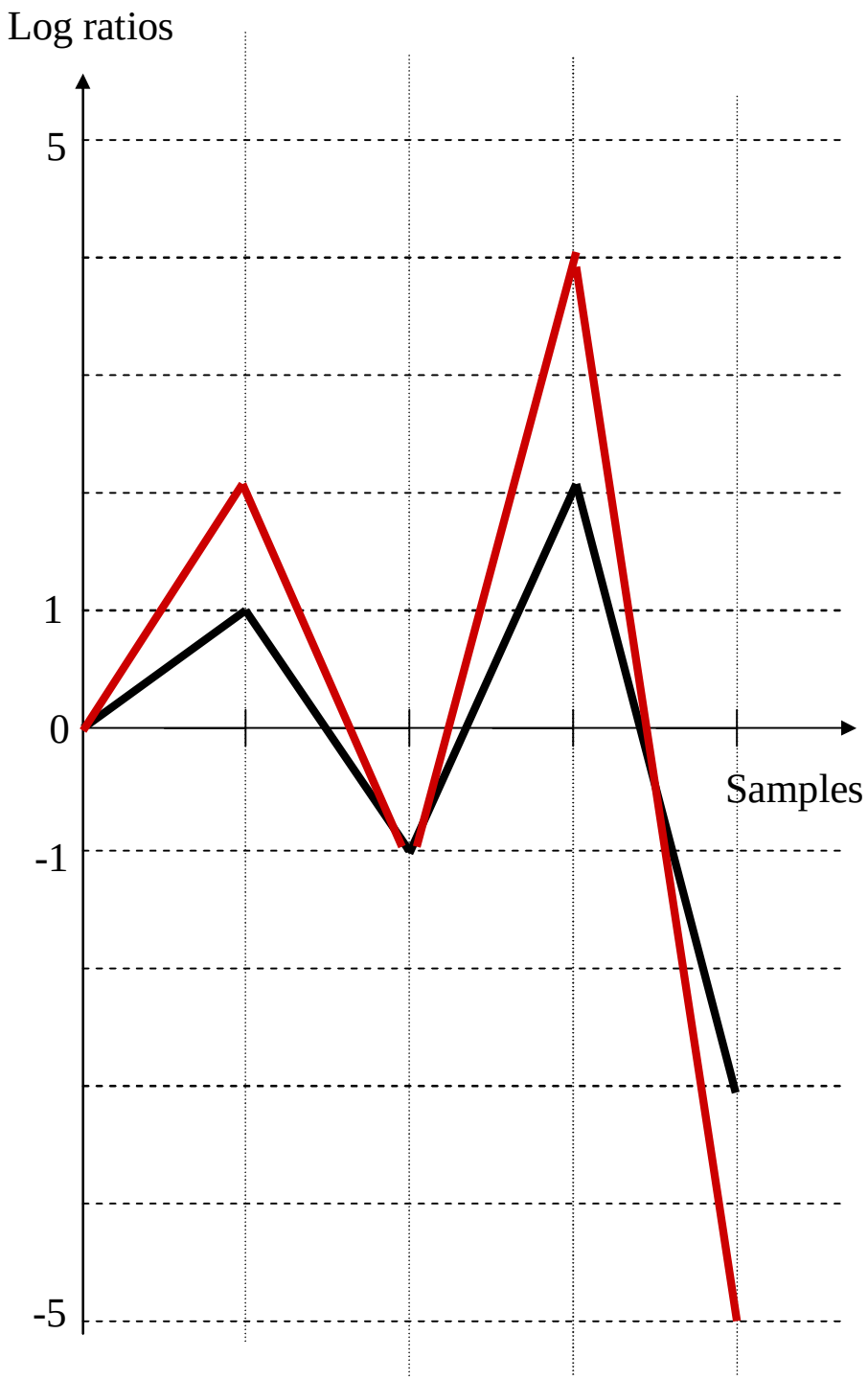


A = (-1, 0,-2,1,-4)

B = (0,2,-1,4,-5)

1. Center all vectors around 0


Log ratios
5
1
0
-1
-5
Samples
5
1
0
-1
-5
Samples



$\mathbf{A'} = (0, 1, -1, 2, -3)$

$\mathbf{B} = (0, 2, -1, 4, -5)$

Chord distance =

Make the length of both equal to 1

$|\mathbf{A'}| = (0+1+1+4+9)^{1/2} \approx 4$

$|\mathbf{B}| = (0+4+1+16+25)^{1/2} \approx 7$

The length of a vector

Given a vector *A=(a1, …, ak)*, we define its length |*A*| as

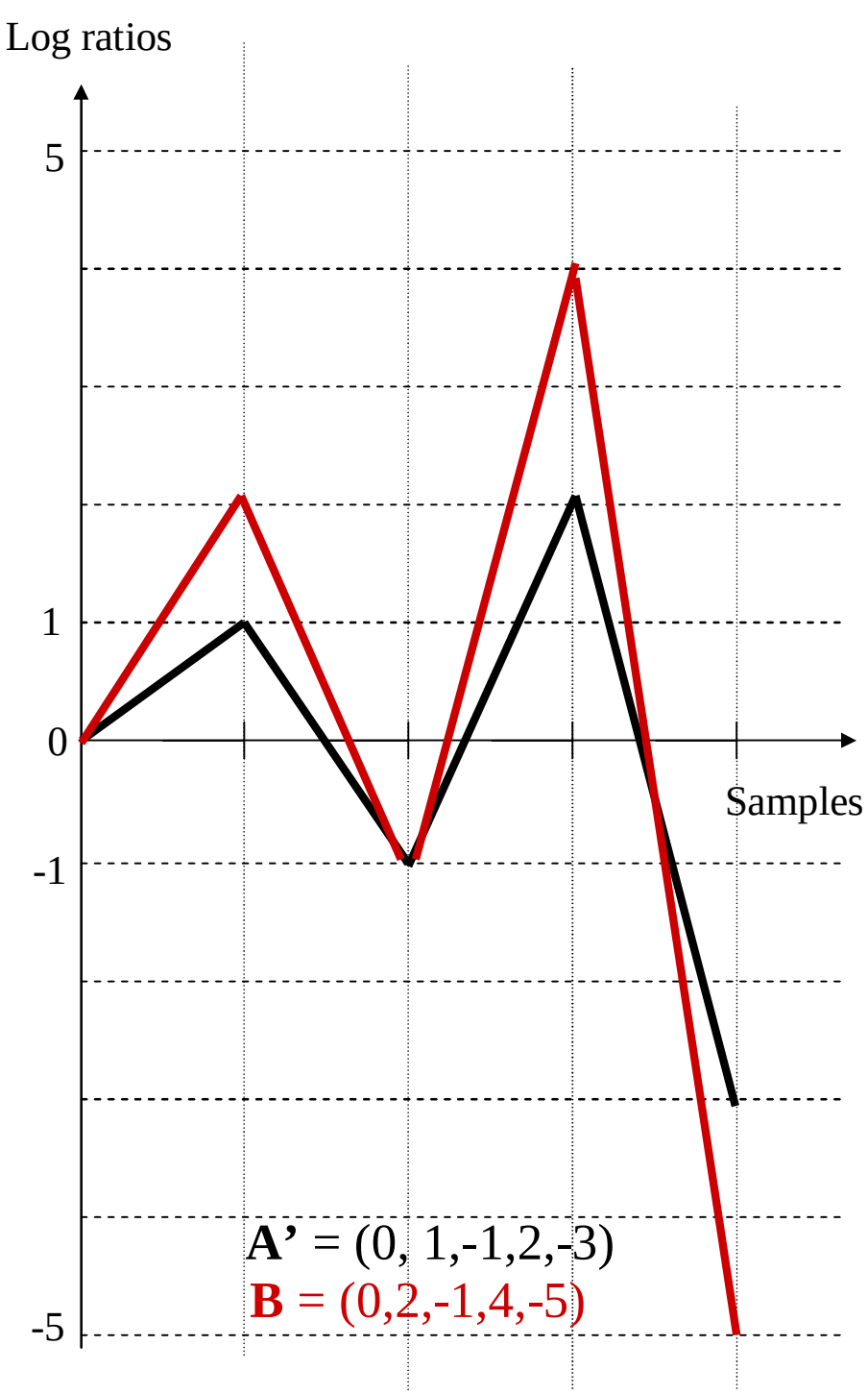
$$\left|A\right| = \sqrt{a_1^2 + \ldots + a_k^2}$$

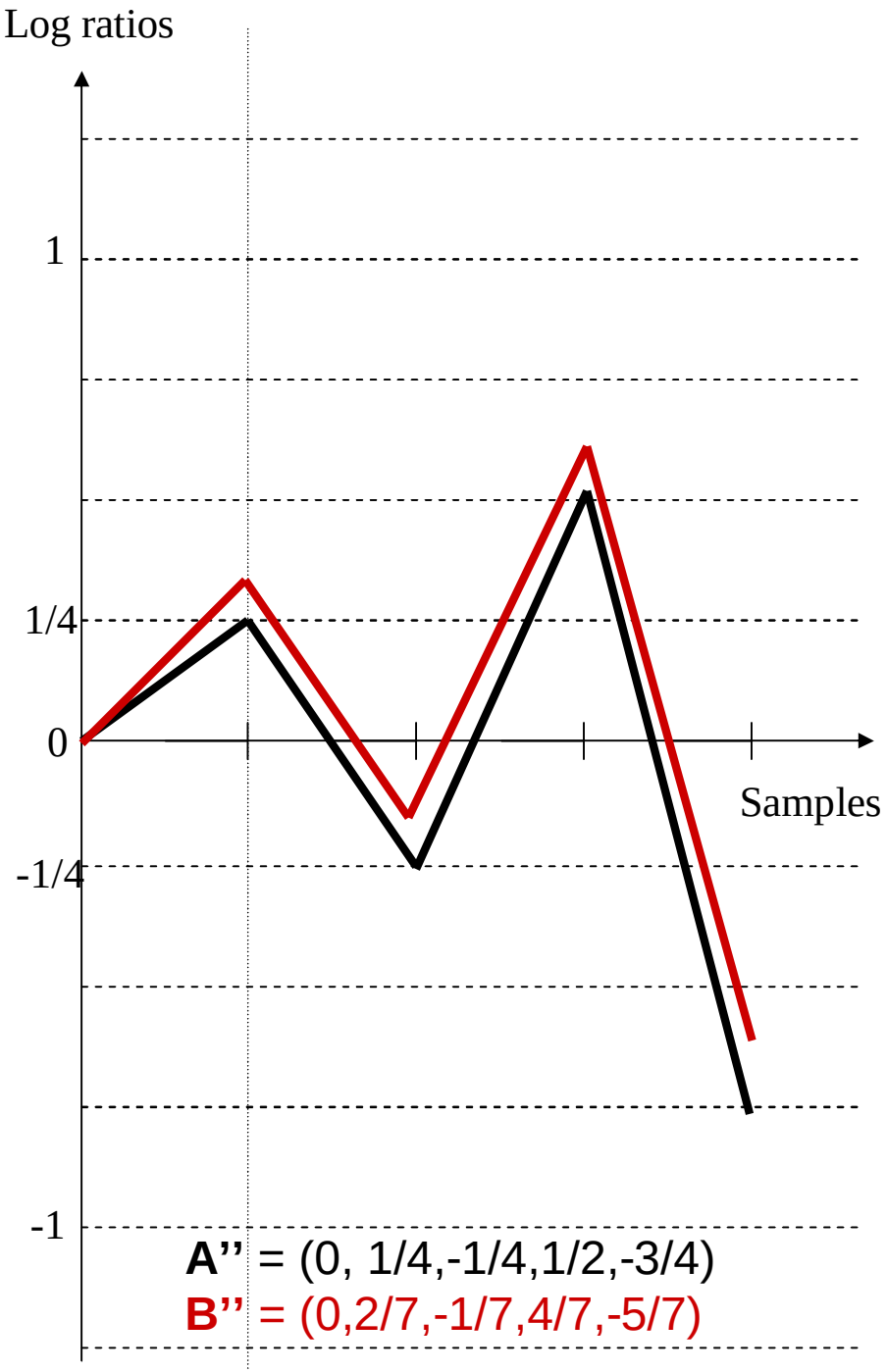
|A'| = $(0+1+1+4+9)^{1/2} \approx 4$

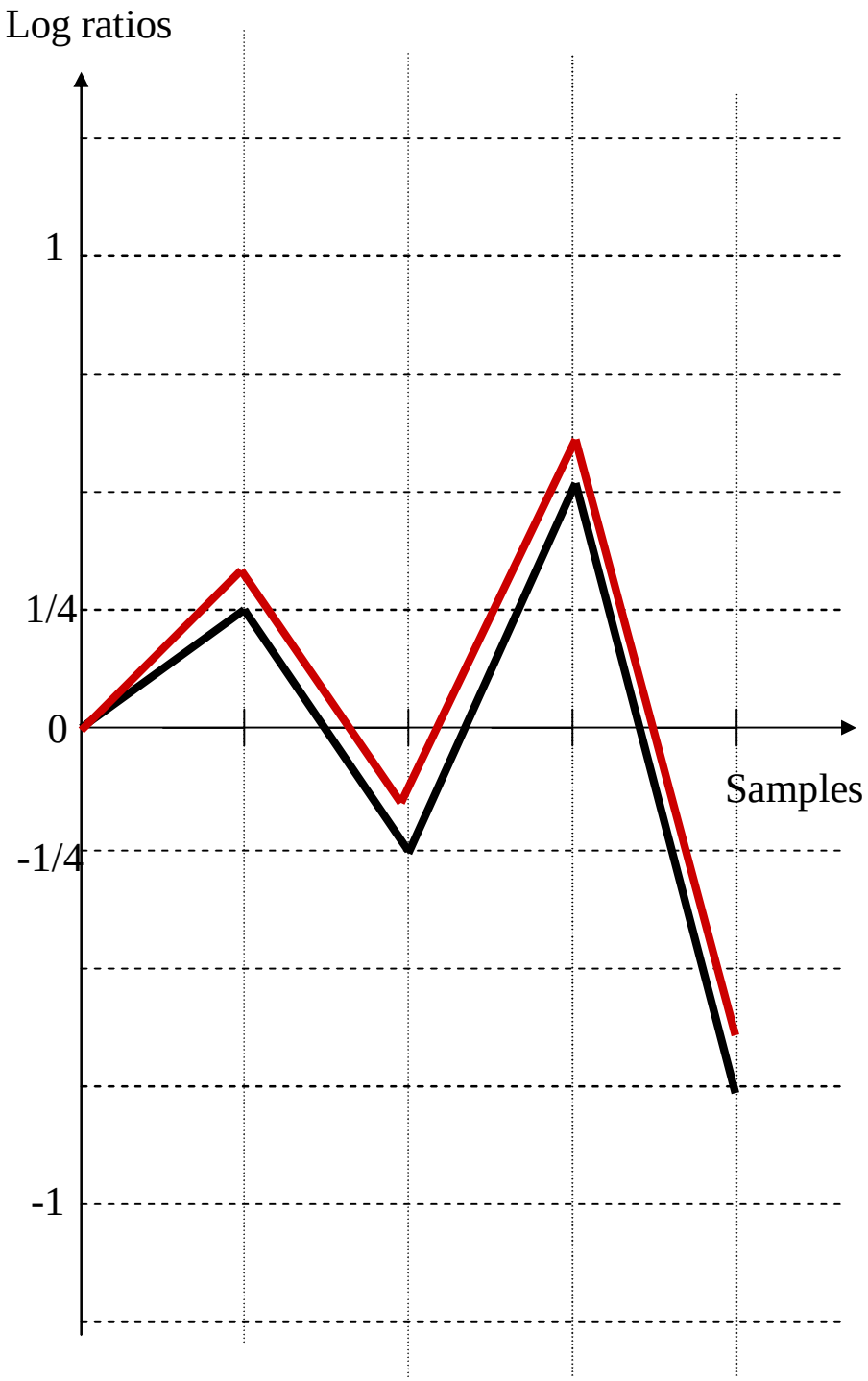
|B'| = $(0+4+1+16+25)^{1/2} \approx 7$

A'' = (0, 1/4,-1/4,1/2,-3/4)

B'' = (0,2/7,-1/7,4/7,-5/7)


Log ratios
5
1
0
-1
-5
Samples
A’ = (0, 1,-1,2,-3)
B = (0,2,-1,4,-5)


Log ratios
1
1/4
0
-1/4
-1
Samples
A’’ = (0, 1/4,-1/4,1/2,-3/4)
B’’ = (0,2/7,-1/7,4/7,-5/7)



$\mathbf{A'} = (0, 1, -1, 2, -3)$

$\mathbf{B'} = (0, 2, -1, 4, -5)$

Chord distance =

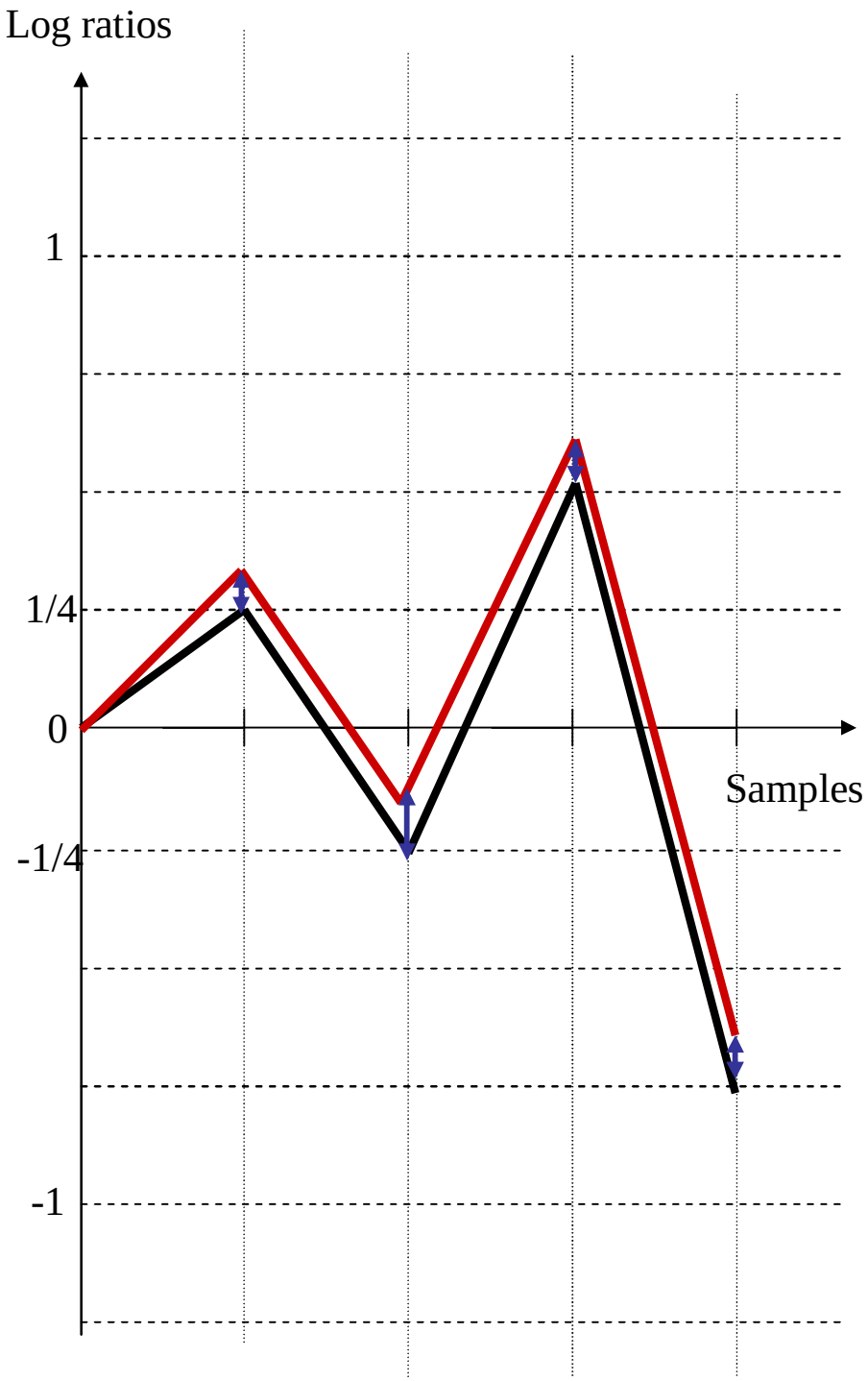
1. Center all vectors around 0
2. Make the length of both equal to 1

$|\mathbf{A'}| = (0+1+1+4+9)^{1/2} \approx 4$

$|\mathbf{B'}| = (0+4+1+16+25)^{1/2} \approx 7$

$\mathbf{A''} = (0, 1/4, -1/4, 1/2, -3/4)$

$\mathbf{B''} = (0, 2/7, -1/7, 4/7, -5/7)$



$|\mathbf{A'}| = (0+1+1+4+9)^{1/2} \approx 4$

$|\mathbf{B'}| = (0+4+1+16+25)^{1/2} \approx 7$

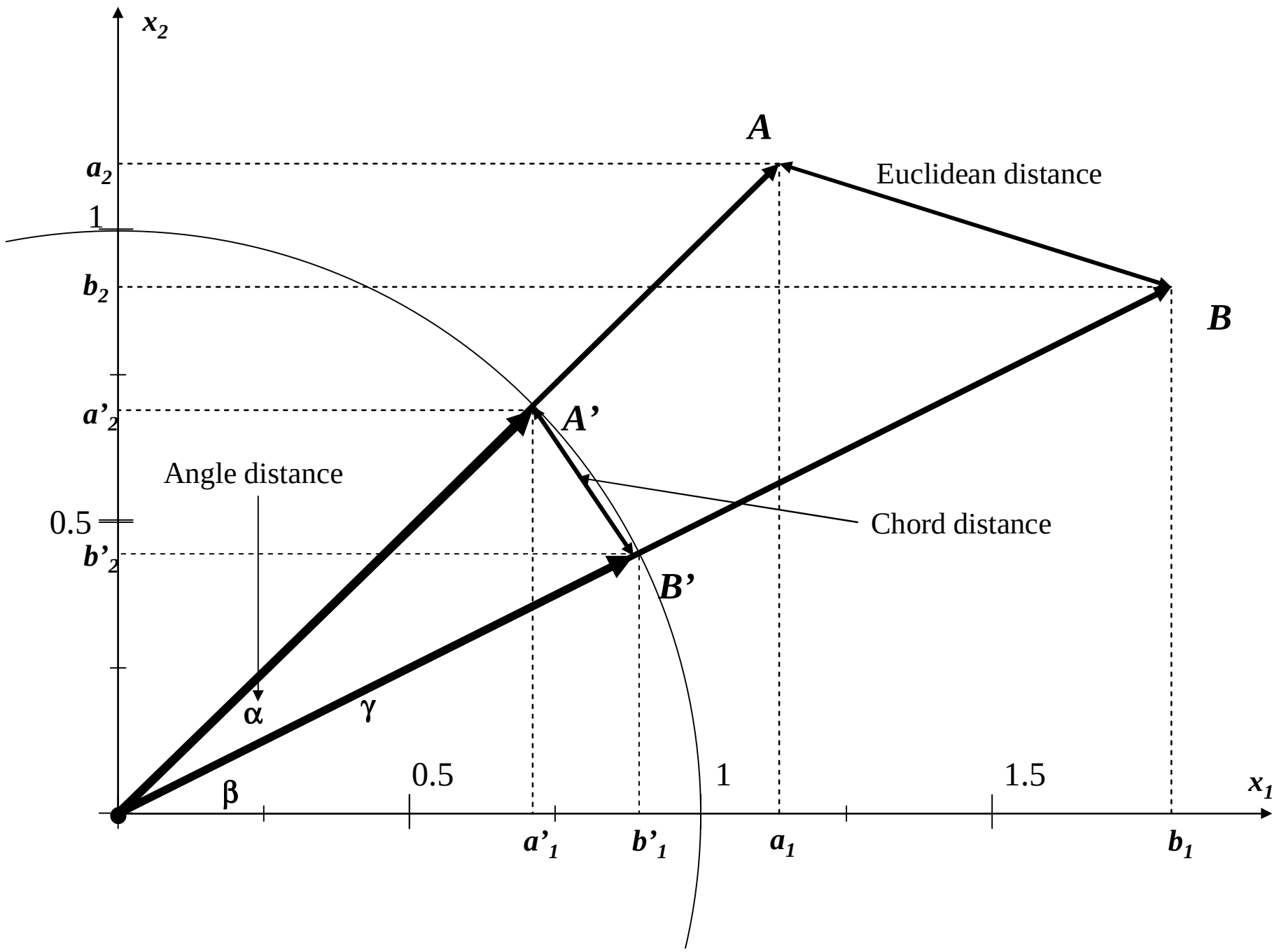
$\mathbf{A''} = (0, 1/4, -1/4, 1/2, -3/4)$

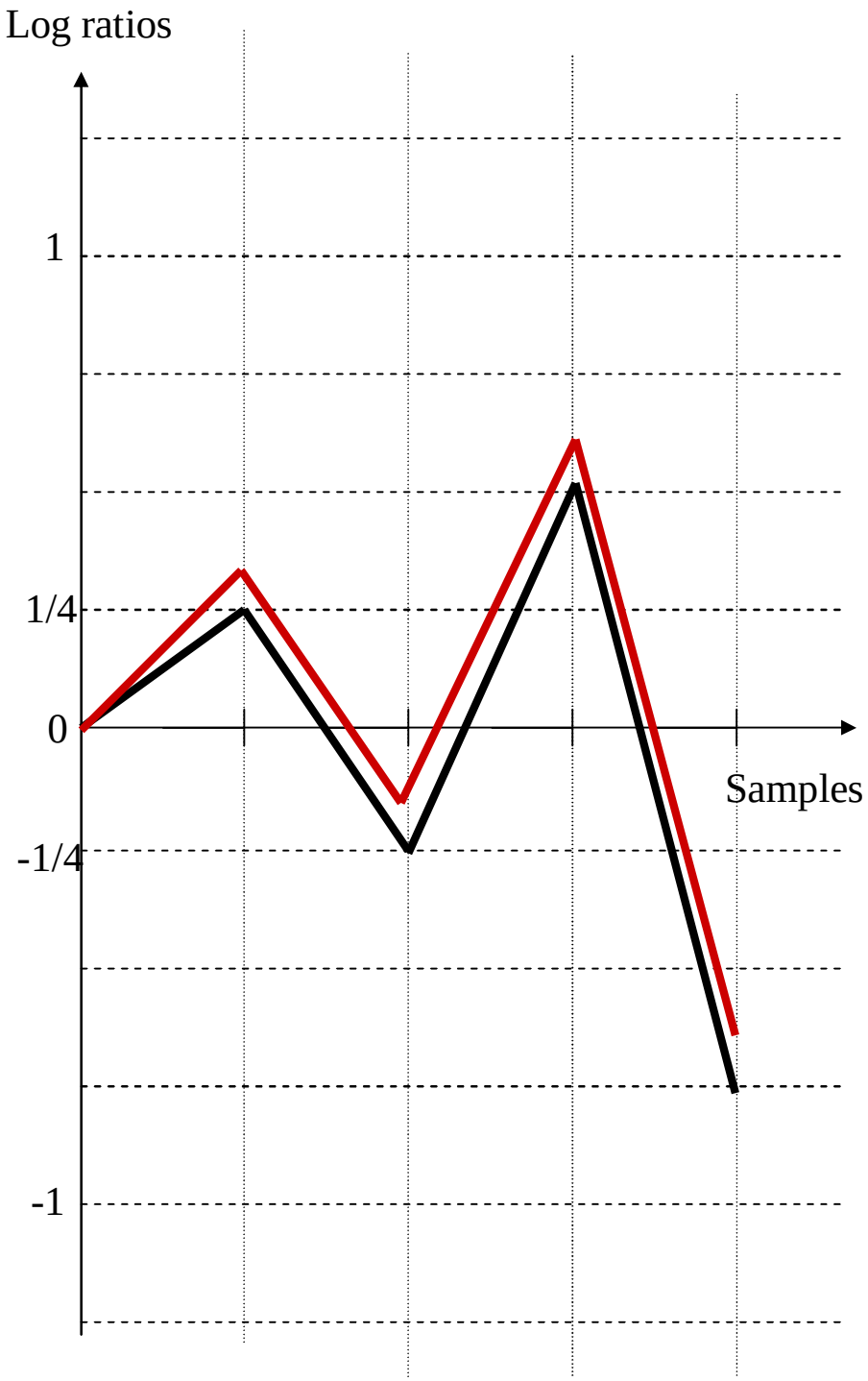
$\mathbf{B''} = (0, 2/7, -1/7, 4/7, -5/7)$

Chord distance =

1. Center all vectors around 0
2. Make the length of both equal to 1
3. Calculate Euclidean distance between the centered and scaled vectors

(see that the chord distance in this case is about 0.02)


x_2
a_2
1
b_2
a'_2
0.5
b'_2
A
B
A'
B'
Euclidean distance
Angle distance
Chord distance
α
β
γ
0.5
1
1.5
x_1
a'_1
b'_1
a_1
b_1



$|\mathbf{A'}| = (0+1+1+4+9)^{1/2} \approx 4$

$|\mathbf{B'}| = (0+4+1+16+25)^{1/2} \approx 7$

A'' = (0, 1/4,-1/4,1/2,-3/4)

B'' = (0,2/7,-1/7,4/7,-5/7)

Correlation distance =

1. Very similar to Chord distance – calculate the cos between the two vectors:
   cos(A'',B'')=
   =0*0+1/4*2/7+(-1/4)*(-1/7)+
   +1/2*4/7+(-3/4)*(-5/7) =
   =11/14
2. Cor_dist = 1-cos(A'',B'') =
   =3/14 ≈ 1/5

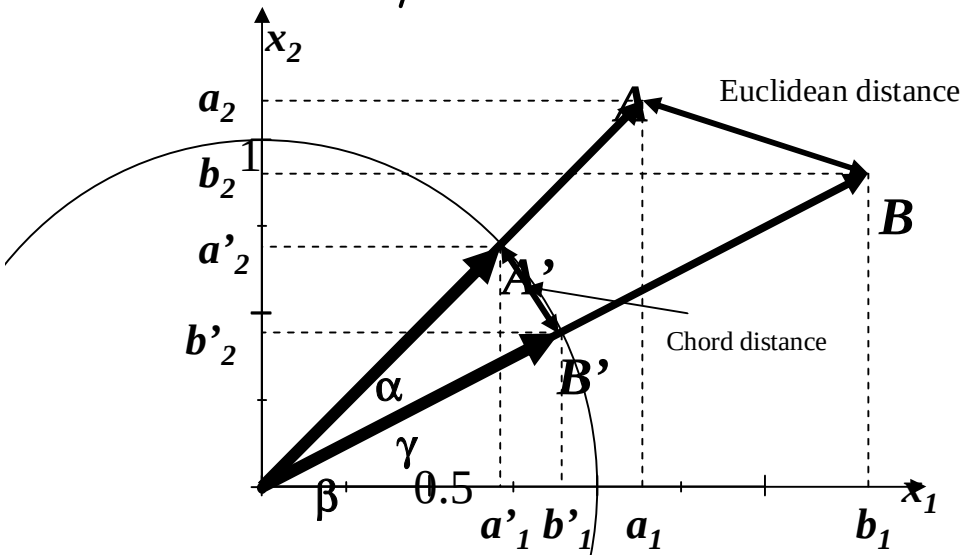
Relationships between chord and correlation distances

$$D_{chord}(A,B) = \sqrt{(a'_1 - b'_1)^2 + (a'_2 - b'_2)^2} = \sqrt{\left(\frac{a_1}{|A|} - \frac{b_1}{|B|}\right)^2 + \left(\frac{a_2}{|A|} - \frac{b_2}{|B|}\right)^2}$$

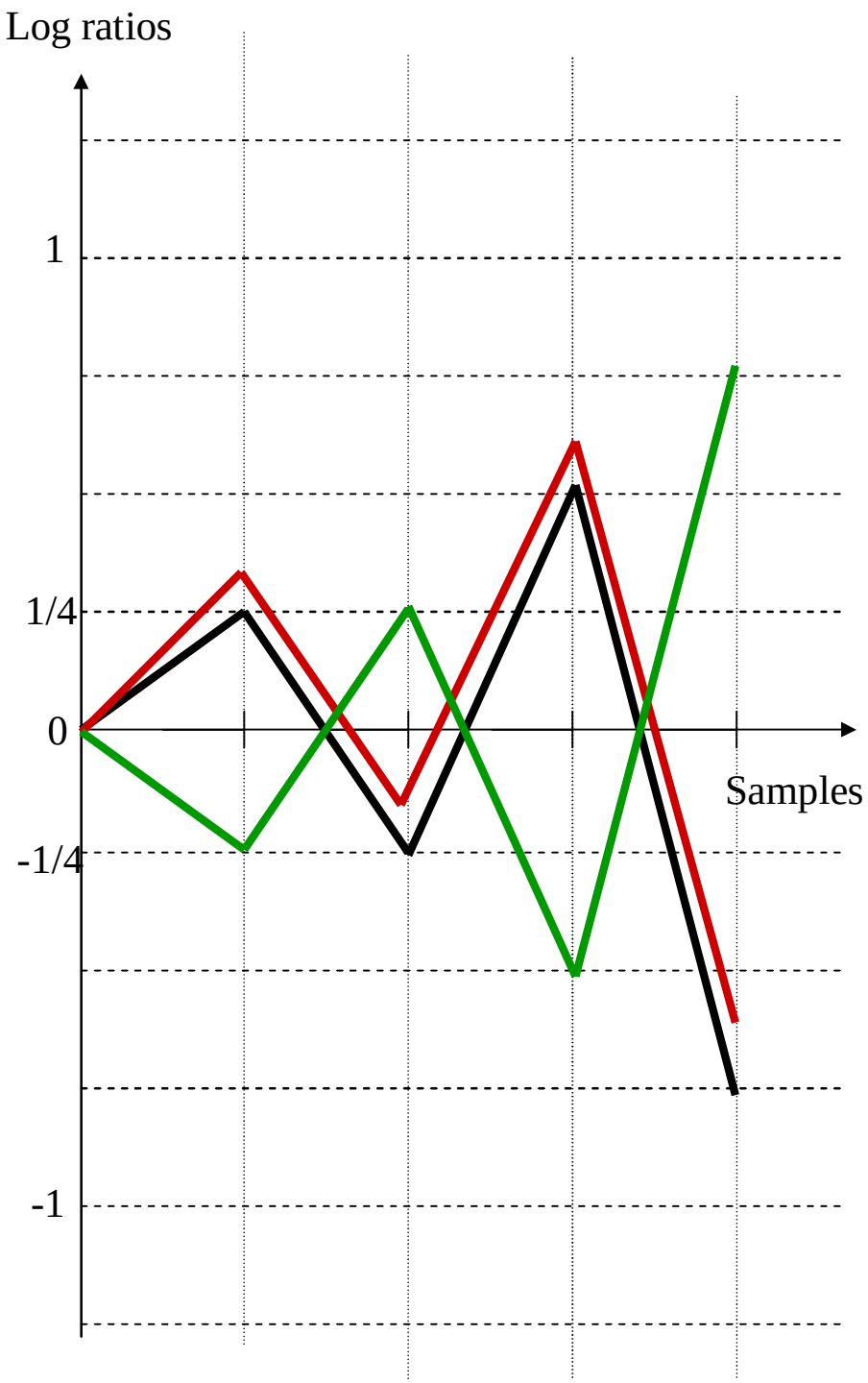
$$D_{chord}(A,B) = \sqrt{2\left(1 - \left(a'_1 b'_1 + a'_2 b'_2\right)\right)} = \sqrt{2(1 - \cos\alpha)}$$

$$D_{chord}(A,B) = 2\sin\alpha/2$$

$$2\sin^2\alpha/2 = 1 - \cos\alpha$$



Cor(A,B) = cos(AB), if A and B are means centered normalised (length 1) vectors

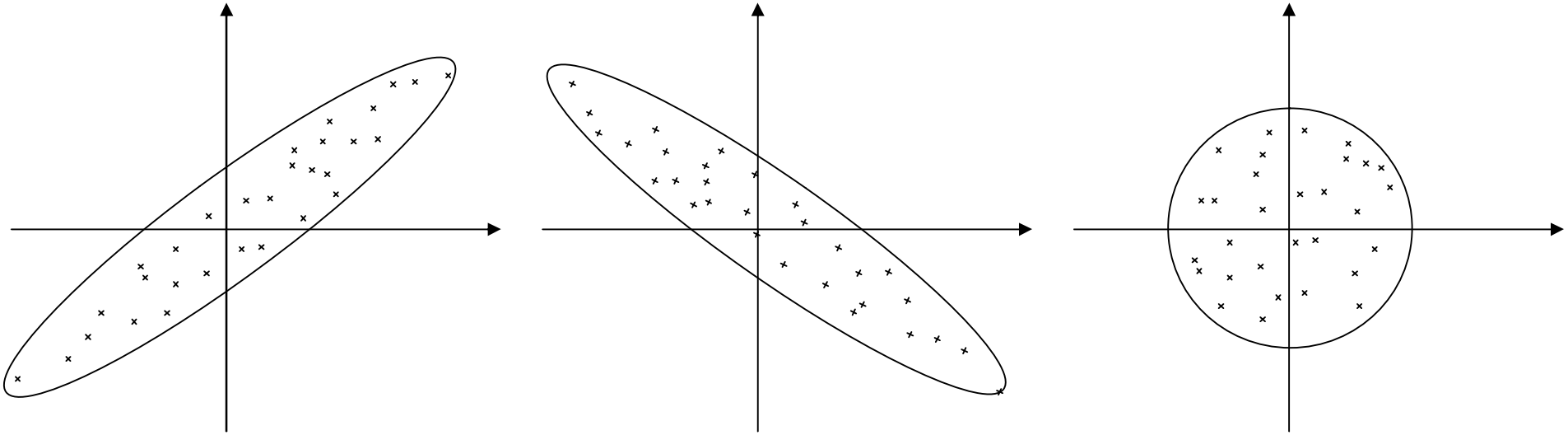


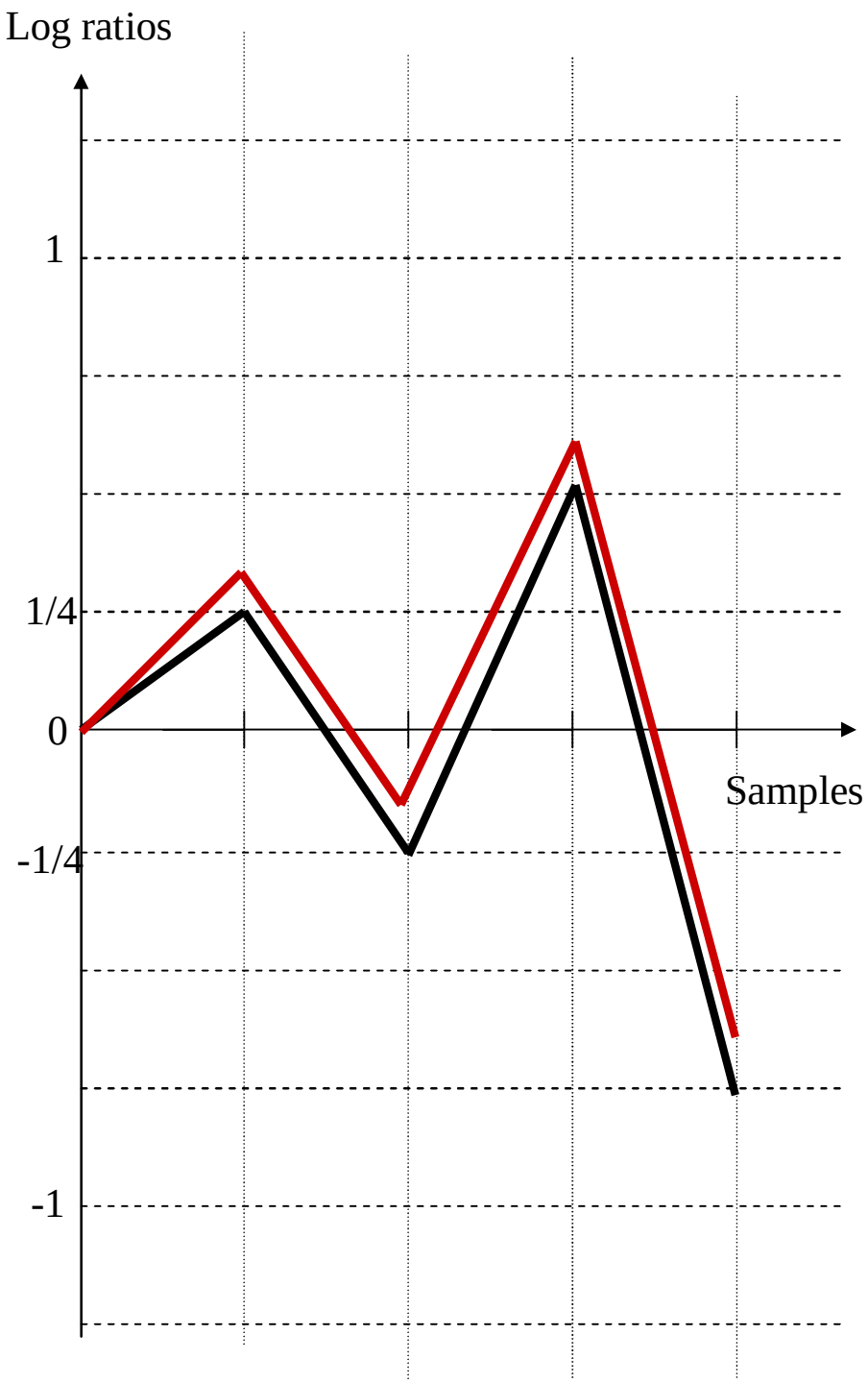
Correlation and anticorrelation

1 – cos x – perfect correlation has distance 0, anticorrelation has max distance 2

1 - |cos x| - both perfect correlation and perfect anticorrelation distances are 0

Correlation, anticorrelation and no correlation





Rank correlation

A’’ = (0, 1/4,-1/4,1/2,-3/4)
B’’ = (0,2/7,-1/7,4/7,-5/7)

Transform the values to ranks

A’’’ = (0,1,-1,2,-2)
B’’’= (0,1,-1,2,-2)

Compute the (correlation) distance between them (for that first normalise them to length 1).

Advantages and disadvantages of rank correlation based distances

- Advantages – does not depend on the precise values
- Disadvantages – ranks depend on the precise values in the large density arrears, e.g., when the expression values are very close to each other (closer than the error bars), their relative order (ranks) are very prone to error

Distance measures

A distance measure D(*A,B*) is said to be *metric*, if it satisfies the following properties:

- if *A*=*B*, then D(*A,B*) = 0, *i.e*., the distance of an object to itself is 0;
- if *A*≠*B*, then D(*A,B*) ≥ 0, *i.e*., the distance is always nonnegative;
- D(*A,B*) = D(*B,A*), *i.e*., it does not matter in which order we measure the distance;
- D(*A,B*) + D(*B,C*) ≥ D(*A,C*), *i.e*., given three objects, the length of a direct path from the first to the third objects cannot be greater than the length of the path through the second object.

Missing data points

- Why they arise?
  - Bad quality spot – e.g. flagged as bad by the image analysis software (e.g, so-called half moon spots, empty circles, …)
  - Very low intensity signal in one or both channels (may be 0 or infinity ratio)
  - Inconsistency between replicates (on the same or different arrays)

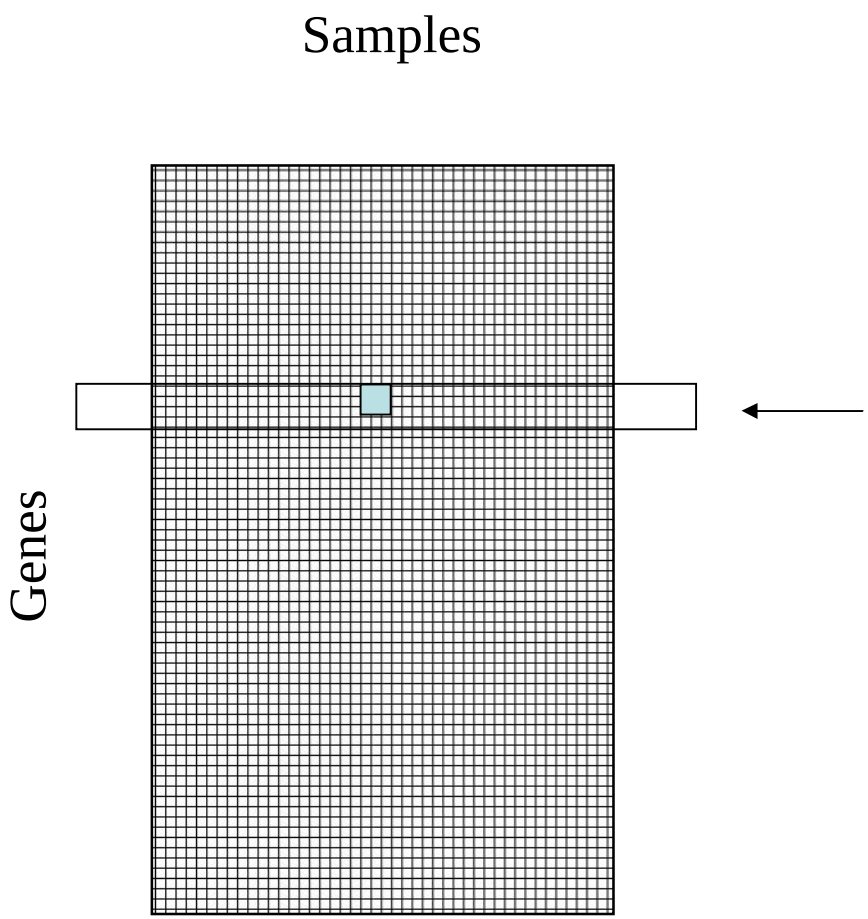
Missing data points

- Why they are a nuisance?
  - How to compute distance between vectors with missing data points – ignore the dimension
  - If many comparisons have to be made, missing dimensions may start to accumulate
- How to deal with them?
  - If replicates are available, they can be used
  - Replace missing values by 0
  - Replace by the row average value
  - K nearest neighbour imputation (KNN imputation)

KNN imputation

- We are given a gene expression matrix **M**
- Let **X**=(X1, X2, …, Xi, …, Xn) be a vector in the matrix **M** with a missing value at Xi at the dimension i
- Find in the gene expression data matrix matrix vectors $\mathbf{X^1}$, $\mathbf{X^2}$, …, $\mathbf{X^k}$, such that they are the k closest vectors to X in M (in the sense of a chosen distance measure) among the vectors that do not have a missing value at dimension i
- Replace the missing value Xi with the mean (or median) of X^1i, X^2i, …, X^ki , i.e., mean (median) of the values at dimension i of vectors $\mathbf{X^1}$, $\mathbf{X^2}$, …, $\mathbf{X^k}$

Gene Expression Profile



A gene expression profile: **X**=(X1,… Xi, … , Xn) – a vector of real numbers, Xi a missing data point

KNN imputation

- We are given a gene expression matrix **M**
- Let **X**=(X1, X2, …, Xi, …, Xn) be a vector in the matrix **M** with a missing value at Xi at the dimension i
- Find in the gene expression data matrix matrix vectors $\mathbf{X^1}$, $\mathbf{X^2}$, …, $\mathbf{X^k}$, such that they are the k closest vectors to X in M (in the sense of a chosen distance measure) among the vectors that do not have a missing value at dimension i
- Replace the missing value Xi with the mean (or median) of X^1i, X^2i, …, X^ki , i.e., mean (median) of the values at dimension i of vectors $\mathbf{X^1}$, $\mathbf{X^2}$, …, $\mathbf{X^k}$

Practical

- See http://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/06-07/pcmda/ for instructions