Clustering algorithms

Konstantinos Koutroumbas

Unit 5

– k-medoids clustering algorithms (PAM, CLARA, CLARANS)

– Probabilistic CFO clustering algorithms (EM)

koutroum@noa.gr

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

k-Medoids Algorithms

- Each cluster is represented by a vector selected among the elements of X (medoid).

- A cluster contains
  - Its medoid
  - All vectors in X that
    - Are not used as medoids in other clusters
    - Lie closer to its medoid than the medoids representing other clusters.

In the k-medoids algorithms ONLY **point representatives** can be used to represent clusters

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

k-Medoids Algorithms

Let

- Θ be the set of medoids of all clusters,
- I_{Θ} the set of indices of the points in X that constitute Θ and
- $I_{X\text{-}\Theta}$ the set of indices of the points that are not medoids.

Obtaining the set of medoids Θ that best represents the data set, X is equivalent to minimizing the following cost function

$$J(\Theta, U) = \sum_{i \in I_{X-\Theta}} \sum_{j \in I_{\Theta}} u_{ij} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

with

$$u_{ij} = \begin{cases} 1, & if\ d(x_i, x_j) = min_{q \in I_{\Theta}} d(x_i, x_q) \\ 0, & otherwise \end{cases}, \qquad i = 1, \ldots, N$$

CFO hard clustering algorithms

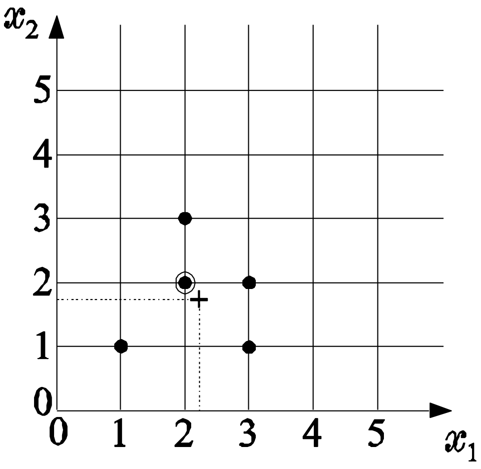
Generalized Hard Algorithmic Scheme (GHAS)

k-Medoids Algorithms

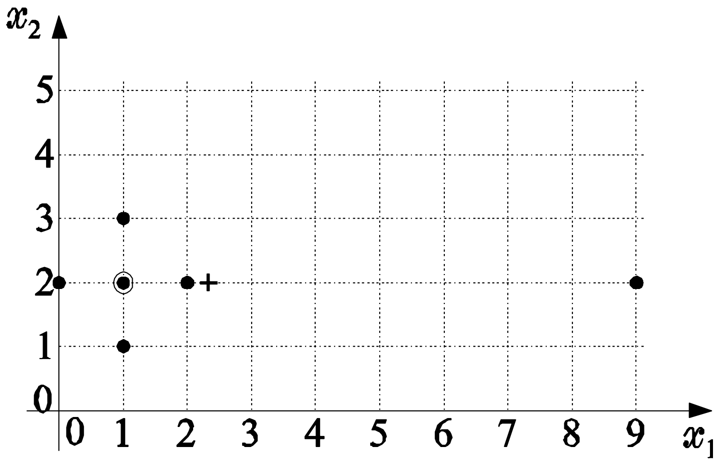
Example 3:

(a) The five-point two-dimensional set stems from the discrete domain $D = \{1,2,3,4, \ldots\} \times \{1,2,3,4, \ldots\}$. Its medoid is the circled point and its mean is the “$+$” point, which does not belong to D.

(b) In the six-point two-dimensional set , the point $(9,2)$ can be considered as an outlier. While the outlier affects significantly the mean of the set, it does not affect its medoid.



(a)



(b)

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

Representing clusters with mean values **vs** representing clusters with medoids

Mean Values	Medoids
1. Suited only for continuous domains	**1. Suited for either cont. or discrete domains**
2. Algorithms using means are sensitive to outliers	**2. Algorithms using medoids are less sensitive to outliers**
3. The mean possess a clear geometrical and statistical meaning	3. The medoid has not a clear geometrical meaning
4. Algorithms using means are not computationally demanding	4. Algorithms using medoids are more computationally demanding

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

k-Medoids Algorithms

Algorithms to be considered

- PAM (Partitioning Around Medoids)
- CLARA (Clustering LARge Applications)
- CLARANS (Clustering Large Applications based on RANdomized Search)

The PAM algorithm

- The number of clusters m is **required** *a priori*.

Definitions-preliminaries

- Two *sets* of medoids Θ and Θ', each one consisting of m elements, are called neighbors if they **share** $m-1$ elements.

- A set Θ of medoids with m elements can have $m(N-m)$ neighbors.

- Let Θ_{ij} denote the neighbor of Θ that results if $\boldsymbol{x}_j$, $j \in I_{X-\Theta}$ **replaces** $\boldsymbol{x}_i$, $i \in I_{\Theta}$.

- Let $\Delta J_{ij} = J(\Theta_{ij}, U_{ij}) - J(\Theta, U)$.

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The PAM algorithm

- *Determination of Θ that best represents the data*
  - Generate a set Θ of m medoids, randomly selected out of X.
  - (A) Determine the neighbor Θ_{qr}, $q \in I_\Theta$, $r \in I_{X-\Theta}$ among the $m(N-m)$ neighbors of Θ for which $\Delta J_{qr} = min_{i \in I_\Theta,\, j \in I_{X-\Theta}} \Delta J_{ij}$.
  - If $\Delta J_{qr} < 0$ then
    - Replace Θ by Θ_{qr}
    - Go to (A)
  - End

$$\Delta J_{qr} < 0 \Leftrightarrow J(\Theta_{qr}, U_{qr}) < J(\Theta, U)$$

- *Assignment of points to clusters*
  - Assign each $\boldsymbol{x} \in I_{X-\Theta}$ to the cluster represented by the closest to $\boldsymbol{x}$ medoid.

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The PAM algorithm

Computation of ΔJ_{ij}.

It is defined as:

$$\Delta J_{ij} = J(\Theta_{ij}, U_{ij}) - J(\Theta, U) = \sum_{s \in I_{X-\Theta_{ij}}} \sum_{t \in I_{\Theta_{ij}}} u_{st} d(\boldsymbol{x}_s, \boldsymbol{x}_t) - \sum_{s \in I_{X-\Theta}} \sum_{t \in I_{\Theta}} u_{st} d(\boldsymbol{x}_s, \boldsymbol{x}_t)$$

$$\equiv \sum_{h \in I_{X-\Theta}} C_{hij}$$

where C_{hij} is the *difference in J, resulting from the (possible) assignment of the vector $\boldsymbol{x}_h \in X - \Theta$ from the cluster it currently belongs to another, as a consequence of the replacement of $\boldsymbol{x}_i \in \Theta$ by $\boldsymbol{x}_j \in X - \Theta$*.

For the computation of C_{hij} associated with a specific each $\boldsymbol{x}_h \in X - \Theta$ it is required

- The **distance** of $\boldsymbol{x}_h$ from its **closest medoid** in Θ
- The **distance** of $\boldsymbol{x}_h$ from its **next to closest medoid** in Θ.
- The **distance** of $\boldsymbol{x}_h$ from the **newly inserted medoid** in Θ_{ij}.

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The PAM algorithm (cont.)

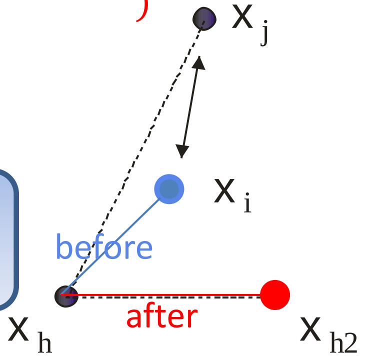
Computation of C_{hij}:

$\boldsymbol{x}_h$ **belongs** to the cluster represented by $\boldsymbol{x}_i$ ($\boldsymbol{x}_{h2}$ Θ denotes the second closest to $\boldsymbol{x}_h$ **representative**) **and** $d(\boldsymbol{x}_h, \boldsymbol{x}_j) \geq d(\boldsymbol{x}_h, \boldsymbol{x}_{h2})$. Then

$$C_{hi\,j} = d(\boldsymbol{x}_h, \boldsymbol{x}_{h2}) - d(\boldsymbol{x}_h, \boldsymbol{x}_i) \geq 0$$

Contribution of $\boldsymbol{x}_h$ to $J(\Theta_{ij}, U_{ij})$

Contribution of $\boldsymbol{x}_h$ to $J(\Theta, U)$

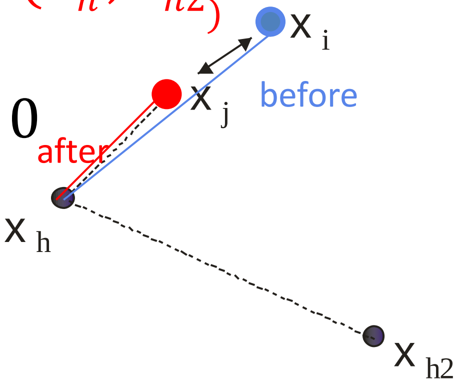


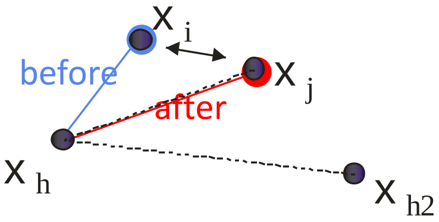
$\boldsymbol{x}_h$ **belongs** to the cluster represented by $\boldsymbol{x}_i$ ($\boldsymbol{x}_{h2}$ Θ denotes the second closest to $\boldsymbol{x}_h$ representative) **and** $d(\boldsymbol{x}_h, \boldsymbol{x}_j) \leq d(\boldsymbol{x}_h, \boldsymbol{x}_{h2})$. Then

$$C_{hij} = d(\boldsymbol{x}_h, \boldsymbol{x}_j) - d(\boldsymbol{x}_h, \boldsymbol{x}_i)\ (><)\ 0$$

Contribution of $\boldsymbol{x}_h$ to $J(\Theta_{ij}, U_{ij})$

Contribution of $\boldsymbol{x}_h$ to $J(\Theta, U)$





CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The PAM algorithm (cont.)

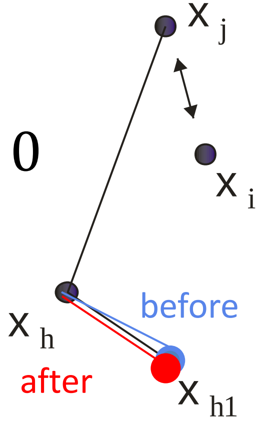
Computation of C_{hij} (cont.):

$\boldsymbol{x}_h$ is not represented by $\boldsymbol{x}_i$ ($\boldsymbol{x}_{h1}$ denotes the closest to $\boldsymbol{x}_h$ medoid) **and** $d(\boldsymbol{x}_h, \boldsymbol{x}_{h1_{)}} \leq d(\boldsymbol{x}_h, \boldsymbol{x}_j)$. Then

$$C_{hij} = d(\boldsymbol{x}_h, \boldsymbol{x}_j) - d(\boldsymbol{x}_h, \boldsymbol{x}_{h1}) = 0$$

Contribution of $\boldsymbol{x}_h$ to $J(\Theta_{ij}, U_{ij})$

Contribution of $\boldsymbol{x}_h$ to $J(\Theta, U)$



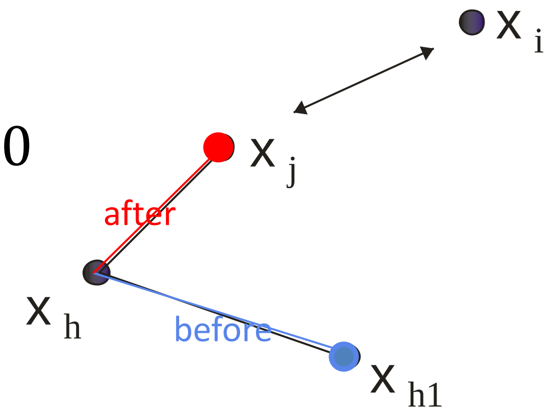
Most frequent scenario

$\boldsymbol{x}_h$ is not represented by $\boldsymbol{x}_i$ ($\boldsymbol{x}_{h1}$ denotes the closest to $\boldsymbol{x}_h$ medoid) **and** $d(\boldsymbol{x}_h, \boldsymbol{x}_{h1_{)}} > d(\boldsymbol{x}_h, \boldsymbol{x}_j)$. Then

$$C_{hij} = d(\boldsymbol{x}_h, \boldsymbol{x}_j) - d(\boldsymbol{x}_h, \boldsymbol{x}_{h1}) < 0$$

Contribution of $\boldsymbol{x}_h$ to $J(\Theta_{ij}, U_{ij})$

Contribution of $\boldsymbol{x}_h$ to $J(\Theta, U)$



CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The PAM algorithm (cont.)

Remarks:

- Experimental results show the PAM works satisfactorily with small data sets.

- Its computational complexity is $O(m(N-m)^2)$. Unsuitable for large data sets.

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The PAM algorithm (Example)

Data set: $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6\}$, with

$\boldsymbol{x}_1 = [0{,}3]^T, \boldsymbol{x}_2 = [1{,}3]^T, \boldsymbol{x}_3 = [2{,}3]^T, \boldsymbol{x}_4 = [0{,}0]^T, \boldsymbol{x}_5 = [1{,}0]^T, \boldsymbol{x}_1 = [2{,}0]^T.$

Set of medoids: $\Theta = \{\boldsymbol{x}_4, \boldsymbol{x}_5\}$

Computation of $J(\Theta, U)$ (**Squared Euclidean distance** is considered):

$$\boldsymbol{x}_1 \longrightarrow d(\boldsymbol{x}_1, \boldsymbol{x}_4) = 9 < 10 = d(\boldsymbol{x}_1, \boldsymbol{x}_5) \longrightarrow u_{14} = 1, u_{15} = 0$$

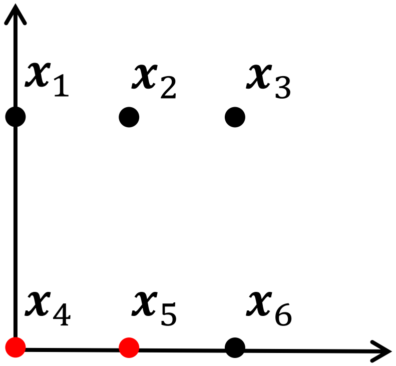
$$\boldsymbol{x}_2 \longrightarrow d(\boldsymbol{x}_2, \boldsymbol{x}_4) = 10 > 9 = d(\boldsymbol{x}_2, \boldsymbol{x}_5) \longrightarrow u_{24} = 0, u_{25} = 1$$

$$\boldsymbol{x}_3 \longrightarrow d(\boldsymbol{x}_3, \boldsymbol{x}_4) = 13 > 10 = d(\boldsymbol{x}_3, \boldsymbol{x}_5) \longrightarrow u_{34} = 0, u_{35} = 1$$

$$\boldsymbol{x}_4 \longrightarrow d(\boldsymbol{x}_4, \boldsymbol{x}_4) = 0 < 1 = d(\boldsymbol{x}_4, \boldsymbol{x}_5) \longrightarrow u_{44} = 1, u_{45} = 0$$

$$\boldsymbol{x}_5 \longrightarrow d(\boldsymbol{x}_5, \boldsymbol{x}_4) = 1 > 0 = d(\boldsymbol{x}_5, \boldsymbol{x}_5) \longrightarrow u_{54} = 0, u_{55} = 1$$

$$\boldsymbol{x}_6 \longrightarrow d(\boldsymbol{x}_6, \boldsymbol{x}_4) = 2 > 1 = d(\boldsymbol{x}_6, \boldsymbol{x}_5) \longrightarrow u_{64} = 0, u_{65} = 1$$

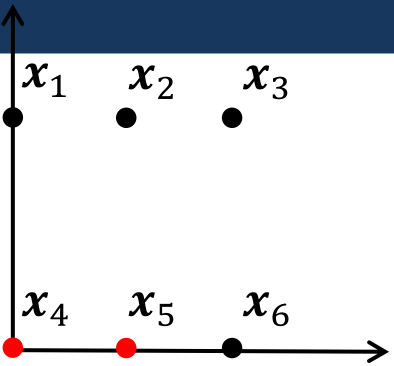


$$J(\Theta, U) = \begin{array}{ll} u_{14} d(\boldsymbol{x}_1, \boldsymbol{x}_4) + & u_{15} d(\boldsymbol{x}_1, \boldsymbol{x}_5) + \\ u_{24} d(\boldsymbol{x}_1, \boldsymbol{x}_4) + & u_{25} d(\boldsymbol{x}_1, \boldsymbol{x}_5) + \\ u_{34} d(\boldsymbol{x}_1, \boldsymbol{x}_4) + & u_{35} d(\boldsymbol{x}_1, \boldsymbol{x}_5) + \\ u_{44} d(\boldsymbol{x}_1, \boldsymbol{x}_4) + & u_{45} d(\boldsymbol{x}_1, \boldsymbol{x}_5) + \\ u_{54} d(\boldsymbol{x}_1, \boldsymbol{x}_4) + & u_{55} d(\boldsymbol{x}_1, \boldsymbol{x}_5) + \\ u_{64} d(\boldsymbol{x}_1, \boldsymbol{x}_4) + & u_{65} d(\boldsymbol{x}_1, \boldsymbol{x}_5) \end{array} = \begin{array}{ll} 1 \cdot 9 + & 0 \cdot 10 + \\ 0 \cdot 10 + & 1 \cdot 9 + \\ 0 \cdot 13 + & 1 \cdot 10 + \\ 1 \cdot 0 + & 0 \cdot 1 + \\ 0 \cdot 1 + & 1 \cdot 0 + \\ 0 \cdot 2 + & 1 \cdot 1 \end{array} = \mathbf{29}$$

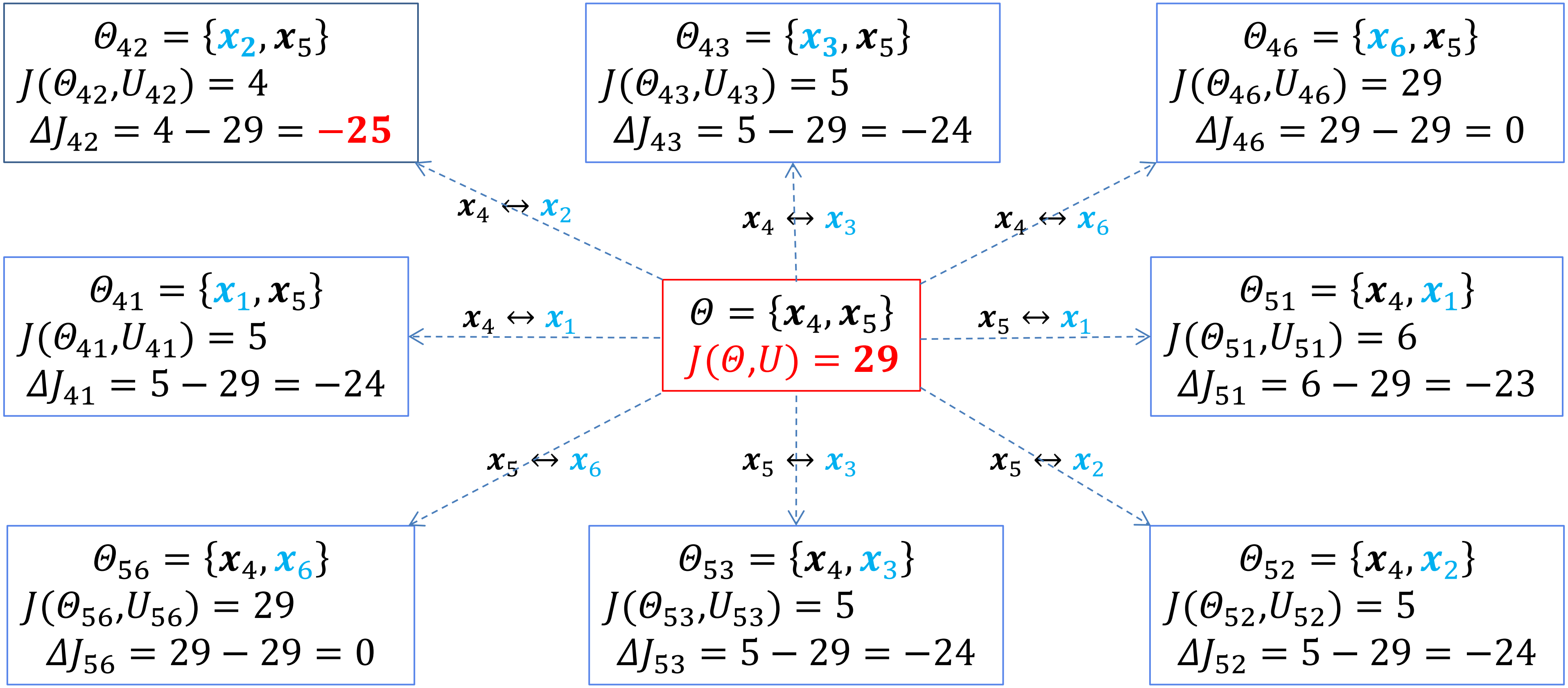
CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The PAM algorithm (Example)



Data set: $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6\}$, with
$\boldsymbol{x}_1 = [0,3]^T, \boldsymbol{x}_2 = [1,3]^T, \boldsymbol{x}_3 = [2,3]^T, \boldsymbol{x}_4 = [0,0]^T, \boldsymbol{x}_5 = [1,0]^T, \boldsymbol{x}_1 = [2,0]^T.$
Set of medoids: $\Theta = \{\boldsymbol{x}_4, \boldsymbol{x}_5\}$



$\Theta_{42} = \{\boldsymbol{x}_2, \boldsymbol{x}_5\}$
$J(\Theta_{42}, U_{42}) = 4$
$\Delta J_{42} = 4 - 29 = \mathbf{-25}$

$\Theta_{43} = \{\boldsymbol{x}_3, \boldsymbol{x}_5\}$
$J(\Theta_{43}, U_{43}) = 5$
$\Delta J_{43} = 5 - 29 = -24$

$\Theta_{46} = \{\boldsymbol{x}_6, \boldsymbol{x}_5\}$
$J(\Theta_{46}, U_{46}) = 29$
$\Delta J_{46} = 29 - 29 = 0$

$\boldsymbol{x}_4 \leftrightarrow \boldsymbol{x}_2$

$\boldsymbol{x}_4 \leftrightarrow \boldsymbol{x}_3$

$\boldsymbol{x}_4 \leftrightarrow \boldsymbol{x}_6$

$\Theta_{41} = \{\boldsymbol{x}_1, \boldsymbol{x}_5\}$
$J(\Theta_{41}, U_{41}) = 5$
$\Delta J_{41} = 5 - 29 = -24$

$\boldsymbol{x}_4 \leftrightarrow \boldsymbol{x}_1$

$\Theta = \{\boldsymbol{x}_4, \boldsymbol{x}_5\}$
$J(\Theta, U) = \mathbf{29}$

$\boldsymbol{x}_5 \leftrightarrow \boldsymbol{x}_1$

$\Theta_{51} = \{\boldsymbol{x}_4, \boldsymbol{x}_1\}$
$J(\Theta_{51}, U_{51}) = 6$
$\Delta J_{51} = 6 - 29 = -23$

$\boldsymbol{x}_5 \leftrightarrow \boldsymbol{x}_6$

$\boldsymbol{x}_5 \leftrightarrow \boldsymbol{x}_3$

$\boldsymbol{x}_5 \leftrightarrow \boldsymbol{x}_2$

$\Theta_{56} = \{\boldsymbol{x}_4, \boldsymbol{x}_6\}$
$J(\Theta_{56}, U_{56}) = 29$
$\Delta J_{56} = 29 - 29 = 0$

$\Theta_{53} = \{\boldsymbol{x}_4, \boldsymbol{x}_3\}$
$J(\Theta_{53}, U_{53}) = 5$
$\Delta J_{53} = 5 - 29 = -24$

$\Theta_{52} = \{\boldsymbol{x}_4, \boldsymbol{x}_2\}$
$J(\Theta_{52}, U_{52}) = 5$
$\Delta J_{52} = 5 - 29 = -24$

It is $\Delta J_{42} = min_{i \in I_{\Theta},\, j \in I_{X-\Theta}} \Delta J_{ij} = -25 < 0$
Thus, according to **PAM**, Θ will be **replaced** by Θ_{42}.

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The CLARA algorithm

- It is more suitable for large data sets.
- The strategy:
  - **Draw** randomly a sample X' of size N' from the entire data set.
  - **Run** the PAM algorithm to **determine** Θ' that best represents X'.
  - Use Θ' in the place of Θ to represent the entire data set X.
- The rationale:
  - Assuming that X' has been selected in a way representative of the statistical distribution of the data points in X, Θ' is expected to be a good approximation of Θ, which would have been produced if PAM were run on the entire X.
- The algorithm:
  - Draw s sample subsets of size N' from X, denoted by $X'_1, \dots, X'_s$ (typically $s = 5, N' = 40 + 2m$).
  - Run PAM on each one of them and identify $\Theta'_1, \dots, \Theta'_s$.
  - Choose the set Θ'_j that minimizes

$$J(\Theta, U) = \sum_{i \in I_{X-\Theta'}} \sum_{j \in I_{\Theta'}} u_{ij} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

based on the entire data set X.

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The CLARANS algorithm

- It is more suitable for large data sets.
- It follows the philosophy of PAM with the difference that only a randomly selected fraction $q(< m(N-m))$ of the neighbors of the current medoid set is considered.
- It performs several runs (s) starting from different initial choices for Θ.

The algorithm:

```
– For $i = 1$ to $s$
    o **Initialize** randomly $\Theta$.
    o (A) **Select** randomly $q$ neighbors of $\Theta$.
    o For $j = 1$ to $q$
       * **If** the present neighbor of $\Theta$ is **better** than $\Theta$ (in terms of $J(\Theta, U)$) then
          -- **Set** $\Theta$ equal to its neighbor
          -- Go to (A)
       * End If
    o End For
    o Set $\Theta^i = \Theta$
– End For
– **Select** the best $\Theta^i$ with respect to $J(\Theta, U)$.
– Based on $\Theta^i$, assign each $\boldsymbol{x} \in X - \Theta$ to the cluster whose representative is closest to $\boldsymbol{x}$
```

CFO hard clustering algorithms

Generalized Hard Algorithmic Scheme (GHAS)

The CLARANS algorithm (cont.)

Remarks:

- CLARANS **depends** on q and s. Typically, $s = 2$ and
  $$q = \max(0.125m(N-m), 250)$$
- As q approaches $m(N-m)$ CLARANS approaches PAM and the complexity increases.
- CLARANS can also be described in terms of graph theory concepts.
- CLARANS unravels better quality clusters than CLARA.
- In some cases, CLARA is significantly faster than CLARANS.
- CLARANS retains its quadratic computational nature and thus it is not appropriate for very large data sets.

Probability and statistics: a brief review

Random variable (RV): It models the output of an experiment.

RV types:

- Discrete
- continuous

Discrete random variables:

- A **discrete RV** *x* can take any value *x* from a finite or countably infinite set *X.*
- *X*: sample space or state space.
- **Event:** Any subset of *X.*
- **Elementary** or **simple event**: A single element subset of *X*.
- **Example:** Consider the die roll experiment. X={1,2,3,4,5,6}
- Events: “Odd number”, “number>3”, “2”, “5”

Elementary events

Probability and statistics: a brief review

Discrete random variables (cont.)**:**

•**Notation:** Probability of the event $x\!=\!x \in X$: $P(x\!=\!x) \equiv P(x)$

•*P*(.):A function called probability mass function (pmf) satisfying

- ✓ $P(x) \geq 0, \forall x \in X$
- ✓ $\sum_{x \in X} P(x) = 1$

Probability and statistics: a brief review

Discrete random variables (cont.):

The case of more than one random variables: Definitions

Discrete RV	x	y
Sample space	$X=\{x_1,\ldots,x_{nx}\}$	$Y=\{y_1,\ldots,y_{ny}\}$

Joint probability: $P(x_i, y_j) \equiv P(x=x_i \text{ AND } y=y_j)$

- It corresponds to the case where x takes the value x_i **AND** y takes the value y_j, **simultaneously**.

Marginal probabilities: $P(x_i) \equiv P(x=x_i)$, $P(y_j) = P(y=y_j)$

- This terminology is used only when more than one rvs are involved.

Conditional probability: $P(x_i | y_j) \equiv P(x=x_i \mid y=y_j) = P(x_i,y_j) / P(y_j)$

- It corresponds to the case where x takes the value x_i **given that** y takes the value y_j.

Probability and statistics: a brief review

Discrete random variables (cont.):

The case of more than one variables: Properties

Discrete RV	x	y
Sample space	$X=\{x_1,\ldots,x_{nx}\}$	$Y=\{y_1,\ldots,y_{ny}\}$

Sum rule: $P(x)=\sum_{y\in Y}P(x,y), \quad \forall x\in X$

Product rule: $P(x,y)=P(x\,|\,y)P(y)$

Statistical independence: $P(x,y)=P(x)P(y)$

A consequence: $P(x\,|\,y)=P(x) \quad P(y\,|\,x)=P(y)$

Bayes rule: $$P(y\,|\,x)=\frac{P(x\,|\,y)P(y)}{P(x)}$$

It plays a key role in ML.

or $$P(y\,|\,x)=\frac{P(x\,|\,y)P(y)}{\sum_{y\in Y}P(x\,|\,y)P(y)}$$

Probability and statistics: a brief review

Continuous random variables:

•A **continuous RV** x can take any value $x \in R$.

•Sample space or state space: R

Corresponds to the probability mass function from the discrete case.

•**Events:** $\{x \le x\}$, $\{x_1 < x \le x_2\}$, $\{x \ge x\}$

•**Cumulative distribution function** (**cdf**): $F_x(x) = P(x \le x)$

It assigns “mass” to events.

•It is $F_x(\infty) = P(x < \infty) = 1$

•**Probability of events** in terms of **cdf**:

- $P(x \le x) = F_x(x)$
- $P(x_1 < x \le x_2) = P(x \le x_2) - P(x \le x_1) = F_x(x_2) - F_x(x_1)$
- $P(x \ge x) = = P(x \le \infty) - P(x \le x) = 1 - P(x \le x) = 1 - F_x(x)$

Probability and statistics: a brief review

Continuous random variables (cont.)**:**

- **Assumption:** $F_x(x)$ is *continuous* and *differentiable*.

- **Probability density function (pdf)**:

$$p_x(x) = \frac{dF_x(x)}{dx}$$

It assigns “mass” to values.

- **cdf** in terms of **pdf:**

$$F_x(x) = \int_{-\infty}^{x} p_x(z)dz$$

- **Probability of events** in terms of **pdf**:

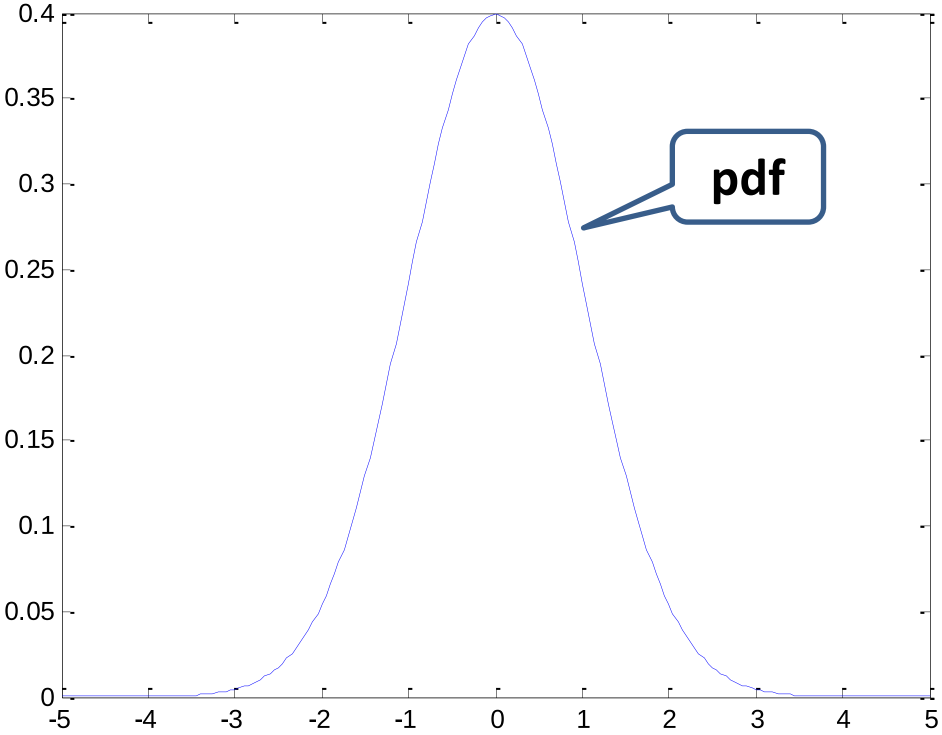
➢ $P(x \le x) = F_x(x) = \int_{-\infty}^{x} p_x(z)dz$

➢ $P(x_1 < x \le x_2) = P(x \le x_2) - P(x \le x_1) = F_x(x_2) - F_x(x_1) = \int_{x_1}^{x_2} p_x(x)dx$

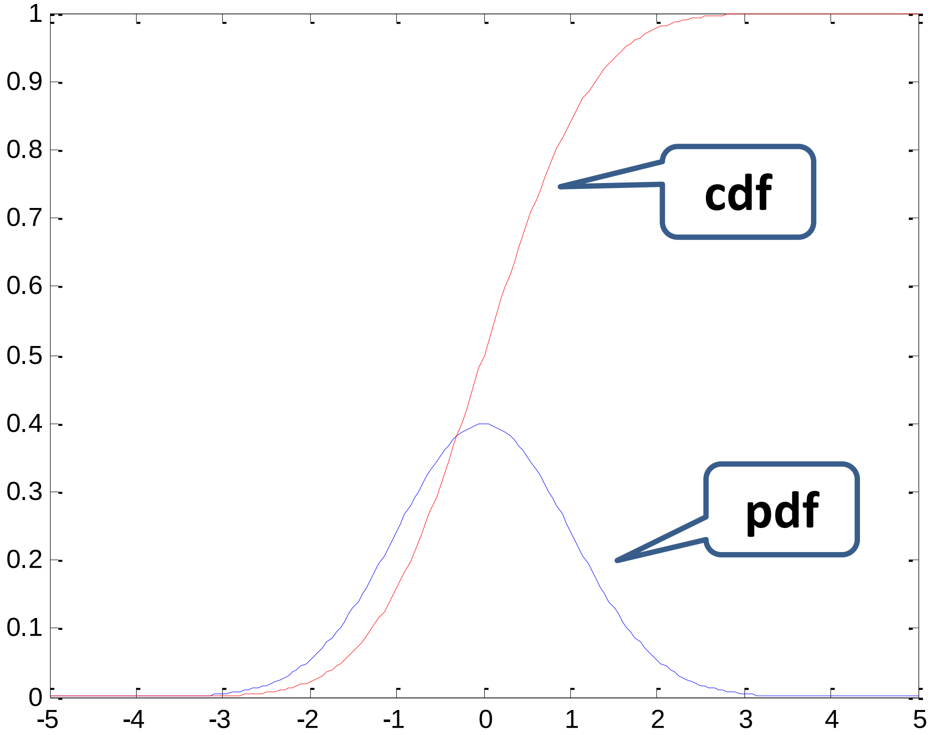
➢ $P(x \ge x) = = P(x \le \infty) - P(x \le x) = 1 - P(x \le x) = 1 - F_x(x) = \int_{-\infty}^{x} p_x(z)dz$

Probability and statistics: a brief review

Continuous random variables (cont.):



pdf



cdf

pdf

Probability and statistics: a brief review

Continuous random variables (cont.):

- Since $P(-\infty<\mathrm{x}<+\infty)=1$ it is: $\int_{-\infty}^{+\infty} p_{\mathrm{x}}(x)dx=1$

- It is $P(x<\mathrm{x}\leq x+\Delta x)=\int_{x}^{x+\Delta x} p_{\mathrm{x}}(z)dz \approx p_{\mathrm{x}}(x)\Delta x$

As $\Delta x \to 0$, $P(x<\mathrm{x}<x+\Delta x)=P(\mathrm{x}=x)=0$.

The probability of a continuous rv to take a single value is zero.

The case of more than one variables:

Continuous RV	x	y
Sample space	R	R

NOTE: All rules stated for the probability mass function in the discrete case are stated for the pdf in the continuous case.

Product rule

$$p(x,y)=p(x\,|\,y)\,p(y)$$

We drop the name of rv from the subscript of *p*.

Sum rule

$$p(x)=\int_{-\infty}^{+\infty} p(x,y)dy$$

Probability and statistics: a brief review

Useful quantities related to (continuous) rvs:

For **discrete** rv's, the integrals become summations.

- **Mean** (**expected**) **value** of a **rv** *x*: $\mathrm{E}[\mathrm{x}] = \int_{-\infty}^{+\infty} xp(x)dx$

- **Variance** of a **rv** *x* : $\sigma_{\mathrm{x}}^2 = \int_{-\infty}^{+\infty} (x - \mathrm{E}[\mathrm{x}])^2 p(x)dx = \mathrm{E}[(\mathrm{x} - \mathrm{E}(\mathrm{x}))^2]$

- **Mean** (**expected**) **value** of a **function** of an **rv** *x* : $\mathrm{E}[f(\mathrm{x})] = \int_{-\infty}^{+\infty} f(x)p(x)dx$

- **Mean** of a **function** of two **rv's** ***x*, *y***: $\mathrm{E}_{\mathrm{x,y}}[f(\mathrm{x, y})] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)p(x,y)dxdy$

- **Conditional mean** of an rv *y* given *x = x*: $\mathrm{E}[\mathrm{y} \mid x] = \int_{-\infty}^{+\infty} yp(y \mid x)dy$

- It is $\mathrm{E}_{\mathrm{x,y}}[f(\mathrm{x, y})] = E_{\mathrm{x}}[E_{\mathrm{y|x}}[f(\mathrm{x, y})]]$

- **Covariance** between two **rvs** *x* and *y*: $\mathrm{cov}(\mathrm{x, y}) = \mathrm{E}[(\mathrm{x} - \mathrm{E}[\mathrm{x}])(\mathrm{y} - \mathrm{E}[\mathrm{y}])]$

- **Correlation** between two **rv's** *x* and *y*: $r_{\mathrm{xy}} \equiv \mathrm{E}(\mathrm{xy}) = \mathrm{cov}(x, y) + \mathrm{E}[\mathrm{x}]\mathrm{E}[\mathrm{y}]$

- **Correlation coefficient** $r_{xy} = \frac{E[x-E[x])(y-E[y])]}{\sigma_x\sigma_y}$

Probability and statistics: a brief review

Random vectors

- A **collection** of **rvs:** $\boldsymbol{x}=[x_1,x_2,\ldots x_l]^T$

- **Probability density function** (**pdf**) of $\boldsymbol{x}$: The joint pdf of $x_1,x_2,\ldots x_l$.

  $p(\boldsymbol{x})=p(x_1,x_2,\ldots x_l)$

- **Covariance matrix** of $\boldsymbol{x}$:

$$\mathrm{cov}(\mathbf{x}) = \mathrm{E}[(\mathbf{x}-\mathrm{E}[\mathbf{x}])(\mathbf{x}-\mathrm{E}[\mathbf{x}])^{\mathrm{T}}] = \begin{bmatrix} \mathrm{cov}(\mathrm{x}_1,\mathrm{x}_1) & \cdots & \mathrm{cov}(\mathrm{x}_1,\mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathrm{x}_l,\mathrm{x}_1) & \cdots & \mathrm{cov}(\mathrm{x}_l,\mathrm{x}_l) \end{bmatrix}$$

- **Correlation matrix** of $\boldsymbol{x}$: $R_{\mathbf{x}} = \mathrm{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \begin{bmatrix} \mathrm{E}(\mathrm{x}_1\mathrm{x}_1) & \cdots & \mathrm{E}(\mathrm{x}_1\mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \mathrm{E}(\mathrm{x}_l\mathrm{x}_1) & \cdots & \mathrm{E}(\mathrm{x}_l\mathrm{x}_l) \end{bmatrix}$

- It is $R_{\mathbf{x}} \equiv \mathrm{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \mathrm{cov}(\mathbf{x}) + \mathrm{E}[\mathbf{x}]\mathrm{E}[\mathbf{x}^{\mathrm{T}}]$

Exercise: Prove this identity

Probability and statistics: a brief review

Random vectors (cont.)

Exercise: Prove these statements

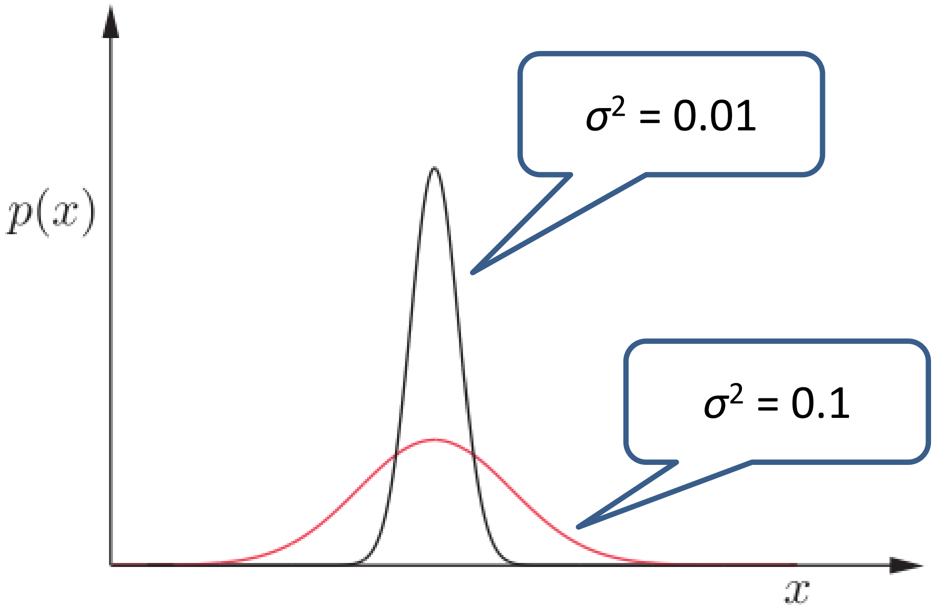
- **Remark:** Both $R_{\boldsymbol{x}}$ and cov($\boldsymbol{x}$) are symmetric and positive definite lxl matrices.

A square matrix A is symmetric iff $A^{\mathrm{T}}=A$.

A square matrix A is positive definite iff $\boldsymbol{z}^{\mathrm{T}}A\boldsymbol{z}>0,\ \forall \boldsymbol{z}\in \mathrm{R}^{l}$.

Probability and statistics: a brief review

- **One dim. normal (Gaussian) distribution** $x \sim N(\mu,\sigma^2)$ **or** $N(x|\mu,\sigma^2)$ **:**
  - Sample space: R
  - It is
    - $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$
    - $E[x]=\mu$
    - $\sigma_x^2 = \sigma^2$.



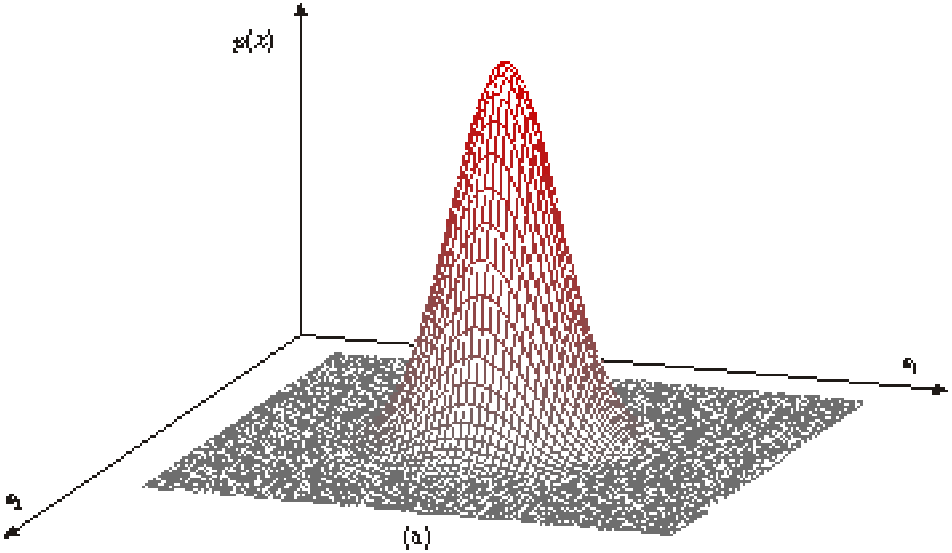
Probability and statistics: a brief review

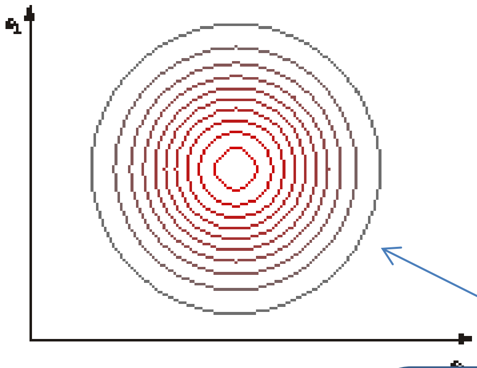
- **Multi dim. normal (Gaussian) distribution $\boldsymbol{x}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ or $N(\boldsymbol{x}|\ \boldsymbol{\mu},\boldsymbol{\Sigma})$:**

  - Sample space: R^l
  - It is
    - $p(\boldsymbol{x}) = \dfrac{1}{(2\pi)^{l/2}\,|\Sigma|^{1/2}}\exp\left(-\dfrac{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$
    - E[**x**]=$\boldsymbol{\mu}$
    - *cov*(**x**) = *Σ*.

Probability and statistics: a brief review

- **Multi dim. normal (Gaussian) distribution $\boldsymbol{x} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ or $N(\boldsymbol{x}|\ \boldsymbol{\mu},\boldsymbol{\Sigma})$:**

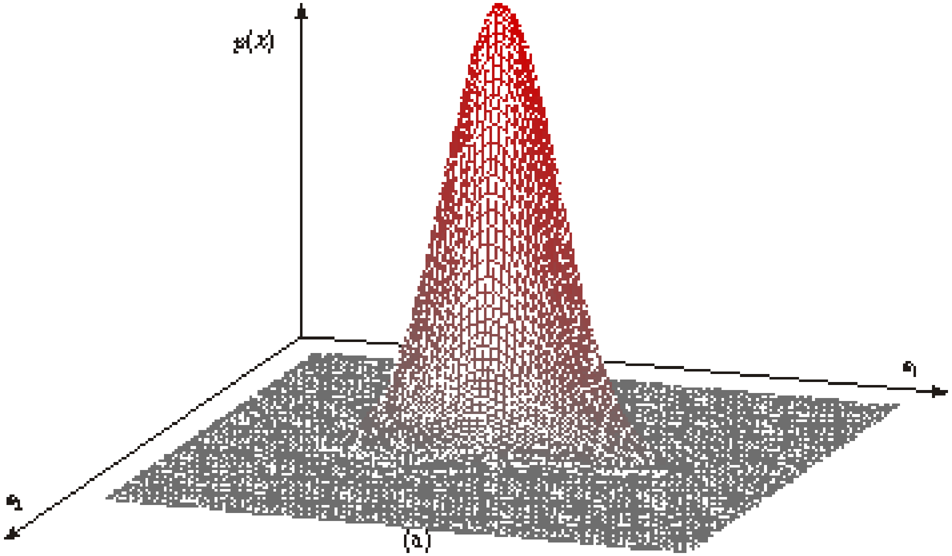


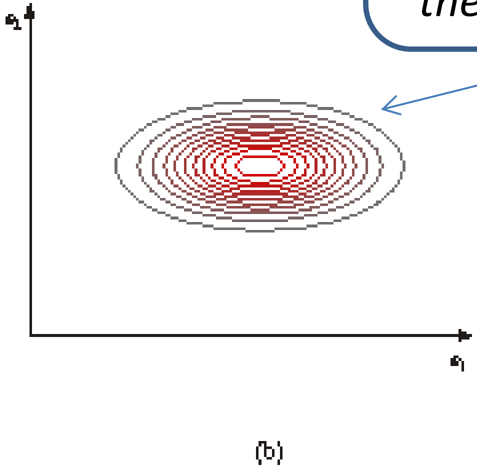


Σ: diagonal with equal diagonal entries

Isovalued curves:

- $(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\mu)=const.$
- *All points on it share the value $p(\boldsymbol{x})$*

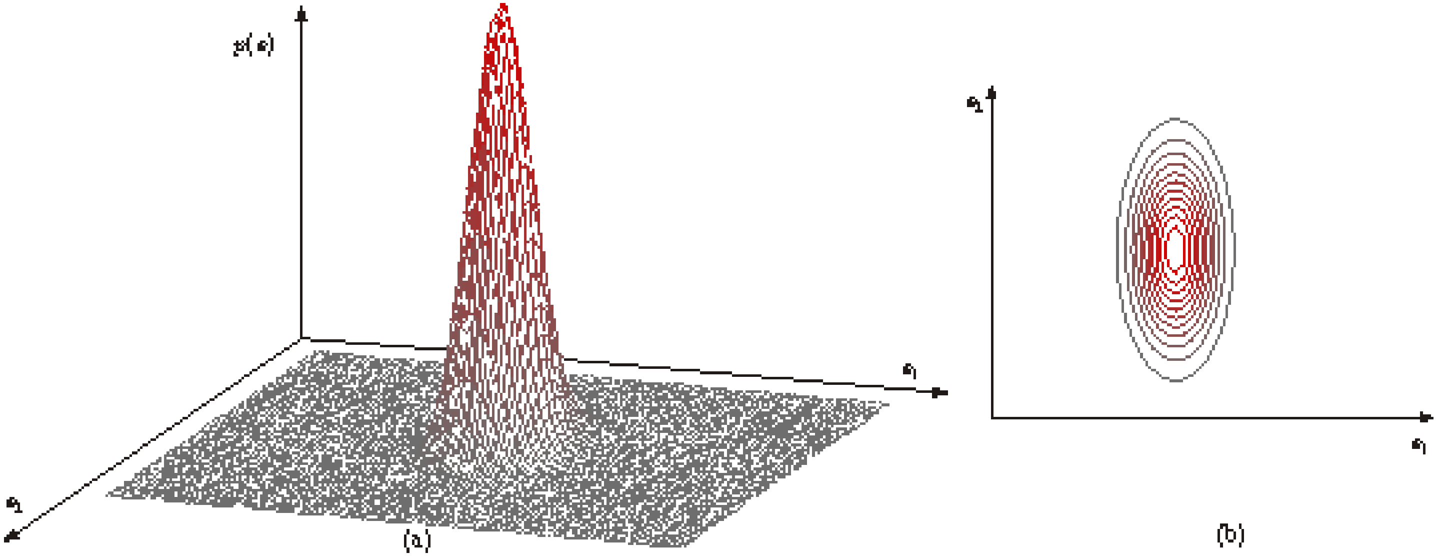




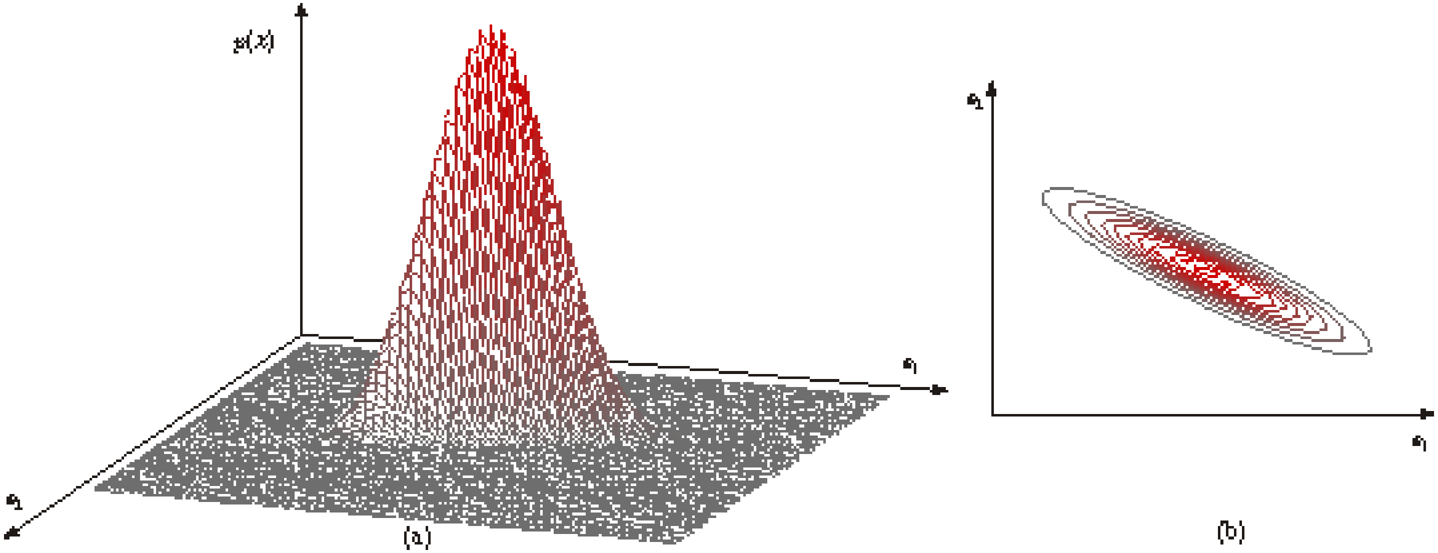
Σ: diagonal with $\sigma_1^2 >> \sigma_2^2$

Probability and statistics: a brief review

- **Multi dim. normal (Gaussian) distribution $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $N(\boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$:**

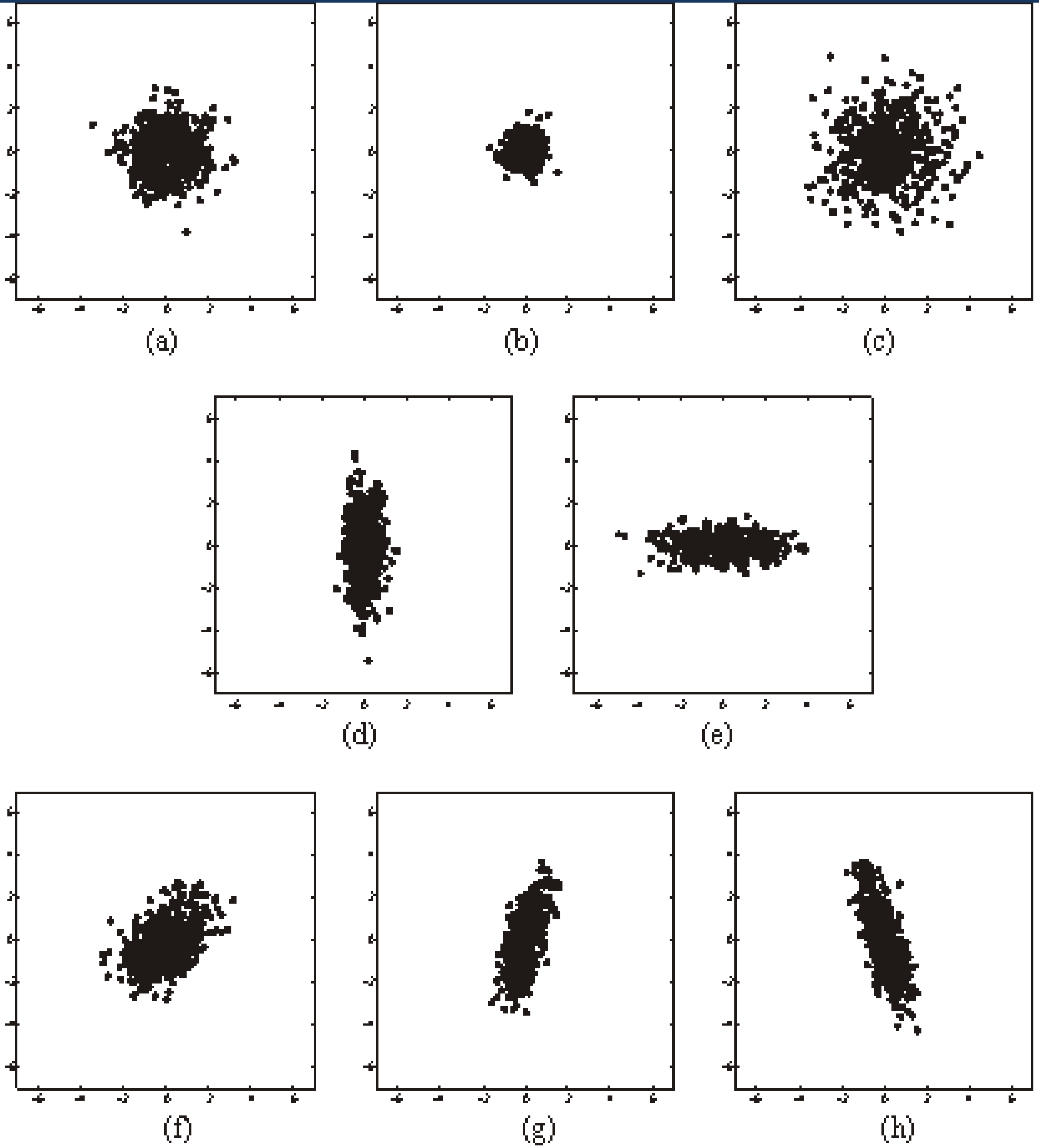


Σ: diagonal with $\sigma_1^2 << \sigma_2^2$



Σ: non diagonal

Probability and statistics: a brief review



$$\Sigma=\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

(α) $\sigma_1^2=\sigma_2^2=1, \sigma_{12}=0$

(β) $\sigma_1^2=\sigma_2^2=0.2, \sigma_{12}=0$

(γ) $\sigma_1^2=\sigma_2^2=2, \sigma_{12}=0$

(δ) $\sigma_1^2=0.2, \sigma_2^2=2, \sigma_{12}=0$

(ε) $\sigma_1^2=2, \sigma_2^2=0.2, \sigma_{12}=0$

(στ) $\sigma_1^2=\sigma_2^2=1, \sigma_{12}=0.5$

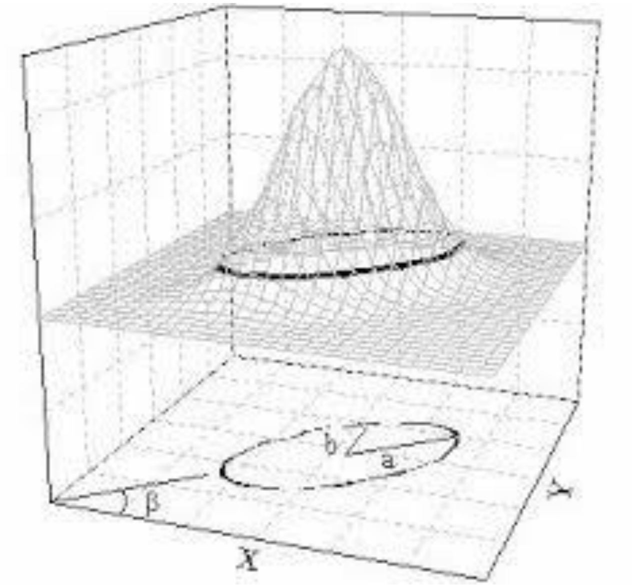
(ζ) $\sigma_1^2=0.3, \sigma_2^2=2, \sigma_{12}=0.5$

(η) $\sigma_1^2=0.3, \sigma_2^2=2, \sigma_{12}=-0.5$

Probability and statistics: a brief review

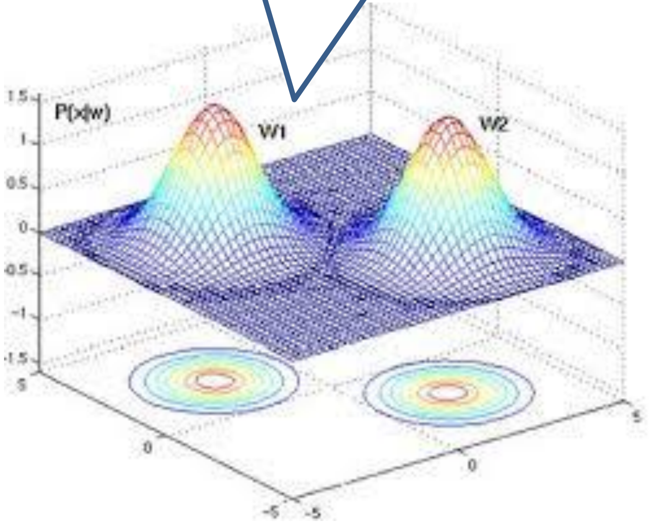
Continuous RV distributions (cont.)

- **Other examples of multi-dimensional pdfs**



Two-dim. pdfs

Bimodal distribution



Probability and statistics: a brief review

Likelihood function

- Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ a set of independent data vectors
- Let $p_{\boldsymbol{\theta}}(\cdot)$ be a pdf belonging to a known parametric set of pdf functions of parameter vector $\boldsymbol{\theta}$.
- $p(\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}) \equiv p(\boldsymbol{x}; \boldsymbol{\theta})$.

  Examples:

  ➢If $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is normal distribution parameterized on the mean vector $\boldsymbol{\mu}$, $\boldsymbol{\theta}$ will simply be $\boldsymbol{\mu}$.

  ➢If $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is normal distribution parameterized on both the mean vector $\boldsymbol{\mu}$ and the cov. matrix Σ, $\boldsymbol{\theta}$ will contain the coordinates of both $\boldsymbol{\mu}$ and Σ.

Likelihood function of $\boldsymbol{\theta}$ wrt X: $p(X; \boldsymbol{\theta}) = p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N; \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i; \boldsymbol{\theta})$

Log-likelihood function of $\boldsymbol{\theta}$ wrt X:

$$L(\boldsymbol{\theta}) = \ln p(X; \boldsymbol{\theta}) = \ln p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N; \boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i; \boldsymbol{\theta})$$

Probability and statistics: a brief review

Likelihood function

Example:

- $X = \{-2, -1, 0, 1, 2\}$

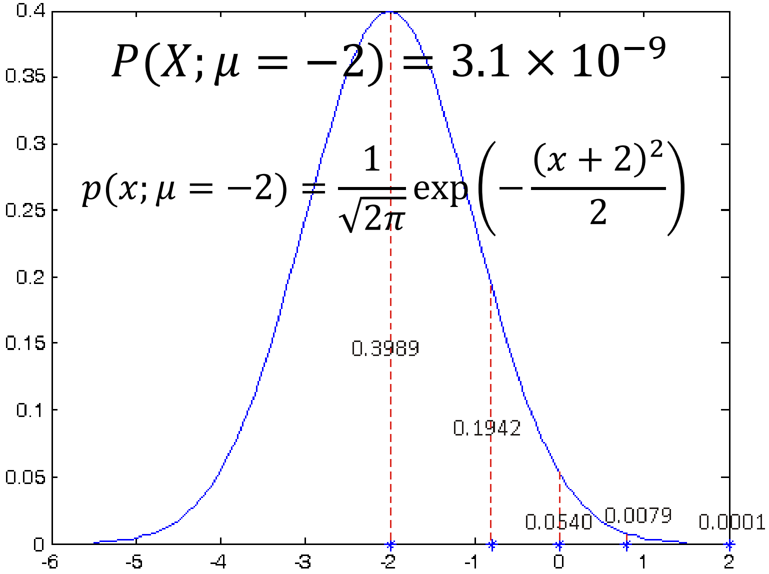
$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2})$$

- Consider the parametric set of normal distributions of unit variance, parameterized on μ.
- The likelihood of μ wrt X is

$$p(X;\mu) = p(-2,-1,0,1,2;\mu) =$$
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-2-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-1-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(0-\mu)^2}{2}\right)$$
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(2-\mu)^2}{2}\right)$$

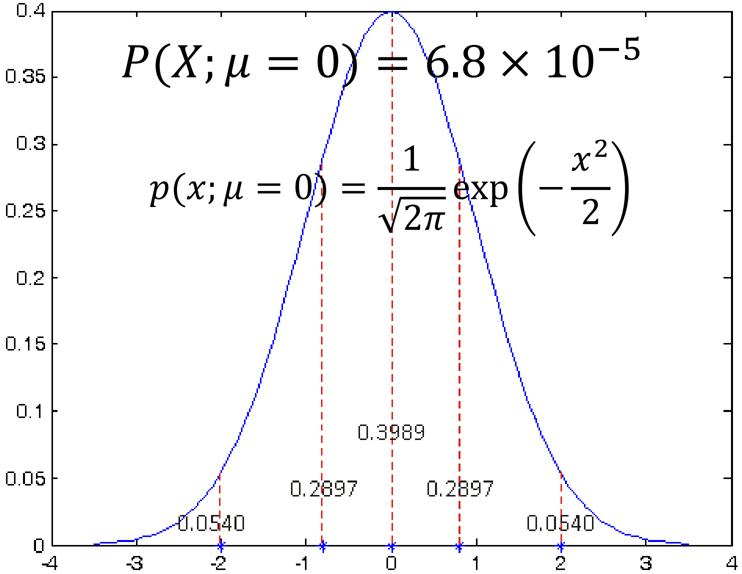
Probability and statistics: a brief review

Likelihood function



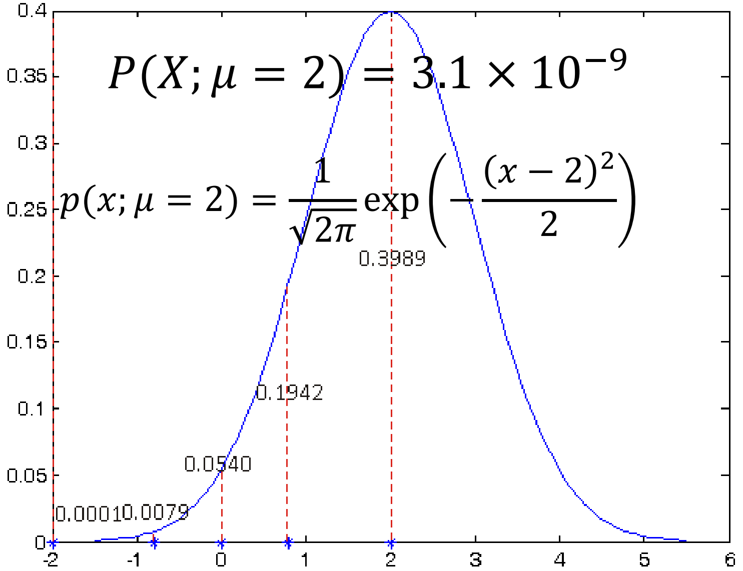
$$P(X;\mu = -2) = 3.1 \times 10^{-9}$$

$$p(x;\mu = -2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+2)^2}{2}\right)$$



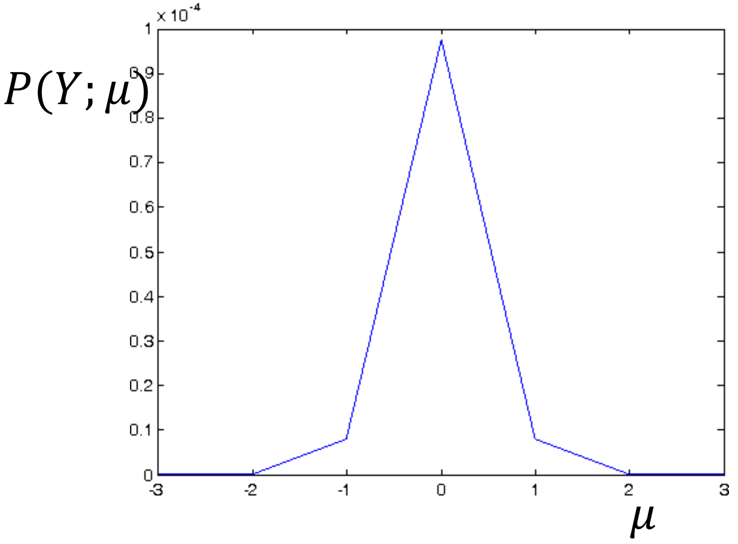
$$P(X;\mu = 0) = 6.8 \times 10^{-5}$$

$$p(x;\mu = 0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$



$$P(X;\mu = 2) = 3.1 \times 10^{-9}$$

$$p(x;\mu = 2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-2)^2}{2}\right)$$



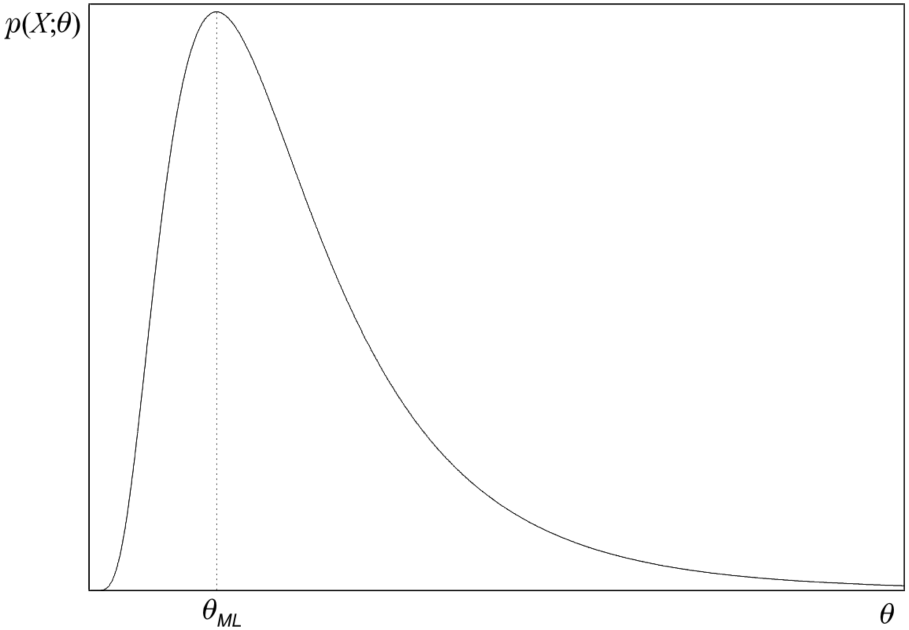
Probabilistic CFO clustering algorithms

Maximum likelihood (ML) method:

Given a set of independent data vectors $Y = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$, estimate the parameter vector $\boldsymbol{\theta}$ as the maximum of the likelihood ($p(Y;\boldsymbol{\theta})$) or the log-likelihood ($L(\boldsymbol{\theta})$) function.

$$\boxed{\hat{\boldsymbol{\theta}}_{ML} = argmax_{\boldsymbol{\theta}}\, p(Y;\boldsymbol{\theta})} \rightarrow \boxed{\hat{\boldsymbol{\theta}}_{ML}: \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{\kappa=1}^{\mathrm{N}} \frac{1}{p(\boldsymbol{x}_k;\theta)} \frac{\partial p(\boldsymbol{x}_k;\theta)}{\partial \boldsymbol{\theta}} = \mathbf{0}}$$

Since $\ln(\cdot)$ is an increasing function, $p(Y;\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$ share the same maxima.



Probabilistic CFO clustering algorithms

Maximum likelihood (ML) method:

Assuming that

- the chosen model $p(\boldsymbol{x};\ \boldsymbol{\theta})$ is correct and
- there exists a true parameter $\boldsymbol{\theta}_{\boldsymbol{o}}$,

the ML estimator

(a) is asymptotically **unbiased** $lim_{N\to\infty}E\left[\widehat{\boldsymbol{\theta}}_{ML}\right]=\boldsymbol{\theta}_{o}$

(b) is asymptotically **consistent** $lim_{N\to\infty}Prob\left\{\left\|\widehat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_{o}\right\|\right\}=0$

(c) is asymptotically **efficient** (it achieves the Cramer-Rao lower bound)

The **pdf** of the ML estimator approaches the normal distribution with mean $\boldsymbol{\theta}_{\boldsymbol{o}}$, as $N\to\infty$.

Maximum likelihood method

Example 1:

-Let Y be a set of N (independent from each other) data points, $\boldsymbol{x}_i$, $i = 1, \dots, N$, generated by a normal distribution $p(\boldsymbol{x};\ \boldsymbol{\theta})$ of known covariance matrix and unknown mean.

-Determine the ML estimate of the mean $\boldsymbol{\mu}$ of $p(\boldsymbol{x};\boldsymbol{\theta})$, based on Y.

Solution:

-The unknown parameter vector in this case is the mean vector $\boldsymbol{\mu}$, i.e. $\boldsymbol{\theta} \equiv \boldsymbol{\mu}$.

-It is

$$p(\boldsymbol{x};\boldsymbol{\theta}) \equiv p(\boldsymbol{x};\boldsymbol{\mu}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \cdot exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \Rightarrow$$

$$\ln p(\boldsymbol{x};\boldsymbol{\mu}) = \ln\frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = C - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

Then

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i;\boldsymbol{\mu}) = NC - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

Maximum likelihood method

Example 1 (cont.):

Setting the gradient of $L(\boldsymbol{\mu})$ wrt $\boldsymbol{\mu}$ equal to $\mathbf{0}$ we have

$$\frac{\partial L(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \frac{\partial}{\partial \boldsymbol{\mu}}\left(NC - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right) = \mathbf{0} \Leftrightarrow$$

$$\sum_{i=1}^{N}\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) = \mathbf{0} \Leftrightarrow \sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu}) = \mathbf{0} \Leftrightarrow \sum_{i=1}^{N}\boldsymbol{x}_i - N\boldsymbol{\mu} = \mathbf{0}$$

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

Remark: The ML estimate for the covariance matrix is

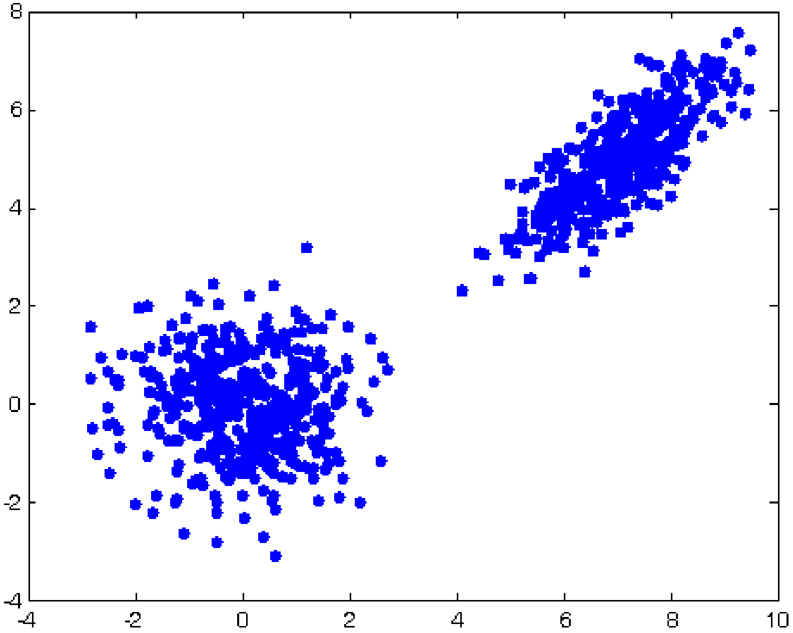
$$\Sigma_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

Probabilistic CFO clustering algorithms

Mixture models - The Expectation – Maximization (EM) algorithm

Mixture model: A weighted sum of known parametric form pdfs.

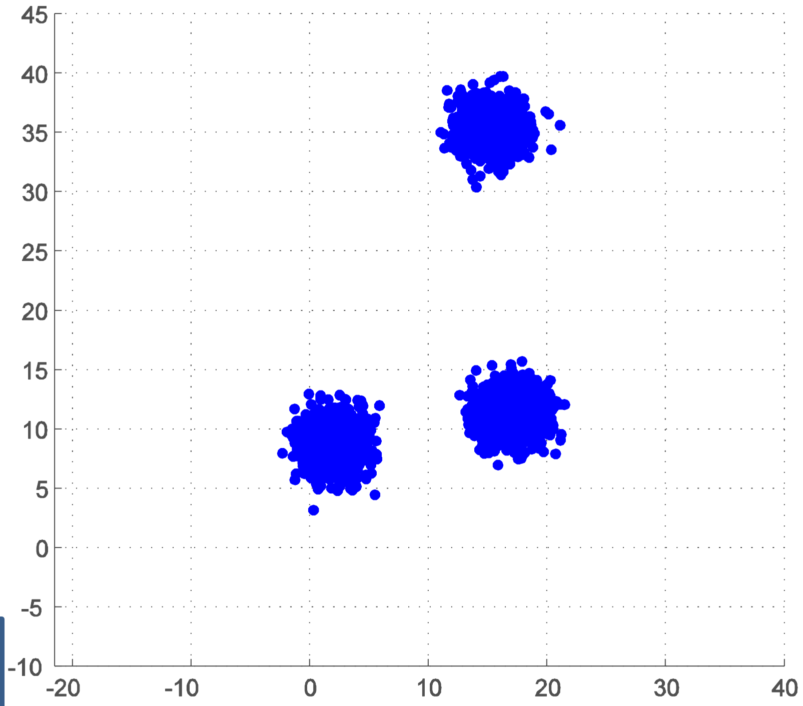
$$p(\boldsymbol{x}) = \sum_{j=1}^{m} P_j \, p(\boldsymbol{x} \mid j), \quad \sum_{j=1}^{m} P_j = 1, \quad \int_{-\infty}^{+\infty} p(\boldsymbol{x} \mid j) = 1$$



- Assume that $p(\boldsymbol{x})$ models the distribution of the data in X (each pdf models a cluster).

- The **aim** is to **move** each pdf so that to "cover" the area in the data space where the vectors of each cluster lie (**mixture decomposition**).

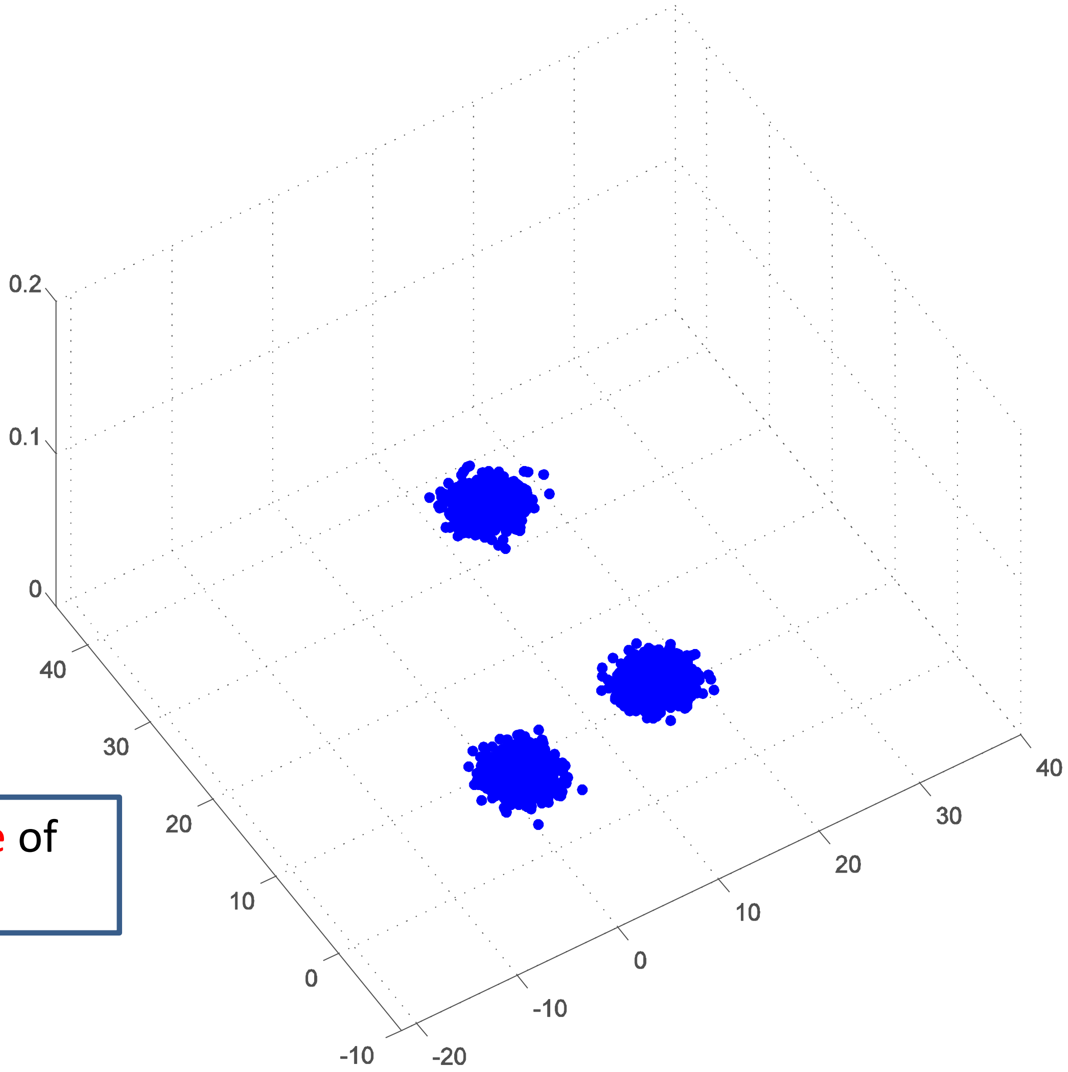
Probabilistic CFO clustering algorithms



Prerequisite: Knowledge of the number of clusters.

- **Adopt** a parametric mixture of distributions, each one corresponding to a cluster (e.g., mixture of Gaussians), initialized randomly.
- **Move** iteratively the distributions each one above a cluster, **optimizing** a criterion.

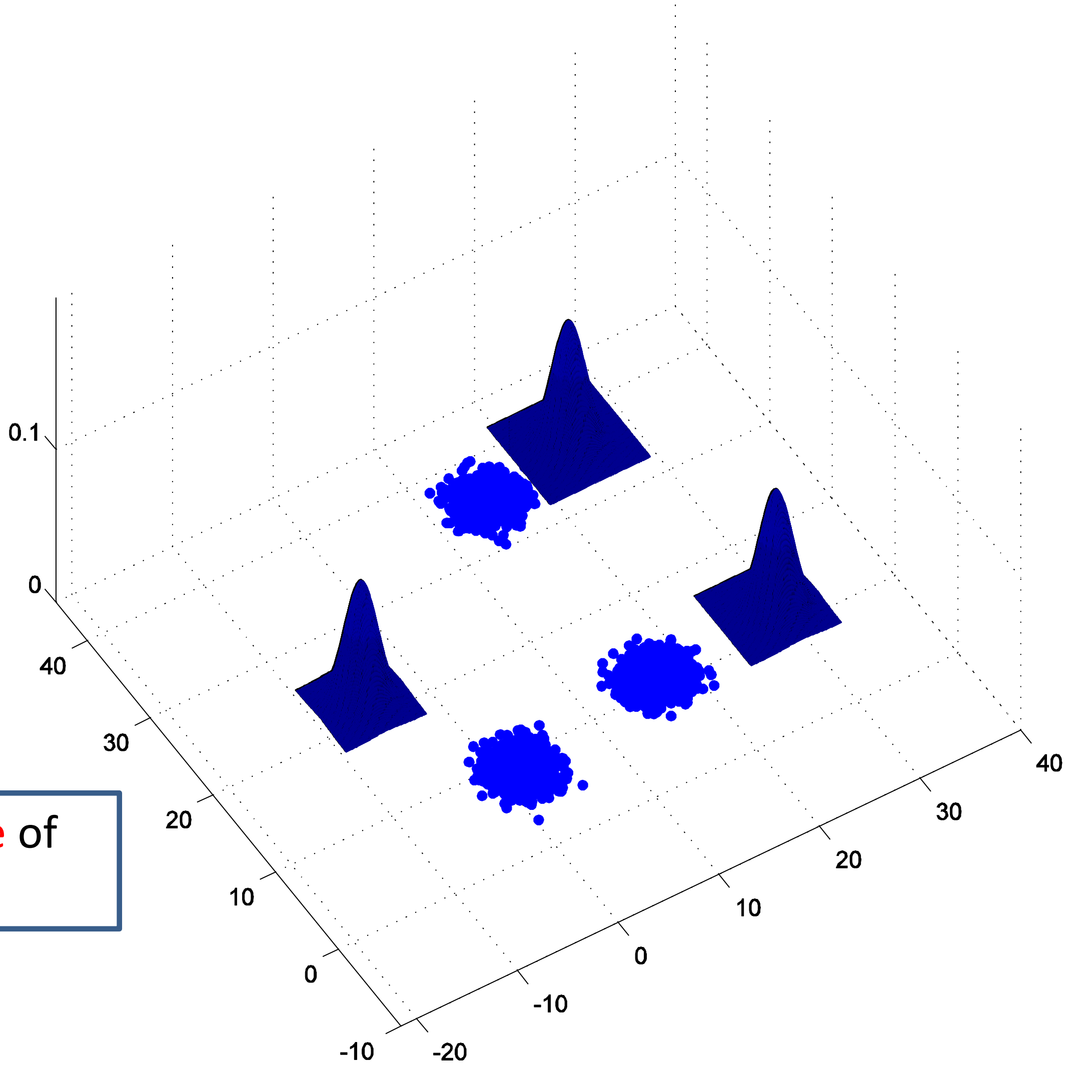
Probabilistic CFO clustering algorithms



Prerequisite: Knowledge of the number of clusters.

- **Adopt** a parametric mixture of distributions, each one corresponding to a cluster (e.g., mixture of Gaussians), initialized randomly.
- **Move** iteratively the distributions each one above a cluster, **optimizing** a criterion.

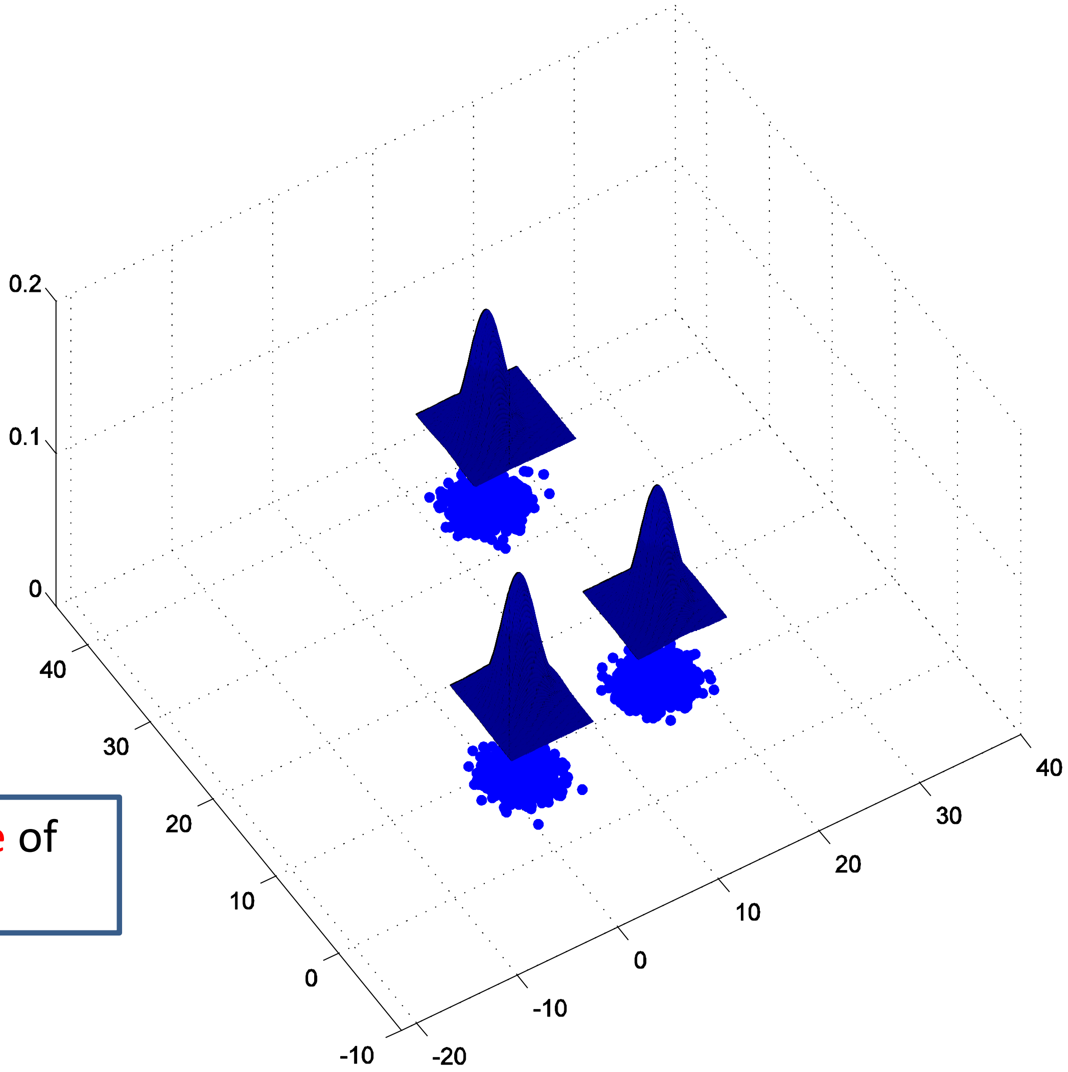
Probabilistic CFO clustering algorithms



Prerequisite: Knowledge of the number of clusters.

- **Adopt** a parametric mixture of distributions, each one corresponding to a cluster (e.g., mixture of Gaussians), initialized randomly.
- **Move** iteratively the distributions each one above a cluster, **optimizing** a criterion.

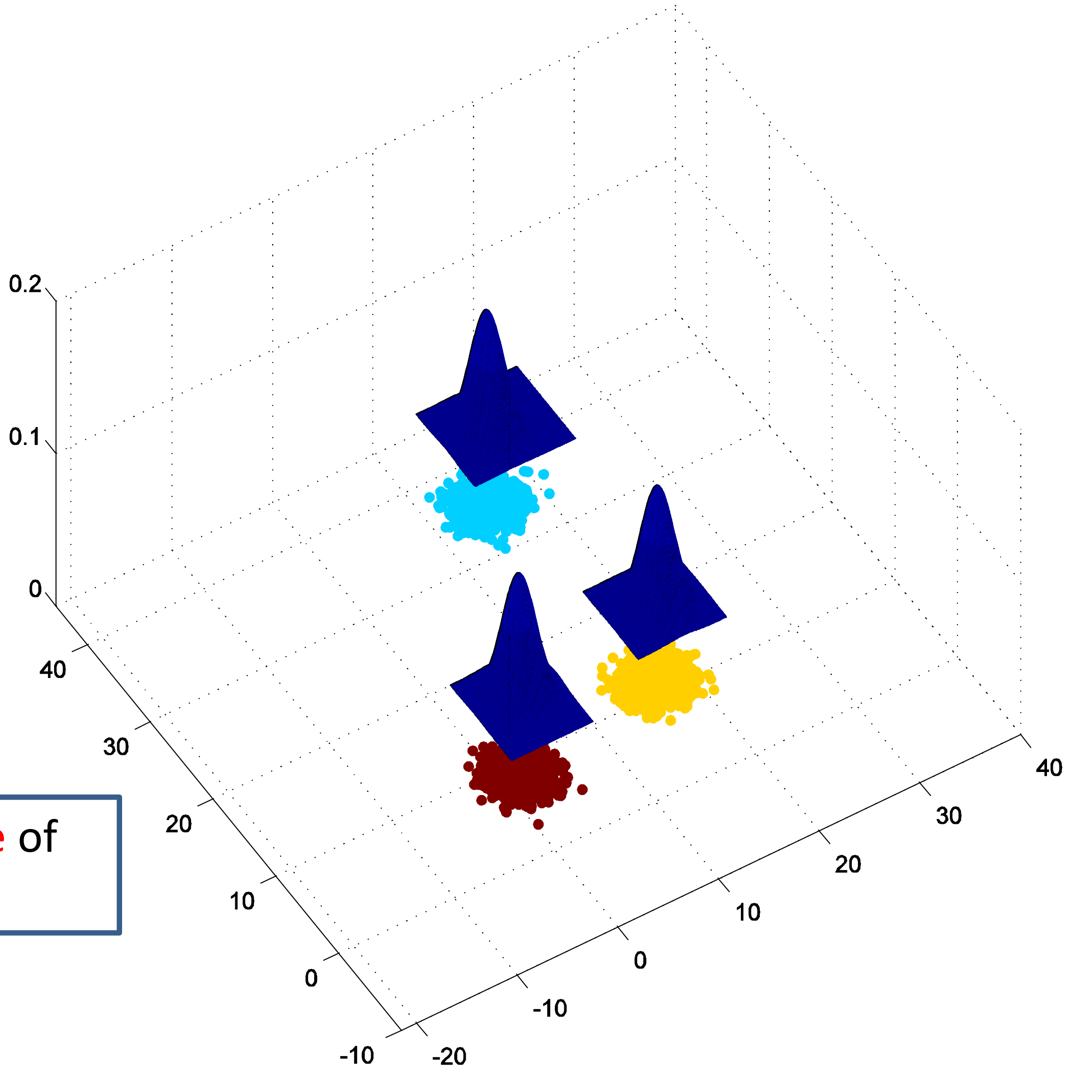
Probabilistic CFO clustering algorithms



Prerequisite: Knowledge of the number of clusters.

- **Adopt** a parametric mixture of distributions, each one corresponding to a cluster (e.g., mixture of Gaussians), initialized randomly.
- **Move** iteratively the distributions each one above a cluster, **optimizing** a criterion.

Probabilistic CFO clustering algorithms



Prerequisite: Knowledge of the number of clusters.

- **Adopt** a parametric mixture of distributions, each one corresponding to a cluster (e.g., mixture of Gaussians), initialized randomly.
- **Move** iteratively the distributions each one above a cluster, **optimizing** a criterion.

Probabilistic CFO clustering algorithms

Let $X = \{x_1, x_2, \dots, x_N\}$ be a set of data points.

Each vector belongs exclusively to a single cluster, with a certain probability.

Each cluster is **modeled** by a pdf $p(\boldsymbol{x}|j)$, parameterized by the vector $\boldsymbol{\theta}_j$.

Let:

$\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_m\}$

$P = \{P_1, P_2, \dots, P_m\}$, the set of a priori probabilities of the clusters.

$P(j|\boldsymbol{x}) \equiv P\left(j\middle|\boldsymbol{x}; \boldsymbol{\theta}_j\right)$ the (a posteriori) probability of cluster j, given $\boldsymbol{x}$.

$p(\boldsymbol{x}|j) \equiv p\left(\boldsymbol{x}\middle|j; \boldsymbol{\theta}_j\right)$ the pdf that models cluster j.

It is $p(\boldsymbol{x}) = \sum_{j=1}^{m} p(\boldsymbol{x}, j) = \sum_{j=1}^{m} p(\boldsymbol{x}|j)\, P_j$

Bayes rule $P(j|\boldsymbol{x}) = \frac{p(\boldsymbol{x},j)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|j)\boldsymbol{P_j}}{p(\boldsymbol{x})}$

Probabilistic CFO clustering algorithms

It is

- $\sum_{j=1}^{m} P(j|\boldsymbol{x}_i) = 1\,, i = 1, \ldots, N$
- $\sum_{j=1}^{m} P_j = 1\,.$

ML: $L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln(p(\boldsymbol{x}_i; \boldsymbol{\theta}))$

Define the cost function

$$\ln p(X; \Theta, P) = \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln p(\boldsymbol{x}_i, j; \boldsymbol{\theta}_j)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln\left(p\left(\boldsymbol{x}_i \middle| j; \boldsymbol{\theta}_j\right) P_j\right)$$

When $\ln p(X; \Theta, P)$ is **maximized**?

When large $P(j|\boldsymbol{x}_i)$’s are **multiplied** by large $\ln p(\boldsymbol{x}_i, j; \boldsymbol{\theta}_j)$ ‘s.

Probabilistic CFO clustering algorithms

For **fixed $\boldsymbol{\theta}_j$'s:** Use the Bayes rule $P(j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|j;\boldsymbol{\theta}_j)\boldsymbol{P_j}}{p(\boldsymbol{x};\boldsymbol{\Theta})}$

For **fixed $P(j|\boldsymbol{x})$'s:** Solve the following maximization problem

$$max_{\Theta,P} \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln\left(p\left(\boldsymbol{x}_i\middle|j;\boldsymbol{\theta}_j\right)P_j\right)$$

$$= max_{\Theta,P} \left[\sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln\left(p\left(\boldsymbol{x}_i\middle|j;\boldsymbol{\theta}_j\right)\right) + \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln P_j\right]$$

Subject to the constraint $\sum_{j=1}^{m} P_j = 1$.

Mixture models – Expectation-Maximization (EM) algorithm

For **fixed $\boldsymbol{\theta}_j$‘s:** Use the Bayes rule $P(j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|j;\boldsymbol{\theta}_j)\boldsymbol{P_j}}{p(\boldsymbol{x};\boldsymbol{\Theta})}$

For **fixed $P(j|\boldsymbol{x})$‘s:** Solve the following maximization problem

$$max_{\Theta,P} \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln\left(p(\boldsymbol{x}_i|j;\boldsymbol{\theta}_j)P_j\right) =$$

$$max_{\Theta} \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln\left(p(\boldsymbol{x}_i|j;\boldsymbol{\theta}_j)\right) + max_P \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln P_j$$

$$= max_{\Theta} \sum_{j=1}^{m} \sum_{i=1}^{N} P(j|\boldsymbol{x}_i) \ln\left(p(\boldsymbol{x}_i|j;\boldsymbol{\theta}_j)\right) + max_P \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x}_i) \ln P_j$$

Subject to the **constraint** $\sum_{j=1}^{m} P_j = 1$.

The above maximization problem is equivalent to the following maximization sub-problems

– $\boldsymbol{\theta}_j = argmax_{\boldsymbol{\theta}_j} \sum_{i=1}^{N} P(j|\boldsymbol{x_i}) \ln\left(p(\boldsymbol{x}_i|j;\boldsymbol{\theta}_j)\right), j = 1, \dots, m$

– $P \equiv \{P_1, P_2, \dots, P_m\} = argmax_P \sum_{i=1}^{N} \sum_{j=1}^{m} P(j|\boldsymbol{x_i}) \ln P_j, s.t. \sum_{j=1}^{m} P_j = 1 \Leftrightarrow$

$$P_j = \frac{1}{N} \sum_{i=1}^{N} P(j|\boldsymbol{x_i}), j = 1, \dots, m$$

Probabilistic CFO clustering algorithms

Generalized probabilistic Algorithmic Scheme (GPrAS)

```
• Choose $\boldsymbol{\theta}_j(0)$, $P_j(0)$ as initial estimates for $\boldsymbol{\theta}_j, P_j,$ respectively, $j = 1, \dots, m$
• t=0
• Repeat
    – For i=1 to N  % Expectation step
        o For j=1 to m
            $P(j|\boldsymbol{x}_i;\Theta^{(t)},P^{(t)}) = \frac{p(x_i|j;\theta_j{}^{(t)})P_j{}^{(t)}}{\sum_{q=1}^{m} p(x_i|q;\theta_q{}^{(t)})P_q{}^{(t)}} \equiv \gamma_{ji}{}^{(t)}$
        o End {For-j}
    – End {For-i}
    – t=t+1
    – For j=1 to m % Parameter updating – Maximization step
        o Set
            $\boldsymbol{\theta}_j{}^{(t)} = argmax_{\boldsymbol{\theta}_j} \sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)} \ln\left(p\left(\boldsymbol{x}_i\middle|j;\boldsymbol{\theta}_j\right)\right), j = 1, \dots, m$
            $P_j{}^{(t)} = \frac{1}{N}\sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)}, j = 1, \dots, m$
    - End {For-j}
• Until a termination criterion is met.
```

Probabilistic CFO clustering algorithms

Remark: The above algorithm is an instance of the more general Expectation-Maximization (EM) framework.

<u>GPrAS – The case of normal pdfs</u>

Each cluster is **modeled** by a normal distribution

$$p(\boldsymbol{x}|j;\mu_j,\Sigma_j)=\frac{1}{(2\pi)^l|\Sigma_j|^{1/2}}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_j)^T{\Sigma_j}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)}{2}\right), j=1,\dots m$$

In this case $\boldsymbol{\theta}_j=\{\boldsymbol{\mu}_j,\Sigma_j\}$.

$$\{\boldsymbol{\mu}_j,\Sigma_j\}=argmax_{\{\boldsymbol{\mu}_j,\Sigma_j\}}\sum_{i=1}^{N}P(j|\boldsymbol{x}_i)\ln\left(p(\boldsymbol{x}_i|j;\boldsymbol{\mu}_j,\Sigma_j)\right)$$

Equating the gradient of the above function wrt $\boldsymbol{\mu}_j,\Sigma_j$ to **0** and *O*, respectively, we have

$$\boldsymbol{\mu}_j=\frac{\sum_{i=1}^{N}P(j|\boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_{i=1}^{N}P(j|\boldsymbol{x}_i)}$$

$$\Sigma_j=\frac{\sum_{i=1}^{N}P(j|\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_j)\,(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^{\boldsymbol{T}}}{\sum_{i=1}^{N}P(j|\boldsymbol{x}_i)}$$

Probabilistic CFO clustering algorithms

GPrAS – The normal pdfs case

```
•  Choose $\boldsymbol{\mu}_j(0)$, $\Sigma_j(0)$, $P_j(0)$ as initial estimates for $\boldsymbol{\mu}_j, \Sigma_j, P_j$, resp., $j = 1, \dots, m$
•  $t$=0
•  Repeat
     – For $i$=1 to $N$  % Expectation step
          o For $j$=1 to $m$
               $P(j|\boldsymbol{x}_i; \Theta^{(t)}, P^{(t)}) = \frac{p(x_i|j;\theta_j^{(t)})P_j^{(t)}}{\sum_{q=1}^{m} p(x_i|q;\theta_q^{(t)})P_q^{(t)}} \equiv \gamma_{ji}^{(t)}$
          o End {For-$j$}
     – End {For-$i$}
     –$t$=$t$+1
     – For $j$=1 to $m$ % Parameter updating – Maximization step
          o Set
               $\boldsymbol{\mu}_j^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{ji}^{(t-1)} \boldsymbol{x}_i}{\sum_{i=1}^{N} \gamma_{ji}^{(t-1)}}, \quad \Sigma_j^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{ji}^{(t-1)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^{\boldsymbol{T}}}{\sum_{i=1}^{N} \gamma_{ji}^{(t-1)}} \quad j = 1, \dots, m$
               $P_j^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ji}^{(t-1)}, j = 1, \dots, m$
     - End {For-$j$}
•  Until a termination criterion is met.
```

Probabilistic CFO clustering algorithms

GPrAS – The normal pdfs case

```
• Choose $\boldsymbol{\mu}_j(0), \Sigma_j(0), P_j(0)$ as initial estimates for $\boldsymbol{\mu}_j, \Sigma_j, P_j$, resp., $j = 1, \ldots, m$
• t=0
• Repeat
    – For i=1 to N  % Expectation step
        $P(C_j|\boldsymbol{x};\Theta(t)) = \dfrac{|\Sigma_j(t)|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_j(t))^T\Sigma_j^{-1}(t)(\boldsymbol{x}-\mu_j(t))\right)P_j(t)}{\sum_{k=1}^{m}|\Sigma_k(t)|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_k(t))^T\Sigma_k^{-1}(t)(\boldsymbol{x}-\mu_k(t))\right)P_k(t)}$
        o End {For-j}
    – End {For-i}
    –t=t+1
    – For j=1 to m % Parameter updating – Maximization step
        o Set
            $\boldsymbol{\mu}_j{}^{(t)} = \dfrac{\sum_{i=1}^{N}\gamma_{ji}{}^{(t-1)}\boldsymbol{x}_i}{\sum_{i=1}^{N}\gamma_{ji}{}^{(t-1)}}, \qquad \Sigma_j{}^{(t)} = \dfrac{\sum_{i=1}^{N}\gamma_{ji}{}^{(t-1)}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^T}{\sum_{i=1}^{N}\gamma_{ji}{}^{(t-1)}} \quad j = 1, \ldots, m$
            $P_j{}^{(t)} = \dfrac{1}{N}\sum_{i=1}^{N}\gamma_{ji}{}^{(t-1)}, j = 1, \ldots, m$
    - End {For-j}
• Until a termination criterion is met.
```

Probabilistic CFO clustering algorithms

Remark:

- The above scheme is more computationally demanding since it requires the inversion of the m covariance matrices at each iteration step. Two ways to deal with this problem are:
    - The use of a single covariance matrix for all clusters.
    - The use of different diagonal covariance matrices.

Example: (a) Consider three two-dimensional normal distributions with mean values:

$$\boldsymbol{\mu}_1=[1,\ 1]^T,\ \boldsymbol{\mu}_2=[3.5,\ 3.5]^T,\ \boldsymbol{\mu}_3=[6,\ 1]^T$$

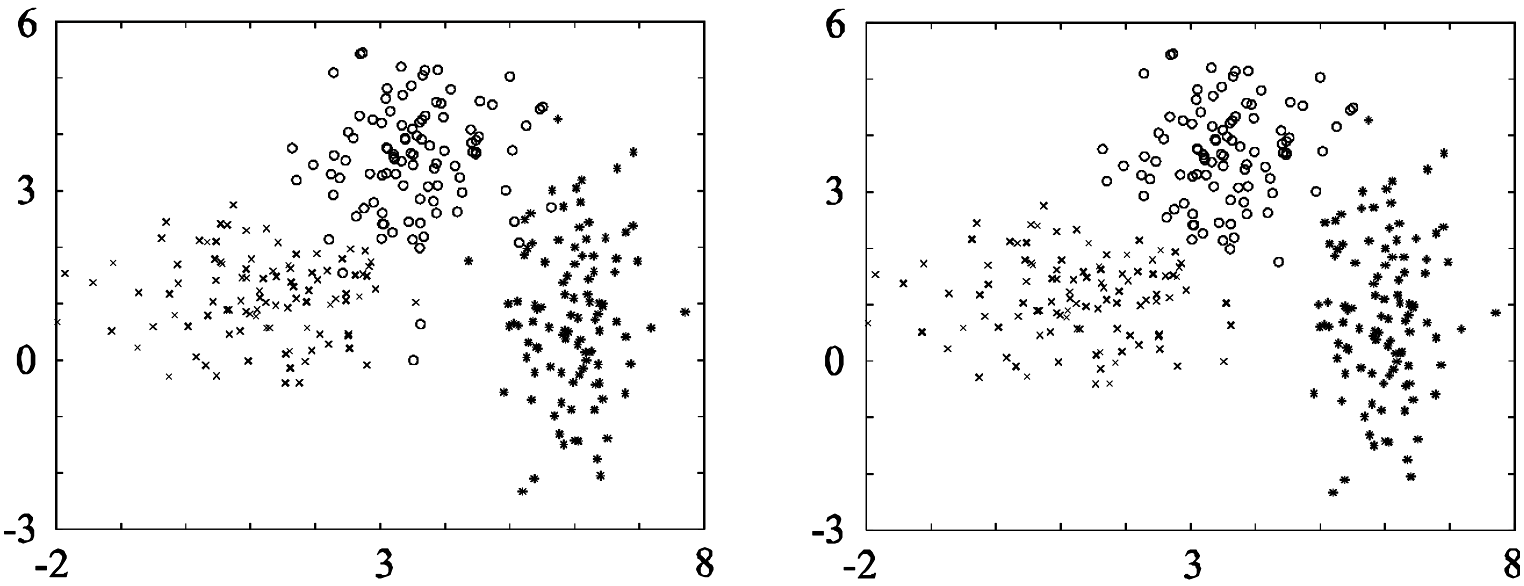
and covariance matrices

$$\Sigma_1=\begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix},\quad \Sigma_2=\begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix},\quad \Sigma_3=\begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix},$$

respectively.

A group of 100 vectors stem from each distribution. These form the data set X.

Probabilistic CFO clustering algorithms



(a) The data set

(b) Results of GMDAS

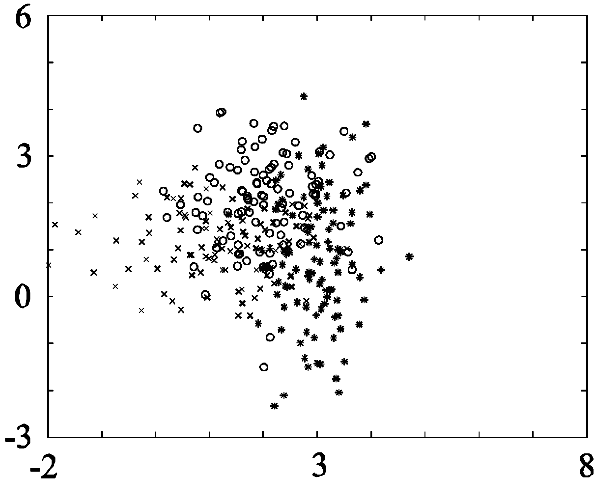
Confusion matrix:

	Cluster 1	*Cluster 2*	*Cluster 3*
1st distribution	99	0	1
2nd distribution	0	100	0
3rd distribution	3	4	93

The algorithm reveals accurately the underlying structure.

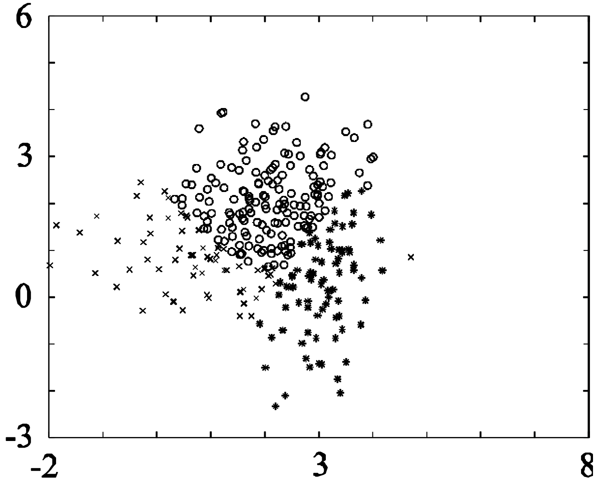
Probabilistic CFO clustering algorithms

(b) The same as (a) but now $\underline{\mu}_1=[1, 1]^T$, $\underline{\mu}_2=[2, 2]^T$, $\underline{\mu}_3=[3, 1]^T$ (The clusters are closer).



(a)

The data set



(b)

Results of GMDAS

Confusion matrix:

	Cluster 1	*Cluster 2*	*Cluster 3*
1st distribution	85	4	11
2nd distribution	35	56	9
3rd distribution	26	0	74

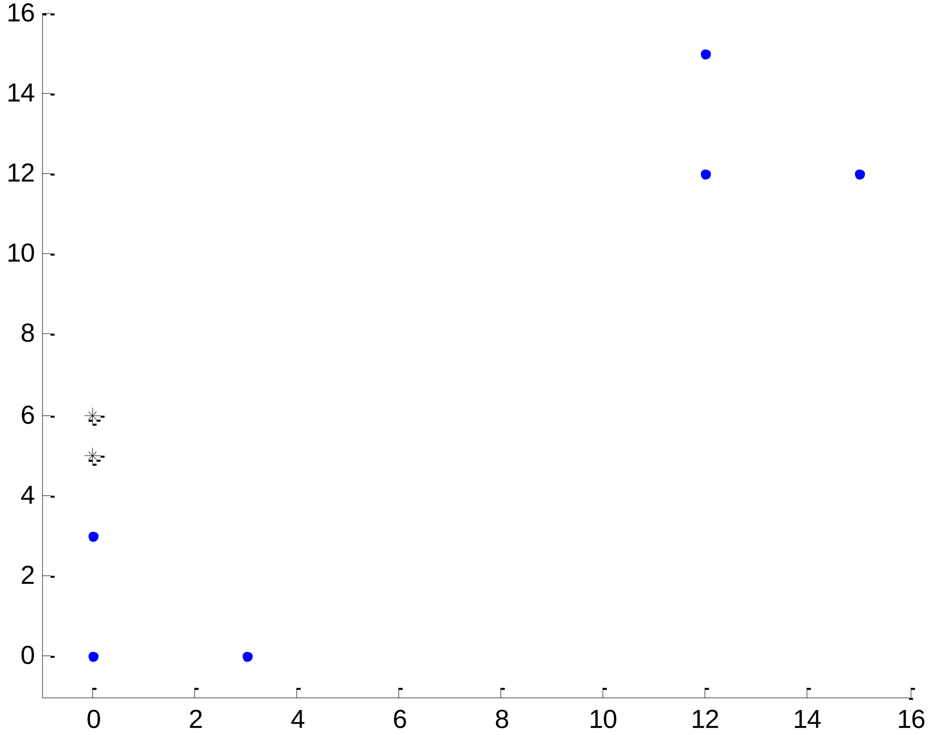
The algorithm reveals the underlying structure less accurately.

Probabilistic CFO clustering algorithms

<u>*Example*</u> $\boldsymbol{x}_1 = [0\ 0]^T, \boldsymbol{x}_2 = [3\ 0]^T, \boldsymbol{x}_3 = [0\ 3]^T, \boldsymbol{x}_4 = [12\ 12]^T, \boldsymbol{x}_5 = [15\ 12]^T, \boldsymbol{x}_6 = [12\ 15]^T$

Initially:

$$\theta_1(0) = [0, 5]^{\mathrm{T}}$$
$$\theta_2(0) = [0, 6]^{\mathrm{T}}$$
$$\mathrm{P}_1(0) = 0.1$$
$$P_2(0) = 0.9$$



$$p(\boldsymbol{x}|1) = \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_1\|^2), \qquad P(1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|1)P_1}{p(\boldsymbol{x})}$$

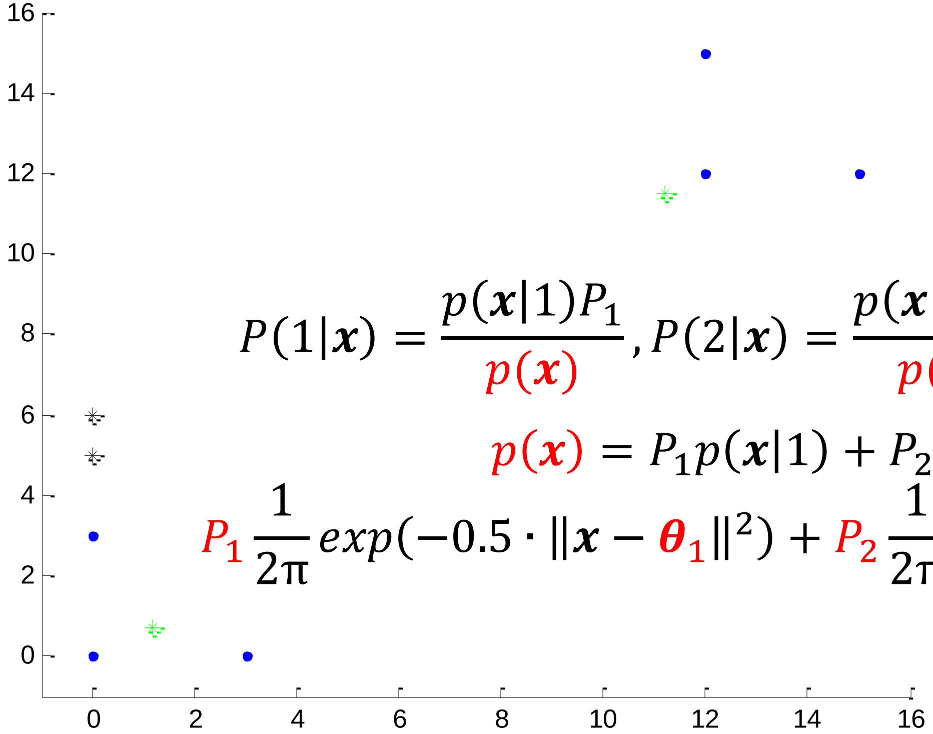
$$p(\boldsymbol{x}|2) = \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_2\|^2), \qquad P(2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|2)P_2}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = P_1 p(\boldsymbol{x}|1) + P_2 p(\boldsymbol{x}|2) = P_1 \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_1\|^2) + P_2 \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_2\|^2)$$

$$\ln p(X; \Theta, P) = \sum_{i=1}^{N} [P(1|\boldsymbol{x}_i) \ln(p(\boldsymbol{x}_i|1; \boldsymbol{\theta}_1)P_1) + P(2|\boldsymbol{x}_i) \ln(p(\boldsymbol{x}_i|2; \boldsymbol{\theta}_2)P_2)]$$

Probabilistic CFO clustering algorithms

<u>*Example*</u> $\boldsymbol{x}_1 = [0\ 0]^T, \boldsymbol{x}_2 = [3\ 0]^T \boldsymbol{x}_3 = [0\ 3]^T, \boldsymbol{x}_4 = [12\ 12]^T, \boldsymbol{x}_5 = [15\ 12]^T, \boldsymbol{x}_6 = [12\ 15]^T$



$$P(1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|1)P_1}{p(\boldsymbol{x})}, P(2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|2)P_2}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = P_1 p(\boldsymbol{x}|1) + P_2 p(\boldsymbol{x}|2) =$$

$$P_1 \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_1\|^2) + P_2 \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_2\|^2)$$

1st iteration:

A posteriori probs

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$
$P(1\|x)$	0.9645	0.9645	0.5751	0.0002	0.0002	0.0000
$P(2\|x)$	0.0355	0.0355	0.4249	0.9998	0.9998	1.0000

$\boldsymbol{\theta}_1(1) = [1.1572 \quad 0.6906]^{\mathrm{T}}$

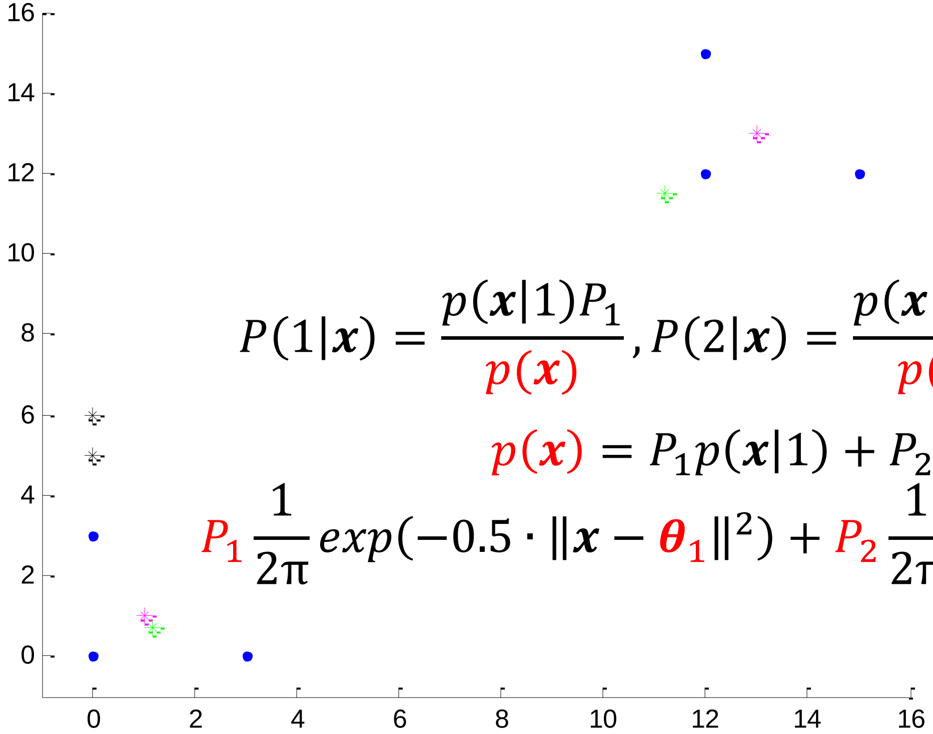
$\boldsymbol{\theta}_2(1) = [11.1864 \quad 11.5207]^{\mathrm{T}}$

$\mathrm{P}_1(1) = 0.4174$

$P_2(1) = 0.5826$

Probabilistic CFO clustering algorithms

<u>*Example*</u> $\boldsymbol{x}_1 = [0\ 0]^T, \boldsymbol{x}_2 = [3\ 0]^T \boldsymbol{x}_3 = [0\ 3]^T, \boldsymbol{x}_4 = [12\ 12]^T, \boldsymbol{x}_5 = [15\ 12]^T, \boldsymbol{x}_6 = [12\ 15]^T$



$$P(1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|1)P_1}{p(\boldsymbol{x})}, P(2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|2)P_2}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = P_1 p(\boldsymbol{x}|1) + P_2 p(\boldsymbol{x}|2) =$$

$$P_1 \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_1\|^2) + P_2 \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_2\|^2)$$

2nd iteration:

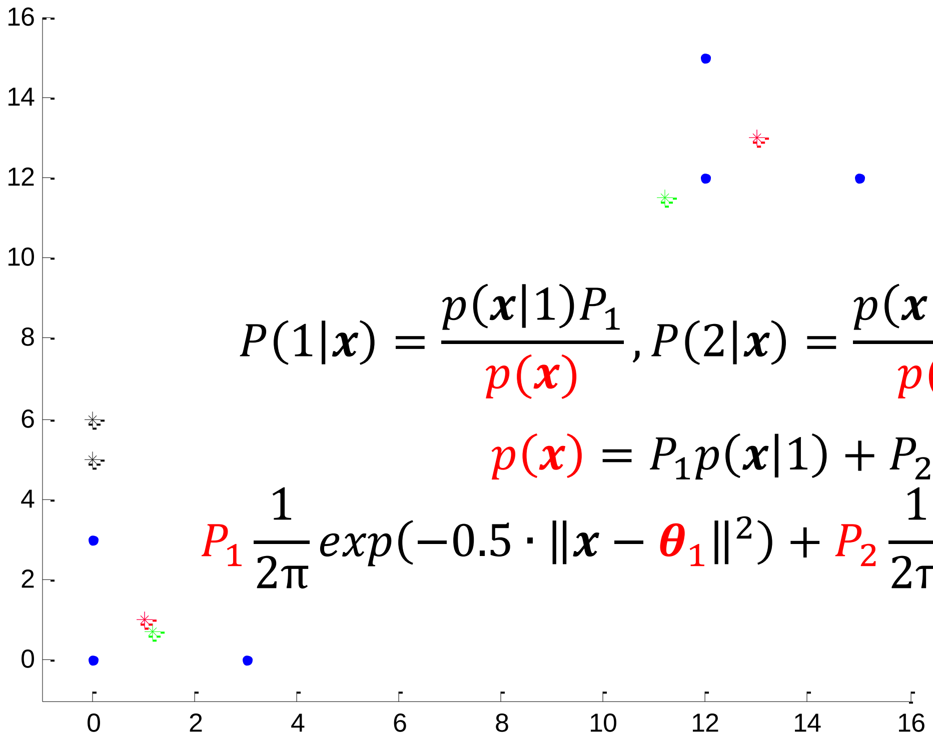
A posteriori probs

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$
$P(1\|x)$	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
$P(2\|x)$	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000

$$\boldsymbol{\theta}_1(2) = [1\ 1]^{\mathrm{T}}$$
$$\boldsymbol{\theta}_2(2) = [13\ 13]^{\mathrm{T}}$$
$$\mathrm{P}_1(2) = 0.5$$
$$P_2(2) = 0.5$$

Probabilistic CFO clustering algorithms

<u>*Example*</u> $\boldsymbol{x}_1 = [0\ 0]^T, \boldsymbol{x}_2 = [3\ 0]^T \boldsymbol{x}_3 = [0\ 3]^T, \boldsymbol{x}_4 = [12\ 12]^T, \boldsymbol{x}_5 = [15\ 12]^T, \boldsymbol{x}_6 = [12\ 15]^T$



$$P(1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|1)P_1}{p(\boldsymbol{x})}, P(2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|2)P_2}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = P_1 p(\boldsymbol{x}|1) + P_2 p(\boldsymbol{x}|2) =$$

$$P_1 \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_1\|^2) + P_2 \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_2\|^2)$$

3rd iteration:

A posteriori probs

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$
$P(1\|x)$	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
$P(2\|x)$	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000

$$\boldsymbol{\theta}_1(3) = [1\ 1]^{\mathrm{T}}$$
$$\boldsymbol{\theta}_2(3) = [13\ 13]^{\mathrm{T}}$$
$$\mathrm{P}_1(3) = 0.5$$
$$P_2(3) = 0.5$$