CRV 2010

A short tutorial on
Gaussian Mixture Models

By: Mohand Saïd Allili
Université du Québec en Outaouais

Plan

-Introduction

-What is a Gaussian mixture model?

-The Expectation-Maximization algorithm

-Some issues

-Applications of GMM in computer vision

Introduction

A few basic equalities that are often used:

- Conditional probabilities:

$$p(A \cap B) = p(A \mid B)\, p(B)$$

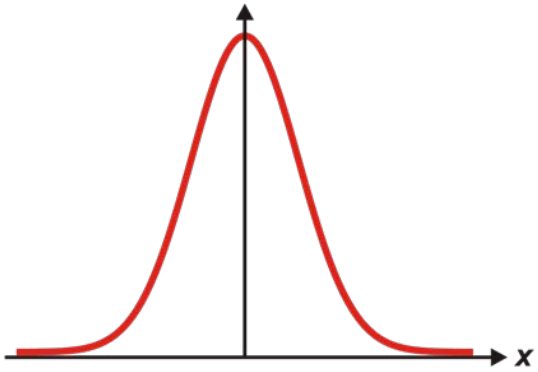
- Bayes rule:

$$p(B \mid A) = \frac{p(A \mid B)\, p(B)}{P(A)}$$

- if $\bigcup_{i=1}^{K} B_i = \Omega$ and $\forall i \neq j : B_i \cap B_j = \phi$, then :

$$p(A) = \sum_{i=1}^{K} p(A \cap B_i)$$

Introduction



Carl Friedrich Gauss invented the normal distribution in 1809 as a way to rationalize the method of least squares.

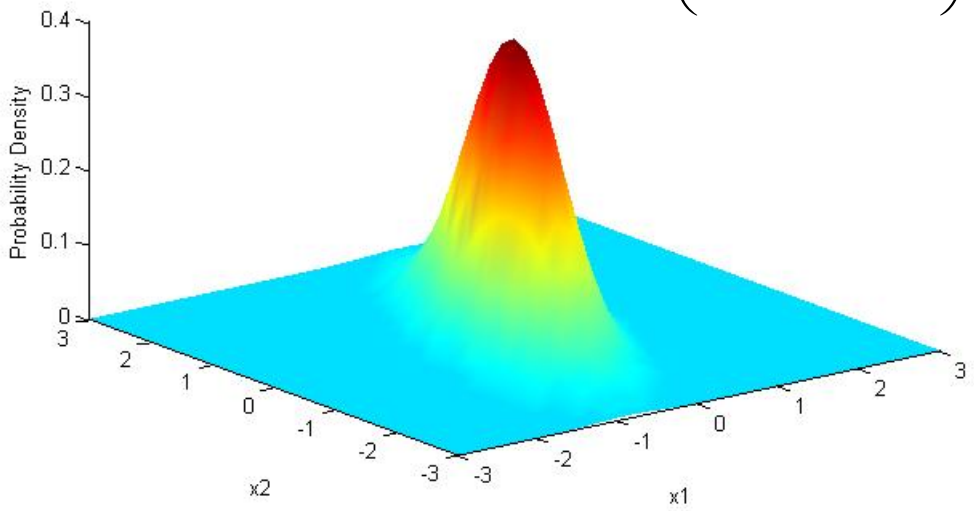
What is a Gaussian?

For ***d* dimensions**, the Gaussian distribution of a vector $x = (x^1, x^2, \ldots, x^d)^T$ is defined by:

$$N(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

where μ is the mean and Σ is the covariance matrix of the Gaussian.

Example: $\mu = (0,0)^T$ $\Sigma = \begin{pmatrix} 0.25 & 0.30 \\ 0.30 & 1.00 \end{pmatrix}$



What is a Gaussian mixture model?

The probability given in a mixture of *K* Gaussians is:

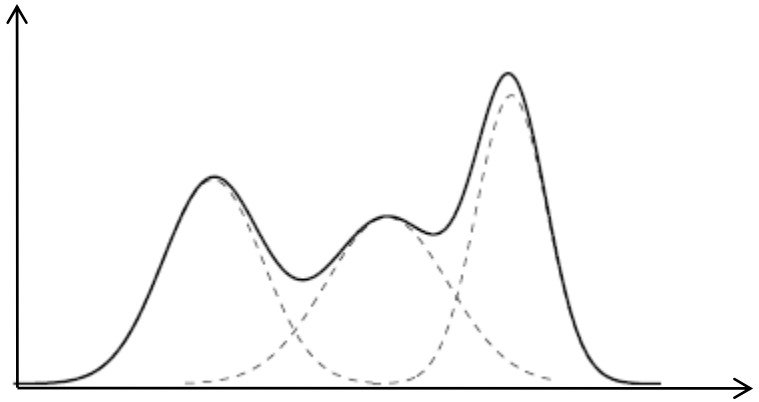
$$p(x) = \sum_{j=1}^{K} w_j \cdot N(x \mid \mu_j, \Sigma_j)$$

where w_j is the prior probability (weight) of the jth Gaussian.

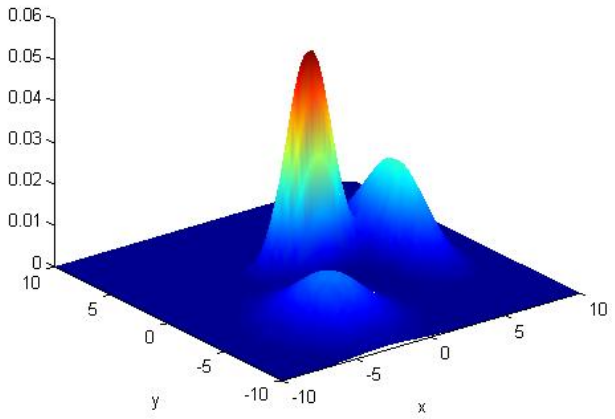
$$\sum_{j=1}^{K} w_j = 1 \qquad \text{and} \qquad 0 \le w_j \le 1$$

Examples:

d=1:



d=2:



What is a Gaussian mixture model?

- **Problem**:

  Given a set of data $X = \{x_1, x_2, ..., x_N\}$ drawn from an unknown distribution (probably a GMM), estimate the parameters θ of the GMM model that fits the data.

- **Solution**:

  Maximize the likelihood $p(X \mid \theta)$ of the data with regard to the model parameters?

$$\theta^* = \arg\max_{\theta} p(X \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{N} p(x_i \mid \theta)$$

The Expectation-Maximization algorithm

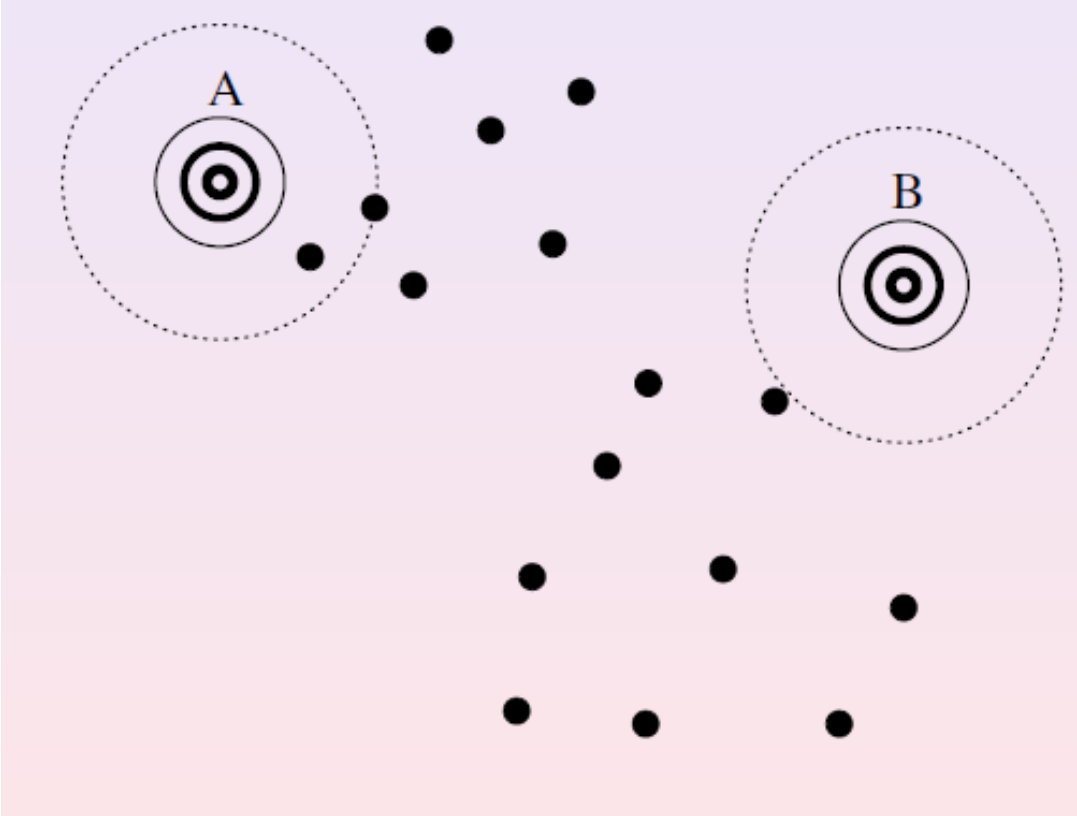
One of the most popular approaches to maximize the likelihood is to use the Expectation-Maximization (EM) algorithm.

- **Basic ideas of the EM algorithm:**

- Introduce a hidden variable such that its knowledge would simplify the maximization of the likelihood.
- At each iteration:
    - E-Step: Estimate the distribution of the hidden variable given the data and the current value of the parameters.
    - M-Step: Maximize the joint distribution of the data and the hidden variable.

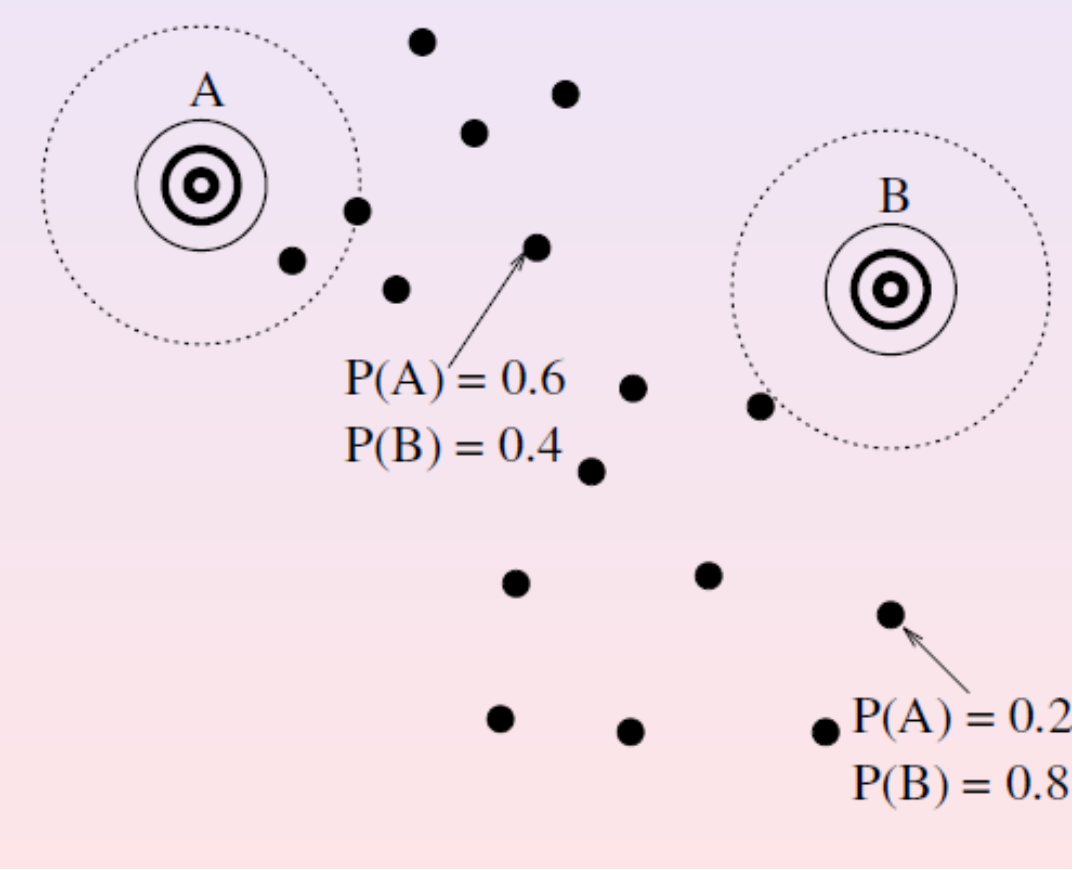
The EM for the GGM (graphical view 1)

Hidden variable: for each point, which Gaussian generated it?



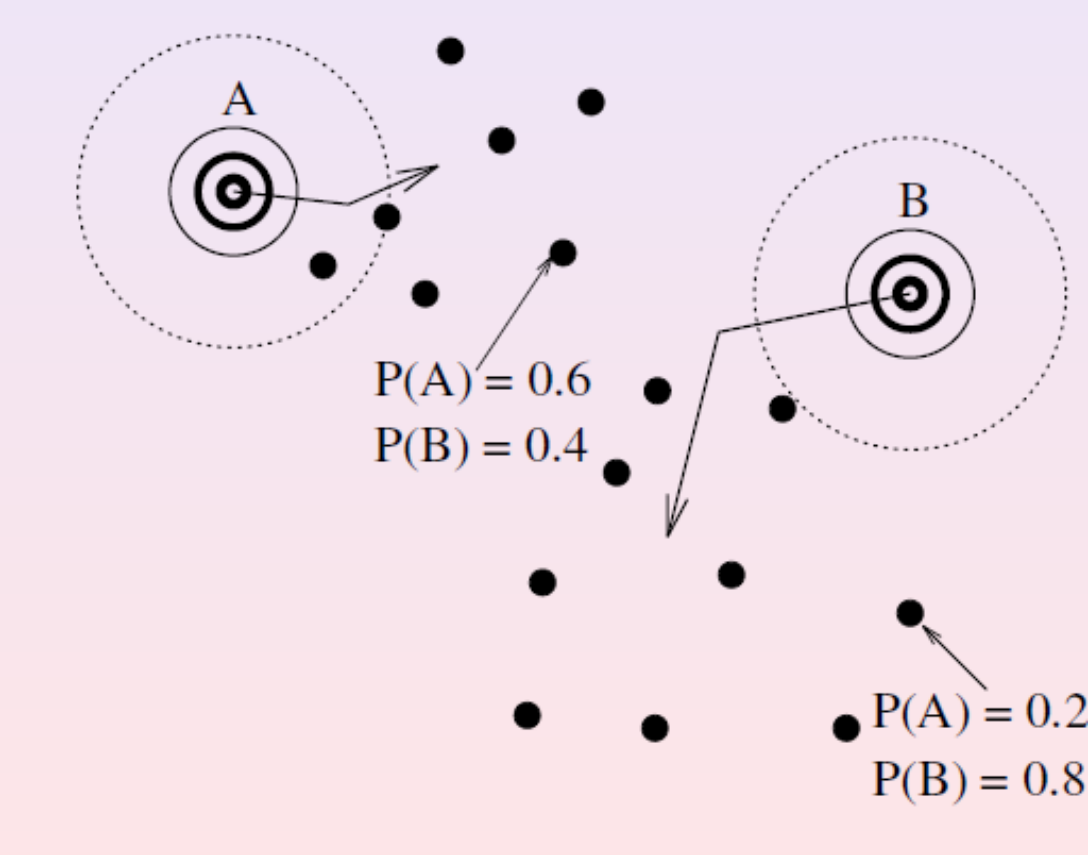
The EM for the GGM (graphical view 2)

E-Step: for each point, estimate the probability that each Gaussian generated it.



The EM for the GGM (graphical view 3)

M-Step: modify the parameters according to the hidden variable to maximize the likelihood of the data (and the hidden variable).



General formulation of the EM

(In what follows the hidden variable name is Z)

Consider the following "auxillary" function:

$$Q(\theta, \theta^t) = E_Z\left[\log(p(X, Z \mid \theta^t))\right]$$

It can been shown that maximizing $Q(\theta, \theta^t)$:

$$\theta^{t+1} = \arg\max_{\theta}\left[Q(\theta, \theta^t)\right]$$

always maximizes the liekelihood of the data.

Proof of EM convergence

$$Q(\theta,\theta^t) = E_Z\left[\log(p(X,Z\mid\theta^t))\right]$$

$$= \sum_z p(z\mid X,\theta^t)\log(p(X,Z\mid\theta))$$

$$= \sum_z p(z\mid X,\theta^t)\log[p(Z\mid X,\theta)\cdot p(X\mid\theta)]$$

$$= \left[\sum_z p(z\mid X,\theta^t)\log[p(Z\mid X,\theta)]\right] + \log[p(X\mid\theta)] \qquad (1)$$

Proof of EM convergence

If $\theta = \theta^t$, we have:

$$Q(\theta^t, \theta^t) = \left[\sum_z p(z \mid X, \theta^t) \log[p(Z \mid X, \theta^t)] \right] + \log[p(X \mid \theta^t)] \qquad (2)$$

From (1) and (2), we have:

$$\log[p(X \mid \theta)] - \log[p(X \mid \theta^t)] =$$

$$Q(\theta, \theta^t) - Q(\theta^t, \theta^t) + \underbrace{\left[\sum_z p(z \mid X, \theta^t) \log\left[\frac{p(Z \mid X, \theta)}{p(Z \mid X, \theta^t)} \right] \right]}_{\geq 0}$$

Conclusion:

If $Q(\theta, \theta^t)$ increases then the likelihood $p(X \mid \theta)$ increases.

The maximum of $Q(\theta, \theta^t)$ is the same as the maximum likelihood.

EM for the GMM

- Note that if Z is observed, then the parameters can be estimated Gaussian by Gaussian. Unfortunately, Z is not always known.

- Let us write the mixture of Gaussians for a data point $x_i, i = 1,\ldots,N$:

$$p(x_i) = \sum_{j=1}^{K} w_j \cdot N(x_i \mid \mu_j, \Sigma_j)$$

- We introduce the indicator variable:

$$z_{ij} = \begin{cases} 1 & \text{if Gaussian } j \text{ emitted } x_i. \\ 0 & \text{otherwise.} \end{cases}$$

EM for the GMM

We can now write the joint likelihood of all X and Z as follows:

$$
\begin{aligned}
L(X, Z, \theta) &= p(X, Z \mid \theta) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{K} \left[p(x_i, j \mid \theta) \right]^{z_{ij}} \\
&= \prod_{i=1}^{N} \prod_{j=1}^{K} \left[p(x_i \mid j, \theta) \right]^{z_{ij}} \left[p(j \mid \theta) \right]^{z_{ij}}
\end{aligned}
$$

Using "log" function gives:

$$
\log\left[p(X, Z) \right] = \sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij} \log\left[p(x_i \mid j, \theta) \right] + z_{ij} \log\left[p(j \mid \theta) \right]
$$

EM for the GMM

The auxillary function is given by :

$$Q(\theta,\theta^t) = E_Z\left[\log(p(X,Z)) \mid \theta^t\right]$$

$$= E_Z\left[\sum_{i=1}^{N}\sum_{j=1}^{K} z_{ij}\log\left[p(x_i \mid j,\theta)\right] + z_{ij}\log\left[p(j \mid \theta)\right] \mid \theta^t\right]$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{K} E_Z(z_{ij} \mid \theta^t)\log\left[p(x_i \mid j,\theta)\right] + E_Z(z_{ij} \mid \theta^t)\log\left[p(j \mid \theta)\right]$$

We have then the E-Step given by :

$$E_Z(z_{ij} \mid \theta^t) = 1 \times p(z_{ij} = 1 \mid \theta^t) + 0 \times p(z_{ij} = 0 \mid \theta^t)$$

$$= p(j \mid x_i, \theta^t)$$

$$= \frac{p(j,\theta^t)p(x_i \mid j,\theta^t)}{p(x_i,\theta^t)}$$

(Posterior distribution)

EM for the GMM

And the M-Step given by:

$$\frac{\partial Q(\theta, \theta^t)}{\partial \theta} = 0$$

where $\theta = (w_j, \mu_j, \Sigma_j, j = 1, \ldots, K)$ and $\sum_{j=1}^{K} w_j = 1.$

After straightforward manipulations, we obtain:

$$\mu_j = \frac{\sum_{i=1}^{N} p(j \mid x_i, \theta^t) x_i}{\sum_{i=1}^{N} p(j \mid x_i, \theta^t)}$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} p(j \mid x_i, \theta^t)\left[(x_i - \mu_j)(x_i - \mu_j)^T\right]}{\sum_{i=1}^{N} p(j \mid x_i, \theta^t)}$$

EM for the GMM

Incorporating the constraint $\sum_{j=1}^{K} w_j = 1$ using Lagrange multiplier gives:

$$J(\theta, \theta^t) = Q(\theta, \theta^t) - \lambda\left(\sum_{j=1}^{K} w_j - 1\right)$$

Then:

$$\frac{\partial J(\theta, \theta^t)}{\partial w_j} = \frac{\partial Q(\theta, \theta^t)}{\partial w_j} - \lambda$$

$$= \sum_{i=1}^{N} \frac{p(j \mid x_i, \theta^t)}{w_j} - \lambda = 0$$

which gives:

$$w_j = \frac{\sum_{i=1}^{N} p(j \mid x_i, \theta^t)}{\lambda} \qquad (3)$$

EM for the GMM

Also, we have:

$$\frac{\partial J(\theta,\theta^t)}{\partial \lambda} = \sum_{j=1}^{K} w_j - 1 = 0 \qquad (4)$$

From (3) and (4), we obtain:

$$\frac{1}{\lambda}\sum_{j=1}^{K}\sum_{i=1}^{N} p(j \mid x_i,\theta^t) = 1$$

It follows that: $\lambda = N$

and finally:

$$w_j = \frac{1}{N}\sum_{i=1}^{N} p(j \mid x_i,\theta^t)$$

Some issues

1- Initialization:

- EM is an iterative algorithm which is very sensitive to initial conditions:

$$\text{Start from trash} \rightarrow \text{end up with trash}$$

- Usually, we use the K-Means to get a good initialization.

2- Number of Gaussians:

- Use information-theoretic criteria to obtain the optima *K*.

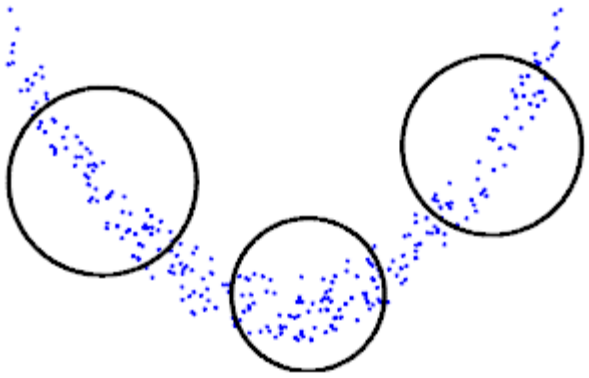
Ex: $MDL = -\log(L(X, Z, \theta)) + \left\{(K-1) + K\left[D + \frac{1}{2}D(D+1)\right]\right\}\log(N)$

Other criteria : $AIC, BIC, MML, etc.$

Some issues

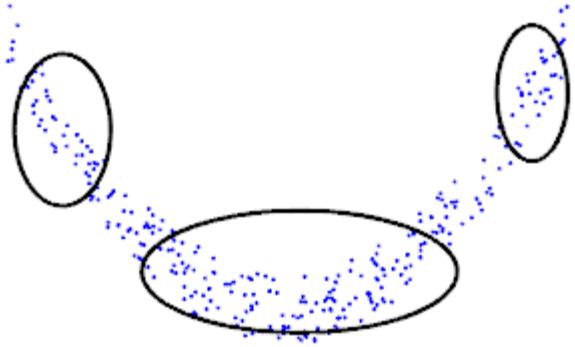
3- Simplification of the covariance matrices:

Case 1: Spherical covariance matrix $\Sigma_j = diag(\sigma_j^2, \sigma_j^2, ..., \sigma_j^2) = \sigma_j^2 I$



-Less precise.
-Very efficient to compute.

Case 2: Diagonal covariance matrix $\Sigma_j = diag(\sigma_{j1}^2, \sigma_{j2}^2, ..., \sigma_{jd}^2)$



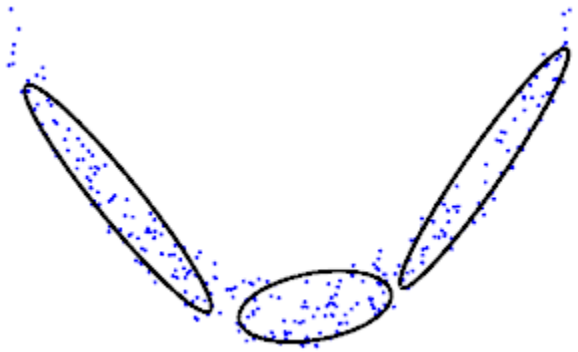
-More precise.
-Efficient to compute.

Some issues

3- Simplification of the covariance matrices:

Case 1: Full covariance matrix

$$\Sigma_j = \begin{bmatrix} \sigma_{j1}^2 & \text{cov}_j(x^1, x^2) & \cdots & \text{cov}_j(x^1, x^d) \\ \text{cov}_j(x^2, x^1) & \sigma_{j2}^2 & \cdots & \text{cov}_j(x^2, x^d) \\ \vdots & \vdots & \vdots & \vdots \\ \text{cov}_j(x^d, x^1) & \text{cov}_j(x^d, x^2) & \cdots & \sigma_{jd}^2 \end{bmatrix}$$

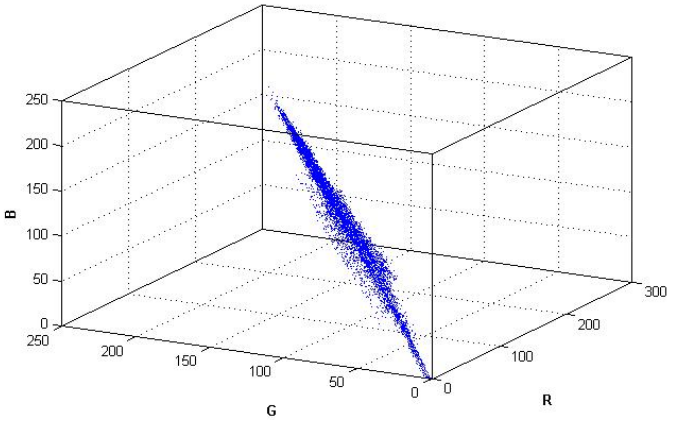


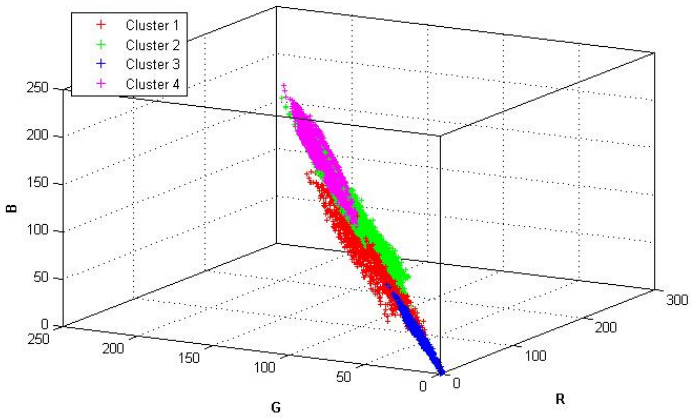
-Very precise.
-Less efficient to compute.

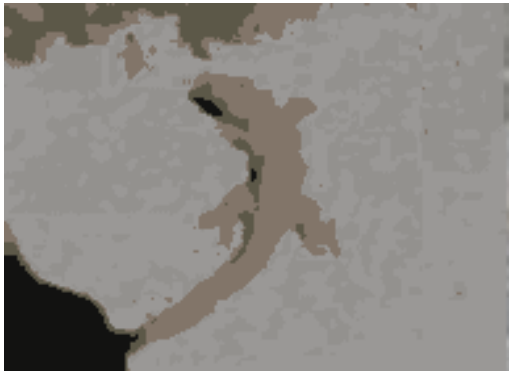
Applications of GMM in computer vision

1- Image segmentation:

$$X = (R, G, B)^T$$



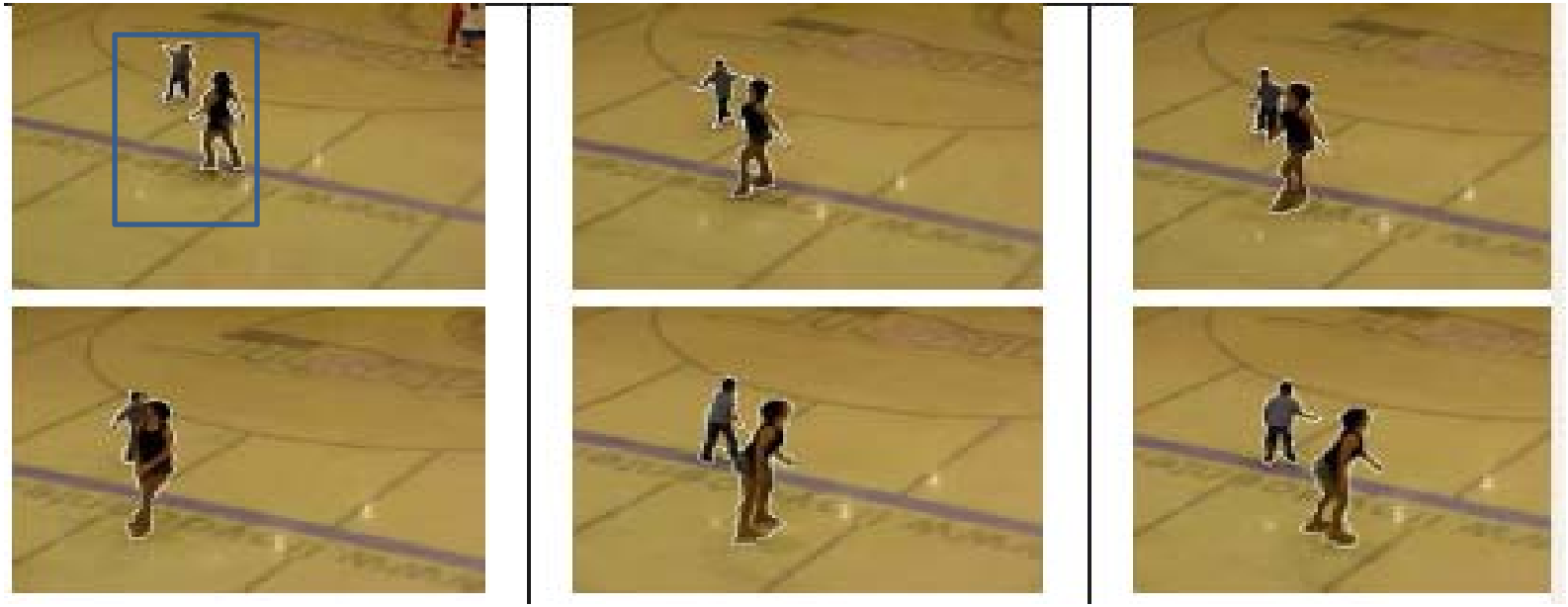




Applications of GMM in computer vision

2- Object tracking:

Knowing the moving object distribution in the first frame, we can localize the objet in the next frames by tracking its distribution.



Some references

1. Christopher M. Bishop*. Pattern Recognition and Machine Learning*, Springer, 2006.

2. Geoffrey McLachlan and David Peel. *Finite Mixture Models.* Wiley Series in Probability and Statistics, 2000.

3. Samy Bengio. *Statistical Machine Leaning from Data: Gaussian Mixture Models*. http://bengio.abracadoudou.com/lectures/gmm.pdf

4. T. Hastie et al. *The Elements of of Statistical Learning*. Springer, 2009.

Questions?