Systèmes Intelligents : Raisonnement et Reconnaissance

James L. Crowley

Deuxième Année ENSIMAG Deuxième Sémestre 2006/2007

Séance 10 25 avril 2007

Mélange de Gaussiennes et L'Algorithme EM (Expectation-Maximization)

Sources Bibliographiques :

"Neural Networks for Pattern Recognition", C. M. Bishop, Oxford Univ. Press, 1995.

"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

L'Analyse en Composantes Principales

L'analyse en composant principales est une méthodes de déterminer un sous-espace "optimale" pour la reconstruction. Il est souvent utilisé pour la discrimination dans la cas où le vecteur X est composé d'une grande nombre de caractéristiques.

Voici son application pour la reconnaissance d'images de visages.

Soit une ensemble de M imagettes, toutes classes confondus, $W_m(i, j)$, composé de N pixels :

$W_m(i, j)$ pour $m \in \{1, M\}$, tel que $i \in [1, I]$, $j \in [1, J]$, $I \times J = N$

On les exprime sous forme de vecteur :

Ensemble d'imagettes $X_m(n) = W_m(i, j)$ ou $n = j*I + i$.

On cherche une base orthogonale $\vec{\phi}(n) = \{\phi_d(n)\}$ $d = 1, 2, \ldots D$
pour représenter les X(n). telles que $D << M$.

Imagette moyenne $$\mu(n) = \frac{1}{M} \sum_{m=1}^{M} X_m(n)$$

Imagette moyenne zéro $V_m(n) = X_m(n) - \mu(n)$

Base orthogonal $\vec{\phi}(n) = \{ \phi_d(n) \}$ $d = 1, 2, \ldots D$
avec $D \leq N$.

Vecteur $\vec{\alpha} = \langle V(n), \vec{\phi}(n) \rangle$

Image reconstruite : $$\hat{X}(n) = \mu(n) + \sum_{n=1}^{D} \alpha_n \vec{\phi}(n)$$

Image de résidu : $R(n) = X(n) - \hat{X}(n)$

Energie de résidu : $$\varepsilon_r^2 = \sum_{n=1}^{D} R^2(n)$$

La covariance, **C**, est composée de N x N = N^2 termes

$$\mathbf{C} \equiv E\{ \vec{V}\ \vec{V}^T \} \equiv E\{[\vec{\mathbf{X}} - \vec{\mu})(\vec{\mathbf{X}} - \vec{\mu})^T\}$$

(pour une image de taille D= 2^N, il y a 2^{2N} pixels et 2^{4N} termes dans la covariance.)

$$\sigma_{ij}^2 = \frac{1}{M} \sum_{m=1}^{M} (V_m(i) \cdot V_m(j))$$

On peut faire une matrix V composé de M colones $\{\vec{V}_m\}$

$$V = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1M} \\ v_{21} & v_{22} & \cdots & v_{2M} \\ \cdots & \cdots & & \cdots \\ v_{N1} & v_{N2} & \cdots & v_{NM} \end{pmatrix}$$

$$\mathbf{C}_X \equiv V\ V^T = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

V est N par M. Chaque colonne de V est une image.
Chaque coefficient, σ_{ij}^2, est une covariance de la pixel i et j pour l'ensemble de M images

Pour une image de 32 x 32, le matrice C= V V^T est de taille 1024 x 1024.
Il y a un coefficient par pair de pixels. Chaque terme est la covariance d'une paire de pixels.

C = V V^T est de taille N x N

On cherche une ensemble orthogonales $\vec{\phi}(n) = \{\phi_d(n)$ d = 1, 2, ... N
tels que : $\phi^T\ V\ V^T\ \phi = \lambda_C$

λ_c est une matrice diagonale des valeurs principales de C.
Chaque colonne de ϕ est un vecteur directeur $\phi_d(n)$. Les colonnes ϕ_d sont orthogonales. Nous allons choisir une sous ensemble D des N colones de ϕ
Une telle matrice de rotation est fournie par une procedure d'analyse en composants principales.

$$(\phi_d(n),\ \lambda_c) \leftarrow \text{PCA}(C).$$

Pour une image de 32 x 32, le matrice $C= V V^T$ est de taille $2^{10} \times 2^{10} = 2^{20}$.
Pour une image de 512 x 512, le matrice $C= V V^T$ est de taille $2^{18} \times 2^{18} = 2^{36}$.

Heureusement, il y a une astuce pour éviter la matrice de covariance de N x N coefficients. Le rank de VV^T est M, ou M est le nombre d'images. M << N·

Noter que $B = V^T V$ est de taille M^2, M est le nombre d'images.
Chaque coefficient est une produit de deux images !

$$V^T V = B = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}$$

Les coefficients sont les covariances des images (et non pas les pixels!).

$$b_{ij}^2 = \sum_{n=1}^{N} V_i(n)\ V_j(n)$$

Pour M < 512 on peut facilement calculer une matrice R de rotation tels que chaque colonne est un vecteur directeur orthogonal.
Soit R les composantes principales de $\tilde{X}^T \tilde{X}$

$$\mathbf{R}^T V^T V\ \mathbf{R} = \lambda_b$$

On multiplie les deux cotés par R :

$$R\,R^T\,V^T V\,R = V^T V\,R = R\,\lambda_b$$

$R^T R = I$. Maintenant on multiplie par A

$$V\,V^T V\,R = V\,R\,\lambda_b$$
$$= (VV^T)(V\,R) = (V\,R)\,\lambda_b$$
$$= (VV^T)\,\phi = \phi\,\lambda_c$$

Donc, par inspection : $\phi = VR$.

$$\phi = V \quad R$$

$$\begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}$$

Raison : Les mêmes images ont été utilisé pour ϕ et pour R.
Les premiers M vecteurs propres de $V^T V$ sont aussi les premiers M vecteurs propres de VV^T ,le Rank de ϕ est le même que R, et λ_b sont les premiers M valeurs propre de λ_c et Donc, les vecteurs propres : $\phi_d(n) = V\,R$ triés par λ_b

$$\phi = VR$$

Chaque colonne est un vecteur dans une base orthogonale.

$$\hat{\phi}(n) = \sum_{m-1}^{M} V_m(n)\,R(m,n)$$

$\hat{\phi}(n)$ fournit une base "ortho-normal" pour $X_m(n)$.

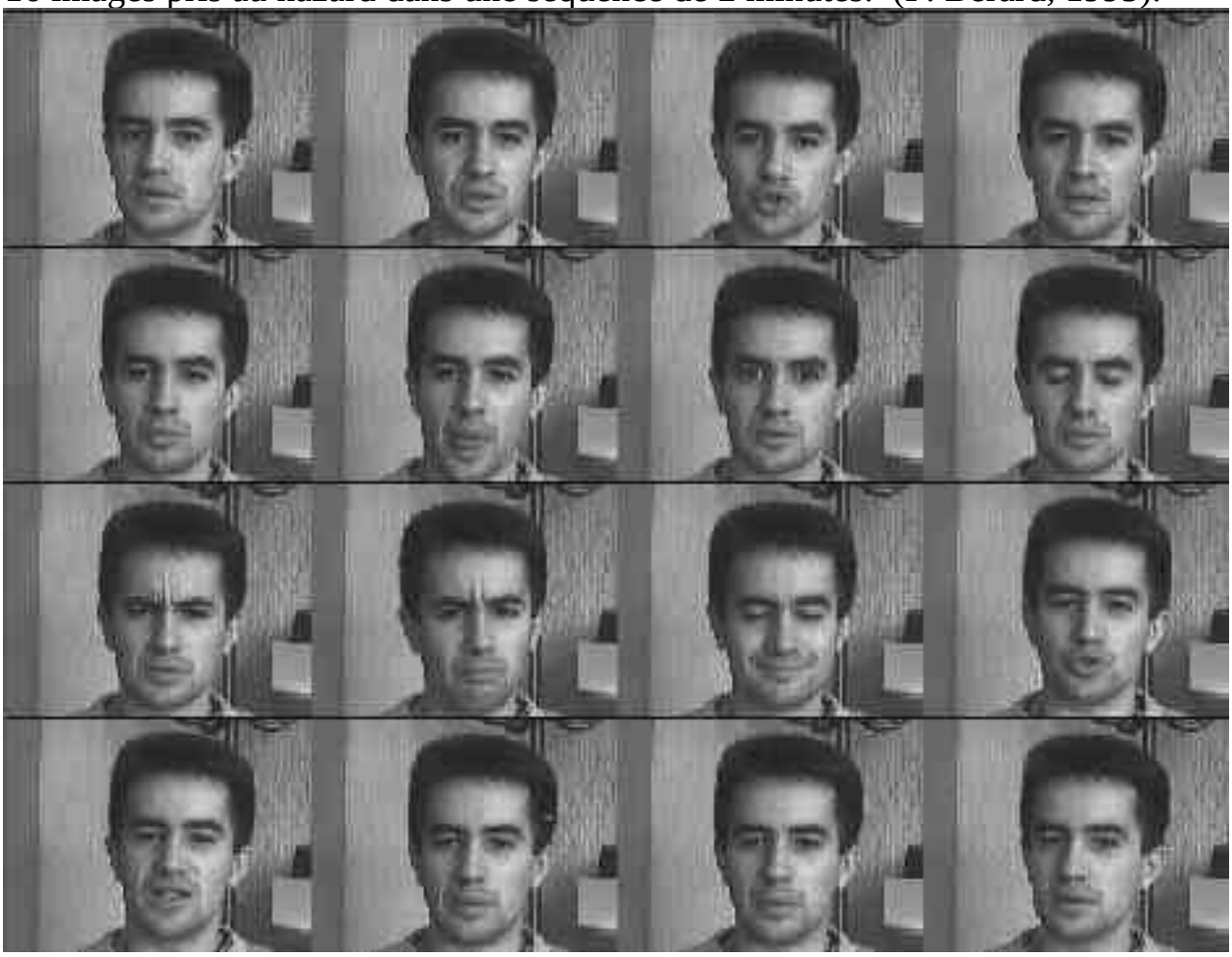
Une image (normalisée) peut être exprimé par

$$\alpha_d = <X(n),\phi_d(n)> = \sum_{d=1}^{M} X(n)\phi_d(n)$$

Les valeurs α_d sont un "code" qui représente X(n) pour la reconnaissance ou la transmission.

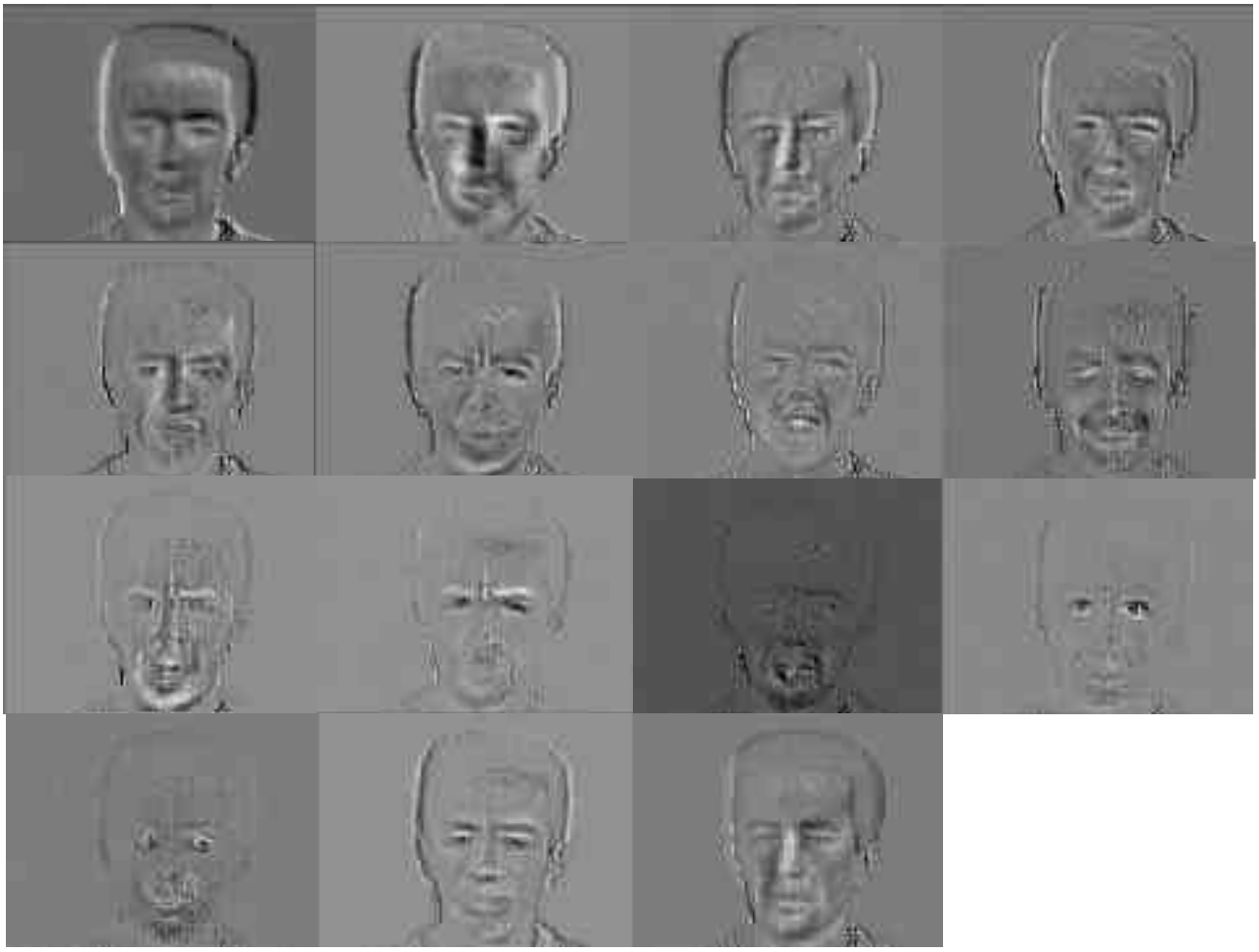
Exemple :

16 images pris au hazard dans une séquence de 2 minutes. (F. Bérard, 1995).

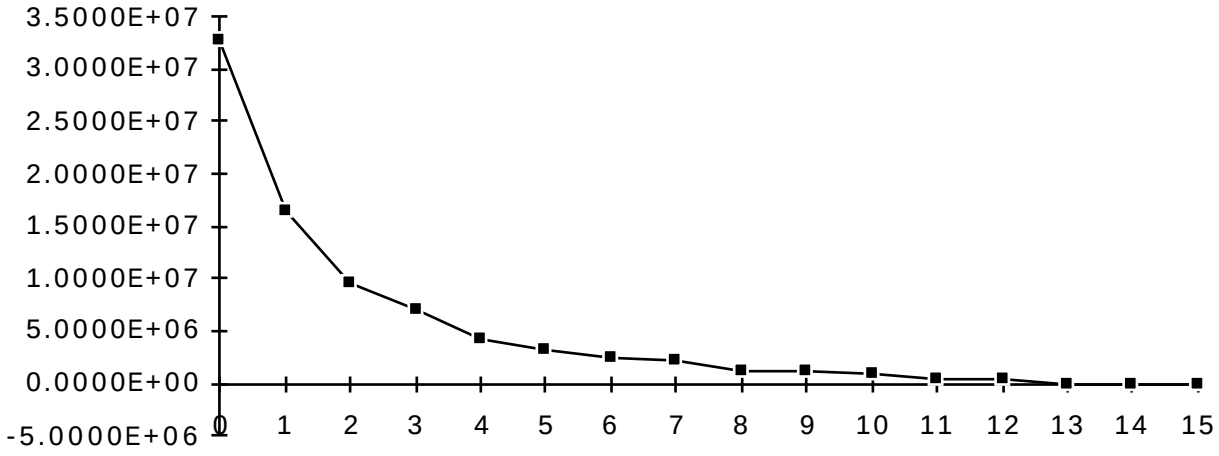


Average Image

Components Principales :



Eigen Values.



Reconstruction

Image Reconstructed image (120 bytes) Error Image.



Reconstruction (120 bytes)



Image Error

Reconnaissance de visages avec PCA

Dans l'espace PCA, pour chaque individu, on fait une ensemble de M_k images.

Ensemble d'imagettes de la classe ω_k : $X_{mk}(n) = W_{mk}(i, j)$ ou $n = j*32 + i$.

Imagette moyenne de la classe T_k : $\mu_k(n) = \frac{1}{M} \sum_{m=1}^{M} X_{mk}(n)$

Imagette moyenne zéro $V_{mk}(n) = X_{mk}(n) - \mu_k(n)$

Projection sur $\vec{\phi}$: $\vec{Z}_{mk} = \langle V_{mk}, \vec{\phi} \rangle$

La covariance, C_k, est composée de D_y x D_y termes

$$C_k = E\{ \vec{Z}_{mk} \ \vec{Z}_{mk}^T \}$$

(pour une image de taille 2^n, il y a 2^{2n} pixels et 2^{4n} termes dans la covariance.)

$$\sigma_{ij}^2 = \frac{1}{M} \sum_{m=1}^{M} (Z_{mk}(i) \ Z_{mk}(j))$$

ou bien :

$$\sigma_{ij}^2 = \frac{1}{M} \sum_{m=1}^{M} (Z_{mk}(i) - \mu_k(i))(Z_{mk}(j) - \mu_k(j))$$

$$\omega_k = \arg\max_k \{g_k(\vec{Z})\}$$

ou $g_k(\vec{Z}) = \vec{Z}^T (B_k) \vec{Z} + \vec{b}_k^{\ T} \vec{Z} + b_{ko}$.

avec $\mathbf{B}_k = \frac{1}{2} \mathbf{C}_k^{-1}$

$\vec{b}_k = C_k^{-1} \vec{\mu}_k$

$b_{ko} = -\frac{1}{2}(\vec{\mu}_k^{\ T} C_k^{-1} \vec{\mu}_k) - \frac{1}{2} \text{Log}\{\det(C_k)\} + \text{Log}\{p(\omega_k)\}$

Quelle sont les sources de bruit? $\det(B_o) >> \det(B_k)$ est dominé par

1) L'orientation du visage
2) Les ombres, et
3) Les expressions.

Ceci donne plusieurs densité Gaussiennes. Comment les estimer?

Mélange de Gaussiens.

Si les événements sont issus d'une composition de "N" phénomènes, la densité $p(\vec{X})$ prendra la forme d'une composition de lois Normales.

Dans un tel cas, on peut approximer par une somme pondérée de densités Normales.

$$p(x) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(x;\ \mu_n,\sigma_n)$$

Un tel somme est connu par la terme "mélange de Gaussiens"
pour chaque Gaussien il faut estimer trois paramètres :

$$\vec{\nu}_n = (\alpha_n,\ \mu_n,\sigma_n)$$

En total il y a $3 \cdot N$ paramètres à estimer.

$$\vec{\nu} = (\alpha_1,\ \mu_1,\sigma_1,\ \alpha_2,\ \mu_2,\sigma_2,\ \ldots\ ,\alpha_N,\ \mu_N,\sigma_N)$$

Si toutes les μ et σ etait fixe (et égale), on pouvait calculer les α_n directement.
Mais parce qu'ils sont libres il faut les estimer par un processus iterative.
Un tel processus est composé de deux étapes.

Ces étapes fournissent une estimation des variables cachées.

Pour un mélange de Gaussiens, les variables cachées sont les "sources" pour des événements.

On suppose chaque événement est issu d'un des N sources.
Nous allons construire une table de probabilités. h(m, n)

h(m, n) = Pr{l'événement E_m est issu de la source N}

les probabilités, h(m, n), nous donnera les facteurs de Mélange, α_n, ainsi que μ_n,σ_n.
L'algorithme d'estimation s'appelle "Expectation-Maximization" ou "EM".

Ebauche de l'Algorithme EM

Soit une ensemble ("training set") de M observations $\mathbf{T} = \{X_m\}$.
Fait une premier estimation des paramètres $\vec{\nu}^{(0)}$ et puis on alterne "E" et "M".

E: Faire une estimation des valeurs manquantes, h(m, n), pour les événements.

$h(m, n)^{(i)} = p(h_m | X_1, X_2, ..., X_M, \vec{\nu}^{(i)})$ pour chaque terme "n".

$$h(m, n)^{(i)} = \frac{\alpha_n^{(i)} \mathcal{N}(X_m; \mu_n^{(i)}, \sigma_n^{(i)})}{\sum_{j=1}^{N} \alpha_j^{(i)} \mathcal{N}(X_m; \mu_j^{(i)}, \sigma_j^{(i)})}$$

M: Recalculer $\vec{\nu}^{(i+1)}$ avec $p(h_m | X_1, X_2, ..., X_M, \vec{\nu}^{(i)})$

$$S_n^{(i+1)} = \sum_{m=1}^{M} h(m,n)^{(i)}$$

$$\alpha_n^{(i+1)} = \frac{1}{M} S_n^{(i+1)} = \frac{1}{M} \sum_{m=1}^{M} h(m,n)^{(i)}$$

$$\mu_n^{(i+1)} = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)^{(i)} X_m$$

$$\sigma^2_n{}^{(i+1)} = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)^{(i)} (X_m - \mu_n^{(i+1)})^2$$

Pour la dérivation l'algorithme EM, il faut introduire le concept de "likelihood" (vraisemblance) .

Maximisation de la Vraisemblance

Nous commençons par le cas ou D = 1. (une seule caractéristique)
Soit une ensemble de M exemples de caractéristiques $\{X_m\}$.
Les exemples sont supposés d'être les échantillons independents .

On souhaite estimer une forme paramétrique pour p(x).
Supposons que $p(x) = \mathcal{N}(x; \mu,\sigma)$.

Pour une loi normale avec D = 1 les paramètres sont

$$\vec{\nu} = (\mu,\sigma)$$

On souhait trouve une estimation $\hat{\vec{\nu}} = (\hat{\mu},\hat{\sigma})$ qui minimise la probabilité d'erreur.
Pour ceci on définit le Likelihood $L(\vec{\nu} \mid X_1, X_2, ..., X_M)$ de $\hat{\vec{\nu}}$ étant donnée $\{X_m\}$
Si les X_m sont indépendants,

$$p(X_1, X_2 \mid \vec{\nu}) = p(X_1 \mid \vec{\nu}) \cdot p(X_2 \mid \vec{\nu})$$

En générales pour M événements :

$$p(\{X_m\} \mid \vec{\nu}) = p(X_1, X_2, ..., X_M \mid \vec{\nu}) = \prod_{m=1}^{M} p(X_m \mid \vec{\nu})$$

Nous définissons le "Likelihood" (vraisemblance) de $\vec{\nu}$ étant $\{X_m\}$ comme

$$L(\vec{\nu} \mid \{X_m\}) = L(\vec{\nu} \mid X_1, X_2, ..., X_M) = p(X_1, X_2, ..., X_M \mid \vec{\nu})$$

$$= \prod_{m=1}^{M} p(X_m \mid \vec{\nu})$$

Notre objectif est d'estimer les paramètres $\hat{\vec{\nu}}$ pour maximiser $L(\vec{\nu} \mid S)$.

$$\hat{\vec{\nu}} = \arg\max_{\vec{\nu}} \{ L(\vec{\nu} \mid \{X_m\})\} = \arg\max_{\vec{\nu}} \{ \prod_{m=1}^{M} p(X_m \mid \vec{\nu})\}$$

Pour simplifier l'analyse, on travaille avec la log : $\ell(\vec{\nu}) = \mathrm{Log}\{L(\vec{\nu} \mid S)\}$.

$$\ell(\vec{\nu}) = \text{Log}\{L(\vec{\nu} \mid \{X_m\})\} = \text{Log}\{ p(\{X_m\} \mid \vec{\nu})\} = \sum_{m=1}^{M} \text{Log}\{p(X_m \mid \vec{\nu})\}$$

Si $p(X_m \mid \vec{\nu})$ est une simple loi normale, il suffit de trouver

$$\hat{\nu} = \arg\max_{\nu}\{ \prod_{m=1}^{M} p(X_m \mid \vec{\nu})\}$$

Ceci nous donne les formes habituelles des moments $\hat{\nu} = (\mu,\sigma)$

Maximum de vraisemblance pour le cas univariate

Soit un modèle normale $\mathcal{N}(\mu,\sigma)$. Pour estimer μ,σ :

$$\ell(\vec{\nu}) = \text{Log}\{p(X_m \mid \vec{\nu})\} = -\frac{1}{2} \text{Log}\{2\pi\sigma^2\} - \frac{1}{2\sigma^2}(X_m - \mu)^2$$

$$\frac{\partial \ell(\vec{\nu})}{\partial \mu} = \frac{1}{\sigma^2}(X_m - \mu)$$

$$\frac{\partial \ell(\vec{\nu})}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(X_m - \mu)^2}{2\sigma^4}$$

$$\nabla_{\mu,\sigma}\, \ell(\vec{\nu}) = \begin{pmatrix} \frac{1}{\sigma^2}(X_m - \mu) \\ -\frac{1}{2\sigma^2} + \frac{(X_m - \mu)^2}{2\sigma^4} \end{pmatrix}$$

La maximum se trouve ou le dérivé est nul.

$$\frac{\partial \ell(\vec{\nu})}{\partial \mu} = \sum_{m=1}^{M} \frac{1}{\sigma^2}(X_m - \hat{\mu}) = 0.$$

Avec un peu d'algèbre on a :

$$\sum_{m=1}^{M} \frac{1}{\sigma^2}(X_m - \hat{\mu}) = 0.$$

$$\frac{1}{\sigma^2}\sum_{m=1}^{M} X_m = \frac{1}{\sigma^2}\sum_{m=1}^{M} \hat{\mu}$$

$$\sum_{m=1}^{M} X_m = M\,\hat{\mu}$$

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M} X_m$$

et de la même façon pour

$$\frac{\partial \ell(\vec{\nu})}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(X_m - \mu)^2}{2\sigma^4} = 0$$

$$\sum_{m=1}^{M} -\frac{1}{2\sigma^2} + \frac{(X_m - \mu)^2}{2\sigma^4} = 0$$

$$\sum_{m=1}^{M} \frac{1}{2\sigma^2} = \sum_{m=1}^{M} \frac{(X_m - \mu)^2}{2\sigma^4}$$

$$\frac{1}{2\sigma^2}\sum_{m=1}^{M} 1 = \frac{1}{2\sigma^4}\sum_{m=1}^{M} (X_m - \mu)^2$$

$$M = \frac{1}{\sigma^2}\sum_{m=1}^{M} (X_m - \mu)^2 \qquad \hat{\sigma}^2 = \frac{1}{M}\sum_{m=1}^{M} (X_m - \hat{\mu})^2$$

Le cas multi-variate.

Pour $\vec{x}$ composé de D caractéristiques, avec M exemples d'une classe $\mathbf{T} = \{\vec{X}_m\}$ Issues un modèle normale $\mathcal{N}(\vec{\mu}, \mathbf{C})$.

Le problème est d'estimer les

$$\hat{\nu} = \max_{\nu}\{L(\vec{\nu} \mid \mathbf{T})\} = \max_{\nu}\{\prod_{m=1}^{M} p(\vec{X}_m \mid \vec{\nu})\}$$

Le maximum de $\vec{\nu}$ est trouvé quand le gradient est null.

$$\nabla_{\nu} \ell(\vec{\nu}) = \sum_{m=1}^{M} \nabla_{\nu} \mathrm{Log}\{p(\vec{X}_m \mid \vec{\nu}) = 0$$

où le gradient est $\nabla_{\nu} = \begin{pmatrix} \frac{\partial}{\partial \nu_1} \\ \dots \\ \frac{\partial}{\partial \nu_{2D}} \end{pmatrix}$

Le gradient nous permet de calculer une solution analytique.
Les estimations de $\hat{\mu}, \hat{\mathbf{C}}$ sont obtenus par

$$\mathrm{Log}\{p(\vec{X}_n \mid \vec{\nu})\} = -\frac{1}{2} \mathrm{Log}\{2\pi^n \det(\mathbf{C})\} - \frac{1}{2}(\vec{X}_n - \vec{\mu})^T \mathbf{C}^{-1}(\vec{X}_n - \vec{\mu})$$

$$\sum_{m=1}^{M} \nabla_{\mu,C} \mathrm{Log}\{p(\vec{X}_m \mid \vec{\nu})\} = 0$$

Donne la formule classique :

$$\hat{\mu} = \frac{1}{M} \sum_{m=1}^{M} \vec{X}_m \text{ et } \hat{\mathbf{C}} = \frac{1}{M} \sum_{m=1}^{M} (X_m - \hat{\mu})(X_m - \hat{\mu})^T$$

L'Algorithme EM (Expectation-Maximization)

L'algorithme EM s'applique à l'estimation de données cachés.
Il est utilisé notamment pour l'estimation des Modèles de Markov Cachées (HMM's) et pour l'estimation des Mélanges de Gaussiens.

Pour un mélange de Gaussiens, les variables cachées sont les sources des événements. On suppose que chaque événement est produit par un des N sources.
Pour chaque événement E_m On définit la variable "cachée" h_m

$h_m = N$ si l'événement E_m (avec caractéristique X_m) est issu de la source N.
Nous cherchons à estimer

$$p(x) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(x; \mu_n, \sigma_n) \quad \text{Tels que} \quad \sum_{n=1}^{N} \alpha_n = 1.$$

Il faut estimer N vecteurs : $\nu_n = (\alpha_n, \mu_n, \sigma_n)$
En total il y a $3 \cdot N$ paramètres à estimer.

$$\vec{\nu} = (\alpha_1, \mu_1, \sigma_1, \alpha_2, \mu_2, \sigma_2, \ldots, \alpha_N, \mu_N, \sigma_N)$$

$$\ell(\vec{\nu} \mid \{X_m\}) = \text{Log}\,(L(\vec{\nu} \mid X_1, X_2, \ldots, X_M)) = \text{Log}\ \{ \prod_{m=1}^{M} p(X_m|\vec{\nu})$$

$$= \sum_{m=1}^{M} \text{Log}\ \{p(X_m|\vec{\nu})\}$$

$$= \sum_{m=1}^{M} \text{Log}\left(\sum_{n=1}^{N} \alpha_n\, p(X_m \mid \nu_n) \right)$$

Il faut $\hat{\nu}_S = \max_{\nu} \{ \ell(\vec{\nu} \mid \{X_m\})\}$

Il n'y pas de solution analytique, parce qu'il n'y est pas une dérivé pour la logarithme de la somme.

Soit $\mathbf{T} = \{X_m\}$ h(m,n) le donnée Complèt.
On vas maximiser une mesure de qualité $Q(\vec{\nu}, \vec{\nu}^{(i)})$ définit par une esperence conditionnel :

$$Q(\vec{\nu}, \vec{\nu}^{(i)}) = E\{ \ell(\vec{\nu} \mid \mathbf{T}) \mid \{X_m\}, \vec{\nu}^{(i)}) \}$$

L'etape E

Pour chaque événement E_m avec caractéristique X_m,
On suppose qu'il manque l'information: h_m

$h_m = n$, la source de l'événement E_m.

On ne connaît pas h_m, mais on peut estimer les probabilités $P(h_m = n)$.
par une table h(m,n).

Pour chaque X_m, et son source caché h_m

$$p(h_m, X_m \mid \vec{\nu}) = p(h_m \mid X_m, \vec{\nu})\, p(X_m \mid \vec{\nu})$$

donc

$$p(h_m \mid X_m, \vec{\nu}) = \frac{p(h_m,\ X_m \mid \vec{\nu})}{p(X_m \mid \vec{\nu})}$$

où

$$p(h_m=n, X_m \mid \vec{\nu}) = \alpha_n \mathcal{N}(X_m;\ \mu_n, \sigma_n)$$

$$p(X_m \mid \vec{\nu}) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(x_m;\ \mu_n, \sigma_n)$$

Donc

$$p(h_m=n \mid X_1, X_2, \ldots, X_M,\ \vec{\nu}_j) = \frac{\alpha_n \mathcal{N}(X_m;\ \mu_n, \sigma_n)}{\sum_{j=1}^{N} \alpha_j \mathcal{N}(X_m;\ \mu_j, \sigma_j)}$$

Donc pour chaque itération (i) le premier étape E est :

$$h(m, n)^{(i)} = \frac{\alpha_n^{(i)} \mathcal{N}(X_m;\ \mu_n^{(i)}, \sigma_n^{(i)})}{\sum_{j=1}^{N} \alpha_j^{(i)} \mathcal{N}(X_m;\ \mu_j^{(i)}, \sigma_j^{(i)})}$$

L'etape M

Pour le deuxième étape, M, nous allons maximiser le "likelihood" $\text{Log}\{p(X_m \mid \vec{\nu})\}$
On défini la "Expected Complete Data Log Likelihood" pour **T**. = $\{X_m\}$ U h(m,n)

$$Q(\vec{\nu}, \vec{\nu}^{(i)}) = E\{ \ell(\vec{\nu} \mid \mathbf{T}) \mid \{X_m\}, \vec{\nu}^{(i)}) \}$$

$$Q(\vec{\nu}, \vec{\nu}^{(i)}) = E\{ \ell(\vec{\nu} \mid \{X_m\}, h(m, n)) \mid \{X_m\}, \vec{\nu}^{(i)}) \}$$

$$Q(\vec{\nu}, \vec{\nu}^{(i)}) = \sum_{m=1}^{M} \sum_{n=1}^{N} \text{Log}\{p(h_m=n, X_m \mid \vec{\nu})\}\, p(h_m=n \mid X_m, \vec{\nu}^{(i)})$$

Pour chaque cycle dans l'itération nous allons chercher :

$$\vec{\nu}^{(i+1)} = \underset{\nu}{\text{argmax}}\{Q(\vec{\nu} \mid \vec{\nu}^{(i)})\}$$

Ce maximum est donné par :

$$S_n^{(i+1)} = \sum_{m=1}^{M} p(h_m=n \mid X_m, \vec{\nu}^{(i)})) = \sum_{m=1}^{M} h(m, n)$$

$$\alpha_n^{(i+1)} = \frac{1}{M} \sum_{m=1}^{M} P(h_m=n \mid X_m, \vec{\nu}^{(i)}) = \frac{1}{M} S_n^{(i+1)}$$

$$\mu_n^{(i+1)} = \frac{1}{S_n^{(i)}} \sum_{m=1}^{M} p(h_m=n \mid X_m, \vec{\nu}^{(i)})\, X_m = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,.n)\, X_m$$

$$\sigma^2{}_n^{(i+1)} = \frac{1}{S_n^{(i)}} \sum_{m=1}^{M} p(h_m=n \mid X_m, \vec{\nu}^{(i)})\, (X_m - \mu_n^{(i+1)})^2.$$

$$= \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)\, (X_m - \mu_n^{(i+1)})^2$$

Avec notre table de probabilités h(m, n)

$$h(m, n) = P(h_m=n \mid X_m, \vec{\nu}^{(i)})$$

E (Expectation) :

$$h(m, n)^{(i)} := \frac{\alpha_n^{(i)} \mathcal{N}(X_m;\ \mu_n^{(i)}, \sigma_n^{(i)})}{\sum_{j=1}^{N} \alpha_j^{(i)} \mathcal{N}(X_m;\ \mu_j^{(i)}, \sigma_j^{(i)})}$$

M: (Maximisation)

$$S_n^{(i+1)} := \sum_{m=1}^{M} h(m, n)^{(i)}$$

$$\alpha_n^{(i+1)} := \frac{1}{M} S_n^{(i+1)}$$

$$\mu_n^{(i+1)} := \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,.n)^{(i)} X_m$$

$$\sigma^2{}_n^{(i+1)} := = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)^{(i)} (X_m - \mu_n^{(i+1)})^2$$

Dans le cas Multivariate (D> 1) la covariance C est composée de σ_{jk}^2:

$$\sigma^2{}_{jkn}^{(i+1)} := = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)^{(i)} (X_{jm} - \mu_{jn}^{(i+1)})(X_{km} - \mu_{kn}^{(i+1)})$$