Data analysis and stochastic modeling

Lecture 5 – Mixture models and the EM algorithm

Guillaume Gravier

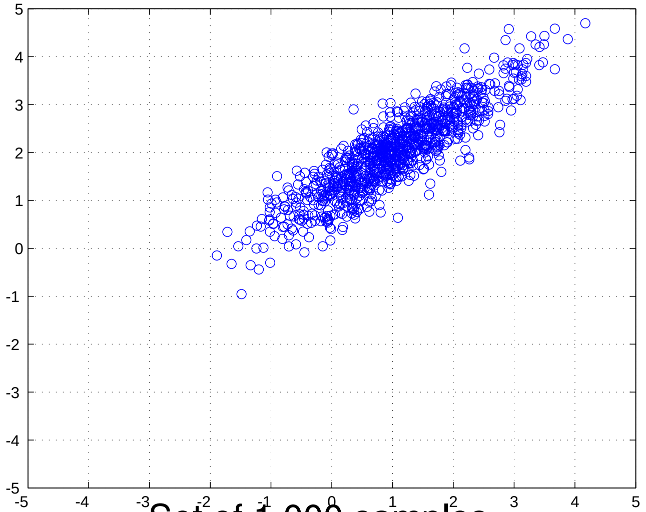
guillaume.gravier@irisa.fr

Multivariate Gaussian density

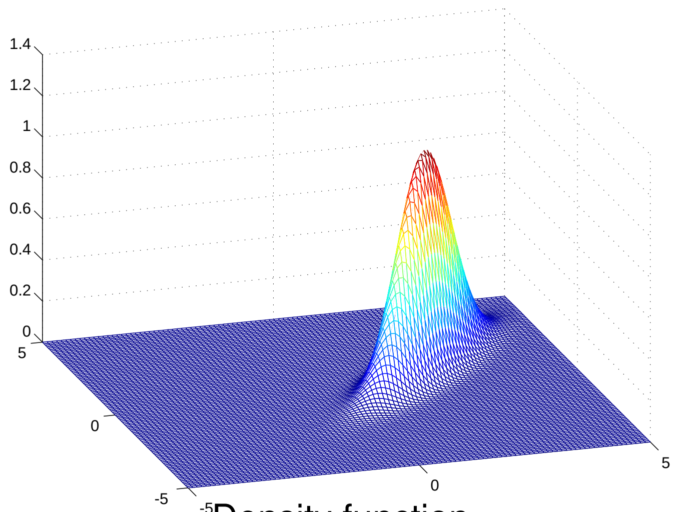
Definition

X is a Gaussian vector of dimension p if any linear combination of its components $a'X$ is a Gaussian in dimension 1.

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}$$



Set of 1,000 samples



Density function

Example of a multivariate Gaussian with $m = [21]$ and $\theta = \pi/6$.

Properties of the covariance matrix

The covariance matrix is symmetric, definite positive,

$$\Sigma = V \; D \, V'$$

where

- V are the eigen vectors defining the principal axes (orientation of the density) and
- D are the eigen values defining the dispertion along the axes.

Theorem

> The components of a Gaussian vector are independent if and only if Σ is a diagonal matrix, *i.e.* if the components are not correlated.

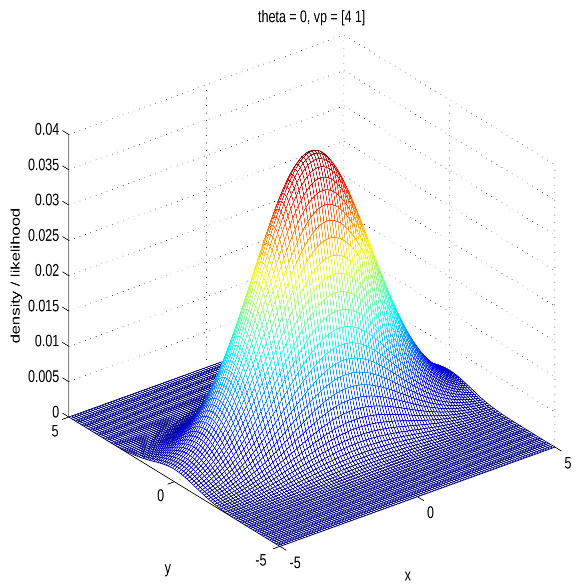
Illustration of 2D Gaussians

From the correlation point of view

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix}$$

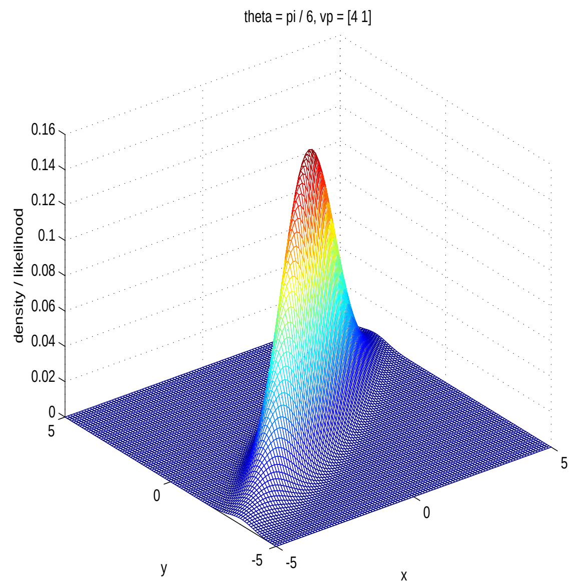
From the geometric point of view

$$V = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$



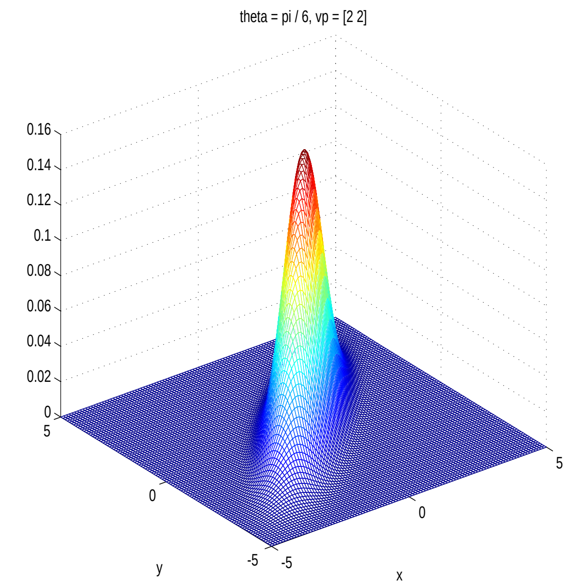
$\theta = 0$

$D =$ diag$(4 \quad 1)$



$\theta = \pi/6$

$D =$ diag$(4 \quad 1)$



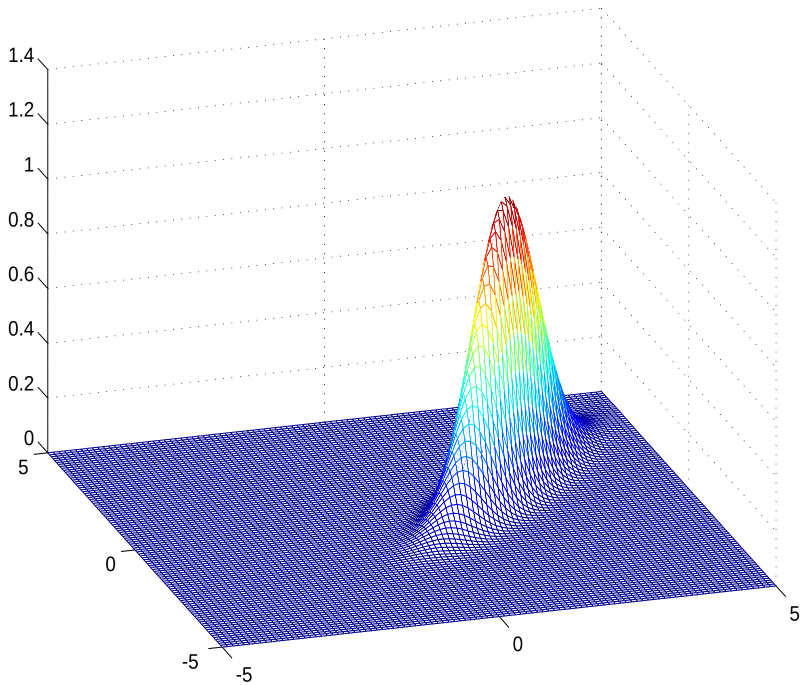
$\theta = \pi/6$

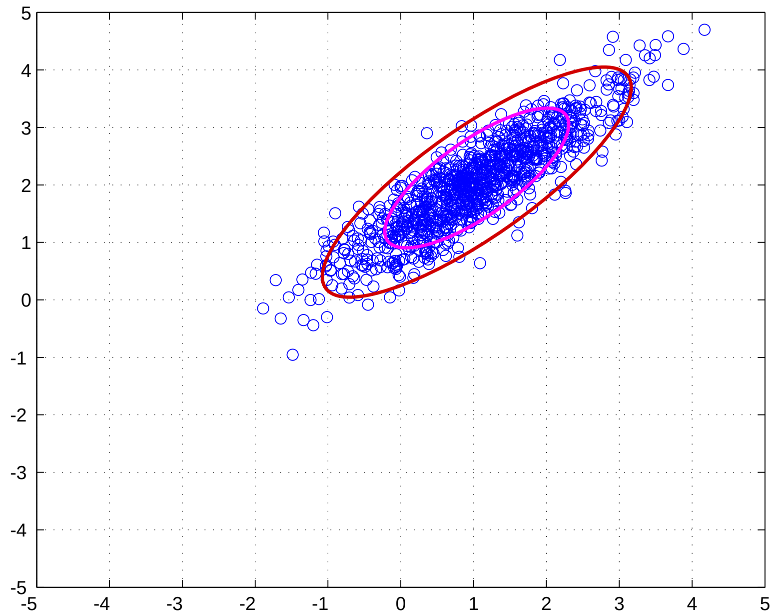
$D =$ diag$(2 \quad 2)$

Isodensity ellipsoids

Isodensity curves are (hyper)ellipsoids whose equation is given by

$$(x-\mu)'\Sigma^{-1}(x-\mu) = c$$

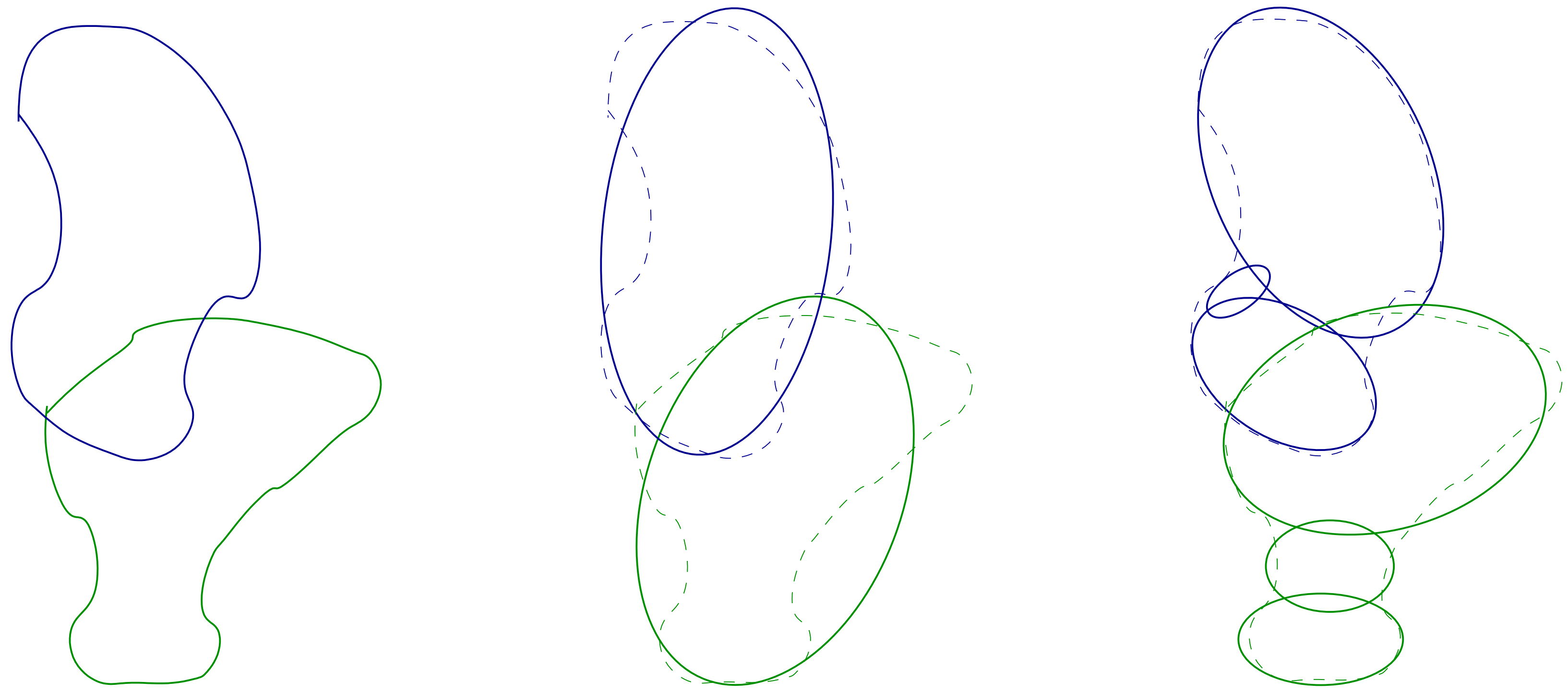




90% of the samples lie whithin the pink ellipse 99% of the samples lie whithin the red ellipse

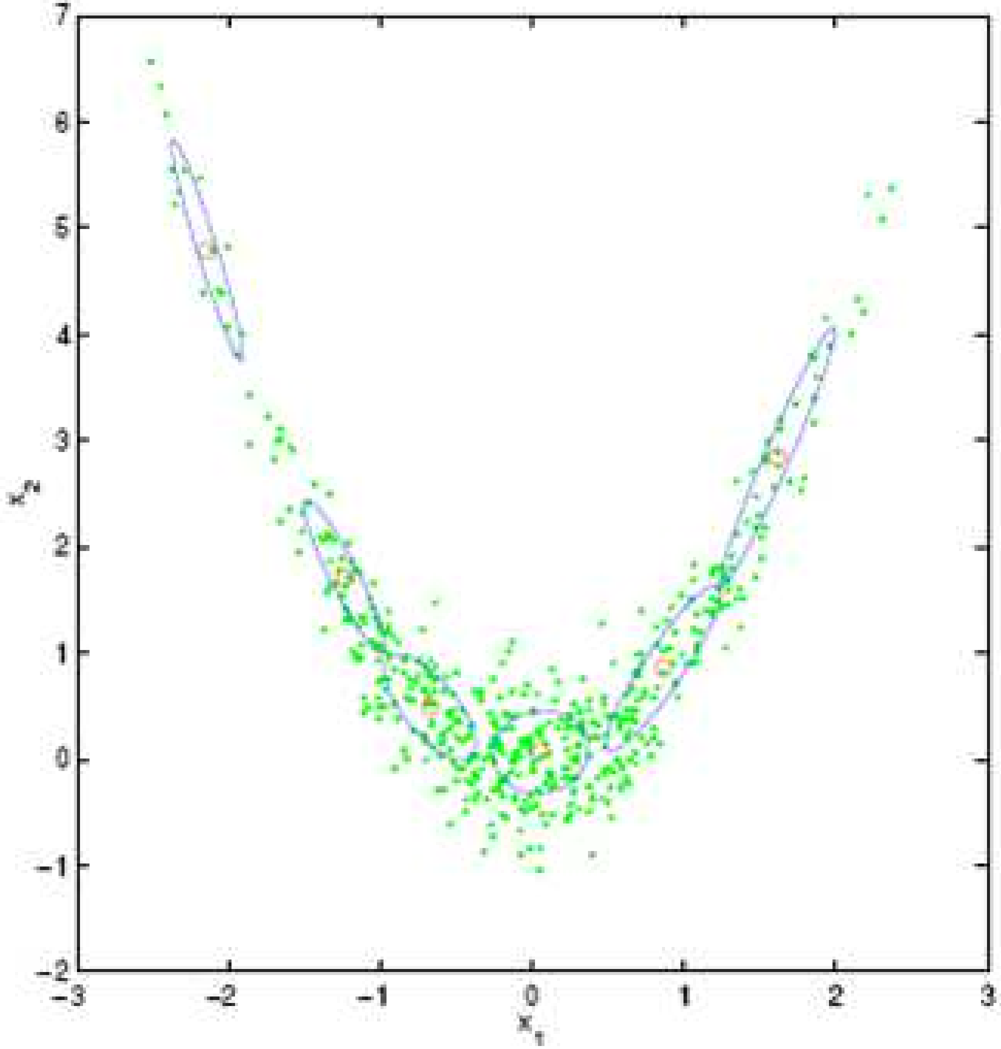
IRISA

Why mixture models?



true distribution Gaussian model Mixture model

Why mixture models?



[courtesy of J-F. Bonastre]

Mixture model – definition

A mixture model is a weighted sum of laws, the likelihood of a sample x being given by

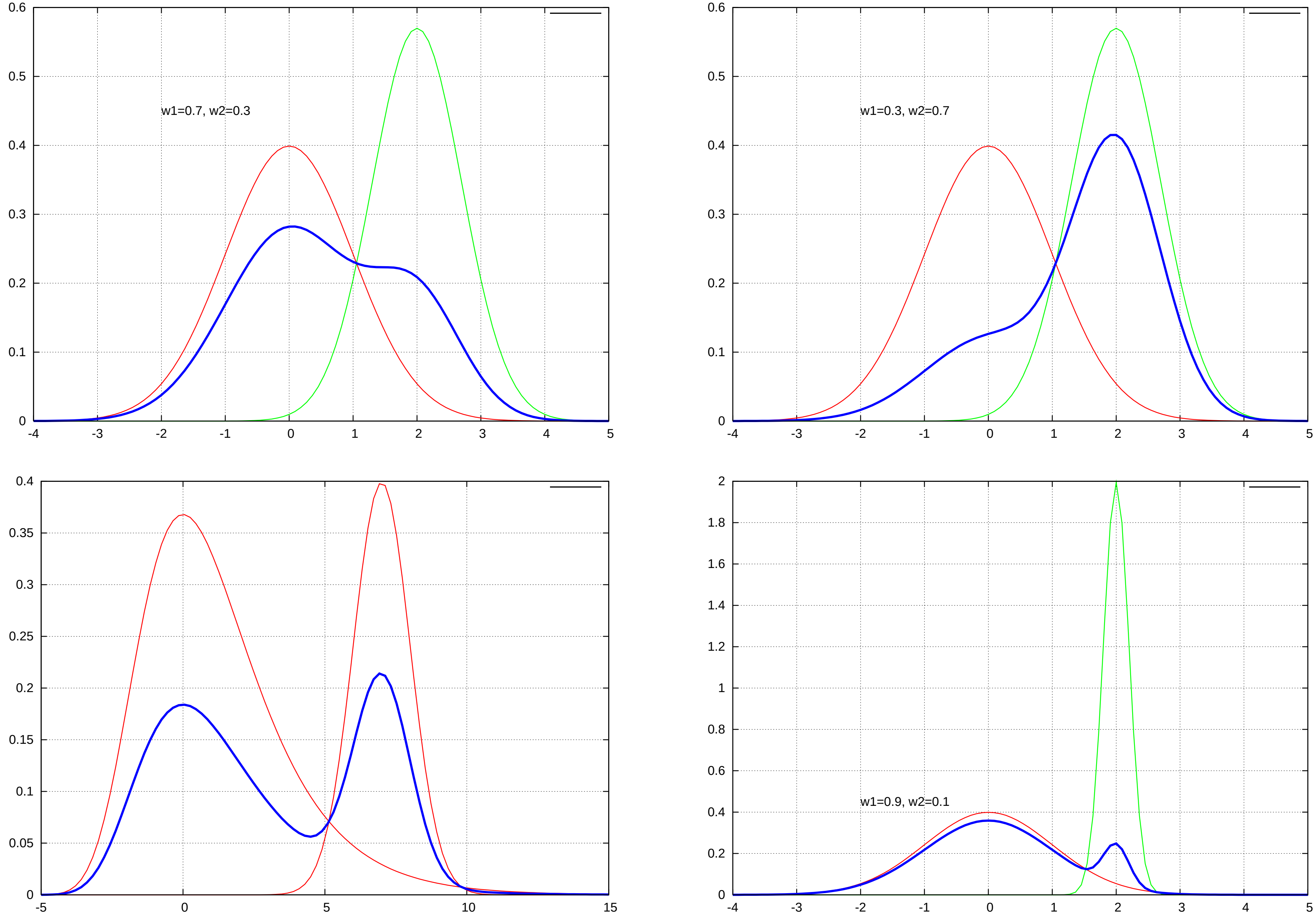
$$f(x) = \sum_{i=1}^{K} \pi_i f_i(x)$$

with the constraint that

$$\sum_{i=1}^{K} \pi_i = 1 \ .$$

- $f_i()$ can be any density and the $f_i()$'s need not be from the same family
- $f()$ is a density since $\int_{-\infty}^{\infty} f(x)dx = 1$
- the model can extend to discrete variables

Examples of mixture model densities



IRISA

Multivariate Gaussian mixture model

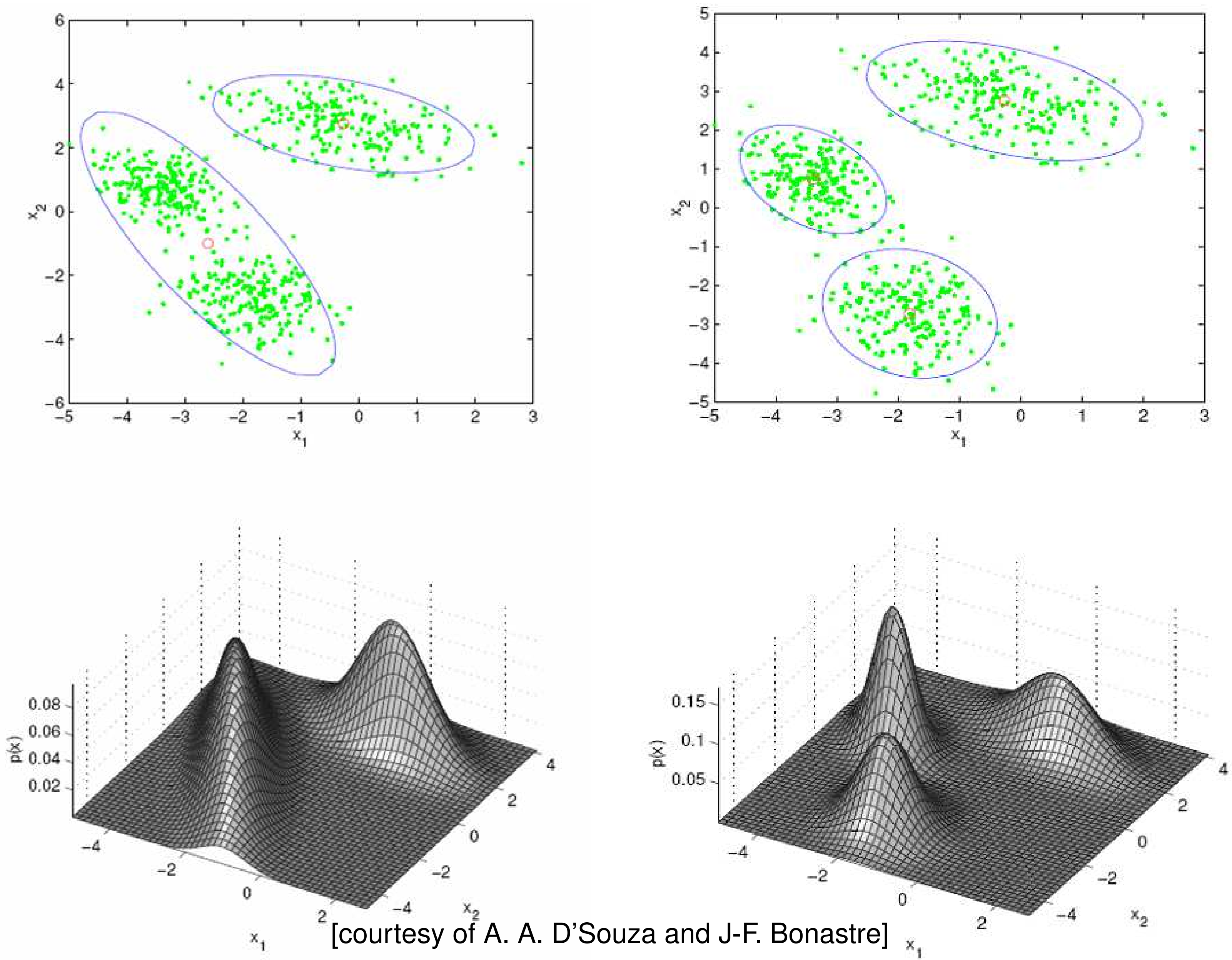
- Each component of the mixture is a multivariate Gaussian density

$$f_i(x) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\}$$

- Parameters of the model
  - weights of each component (K)
  - mean vectors (Kn)
  - covariance matrices $(Kn(n+1)/2)$

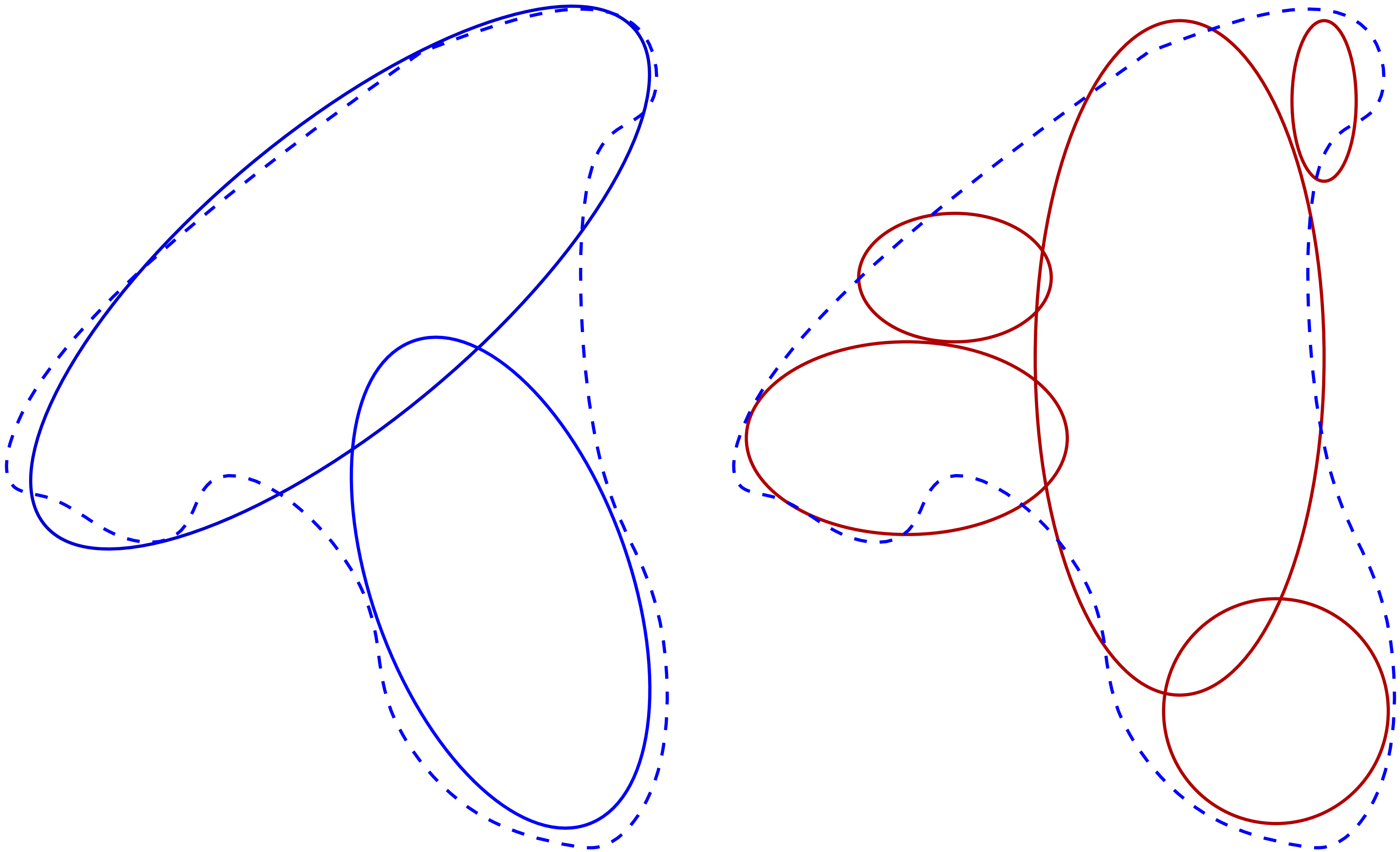
- In practice, we often assume diagonal covariance matrices
  - much less parameters $(Kn << Kn(n+1)/2)$
  - much less computation (saves matrix inversion)
  - can be compensated with more components

Gaussian mixture model



[courtesy of A. A. D'Souza and J-F. Bonastre]

Gaussian mixture model



Mixture models from the generative viewpoint

A sample is drawn according to the law of the component $f_i()$ of the mixture with probabiliy π_i.

Practically, sampling is a two step process

1. choose a component i of the mixture according to the discrete law defined by the weights w_j
2. draw a sample according to the law $f_i()$

Example: $w = [0.3, 0.7], f_1 = \mathcal{N}(0, 1), f_2 = \mathcal{N}(2, 1)$

IRISA

Mixture models from the generative viewpoint

Interpretation: [from a generative point of view,] the samples in a mixture model are drawn from either one of the component of the models with the proportion defined by the weights, *i.e.*

- for each component i, there is a set of samples distributed according to $f_i(x)$
- the proportion of such samples is given by π_i

$\Rightarrow$ hidden variable indicating the component to which a sample belong!

Mixture models and hidden variables

The law of x is the marginal over the hidden variable Z, *i.e.*

$$f(x) = \sum_{i=1}^{K} \underbrace{\pi_i}_{P[Z=i]} \underbrace{f_i(x)}_{p(x|Z=i)}$$

Sampling

$\Downarrow$

(Z, X)

Likelihood

$\Downarrow$

$(X) = \sum_Z (X, Z)$

IRISA

Mixture models and hidden variables

- Conditional density of x given z

$$p(x|z) = f_z(x) = \sum_{i=1}^{K} f_i(x)\mathbb{I}_{(z=i)}$$

- Joint density of (x, z)

$$p(x, z) = \pi_z f_z(x) = \left(\sum_{i=1}^{K} f_i(x)\mathbb{I}_{(z=i)}\right)\left(\sum_{i=1}^{K} \pi_i\mathbb{I}_{(z=i)}\right)$$

- Marginal density of x

$$p(x) = \sum_{z} p(x, z) = \sum_{z}\sum_{i} \pi_i f_i(x)\mathbb{I}_{(z=i)} = \sum_{i} \pi_i f_i(x)$$

Maximum likelihood parameter estimation

- Let $\mathbf{x} = \{x_1, \ldots, x_N\}$ be a set of training samples from which we want to estimate the parameters of a Gaussian mixture model with K components, *i.e.*
  - weights $\{w_1, \ldots, w_K\}$,
  - mean vectors $\{\mu_\mathbf{1}, \ldots, \mu_\mathbf{2}\}$,
  - variance vectors $\{\sigma_\mathbf{1}, \ldots, \sigma_\mathbf{2}\}$.
- Maximum likelihood criterion

$$\ln f(\mathbf{x}) = \sum_{i=1}^{N} \ln \left(\sum_{j=1}^{K} w_j f_j(x_i; \theta_j) \right)$$

$\Rightarrow$ direct maximization (nearly) impossible!

Direct maximum likelihood parameter estimation

- Directly solving the ML equations
  - ▷ they often do not exist!
  - ▷ complex equations when they do exist

- Gradient descent algorithms
  - ▷ non convex likelihood function $\Rightarrow$ non unicity of the solution
  - ▷ need for prior knowledge on the domain of θ

- The Expectation-Maximization algorithm
  - ▷ nice and elegant solution!

Maximum likelihood with complete data

- The set of training samples $\mathbf{x}$ is **incomplete!**
- Assume for each sample x_i, we know the component indicator function z_i
- The set $\{x_1, z_1, \ldots, x_N, z_N\}$ is known as *complete data*
- Maximum likelihood estimates can be obtained from the complete data, *e.g.*

$$\widehat{w}_i = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_{(z_j=i)} \qquad \text{and} \qquad \widehat{\mu}_i = \frac{\sum_{j=1}^{N} x_j \, \mathbb{I}_{(z_j=i)}}{\sum_{j=1}^{N} \mathbb{I}_{(z_j=i)}}$$

$\Rightarrow$ but the variables z_j are not known!

The Expectation-Maximization principle

The Expectation-Maximization (EM) principle compensates for missing (aka latent) data, replacing them by their expectations.

EM Iterative principle

1. estimate the missing variables given a current estimate of the parameters
2. estimate new parameters given the current estimate of the missing variables
3. repeat steps 1 and 2 until convergence

Note: this principle applies to many problems, not only for maximum likelihood parameter estimation!

The auxiliary function

The EM algorithm aims at maximizing an auxiliary function defined as

$$Q(\theta, \widehat{\theta}) = E[\ln f(\mathbf{z}, \mathbf{x}; \theta) | \mathbf{x}; \widehat{\theta}]$$

where $f(z, x; \theta)$ is the likelihood function of the complete data.

Estimation step

E compute the expected quantities in $Q(\theta, \widehat{\theta})$ (given $\widehat{\theta} = \theta_n$)

Maximization step

M maximize the auxiliary function w.r.t. the (true) parameters θ (given the expected quantities) to obtain a new estimate $\widehat{\theta} = \theta_{i+1}$, *i.e.*

$$\theta_{i+1} = \arg\max_{\theta} Q(\theta, \theta_i)$$

Close-up on the auxiliary function

Assume we have a n-sample $\mathbf{x} = \{x_1, \ldots, x_n\}$ and already know a wide guess $\widehat{\theta}$ of the parameters θ that we seek to estimate.

The log-likelihood of the complete data is given by

$$
\begin{aligned}
\ln f_\theta(\mathbf{z}, \mathbf{x}) &= \ln \left(\prod_{i=1}^{n} P_\theta[Z_i = z_i] p_\theta(x_i | z_i) \right) \\
&= \sum_{i=1}^{n} \ln \underbrace{P_\theta[Z_i = z_i]}_{= \pi_{z_i}} + \ln \underbrace{p_\theta(x_i | z_i)}_{\text{e.g., } \mathcal{N}(\mu_{z_i}, \sigma_{z_i})} \\
&= \sum_{j=1}^{K} \sum_{i=1}^{n} \ln(\pi_j) \mathbb{I}_{(j=z_i)} + \ln(p_{\theta_j}(x_i)) \mathbb{I}_{(j=z_i)}
\end{aligned}
$$

Hence the auxiliary function boils down to

$$
Q(\theta, \widehat{\theta}) = \sum_{j=1}^{K} \sum_{i=1}^{n} \ln(\pi_j) E_{\widehat{\theta}}[\mathbb{I}_{(j=z_i)} | \mathbf{x}] + \ln(p_{\theta_j}(x_i)) E_{\widehat{\theta}}[\mathbb{I}_{(j=z_i)} | \mathbf{x}]
$$

Maximizing the auxiliary function

Maximizing

$$Q(\theta, \widehat{\theta}) = \sum_{j=1}^{K} \sum_{i=1}^{n} \ln(\pi_j) E_{\widehat{\theta}}[\mathbb{I}_{(j=z_i)}|\mathbf{x}] + \ln(p_{\theta_j}(x_i)) E_{\widehat{\theta}}[\mathbb{I}_{(j=z_i)}|\mathbf{x}]$$

w.r.t. π_j under the constraints that weights sum to 1 yields

$$\widehat{\pi}_j \leftarrow \frac{\sum_i E_{\widehat{\theta}}[\mathbb{I}_{(z_i=j)}|\mathbf{x}]}{\sum_k \sum_i E_{\widehat{\theta}}[\mathbb{I}_{(z_i=k)}|\mathbf{x}]} .$$

Similarly, maximization w.r.t. the parameters θ_j of the log-likelihood of the j-th component, $\ln(p_{\theta_j}(x_i))$, will yield a function of the expectations $E_{\widehat{\theta}}[\mathbb{I}_{(z_i=j)}|\mathbf{x}]$, e.g., for a Gaussian density — details later

$$\widehat{\mu}_j \leftarrow \frac{\sum_i x_i E_{\widehat{\theta}}[\mathbb{I}_{(z_i=j)}|\mathbf{x}]}{\sum_i E_{\widehat{\theta}}[\mathbb{I}_{(z_i=j)}|\mathbf{x}]} .$$

A simple example

- $\mathbf{x} = \{x_1, \ldots, x_N\}$ from two classes with prior probabilities π and $1 - \pi$
- class indicator function $\mathbf{z} = \{z_1, \ldots, z_N\}$, $z_i \in \{0, 1\}$
- joint distribution of $(\mathbf{x}, \mathbf{z})$

$$\begin{aligned} \ln p_\theta(\mathbf{x}, \mathbf{z}) &= \sum_{i=1}^{N} \ln p_\theta(x_i | z_i) + \ln P_\theta[Z_i = z_i] \\ &= \sum_{i=1}^{N} \sum_{j \in \{0,1\}} (\ln p_{\theta_j}(x_i) + \ln P_\theta[Z_i = j]) \mathbb{I}_{(z_i = j)} \end{aligned}$$

- auxiliary function

$$Q(\theta, \widehat{\theta}) = \sum_{i=1}^{N} \sum_{j \in \{0,1\}} \left(\ln p_{\theta_j}(x_i) + \ln P_\theta[Z_i = j]\right) E_{\widehat{\theta}}[\mathbb{I}_{(z_i = j)} | \mathbf{x}]$$

A simple example: maximization equations

- auxiliary function (cont'd)

$$Q(\theta, \widehat{\theta}) = \sum_{j\in\{0,1\}} \sum_{i=1}^{N} \ln p_{\theta_j}(x_i) E_{\widehat{\theta}}[\mathbb{I}_{(z_i=j)}|\mathbf{x}]$$

$$+ \quad \ln(\pi) \sum_{i=1}^{N} E_{\widehat{\theta}}[\mathbb{I}_{(z_i=0)}|\mathbf{x}] + \ln(1-\pi) \sum_{i=1}^{N} E_{\widehat{\theta}}[\mathbb{I}_{(z_i=1)}|\mathbf{x}]$$

- Maximization w.r.t. π

$$\pi \leftarrow \frac{\sum_i E_{\widehat{\theta}}[\mathbb{I}_{(z_i=-1)}|\mathbf{x}]}{\sum_{j\in\{0,1\}} \sum_i E_{\widehat{\theta}}[\mathbb{I}_{(z_i=j)}|\mathbf{x}]}$$

- Maximization w.r.t. θ_j
  - only depends on the expecation $E_{\theta^n}[\mathbb{I}_{(z_i=j)}|\mathbf{x}]$

Simple example: expectation computation

The maximization (M-step) of $Q(\theta, \widehat{\theta})$ requires the computation of the expectations (E-step)

$$E_{\widehat{\theta}}[\mathbb{I}_{(z_i=j)}|\mathbf{x}] = P_{\widehat{\theta}}[Z_i = j|x_i] = \gamma_j(i) \ ,$$

for $i \in [1, N]$ and $j \in \{0, 1\}$, which is given by (dropping $\widehat{\theta}$ to facilitate reading)

$$P[Z_i = j|x_i] = \frac{p(x_i|Z_i = j)P[Z_i = j]}{\sum\limits_{k\in\{0,1\}} p(x_i|Z_i = k)P[Z_i = k]} \ .$$

In our two class example, for $j = 0$ we have

$$p(x_i|Z_i = j)P[Z_i = j] = \widehat{pi}\ p_{\widehat{\theta}_0}(x_i)$$

where $\widehat{\pi}$ and $\widehat{\theta}_0$ are the current estimates of the parameters.

EM and hidden class

In all generality

$$\gamma_j(i) = P_{\widehat{\theta}}[Z_i = j|x_i] = \frac{\widehat{\pi}_j p_{\widehat{\theta}_j}(x_i)}{\displaystyle\sum_k \widehat{\pi}_k p_{\widehat{\theta}_k}(x_i)} \ .$$

The latent variable $\gamma_j(i)$ indicates membership to a class as estimated from the current estimate $\widehat{\theta}$ of the parameters, we have the following interpretation:

- $P_{\widehat{\theta}}[Z_i = j|x_i]$ = degree ($\in [0, 1]$) of membership to class j of the i'th observation
- maximization relies on standard estimators based on the degree of membership (weighted standard estimators) [see Gaussian mixture example]

EM from an algorithmic viewpoint

1. choose some initial (good) parameters θ_0
2. $n \leftarrow 0$
3. while not happy (with convergence)
   (a) for $i = 1 \rightarrow N$ and $j = 1 \rightarrow K$
   compute the component posterior $\gamma_j^{(n)}(i) = P[Z_i = j | x_i; \theta_n]$
   (b) foreach paramter α in θ
   compute new parameter value α_{n+1} from the quantities $\gamma_j^{(n)}(i)$
   (by maximizing $Q(\theta, \theta_n)$)
   (c) $n \leftarrow n + 1$

Properties of the EM algorithm

Property 1

The serie of estimators $\{\theta_n\}$ is such that the likelihood of the data increases with each iteration of the algorithm.

It can be shown that

$$Q(\theta, \theta_{i+1}) - Q(\theta, \theta_i) =$$

$$\ln p(\mathbf{x}; \theta_{i+1}) - \ln p(\mathbf{x}; \theta_i) + \underbrace{E\left[\ln \frac{p(\mathbf{z}|\mathbf{x}; \theta_{i+1})}{p(\mathbf{z}|\mathbf{x}; \theta_i)} |\mathbf{x}; \theta_i\right]}_{<0}$$

which implies that

$$Q(\theta, \theta_{i+1}) \geq Q(\theta, \theta_i) \implies p(\mathbf{x}; \theta_{i+1}) \geq p(\mathbf{x}; \theta_i)$$

Properties of the EM algorithm

Property 2

The EM algorithm enables the computation of the gradient of the log-likelihood function at the points θ_i.

It can be verified that under some non very restrictive assumptions

$$\left.\frac{\partial Q(\theta,\theta_i)}{\partial\theta}\right|_{\theta=\theta_i} = \left.\frac{\partial \ln f(\mathbf{x};\theta)}{\partial\theta}\right|_{\theta=\theta_i} + \underbrace{\left.\frac{\partial E[\ln p(\mathbf{z}|\mathbf{x};\theta)|\mathbf{x};\theta_i]}{\partial\theta}\right|_{\theta=\theta_i}}_{=0}$$

Convergence of the EM algorithm

Property 1

The serie of estimators $\{\theta^{(n)}\}$ *is such that the likelihood of the data increases with each iteration of the algorithm.*

Property 2

The EM algorithm enables the computation of the gradient of the log-likelihood function at the points θ_i*.*

$$\Downarrow$$

The EM estimate converges toward stationary points of the log-likelihood function $\ln p(\mathbf{x};\theta)$.

Convergence of the EM algorithm in practice

- The convergence is only guaranteed toward a local maximum of the likelihood function $\ln p(\mathbf{x}; \theta)$.
  - need for a good initial guess θ_0
  - need to avoid degenerate solutions!

- In practice, convergence is controled by two factors
  - increase of the log-likelihood of the data
  - fixed number of iterations

- Constraints on the parameter space are often used to avoid bad or degenerated solutions, *e.g.*
  - minimum variance floor
  - initialization based on (segmental) k-means algorithm

EM for Gaussian mixtures

- Joint likelihood of $(\mathbf{x}, \mathbf{z})$

$$\ln f(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{N} \sum_{j=1}^{K} \ln(w_j f(x_i; \mu_j, \sigma_j)) \mathbb{I}_{(z_i = j)}$$

- Auxiliary function

$$\begin{aligned} Q(\theta, \theta_n) \propto & \sum_{j=1}^{K} \sum_{i=1}^{N} \ln(w_j) E[\mathbb{I}_{(z_i=j)} | \mathbf{x}; \theta_n] \\ - & \frac{1}{2} \sum_{j=1}^{K} \sum_{i=1}^{N} \left(\sum_{k=1}^{d} \ln(\sigma_{jk}^2) + \frac{(x_{ik} - \mu_{jk})^2}{\sigma_{jk}^2} \right) E[\mathbb{I}_{(z_i=j)} | \mathbf{x}; \theta_n] \end{aligned}$$

EM for Gaussian mixtures (cont'd)

- Compute the expectations at iteration n

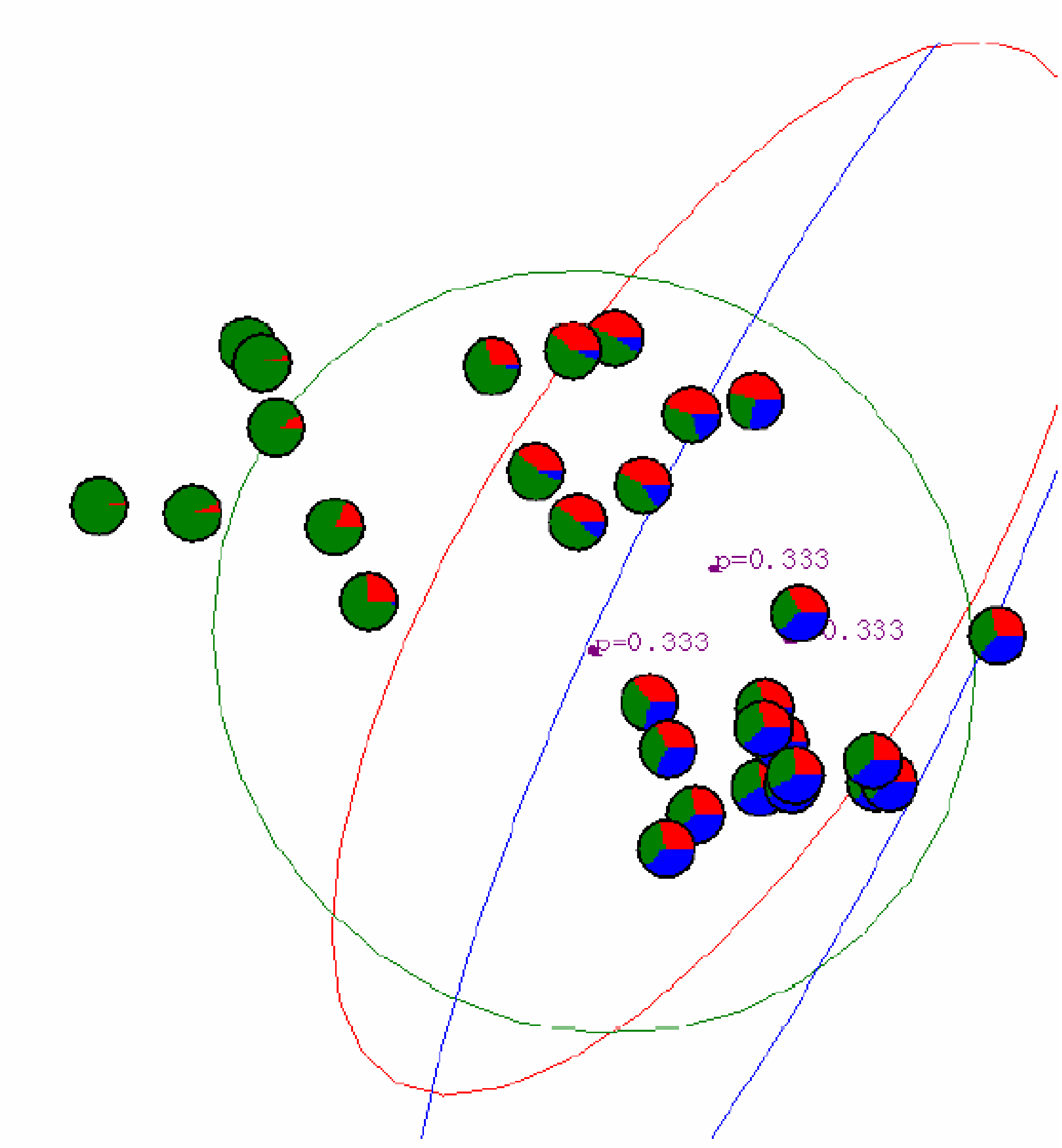
$$\gamma_j^{(n)}(i) = \frac{w_j^{(n)} \; f(x_i; \mu_j^{(n)}, \sigma_j^{(n)})}{\sum_k w_k^{(n)} \; f(x_i; \mu_k^{(n)}, \sigma_k^{(n)})}$$

  where the parameters correspond to the current estimate $\theta^{(n)}$.

- Maximization

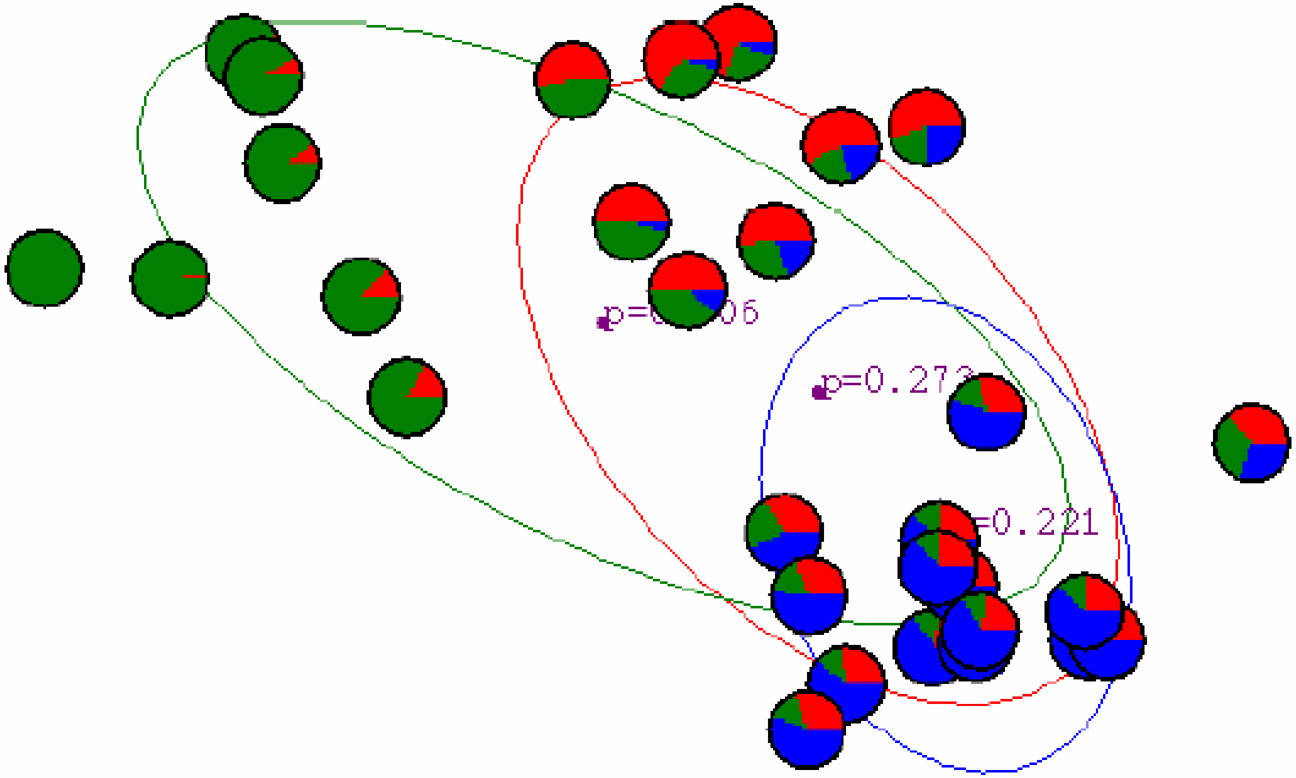
$$w_j^{(n+1)} = \frac{\sum_{i=1}^{N} \gamma_j^{(n)}(i)}{\sum_{k=1}^{K} \sum_{i=1}^{N} \gamma_k^{(n)}(i)} \qquad \mu_{jk}^{(n+1)} = \frac{\sum_{i=1}^{N} \gamma_j^{(n)}(i) \; x_{ik}}{\sum_{i=1}^{N} \gamma_j^{(n)}(i)}$$

The EM at work: initialization



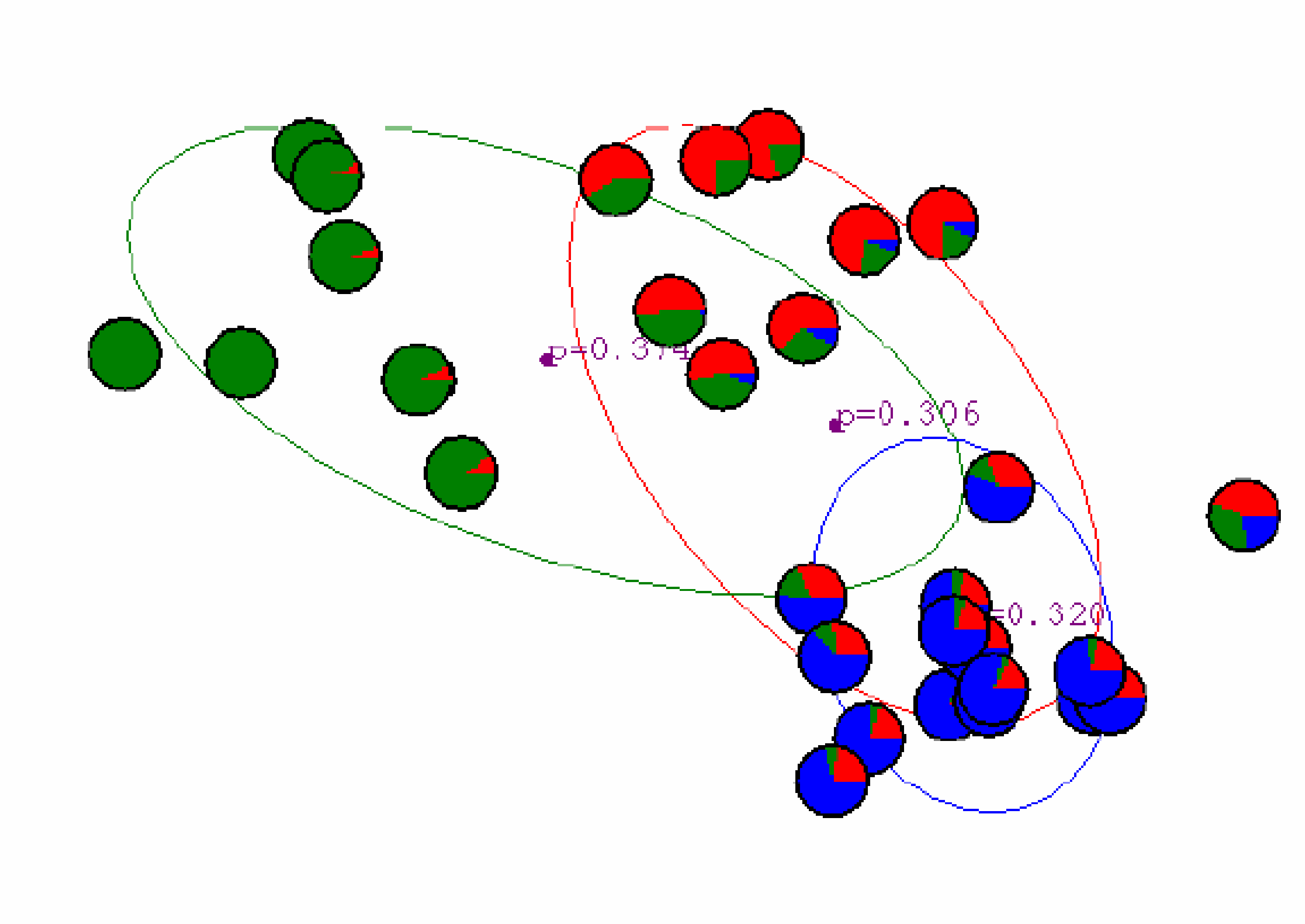
[courtesy of A. W. Moore and J-F. Bonastre]

The EM at work: iteration 1



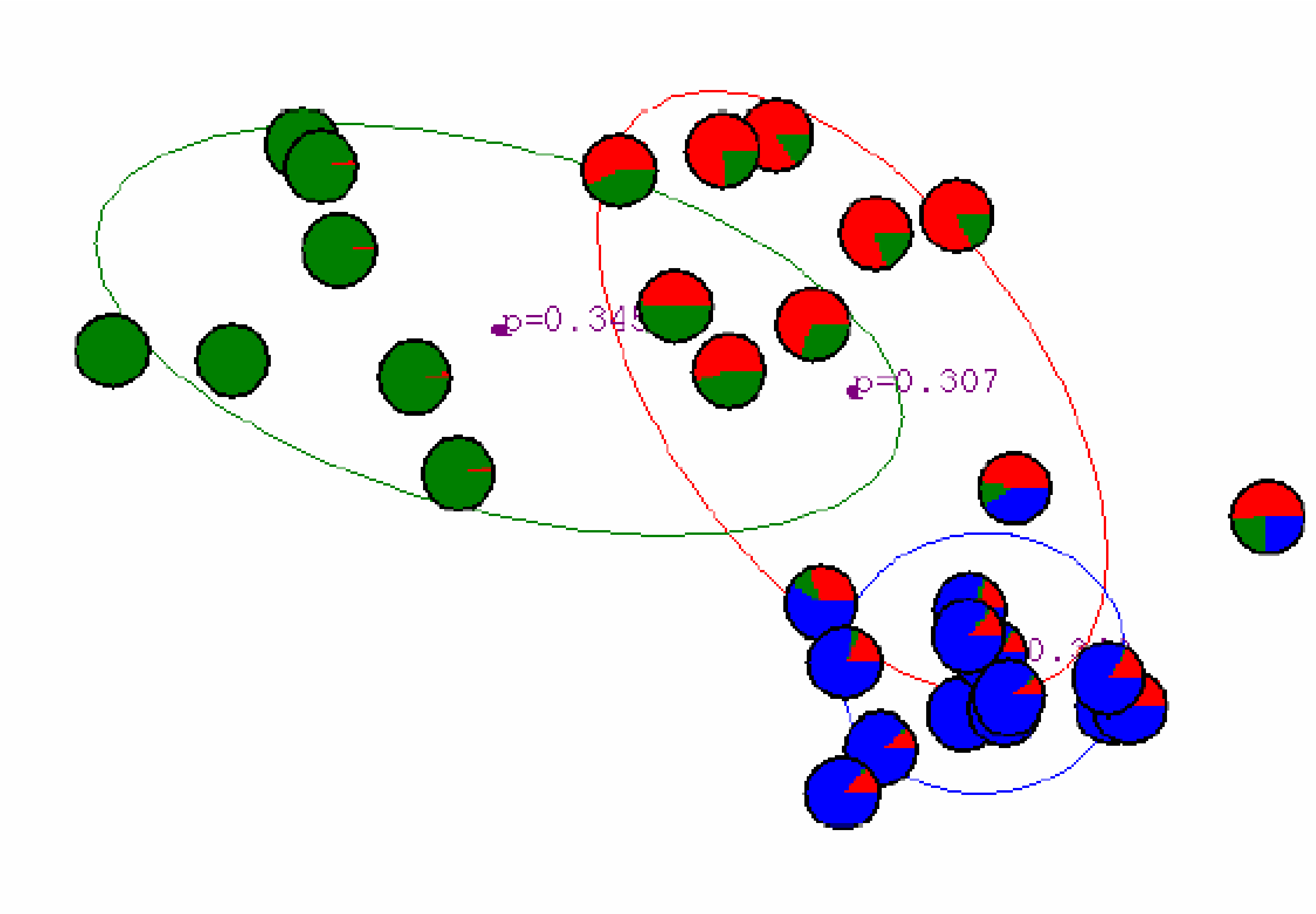
[courtesy of A. W. Moore and J-F. Bonastre]

The EM at work: iteration 2



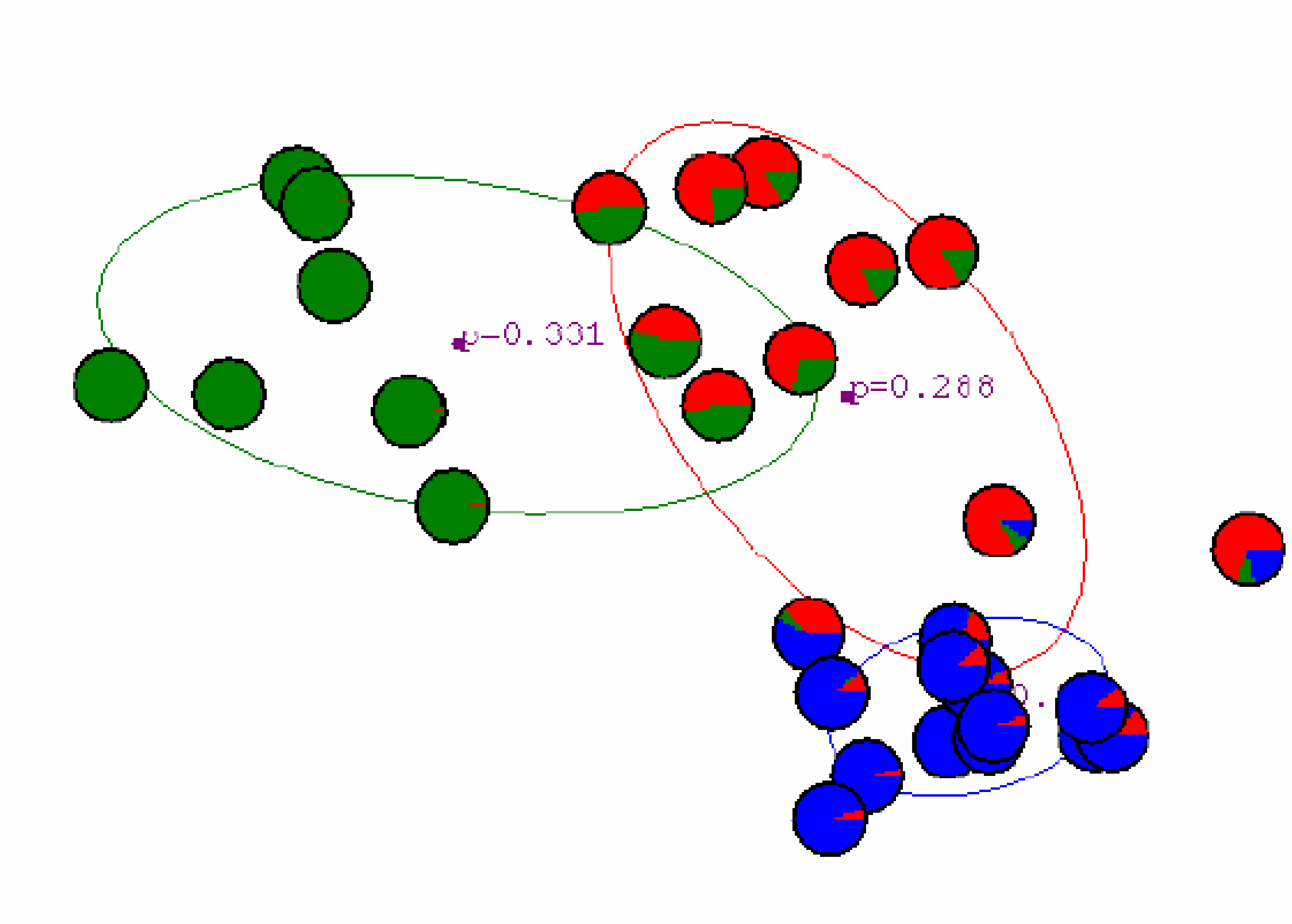
[courtesy of A. W. Moore and J-F. Bonastre]

The EM at work: iteration 3



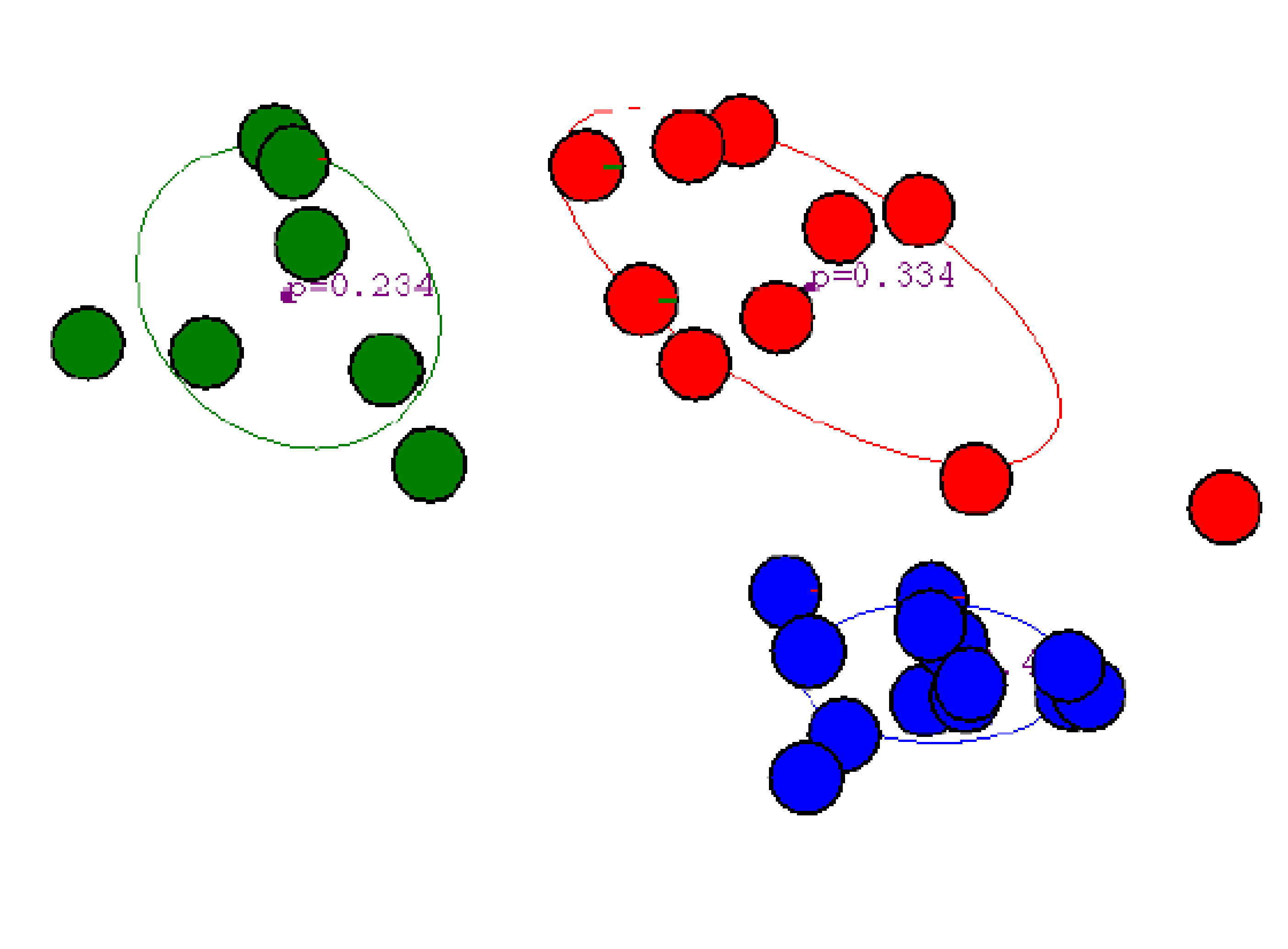
[courtesy of A. W. Moore and J-F. Bonastre]

The EM at work: iteration 4



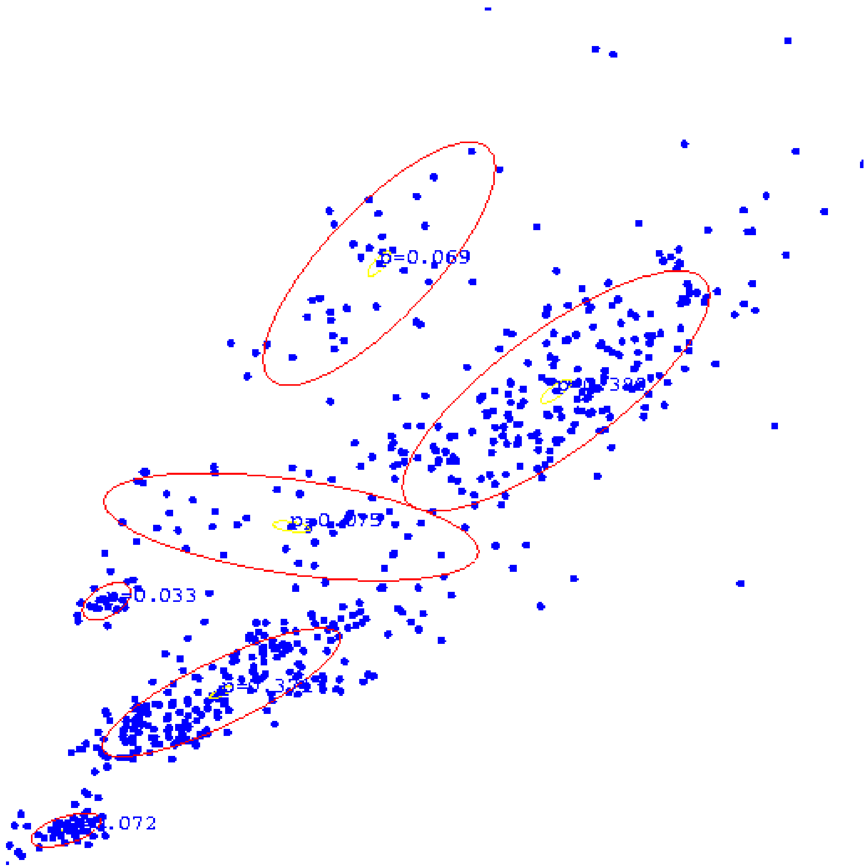
[courtesy of A. W. Moore and J-F. Bonastre]

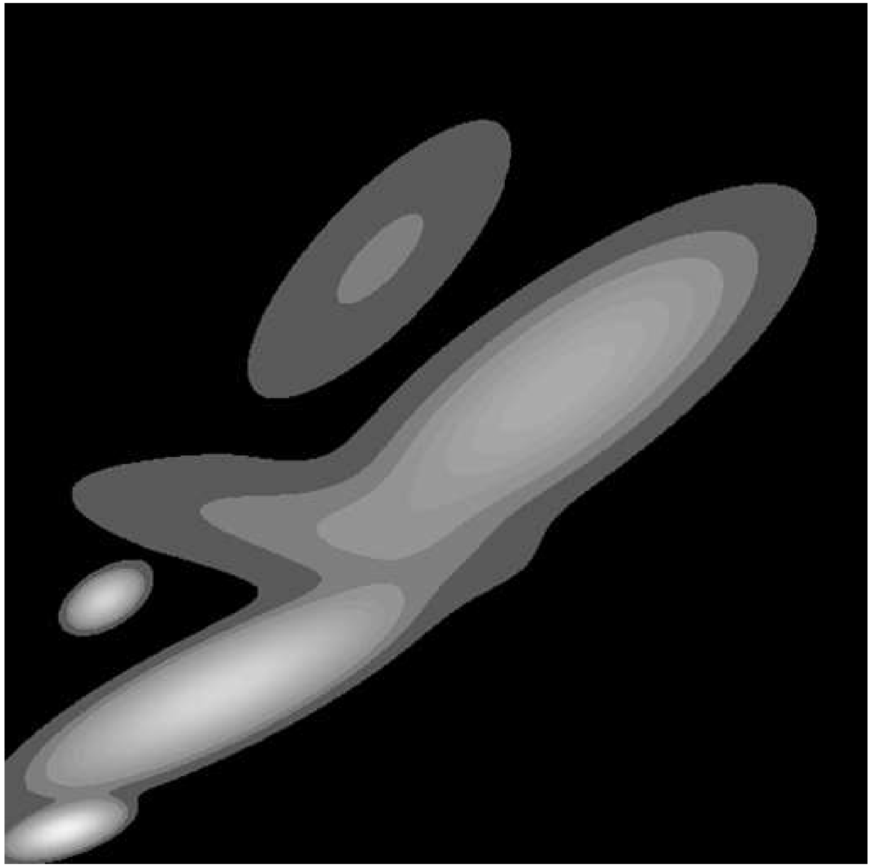
The EM at work: iteration 20



[courtesy of A. W. Moore and J-F. Bonastre]

The EM at work: another example





[courtesy J-F. Bonastre]

EM and k-means clustering

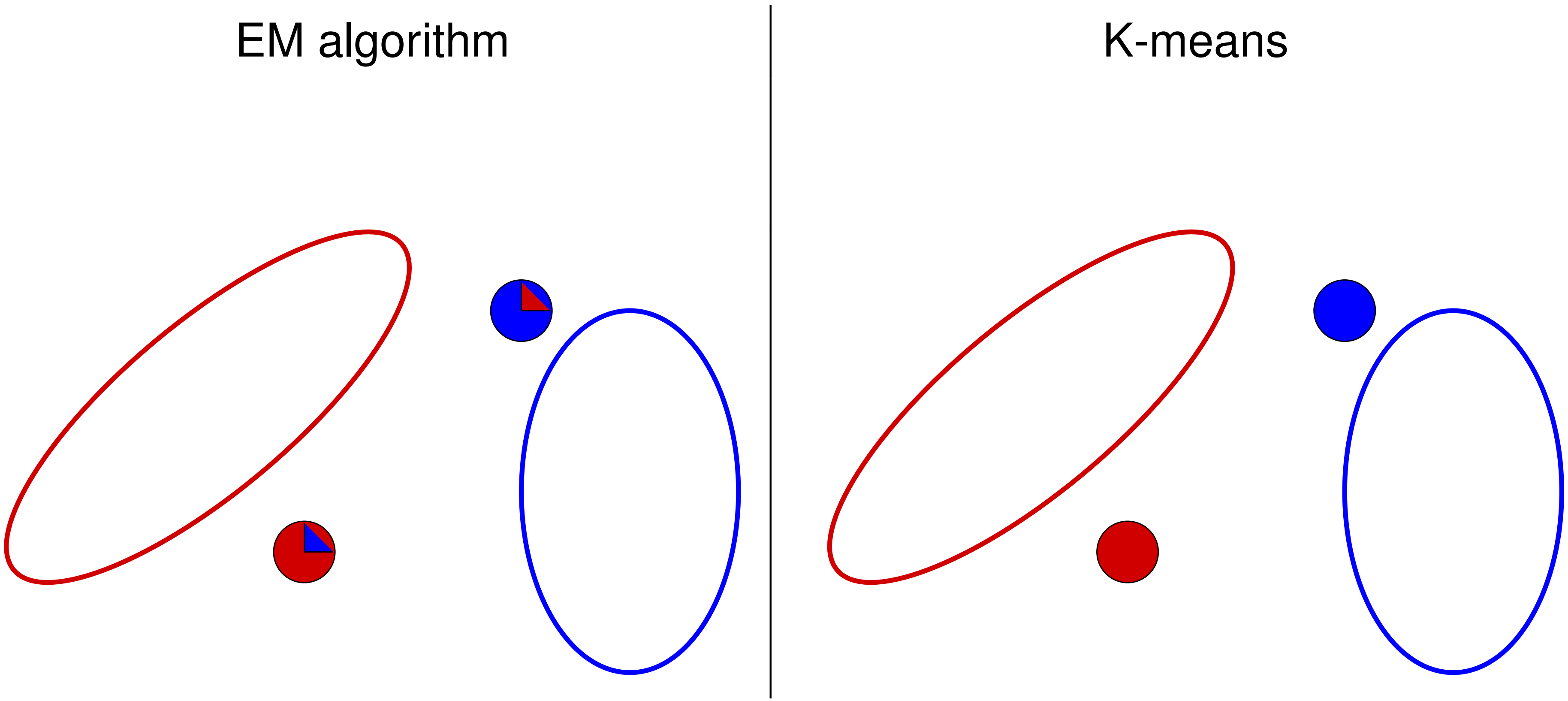
- Maximum likelihood estimates with known class membership

$$\widehat{w}_i = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_{(z_j=i)} \qquad \widehat{\mu}_i = \frac{\sum_{j=1}^{N} x_j \; \mathbb{I}_{(z_j=i)}}{\sum_{j=1}^{N} \mathbb{I}_{(z_j=i)}}$$

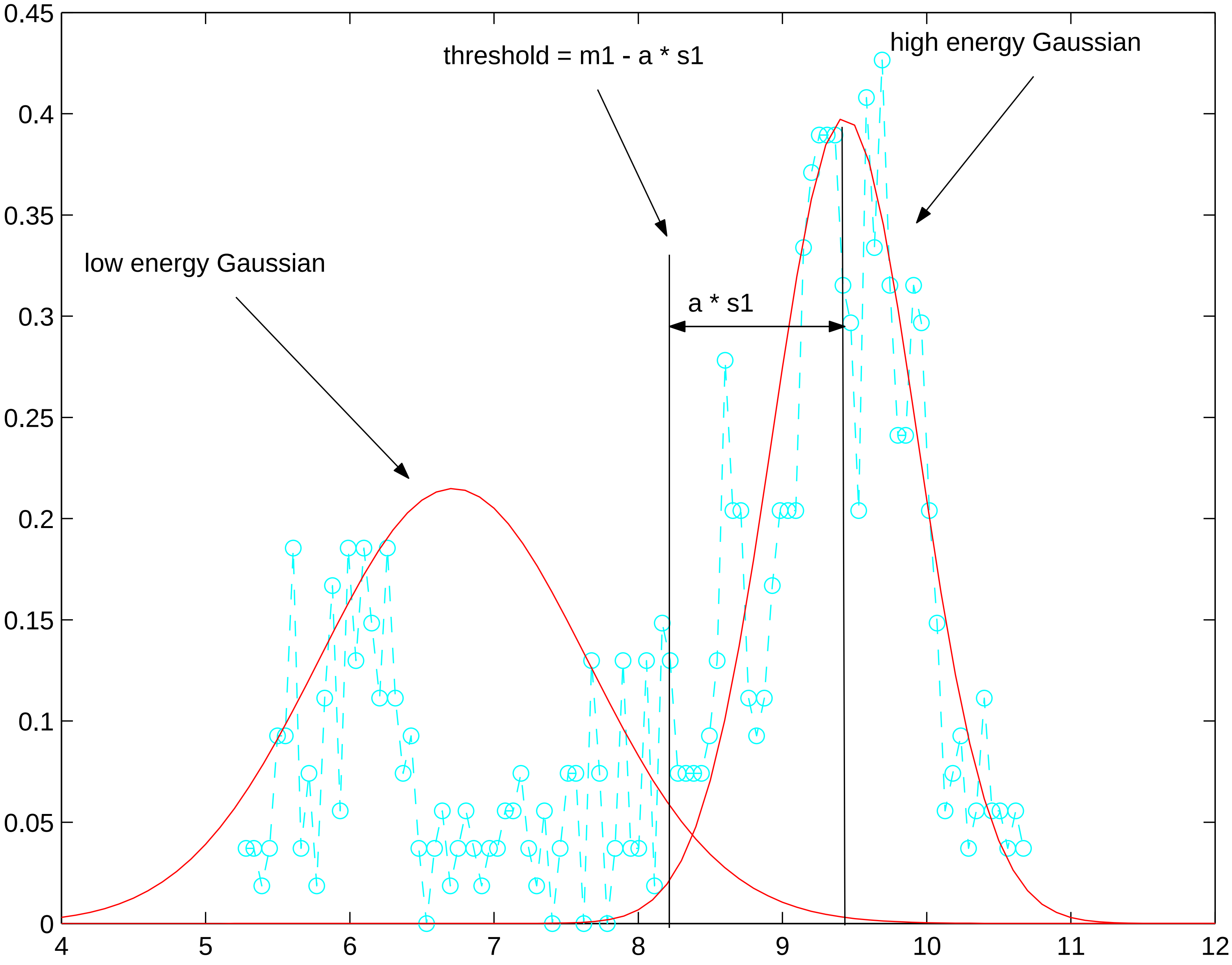
- EM estimates with unknown class membership

$$\widehat{w}_i = \frac{1}{N} \sum_{j=1}^{N} E[\mathbb{I}_{(z_j=i)} | \mathbf{x}, \theta_n] \qquad \widehat{\mu}_i = \frac{\sum_{j=1}^{N} x_j \; E[\mathbb{I}_{(z_j=i)} | \mathbf{x}, \theta_n]}{\sum_{j=1}^{N} E[\mathbb{I}_{(z_j=i)} | \mathbf{x}, \theta_n]}$$

EM and k-means clustering



A practical use of the Gaussian law



IRISA

EM, sufficient statistics and the exponential family

- Joint density is from the exponential family

$$f(\mathbf{x}, \mathbf{z}; \theta) = \exp\left(\alpha(\theta)' a(\mathbf{x}, \mathbf{z}) + b(\mathbf{x}, \mathbf{z}) - \beta(\theta)\right)$$

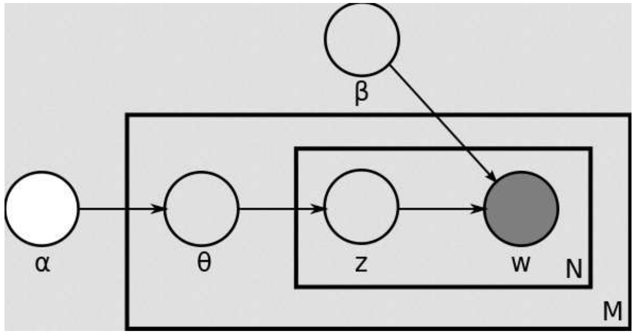
- E-step $\Rightarrow$ estimate the sufficient statistic $a(\mathbf{x}, \mathbf{z})$ by computing its expectation under the posterior law given a current estimation of the parameters

- Examples:

$$\begin{aligned}
\sum_j \mathbb{I}_{(z_j=i)} \quad &\longrightarrow \quad \sum_j E[\mathbb{I}_{(z_j=i)} | \mathbf{x}; \theta_n] \\
\sum_j x_j \; \mathbb{I}_{(z_j=i)} \quad &\longrightarrow \quad \sum_j x_j \; E[\mathbb{I}_{(z_j=i)} | \mathbf{x}; \theta_n] \\
\sum_j (x_j - \widehat{\mu}_j)^2 \; \mathbb{I}_{(z_j=i)} \quad &\longrightarrow \quad \sum_j (x_j - \widehat{\mu}_j)^2 \; E[\mathbb{I}_{(z_j=i)} | \mathbf{x}; \theta_n]
\end{aligned}$$

The LDA topic mixture model

I eat fish and vegetables.
Fishes are pets.
My kitten eats fish.



Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

Variants

The EM principle enables many variants when the E-step and/or the M-step are intractable

- Monte-Carlo EM: replace the exact computation of the expected quantities by some Monte-Carlo approximations obtained using the current parameters
- Generalized EM: simply increase the auxiliary function rather than maximizing it, *e.g.* using a gradient algorithm
- Variational EM: replace the auxiliary function Q by a more simple variational approximation based on factorial distribution $Q \simeq \prod_i Q_i$.
- ...

Choosing the number of components

- Experimentations...
- Information criterion

$$\mathcal{I}(\mathbf{x}, \theta) = \ln p(\mathbf{x}; \theta) - g(\#\text{parameters}, \#\text{data})$$

  - Akaike
  - Bayesian Information criterion (BIC)
  - ...

Bibliography

- A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM algorithm. Journal of the Royal Statistical Society. Series B, Vol. 39, No. 1, pp. 1–38, 1977.
- T. K. Moon. The Expectation-Maximization Algorithm. IEEE Signal Processing Magazine, pp. 46–60, November, 1996.
- G. J. McLachlan, T. Krishnan. The EM algorithm and extensions. Wiley Series in Probability and Statistics, 1997.