Sommaire

Chapitre III : Résultats

Introduction Générale

Introduction générale

Introduction générale

La Reconnaissance De Formes (RDF) est la discipline dont le but est la classification d'objets en nombre de catégories ou classes. Selon l'application, ces objets peuvent êtres des images, des signaux ou tout autres mesures ayant besoin d'être classés.
La reconnaissance de formes possède une longue histoire mais avant 1960 elle était le résultat de recherches théoriques dans le domaine des statistiques.

Comme pour d'autres disciplines, l'avènement des calculateurs a augmenté la demande d'applications pratiques de reconnaissance de formes, qui en retour a imposé de nouvelles exigences pour les développements théoriques.
Comme notre société évolue depuis sa phase industrielle à sa phase postindustrielle, l'automatisation en production industrielle ainsi que le besoin de retrouver et de traiter l'information devient d'une importance capitale. Cette tendance a poussé la reconnaissance de formes à la pointe des applications et de la recherche en ingénierie.

La reconnaissance de formes est une partie intégrale des systèmes d'intelligence machine construits pour la prise de décision. La vision machine est un domaine où la reconnaissance de formes possède une très grande importance. La reconnaissance de caractères est un domaine important qui fait appel à la reconnaissance de formes. Le diagnostic assisté par ordinateur est une autre application importante de la reconnaissance de formes visant à assister les médecins à diagnostiquer et à prendre les bonnes décisions. La reconnaissance de la parole, elle aussi intègre la reconnaissance de forme par un traitement de la voix.

La reconnaissance des formes est utilisée dans de nombreux domaines d'applications, tel que :
Analyse de signaux sismiques,
Analyse d'électrocardiogrammes,
Reconnaissance de la parole,
Reconnaissance de l'écriture,
Analyse de documents,
Imagerie médicales (microscope, radiographie, RMN, ...),

Biométrie (reconnaissance d'empreintes, de faces, ...),
Vision par ordinateur (analyse de scènes 3D)
Imagerie satellitaires,
Applications militaires (observation, guidage, ...).

En général, l'opération de reconnaissance comprend deux étapes :

- Extraction de caractéristiques (attributs ou encore primitives), dont le but est de réduire la quantité de données,
- Classification, qui consiste à associer une description symbolique à l'objet, sur la base de ses caractéristiques.

Les caractéristiques sont choisies de manière à ce qu'elles soient semblables pour les formes d'une même classe et dissemblables pour des formes de classes différentes.
Ce mémoire s'inscrit dans la perspective de la partie classification où la théorie des probabilités prend une place prépondérante.
A ce stade, le premier chapitre présentera quelques définitions de base rencontrées dans la classification, dont les caractéristiques et les vecteurs de caractéristiques. Nous parlerons aussi la distance minimale euclidienne et celle de Mahalanobis et de la décision, et sur quelles bases elle pourra être prise afin de classer optimalement un quelconque objet.

Le chapitre deux fait le point sur les types de classificateurs ainsi que les techniques d'estimation des fonctions de densité de probabilité inconnues. Nous allons nous pencher particulièrement sur la classification bayésienne. Nous allons aussi parler de l'algorithme Espérance-Maximisation (en anglais Expectation-Maximisation algorithm, souvent abrégé par EM) qui permet de retrouver, par le maximum de vraisemblance, les paramètres de modèle probabiliste, lorsque le modèle dépend de variables latentes non observables.

Le troisième chapitre sera consacré aux résultats des simulations réalisées sur la base de l'algorithme EM. En effet nous allons visionner la convergence des paramètres optimaux, à savoir : les statistiques de premier ordre, les statistiques de deuxième ordre

et les probabilités a priori ou poids. Nous allons aussi comparer les valeurs estimées avec celle utilisées pour la création des données elles-mêmes.

Nous terminerons par une conclusion générale qui résumera le travail accompli ainsi que certaines perspectives qui seront à la base de travaux futures. Ceux-ci viendront enrichir notre connaissance ainsi que celle du lecteur quant à la théorie de la classification et celle de la décision.

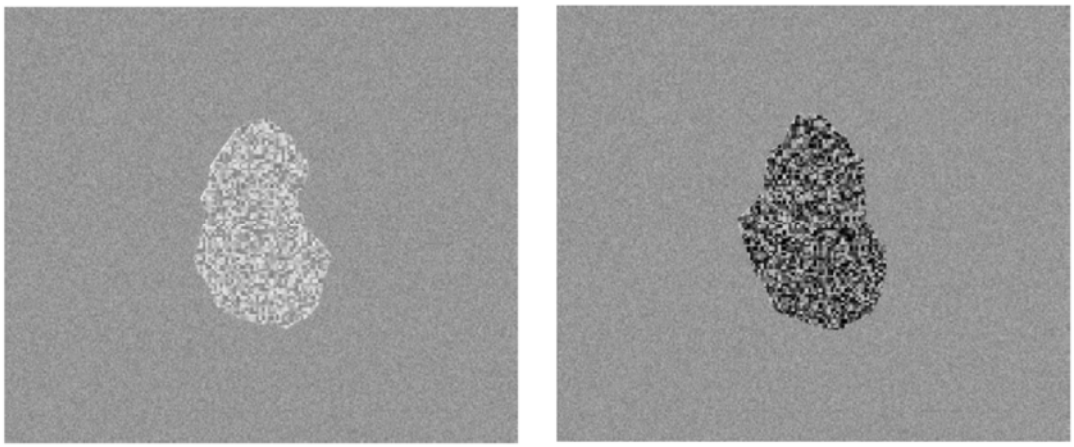
Chapitre I : Définitions de Base pour la Classification

I. 1. Introduction

Ce premier chapitre traite des notions de base qui entrent dans la conception d'un classificateur dans un système de reconnaissance de formes (RDF). L'approche à suivre est fondée sur des arguments probabilistiques découlant de la nature statistique des caractéristiques (ou attributs) générées. Cette nature statistique est due à la variation statistique des formes ainsi que le bruit dans les capteurs de mesure. Ce raisonnement est adopté comme point de départ pour la conception des classificateurs qui classent une forme inconnue dans la plus probable des classes. Ainsi nous définirons la signification de la notion du : "plus probable". Nous allons parler des caractéristiques et des vecteurs de caractéristiques ainsi que de la théorie de décision de Bayes. Celle-ci joue un rôle très important dans la classification des formes par décision fondée sur la notion du plus probable ou vraisemblable.

I. 2. Caractéristiques, vecteurs caractéristiques et classificateurs

Nous allons dans un premier temps simuler un cas simplifié imitant une tâche de classification d'image médicale. La figure I. 1 montre deux images, chacune ayant une région distincte à l'intérieur. Les images elles-mêmes sont visuellement différentes. Nous pouvons dire que la région de la figure I. 1. a, résulte d'une lésion bénigne (classe A) et que celle de la figure I. 1. b, d'une lésion maligne (classe B). De plus nous allons supposer que ce ne sont pas les seules formes (images dans ce cas) disponibles, mais que nous avons aussi accès à une base de données imagerie avec un nombre de formes dont certaines sont originaires de la classe A et d'autres de la classe B.



(a) (b)

Figure I. 1. Exemple de régions d'image correspondant
à la classe A (a) et à la classe B (b)

La première étape est d'identifier les quantités mesurables qui rendent ces deux régions distinctes [1]-[4]. La figure I. 2 montre un tracé de la valeur moyenne de l'intensité dans chaque région en fonction de la déviation standard correspondante autours de cette moyenne. Chaque point correspond à une image différente de la base de données disponible. Il en découle que les formes de la classe **A** tendent à se répartir en une surface différente de celles de la classe **B**. La ligne droite parait être un bon candidat pour séparer les deux classes.

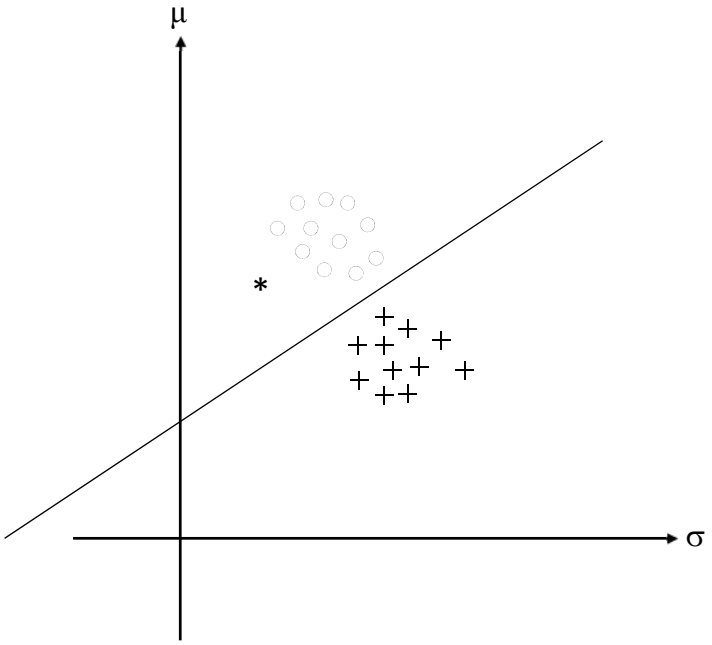


Figure I. 2. Tracé de la valeur moyenne en fonction de la déviation standard pour un nombre de différente images originaires de la classe A (○) et de la classe B (+). Dans ce cas, la ligne droite sépare les deux classes.

Supposons maintenant que nous avons une nouvelle image avec une région à l'intérieur, que nous ne savons pas à quelle classe elle appartient. Il est raisonnable de dire que nous mesurons la moyenne de l'intensité et la déviation standard dans la région qui nous intéresse et que nous tracions le point correspondant. Ceci est montré par l'astérisque (∗) sur la figure I. 2. Après, il serait sensé de supposer que la forme inconnue appartienne vraisemblablement à la classe A plutôt qu'à la classe B.

La tâche de classification artificielle précédente expose brièvement les raisons d'une grande classe de problèmes de reconnaissance de formes. Les mesures utilisées pour la classification, la valeur moyenne et la déviation standard dans ce cas, sont connues comme caractéristiques [1]-[4]. Pour le cas plus général, **L** caractéristiques $\mathbf{x_i}$, i = 1, 2, . . ., L, sont utilisées et elles forment le vecteur caractéristique : $\mathbf{x = [x_1, x_2, \ldots, x_L]^T}$ où T dénotes la transposition. Chacun des vecteurs caractéristiques identifie uniquement une seule forme (objet). Les caractéristiques et les vecteurs caractéristiques seront traités comme variables et vecteurs aléatoires, respectivement. Ceci est naturel puisque les mesures résultantes des formes différentes présentent une

variation aléatoire. Cette variation aléatoire est due en partie au bruit de mesure des équipements et en partie aux caractéristiques distinctes de chaque forme [1]-[4].
Par exemple, en imagerie rayons X, de grandes variations sont prévues à cause des différences en physiologie entre les individus. C'est la raison de dispersion des points dans chaque classe montrée sur la figure I. 1.
La ligne droite sur la figure I. 2 est connue comme ligne de décision, et constitue le classificateur dont le rôle est de diviser l'espace de caractéristiques en régions qui correspondent soit à la classe A ou à la classe B. Si un vecteur caractéristique **x**, correspondant à une forme inconnue, tombe dans la région de la classe A, il sera classé comme classe A, sinon comme classe B. Ceci ne signifie pas nécessairement que la décision est correcte [1]-[4]. Si elle n'est pas correcte, une erreur de classification s'est produite. Afin de tracer la ligne droite sur la figure I. 2, nous exploitons le fait que nous connaissons les labels ou étiquettes (classe A ou classe B) pour chaque point sur la figure. Les formes (vecteurs caractéristiques ou de caractéristiques) dont la vraie classe est connue et qui sont utilisés pour la conception du classificateur sont connue comme formes d'apprentissage (vecteurs caractéristiques d'apprentissage).
Ayant donné les grandes lignes des définitions ainsi que les raisons, nous signalons les questions de base qui se posent dans une tâche de classification, à savoir :

- Comment sont générées les caractéristiques ? Dans l'exemple précédent, nous avons utilisé la moyenne et la déviation standard, parce que nous savons comment les images ont été générées. Dans la pratique, ceci est loin d'être évident. Ceci dépend du problème, et concerne l'étape de génération de caractéristiques dans la conception du système de classification qui exécute une tâche de reconnaissance de forme donnée.

- Quel est le meilleur nombre **L** de caractéristiques à utiliser ? Ceci est aussi une tâche très importante et concerne l'étape de sélection de caractéristiques dans le système de classification. Dans la pratique, des nombres plus grands que le nombre nécessaire de caractéristiques sont générés, après quoi le meilleur d'entre eux est adopté.

- Ayant adopté, pour une tâche spécifique, les caractéristiques appropriées, comment peut-on concevoir le classificateur ? Dans l'exemple précédent la ligne droite a été tracée empiriquement juste pour satisfaire la curiosité du lecteur. Dans la pratique, ceci ne peut être le cas, et la ligne doit être tracée optimalement, c'est-à-dire, relativement à un critère d'optimalité. De plus, les problèmes pour lesquels un classificateur linéaire (ligne droite ou hyperplan dans l'espace à L-dimensions) peut donner une performance acceptable, ne sont pas la règle. En général, les surfaces divisant l'espace dans les différentes régions de classe sont non-linéaires. Quel type de non-linéarité doit-on adopter, et quel type de critère d'optimisation doit-être utiliser afin de situer une surface au bon endroit dans l'espace caractéristique à L-dimensions ? Ces questions concernent l'étape de conception du classificateur.

- Finalement, une fois le classificateur conçue, comment peut-on évaluer sa performance ? C'est-à-dire, quel est le taux d'erreur de la classification ? Ceci est la tâche de l'étape d'évaluation du système.

La figure I. 3. Montre les différentes étapes suivies pour la conception d'un système de classification. Comme il est évident sur les flèches de rétroaction, ces étapes ne sont pas indépendantes. Au contraire, elles sont inter-liées et dépendent des résultats, et l'on doit reconcevoir les étapes précédentes afin d'améliorer la performance globale.

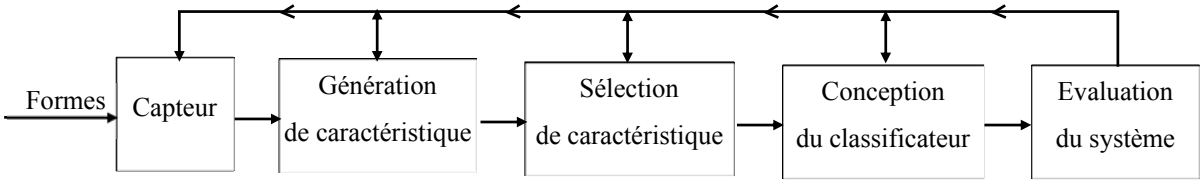


Figure I. 3. Les étapes de base incluses
dans la conception du système de classification

I. 3. Théorie de décision de Bayes

Supposons que nous avons une tâche de classification pour **K** classes, ω_1, ω_2, . . ., ω_K, et une forme inconnue qui est représentée par un vecteur de caractéristiques **x**. Nous créons **K** probabilités conditionnelles $\mathbf{P(\omega_i \mid x)}$, i = 1, 2, . . . , K, appelées parfois probabilités a posteriori. Chacune d'entre-elles représente la probabilité qu'une forme appartienne à la classe respective ω_i sachant que le vecteur de caractéristique correspondant prends la valeur **x** [1]-[4].

En effet, les classificateurs à considérer, calculent le maximum pour les **K** valeurs. La forme inconnue sera assignée à la classe correspondant à ce maximum.

Nous allons initialement nous concentrer sur le cas de deux classes. Soit ω_1, ω_2 les deux classes auxquelles nos formes appartiennent. En conséquence, nous supposons que les probabilités a priori $\mathbf{P(\omega_1)}$, $\mathbf{P(\omega_1)}$ sont connues. C'est une hypothèse très raisonnable puisque même si elles ne sont pas connues, elles peuvent être facilement estimées depuis les vecteurs de caractéristiques d'apprentissage disponibles. En effet, si **N** est le nombre total de formes d'apprentissage disponibles et $\mathbf{N_1}$, $\mathbf{N_2}$ sont ceux de celles appartenant aux classes ω_1 et ω_2 respectivement, alors :

$\mathbf{P(\omega_1) \approx N_1/N}$ et $\mathbf{P(\omega_2) \approx N_2/N}$.

Les autres quantités statistiques supposées connues sont les fonctions de densité de probabilité conditionnelle $\mathbf{p(x|\omega_i)}$, i = 1, 2, décrivant la distribution des vecteurs de caractéristiques dans chacune des classes. Si elles ne sont pas connues, elles aussi peuvent être estimées depuis les données d'apprentissage disponibles.

La PDF $p(x|\omega_i)$ est parfois référencée comme la fonction de vraisemblance de ω_i par rapport à x [1]-[6]. Ici nous soulignons le fait qu'une hypothèse implicite a été faite. C'est-à-dire que les vecteurs de caractéristiques peuvent prendre n'importe quelle valeur dans l'espace de caractéristiques à L-dimensions. Dans le cas où les vecteurs de caractéristiques ne peuvent prendre que des valeurs discrètes, alors les fonctions $p(x|\omega_i)$ deviennent des probabilités et sont notées par $\mathbf{P(x|\omega_i)}$.

Maintenant, rappelons la règle de Bayes :

$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i)\, P(\omega_i)}{p(x)} \qquad (I.1)$$

Où p(x) est la PDF de x et pour laquelle nous avons :

$$p(x) = \sum_{i=1}^{i=2} p(x \mid \omega_i)\, P(\omega_i) \qquad (I.2)$$

La règle de classification de Bayes peut être énoncée par :

Si $P(\omega_1 \mid x) > P(\omega_2 \mid x)$, x est classé dans ω_1

Si $P(\omega_1 \mid x) < P(\omega_2 \mid x)$, x est classé dans ω_2

Le cas d'égalité est nuisible et la forme peut être assignée à l'une ou à l'autre des deux classes. En utilisant la règle de Bayes, la décision peut également être basée sur les inégalités :

$$p(x|\omega_1)P(\omega_1) \lessgtr p(x|\omega_2)P(\omega_2)$$

p(x) n'est pas prise en compte, puisqu'elle est la même pour toutes les classes et n'affecte pas la décision. De plus, si les probabilités a priori sont égales, c'est à dire, $P(\omega_1) = P(\omega_2) = 1/2$, alors on aura :

$$p(x|\omega_1) \lessgtr p(x|\omega_2)$$

Ainsi, la recherche du maximum réside dans les valeurs des PDF conditionnelles évaluées à x [1]-[6]. La figure I. 4. présente un exemple de deux classes équiprobables et montre les variations de $p(x|\omega_i)$, i = 1, 2, en fonction de x pour le cas simple d'une seule caractéristique (l = 1). La ligne pointillée à x_0 est un seuil qui divise l'espace de caractéristiques en deux régions, R_1 et R_2. Selon la règle de décision de Bayes, pour toutes les valeurs de x dans R_1, le classificateur décide ω_1 et pour toutes les valeurs dans R_2, il décide ω_2. Cependant, il est clair d'après la figure que des erreurs de décision sont inévitables. En effet, il y a une probabilité pour un x de se trouver dans la région R_2 et au même moment d'appartenir à la classe ω_1. Ainsi, la décision est en erreur. Ceci reste vrai pour les points émanant de la classe ω_2.

La probabilité totale $\mathbf{P_e}$, de commettre une erreur de décision pour le cas de deux classes équiprobables est donnée par :

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)\,dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega_1)\,dx \qquad (I.3)$$

Qui est égale à la surface hachurée totale sous les courbes dans la figure I. 4.

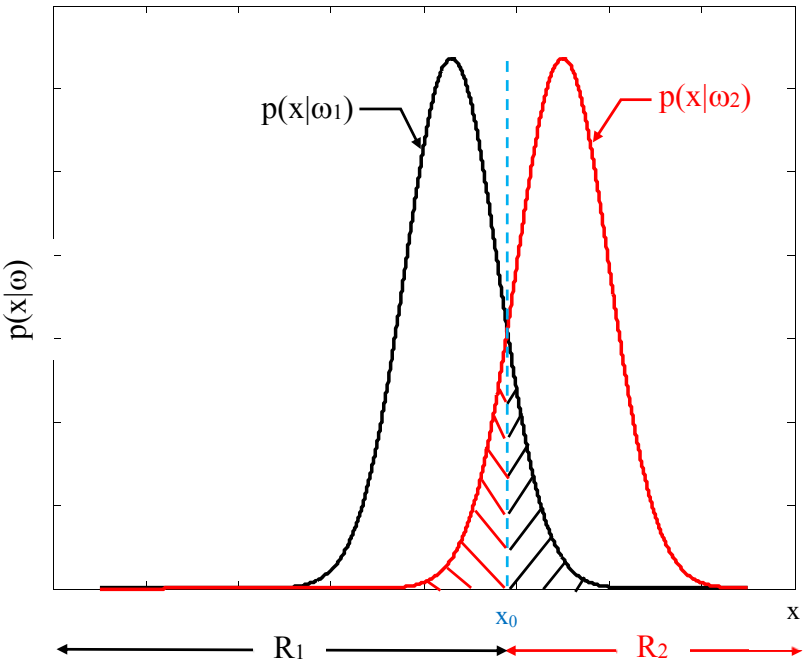


Figure I. 4. Exemple des deux régions R_1 et R_2 formées
par le classificateur de Bayes pour le cas de deux classes équiprobables

I. 4. Classification de Bayes pour distribution gaussiennes ou normales

I. 4. 1. Fonction de densité de probabilité de Gauss

L'une des densités de probabilité les plus rencontrées dans la pratique est la densité de probabilité de Gauss. La raison principale est sa simplicité (statistique d'ordre un et d'ordre deux seulement) et qu'elle modèle parfaitement un bon nombre de cas réels.

L'un des théorèmes les plus célèbre en statistique est le théorème de la limité centrale. Ce théorème stipule que si une variable aléatoire est le résultat de la sommation d'un nombre de variables aléatoires indépendantes, alors sa PDF approche une gaussienne au fur et à mesure que le nombre d'opérandes tend vers l'infinie [1]-[6].

En pratique, il est plus commun d'assumer que la somme de variables aléatoires suit une distribution gaussienne pour un nombre assez grand et suffisant d'opérandes.

Une gaussienne unidimensionnelle ou uni-variable est définie par :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (I.4)$$

Les paramètres **μ** et **σ^2** possèdent une signification spécifique. La valeur moyenne d'une variable aléatoire **x** est égale à **μ**, ceci dit :

$$\mu = E[x] = \int_{-\infty}^{+\infty} x\, p(x)\, dx \qquad (I.5)$$

Où :

E [·] : dénotes la valeur moyenne ou espérance d'une variable aléatoire.

σ^2 : est la variance de x, ceci dit :

$$\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2\, p(x)\, dx \qquad (I.6)$$

La généralisation multi-variables de la PDF gaussienne dans un espace L-dimensions est donnée par :

$$p(x) = \frac{1}{(2\pi)^{\frac{L}{2}}\, |\sum|^{\frac{1}{2}}}\, e^{\left(-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu)\right)} \qquad (I.7)$$

Où :

μ = E[x] est la valeur moyenne et **Σ** est la matrice de covariance d'ordre (**L x L**), définie par :

$\sum = E[(x-\mu)\,(x-\mu)^T]$

|Σ| est le déterminant de **Σ**.

I. 4. 2. Classificateur de Bayes pour classes à distribution normal

Notre but dans cette partie est d'étudier le classificateur de Bayes optimal quand les PDF impliquées, $p(x|\omega_i)$, i = 1, 2, . . . , K (fonctions de vraisemblance de ω_i en fonction de x), décrivant la distribution des données dans chacune des classes, sont des distributions normales multi-variables. C'est à dire **$\mathcal{N}(\mu_i, \sum_i)$**, i = 1, 2, . . ., K.

A cause de la forme exponentielle des densités impliquées, il est préférable de travailler avec des fonctions discriminantes de type logarithmique et monotone. Celles-ci sont données par [1]-[6] :

$$g_i(x) = \ln\big(p(x|\omega_i)P(\omega_i)\big) = \ln\Big(p(x|\omega_i) + \ln\big(P(\omega_i)\big)\Big)$$

Ou

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i) + \ln\big(P(\omega_i)\big) + c_i \qquad (I.8)$$

Où c_i est une constante égale à :

$$c_i = -\frac{l}{2}\ln(2\pi) - \frac{1}{2}\ln(|\sum_i|) \qquad (I.9)$$

En développant, nous obtenons :

$$g_i(x) = -\frac{1}{2}x^T\sum_i^{-1}x + \frac{1}{2}x^T\sum_i^{-1}\mu_i - \frac{1}{2}\mu_i^T\sum_i^{-1}\mu_i + \frac{1}{2}\mu_i^T\sum_i^{-1}x + \ln\big(P(\omega_i)\big) + c_i$$

En général, celle-ci est une forme quadratique non-linéaire. Prenons par exemple le cas de L = 2 et supposons que :

$$\sum_i = \begin{bmatrix}\sigma_i^2 & 0\\ 0 & \sigma_i^2\end{bmatrix}$$

Alors l'équation deviendra :

$$g_i(x) = -\frac{1}{2\sigma_i^2}(x_1^2 + x_2^2) + \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln\big(P(\omega_i)\big) + c_i$$

I. 4. 3. Classificateur à distance minimale

Si nous supposons des classes équiprobables avec la même matrice de covariance, nous aurons :

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum^{-1} (x-\mu_i)$$

Si $\Sigma = \sigma^2 I$, c'est à dire une matrice diagonale, alors dans ce cas une $g_i(x)$ maximale implique une distance Euclidienne minimale :

$$d_e = \|x - \mu_i\|$$

Ainsi, les vecteurs de caractéristiques seront assignés aux classes selon leurs distances Euclidiennes sur base de leurs moyennes.

Si $\Sigma \neq \sigma^2 I$, c'est à dire une matrice non-diagonale, alors dans ce cas la maximisation de $g_i(x)$ implique la minimisation de la norme de Σ^{-1}, connue sous le nom de distance de Mahalanobis :

$$d_m = \left((x - \mu_i)^T \sum^{-1} (x - \mu_i)\right)^{\frac{1}{2}}$$

Exemple

Dans une tâche de classification 2-D à deux classes, les vecteurs de caractéristiques sont générés par deux distributions normales partageant la même matrice de covariance :

$$\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$$

Et les vecteurs de moyennes sont $\mu_1 = [0, 0]^T$, $\mu_2 = [3, 3]^T$, respectivement.
Nous aurons la tâche de classer le vecteur $[1.0, 2.2]^T$ selon le classificateur de Bayes. Pour cela, il suffit de calculer la distance de Mahalanobis de $[1.0, 2.2]^T$ à partir des deux vecteurs de moyennes. Ainsi :

$$d_m^2(\mu_1, x) = (x - \mu_1)^T \sum^{-1} (x - \mu_1) = [1.0 - 0, 2.2 - 0] \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}^{-1} \begin{bmatrix} 1.0 - 0 \\ 2.2 - 0 \end{bmatrix}$$

$$d_m^2(\mu_1, x) = [1.0, 2.2] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952$$

De même :

$$d_m^2(\mu_2, x) = [1.0 - 3, 2.2 - 3] \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}^{-1} \begin{bmatrix} 1.0 - 3 \\ 2.2 - 3 \end{bmatrix}$$

$$d_m^2(\mu_2, x) = [-2.0, -0.8] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$$

Ainsi, le vecteur est assigné à la classe dont le vecteur de moyennes est $\mu_1 = [0, 0]^T$. Notons que le vecteur donné $[1.0, 2.2]^T$ est plus proche de $\mu_2 = [3, 3]^T$ selon la distance Euclidienne !

I. 5. Conclusion

En général, l'opération de reconnaissance comprend deux étapes :

Etape d'extraction de caractéristiques (ou primitives), dont le but est de réduire la quantité de données.

Etape de classification qui consiste à associer une description symbolique à l'objet, sur la base de ses caractéristiques.

Les caractéristiques sont choisies de manière à ce qu'elles soient semblables pour les formes d'une même classe et dissemblables pour des formes de classes différentes.

Au cours de ce chapitre, quelques notions de bases pour la classification ont bien été exposées. Nous pouvons conclure ce chapitre par cette distinction :

- L'espace d'observation (ou espace des données). Il contient beaucoup d'information redondante. L'information originale (image numérique, signal, etc.), fait partie de cet espace.
- L'espace de représentation (ou espace des caractéristiques). Il permet de représenter les caractéristiques. Il contient l'information jugée pertinente.
- L'espace d'interprétation (ou espace des classes ou catégories). Il est le domaine qui permet de représenter le résultat de la reconnaissance de formes. Il est en général très petit, souvent un ensemble fini d'étiquettes ou labels.

Maintenant, nous allons poursuivre avec le chapitre deux, où nous allons exposer les techniques d'estimation des paramètres optimaux pour fonctions de densité de probabilité inconnues.

Chapitre II : Estimation des Fonctions de Densité de Probabilité Inconnues

II. 1. Introduction

Le chapitre deux fait le point sur la classification de Bayes et les techniques pour l'estimation des fonctions de densité de probabilité inconnues.

Dans plusieurs applications, on suppose que les fonctions de densité de probabilité sont connues, or ceci n'est pas le cas en général.

Dans plusieurs problèmes, les PDF doivent êtres estimées à partir de données disponibles. Il existe différentes approches pour résoudre le problème. Parfois, nous connaissons la PDF (e. g. Gauss, Rayleigh), mais nous ne connaissons pas certains paramètres comme la moyenne et la variance. Par contre, dans d'autres cas, nous ne connaissons pas la nature de la PDF, mais nous connaissons les statistiques d'ordre un (moyenne) et d'ordre deux (variance). Selon, les informations disponibles, différentes approches peuvent être adoptées. Nous allons ainsi, explorer un bon nombre de méthodes d'estimation des fonctions de densité de probabilité, étape incontournable pour la classification.

II. 2. Estimation de paramètres par le maximum de vraisemblance

Considérons un problème à **K** classes avec les vecteurs de caractéristiques distribués selon $p(x|\omega_i)$, i = 1, 2, . . ., K. Nous supposons que ces fonctions de vraisemblance sont données sous forme paramétrique et que les paramètres correspondants sont données par $\boldsymbol{\theta_i}$ qui est l'inconnu.

Notre but est d'estimer les paramètres inconnus en utilisant un ensemble de vecteurs caractéristiques dans chaque classe. Si en plus, nous supposons que les données depuis une classe n'affectent pas l'estimation des paramètres des autres classes, alors nous pouvons formuler le problème indépendamment des classes. Enfin, nous pourrons ainsi, résoudre le problème pour chaque classe indépendamment [7].

Posons $\mathbf{x_1}$, $\mathbf{x_2}$, . . ., $\mathbf{x_N}$; comme étant des échantillons aléatoires obtenues depuis la PDF $p(x; \theta)$. Nous formons la PDF associée $p(X; \theta)$, où $X = \{x_1, \ldots, x_N\}$ est l'ensemble des échantillons. En supposant, une indépendance statistique entre les différents échantillons, nous aurons :

$$p(X;\theta) = p(x_1, x_2, \ldots, x_N;\ \theta) = \prod_{k=1}^{k=N} p(x_k;\ \theta) \qquad \text{(II. 1)}$$

C'est une fonction de $\boldsymbol{\theta}$ et connue sous le nom de fonction de vraisemblance de θ relative à **X**. La méthode du maximum de vraisemblance (ML) estimes θ, de sorte que la fonction de vraisemblance prend son maximum, c'est-à-dire :

$$\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{k=1}^{k=N} p(x_k;\ \theta) \qquad (II.2)$$

Une condition nécessaire que $\hat{\theta}_{ML}$ doit satisfaire afin qu'elle soit maximale est que le gradient de la fonction de vraisemblance relative à θ soit égale à zéro, c'est-à-dire :

$$\frac{\partial \prod_{k=1}^{k=N} p(x_k;\ \theta)}{\partial \theta} = 0$$

A cause de la monotonie de la fonction logarithmique, nous définissons la fonction de vraisemblance par :

$$L(\theta) = \ln \prod_{k=1}^{k=N} p(x_k;\ \theta) \qquad (II.3)$$

Et ceci est équivalent à :

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{k=1}^{k=N} \frac{\partial \ln p(x_k;\ \theta)}{\partial \theta} = \sum_{k=1}^{k=N} \frac{1}{p(x_k;\ \theta)} \frac{\partial\, p(x_k;\ \theta)}{\partial \theta} = 0$$

L'estimation du maximum de vraisemblance possède certaines propriétés intéressantes. Si $\boldsymbol{\theta_0}$ est la valeur exacte du paramètre inconnu dans $p(x;\ \theta)$, il est démontré, sous certaines conditions généralement vérifiables, que :

L'estimé par le ML est asymptotiquement non biaisé, qui par définition signifie que :

$$\lim_{N\to\infty} E\left[\hat{\theta}_{ML}\right] = \theta_0$$

L'estimé par le ML est asymptotiquement cohérent, c'est-à-dire qu'il satisfait :

$$\lim_{N\to\infty} \text{prob}\left\{\left\|\hat{\theta}_{ML} - \theta_0\right\| \leq \varepsilon\right\} = 1$$

Où ε est une valeur très faible.

Une autre condition pour la cohérence du résultat est obtenue pour :

$$\lim_{N\to\infty} E\left[\left\|\hat{\theta}_{ML} - \theta_0\right\|^2\right] = 0$$

Exemple 1

Supposons que N points donnés, x_1, x_2, . . ., x_N, ont été générés par une PDF unidimensionnelle gaussienne de moyenne connue μ mais de variance inconnue. Nous allons dériver l'estimé de la variance par le maximum de vraisemblance (ML).

La fonction log-vraisemblance pour ce cas-ci est donnée par :

$$L(\sigma^2) = \ln \prod_{k=1}^{k=N} p(x_k; \sigma^2) = \ln \prod_{k=1}^{k=N} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}$$

Ou :

$$L(\sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{k=N}(x_k - \mu)^2$$

En prenant la dérivée de cette dernière équation par rapport à σ^2 et en mettant le résultat égal à zéro, nous obtenons :

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{k=1}^{k=N}(x_k - \mu)^2 = 0$$

Finalement, l'estimé de σ^2 par la ML sera la solution de l'équation ci-dessus :

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{k=1}^{k=N}(x_k - \mu)^2$$

Qui est l'expression même de la variance.

Exemple 2

Soit x_1, x_2, . . ., x_N des vecteurs découlant d'une distribution normal avec matrice de covariance connue et moyenne inconnue, c'est-à-dire :

$$p(x_k; \mu) = \frac{1}{(2\pi)^{\frac{L}{2}} |\sum|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x_k-\mu)^T \sum^{-1} (x_k-\mu)\right)}$$

Nous allons donner l'estimé du vecteur moyenne par la ML.

Pour N échantillons disponibles, nous avons :

$$L(\mu) = \ln \prod_{k=1}^{k=N} p(x_k; \mu) = -\frac{N}{2}\left((2\pi)^l|\sum|\right) - \frac{1}{2}\sum_{k=1}^{k=N}(x_k - \mu)^T \sum^{-1} (x_k - \mu)$$

En prenant le gradient par rapport à μ, nous obtenons :

$$\frac{\partial L(\mu)}{\partial \mu} = \begin{bmatrix} \frac{\partial L}{\partial \mu_1} \\ \frac{\partial L}{\partial \mu_2} \\ \vdots \\ \vdots \\ \frac{\partial L}{\partial \mu_l} \end{bmatrix} = \sum_{k=1}^{k=N} \Sigma^{-1} (x_k - \mu) = 0$$

Ou nous pouvons déduire facilement :

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{k=1}^{k=N} x_k$$

Qui est l'expression même de la moyenne.

II. 3. Estimation de la probabilité a posteriori maximale (MAP)

Pour la détermination de l'estimé par le ML, nous avons considéré θ comme un paramètre inconnu. Ici, nous allons le considérer comme vecteur aléatoire, et nous allons estimer sa valeur sous la condition que les échantillons $x_1, \ldots, x_N$ se sont produits. Mettons $X = \{x_1, \ldots, x_N\}$.

Notre point de départ est $p(\theta \mid X)$. D'après le théorème de Bayes, nous avons :

$$p(\theta)\, p(X \mid \theta) = p(X)\, p(\theta \mid X)$$

Ou :

$$p(\theta \mid X) = \frac{p(\theta)\, p(X \mid \theta)}{p(X)}$$

L'estimé $\hat{\theta}_{MAP}$ de la probabilité a posteriori maximum (MAP) est défini au point où $p(\theta \mid X)$ devient maximale :

$$\hat{\theta}_{MAP} : \frac{\partial}{\partial \theta}\, p(\theta \mid X) = 0$$

Ou :

$$\hat{\theta}_{MAP} : \frac{\partial}{\partial \theta} \left(p(\theta)\, p(X \mid \theta) \right) = 0$$

On note que p(X) n'est pas impliqué puisqu'elle est indépendante de θ. La différence entre les estimés de la ML et de la MAP, réside dans l'implication de p(θ) dans cette dernière [7].

Exemple

Supposons pour l'exemple 2 précédent que le vecteur des moyennes μ est distribué normalement, c'est à dire :

$$p(\mu) = \frac{1}{(2\pi)^{l/2}\,\sigma_\mu^l} e^{-\frac{\|\mu-\mu_0\|^2}{2\,\sigma_\mu^2}}$$

L'estimé MAP sera donné par la solution de :

$$\frac{\partial}{\partial\mu} \ln\left(\prod_{k=1}^{k=N} p(x_k \mid \mu) p(\mu)\right) = 0$$

Pour $\sum = \sigma^2 I$

$$\sum_{k=1}^{k=N} \frac{1}{\sigma^2}(x_k - \hat{\mu}) - \frac{1}{\sigma_\mu^2}(\hat{\mu} - \mu_0) = 0$$

$$\hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_\mu^2}{\sigma^2}\sum_{k=1}^{k=N} x_k}{1 + \frac{\sigma_\mu^2}{\sigma^2}\,N}$$

Nous constatons que si $\frac{\sigma_\mu^2}{\sigma^2} >> 1$, c'est à dire la variance σ_μ^2 est très grande et que la gaussienne correspondante est très large avec peu de variation sur l'intervalle d'intérêt, alors :

$$\hat{\mu}_{MAP} \approx \hat{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{k=N} x_k$$

De plus, nous observons que ceci reste vraie pour N→ ∞, quel que soit les valeurs des variances. Ainsi l'estimé de la MAP tends asymptotiquement vers celui de la ML.

II. 4. Inférence Bayésienne

Les deux méthodes considérées précédemment, calculent un estimé spécifique du vecteur θ des paramètres inconnus. Dans cette méthode, une approche différente est adoptée. Étant donné l'ensemble des N vecteurs d'apprentissage ainsi que l'information a priori à propos de la PDF p(θ), le but est de calculer la PDF conditionnelle p(x|X) [7], [8]. A cette fin, et faisant usage d'entités connues depuis les bases des statistiques, nous avons à notre disposition l'ensemble des relations suivantes :

$$p(x|X) = \int p(x|\theta)\, p(\theta|X)\, d\theta \qquad (II.4)$$

Avec :

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)\, p(\theta)\, d\theta} \qquad (II.5)$$

$$p(X|\theta) = \prod_{k=1}^{k=N} p(x_k|\theta)$$

La densité conditionnelle p(θ|X) est aussi connue sous le nom de l'estimé de la PDF a posteriori, parce qu'elle met à jour notre connaissance à propos des propriétés statistiques de θ après observation de l'ensemble de données X.

En général, le calcul de p(x|X) exige l'intégration du terme à droite de l'équation (II. 4). Cependant, des solutions analytiques sont possibles, seulement pour des formes spéciales des fonctions concernées. Pour la majorité des cas, les solutions analytiques de même que le dénominateur restent impossibles, et l'on aura recours aux approximations numériques [7], [8].

Si nous regardons de près l'équation de p(x|X) et en supposant que p(θ|X) est connue, alors p(x|X) n'est rien d'autre que la moyenne de p(x|θ) par rapport à θ, c'est-à-dire :

$$p(x|X) = E_\theta[p(x|\theta)]$$

Si nous supposons un nombre d'échantillons assez grand θ_i, i = 1, 2 . . ., M, du vecteur aléatoire θ sont disponibles, nous pouvons calculer les valeurs correspondante $p(x|\theta_i)$ et ainsi approximer l'espérance par la valeur moyenne :

$$p(x|X) \approx \frac{1}{M} \sum_{i=1}^{i=M} p(x|\theta_i)$$

Maintenant, le problème devient la génération d'un ensemble d'échantillons, θ_i, i = 1, 2 . . ., M. Par exemple, si $p(\theta \mid X)$ est une PDF gaussienne, nous pouvons utiliser un générateur pseudo-aléatoire gaussien pour générer les M échantillons. Dans ce cas et en générale, la difficulté est que la forme exacte de $p(\theta \mid X)$ est inconnue et son calcul présuppose l'intégration numérique de la constante normalisante dans le dénominateur de l'équation (II. 5).

Exemple

Soit $p(x|\mu)$ une gaussienne uni-variable $N(\mu, \sigma^2)$ avec comme paramètre inconnu, la moyenne qui est supposé suivre aussi une gaussienne $N(\mu_0, \sigma_0^2)$. A partir de la théorie exposée précédemment, nous aurons :

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} = \frac{1}{\alpha}\prod_{k=1}^{k=N} p(x_k|\mu)\, p(\mu)$$

Où pour un ensemble de données d'apprentissage donné X, p(X) est une constante notée par α, tel que :

$$p(\mu|X) = \frac{1}{\alpha}\prod_{k=1}^{k=N} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\,\sigma_0} e^{-\frac{(\mu-\mu_0)^2}{2\,\sigma_0^2}}$$

Ce n'est qu'une question d'algèbre pour démontrer que pour un nombre donné d'échantillons N, $p(\mu|X)$ s'avère aussi une gaussienne, c'est-à-dire :

$$p(\mu|X) = \frac{1}{\sqrt{2\pi}\,\sigma_N} e^{-\frac{(\mu-\mu_N)^2}{2\,\sigma_N^2}}$$

Avec une valeur moyenne :

$$\mu_N = \frac{N\sigma_0^2\bar{x}_N + \sigma^2\mu_0}{N\sigma_0^2 + \sigma^2}$$

Et une variance :

$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

Aussi nous aurons :

$$\bar{x}_N = \frac{1}{N}\sum_{k=1}^{k=N} x_k$$

Si nous laissons N varier de 1 à ∞, nous générons une séquence de gaussiennes $\mathcal{N}(\mu_N, \sigma_N^2)$, dont les valeurs des moyennes partent de μ_0 et tendent en limite à la moyenne de l'échantillon, qui asymptotiquement devient égal à la vraie valeur moyenne. De plus, leurs variances décroissent au taux de 2/N pour N très grand. Ainsi, pour de grandes valeurs de N, p (μ|X) devient très étroite autours de la moyenne de l'échantillon.

Une fois p (μ|X) calculée, il peut être démontré que :

$$p(x|X) = \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_N^2)}} e^{-\frac{(x-\mu_N)^2}{2(\sigma^2+\sigma_N^2)}}$$

Qui est une PDF gaussienne de moyenne μ_N et de variance $\sigma^2 + \sigma_N^2$.

II. 5. Estimation de l'entropie maximum

Le concept d'entropie est connu à partir de la théorie de l'information de Shannon. C'est la mesure de l'incertitude concernant un évènement, et sous un autre point de vue, la mesure de la stochasticité de messages (vecteurs de caractéristiques par exemple) se produisant à la sortie d'un système. Si p(x) est la fonction, de densité de probabilité, alors l'entropie associée est donnée par [9]-[10] :

$$H = -\int_x p(x) \ln\big(p(x)\big)\, dx \qquad (II.6)$$

Supposons maintenant que p(x) est inconnue mais que nous connaissons un nombre de contraintes relatives : moyenne, variance, etc. L'estimé de l'entropie maximale de la PDF est celui qui maximise l'entropie soumise aux contraintes données. Selon le principe de l'entropie maximale, un tel estimé corresponds à la distribution qui révèle la stochasticité possible la plus élevée soumise aux contraintes disponibles.

Exemple

La variable aléatoire **x** est non-nulle pour $x_1 \leq x \leq x_2$ et nulle autrement. Nous allons calculer l'estimé de l'entropie maximale de sa PDF. Nous avons à maximiser l'entropie sous la contrainte :

$$\int_{x_1}^{x_2} p(x)\, dx = 1$$

Ceci est équivalent à la maximisation de :

$$H_L = -\int_{x_1}^{x_2} p(x)(\ln p(x) - \lambda)\, dx$$

Si nous prenons la dérivée par rapport à p(x), nous obtenons :

$$\frac{\partial H_L}{\partial p(x)} = -\int_{x_1}^{x_2} \{(\ln p(x) - \lambda) + 1\}\, dx$$

Si nous mettons la dérivée égale à zéro, nous obtenons :

$$\hat{p}(x) = e^{(\lambda-1)}$$

Afin de calculer λ, nous le remplaçons dans l'équation de la contrainte et nous obtenons :

$$e^{(\lambda-1)} = \frac{1}{x_2 - x_1}$$

Ainsi :

$$\hat{p}(x) = \begin{cases} \dfrac{1}{x_2 - x_1} & \text{si } x_1 \leq x \leq x_2 \\ 0 & \text{autrement} \end{cases}$$

C'est dire que l'estimé de l'entropie maximale de la PDF inconnue est une distribution uniforme. Ceci va dans le sens de l'entropie maximale puisque nous n'avons imposé aucune autre contrainte à part celle qui est évidente. Toutefois, si la moyenne et la variance sont données comme deuxième et troisième contraintes, l'estimé de l'entropie maximale résultante pour $-\infty < x < +\infty$ est une gaussienne.

II. 6. Modèles mixtes ou mélanges

Une alternative pour modéliser une PDF inconnue p(x) se fait à travers une combinaison linéaire de foncions de densité sous la forme :

$$p(x) = \sum_{j=1}^{j=J} p(x \mid j)\, P_j$$

Où :

$$\sum_{j=1}^{j=J} P_j = 1$$

$$\int_{x} p(x \mid j)\, dx = 1$$

En d'autres termes, il est supposé que **J** distributions contribuent à la formation de p(x). Ainsi, cette modélisation implicite, suppose que chaque point x peut être prie depuis n'importe laquelle des J distributions avec la probabilité P_j, j = 1, 2, . . ., J. il peut être démontré, que cette modélisation, peut approximer aléatoirement une quelconque fonction de densité pour un nombre suffisant de mixantes J et des paramètres de modèle approprié [11]-[16].

La première étape de la procédure implique le choix de l'ensemble des composants de la densité p(x|j) sous forme paramétrique, c'est à dire p(x|j ; θ), et après ce sera le calcul des paramètres inconnus θ ainsi que P_j, j = 1, 2, . . . , J, en se basant sur l'ensemble des échantillons d'apprentissage disponibles x_k. Il existe plusieurs manières pour accomplir ceci. Une formulation typique de la vraisemblance maximale, maximisant la fonction de vraisemblance Π_k p(x_k; θ, P_1, P_2, . . . , P_J) par rapport à θ et les P_j, est une première réflexion. La difficulté ici vient du fait que les paramètres inconnus contribuent à la tâche de maximisation d'une manière non-linéaire ; ainsi des techniques itératives d'optimisation non-linéaire doivent être adoptées. La source de complication qui se présente est le manque d'information à propos des échantillons d'apprentissage disponibles [11]-[16].

II. 7. Algorithme d'espérance-maximisation (EM)

Cet algorithme est idéalement adapté aux cas où l'ensemble les données disponibles est incomplet [11]-[16].

En premier lieu, nous allons exposer le problème en termes plus généraux. Notons par **y** l'ensemble des échantillons, avec $\mathbf{y} \in \mathbf{Y} \subseteq \mathbf{R^m}$, et la PDF correspondante $p_y(y; \theta)$, où θ est le vecteur des paramètres inconnus. Cependant les échantillons **y** ne peuvent être observés directement. Par contre, ceux que nous pouvons observer sont les échantillons $\mathbf{x = g(y) \in X_{ob} \subseteq R^l}$, $\mathbf{l < m}$. Nous notons par $p_x(x; \theta)$ la PDF correspondante. C'est une application équivoque. Notons par $\mathbf{Y\ (x) \subseteq Y}$ le

sous-ensemble de tous les **y** correspondant à un **x** spécifique. Alors, la PDF pour les données incomplètes sera donnée par :

$$p_x(x;\ \theta) = \int_{Y(x)} p_y(y;\ \theta)\, dy \qquad (II.7)$$

L'estimé de la vraisemblance maximale de θ sera donné par :

$$\hat{\theta}_{ML} : \sum_{k=1}^{k=N} \frac{\partial \ln\left(p_y(y_k;\ \theta)\right)}{\partial \theta} = 0$$

Cependant, les **y** ne sont pas disponibles. Ainsi, l'algorithme EM maximise l'espérance de la fonction log-vraisemblance, conditionnée par les échantillons observés et l'estimé de θ à l'itération en cours.

Les deux étapes de l'algorithme sont :

Étape E :

A l'étape (t + 1) de l'itération, où θ (t) est disponible, calculer la valeur espérée de :

$$Q\big(\theta; \theta(t)\big) = E\left[\sum_k \ln\left(p_y\big(y_k;\ \theta \mid X; \theta(t)\big)\right)\right]$$

Ceci est appelé l'étape d'espérance de l'algorithme.

Étape M :

Calculer l'estimé (t + 1) suivant de θ en maximisant Q(θ; θ (t)), c'est-à-dire :

$$\theta(t+1) : \frac{\partial Q\big(\theta; \theta(t)\big)}{\partial \theta} = 0$$

Ceci est l'étape de maximisation de l'algorithme.

Afin d'appliquer l'algorithme EM, nous commençons depuis un estimé initial θ (0), et le processus itératif prend fin si $\|\theta\ (t+1) - \theta\ (t)\| \leq \varepsilon$ pour une norme vectorielle appropriée et une bonne valeur de ε.

II. 8. Application au problème de modélisation mixte

Dans ce cas l'ensemble de données complètes consiste en les évènements associés (x_k, j_k), k = 1, 2, . . ., N, et j_k prends des valeurs entières dans l'intervalle [1, J], et indique la mixtante d'où x_k est généré. En employant la règle familière, nous obtenons :

$$p(x_k , j_k ; \theta) = p(x_k \mid j_k ; \theta)\, P_{jk}$$

Si nous supposons une indépendance mutuelle entre les échantillons de l'ensemble de données, alors la fonction log-vraisemblance devient :

$$L(\theta) = \sum_{k=1}^{k=N} \ln \left(p(x_k \mid j_k ; \theta)\, P_{jk} \right)$$

Mettons $P = [P_1, P_2, \ldots, P_J]^T$. Dans la configuration actuelle, le vecteur paramètre inconnu est $\Theta^T = [\theta^T, P^T]^T$. En prenant l'espérance sur les données non-observées, conditionnées sur les échantillons d'apprentissage et les estimés en cours Θ (t) des paramètres inconnus, alors nous avons :

Etape E :

$$Q(\Theta; \Theta(t)) = E\left[\sum_{k=1}^{k=N} \ln \left(p(x_k \mid j_k ; \theta)\, P_{jk} \right) \right]$$

$$Q(\Theta; \Theta(t)) = \sum_{k=1}^{k=N} E\left[\ln \left(p(x_k \mid j_k ; \theta)\, P_{jk} \right) \right]$$

$$Q(\Theta; \Theta(t)) = \sum_{k=1}^{k=N} \sum_{j_k=1}^{j_k=J} P(j_k \mid x_k ; \Theta(t)) \ln \left(p(x_k \mid j_k ; \theta)\, P_{jk} \right)$$

La notation peut maintenant être simplifiée en laissant tomber l'indice k de j_k, parce que pour chaque k, nous sommons sur toutes les valeurs de j_k et ceci pour toutes les valeurs de k.

Nous allons démontrer l'algorithme pour le cas de mixtures gaussiennes avec une matrice de covariance diagonale sous la forme $\sum_j = \sigma_j^2$ I, c'est-à-dire :

$$p(x_k \mid j\,;\,\theta) = \frac{1}{\left(2\pi\sigma_j^2\right)^{l/2}} e^{\left(-\frac{\|x_k-\mu_j\|^2}{2\sigma_j^2}\right)}$$

Supposons en plus des probabilités P_j, que les moyennes respectives μ_j ainsi que les variances σ_j^2, j =1, 2, . . , J, pour les gaussiennes sont aussi connues. En combinant les deux dernières équations et en négligeant les constantes, nous obtenons :

Étape E :

$$Q(\Theta;\Theta(t)) = \sum_{k=1}^{k=N}\sum_{j=1}^{j=J} P(j \mid x_k\,;\,\Theta(t))\left(-\frac{l}{2}\ln(\sigma_j^2) - \frac{1}{2\sigma_j^2}\|x_k-\mu_j\|^2 + \ln(P_j)\right)$$

Étape M : en maximisant l'équation ci-dessus par rapport à μ_j, σ_j^2 et P_j, nous obtenons comme résultat :

$$\mu_j(t+1) = \frac{\sum_{k=1}^{k=N} P(j \mid x_k\,;\,\Theta(t))x_k}{\sum_{k=1}^{k=N} P(j \mid x_k\,;\,\Theta(t))}$$

$$\sigma_j^2(t+1) = \frac{\sum_{k=1}^{k=N} P(j \mid x_k\,;\,\Theta(t))x_k\,\|x_k-\mu_j(t+1)\|^2}{l\sum_{k=1}^{k=N} P(j \mid x_k\,;\,\Theta(t))}$$

$$P_j(t+1) = \frac{1}{N}\sum_{k=1}^{k=N} P(j \mid x_k\,;\,\Theta(t))$$

Pour que les itérations soient complètes, nous avons besoin seulement de calculer $P(j \mid x_k\,;\,\Theta(t))$. Ceci est facilement obtenu depuis :

$$P(j \mid x_k\,;\,\Theta(t)) = \frac{p(x_k \mid j\,;\,\theta(t))\,P_j(t)}{p(x_k\,;\,\Theta(t))}$$

$$p\left(x_k\,;\,\Theta(t)\right) = \sum_{j=1}^{j=J} p\left(x_k \mid j\,;\,\theta(t)\right) P_j(t)$$

Les cinq équations constituent l'algorithme Espérance-Maximisation (EM) pour l'estimation des paramètres inconnus des mixtures gaussiennes. L'algorithme commence par un estimé initial valide, c'est-à-dire que la somme des probabilités doit être égale à 1.

L'algorithme Espérance-Maximisation sera au cœur même de notre module de calcul pour l'estimation des paramètres optimaux d'un modèle gaussien mixte. C'est dire que nous allons le détailler encore plus dans le chapitre trois qui va suivre.

II. 9. Conclusion

Dans ce deuxième chapitre, nous avons exposé un bon nombre de méthodes pour l'estimation des fonctions de densité de probabilité inconnues et leurs statistiques d'ordre un et d'ordre deux (moyennes et variances) pour le cas spécial de distribution gaussienne qui représente grandement un grand nombre de réalisation dans la réalité.

L'algorithme espérance-maximisation sera intégré au module de calcul des paramètres optimaux pour les densités de probabilité ainsi que les proportions de mélange de celles-ci dans un modèle mixte. Les résultats de simulation seront au cœur du chapitre trois suivant.

Chapitre III : Résultats

III. 1. Introduction

Au chapitre deux, nous avons évoqué l'intégration de l'algorithme d'espérance-maximisation dans un module de calcul itératif afin d'extraire les paramètres optimaux pour un modèle mixte. Il s'agit en effet des probabilités a priori de chaque mixante et classe, la matrice de covariance, le vecteur des moyennes pour le cas spécial d'une distribution gaussienne dans un espace L-dimensions.

Le chapitre trois sera consacré au résultat des différentes simulations, où dans un premier temps, nous allons générer N échantillons à L-dimensions et pour un nombre de classes K avec comme données la matrice de covariance et le vecteur de moyennes ainsi que les probabilités des composantes de la mixture (a priori).

Dans un deuxième temps, nous allons résoudre le problème inverse et nous nous poserons dans la situation où nous avons N échantillons à L-dimensions observés et nous appliquerons l'algorithme EM pour trouver les matrices de covariance et les vecteurs de moyennes ainsi que les probabilités a priori. Il faut rappeler qu'il s'agit d'un problème d'optimisation et que la solution finale n'est jamais garantie pour être la meilleure à cause du problème des optima locaux qui fausse la convergence d'où l'intérêt d'un bon estimé initial.

III. 2. Modèles mixtes

Nous avons un ensemble de données ou échantillons $X= [x_1, x_2,\ldots, x_N]$, où x_i est un vecteur de mesure ou d'observations à L-dimensions. Nous supposons que les points ont été générés à partir de la densité p(x) et que p(x) est définit comme étant un modèle mixte ou mélange finie à K composantes, c'est-à-dire :

$$p(x\,|\Theta) = \sum_{k=1}^{k=K} P_k\, p_k(x \mid z_k, \theta_k) \qquad \text{(III. 1)}$$

Où :

Les $p_k(x|z_k, \theta_k)$ sont les composantes de la mixture, $1 \leq k \leq K$. Chacune est une densité définie sur p(x), avec les paramètres θ_k.

$z = [z_1, z_2,\ldots, z_K]$ est un vecteur de K indicateurs binaires qui sont mutuellement exclusive, c'est-à-dire un seul et seulement un seul des z_k est égal à 1 et les autres sont nuls. En effet, z est une variable aléatoire représentant l'identité de la composante de la mixture qui a générée x [11]-[16].

Les $P_k = p(z_k)$ sont les poids de la mixture (probabilités a priori), représentant la probabilité qu'un **x** choisie aléatoirement ait été généré par la composante k et où les conditions suivantes doivent être respectées :

$$\sum_{k=1}^{k=K} P_k = 1$$

$$\int_x p_k(x \mid z_k, \theta_k)\, dx = 1$$

L'ensemble complet des paramètres pour un modèle mixte à K composantes est donné par :

$$\Theta = \{\alpha_1, \alpha_2, \dots, \alpha_K, \theta_1, \theta_2, \dots, \theta_K\}$$

III. 2. 1. Poids de membres

Nous pouvons calculer les poids de membres pour les données x_i dans le groupe k, étant donné les paramètres Θ, par :

$$\omega_{ik} = p(z_{ik} = 1 \mid x_i, \Theta) = \frac{p_k(x_i \mid z_k, \theta_k)\, \alpha_k}{\sum_{m=1}^{m=K} \alpha_m\, p_m(x_i \mid z_m, \theta_m)} \qquad \text{(III. 2)}$$

$1 \leq k \leq K$ et $1 \leq i \leq N$

N : nombre d'échantillons ou points de données.

K : nombre de composantes dans le mélange et de classes.

III. 2. 2. Modèles mixtes gaussiens

Pour $x \in R^L$, nous pouvons définir un modèle mixte gaussien en mettant chacune des K composantes comme densité gaussienne avec les paramètres μ_k et $\sum_k$. Chaque composante est une densité gaussienne multi-variables, définie par :

$$p_k(x \mid \theta_k) = \frac{1}{(2\pi)^{\frac{L}{2}} |\sum_k|^{\frac{1}{2}}}\, e^{\left(-\frac{1}{2}(x-\mu_k)^T \sum_k^{-1} (x-\mu_k)\right)} \qquad \text{(III. 3)}$$

Avec ses propres paramètres :

$$\theta_k = \{\mu_k, \sum_k\}$$

μ_k : moyenne des échantillons apparentant à la composante k.

$\sum_k$: diagonale de la matrice de covariance.

L : nombre de dimensions

III. 3. Algorithme Espérance Maximisation pour modèles mixtes gaussiens

Nous définissons l'algorithme Espérance-Maximisation (EM) pour les mélanges gaussiens comme suit :

L'algorithme est un algorithme itératif qui commence par un estimé initial de Θ (qui peut être aléatoire), et puis procède à une mise à jour itérative de Θ jusqu'à détection d'une convergence (faible variations). Chaque itération consiste en une étape E et une étape M [11]-[16].

Étape E : notons par Θ la valeur en cours du paramètre. Calculer ω_{ik} en utilisant l'équation de poids de membres (III. 2) pour tous les points x_i, $1 \leq i \leq N$ et pour toutes les composantes de la mixture $1 \leq k \leq K$. Notons que pour chaque point x_i, les poids de membres sont définies de sorte que :

$$\sum_{k=1}^{k=K} \omega_{ik} = 1$$

Ceci donne une matrice de taille (N, K) pour les poids de membres, où la somme sur chaque ligne est égale à 1.

Étape M :

Ici nous utiliserons les poids et les données ou échantillons pour calculer les nouvelles valeurs de paramètres. Nous posons :

$$\sum_{i=1}^{i=N} \omega_{ik} = N_k \qquad \text{(III. 4)}$$

Spécifiquement, nous aurons :

$$P_k^{t+1} = \frac{N_k}{N} \qquad \text{(III. 5)}$$

Qui sont les nouveaux poids à l'itération (t+1).

La moyenne mise à jour à l'itération (t+1) est calculée d'une manière similaire au calcul d'une moyenne empirique standard, sauf que l'i^{ème} vecteur de données x_i possède un poids fractionnel ω_{ik}. Cette moyenne est décrite par l'équation suivante :

$$\mu_k^{t+1} = \left(\frac{1}{N_k}\right)\sum_{i=1}^{i=N} \omega_{ik}\, x_i \qquad \text{(III. 6)}$$

Notons que cette dernière équation est une équation vecteur puisque μ_k^{t+1} et x_i sont tous deux des vecteurs à L-dimensions.

Les éléments diagonaux des matrices de covariance à l'itération (t+1), prendrons pour expression :

$$\sum_k^{t+1} = \left(\frac{1}{N_k}\right)\sum_{i=1}^{i=N} \omega_{ik}\left(x_i - \mu_k^{t+1}\right)\left(x_i - \mu_k^{t+1}\right)^T \qquad \text{(III. 7)}$$

Encore une fois nous obtenons une équation qui est similaire en forme à ce que l'on devrait trouver par calcul empirique de la matrice de covariance, sauf que la contribution de chaque point de données est pondéré par ω_{ik}. Notons que cette dernière équation et sur ses deux côtés, est une équation matrice de dimensions (L * L).

Remarque

Les équations dans l'étape M doivent être calculées dans cet ordre, c'est-à-dire que nous devrons commencer par calculer les K nouveaux poids P, puis les K nouvelles moyennes μ et finalement les K nouveaux éléments diagonaux $\sum$.

Après avoir calculer tous les nouveaux paramètres, l'étape M est terminé et nous pouvons revenir en arrière et recalculer les poids de membres à l'étape E, et ainsi continuer la mise à jour des paramètres de cette manière. Chaque paire d'étapes E et M est considéré comme étant une itération [11]-[16].

III. 4. Initialisation et convergence pour l'algorithme

L'algorithme EM peut débuter soit par initialisation de l'algorithme par un ensemble de paramètres initiaux et conduire une étape E, soit en commençant par un ensemble de poids initiaux et conduire une première étape M.

Les paramètres initiaux peuvent être choisis aléatoirement. La convergence est généralement détectée par calcul de la valeur du log-vraisemblance après chaque

itération et stopper lorsqu'il n'y a plus de changement significatif entre l'itération actuelle et la précédente. Nous notons que la fonction de log-vraisemblance est définie par :

$$\log l(\Theta) = \sum_{i=1}^{i=N} \log p(x_i, \Theta) = \sum_{i=1}^{i=N} \left(\log \sum_{k=1}^{k=K} \alpha_k p_k(x_i \mid z_k, \theta_k) \right) \quad \text{(III. 8)}$$

Où :

$p_k(x_i \mid z_k, \theta_k)$ est la densité gaussienne pour la $k^{ème}$ composante de la mixture.

III. 5. Fonction de densité de probabilité mixte

Nous donnons ici les tracés de la PDF multimodale pour modèle mixte unidimensionnel à trois composantes. En effet les figures III. 1, III. 2, III. 3, représentent de suite les PDF pour gaussiennes ayant, respectivement et en exemple, les statistiques (moyennes et variances) suivantes :

$\mu = 3, \sigma^2 = 4.$

$\mu = 10, \sigma^2 = 8.$

$\mu = 5, \sigma^2 = 30.$

La figure III. 4, quant à elle, représente la somme pondérée des trois gaussiennes par les poids (probabilités a priori) respectifs : 0.4, 0.4, 0.2.

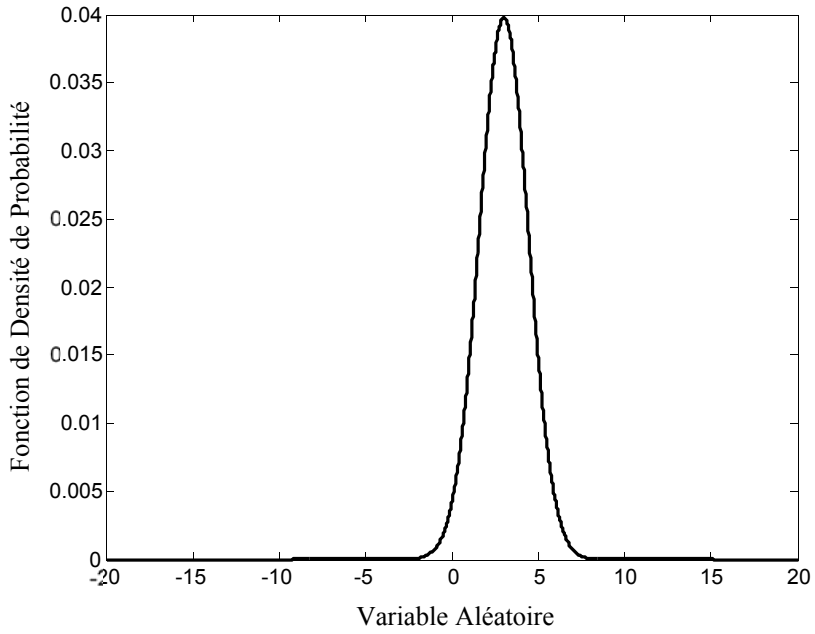


Figure III. 1. PDF en fonction de la variable aléatoire

$\mu = 3, \sigma^2 = 4$

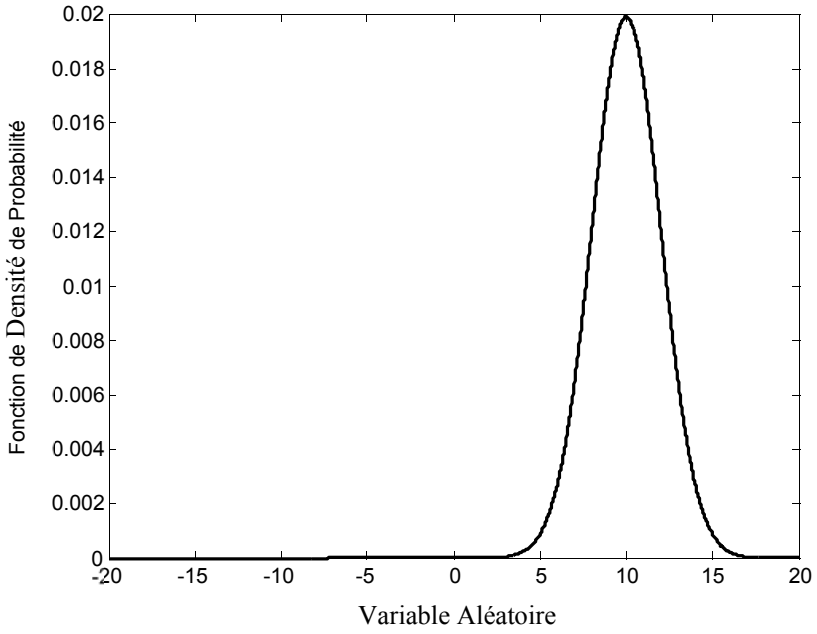


Figure III. 2. PDF en fonction de la variable aléatoire

$\mu = 10, \sigma^2 = 8$

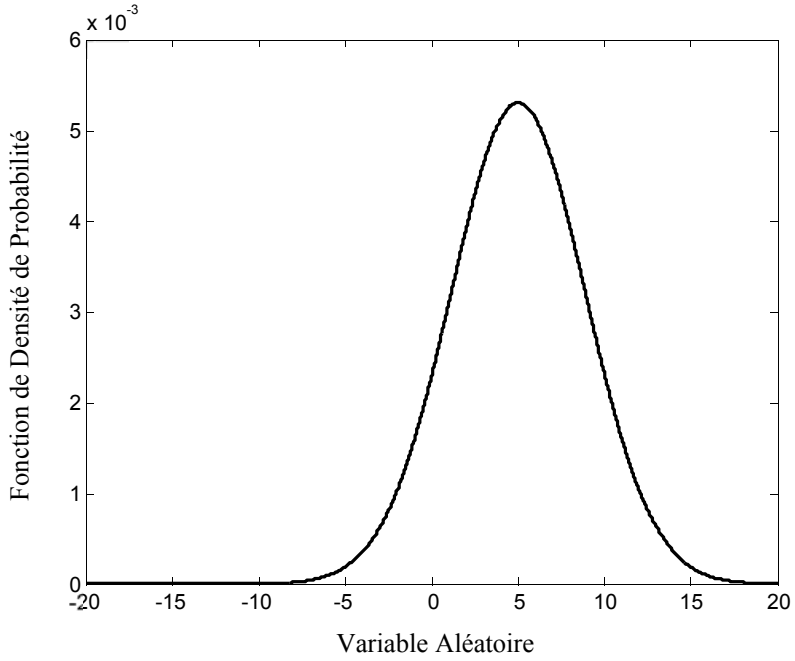


Figure III. 3. PDF en fonction de la variable aléatoire

$\mu = 5, \sigma^2 = 30$

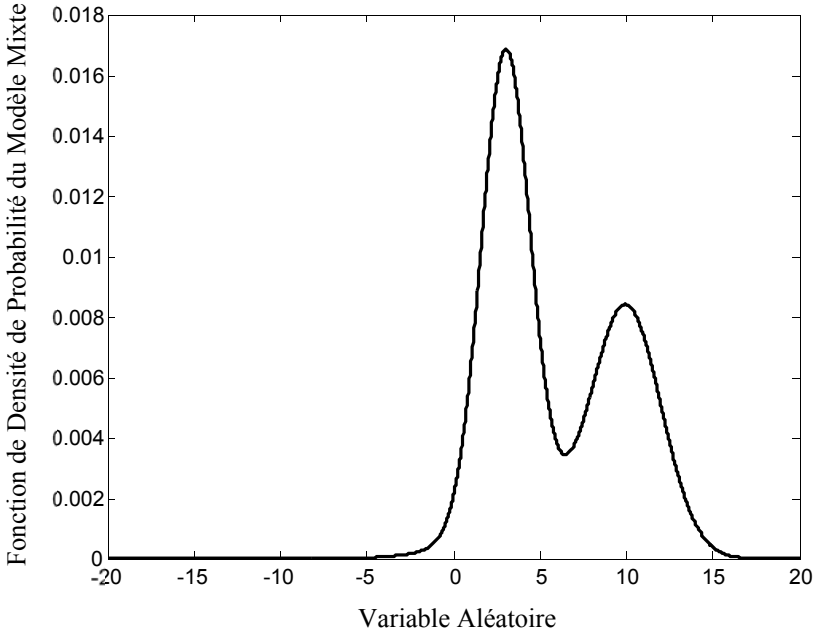


Figure III. 4. PDF du modèle mixte avec :

$P_1 = 0.4, P_2 = 0.4, P_3 = 0.2$

III. 6. Résultats et commentaires

Notre module de calcul à base d'algorithme espérance-maximisation est composé de deux partie ; la partie génération des données aléatoires, obéissant à la distribution pondérée du modèle mixte, et la partie estimation, par optimisation, des paramètres des composantes gaussiennes du modèle mixte. Nous étalerons les résultats pour deux applications afin de confirmer la capacité de notre module de calcul à retrouver les paramètres optimaux du modèle mixte et le valider par voie de conséquence.

Les résultats de la première partie, et pour la première application, seront représentés par les différentes PDF du modèle mixte et les formes en grappes qui donnent pour le cas particulier de L=2, le tracé de la deuxième dimension en fonction de la première dimension afin de visionner la forme et la localisation des différents échantillons dans l'espace 2-dimensions.

Les résultats de la deuxième partie, et pour la première application, concernerons la convergence des paramètres à travers les itérations ainsi que leurs valeurs optimales pour les K gaussiennes du modèle mixte.

Pour la deuxième application et pour raisons de commodité, il sera question uniquement de valeurs optimales pour les paramètres du modèle mixte. Bien sûr, nous entendons par commodité, la non-redondance ainsi que la capacité du lecteur à visualiser et interpréter uniquement les tracés en 2D.

III. 6. 1. Première application

Dans cette première application, nous allons donner les résultats sur les différentes PDF du modèle, les formes en grappes, la convergence, les moyennes optimales, les matrices de covariances optimales, ainsi que les poids optimaux pour le modèle mixte.

Tout d'abords et dans la première partie, nous allons générer grâce à notre module de calcul N échantillons à L-dimensions appartenant à K classes différentes, avec :

N = 1000 : nombre d'échantillons.

L = 2 : nombre de dimensions.

K = 3 : nombre de classes et de mixantes.

III. 6. 1. 1. Génération des échantillons

Nous allons, dans un premier temps générer les N échantillons à L-dimensions appartenant à K classes différentes. Les statistiques d'ordre un et d'ordre deux ainsi que les poids pour les gaussiennes générées sont de suite :

Moyennes :

$\mu_1 = [\mu_{11}, \mu_{12}] = [1, 2]$,

$\mu_2 = [\mu_{21}, \mu_{22}] = [4, 3]$,

$\mu_3 = [\mu_{31}, \mu_{32}] = [0, 6]$.

Matrices de covariances :

$$\sum_1 = \begin{bmatrix} (\sigma_{11}^2)_1 & (\sigma_{12}^2)_1 \\ (\sigma_{21}^2)_1 & (\sigma_{22}^2)_1 \end{bmatrix} = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}$$

$$\sum_2 = \begin{bmatrix} (\sigma_{11}^2)_2 & (\sigma_{12}^2)_2 \\ (\sigma_{21}^2)_2 & (\sigma_{22}^2)_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\sum_3 = \begin{bmatrix} (\sigma_{11}^2)_3 & (\sigma_{12}^2)_3 \\ (\sigma_{21}^2)_3 & (\sigma_{22}^2)_3 \end{bmatrix} = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$$

Poids :

$P_1 = 0.5$, $P_2 = 0.3$, $P_3 = 0.2$

III. 6. 1. 2. Tracé des différentes PDF du modèle mixte

Nous avons engagé une simulation à base de K = 3 PDF qui constituent le modèle mixte d'où découlent les N échantillons. Les trois PDF sont ainsi représentées, de suite, sur les figures III. 5, III. 6 et III. 7. La PDF multimodale résultante est quant à elle représentée sur la figure III. 8. On s'aperçoit vite qu'elle comprend deux maxima d'où la notion de multimodale.

Les tracés des PDF sont réalisés en coupe à 45°, c'est-à-dire pour $x_1 = x_2 = x$ (vecteur de taille N) en abscisse et les éléments diagonaux des PDF (matrices de taille N * N) en ordonnées (voir équation (III. 3)).

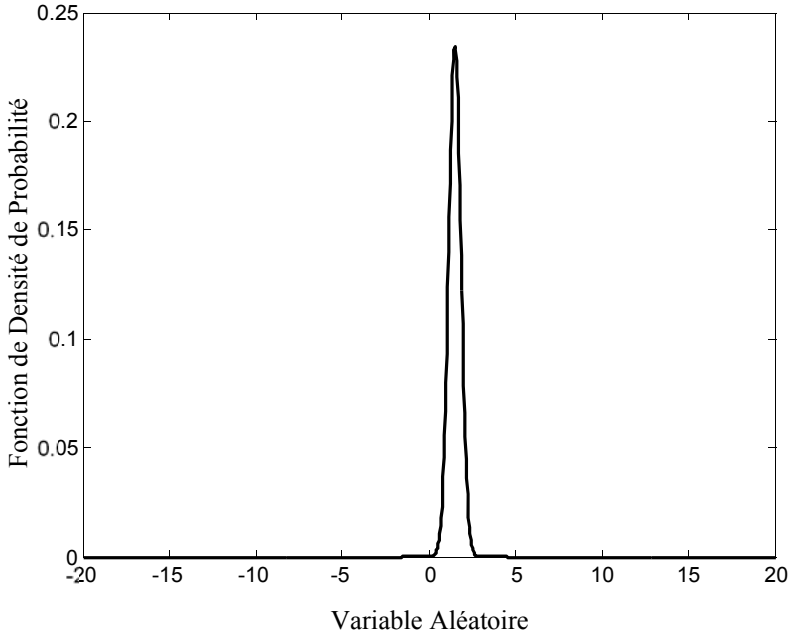


Figure III. 5. Eléments diagonaux de la première PDF
en fonction de la variable aléatoire

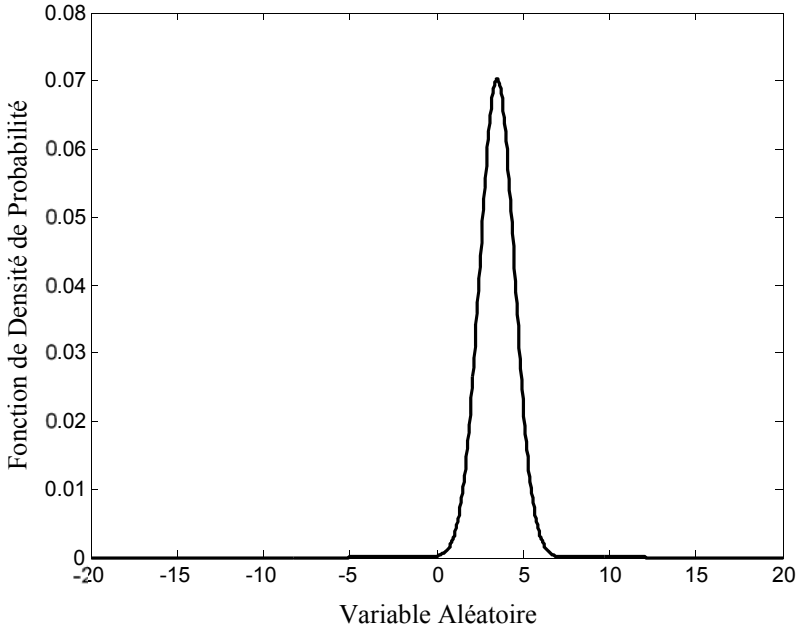


Figure III. 6. Eléments diagonaux de la deuxième PDF
en fonction de la variable aléatoire

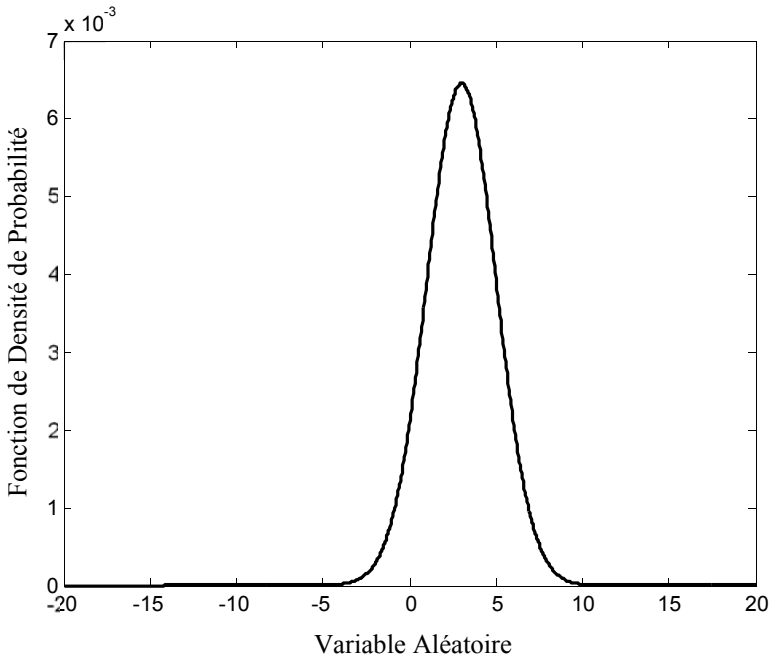


Figure III. 7. Eléments diagonaux de la troisième PDF
en fonction de la variable aléatoire

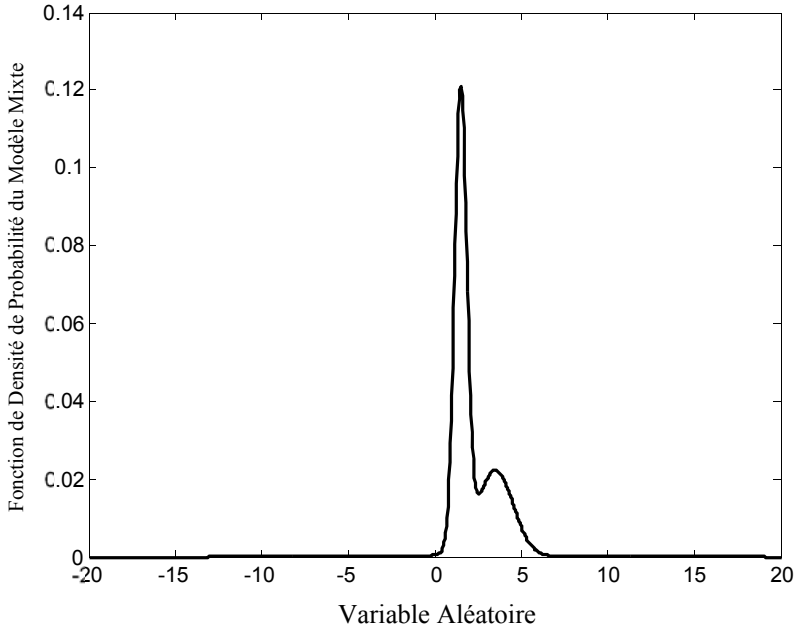


Figure III. 8. Eléments diagonaux de la PDF mixte
en fonction de la variable aléatoire

III. 6. 1. 3. Estimation des statistiques et poids optimaux

Nous allons lancer la simulation en injectant des valeurs initiales aléatoires pour les moyennes, les matrices de covariances et les poids :

Moyennes initiales :

$\mu_{i1} = [0, 2]$,

$\mu_{i2} = [5, 2]$,

$\mu_{i3} = [5, 5]$.

Matrices de covariances initiales :

$$\sum_{i1} = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}$$

$$\sum_{i2} = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}$$

$$\sum_{i3} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

Poids initiaux :

$P_{i1} = 0.33$, $P_{i2} = 0.33$, $P_{i3} = 0.33$

Erreur relative minimale exigée : 10^{-15}

Simulation

Après simulation et 86 itérations nécessaires, nous avons trouvé les valeurs optimales suivantes :

Moyennes optimales :

$\mu_{f1} = [\mu_{f11}, \mu_{f12}] = [1.0130, 1.9901]$,

$\mu_{f2} = [\mu_{f21}, \mu_{f22}] = [4.0565, 2.9439]$,

$\mu_{f3} = [\mu_{f31}, \mu_{f32}] = [0.2183, 6.0806]$.

Matrices de covariances optimales :

$$\sum_{f1} = \begin{bmatrix} (\sigma_{11}^2)_{f1} & (\sigma_{12}^2)_{f1} \\ (\sigma_{21}^2)_{f1} & (\sigma_{22}^2)_{f1} \end{bmatrix} = \begin{bmatrix} 0.2578 & 0 \\ 0 & 0.2578 \end{bmatrix}$$

$$\sum_{f2} = \begin{bmatrix} (\sigma_{11}^2)_{f2} & (\sigma_{12}^2)_{f2} \\ (\sigma_{21}^2)_{f2} & (\sigma_{22}^2)_{f2} \end{bmatrix} = \begin{bmatrix} 2.0505 & 0 \\ 0 & 2.0505 \end{bmatrix}$$

$$\sum_{f3} = \begin{bmatrix} (\sigma_{11}^2)_{f3} & (\sigma_{12}^2)_{f3} \\ (\sigma_{21}^2)_{f3} & (\sigma_{22}^2)_{f3} \end{bmatrix} = \begin{bmatrix} 8.0960 & 0 \\ 0 & 8.0960 \end{bmatrix}$$

Poids optimaux :

$P_{f1} = 0.4976$, $P_{f2} = 0.2893$, $P_{f3} = 0.2131$

Sur les figures III. 9, III. 10 et III. 11, nous avons représenté la dispersion des échantillons dans l'espace (1ère dimension, 2ème dimension), c'est-à-dire le tracé en grappe de x_2 en fonction de x_1 pour cette application. Il va sans dire que pour L>2, cette représentation ne serait plus perceptible pour le lecteur. Nous pouvons remarquer que sur la figure III. 9, la forme est presque un disque, tandis que sur les deux autres c'est presque une forme d'ellipse à grande axe horizontale pour la figure III. 10 et verticale pour la figure III. 11.

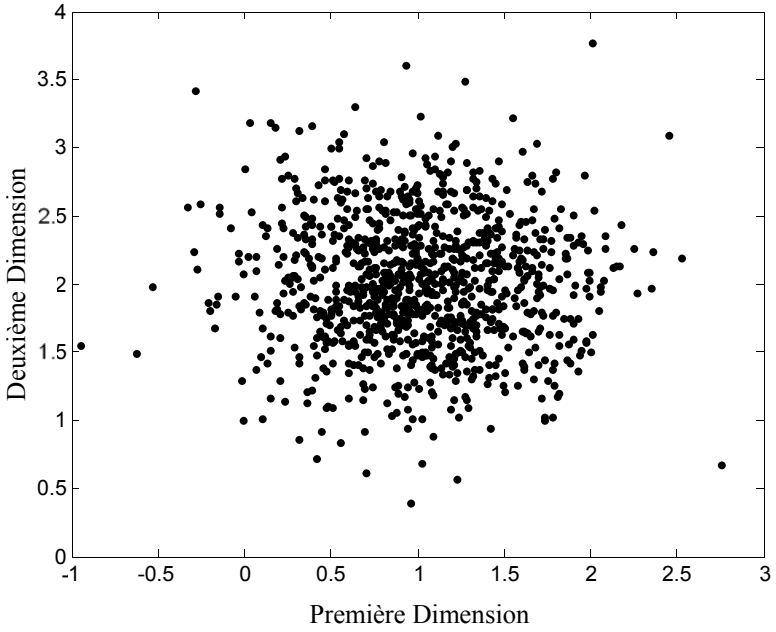


Figure III. 9. Forme de grappe pour la première classe

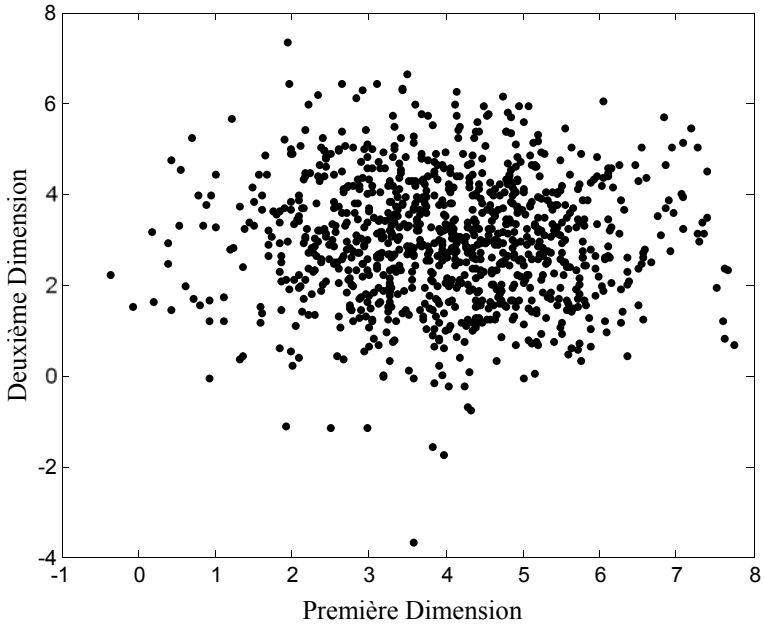


Figure III. 10. Forme de grappe pour la deuxième classe

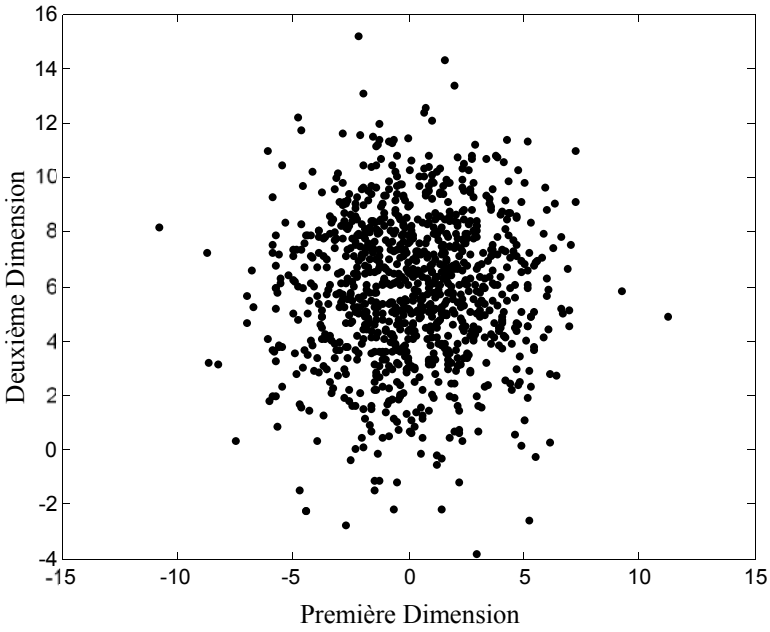


Figure III. 11. Forme de grappe pour la troisième classe

La convergence pour les poids optimaux est représentée sur les figures III. 12, III. 13 et III. 14. Nous pouvons voir que la convergence est rapide, vers la 86ème itération pour l'erreur relative exigée. Nous pouvons remarquer aussi que sur la figure III. 12, cette convergence était au début ascendante depuis la 1ère jusqu'à la 3ème itération puis descendante pour le reste, de la 3éme à la 86ème itération. Pour la figure III. 13 la convergence est strictement ascendante et pour la figure III. 14 elle est descendante.

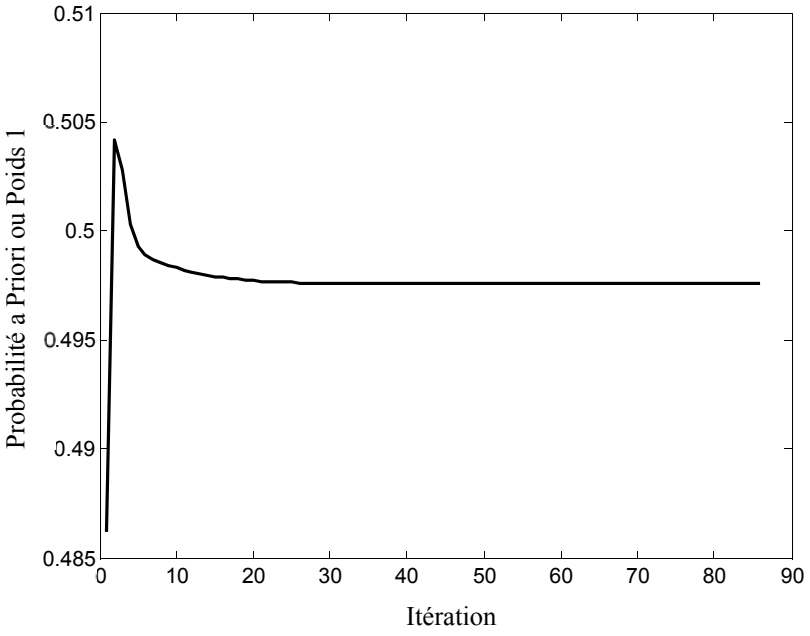


Figure III. 12. Convergence du poids optimisé P_{f1}
de la première PDF du modèle mixte

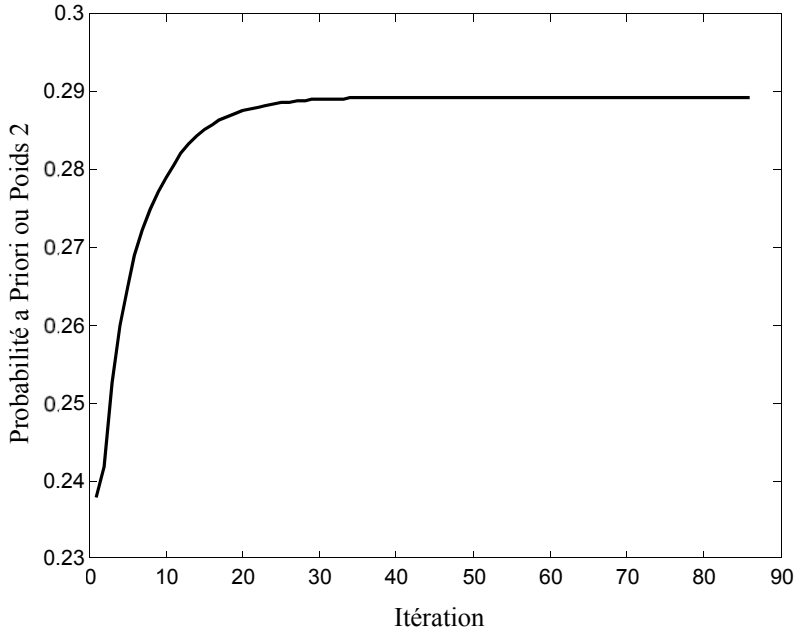


Figure III. 13. Convergence du poids optimisé P_{f2}
de la deuxième PDF du modèle mixte

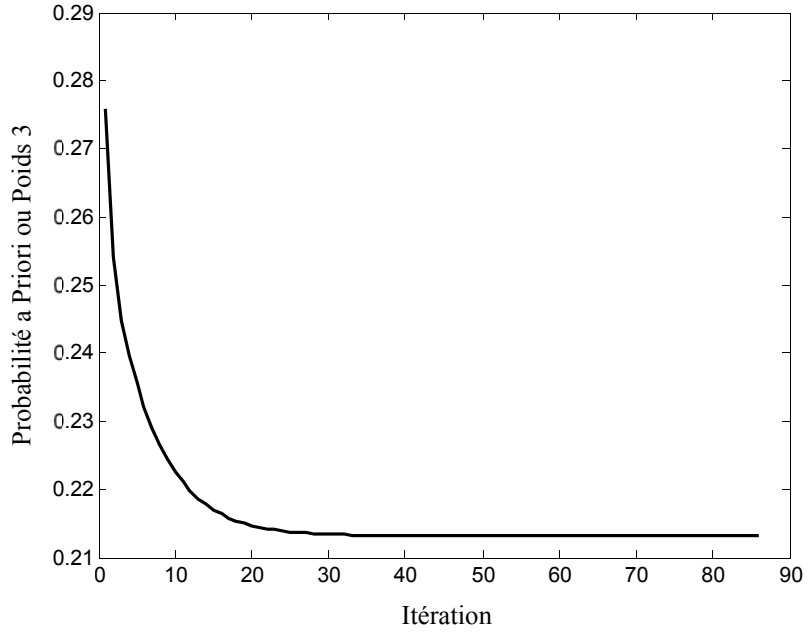


Figure III. 14. Convergence du poids optimisé P_{f3}
de la troisième PDF du modèle mixte

De la même manière nous donnons maintenant, la convergence pour les (K = 3) moyennes et ce pour leurs deux composantes x_1 et x_2 (L = 2). Nous pouvons dire que sur l'ensemble la convergence est très rapide en forme exponentielle et qu'elle est parfois ascendante, parfois descendante et parfois mixte. Les figures III. 15 à III. 20 sont dédiées aux convergences des différentes moyennes composées en fonction des itérations.

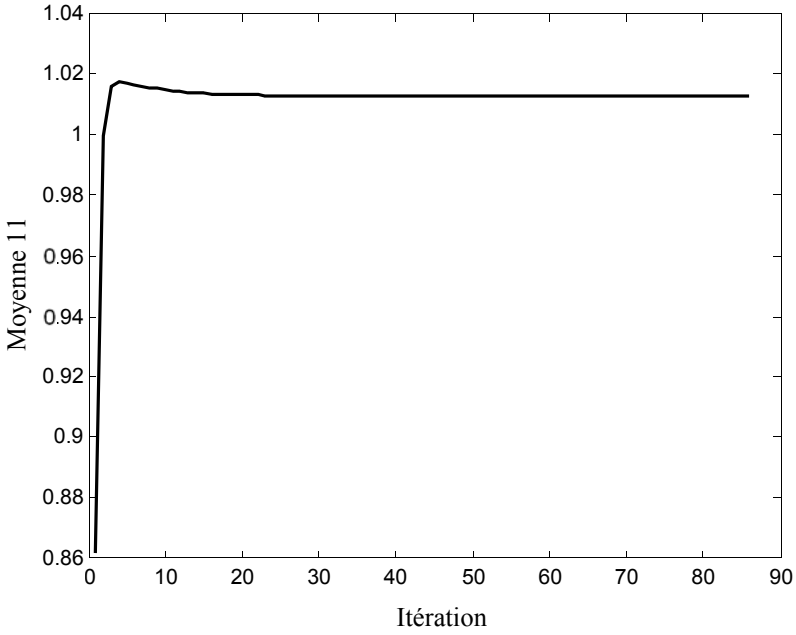


Figure III. 15. Convergence pour μ_{f11}, première composante de la moyenne pour la première PDF du modèle mixte

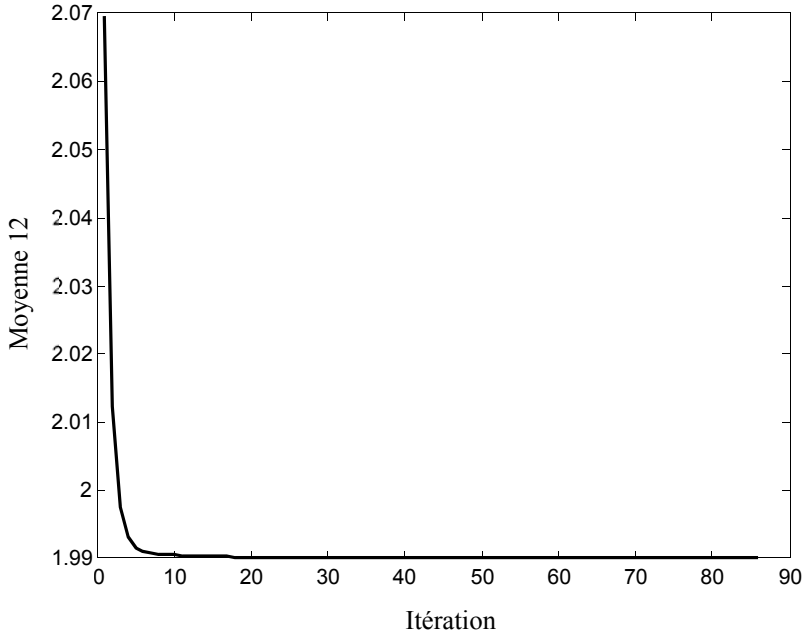


Figure III. 16. Convergence pour μ_{f12}, deuxième composante de la moyenne pour la première PDF du modèle mixte

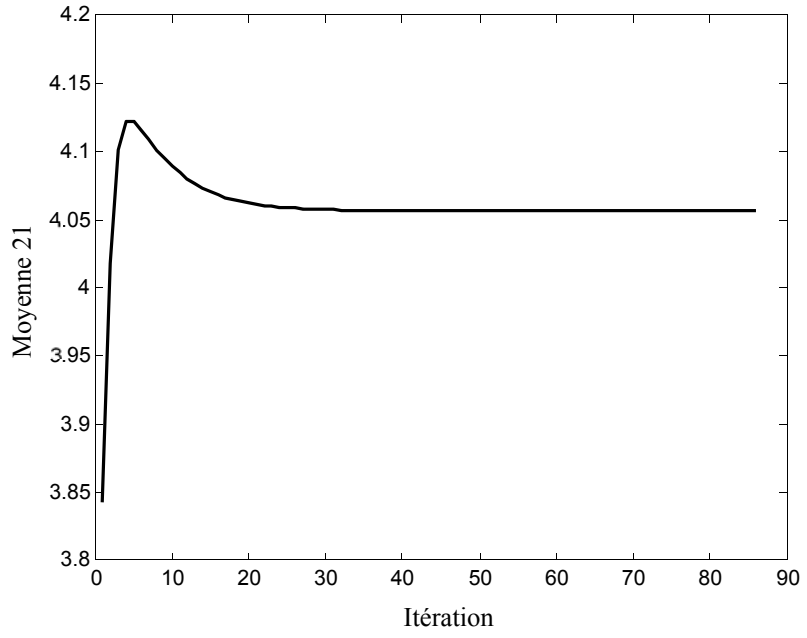


Figure III. 17. Convergence pour μ_{f21}, première composante de la moyenne pour la deuxième PDF du modèle mixte

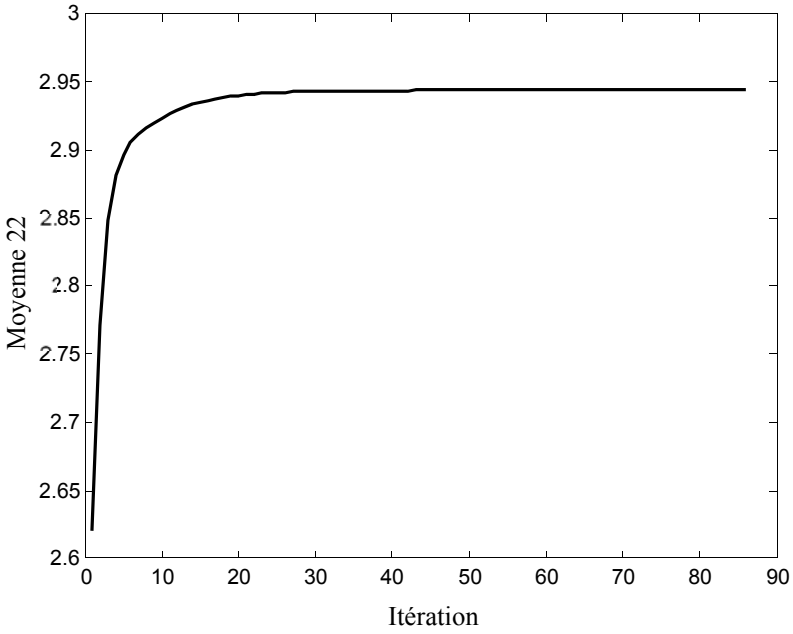


Figure III. 18. Convergence pour μ_{f22}, deuxième composante de la moyenne pour la deuxième PDF du modèle mixte

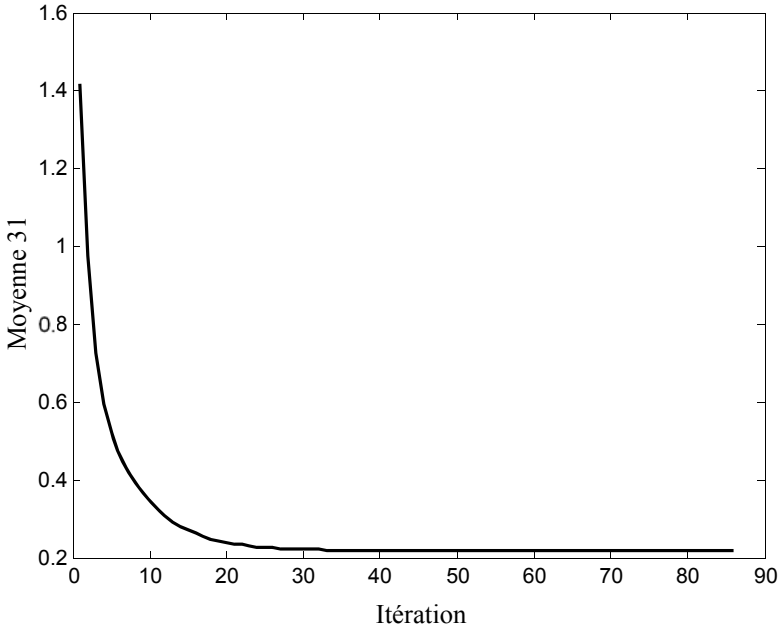


Figure III. 19. Convergence pour μ_{f31}, première composante de la moyenne pour la troisième PDF du modèle mixte

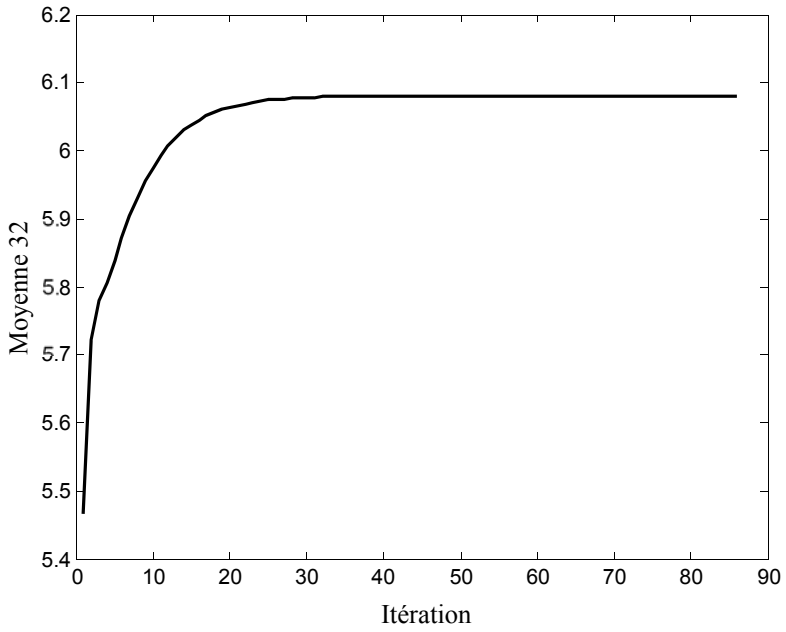


Figure III. 20. Convergence pour μ_{f32}, deuxième composante de la moyenne pour la troisième PDF du modèle mixte

Dans la suite logique, la convergence des éléments diagonaux des matrices de covariance pour les trois PDF est représentée sur les figures III. 21, III. 22 et III. 23. Là aussi et sur la figure III. 21, la convergence est descendante alors que sur les deux autres, elle est mixte. Pour les trois figures, la convergence est exponentielle et donc très rapide.

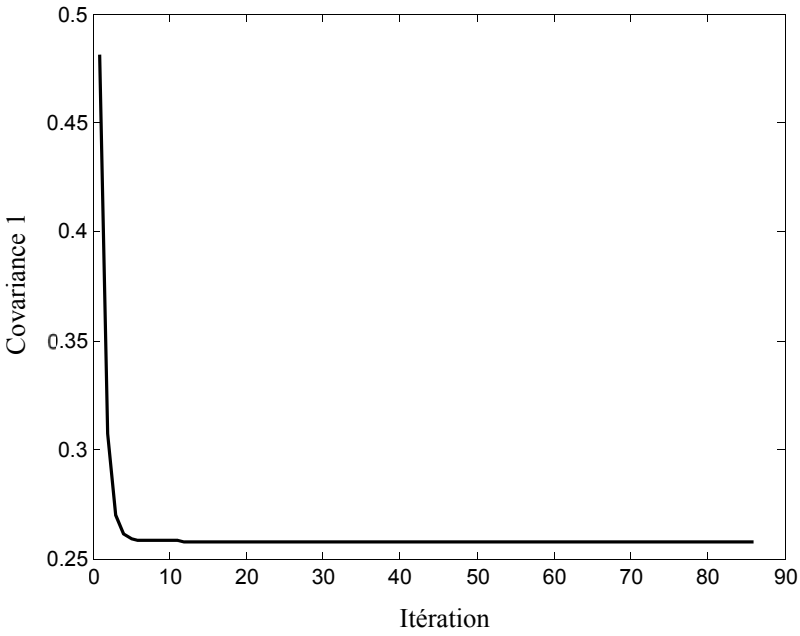


Figure III. 21. Convergence pour $(\sigma_{11}^2)_{f1}$ et $(\sigma_{22}^2)_{f1}$, éléments diagonaux de la matrice de covariance pour la première PDF du modèle mixte

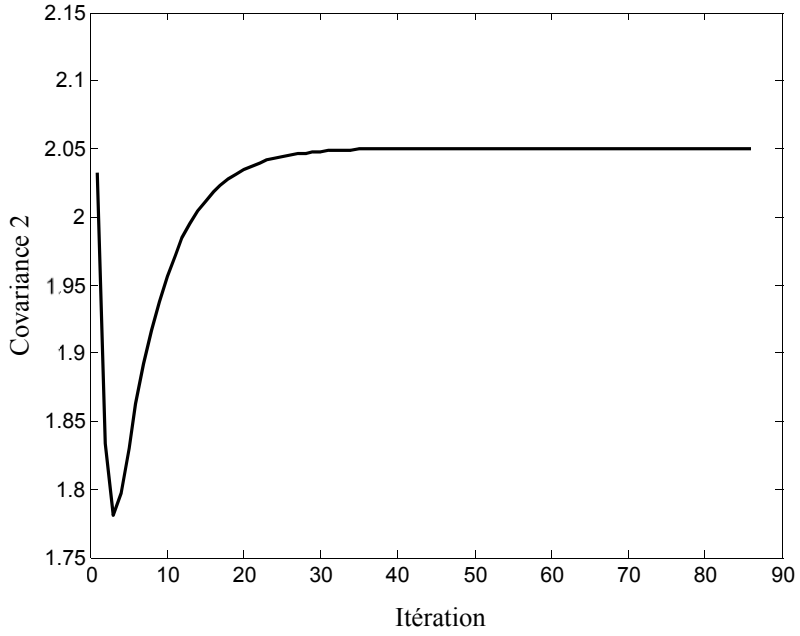


Figure III. 22. Convergence pour $(\sigma_{11}^2)_{f2}$ et $(\sigma_{22}^2)_{f2}$, éléments diagonaux de la matrice de covariance pour la deuxième PDF du modèle mixte

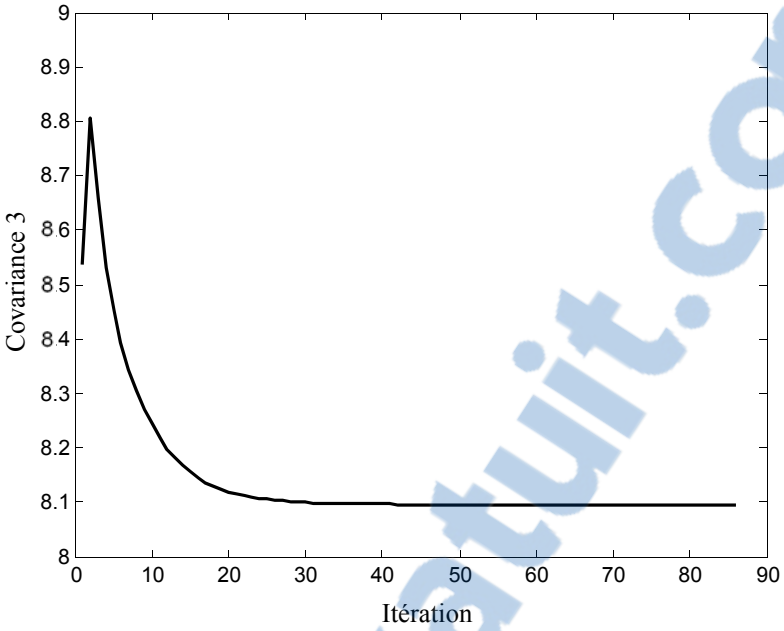


Figure III. 23. Convergence pour $(\sigma_{11}^2)_{f3}$ et $(\sigma_{22}^2)_{f3}$, éléments diagonaux de la matrice de covariance pour la troisième PDF du modèle mixte

En dernier et sur la figure III. 24, nous représentons l'évolution de l'erreur relative, basée sur l'équation (III. 8), à travers les itérations. L'évolution est exponentielle et confirme bien la convergence vers une erreur minimale.

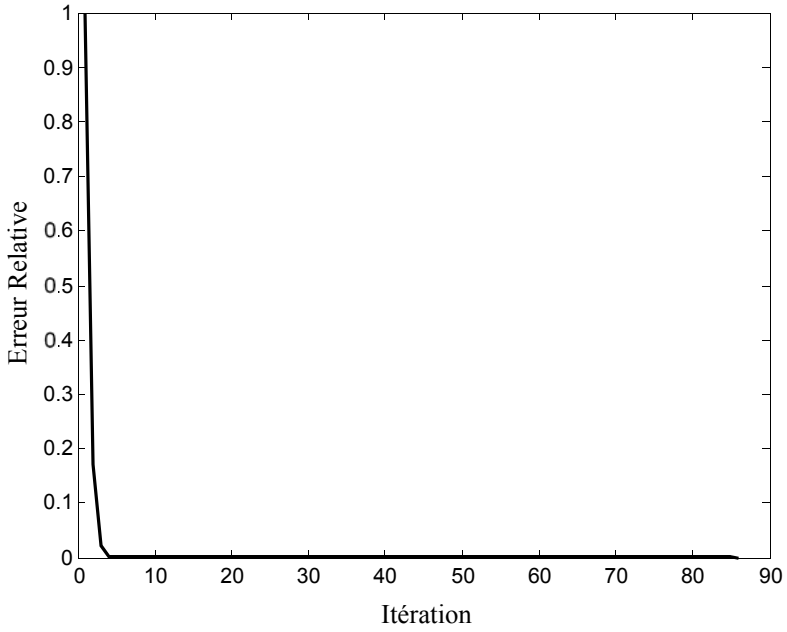


Figure III. 24. Convergence pour l'erreur relative

Pour cette première application, et bien que les valeurs initiales injectées pour le recouvrement des paramètres optimaux (poids, moyennes et covariances) soient très différentes des valeurs injectées pour la création des échantillons, mais les valeurs optimisées au final, par l'algorithme d'espérance-maximisation, sont très proches des valeurs créatrices des échantillons.

Dans le souci de valider encore plus notre module de calcul, nous donnons ci-après le résultat d'une deuxième simulation réalisée à base de données différentes. Il va sans dire que le nombre d'applications reste infini, mais nous avons juste rapporté deux d'entre-elles. Toutefois, le module de calcul et sur ces deux parties :

1- partie création d'échantillons obéissant au modèle mixte, et

2- partie estimation des paramètres optimaux pour le modèle mixte

fonctionne parfaitement et les résultats obtenus sont très concluants.

III. 6. 2. Deuxième application

Dans cette deuxième application, nous allons nous contenter des valeurs des paramètres optimaux pour le modèle mixte gaussien, à savoir les moyennes optimales, les matrices de covariance optimales ainsi que les poids optimaux.

Tout d'abords, et comme d'habitude nous allons générer, grâce à notre module de calcul, N échantillons à L-dimensions appartenant à K classes différentes, avec cette fois-ci :

N = 500 : nombre d'échantillons.

L = 3 : nombre de dimensions.

K = 4 : nombre de classes.

III. 6. 2. 1. Génération des échantillons

Comme pour la première application, nous allons générer N échantillons en L-dimensions appartenant à K classes différentes. Les vecteurs de moyennes, les matrices de covariances et les poids de pondération pour le modèle mixte seront comme suit :

Moyennes :

$\mu_1 = [\mu_{11}, \mu_{12}, \mu_{13}] = [1, 3, 2]$,

$\mu_2 = [\mu_{21}, \mu_{22}, \mu_{23}] = [3, 6, 4]$,

$\mu_3 = [\mu_{31}, \mu_{32}, \mu_{33}] = [2.5, 3.5, 3]$,

$\mu_4 = [\mu_{41}, \mu_{42}, \mu_{43}] = [1, 4, 2]$.

Matrices de covariances :

$$\sum_1 = \begin{bmatrix} (\sigma_{11}^2)_1 & (\sigma_{12}^2)_1 & (\sigma_{13}^2)_1 \\ (\sigma_{21}^2)_1 & (\sigma_{22}^2)_1 & (\sigma_{23}^2)_1 \\ (\sigma_{31}^2)_1 & (\sigma_{32}^2)_1 & (\sigma_{33}^2)_1 \end{bmatrix} = \begin{bmatrix} 0.25 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}$$

$$\sum_2 = \begin{bmatrix} (\sigma_{11}^2)_2 & (\sigma_{12}^2)_2 & (\sigma_{13}^2)_2 \\ (\sigma_{21}^2)_2 & (\sigma_{22}^2)_2 & (\sigma_{23}^2)_2 \\ (\sigma_{31}^2)_2 & (\sigma_{32}^2)_2 & (\sigma_{33}^2)_2 \end{bmatrix} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

$$\sum_3 = \begin{bmatrix} (\sigma_{11}^2)_3 & (\sigma_{12}^2)_3 & (\sigma_{13}^2)_3 \\ (\sigma_{21}^2)_3 & (\sigma_{22}^2)_3 & (\sigma_{23}^2)_3 \\ (\sigma_{31}^2)_3 & (\sigma_{32}^2)_3 & (\sigma_{33}^2)_3 \end{bmatrix} = \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 1.5 \end{bmatrix}$$

$$\sum_4 = \begin{bmatrix} (\sigma_{11}^2)_4 & (\sigma_{12}^2)_4 & (\sigma_{13}^2)_4 \\ (\sigma_{21}^2)_4 & (\sigma_{22}^2)_4 & (\sigma_{23}^2)_4 \\ (\sigma_{31}^2)_4 & (\sigma_{32}^2)_4 & (\sigma_{33}^2)_4 \end{bmatrix} = \begin{bmatrix} 2.5 & 0 & 0 \\ 0 & 2.5 & 0 \\ 0 & 0 & 2.5 \end{bmatrix}$$

Poids :

$P_1 = 0.15$, $P_2 = 0.4$, $P_3 = 0.25$, $P_4 = 0.2$

III. 6. 2. 2. Estimation des statistiques et poids optimaux

Pour la phase estimation et pour le lancement de la simulation, nous donnerons des valeurs initiales aléatoires et ce pour les moyennes, les covariances et les probabilités a priori :

Moyennes initiales :

$\mu_{i1} = [1, 2.2, 1.3]$,

$\mu_{i2} = [4.5, 4.4, 2]$,

$\mu_{i3} = [2.4, 2.3, 2.2]$,

$\mu_{i4} = [1.6, 3.2, 1.4]$.

Matrices de covariances initiales :

$$\sum_{i1} = \begin{bmatrix} (\sigma_{11}^2)_{i1} & (\sigma_{12}^2)_{i1} & (\sigma_{13}^2)_{i1} \\ (\sigma_{21}^2)_{i1} & (\sigma_{22}^2)_{i1} & (\sigma_{23}^2)_{i1} \\ (\sigma_{31}^2)_{i1} & (\sigma_{32}^2)_{i1} & (\sigma_{33}^2)_{i1} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\sum_{i2} = \begin{bmatrix} (\sigma_{11}^2)_{i2} & (\sigma_{12}^2)_{i2} & (\sigma_{13}^2)_{i2} \\ (\sigma_{21}^2)_{i2} & (\sigma_{22}^2)_{i2} & (\sigma_{23}^2)_{i2} \\ (\sigma_{31}^2)_{i2} & (\sigma_{32}^2)_{i2} & (\sigma_{33}^2)_{i2} \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

$$\sum_{i3} = \begin{bmatrix} (\sigma_{11}^2)_{i3} & (\sigma_{12}^2)_{i3} & (\sigma_{13}^2)_{i3} \\ (\sigma_{21}^2)_{i3} & (\sigma_{22}^2)_{i3} & (\sigma_{23}^2)_{i3} \\ (\sigma_{31}^2)_{i3} & (\sigma_{32}^2)_{i3} & (\sigma_{33}^2)_{i3} \end{bmatrix} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

$$\sum_{i4} = \begin{bmatrix} (\sigma_{11}^2)_{i4} & (\sigma_{12}^2)_{i4} & (\sigma_{13}^2)_{i4} \\ (\sigma_{21}^2)_{i4} & (\sigma_{22}^2)_{i4} & (\sigma_{23}^2)_{i4} \\ (\sigma_{31}^2)_{i4} & (\sigma_{32}^2)_{i4} & (\sigma_{33}^2)_{i4} \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

Poids initiaux :

$P_{i1} = 0.2$, $P_{i2} = 0.15$, $P_{i3} = 0.15$, $P_{i4} = 0.5$

Erreur relative minimale exigée : 10^{-15}

Simulation

Pour l'erreur relative minimale exigée, nous avons accomplie 221 itérations et nous avons trouvé les valeurs optimales suivantes :

Moyennes optimales :

$\mu_{f1} = [\mu_{f11}, \mu_{f12}, \mu_{f13}] = [0.9651, 3.0698, 2.0015]$,

$\mu_{f2} = [\mu_{f21}, \mu_{f22}, \mu_{f23}] = [2.9810, 5.9681, 4.0656]$,

$\mu_{f3} = [\mu_{f31}, \mu_{f32}, \mu_{f33}] = [2.6699, 3.5192, 3.1340]$,

$\mu_{f4} = [\mu_{f41}, \mu_{f42}, \mu_{f43}] = [0.6001, 3.7974, 1.9846]$.

Matrices de covariances optimales :

$$\sum_{f1} = \begin{bmatrix} (\sigma_{11}^2)_{f1} & (\sigma_{12}^2)_{f1} & (\sigma_{13}^2)_{f1} \\ (\sigma_{21}^2)_{f1} & (\sigma_{22}^2)_{f1} & (\sigma_{23}^2)_{f1} \\ (\sigma_{31}^2)_{f1} & (\sigma_{32}^2)_{f1} & (\sigma_{33}^2)_{f1} \end{bmatrix} = \begin{bmatrix} 0.2809 & 0 & 0 \\ 0 & 0.2809 & 0 \\ 0 & 0 & 0.2809 \end{bmatrix}$$

$$\sum_{f2} = \begin{bmatrix} (\sigma_{11}^2)_{f2} & (\sigma_{12}^2)_{f2} & (\sigma_{13}^2)_{f2} \\ (\sigma_{21}^2)_{f2} & (\sigma_{22}^2)_{f2} & (\sigma_{23}^2)_{f2} \\ (\sigma_{31}^2)_{f2} & (\sigma_{32}^2)_{f2} & (\sigma_{33}^2)_{f2} \end{bmatrix} = \begin{bmatrix} 0.5190 & 0 & 0 \\ 0 & 0.5190 & 0 \\ 0 & 0 & 0.5190 \end{bmatrix}$$

$$\sum_{f3} = \begin{bmatrix} (\sigma_{11}^2)_{f3} & (\sigma_{12}^2)_{f3} & (\sigma_{13}^2)_{f3} \\ (\sigma_{21}^2)_{f3} & (\sigma_{22}^2)_{f3} & (\sigma_{23}^2)_{f3} \\ (\sigma_{31}^2)_{f3} & (\sigma_{32}^2)_{f3} & (\sigma_{33}^2)_{f3} \end{bmatrix} = \begin{bmatrix} 1.5170 & 0 & 0 \\ 0 & 1.5170 & 0 \\ 0 & 0 & 1.5170 \end{bmatrix}$$

$$\sum_{f4} = \begin{bmatrix} (\sigma_{11}^2)_{f4} & (\sigma_{12}^2)_{f4} & (\sigma_{13}^2)_{f4} \\ (\sigma_{21}^2)_{f4} & (\sigma_{22}^2)_{f4} & (\sigma_{23}^2)_{f4} \\ (\sigma_{31}^2)_{f4} & (\sigma_{32}^2)_{f4} & (\sigma_{33}^2)_{f4} \end{bmatrix} = \begin{bmatrix} 2.6349 & 0 & 0 \\ 0 & 2.6349 & 0 \\ 0 & 0 & 2.6349 \end{bmatrix}$$

Poids optimaux :

$P_{f1} = 0.1662$, $P_{f2} = 0.3999$, $P_{f3} = 0.2507$, $P_{f4} = 0.1832$.

Sur l'ensemble des valeurs calculées et en moyenne, ces valeurs sont très proches des valeurs injectées pour la création des échantillons. Cela dit que notre module de calcul arrive bien à optimiser les paramètres du modèle mixte et permettra ainsi par la suite de classifier chaque échantillon dans la classe qui lui est proche. Nous entendons par proche, le maximum de vraisemblance et une distance minimale.

Il reste à noter que :
Les échantillons générés sont aléatoires et que de ce fait, les statistiques utilisées pour leur créations seront légèrement différentes des statistiques qui en découlent. De plus, cette différence n'est pas constante, c'est-à-dire qu'après chaque simulation, cette différence change, d'où une complication supplémentaire. Tout çà pour dire que deux simulations consécutives ou pas, ne donnerons jamais le même résultat mais leurs résultats resteront très proches.
Une deuxième remarque, concerne cette fois-ci le problème des optima locaux. Il est vrai que pour les problèmes d'optimisation, le choix de l'estimé initial influe sur le résultat final. Donc si le résultat est piégé sur un optimum local, il le restera. D'où la différence entre résultats partants d'estimés initiaux différents. Le choix de l'estimé initial qui mènera à l'optimum global est totalement hasardeux sinon d'autres approches pourrons être utilisées pour sa détermination.

III. 7. Conclusion

Au chapitre trois, nous avons parlé du modèle mixte gaussien et rappelé les étapes de l'algorithme d'espérance-maximisation. Au fait, le modèle mixte est multimodal et peut être approché par des gaussiennes pondérées par des poids ou probabilités a priori. Le modèle mixte concerne les phénomènes dont la distribution n'est pas gaussienne. Cela n'empêche pas son expansion en gaussiennes pondérées afin de faciliter les calculs et profiter des caractéristiques des gaussiennes. Le nombre de gaussiennes dans son expansion est le nombre de classes qui le constituent.
Les échantillons crées à base du modèle mixte seront mélangés et la classification consiste à ramener chaque échantillon à la classe qui vraisemblablement l'a créé.
Maintenant, si nous disposons d'échantillons, alors il faudra déterminer les gaussiennes qui les ont vraisemblablement crées. Déterminer les gaussiennes revient

à déterminer leurs statistiques, à savoir les moyennes et les covariances. En plus nous devrons déterminer leurs proportions, c'est-à-dire les poids ou probabilités a priori. Cette tâche sera réalisée par le biais de l'algorithme d'espérance-maximisation. Cet algorithme est un algorithme itératif partant d'estimés initiaux pour les paramètres des gaussiennes, et enchaine les opérations dans un ordre bien établie comprenant l'étape espérance E et l'étape maximisation M. Cet enchainement aboutira après un certain nombre d'itérations, nécessaire pour la précision, aux valeurs optimales qui ne seront jamais égales aux valeurs injectées pour la création des échantillons. Cette inégalité est due au caractère aléatoire des échantillons et au problème des optima locaux.

Nous avons réalisé un module de calcul composé de deux parties. Une première partie pour la génération des échantillons découlant du model mixte. Une deuxième partie pour l'estimation des paramètres du modèle mixte qui a généré ces mêmes échantillons.

Les résultats des différentes simulations révèlent une excellente concordance entre les paramètres a priori, c'est-à-dire les paramètres de création du modèle mixte dans la première partie et les paramètres a posteriori, c'est-à-dire les paramètres probables et estimés du modèle. Toutefois, une différence substantielle existe à cause toujours du caractère aléatoire des échantillons et de celui des optima locaux.

Nous avons donc validé notre module de calcul par deux applications différentes en nombre d'échantillons, de dimensions et de classes. Ce même module servira sans aucun doute dans une suite concernant la classification de formes, qu'elles soient signaux, images, ou toutes autres entités mesurables.

Conclusion Générale

Conclusion générale

Conclusion générale

Il n'existe pas de méthode universelle de reconnaissance de formes. Le développement de systèmes de reconnaissance de formes, est le sujet de recherches actives qui mettent en jeu des connaissances extrêmement diverses. Informatique, traitement du signal, statistiques, intelligence artificielle, et mathématiques ne sont que quelques-unes des disciplines impliquées dans la conception de systèmes de reconnaissance de formes. Chacun des acteurs de ces domaines de recherche a sa propre conception de la reconnaissance de formes. Il en résulte un foisonnement de méthodes qui, malgré d'apparentes oppositions, possèdent des propriétés communes.

La classification de formes consiste à classer les objets dans une catégorie donnée appelée classe. Pour un problème spécifique de classification de formes, un classificateur, qui est un programme informatique, est développé de sorte que les objets soient classifiés correctement avec une précision raisonnablement bonne.
Les entrées du classificateur sont les caractéristiques qui sont déterminées de manière à ce qu'elles représentent bien chaque classe ou encore à ce que les données appartenant à différentes classes soient bien séparées.

Nous avons voulu à travers cette modeste étude et par le biais de ce mémoire, contribuer à l'engouement pour la reconnaissance de forme et la classification. Ceci a pris forme dans le chapitre premier, où nous avons présenté quelques définitions de base rencontrées dans la classification. Ces définitions ont concerné entre-autres : les caractéristiques, la décision, la probabilité d'erreur sur la décision, etc.

Au chapitre deux, nous avons poussé un peu plus notre intérêt pour aborder enfin les types de classificateurs ainsi que les techniques d'estimation des fonctions de densité de probabilité inconnues. Nous avons aussi parlé de la classification bayésienne et de l'algorithme Espérance-Maximisation qui permet de retrouver, par le maximum de vraisemblance, les paramètres optimaux des modèles mixtes. Cet algorithme constitue le lien avec le chapitre trois auquel nous avons attribué la partie résultats obtenus et ce après simulation sur notre module de calcul fondé sur l'algorithme EM.

En effet au chapitre trois nous avons bien détaillé les étapes de cet algorithme itératif, à savoir l'étape E ou d'espérance et l'étape M ou de maximisation.

Nous avons donné les expressions itératives qui mettent à jour les paramètres optimaux du modèle mixte. Il s'agit des probabilités a priori, du vecteurs des moyennes et des matrices de covariance. Par la suite, nous avons testé notre module de calcul à travers deux applications qui, rappelons-le, ont donné des résultats très probants, malgré le caractère aléatoire des échantillons générés ainsi que le problème des optima locaux.

Nous avons tracé la convergence des paramètres optimaux que nous avons ensuite comparé avec les paramètres injectés pour la création des données elles-mêmes. La comparaison, ne laisse aucun doute, notre module est bien capable d'atteindre les solutions optimales.

Nous projetons comme perspectives d'utiliser ce même module de calcul pour la classification de formes supervisée, où le système connaît les classes auxquelles appartiennent les échantillons, et non supervisée où le système ne connaît pas ces classes.

Références

Références

Références

[1] Bernardo J.M., Smith A.F.M, Bayesian Theory, *John Wiley*, 1994.

[2] Bishop C.M., Pattern Recognition and Machine Learning, *Springer*, 2006.

[3] Neapolitan R.D., Learning Bayesian Networks, *Prentice Hall*, 2004.

[4] Friedman N.,Geiger D.,Goldszmidt M., "Bayesian network classifiers", *Machine Learning*, Vol. 29, pp. 131–163, 1997.

[5] Cramer H., Mathematical Methods of Statistics, *Princeton University Press*, 1941.

[6] Krzyzak A., "Classification procedures using multivariate variable kernel density estimate", *Pattern Recognition Letters*, Vol. 1, pp. 293–298, 1983.

[7] McLachlan G.J., Basford K.A., Mixture Models: Inference and Applications to Clustering, *Marcel Dekker*, 1988.

[8] Cooper G.F., "The computational complexity of probabilistic inference using Bayesian belief networks", *Artificial Intelligence*, Vol. 42, pp. 393–405, 1990.

[9] Aeberhard S., Coomans D., Devel O., "Comparative analysis of statistical pattern recognition methods in high dimensional setting", *Pattern Recognition*, Vol. 27, No. 8, pp. 1065–1077, 1994.

[10] Breiman L., Meisel W., Purcell E., "Variable kernel estimates of multivariate densities" *Technometrics*, Vol. 19, No. 2, pp. 135–144, 1977.

[11] J. A. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", *Technical Report, University of Berkeley*, 1998.

[12] Moon T., "The expectation maximization algorithm", *Signal Processing Magazine*, Vol. 13, No. 6, pp. 47–60, 1996.

[13] Redner R.A., Walker H.F., "Mixture densities, maximum likelihood and the EM algorithm", *SIAM Review*, Vol. 26, No. 2, pp. 195–239, 1984.

[14] Hoffbeck J.P., Landgrebe D.A., "Covariance matrix estimation and classification with limited training data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 7, pp. 763–767, 1996.

[15] Dempster A.P., Laird N.M., Rubin D.B., "Maximum likelihood from incomplete data via the EM algorithm", *Journal of Royal Statistical Society*, Vol. 39, No. 1, pp. 1–38, 1977.

[16] Boyles R.A., "On the convergence of the EM algorithm", *Journal of Royal Statistical Society*, Vol. 45, No. 1, pp. 47–55, 1983.