Clustering algorithms

Konstantinos Koutroumbas

Unit 5

– Hard CFO alg: k-medoids (cont.)
– Probabilistic CFO clust. algorithms

koutroum@noa.gr

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 $X'_1, \ldots, X'_s$ (typically $s = 5, N' = 40 + 2m$).
  - Run PAM on each one of them and identify $\Theta'_1, \ldots, \Theta'_s$.
  - 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 Θ.
  o **(A) Select** randomly q neighbors of Θ.
  o For $j = 1$ to q
    * **If** the present neighbor of Θ is **better** than Θ (in terms of $J(\Theta, U)$) then
      -- **Set** Θ equal to its neighbor
      -- Go to **(A)**
    * End If
  o End For
  o Set $\Theta^i = \Theta$
– End For
– **Select** the best Θ^i with respect to $J(\Theta, U)$.
– Based on Θ^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: $P(x_i \mid y_j) \equiv P(x=x_i \mid y=y_j) = P(x_i,y_j) \,/\, P(y_j)$

- 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: $P(y\,|\,x)=\dfrac{P(x\,|\,y)P(y)}{P(x)}$

It plays a key role in ML.

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

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

- **Events:** $\{x \leq x\}$, $\{x_1 < x \leq x_2\}$, $\{x \geq x\}$

Corresponds to the probability mass function from the discrete case.

- **Cumulative distribution function** (**cdf**): $F_x(x) = P(x \leq x)$

It assigns "mass" to events.

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

- **Probability of events** in terms of **cdf**:
  - $P(x \leq x) = F_x(x)$
  - $P(x_1 < x \leq x_2) = P(x \leq x_2) - P(x \leq x_1) = F_x(x_2) - F_x(x_1)$
  - $P(x \geq x) = = P(x \leq \infty) - P(x \leq x) = 1 - P(x \leq x) = 1 - F_x(x)$

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:**

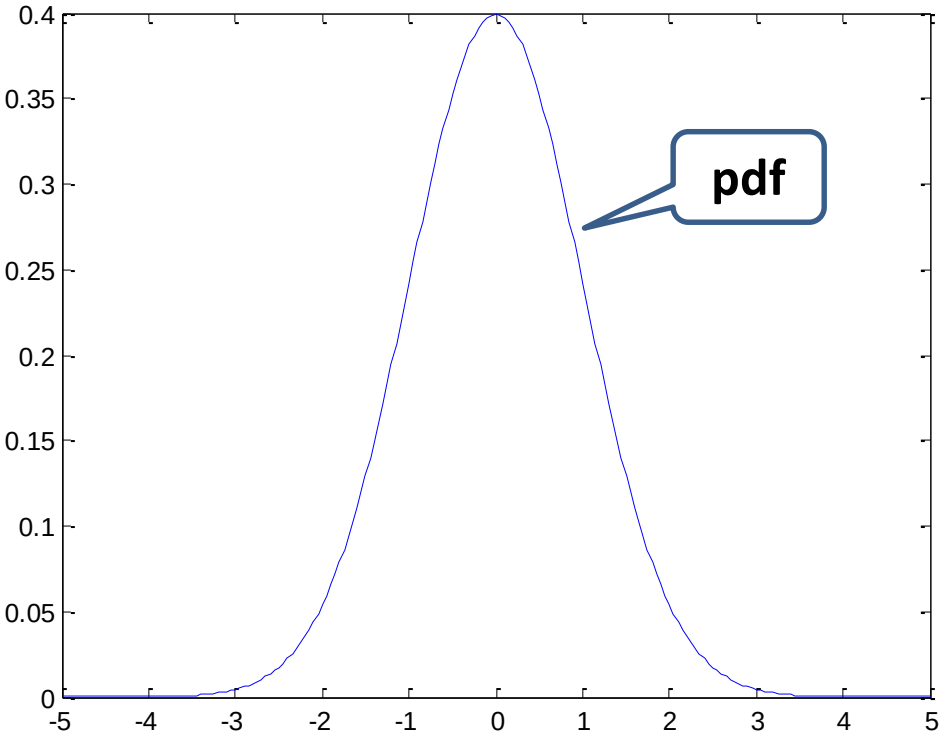
$$F_x(x) = \int_{-\infty}^{x} p_x(z)dz$$

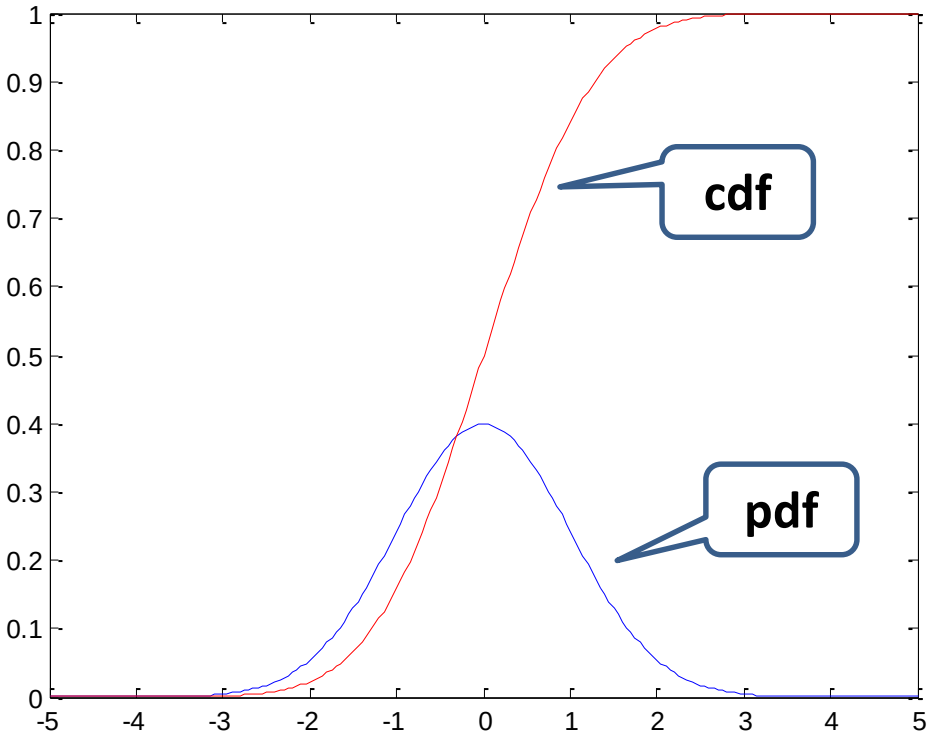
- **Probability of events** in terms of **pdf**:

  - $P(x \leq x) = F_x(x) = \int_{-\infty}^{x} p_x(z)dz$
  - $P(x_1 < x \leq x_2) = P(x \leq x_2) - P(x \leq x_1) = F_x(x_2) - F_x(x_1) = \int_{x_1}^{x_2} p_x(x)dx$
  - $P(x \geq x) = = P(x \leq \infty) - P(x \leq x) = 1 - P(x \leq x) = 1 - F_x(x) = \int_{-\infty}^{x} p_x(z)dz$

Probability and statistics: a brief review

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





Probability and statistics: a brief review

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

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

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

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

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*: $\mathrm{E}[\mathrm{x}] = \int_{-\infty}^{+\infty} x p(x) dx$
- **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*: $\mathrm{E}_{\mathrm{x,y}}[\, f(\mathrm{x}, \mathrm{y})] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) \, p(x, y) dx dy$
- **Conditional mean** of an rv *y* given *x* = *x*: $\mathrm{E}[\mathrm{y} \mid x] = \int_{-\infty}^{+\infty} y p(y \mid x) dy$
- It is $\mathrm{E}_{\mathrm{x,y}}[\, f(\mathrm{x}, \mathrm{y})] = E_{\mathrm{x}}[E_{\mathrm{y|x}}[\, f(\mathrm{x}, \mathrm{y})]]$
- **Covariance** between two **rvs** *x* and *y*: $\mathrm{cov}(\mathrm{x}, \mathrm{y}) = \mathrm{E}[(\mathrm{x} - \mathrm{E}[\mathrm{x}])(\mathrm{y} - \mathrm{E}[\mathrm{y}])]$
- **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

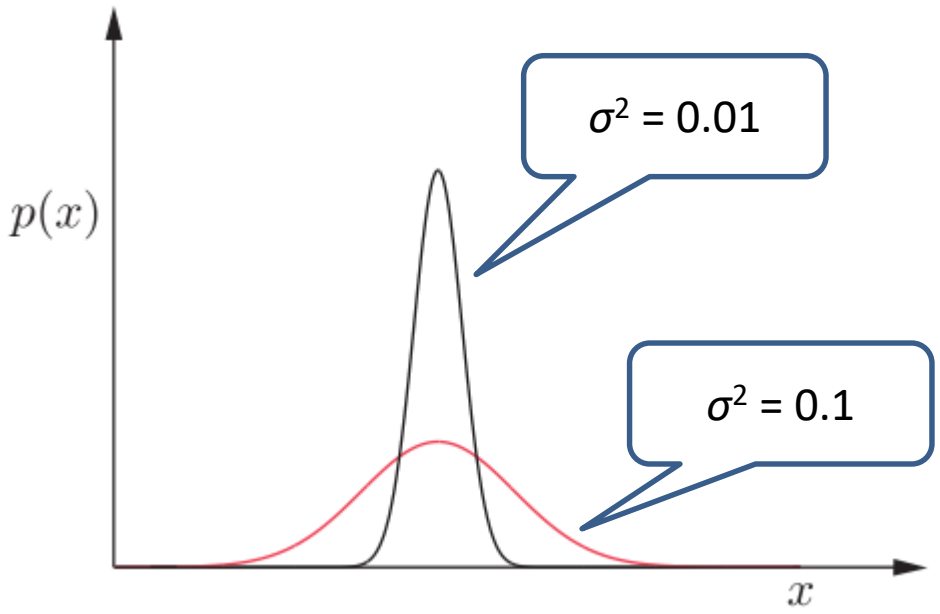
•**Remark:** Both R_x and cov(***x***) are symmetric and positive definite lxl matrices.

A square matrix A is symmetric iff $A^T=A$.

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

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
    - $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
    - $E[x]=\mu$
    - $\sigma_x^2 = \sigma^2$.



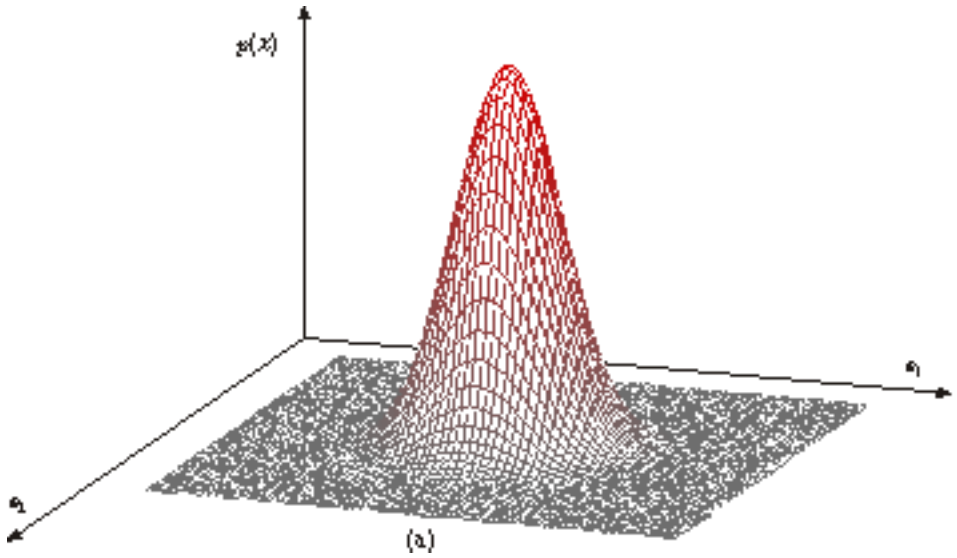
Probability and statistics: a brief review

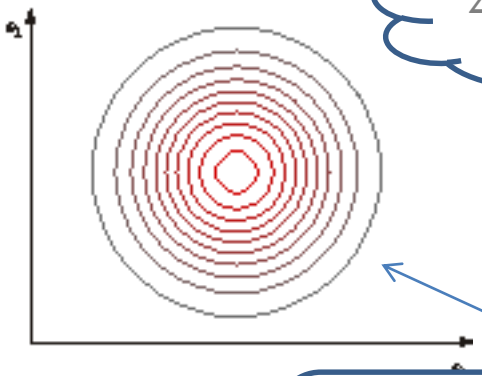
- **Multi dim. normal (Gaussian) distribution $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$:**
  - l-dim. case
  - It is
    - $p(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2})$
    - $E[\mathbf{x}] = \boldsymbol{\mu}$
    - $cov(\mathbf{x}) = \Sigma.$

(*) For the 2-d case $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$

Probability and statistics: a brief review

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



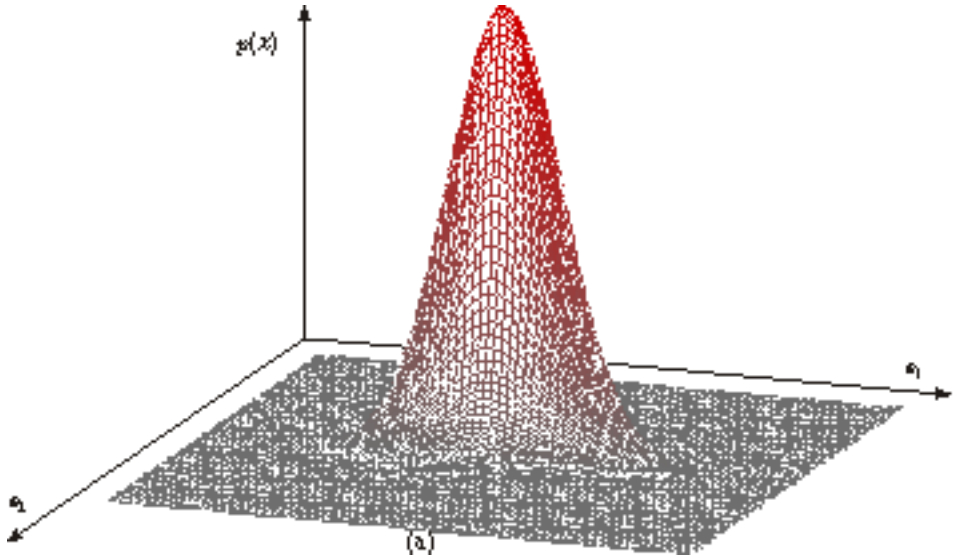


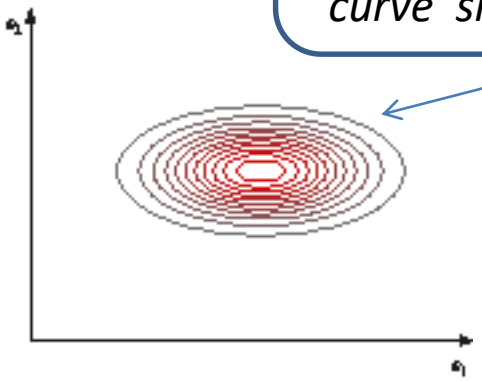
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Σ: **diagonal** with $\sigma_1^2 = \sigma_2^2$

Isovalued curves:

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



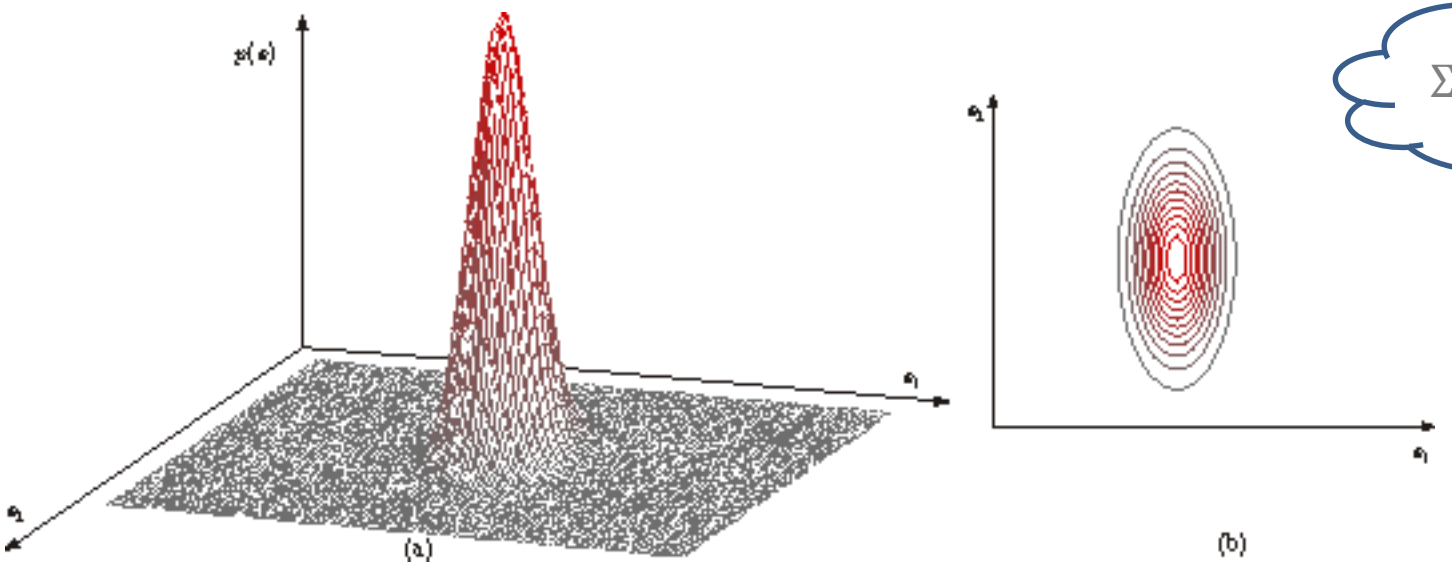


Σ: **diagonal** with $\sigma_1^2 \gg \sigma_2^2$

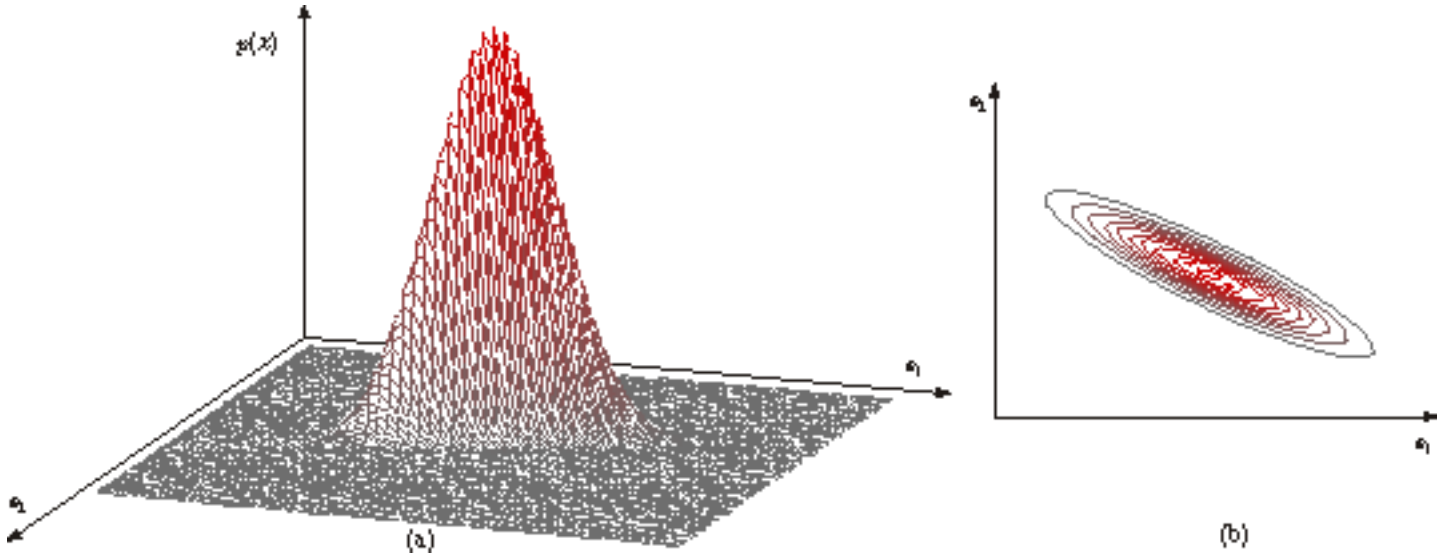
Probability and statistics: a brief review

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

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

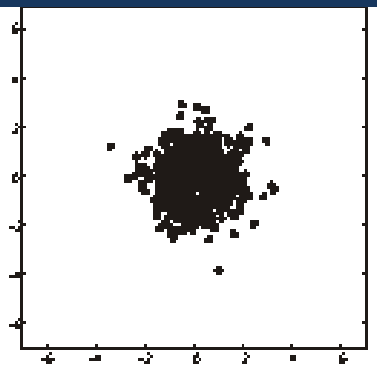


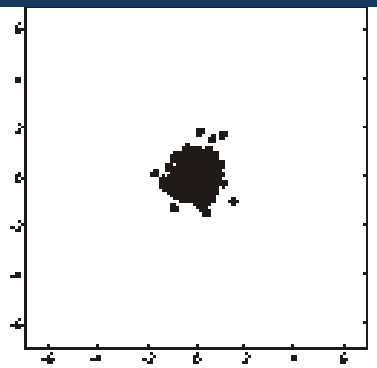
Σ: diagonal with $\sigma_1^2 \ll \sigma_2^2$

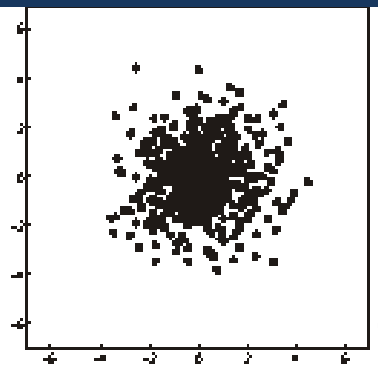


Σ: non diagonal

Probability and statistics: a brief review


(a)


(b)

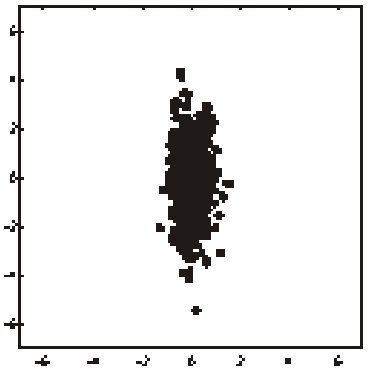

(c)

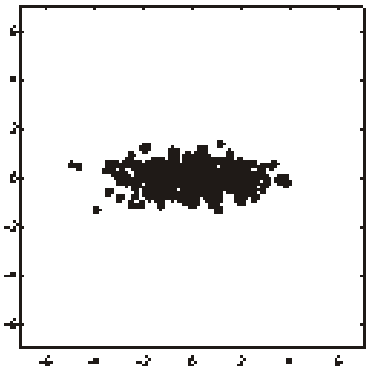
$$\Sigma = \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_2}^2 \end{bmatrix}$$

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

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

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

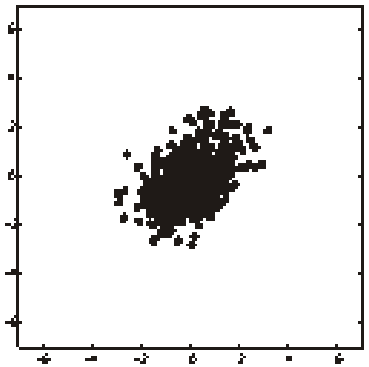

(d)

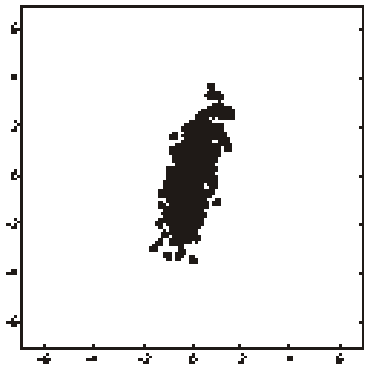

(e)

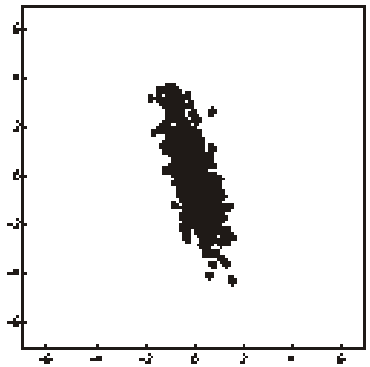
(d) $\sigma_1^2 = 0.2,\ \sigma_2^2 = 2, \sigma_{12} = 0$

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

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


(f)


(g)


(h)

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

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

Probability and statistics: a brief review

Continuous RV distributions (cont.)

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

From 1-dim. → 2-dim. case.

- 1-dim. case: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{(x-\mu)\sigma^{-2}(x-\mu)}{2}\right)$

- A first extension to the 2-dim. case (independent rv's):
- $p(x_1, x_2) = p_1(x_1) \cdot p_2(x_2) =$
- $\frac{1}{(2\pi)^{1/2} \cdot \sigma_1} \exp\left(-\frac{(x_1-\mu_1)\sigma_1^{-2}(x_1-\mu_1)}{2}\right) \cdot \frac{1}{(2\pi)^{1/2} \cdot \sigma_2} \exp\left(-\frac{(x_2-\mu_2)\sigma_2^{-2}(x_2-\mu_2)}{2}\right) =$

$$\frac{1}{(2\pi)|\Sigma|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$
$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

- The final extension to the 2-dim. case (dependent rv's):
- $p(x_1, x_2) = \frac{1}{(2\pi)|\Sigma|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Probability and statistics: a brief review

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

  - **Properties**

  1. If the covariance matrix Σ is diagonal, then, the rv's $\mathrm{x}_1, \ldots, \mathrm{x}_l$ comprising $\mathbf{x}$ are statistically independent. It is

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} p_i(x_i) = \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi\sigma_i^2}} exp\left(-\frac{(x_i - \mu_i)^2}{2{\sigma_i}^2}\right)$$

  2. *Central limit theorem:*

  Let:

  - $\mathrm{x}_1, \ldots, \mathrm{x}_r$ independent rvs following different distributions
  - $\mu_i, {\sigma_i}^2$ mean and variance of x_i.
  - Define $\mathrm{x} = \mathrm{x}_1 + \cdots + \mathrm{x}_r$, $\mu = \mu_1 + \cdots + \mu_r$, $\sigma^2 = \sigma_1^2 + \cdots + \sigma_r^2$.
  - Define $z = (\mathrm{x} - \mu)/\sigma$.

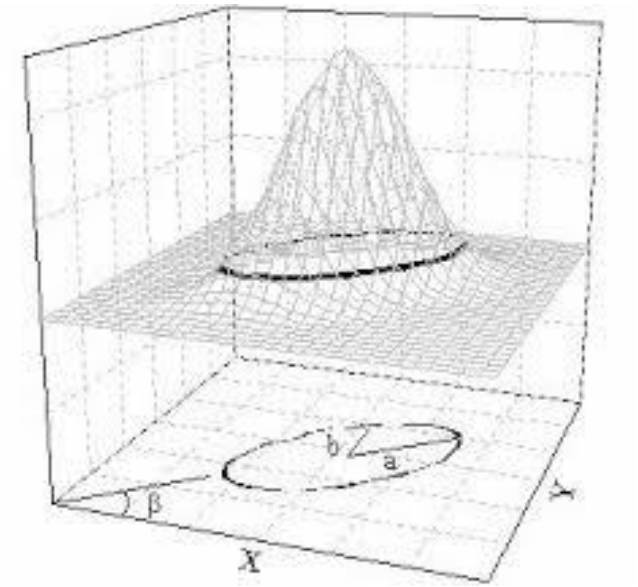
  Then

  - $p(z) \rightarrow N(z|0,1)$, as $r \rightarrow \infty$

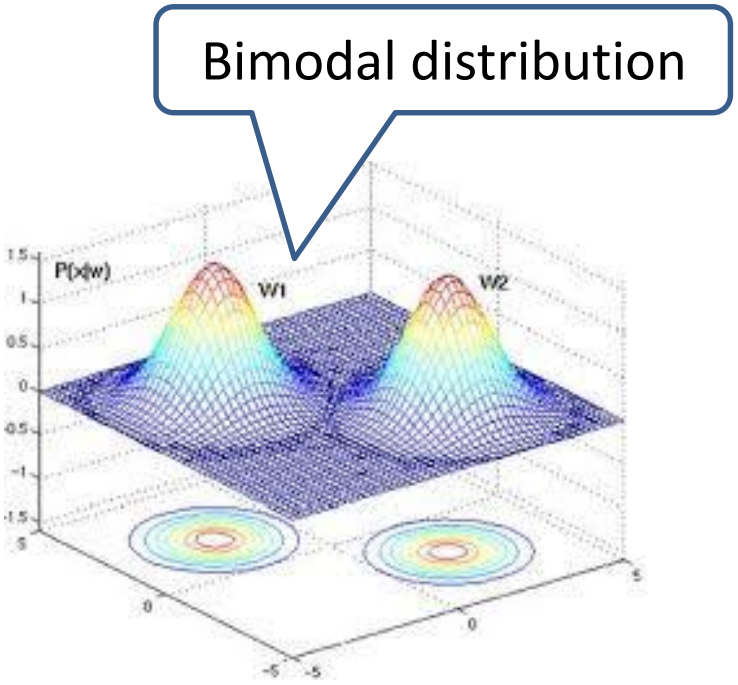
Probability and statistics: a brief review

Continuous RV distributions (cont.)

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



Two-dim. pdfs



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

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

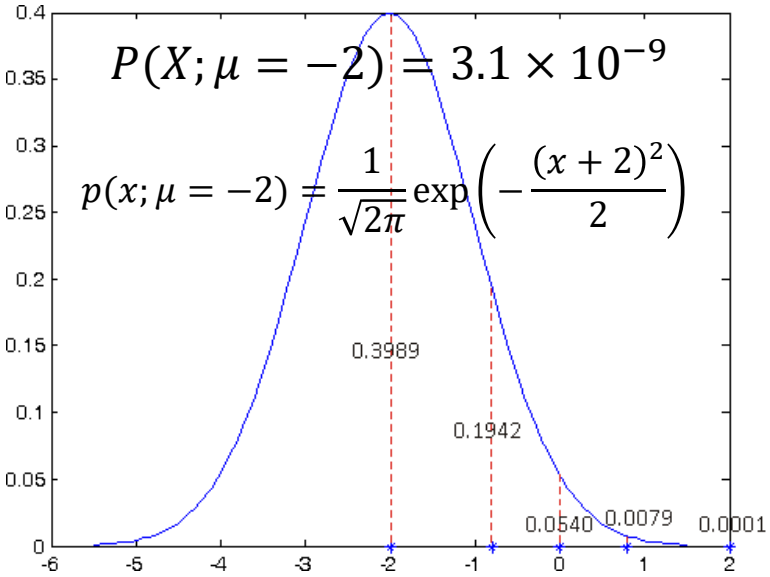
Example:

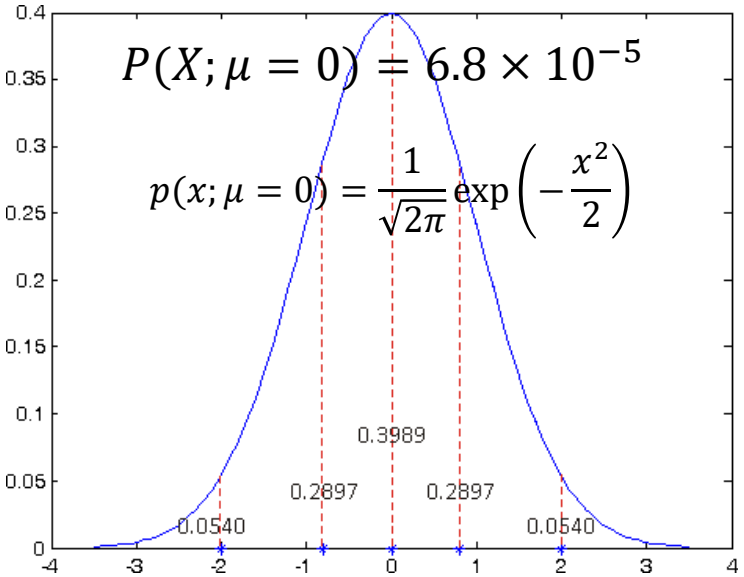
- $X = \{-2, -1, 0, 1, 2\}$
- Consider the parametric set of normal distributions of unit variance, parameterized on μ.
- The likelihood of μ wrt X is

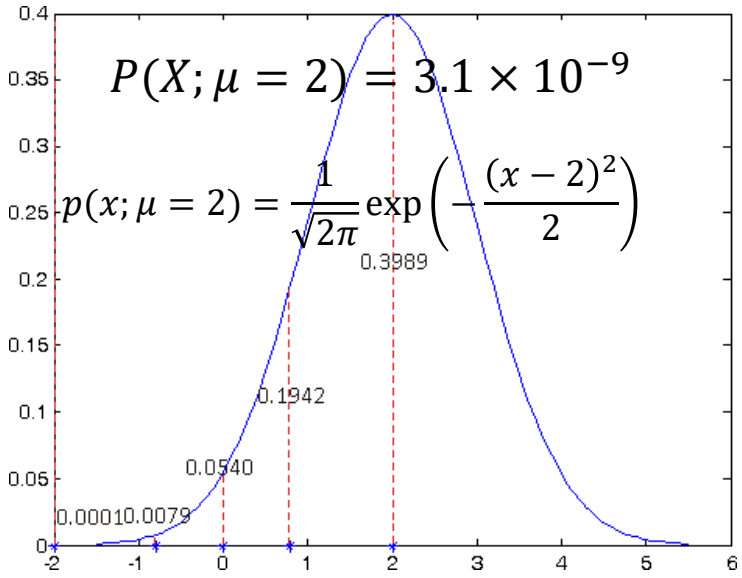
$$p(X;\mu) = p(-2,-1,0,1,2;\mu) =$$
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-2-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-1-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(0-\mu)^2}{2}\right)$$
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}}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

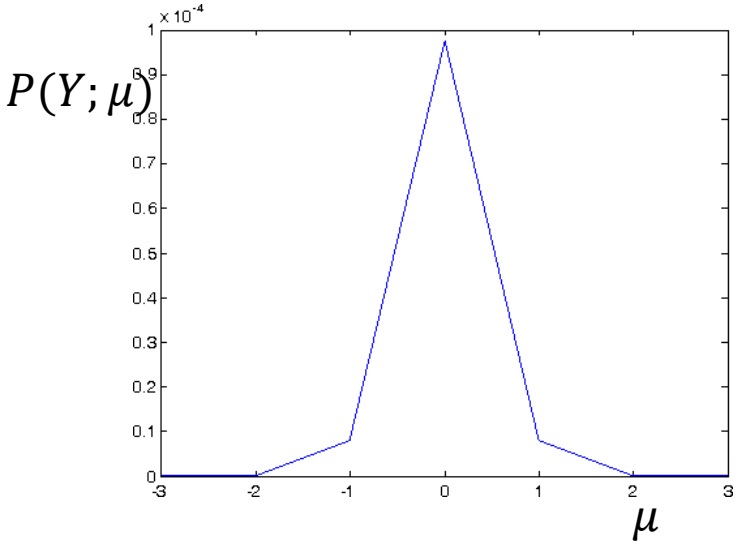
Probability and statistics: a brief review

Likelihood function







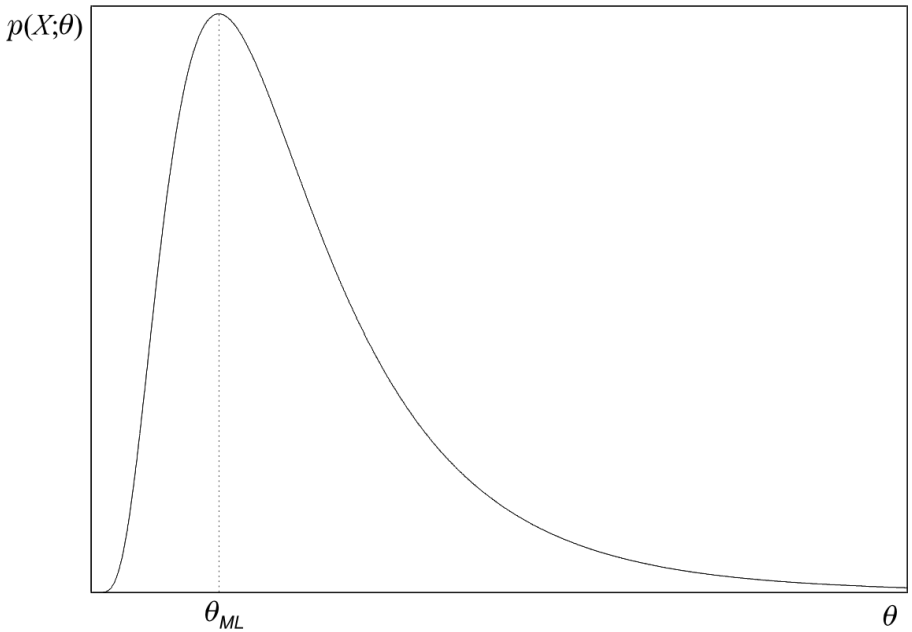


Probabilistic CFO clustering algorithms

Maximum likelihood (ML) method:

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

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



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 $\boldsymbol{\theta}_o$,

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 $\boldsymbol{\theta}_o$, 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

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

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

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

Maximum likelihood (ML) method:

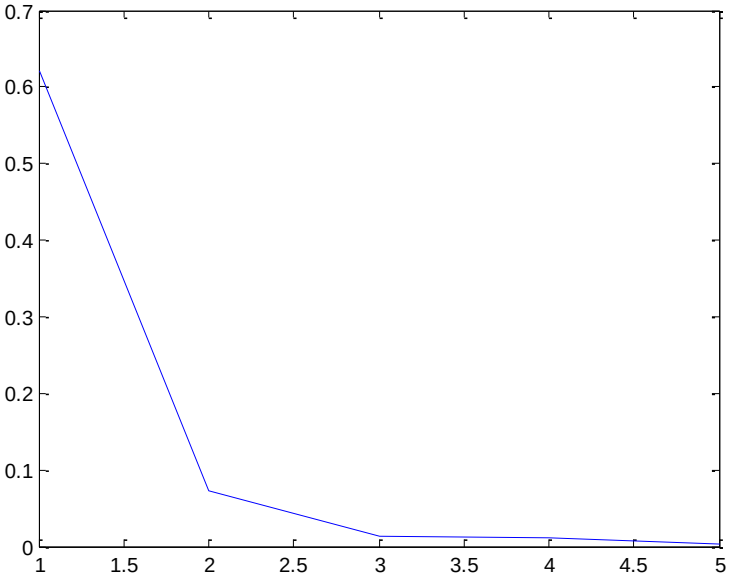
Example:

- Let $N = 10$ points resulting from the 1-dim. **zero** mean, **unit** variance normal distribution, $N(0,1)$.
- Let us **pretend** that we have at our disposal the following:
  - The knowledge that the distribution that generated the data is a normal one with variance equal to 1 and unknown mean vector μ.
  - the 10 points.

➢ The **ML estimate** of the mean vector, $\hat{\mu}$, of the distribution is 0.6210 (the true value is 0).

The more the available data points, the more accurate the estimates for the mean vector.

N	Estimation error
10	$0.6210 - 0$
100	$0.0727 - 0$
1000	$0.0138 - 0$
10000	$0.0118 - 0$
100000	$0.0034 - 0$

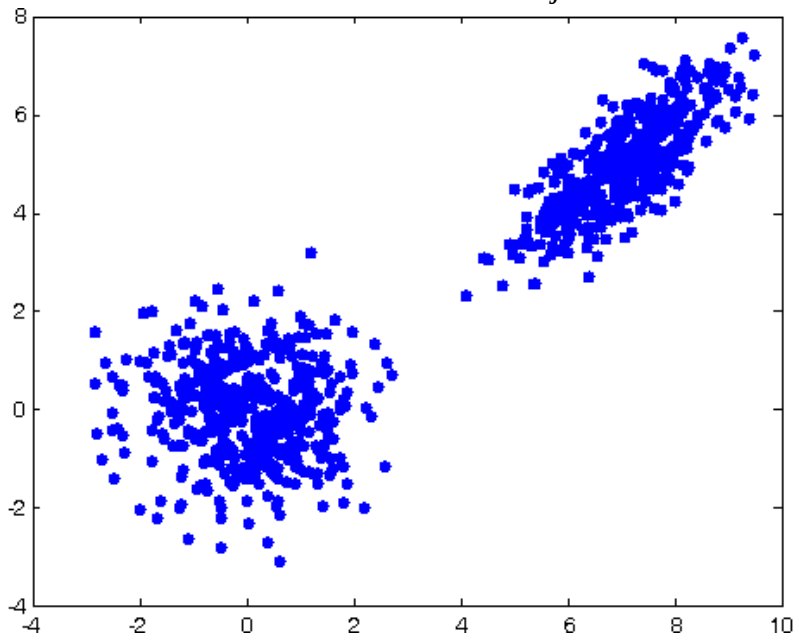


Probabilistic CFO clustering algorithms

Mixture models - The Expectation – Maximization (EM) algorithm

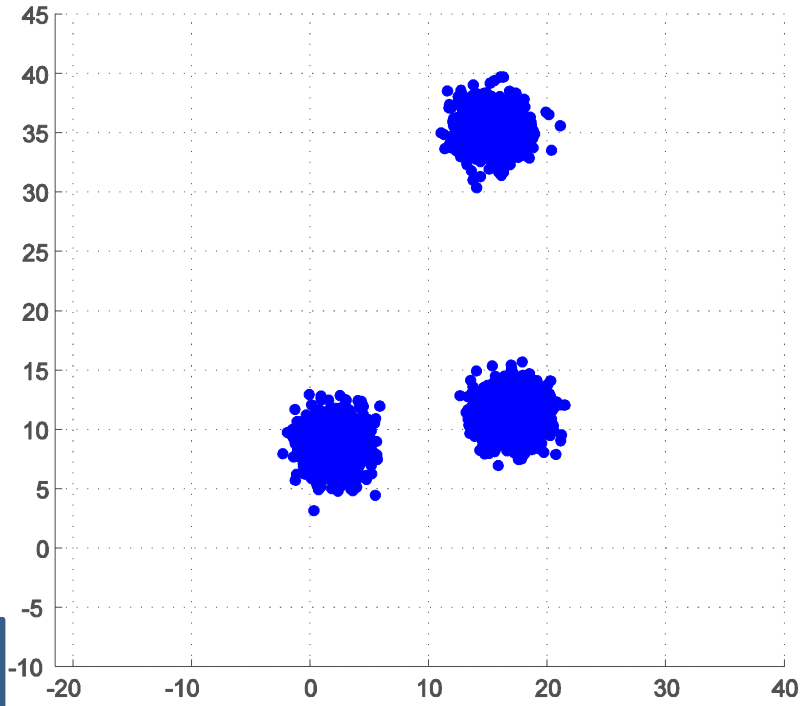
Mixture model: A weighted sum of known parametric form pdfs.

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



- 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**).

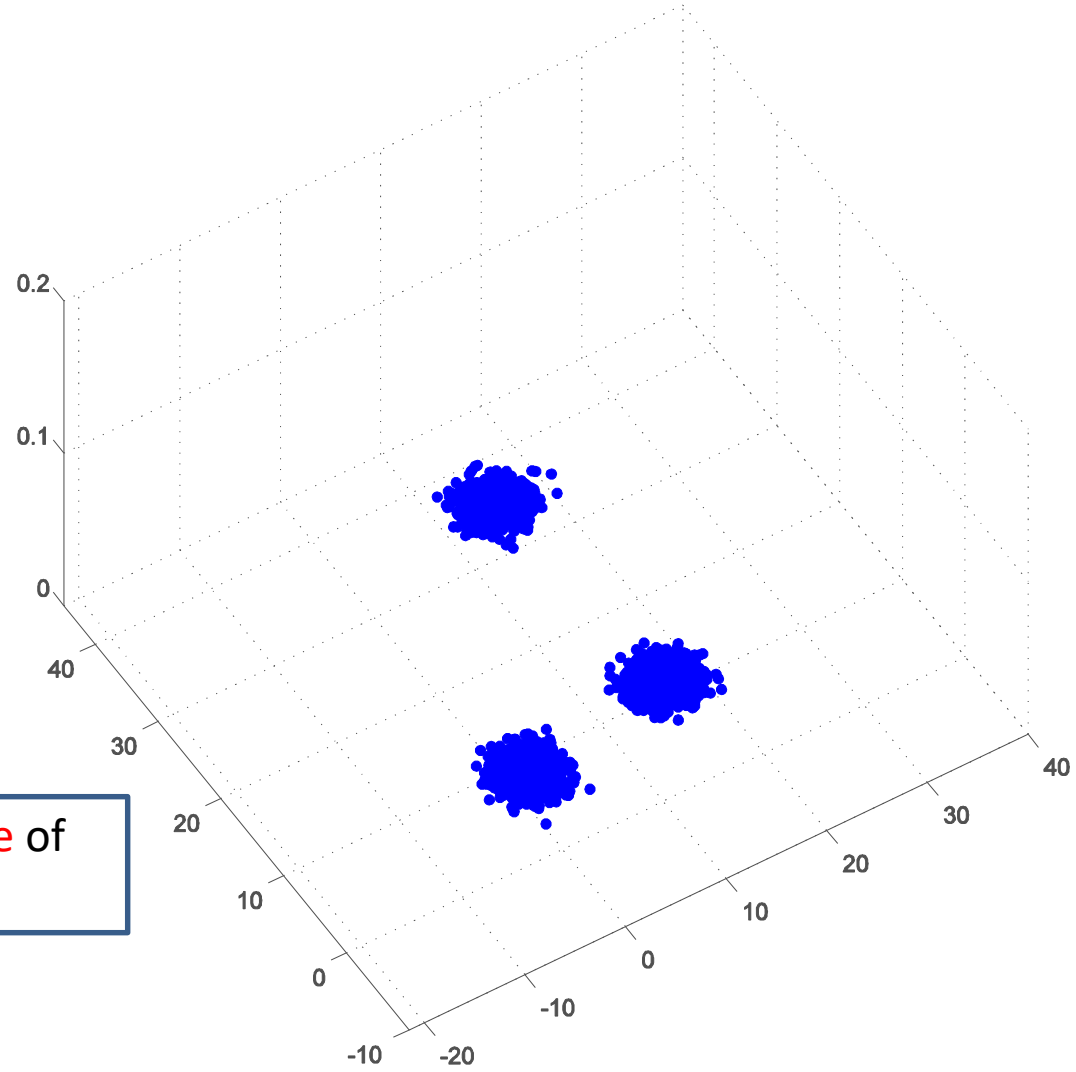
Probabilistic CFO clustering algorithms



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.

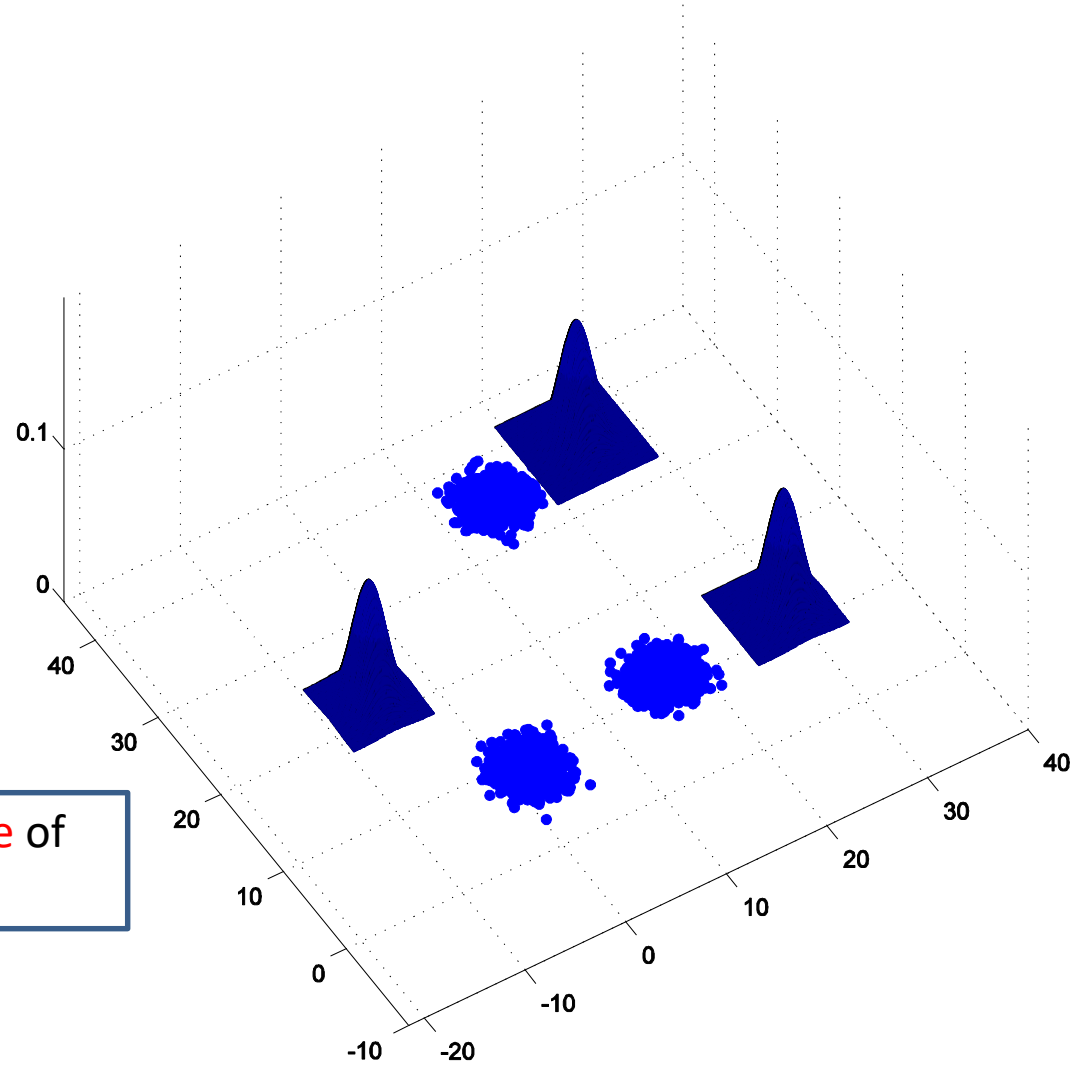
Probabilistic CFO clustering algorithms



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.

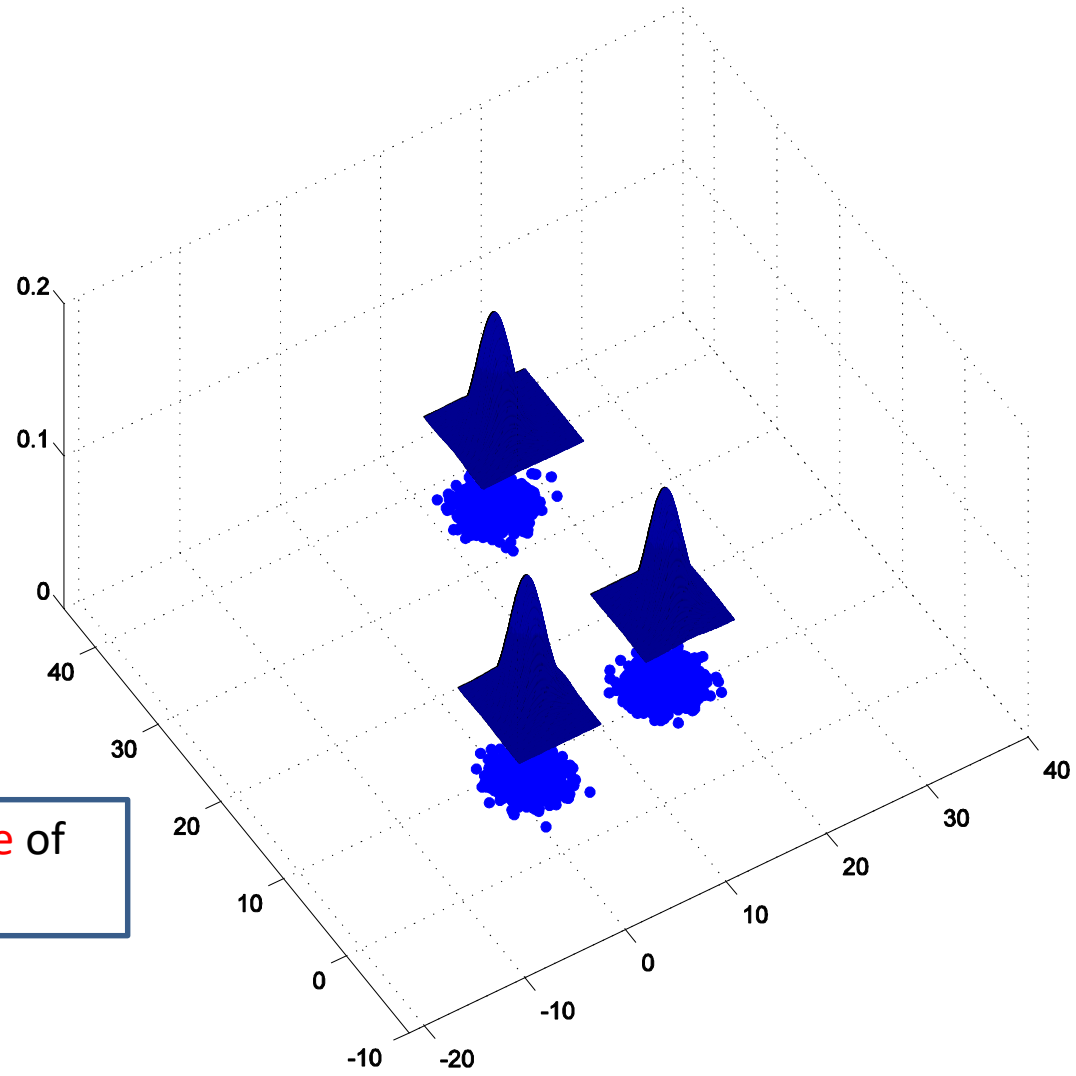
Probabilistic CFO clustering algorithms



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.

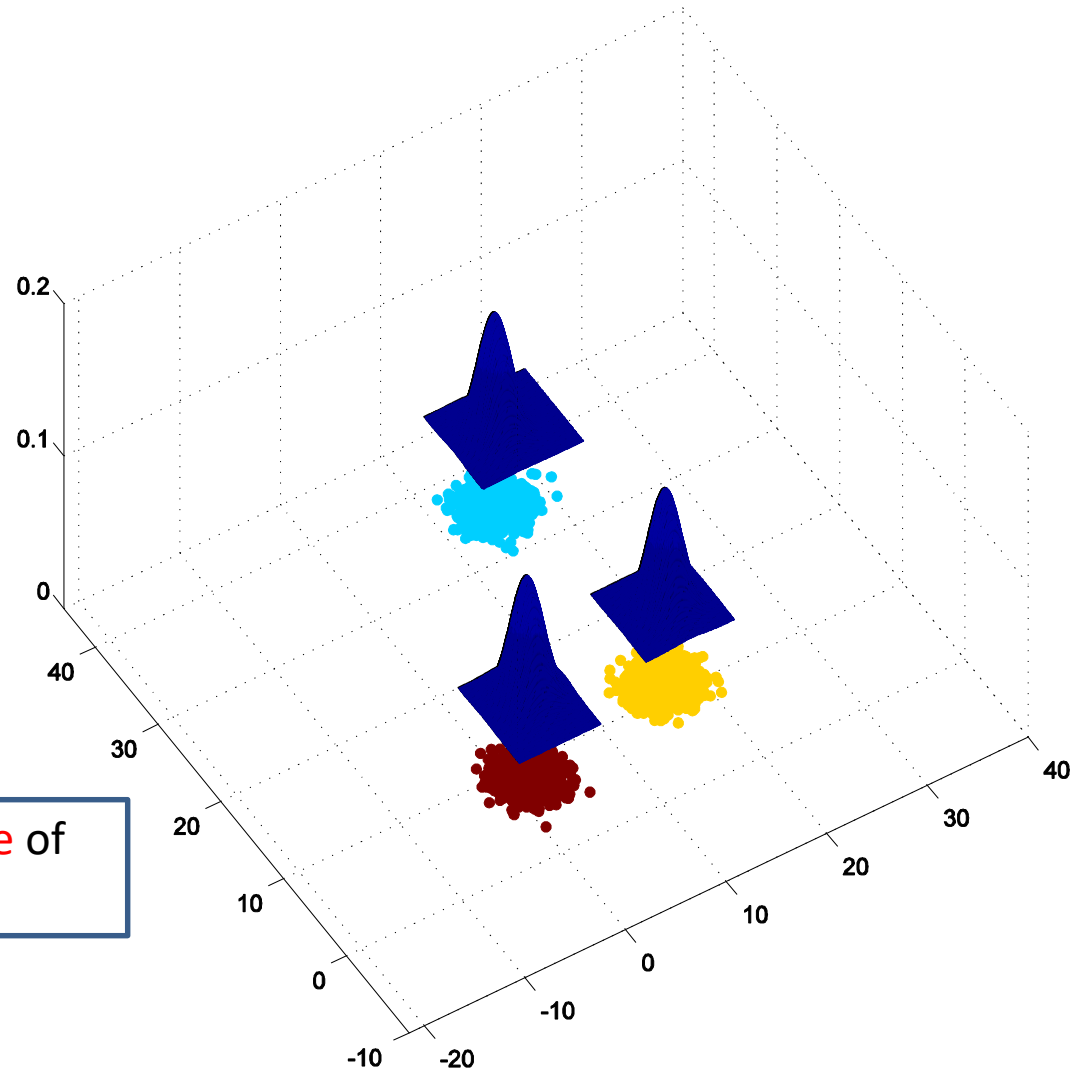
Probabilistic CFO clustering algorithms



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 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ 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:
$\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m\}$

$P = \{P_1, P_2, \ldots, P_m\}$, 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 $P(j|\boldsymbol{x}) = \frac{p(\boldsymbol{x},j)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|j)P_j}{p(\boldsymbol{x})}$

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

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

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

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

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

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

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 $P(j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|j;\boldsymbol{\theta}_j)P_j}{p(\boldsymbol{x};\boldsymbol{\Theta})}$

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, $j = 1, \ldots, m$
- t=0
- **Repeat**
  - For i=1 to N *% Expectation step*
    - For j=1 to m

$$P(j|\boldsymbol{x}_i; \Theta^{(t)}, P^{(t)}) = \frac{p(x_i|j;\theta_j{}^{(t)})P_j{}^{(t)}}{\sum_{q=1}^{m} p(x_i|q;\theta_q{}^{(t)})P_q{}^{(t)}} \equiv \gamma_{ji}{}^{(t)}$$

    - End {For-j}
  - End {For-i}
  - t=t+1
  - For j=1 to m *% Parameter updating – Maximization step*
    - Set

$$\boldsymbol{\theta}_j{}^{(t)} = argmax_{\boldsymbol{\theta}_j} \sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)} \ln\left(p\left(\boldsymbol{x}_i|j;\boldsymbol{\theta}_j\right)\right), j = 1, \ldots, m$$

$$P_j{}^{(t)} = \frac{1}{N}\sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)}, j = 1, \ldots, m$$

  - End {For-j}

- **Until** a termination criterion is met.

Probabilistic CFO clustering algorithms

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

GPrAS – The case of normal pdfs

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

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

$$\Sigma_j=\frac{\sum_{i=1}^{N}P(j|\boldsymbol{x_i})(\boldsymbol{x_i}-\boldsymbol{\mu_j})\,(\boldsymbol{x_i}-\boldsymbol{\mu_j})^T}{\sum_{i=1}^{N}P(j|\boldsymbol{x_i})}$$

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., $j = 1, \ldots, m$
- t=0
- **Repeat**
  - For i=1 to N *% Expectation step*
    - For j=1 to m

$$P(j|\boldsymbol{x}_i; \Theta^{(t)}, P^{(t)}) = \frac{p(x_i|j;\theta_j{}^{(t)})P_j{}^{(t)}}{\sum_{q=1}^{m} p(x_i|q;\theta_q{}^{(t)})P_q{}^{(t)}} \equiv \gamma_{ji}{}^{(t)}$$

    - End {For-j}
  - End {For-i}
  - t=t+1
  - For j=1 to m *% Parameter updating – Maximization step*
    - Set

$$\boldsymbol{\mu}_j{}^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)} \boldsymbol{x_i}}{\sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)}}, \qquad \Sigma_j{}^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)} (\boldsymbol{x_i} - \boldsymbol{\mu_j})(\boldsymbol{x_i} - \boldsymbol{\mu_j})^{\boldsymbol{T}}}{\sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)}} \; j = 1, \ldots, m$$

$$P_j{}^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ji}{}^{(t-1)}, j = 1, \ldots, m$$

  - End {For-j}
- **Until** a termination criterion is met.

Probabilistic CFO clustering algorithms

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*
    - $$P(C_j|\boldsymbol{x}; \Theta(t)) = \frac{|\Sigma_j(t)|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu_j(t))^T \Sigma_j^{-1}(t)(\boldsymbol{x} - \mu_j(t))\right) P_j(t)}{\sum_{k=1}^{m} |\Sigma_k(t)|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu_k(t))^T \Sigma_k^{-1}(t)(\boldsymbol{x} - \mu_k(t))\right) P_k(t)}$$
    - End {For-j}
  - End {For-i}
  - t=t+1
  - For j=1 to m % *Parameter updating – Maximization step*
    - Set

$$\boldsymbol{\mu}_j^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{ji}^{(t-1)} \boldsymbol{x_i}}{\sum_{i=1}^{N} \gamma_{ji}^{(t-1)}}, \qquad \Sigma_j^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{ji}^{(t-1)} (\boldsymbol{x_i} - \boldsymbol{\mu_j})(\boldsymbol{x_i} - \boldsymbol{\mu_j})^T}{\sum_{i=1}^{N} \gamma_{ji}^{(t-1)}} \quad j = 1, \ldots, m$$

$$P_j^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ji}^{(t-1)}, j = 1, \ldots, m$$

  - 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:

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

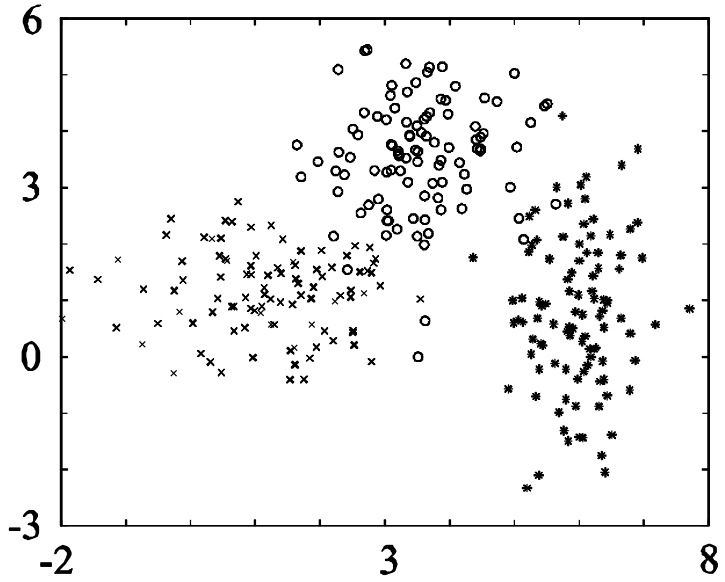
and covariance matrices

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

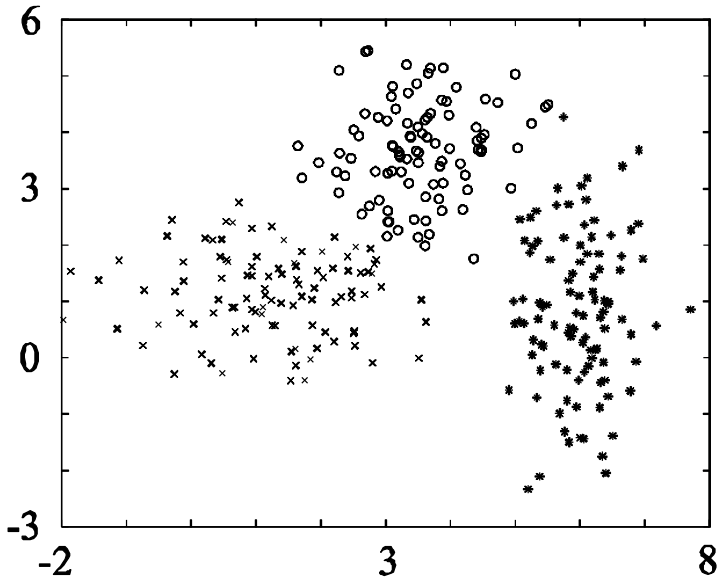
respectively.

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

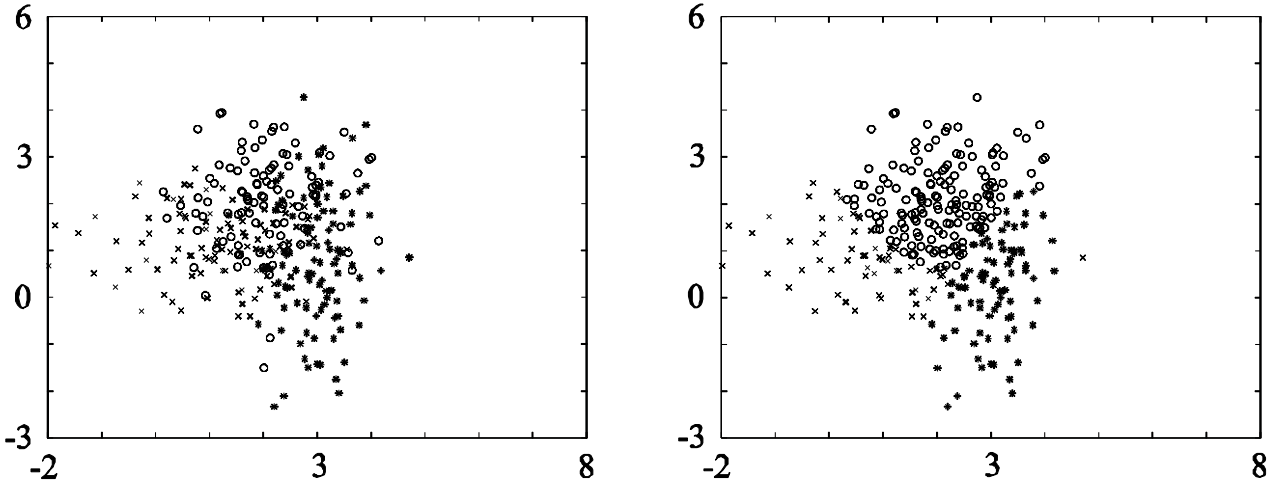
Confusion matrix:

	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

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

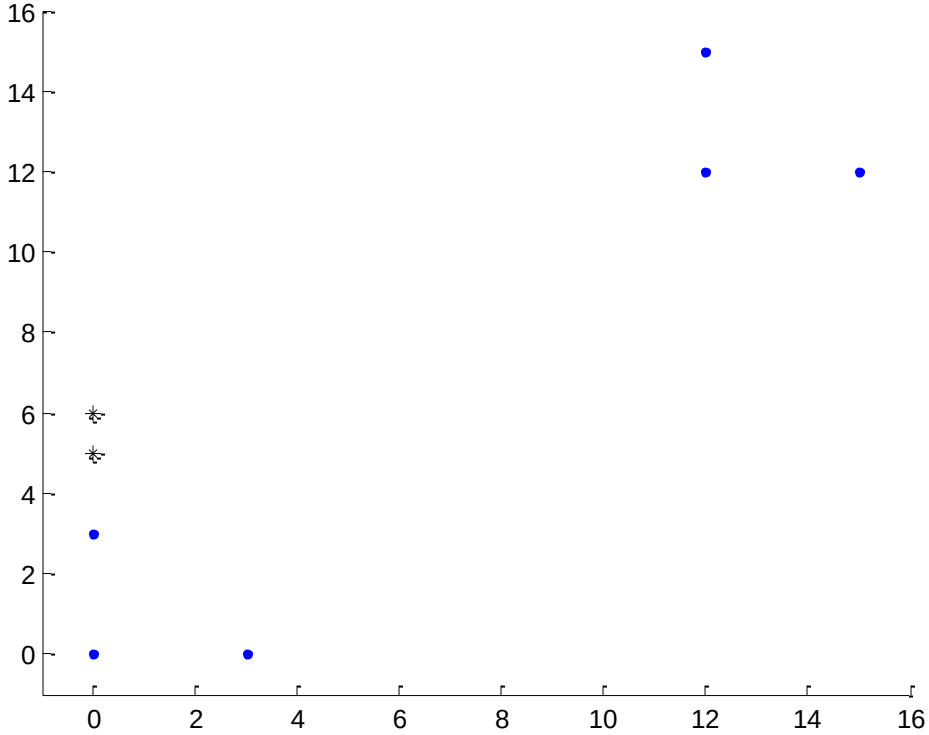
Initially:

$\theta_1(0) = [0, 5]^{\mathrm{T}}$

$\theta_2(0) = [0, 6]^{\mathrm{T}}$

$\mathrm{P}_1(0) = 0.1$

$P_2(0) = 0.9$



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

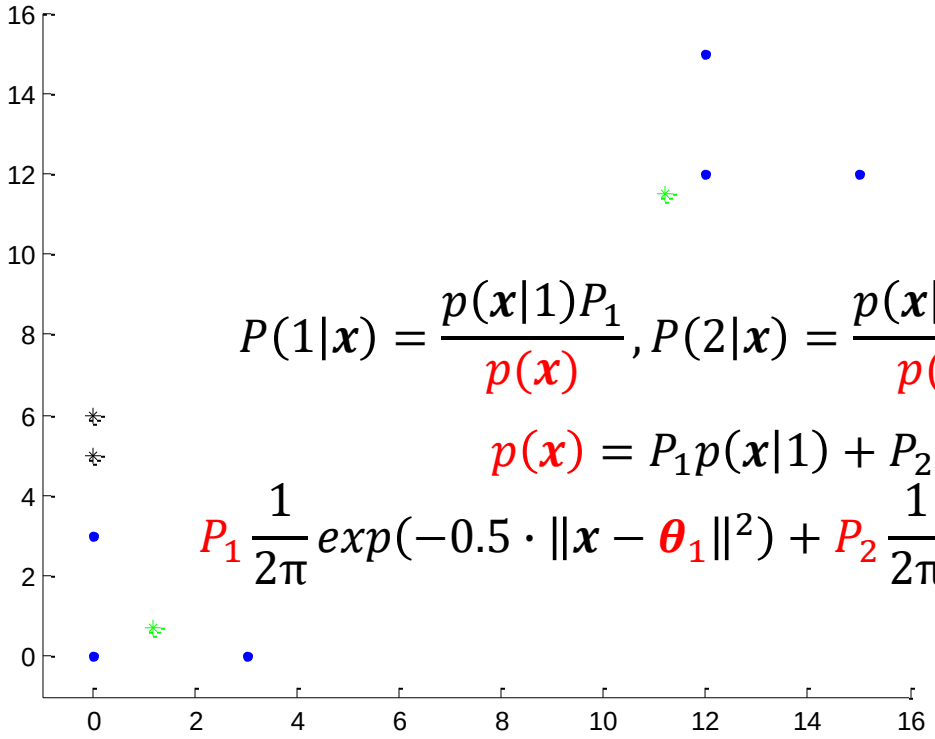
$$p(\boldsymbol{x}|2) = \frac{1}{2\pi} exp(-0.5 \cdot \|\boldsymbol{x} - \boldsymbol{\theta}_2\|^2), \qquad P(2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|2)P_2}{p(\boldsymbol{x})}$$

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

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

Probabilistic CFO clustering algorithms

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



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

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

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

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

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

1st iteration (in more detail):

A. Expectation step - A posteriori probs

$$P(1|\boldsymbol{x}_1) = \frac{P_1(0)\frac{1}{2\pi}exp(-0.5 \cdot \|\boldsymbol{x}_1 - \boldsymbol{\theta}_1(0)\|^2)}{P_1(0)\frac{1}{2\pi}exp(-0.5 \cdot \|\boldsymbol{x}_1 - \boldsymbol{\theta}_1(0)\|^2) + P_2(0)\frac{1}{2\pi}exp(-0.5 \cdot \|\boldsymbol{x}_1 - \boldsymbol{\theta}_2(0)\|^2)} =$$

$$\frac{0.1\frac{1}{2\pi}exp\left(-0.5 \cdot \left\|\begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}0\\5\end{bmatrix}\right\|^2\right)}{0.1\frac{1}{2\pi}exp\left(-0.5 \cdot \left\|\begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}0\\5\end{bmatrix}\right\|^2\right) + 0.9\frac{1}{2\pi}exp\left(-0.5 \cdot \left\|\begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}0\\6\end{bmatrix}\right\|^2\right)} =$$

$$\frac{0.1\frac{1}{2\pi}exp(-12.5)}{0.1\frac{1}{2\pi}exp(-12.5) + 0.9\frac{1}{2\pi}exp(-18)} = \mathbf{0.9645}$$

In a similar way we have

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

Probabilistic CFO clustering algorithms

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

1st iteration (in more detail):

B. Maximization step - Parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{P}_1$, $\boldsymbol{P}_2$

$$\boldsymbol{\theta}_1(1) = \frac{0.9645 \cdot \boldsymbol{x}_1 + 0.9645 \cdot \boldsymbol{x}_2 + 0.5751 \cdot \boldsymbol{x}_3 + 0.0002 \cdot \boldsymbol{x}_4 + 0.0002 \cdot \boldsymbol{x}_5 + 0.0000 \cdot \boldsymbol{x}_6}{0.9645 + 0.9645 + 0.5751 + 0.0002 + 0.0002 + 0.0000} =$$

$$\frac{0.9645 \cdot \begin{bmatrix}0\\0\end{bmatrix} + 0.9645 \cdot \begin{bmatrix}3\\0\end{bmatrix} + 0.5751 \cdot \begin{bmatrix}0\\3\end{bmatrix} + 0.0002 \cdot \begin{bmatrix}12\\12\end{bmatrix} + 0.0002 \cdot \begin{bmatrix}15\\12\end{bmatrix} + 0.0000 \cdot \begin{bmatrix}12\\15\end{bmatrix}}{0.9645 + 0.9645 + 0.5751 + 0.0002 + 0.0002 + 0.0000} =$$

$$= \begin{bmatrix}1.1572\\0.6906\end{bmatrix}$$

In a similar way we have

$$\boldsymbol{\theta}_2(1) = \begin{bmatrix}11.1864\\11.5207\end{bmatrix}$$

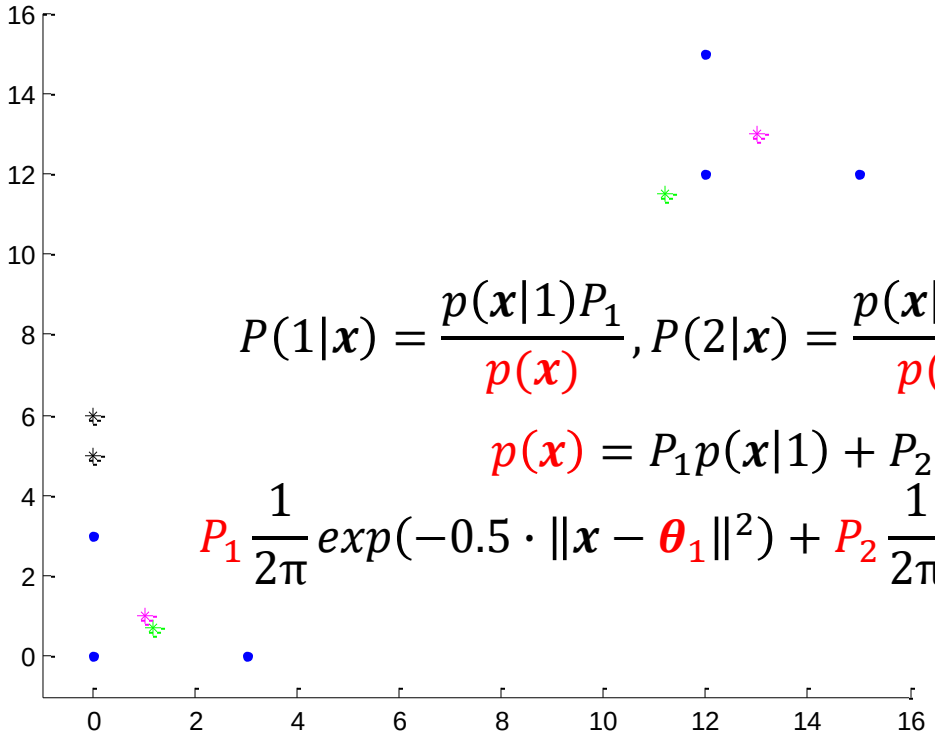
$$P_1(1) = \frac{0.9645 + 0.9645 + 0.5751 + 0.0002 + 0.0002 + 0.0000}{6} = 0.4174$$

In a similar way we have

$$P_2(1) = 0.5826$$

Probabilistic CFO clustering algorithms

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



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

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

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

2nd iteration:

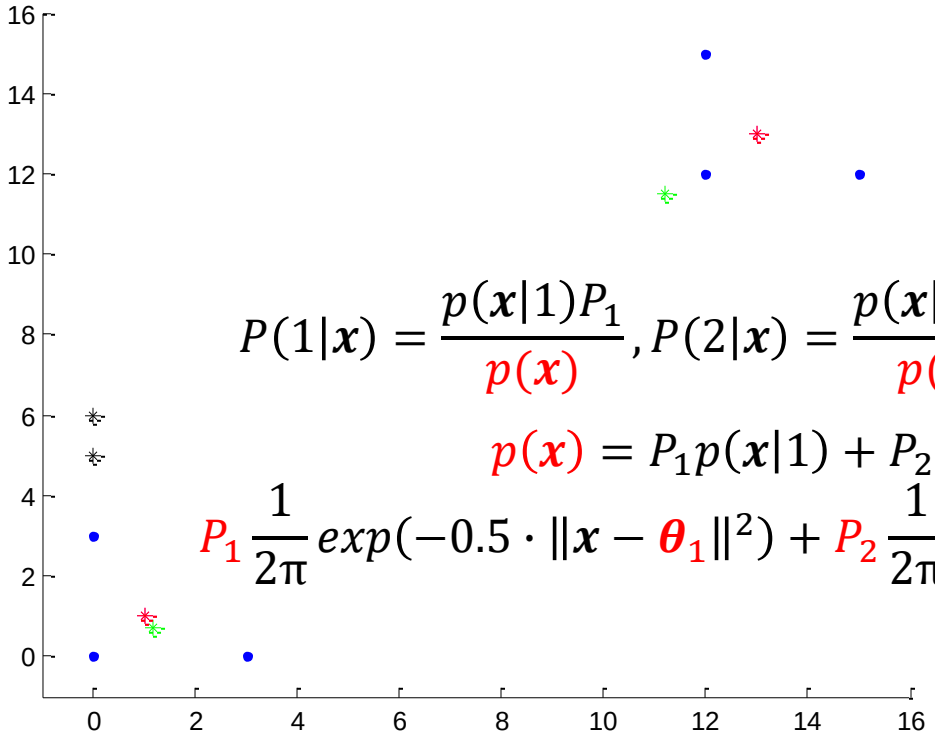
A posteriori probs

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

Example $x_1 = [0\ 0]^T, x_2 = [3\ 0]^T x_3 = [0\ 3]^T, x_4 = [12\ 12]^T, x_5 = [15\ 12]^T, x_6 = [12\ 15]^T$



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

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

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

3rd iteration:

A posteriori probs

	x_1	x_2	x_3	x_4	x_5	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

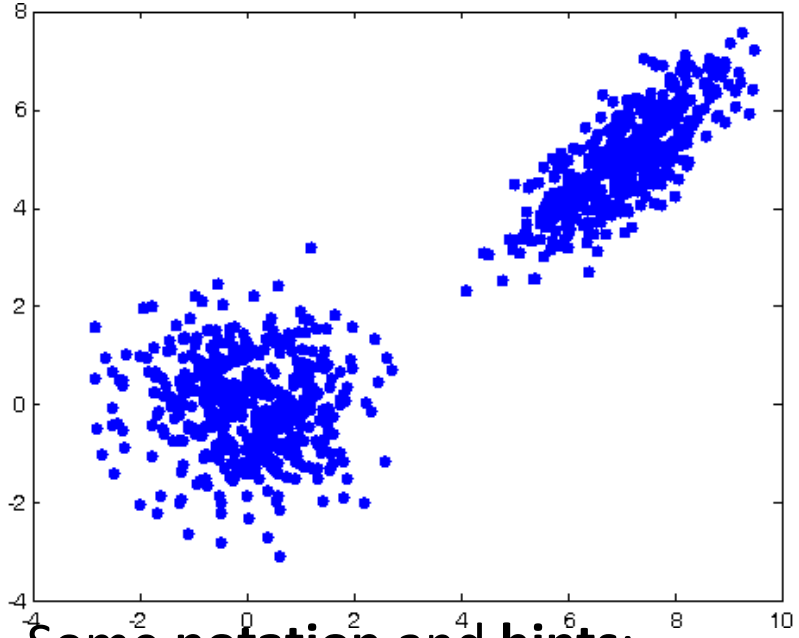
$$\boldsymbol{\theta}_1(3) = [1\ 1]^{\mathrm{T}}$$
$$\boldsymbol{\theta}_2(3) = [13\ 13]^{\mathrm{T}}$$
$$\mathrm{P}_1(3) = 0.5$$
$$P_2(3) = 0.5$$

Probabilistic CFO clustering algorithms

An alternative view: 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**).
- The ML method cannot be used **directly** here due to our ignorance of the pdf label from which each data point stems.
- **Solution:** The **EM** algorithm.

Some **notation** and **hints**:

- $p(\boldsymbol{x}|j) = p(\boldsymbol{x}|j;\boldsymbol{\theta}_j)$
- $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_m)$, $\boldsymbol{P}=[P_1,\ldots,P_m]^T$
- $X=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$: incomplete (**observed**) data set.
- $X^c=\{(\boldsymbol{x}_1,j_1),\ldots,(\boldsymbol{x}_N,j_N)\}$: complete data set (the class labels $j_1, \ldots, j_N$, are unknown)

Probabilistic CFO clustering algorithms

An alternative view: Mixture models - The Expectation – Maximization (EM) algorithm

Aim: Estimate ***Θ*** and ***P*** via the maximization of the expectation of the log-likelihood **conditioned** on the observed data.

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

$$\sum_{n=1}^{N}\sum_{j=1}^{m}P(j\mid \boldsymbol{x}_n;\boldsymbol{\theta}_j)\ln\left(p(\boldsymbol{x}_n\mid j;\boldsymbol{\theta}_j)P_j\right)$$

This is **exactly** the cost function we defined in the beginning!!! The Expectation-Maximization (EM) algorithm is actually the GPrAS

In the special case where individual pdfs are Gaussians, we have mixtures of Gaussians.

$$p(\boldsymbol{x})=\sum_{j=1}^{m}P_j\,p(\boldsymbol{x}\mid j;\boldsymbol{\mu}_j,\Sigma_j),$$

$$p(\boldsymbol{x}\mid j;\boldsymbol{\mu}_j,\Sigma_j)=\frac{1}{(2\pi)^{l/2}\mid\Sigma_j\mid^{1/2}}\exp\left(-0.5\cdot(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)$$