Lecture 12

Metrics, Biclustering and Dimensionality Reduction

Instructor: Ilias Tagkopoulos
iliast@ucdavis.edu

The **perfect** cluster and the **ideal** metric

"There is no single best criterion for obtaining a partition because no precise and workable definition of 'cluster' exists. Clusters can be of any arbitrary shapes and sizes in a multidimensional pattern space. Each clustering criterion imposes a certain structure on the data, and if the data happen to conform to the requirements of a particular criterion, the true clusters are recovered."

Jain, A.K. & Dubes, R.C. *Algorithms for Clustering Data.* (Prentice Hall, Englewood Cliffs, New Jersey, 1988).

Metrics

L^p norm:

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$$

L^1 – Manhattan distance / Taxicab geometry

L^2 – Euclidian norm

L^∞ – Uniform norm; Chebyshev chessboard distance

$max\{|x_i|\}$

Pearson correlation

- Measure of the correlation (linear dependence) between two variables
- Formula:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

- Or

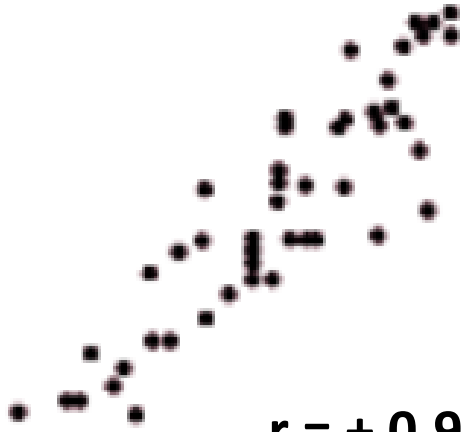
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

- Or

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

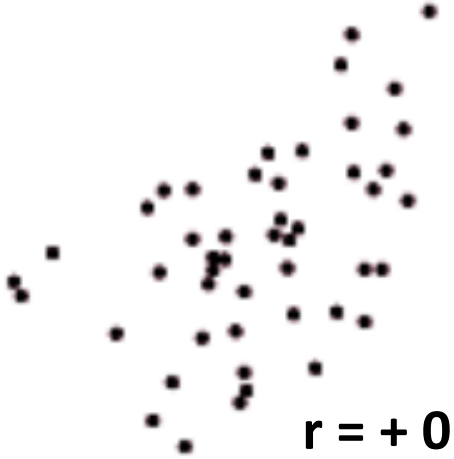
- Values -1 to +1

Pearson correlation



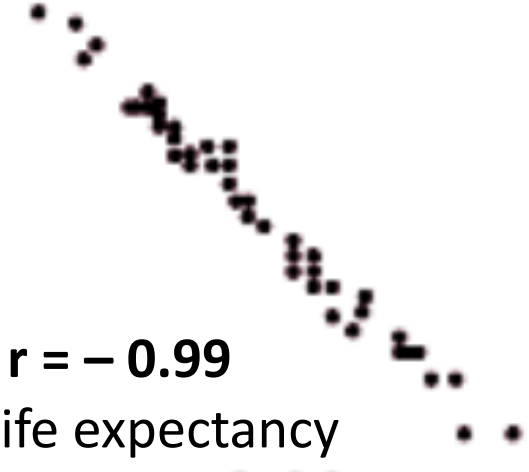
r = + 0.9

e.g: lung cancer probability *vs.* smoking level



r = + 0.5

e.g: prostate cancer probability *vs.* men's age



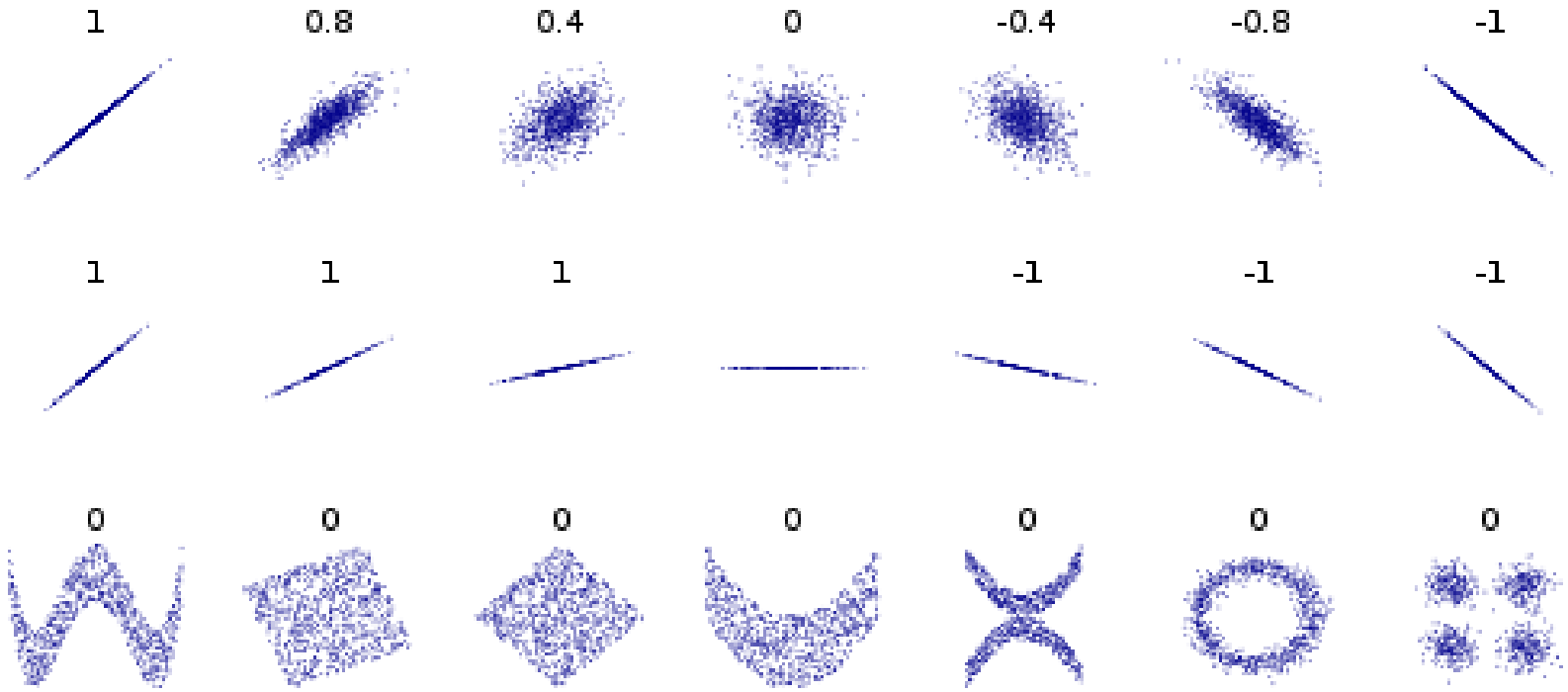
r = – 0.99

e.g: life expectancy *vs.* pancreatic cancer stage

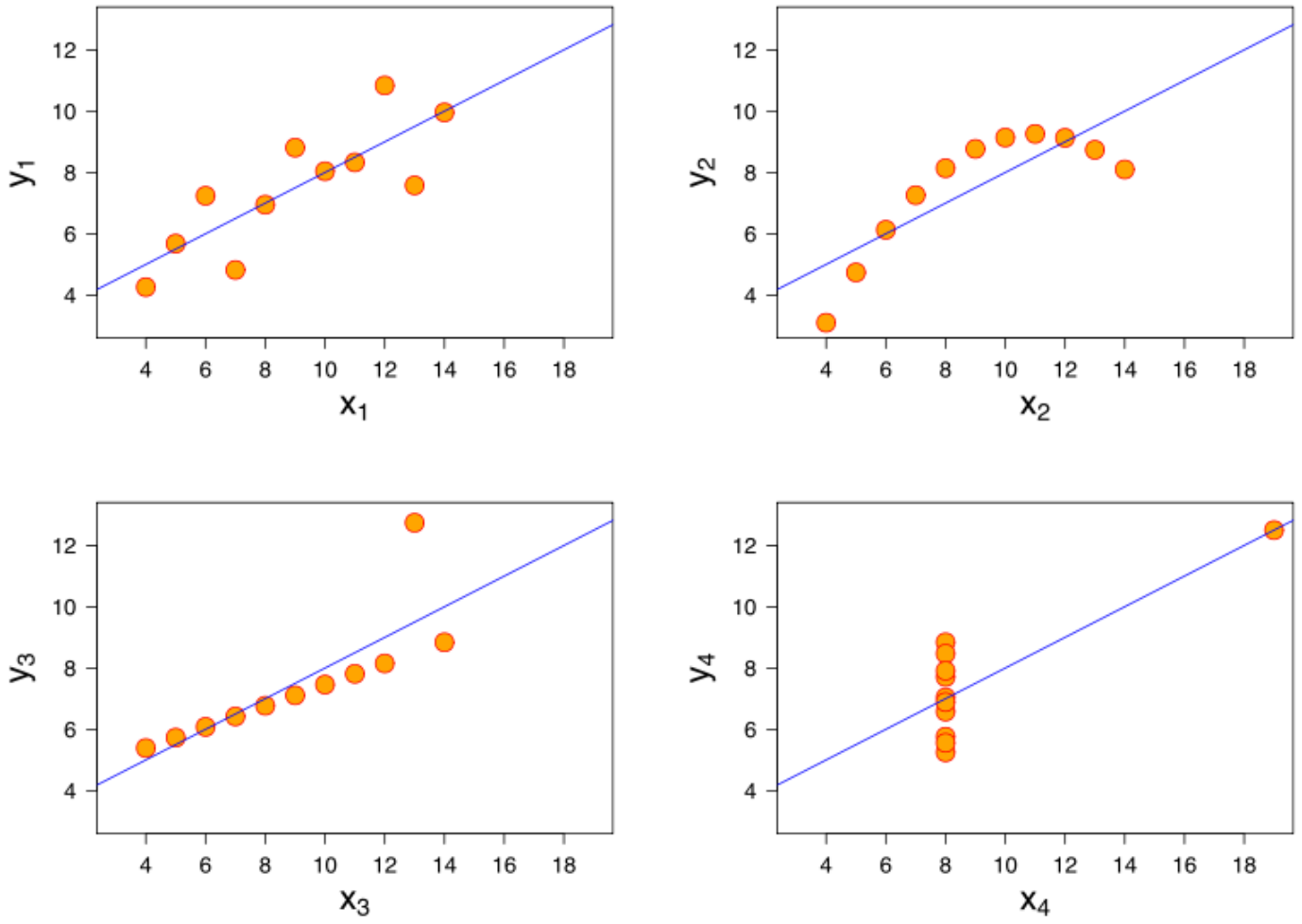
r = 0.0

e.g: lung cancer stage *vs.* eyes color

Pearson correlation



Issues with Pearson Correlation: Outliers

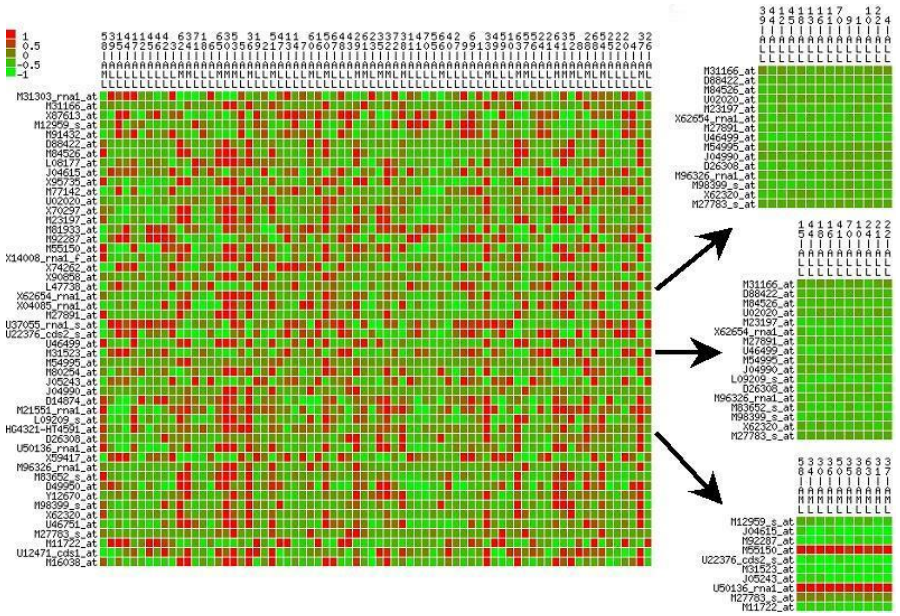


- **Anscombe's quartet:** https://en.wikipedia.org/wiki/Anscombe's_quartet

Biclustering

Simultaneous clustering in both rows and columns

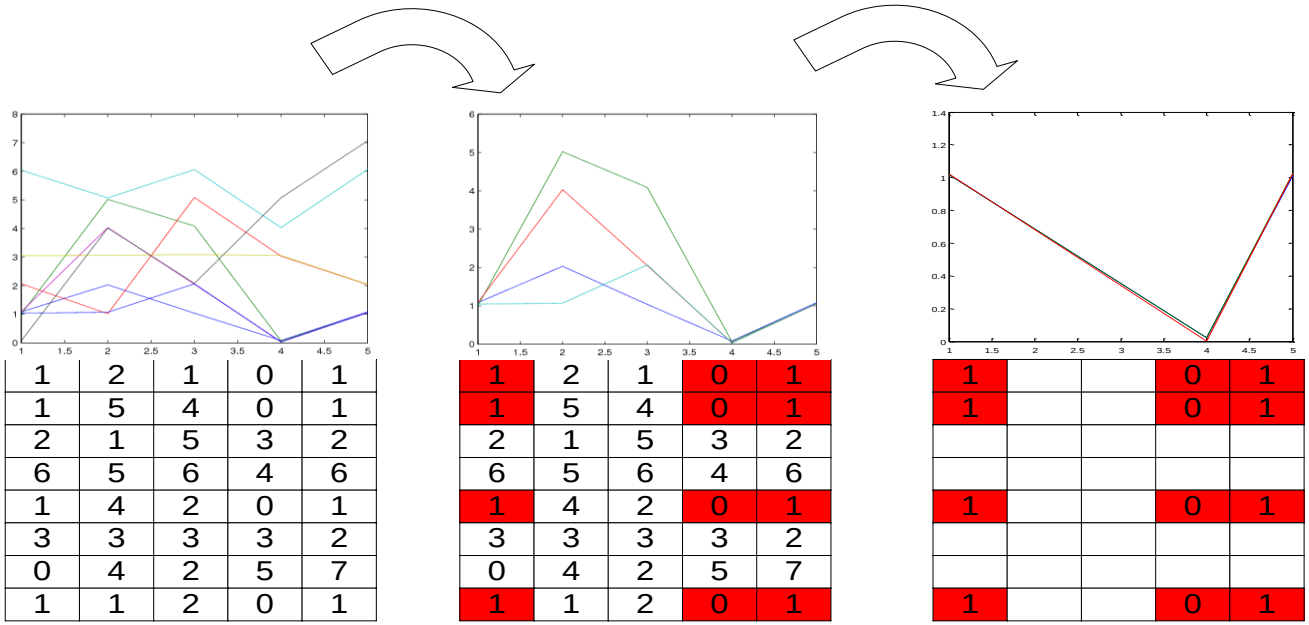
- In many cases, a subset of samples share a strong pattern over a subset of conditions.
  - E.g. Several genes that are up-regulated under stress, but also in individual conditions.
- One way to cluster them: Biclustering
  - "**Simultaneous** Clustering of both row and column sets in a data matrix". (Hartigan, 1972; Mirkin 1996; Church, 2000).



Wu CJ, Kasif S. GEMS: a web server for biclustering analysis of expression data.Nucleic Acids Res. 2005 Jul 1;33.

Clustering based on a subset of columns

- Biclustering vs. Clustering:
  - Subsets of genes may be correlated only under certain experimental conditions.
  - Local vs. Global patterns.



1	2	1	0	1
1	5	4	0	1
2	1	5	3	2
6	5	6	4	6
1	4	2	0	1
3	3	3	3	2
0	4	2	5	7
1	1	2	0	1

1	2	1	0	1
1	5	4	0	1
2	1	5	3	2
6	5	6	4	6
1	4	2	0	1
3	3	3	3	2
0	4	2	5	7
1	1	2	0	1

1			0	1
1			0	1
1			0	1
1			0	1

Heuristic solution

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

Algorithm 0 (Brute-Force Deletion and Addition).

Input: A, a matrix of real numbers, and $\delta \geq 0$, the maximum acceptable mean squared residue score.

Output: A_{IJ}, a δ-bicluster that is a submatrix of A with row set I and column set J, with a score no larger than δ.

Initialization: I and J are initialized to the gene and condition sets in the data and $A_{IJ} = A$.

Iteration:

1. Compute the score H for each possible row/column addition/deletion and choose the action that decreases H the most. If no action will decrease H, or if $H <= \delta$, return A_{IJ}.

Cheng, Church, "Biclustering of expression data", ISMB 2000

- Biclustering is an NP-complete problem
- One heuristic algorithm is normalizing on column, row and overall average:

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \left(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}\right)^2,$$

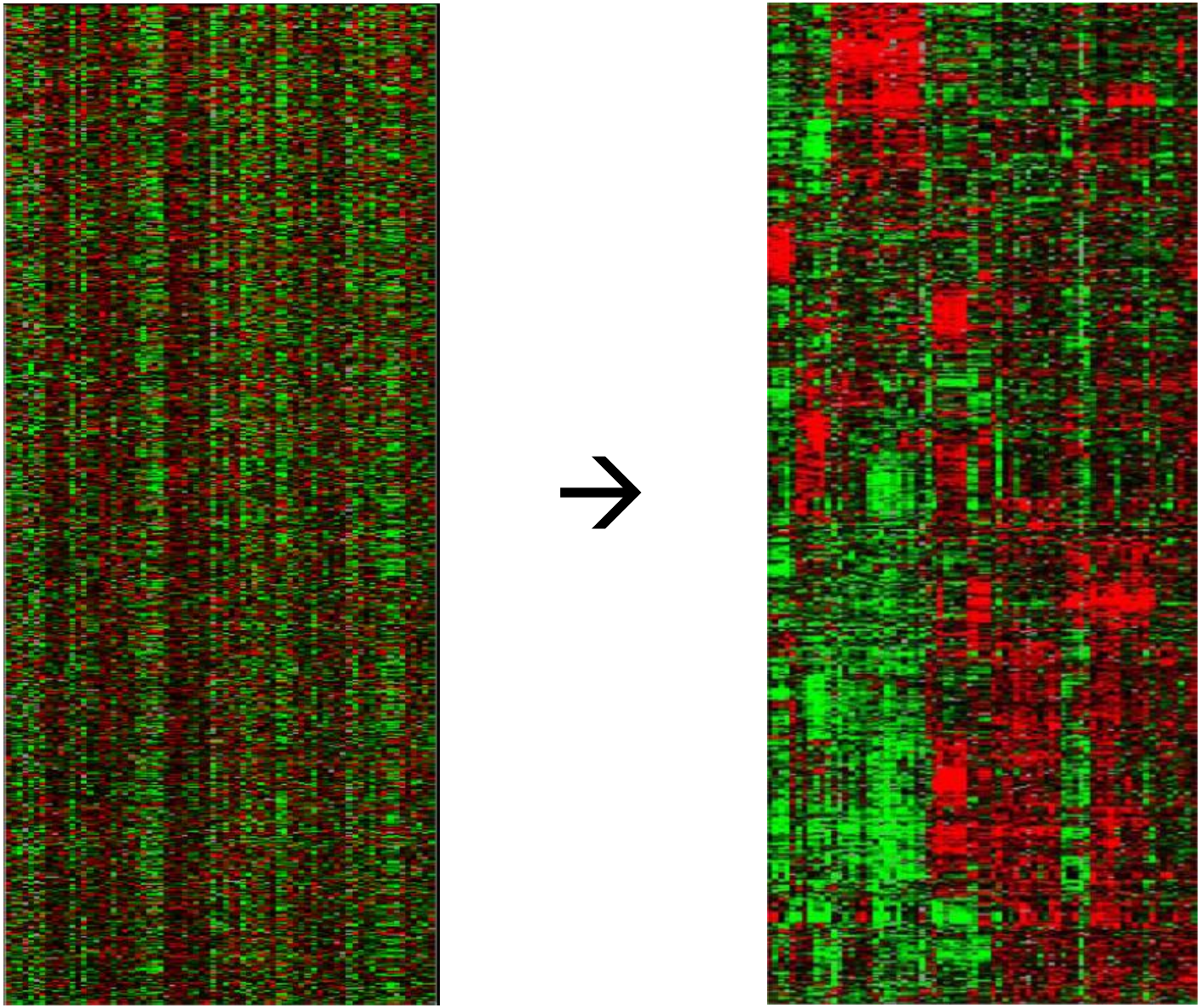
where

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}, \quad a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij},$$

and

$$a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij} = \frac{1}{|I|} \sum_{i \in I} a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{Ij}$$

Other clustering techniques: Biclustering



Measuring the statistical significance of a clustering result

- After we have clustered the data, we need to determined two characteristics:
  - How accurately our cluster represents its members
    - Measure average error (distance of members to the cluster center)
  - How statistically significant is this grouping relative to a random grouping.
    - Calculate the p-value
    - P-value can follow various distributions, but usually the hypergeometric distribution

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

Here N = 15, m=4, n=6, k=3

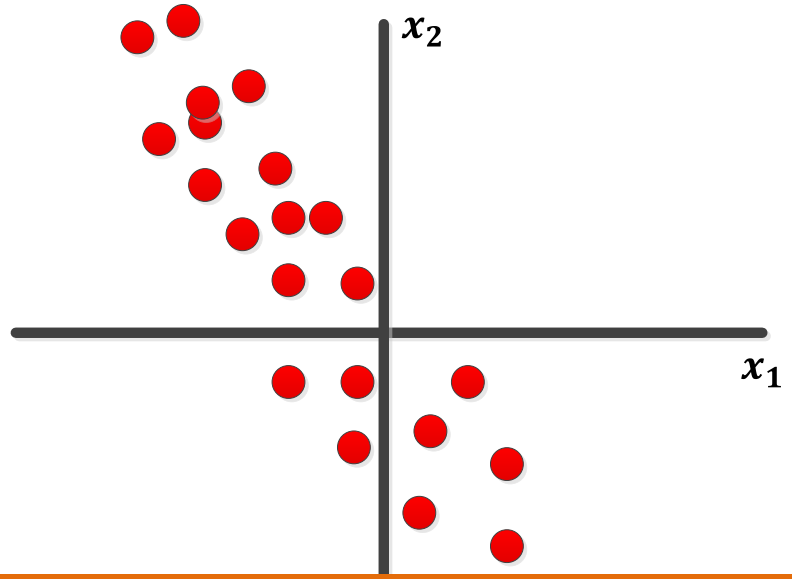
Introduction to Principal Component Analysis

Principal Component Analysis

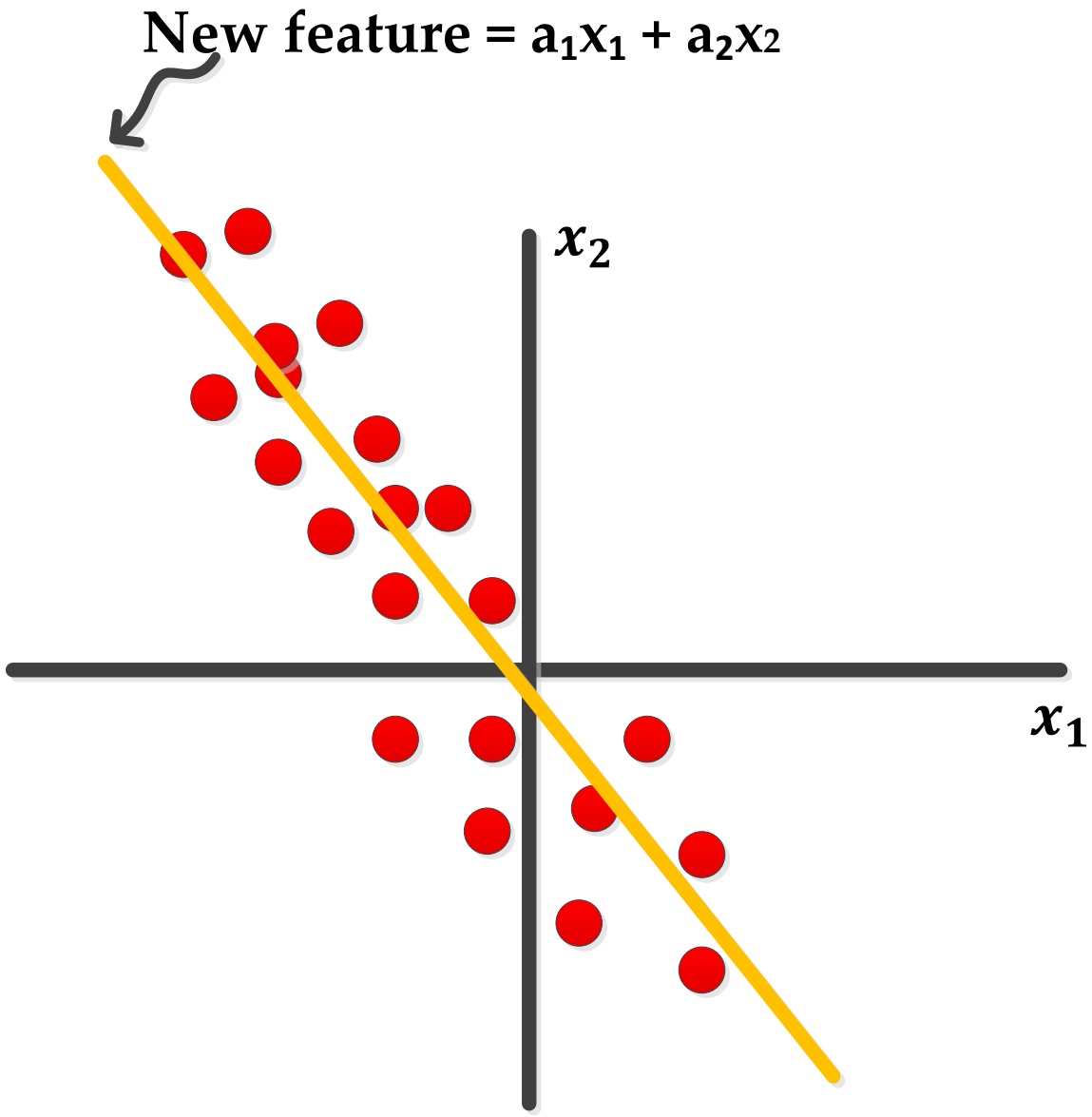
- **Principal Component Analysis (PCA)** is a statistical procedure that uses an *orthogonal transformation* to convert a set of observations of *possibly correlated variables* into a set of values of *linearly uncorrelated variables* called **principal components**.

- **Problem Formulation:** Suppose you are given *m* samples, each having *n* features. How can you reduce the number of features to k (where k<<n) while retaining the feature information?

- Solutions:
  - Maximize the variance in the data
  - Project to the line that minimizes SSE



Principal Component Analysis



Principal Component Analysis

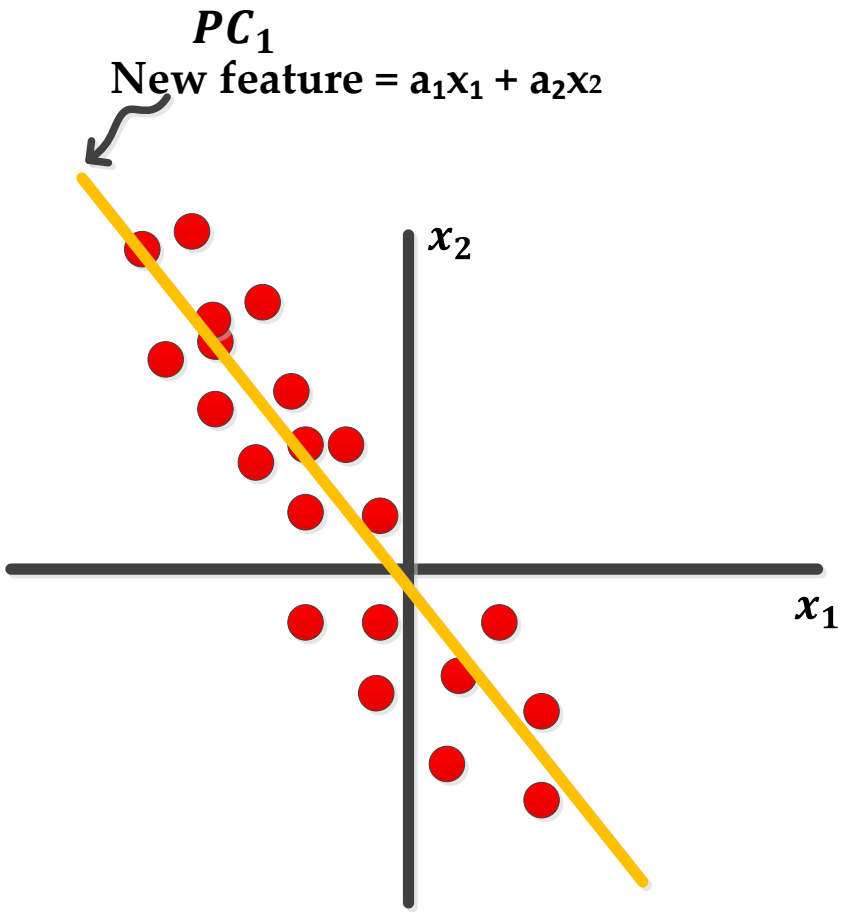
- Find a linear transformation of the original n variables that will produce new variables (Principal Components, PC) that are orthogonal (uncorrelated):

$$PC_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n$$

$$\ldots$$

$$PC_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kn}x_n$$

How ? By calculating eigenvalues.



Principal Component Analysis

- In general for n non-redundant variables, you have n eigenvalues, each with a corresponding eigenvector.

- The eigenvector **provides the direction** (line), the eigenvalue provides the **variance of the projected points to that line.**

- So you want to pick the eigenvectors with **max eigenvalues**

- **General procedure:**
  1. **Normalize data by subtracting the mean (feature mean; center the data to 0,0)**
  2. **Calculate the covariance matrix**
  3. **Find the Eigenvectors and Eigenvalues of the covariance matrix**
  4. **Pick the top K of them as the new (reduced features)**

End of Lecture 12