

Apprentissage statistique : modélisation descriptive et introduction aux réseaux de neurones (RCP208)

Estimation de densité

Michel Crucianu

(prenom.nom@cnam.fr)

<http://cedric.cnam.fr/vertigo/Cours/ml/>

Département Informatique
Conservatoire National des Arts & Métiers, Paris, France

10 novembre 2022

Plan du cours

2 Généralités

3 Estimation non paramétrique

- Estimation par histogramme
- Estimation par noyaux

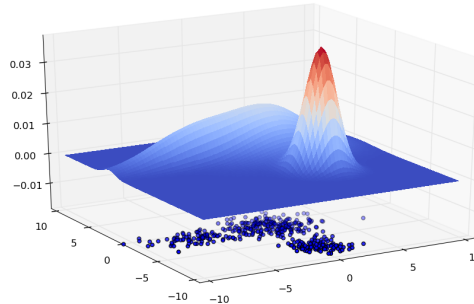
4 Estimation paramétrique

- Estimation par une loi normale
- Modèles de mélange
- Algorithme EM
- Choix du nombre de composantes
- Paramétrique ou non paramétrique ?

Objectifs et utilisations de l'estimation de densités

■ Objectif général :

- Soit $\mathcal{D}_N = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ (en général $\mathcal{D}_N \subset \mathbb{R}^d$) un ensemble de N observations
 ⇒ De quelle densité de probabilité f est issu l'échantillon aléatoire \mathcal{D}_N ?



■ Utilisations :

- Caractériser la distribution des données dans un domaine d'intérêt
 → quelles sont les régions de densité élevée, ces régions diffèrent-elles d'une classe à une autre (si les observations appartiennent à différentes classes), etc.
- Caractériser le **support** de la distribution (la région où la densité ne peut pas être assimilée à 0)
 → comment détecter les *outliers*, où faut-il envisager un rejet de non représentativité, etc.
- Construire un modèle décisionnel sur la base des densités

Typologie des méthodes d'estimation de densités

1 Méthodes non paramétriques : absence d'hypothèses sur la densité f

- 1 Estimation par histogramme
- 2 Estimation par noyaux
- 3 Estimation par k_n plus proches voisins

2 Méthodes paramétriques : hypothèses sur l'appartenance de f à une famille paramétrée → estimation des **paramètres** du modèle

- Loi simple ou mélange (en général additif) de lois simples
- Méthodes d'estimation :
 - Maximisation de la vraisemblance : modèle qui explique le mieux les observations \mathcal{D}_N
 - Maximisation de l'*a posteriori* : tenir compte aussi de connaissances *a priori*

Plan du cours

2 Généralités

3 Estimation non paramétrique

- Estimation par histogramme
- Estimation par noyaux

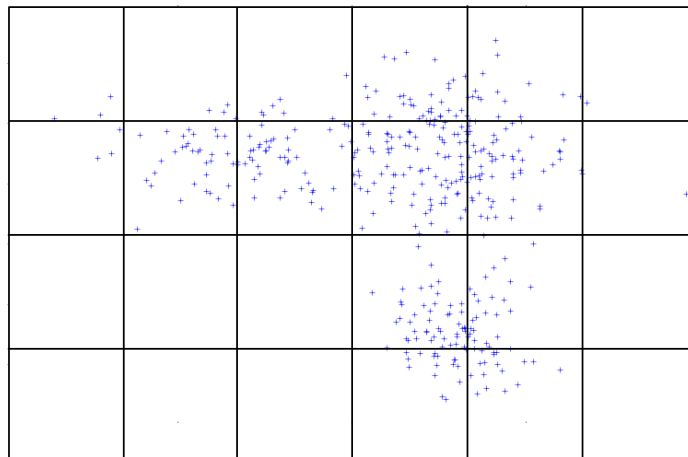
4 Estimation paramétrique

- Estimation par une loi normale
- Modèles de mélange
- Algorithme EM
- Choix du nombre de composantes
- Paramétrique ou non paramétrique ?

Estimation par histogramme

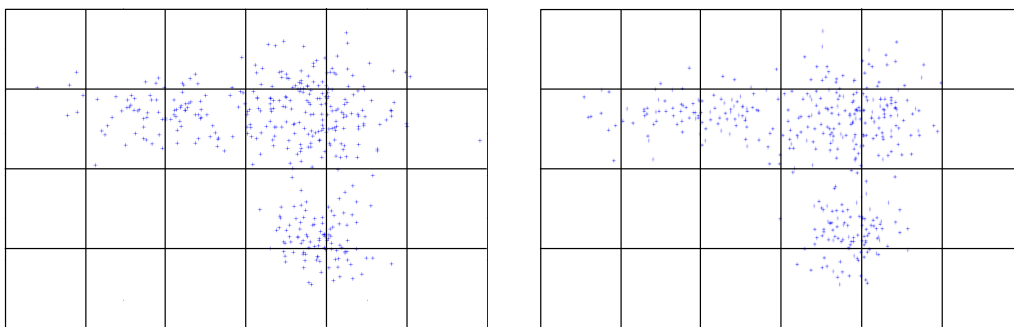
■ Méthode :

- 1 Découpage du domaine en intervalles de même « volume » v
- 2 Si $\mathbf{x} \in \mathbb{R}^d$ se trouve dans l'intervalle i qui contient k_i observations de \mathcal{D}_N , estimation de la densité en $\mathbf{x} \in \mathbb{R}^d$ par $\hat{f}(\mathbf{x}) = \frac{k_i/N}{v}$



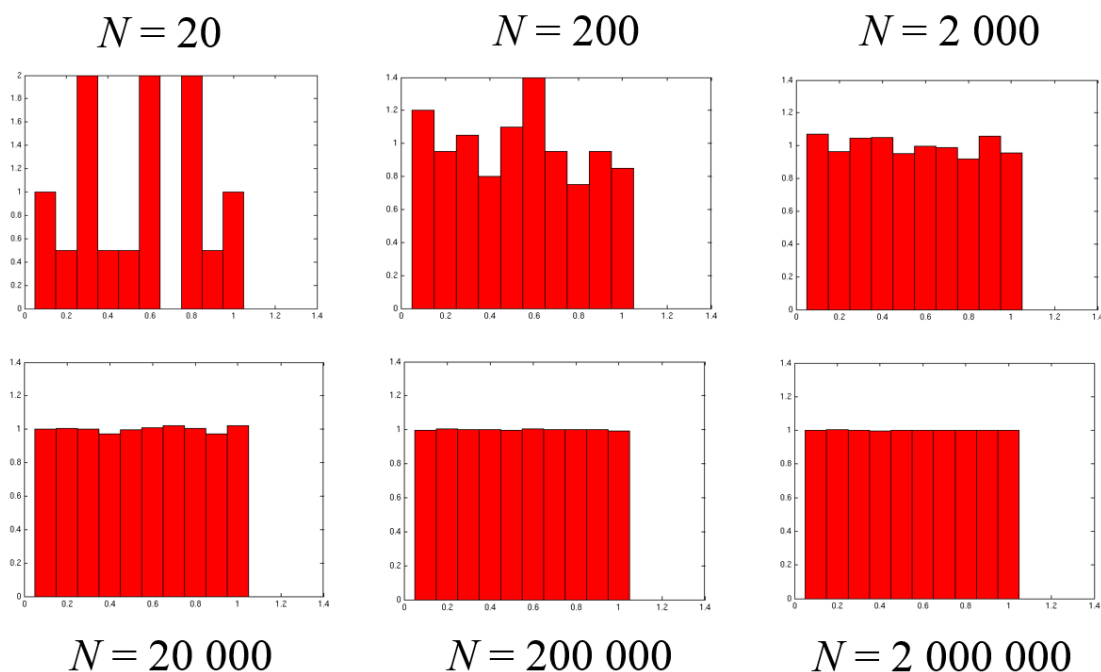
Estimation par histogramme (2)

- Caractérisation de la méthode d'estimation :
 - Qualité d'estimation : résultats proches pour un autre échantillon issu de la **même** densité f (faible variance) ?
 - Résolution : « volume » v auquel s'applique une estimation
- Constats :
 - Variance élevée si changement d'échantillon ou **décalage** des intervalles
 - Avec v fixé, si N augmente \Rightarrow la variance diminue (la qualité d'estimation s'améliore) mais représente une moyenne sur v
 - Avec N fixé, si v diminue \Rightarrow la résolution (précision par rapport à x) s'améliore mais la variance augmente (la qualité d'estimation se dégrade)



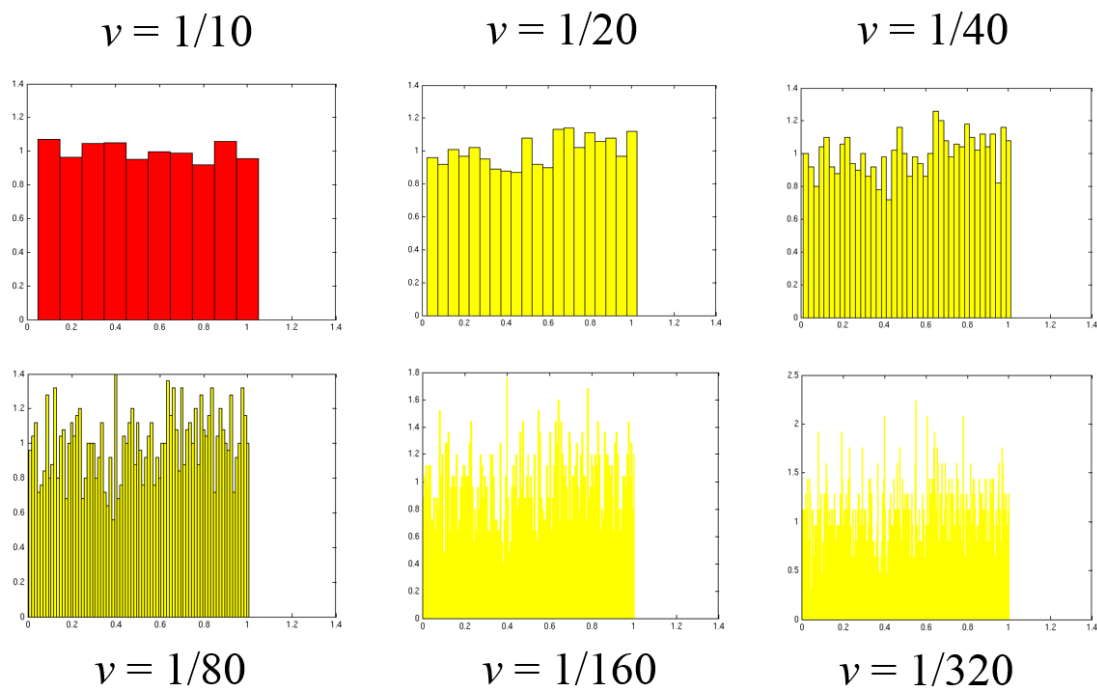
Estimation par histogramme : exemple unidimensionnel

- Données générées suivant une distribution uniforme sur $[0, 1]$
- Nombre d'intervalles fixé, taille d'échantillon augmente \Rightarrow variance diminue



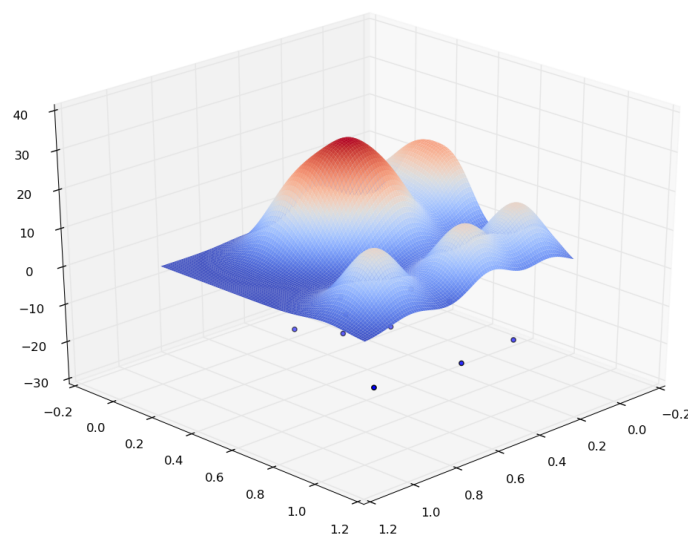
Estimation par histogramme : exemple unidimensionnel (2)

- Échantillon fixé ($N = 2000$), nombre d'intervalles augmente \Rightarrow variance augmente



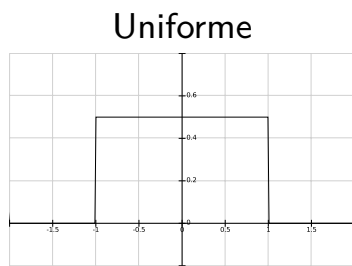
Estimation par noyaux

- Idées : lier le découpage aux données, « lisser » l'estimation de la densité
- Méthode (initiée par Rosenblatt en 1956, développée par Parzen en 1962) :
 - 1 Choisir un type de noyau Φ et un paramètre (de lissage) h
 - 2 Centrer un noyau Φ_h sur chaque observation $\mathbf{x}_i \in \mathcal{D}_N$
 - 3 La densité en un point \mathbf{x} est estimée par $\hat{f}_h(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \Phi_h(\mathbf{x}, \mathbf{x}_i)$
- Exemple :

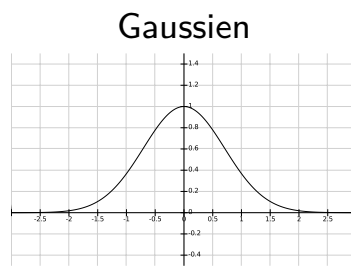


Estimation par noyaux (2)

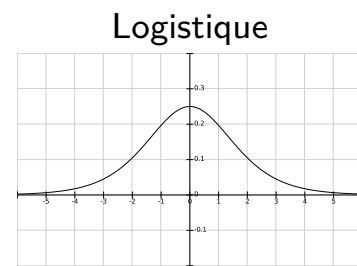
- En général les noyaux $\Phi_h : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$ s'expriment à travers des noyaux unidimensionnels $\phi : \mathbb{R} \rightarrow \mathbb{R}^+$, $\Phi_h(\mathbf{x}, \mathbf{y}) = \frac{1}{h} \phi\left(\frac{\|\mathbf{x}-\mathbf{y}\|}{h}\right)$
- Conditions suffisantes pour que l'estimateur soit une densité de probabilité :
 - 1 $\phi(u) \geq 0, \forall u \in \mathbb{R}$
 - 2 $\int_{\mathbb{R}} \phi(u) du = 1$
- Quelques noyaux :



$$\phi(u) = \frac{1}{2} \mathbf{1}_{|u| \leq 1}$$



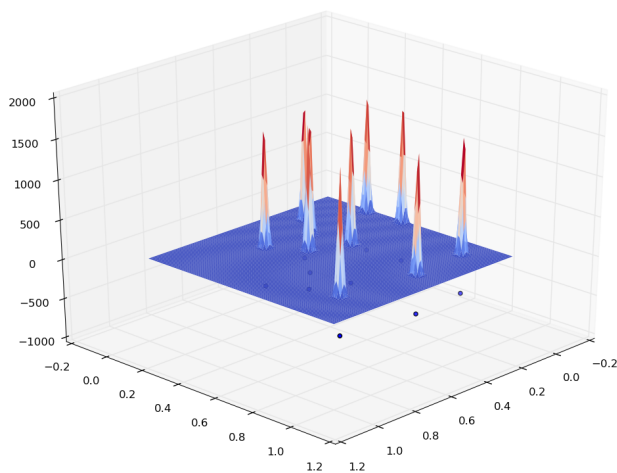
$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$



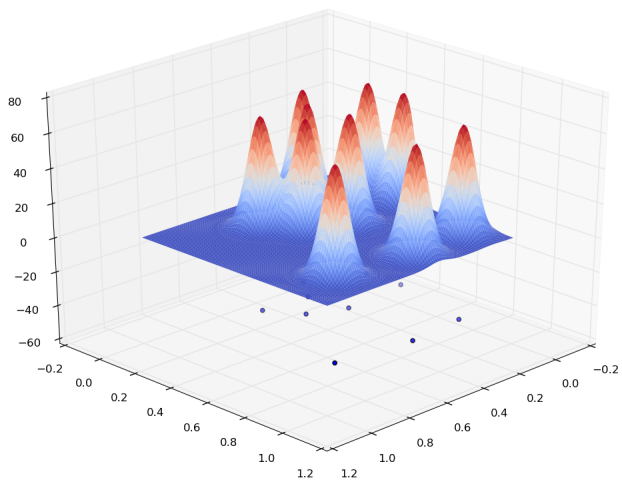
$$\phi(u) = \frac{e^{-|u|}}{(1+e^{-|u|})^2}$$

Estimation par noyaux : exemple bidimensionnel

- $N = 10$ observations dans \mathbb{R}^2 , noyau gaussien de variance indiquée



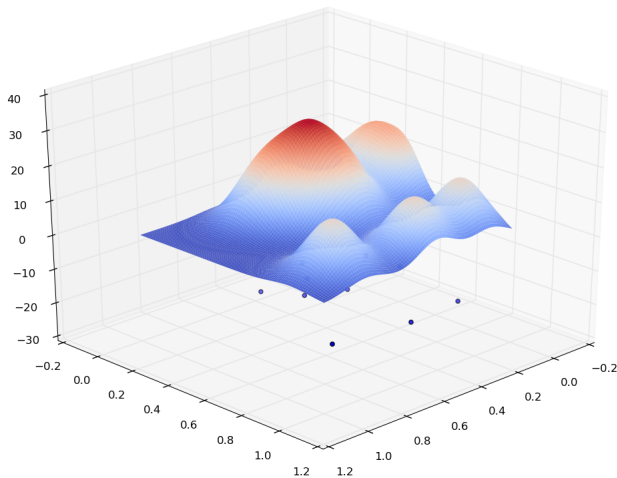
$$h^2 = 0,01$$



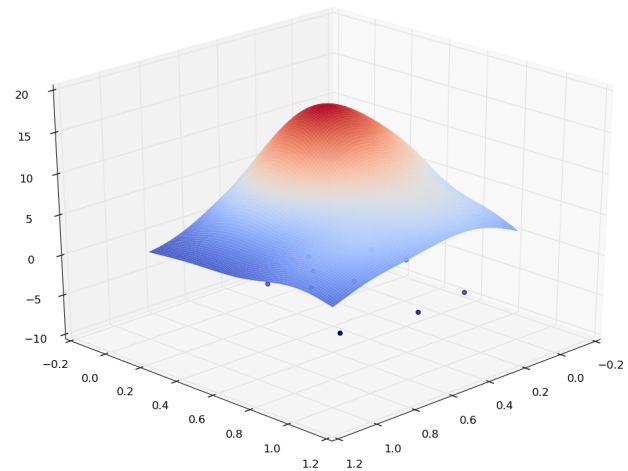
$$h^2 = 0,05$$

Estimation par noyaux : exemple bidimensionnel (2)

- $N = 10$ observations dans \mathbb{R}^2 , noyau gaussien de variance indiquée



$$h^2 = 0,1$$



$$h^2 = 0,2$$

- Choix du paramètre h (via σ pour le noyau gaussien ci-dessus) : voir [1, 3]

Plan du cours

2 Généralités

3 Estimation non paramétrique

- Estimation par histogramme
- Estimation par noyaux

4 Estimation paramétrique

- Estimation par une loi normale
- Modèles de mélange
- Algorithme EM
- Choix du nombre de composantes
- Paramétrique ou non paramétrique ?

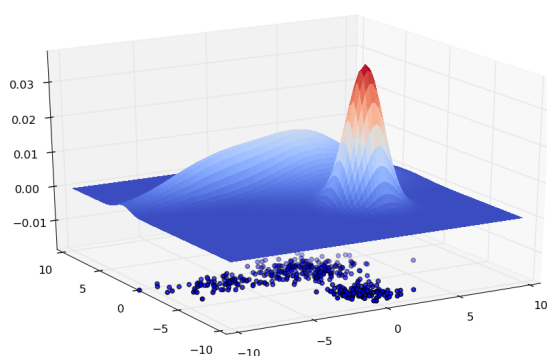
De l'estimation non paramétrique à l'estimation paramétrique

- N très grand \Rightarrow le calcul de $\hat{f}_h(\mathbf{x})$ devient très coûteux
 - \rightarrow Utilisation d'index spatiaux pour réduire la complexité, par ex. *KD tree* ou *Ball tree* dans Scikit-learn
- Et si, plutôt que d'ajouter un noyau pour chaque $\mathbf{x}_i \in \mathcal{D}_N$, on se servait plutôt de \mathcal{D}_N pour trouver les bons **paramètres** d'une densité appartenant à une famille paramétrique? Par exemple, celle des lois normales multidimensionnelles
- Si une seule loi explique mal les données, envisager alors un mélange additif de plusieurs lois de la même famille
 - \rightarrow Peu de lois, chacune estimée à partir de plusieurs données de \mathcal{D}_N , plutôt que grand nombre de lois issue chacune d'une seule donnée (estimation par noyaux)
- Si le choix de la famille paramétrique est bien fondé, il est possible d'obtenir ainsi de meilleurs résultats avec moins de données

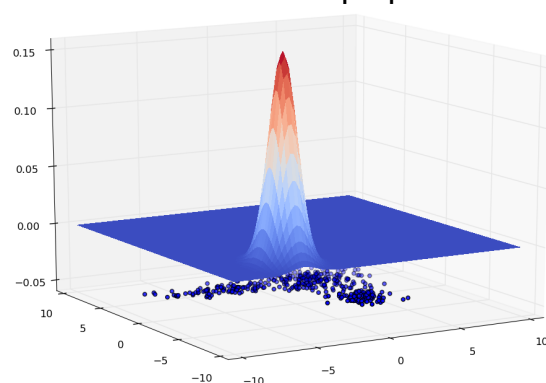
Estimation paramétrique

- On cherche la densité dans une **famille** \mathcal{F} **paramétrée** par un vecteur de paramètres $\theta \in \Omega$ (on indique cela en écrivant f_θ au lieu de f)
- Étant donné \mathcal{D}_N , comment trouver le « bon » vecteur de paramètres?
- Un vecteur θ engendre une densité f_θ qui « explique » plus ou moins bien les observations \mathcal{D}_N :

Données bien expliquées



Données mal expliquées



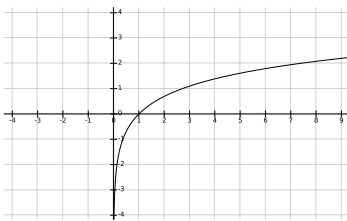
Estimation paramétrique (2)

- Une densité f_θ qui « explique » plus ou moins bien les observations \mathcal{D}_N , tel qu'indiqué par

$$p(\mathcal{D}_N|\theta) = \prod_{i=1}^N f_\theta(\mathbf{x}_i)$$

- Comme fonction de θ , $p(\mathcal{D}_N|\theta)$ est la **vraisemblance** (*likelihood*) de θ par rapport à l'échantillon \mathcal{D}_N
- On préfère en général travailler avec le logarithme de la vraisemblance (*log-likelihood*)

$$L(\theta) \equiv \ln [p(\mathcal{D}_N|\theta)] = \sum_{i=1}^N \ln [f_\theta(\mathbf{x}_i)]$$

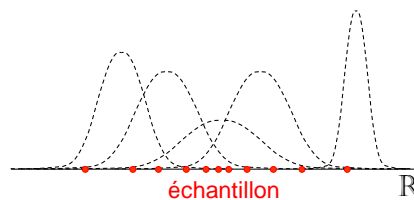


← \ln : fonction strictement croissante

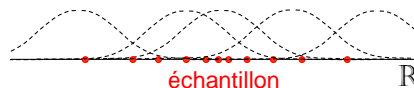
Cas simple : loi normale unidimensionnelle

- Considérons des données unidimensionnelles, $\mathcal{D}_N \subset \mathbb{R}$ ($d = 1$)
- Famille paramétrée : les lois normales $\mathcal{N}(\mu, \sigma)$, donc $\theta = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$

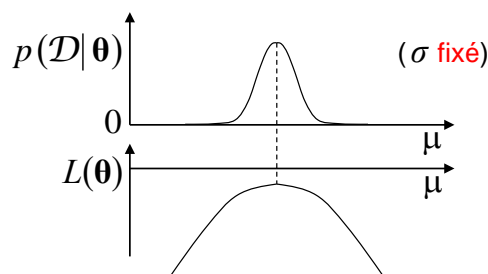
Quelques densités candidates :



Avec σ fixé :



- Variation de la vraisemblance et de la log-vraisemblance pour σ fixé :



Maximum de vraisemblance

- L'estimation $\hat{\theta}$ la plus en accord avec les observations \mathcal{D}_N est celle qui correspond au maximum de la vraisemblance $p(\mathcal{D}_N|\theta)$
 - Logarithme en base e est monotone croissant \Rightarrow maximum de la vraisemblance est atteint pour le même $\hat{\theta}$ que le maximum de la log-vraisemblance $l(\theta)$
- Pour le cas normal unidimensionnel considéré, $f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ donc :

$$\ln [p(\mathcal{D}_N|\theta)] = \sum_{i=1}^N \ln [f_\theta(x_i)] = \sum_{i=1}^N \left[-\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right]$$

- En dérivant par rapport à μ et respectivement σ on obtient

$$\frac{\partial \ln [p(\mathcal{D}_N|\theta)]}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \mu)$$

$$\frac{\partial \ln [p(\mathcal{D}_N|\theta)]}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^N (x_i - \mu)^2$$

Maximum de vraisemblance (2)

- Les dérivées partielles

$$\frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \mu) \quad \text{et} \quad -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^N (x_i - \mu)^2$$

s'annulent pour

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i \quad = \text{moyenne empirique des observations}$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2 \quad = \text{variance empirique biaisée}$$

et on peut vérifier que cette solution correspond bien à un **maximum** de $\ln [p(\mathcal{D}_N|\theta)]$

- Remarques :

- L'estimation $\hat{\mu}$ est sans biais (son espérance sur les échantillons de taille N est égale à la vraie valeur de μ)
- L'estimation $\hat{\sigma}^2$ est en revanche **biaisée** (mais asymptotiquement sans biais, $\lim_{N \rightarrow \infty} E(\hat{\sigma}^2)$ est la vraie valeur de σ); une estimation sans biais est

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

Cas simple : loi normale multidimensionnelle

- Considérons des données multidimensionnelles, $\mathcal{D}_N \subset \mathbb{R}^d$ ($d > 1$)
- Famille paramétrée : les lois normales $\mathcal{N}(\mu, \Sigma)$

$$f_{\theta}(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1}(\mathbf{x}-\mu)}$$

- Ici, le vecteur de paramètres θ regroupe à la fois les composantes de l'espérance multidimensionnelle μ et de la matrice de variance-covariance Σ
- La solution qui maximise la vraisemblance est

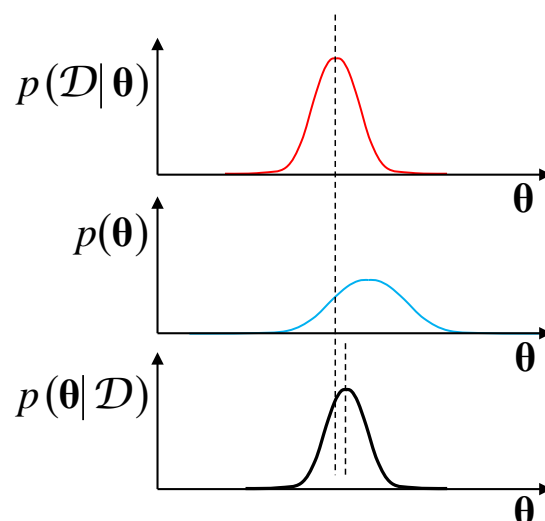
$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \quad \text{et} \quad \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^t$$

- Ici encore, l'estimation $\hat{\Sigma}$ est biaisée et une estimation sans biais est

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^t$$

Maximum a posteriori

- La solution qui maximise la vraisemblance est la plus en accord avec les (seules) données de \mathcal{D}_N
- Et si on dispose de **connaissances a priori** nous incitant à privilégier certains modèles (certaines valeurs des paramètres) par rapport à d'autres ?
- Connaissances \rightarrow densité a priori $p(\theta)$ \rightarrow plutôt **maximiser la probabilité a posteriori** $p(\theta|\mathcal{D}_N) \propto p(\mathcal{D}_N|\theta)p(\theta)$



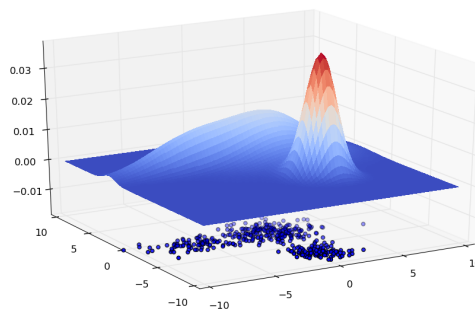
Modèles de mélange

- Une seule loi explique mal les données \Rightarrow mélange additif de plusieurs lois de la même famille

$$f_{\alpha, \theta}(\mathbf{x}) = \sum_{j=1}^m \alpha_j f_j(\mathbf{x}|\theta_j)$$

avec

- m : nombre de composantes (lois individuelles) du mélange
- $f_j(\mathbf{x}|\theta_j)$: densité qui définit une composante, appartient à la famille paramétrée par θ_j
- α_j : coefficients de mélange tels que $\sum_{j=1}^m \alpha_j = 1$ (condition de normalisation)
- α : représentation vectorielle des m coefficients de mélange (pondérations des lois)
- θ : vecteur qui regroupe tous les θ_j



Modèles de mélange : estimation MV

- On cherche à trouver les paramètres $(\hat{\alpha}, \hat{\theta})$ qui maximisent la vraisemblance (MV) des observations de \mathcal{D}_N

$$p(\mathcal{D}_N|\alpha, \theta) = \prod_{i=1}^N f_{\alpha, \theta}(\mathbf{x}_i)$$

- En écrivant $f_{\alpha, \theta}(\mathbf{x}_i)$ comme un mélange additif, on obtient l'expression de la vraisemblance à maximiser

$$\ln p(\mathcal{D}_N|\alpha, \theta) = \sum_{i=1}^N \ln \left[\sum_{j=1}^m \alpha_j f_j(\mathbf{x}_i|\theta_j) \right] \quad (1)$$

sous la contrainte de normalisation $\sum_{j=1}^m \alpha_j = 1$

- Somme sous le logarithme \Rightarrow pas de solution analytique à ce problème de maximisation \Rightarrow méthodes d'**optimisation itérative** pour déterminer $(\hat{\alpha}, \hat{\theta})$

Mélange gaussien : estimation MV

- Pour un mélange additif gaussien, chaque composante est une loi normale :

$$f_j(\mathbf{x}|\theta_j) = (2\pi)^{-\frac{d}{2}} |\Sigma_j|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_j)^t \Sigma_j^{-1} (\mathbf{x}-\mu_j)} \quad (2)$$

et donc la log-vraisemblance $L(\alpha, \theta)$ devient

$$\ln p(\mathcal{D}_N|\alpha, \theta) = -\frac{Nd}{2} \ln(2\pi) + \sum_{i=1}^N \ln \left[\sum_{j=1}^m \alpha_j |\Sigma_j|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_i-\mu_j)^t \Sigma_j^{-1} (\mathbf{x}_i-\mu_j)} \right] \quad (3)$$

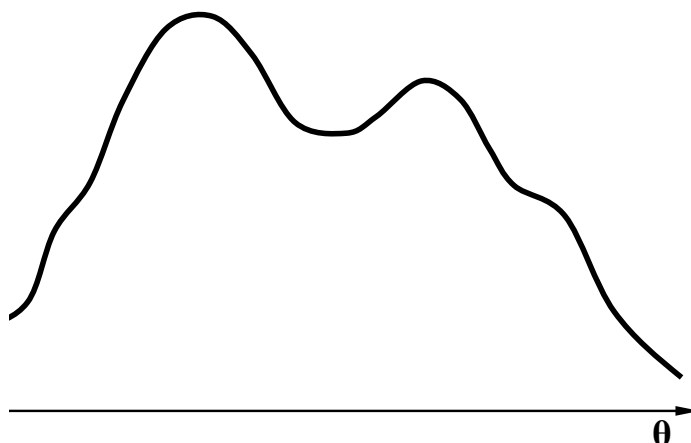
- recherche de maxima : absence de solution analytique
- présence de maxima locaux

- Idée de l'algorithme EM :

- Déterminer une approximation locale $l(\alpha, \theta)$ de $L(\alpha, \theta)$ qui minore $L(\alpha, \theta)$ et peut être maximisée analytiquement
- Calculer $\hat{\alpha}_t, \hat{\theta}_t$ qui maximisent cette approximation locale
- Itérer jusqu'à convergence

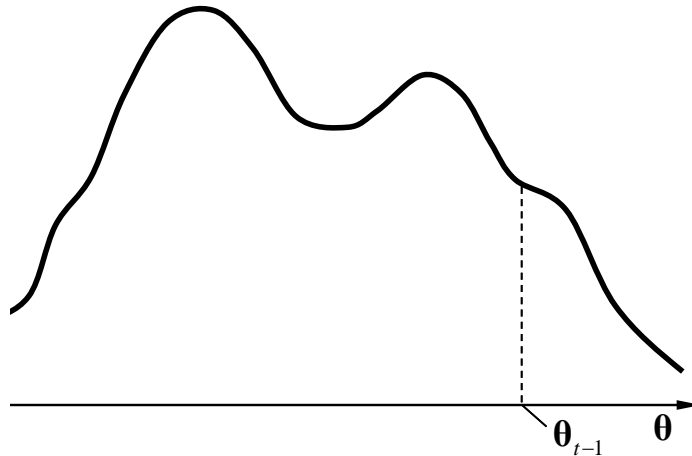
Principe de l'algorithme EM

- Initialiser les paramètres (réduits ici à θ pour éviter de surcharger l'illustration)
- Itérer jusqu'à la convergence
 - 1 Étape E : avec les paramètres actuels, calculer l'approximation locale tangente qui minore la fonction à maximiser
 - 2 Étape M : trouver les paramètres qui maximisent l'approximation locale, en faire les nouveaux paramètres actuels



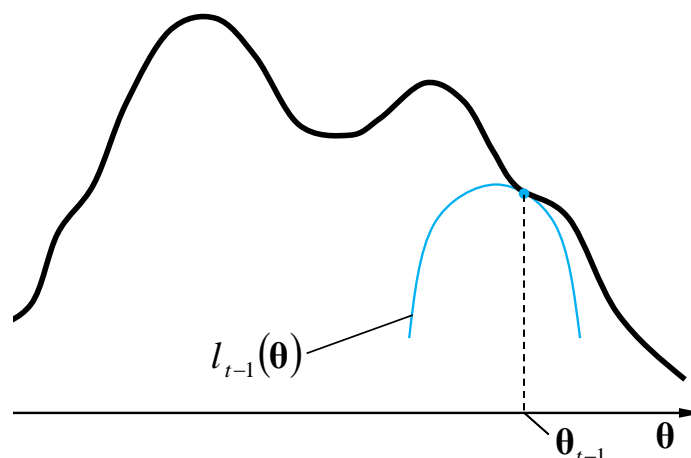
Principe de l'algorithme EM

- Initialiser les paramètres (réduits ici à θ pour éviter de surcharger l'illustration)
- Itérer jusqu'à la convergence
 - 1 Étape E : avec les paramètres actuels, calculer l'approximation locale tangente qui minore la fonction à maximiser
 - 2 Étape M : trouver les paramètres qui maximisent l'approximation locale, en faire les nouveaux paramètres actuels



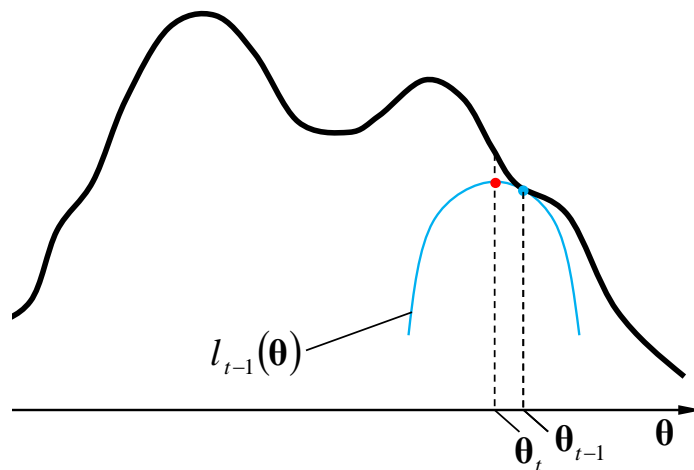
Principe de l'algorithme EM

- Initialiser les paramètres (réduits ici à θ pour éviter de surcharger l'illustration)
- Itérer jusqu'à la convergence
 - 1 Étape E : avec les paramètres actuels, calculer l'approximation locale tangente qui minore la fonction à maximiser
 - 2 Étape M : trouver les paramètres qui maximisent l'approximation locale, en faire les nouveaux paramètres actuels



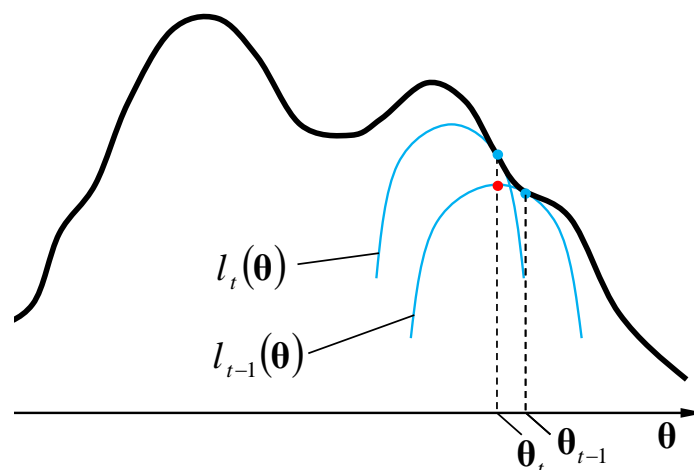
Principe de l'algorithme EM

- Initialiser les paramètres (réduits ici à θ pour éviter de surcharger l'illustration)
- Itérer jusqu'à la convergence
 - 1 **Étape E** : avec les paramètres actuels, calculer l'approximation locale tangente qui minore la fonction à maximiser
 - 2 **Étape M** : trouver les paramètres qui maximisent l'approximation locale, en faire les nouveaux paramètres actuels



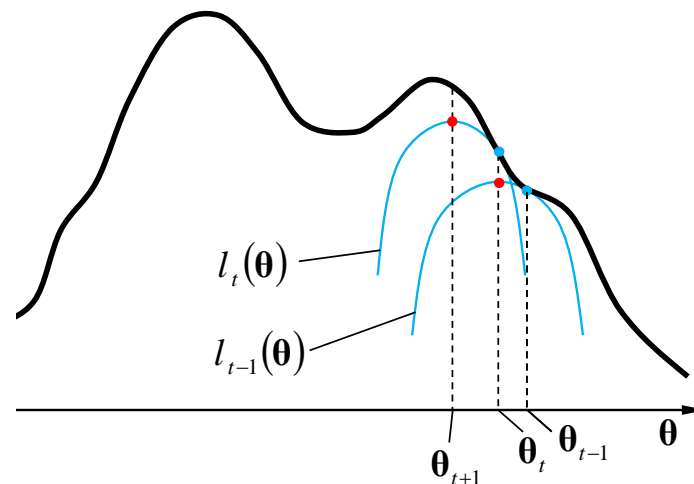
Principe de l'algorithme EM

- Initialiser les paramètres (réduits ici à θ pour éviter de surcharger l'illustration)
- Itérer jusqu'à la convergence
 - 1 **Étape E** : avec les paramètres actuels, calculer l'approximation locale tangente qui minore la fonction à maximiser
 - 2 **Étape M** : trouver les paramètres qui maximisent l'approximation locale, en faire les nouveaux paramètres actuels



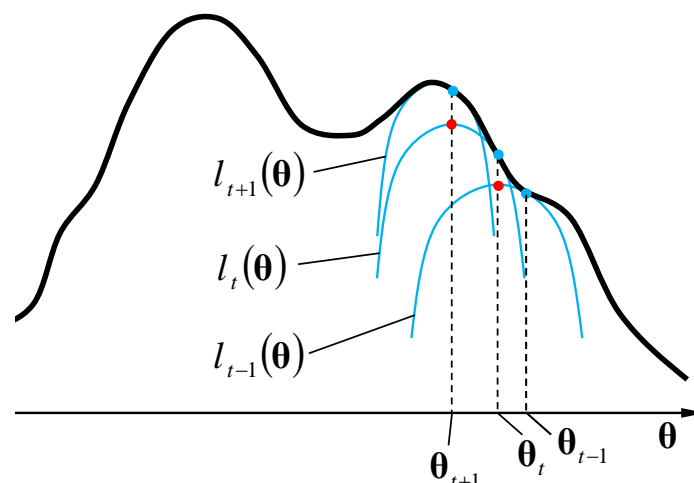
Principe de l'algorithme EM

- Initialiser les paramètres (réduits ici à θ pour éviter de surcharger l'illustration)
- Itérer jusqu'à la convergence
 - 1 **Étape E** : avec les paramètres actuels, calculer l'approximation locale tangente qui minore la fonction à maximiser
 - 2 **Étape M** : trouver les paramètres qui maximisent l'approximation locale, en faire les nouveaux paramètres actuels



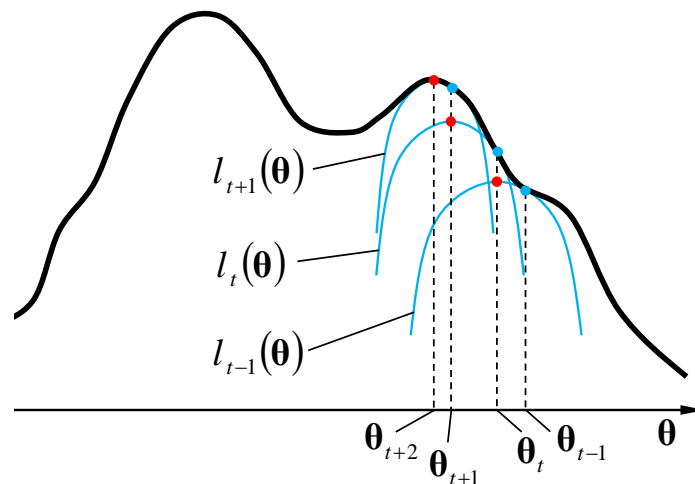
Principe de l'algorithme EM

- Initialiser les paramètres (réduits ici à θ pour éviter de surcharger l'illustration)
- Itérer jusqu'à la convergence
 - 1 **Étape E** : avec les paramètres actuels, calculer l'approximation locale tangente qui minore la fonction à maximiser
 - 2 **Étape M** : trouver les paramètres qui maximisent l'approximation locale, en faire les nouveaux paramètres actuels



Principe de l'algorithme EM

- Initialiser les paramètres (réduits ici à θ pour éviter de surcharger l'illustration)
- Itérer jusqu'à la convergence
 - 1 Étape E : avec les paramètres actuels, calculer l'approximation locale tangente qui minore la fonction à maximiser
 - 2 **Étape M** : trouver les paramètres qui maximisent l'approximation locale, en faire les nouveaux paramètres actuels



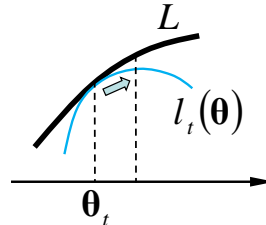
Espérance-Maximisation (EM)

- Algorithme introduit dans [2] ; de nombreuses présentations alternatives existent, voir par ex. celle de <https://arxiv.org/abs/1105.1476v2>
- On considère des données $\mathcal{D} = \{\mathcal{D}_o, \mathcal{D}_m\}$ où
 - \mathcal{D}_o sont des données **observées**, de réalisations notées \mathbf{x} , issues d'une densité de probabilité $f_\theta \in \mathcal{F}$, inconnue et paramétrée par $\theta \in \Omega$
 - \mathcal{D}_m sont des données **manquantes**, de réalisations notées \mathbf{y} , issues d'une densité de probabilité g (inconnue)
- On cherche le paramètre $\hat{\theta}$ qui maximise la log-vraisemblance $\ln p(\mathcal{D}_o|\theta)$ (notée dans la suite de façon simplifiée par $L(\mathbf{x}; \theta)$)
 - $\ln p(\mathcal{D}_o|\theta)$: log-vraisemblance des données incomplètes
 - $\ln p(\mathcal{D}_o, \mathcal{D}_m|\theta)$: log-vraisemblance des données complètes ($L((\mathbf{x}, \mathbf{y}); \theta)$ dans la suite)

Espérance-Maximisation (2)

- Pour tout vecteur de paramètres fixé θ_t , trouvons $l_t(\theta)$ une approximation locale de $L(\mathbf{x}; \theta)$ telle que
 1. $L(\mathbf{x}; \theta) - L(\mathbf{x}; \theta_t) \geq l_t(\theta) - l_t(\theta_t)$
 2. $L(\mathbf{x}; \theta_t) = l_t(\theta_t)$
- Maximiser (ou simplement augmenter) $l_t(\theta)$ permettra alors de s'approcher du

maximum de $L(\mathbf{x}; \theta)$:



1. Nous pouvons écrire

$$\begin{aligned}
 L(\mathbf{x}; \theta) &= \ln p(\mathbf{x}|\theta) && \text{(notation simplifiée de } \ln p(\mathcal{D}_o|\theta)) \\
 &= \ln \int p(\mathbf{x}, \mathbf{y}|\theta) d\mathbf{y} \\
 &= \ln \int p(\mathbf{x}, \mathbf{y}|\theta) \frac{p(\mathbf{y}|\mathbf{x}, \theta_t)}{p(\mathbf{y}|\mathbf{x}, \theta_t)} d\mathbf{y} \\
 &= \ln \int p(\mathbf{y}|\mathbf{x}, \theta_t) \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{p(\mathbf{y}|\mathbf{x}, \theta_t)} d\mathbf{y} \\
 &\geq \underbrace{\int p(\mathbf{y}|\mathbf{x}, \theta_t) \ln \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{p(\mathbf{y}|\mathbf{x}, \theta_t)} d\mathbf{y}}_{\equiv l_t(\theta)} && \text{(par l'inégalité de Jensen)}
 \end{aligned}$$

(\mathbf{y} discret \Rightarrow somme plutôt qu'intégrale, l'inégalité fait passer le log sous la somme !)

Espérance-Maximisation (3)

2. Également, avec cette définition pour $l_t(\theta)$, nous avons

$$\begin{aligned}
 l_t(\theta_t) &= \int p(\mathbf{y}|\mathbf{x}, \theta_t) \ln \frac{p(\mathbf{x}, \mathbf{y}|\theta_t)}{p(\mathbf{y}|\mathbf{x}, \theta_t)} d\mathbf{y} \\
 &= \int p(\mathbf{y}|\mathbf{x}, \theta_t) \ln p(\mathbf{x}|\theta_t) d\mathbf{y} && \text{(car } p(\mathbf{x}, \mathbf{y}|\theta_t) = p(\mathbf{y}|\mathbf{x}, \theta_t)p(\mathbf{x}|\theta_t)) \\
 &= \ln p(\mathbf{x}|\theta_t) && \text{(car } \int p(\mathbf{y}|\mathbf{x}, \theta_t) d\mathbf{y} = 1) \\
 &= L(\mathbf{x}; \theta_t)
 \end{aligned}$$

- L'algorithme EM résultant sera donc :

- 1 Initialiser les paramètres θ_0
- 2 Itérer jusqu'à la convergence
 - 1 Étape E : calculer $l_t(\theta) \leq L(\mathbf{x}; \theta)$, égalité pour $\theta = \theta_t$
 - 2 Étape M : trouver θ_{t+1} qui maximise $l_t(\theta)$

- Utilisations de EM :

- Modélisation à partir de données incomplètes (par ex. difficultés d'observation)
- Introduction artificielle de variable(s) à valeurs manquantes pour faciliter la maximisation de la vraisemblance (variable(s) définie(s) t.q. maximisation des $l_t(\theta)$ successives plus facile que maximisation directe de $L(\mathbf{x}; \theta)$)

Espérance-Maximisation (4)

- Pourquoi le nom « Espérance » (*Expectation*) pour l'étape E ?
- Nous pouvons développer $l_t(\theta)$:

$$\begin{aligned} l_t(\theta) &= \int p(\mathbf{y}|\mathbf{x}, \theta_t) \ln \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{p(\mathbf{y}|\mathbf{x}, \theta_t)} d\mathbf{y} \\ &= \underbrace{\int p(\mathbf{y}|\mathbf{x}, \theta_t) \ln p(\mathbf{x}, \mathbf{y}|\theta) d\mathbf{y}}_{= Q(\theta|\theta_t)} - \underbrace{\int p(\mathbf{y}|\mathbf{x}, \theta_t) \ln p(\mathbf{y}|\mathbf{x}, \theta_t) d\mathbf{y}}_{= \text{constant (non dépendant de } \theta)} \end{aligned}$$

- $Q(\theta|\theta_t)$ est l'**espérance conditionnelle** de la log-vraisemblance des données complètes ($L((\mathbf{x}, \mathbf{y}); \theta)$ calculée pour des données manquantes qui suivent leur densité *a posteriori* $p(\mathbf{y}|\mathbf{x}, \theta_t)$)
- La différence entre $l_t(\theta)$ et $Q(\theta|\theta_t)$ ne dépend pas de θ , la maximisation d'une des fonctions est équivalente à la maximisation de l'autre

L'inégalité de Jensen

- Si $\phi : \mathcal{A} \in \mathbb{R} \rightarrow \mathbb{R}$ est une fonction convexe et X une variable aléatoire à valeurs dans \mathcal{A} , alors

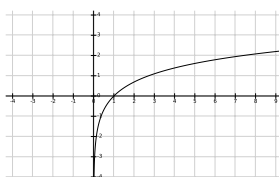
$$\phi(E[X]) \leq E[\phi(X)]$$
- Pour une densité de probabilité : soit g une fonction à valeurs réelles, ϕ convexe sur le co-domaine de g et f une densité de probabilité, alors

$$\phi\left(\int g(x)f(x)dx\right) \leq \int \phi(g(x))f(x)dx$$

- Cas d'un échantillon (fini) : soit $\lambda_i \geq 0$, $\sum_{i=1}^N \lambda_i = 1$, alors

$$\phi\left(\sum_{i=1}^N \lambda_i g(x)\right) \leq \sum_{i=1}^N \lambda_i \phi(g(x))$$

- Si ϕ est **concave**, comme \ln (voir figure), alors le sens des inégalités est **inversé**



EM pour modèle de mélange

- Introduction de variables aléatoires **non observées** $\{Y_i\}_{i=1}^N$ à valeurs $y_i \in \{1, \dots, m\}$, telles que $y_i = j \Leftrightarrow$ l'observation \mathbf{x}_i a été générée par la composante j du mélange
- Pour maximiser $\ln p(\mathcal{D}_o | \alpha, \theta)$, la log-vraisemblance des données incomplètes, le vecteur de paramètres à trouver est $\begin{pmatrix} \hat{\alpha} \\ \hat{\theta} \end{pmatrix}$
- Les variables $\{Y_i\}$ prenant des valeurs discrètes, l'expression de $Q(\theta | \theta_t)$ devient

$$\begin{aligned} Q(\alpha, \theta | \alpha_t, \theta_t) &= \sum_{i=1}^N \sum_{k=1}^m P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t) \ln p(\mathbf{x}_i, y_i = k | \alpha, \theta) \\ &= \sum_{i=1}^N \sum_{k=1}^m P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t) \ln [p(\mathbf{x}_i | y_i = k, \theta_k) P(y_i = k | \alpha, \theta)] \\ &= \sum_{i=1}^N \sum_{k=1}^m P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t) \ln [P(y_i = k | \alpha, \theta) f_k(\mathbf{x}_i | \theta_k)] \end{aligned}$$

en vertu de la définition des données non observées (manquantes) ; ici f_k est la densité de la k -ème composante du mélange (de paramètres θ_k)

- Il n'y a plus de somme sous le logarithme, la maximisation de $Q(\alpha, \theta | \alpha_t, \theta_t)$ admet donc une solution analytique, contrairement à celle de la log-vraisemblance des données incomplètes, expression (1)

EM pour mélange gaussien

- La distribution $P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t)$ est obtenue par normalisation

$$P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t) = \frac{\alpha_k f_k(\mathbf{x}_i | \theta_{kt})}{\sum_{j=1}^m \alpha_j f_j(\mathbf{x}_i | \theta_{jt})}$$

$f_k(\mathbf{x}_i | \theta_{kt})$ étant la densité de la composante k avec les paramètres de l'itération t

- $Q(\alpha, \theta | \alpha_t, \theta_t)$ devient donc :

$$\begin{aligned} Q(\alpha, \theta | \alpha_t, \theta_t) &= \sum_{i=1}^N \sum_{k=1}^m P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t) \ln [\alpha_k f_k(\mathbf{x}_i | \theta_k)] \\ &= \underbrace{\sum_{i=1}^N \sum_{k=1}^m P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t) \ln \alpha_k}_{= \text{constant (non dépendant de } \theta)} + \sum_{i=1}^N \sum_{k=1}^m \frac{\alpha_k f_k(\mathbf{x}_i | \theta_{kt})}{\sum_{j=1}^m \alpha_j f_j(\mathbf{x}_i | \theta_{jt})} \ln f_k(\mathbf{x}_i | \theta_k) \end{aligned}$$

- La maximisation s'intéresse au second terme qui, tenant compte de (2), devient

$$\sum_{i=1}^N \sum_{k=1}^m \frac{\alpha_k f_k(\mathbf{x}_i | \mu_{kt}, \Sigma_{kt})}{\sum_{j=1}^m \alpha_j f_j(\mathbf{x}_i | \mu_{jt}, \Sigma_{jt})} \left[-\frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (\mathbf{x} - \mu_k)^t \Sigma_k^{-1} (\mathbf{x} - \mu_k) \right]$$

EM pour mélange gaussien (2)

- Nous obtenons *in fine* les relations de mise à jour suivantes :

$$P(y_i = k | \mathbf{x}_i, \alpha_t, \mu_t, \Sigma_t) = \frac{\alpha_k f_k(\mathbf{x}_i | \mu_{kt}, \Sigma_{kt})}{\sum_{j=1}^m \alpha_j f_j(\mathbf{x}_i | \mu_{jt}, \Sigma_{jt})} \quad (4)$$

$$\alpha_{k,t+1} = \frac{1}{N} \sum_{i=1}^N P(y_i = k | \mathbf{x}_i, \alpha_t, \mu_t, \Sigma_t) \quad (5)$$

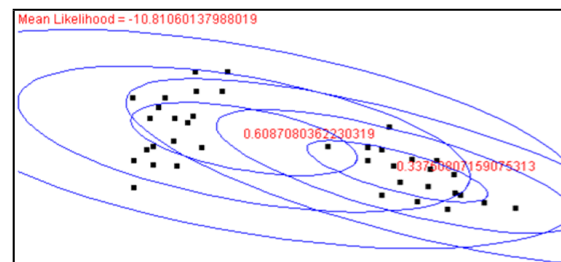
$$\mu_{k,t+1} = \frac{\sum_{i=1}^N P(y_i = k | \mathbf{x}_i, \alpha_t, \mu_t, \Sigma_t) \mathbf{x}_i}{\sum_{i=1}^N P(y_i = k | \mathbf{x}_i, \alpha_t, \mu_t, \Sigma_t)} \quad (6)$$

$$\Sigma_{k,t+1} = \frac{\sum_{i=1}^N P(y_i = k | \mathbf{x}_i, \alpha_t, \mu_t, \Sigma_t) (\mathbf{x}_i - \mu_{k,t+1})(\mathbf{x}_i - \mu_{k,t+1})^t}{\sum_{i=1}^N P(y_i = k | \mathbf{x}_i, \alpha_t, \mu_t, \Sigma_t)} \quad (7)$$

- On peut placer (4) dans l'étape E et (5), (6), (7) dans l'étape M

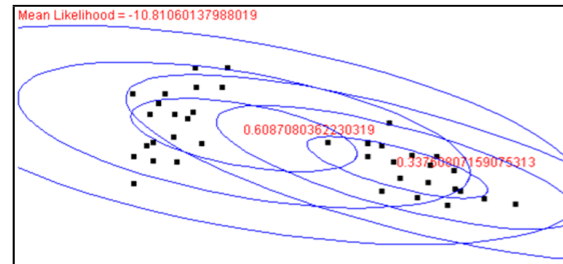
EM pour mélange gaussien : illustration

- Au départ composantes peu spécialisées
 - Valeurs $P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t)$ équilibrées
- Après 1 itération :

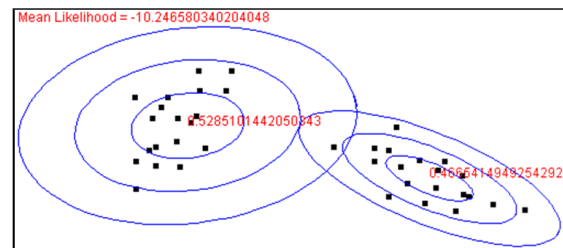


EM pour mélange gaussien : illustration

- Au départ composantes peu spécialisées
 - Valeurs $P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t)$ équilibrées
- Après 1 itération :

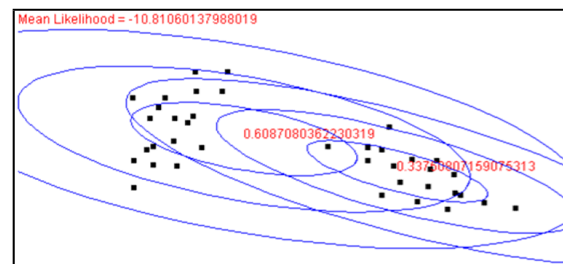


- Spécialisation progressive
 - Les $P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t)$ se différencient
 - Paramètres calculés à partir des données mieux expliquées
- Après 4 itérations :

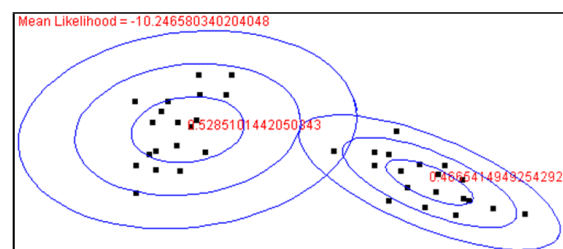


EM pour mélange gaussien : illustration

- Au départ composantes peu spécialisées
 - Valeurs $P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t)$ équilibrées
- Après 1 itération :



- Spécialisation progressive
 - Les $P(y_i = k | \mathbf{x}_i, \alpha_t, \theta_t)$ se différencient
 - Paramètres calculés à partir des données mieux expliquées
- Après 4 itérations :



- Spécialisation complète
 - Paramètres calculés à partir des données expliquées
- Après ≥ 5 itérations :

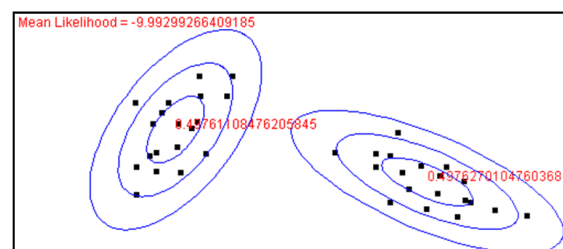
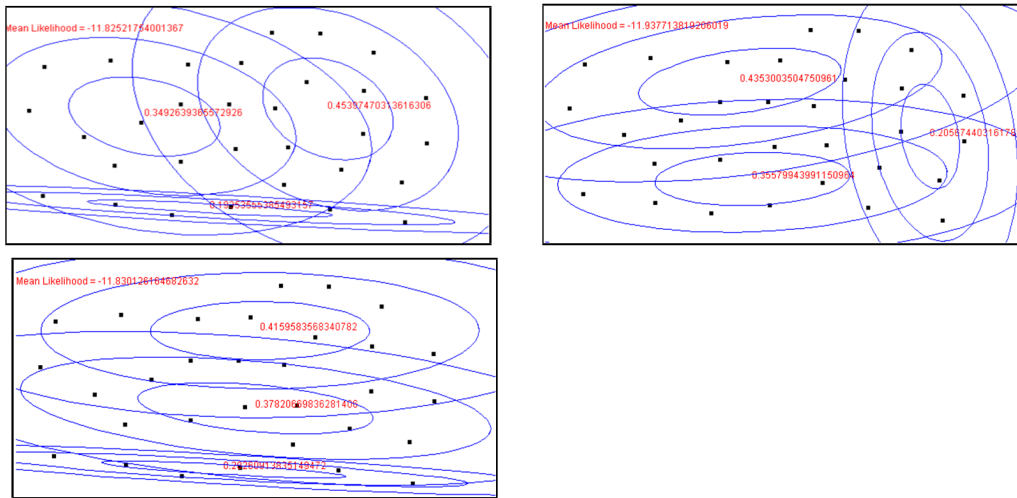


Illustration : données non structurées

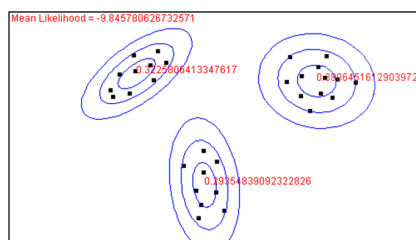
- Résultats obtenus avec des initialisations différentes de l'algorithme EM pour des données issues d'une distribution uniforme :



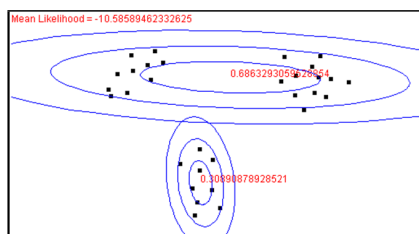
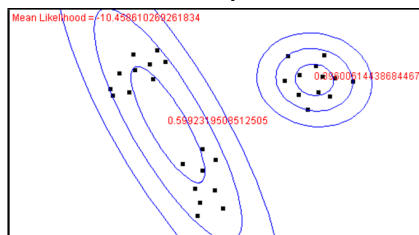
→ lorsque les données ne respectent pas les hypothèses, les résultats présentent une grande variabilité et ne sont pas pertinents

Illustration : nombre variable de composantes

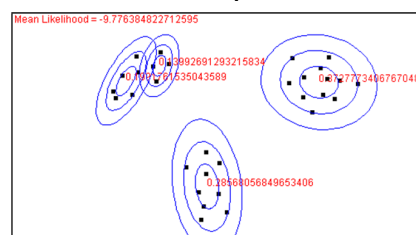
Avec 3 composantes :



Avec 2 composantes :



Avec 4 composantes :



Sélection de modèle en estimation de densités

- Complexité du modèle de mélange :
 - Nombre de composantes : chaque nouvelle composante ajoute comme paramètres son coefficient de mélange et les coefficients qui définissent sa densité
 - Augmentation du nombre de paramètres lors des passages : matrice covariances identité (« sphérique ») → diagonale → quelconque
- Augmentation du nombre de paramètres → en général, augmentation de la valeur maximale que peut atteindre la log-vraisemblance, $\ln p(\mathcal{D}_o|\hat{\theta}) \Rightarrow$ **la log-vraisemblance seule ne peut servir de critère de choix de la complexité!**
- Types de méthodes (voir [4]) :
 - Tests d'hypothèses : inspirés par LRTS (*likelihood ratio test statistic*) pour m composantes vs. $m + 1$ composantes
 - Critères « d'information » : ajout de pénalités (qui augmentent avec le nombre de paramètres libres) à la $-\log$ -vraisemblance
 - Critères de classification : évaluer la capacité du mélange à produire des groupes (1 groupe = données mieux expliquées par une des composantes) bien séparés
 - Autres méthodes : validation croisée, critères graphiques, etc.

Critère d'information d'Akaike

- Ce critère, introduit par Akaike en 1971, ajoute une pénalité qui dépend uniquement, et de façon linéaire, du nombre de paramètres libres :

$$\text{AIC} = -2 \ln p(\mathcal{D}_o|\hat{\theta}) + 2k \quad (8)$$

k : nombre de paramètres libres à estimer ($k \geq 0$)

- Comparaison de modèles : préférer le modèle qui **minimise** AIC
 - k augmente → augmentation de $\ln p(\mathcal{D}_o|\hat{\theta})$ → diminution de $-2 \ln p(\mathcal{D}_o|\hat{\theta})$
 - Ajout de $2k$ → pénalité croissant avec k
- Idée du critère : lorsque k diminue, le **biais** du modèle augmente (terme $-\ln p(\mathcal{D}_o|\hat{\theta})$); lorsque k augmente, la **variance** du modèle augmente (terme k) \Rightarrow additionner les deux pour trouver le meilleur compromis entre biais et variance

Critère d'information bayésien

- Ce critère, introduit par Schwartz en 1978 sur des considérations bayésiennes, tient compte également du nombre d'observations

$$\text{BIC} = -2 \ln p(\mathcal{D}_o | \hat{\theta}) + k \ln N \quad (9)$$

k : nombre de paramètres libres à estimer ($k \geq 0$)

N : nombre d'observations ($N \gg k$); la pénalité augmente avec le (logarithme du) nombre d'observations !

- Des définitions résulte que la pénalité BIC est plus forte que la pénalité AIC dès que $N \geq 8$ ($\ln N > 2$)
- Idée du critère : maximiser la probabilité *a posteriori* du modèle, en considérant une distribution *a priori* non informative sur les modèles et sur les paramètres de chaque modèle

BIC ou AIC ?

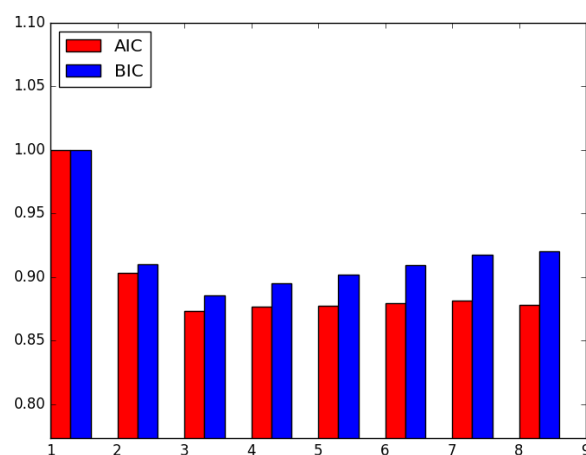


FIG. – AIC et BIC en fonction du nombre de composantes pour un ensemble \mathcal{D}_o

- La pénalité BIC augmente plus vite que la pénalité AIC \Rightarrow modèles en général plus simples (k plus faible) avec BIC (mais pas dans l'exemple ci-dessus)

Paramétrique ou non paramétrique ?

- Avantages des méthodes paramétriques :
 - Si les hypothèses sont valides, bonne estimation avec des échantillons de taille comparativement faible
 - Après construction du modèle, faible coût de l'estimation de la densité en un point précis

- Avantages des méthodes non paramétriques :
 - Généralité due à l'absence d'hypothèses sur le nombre et les types de lois
 - Convergence garantie vers la vraie densité. . . si l'échantillon est suffisant !
 - Paramètre unique. . . mais difficile à choisir

Références I



C. Archambeau, M. Valle, A. Assenza, and M. Verleysen.

Assessment of probability density estimation methods : Parzen window and finite gaussian mixtures.
In *ISCAS*. IEEE, 2006.



A. P. Dempster, N. M. Laird, and D. B. Rubin.

Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society : Series B, 39 :1–38, 1977.



M. C. Jones, J. S. Marron, and S. J. Sheather.

A brief survey of bandwidth selection for density estimation.
Journal of the American Statistical Association, 91(433) :401–407, 1996.



A. Oliveira-Brochado and F. V. Martins.

Assessing the Number of Components in Mixture Models : a Review.
FEP Working Papers 194, Universidade do Porto, Faculdade de Economia do Porto, Nov. 2005.