Master TRIED

ETUDE DE CAS 2

RAPPORT DE PROJET

Sujet :

Classification des signatures par méthodes de K-means et de GMM

Réalisé par :

Karim ASSAAD

Bilal DIAB

Année universitaire : 2016-2017

Table de Contenu

Introduction

La vérification de l'identité est très importante dans la vie quotidienne. Il y a deux manières classiques de vérifier l'identité d'un individu. L'une est basée sur une connaissance (par exemple le mot de passe), et l'autre c'est la biométrie. La 2eme méthode évite la perte, le vol ou l'oublie des informations, contrairement aux systèmes standards de control d'accès, mais ils présentent des limites de point de vue couts et droits. La biométrie est ainsi reliée à la personne, très difficile à usurper et elle ne peut jamais être perdue.

C'est pour ces raisons que la reconnaissance des formes est devenue une des disciplines les plus importantes en informatique. C'est un ensemble de techniques et de méthodes visant à identifier des formes visuelles (empreinte digitale, visage, iris, signature) et qui sont utilisés surtout dans le secteur de sécurité. Cependant, les caractéristiques biométriques dépendent beaucoup de l'environnement de capture, ce qui affecte le bon fonctionnement du système de vérification biométrique. C'est donc ici qu'interviennent les systèmes de vérification.

Objectif et base de données

Le but de cette étude est d'analyser des signatures produites 100 personnes. Ces personnes ont chacun effectué 25 signatures réelles donc nous avons aux totales 2500 signature qui sera la base de données sur laquelle nous allons effectuer nos différentes études. Ces analyses suivant deux approches différentes :

- la première en appliquant des méthodes de classifications sur les 2500 signatures afin de déterminer 3 groupes de signatures mais aussi sur les 100 individus et ceci avec plusieurs méthodes de classification.
- la deuxième en évaluant la performance des différentes méthodes utilisées avec plusieurs paramètres et ceci afin de déterminer la meilleur méthode avec les meilleurs paramètres.

Méthode proposé :

La première étape qui est l'étape de classification des signatures se fait avec différentes méthodes. Le but étant de classer les 2500 signatures et les 25 personnes sur les 3 groupes et ceci avec les deux méthodes suivantes :

- La méthode des K-moyennes qui est une méthode de segmentation des individus en se basant sur les distances minimales entre les individus et les centres des clusters.
- La méthode de GMM (Gaussian Mixture model) qui est un modèle statistique exprimé selon une densité mélangée. Il sert essentiellement à estimer paramétriquement la distribution de variables aléatoires en les modélisant comme une somme de plusieurs gaussiennes. Il s'agit alors de déterminer la moyenne, la variance et l'amplitude de chaque gaussienne.

Ces étapes de classification des signatures et des personnes se font pour différents de nombres de gaussien :

- 4 gaussiennes
- 8 gaussiennes
- 24 gaussiennes

Pour chacun de ces cas, nous avons effectué les méthodes cités ci-dessus et nous allons voir, dans ce qui suit, l'effet de la complexité sur les résultats de la classification.

Mesure de complexité

On s'intéresse dans cette phase à mesurer la complexité des signatures authentiques associées aux 100 individus en appliquant le script de « complexity_GMM ».

La complexité des signatures varie selon le nombre de gaussiennes et diffère légèrement d'un individu à un autre. En effet on remarque différentes valeurs de la complexité pour différent nombre de gaussienne égale à 4, 8 et 24.

Un autre point de vue intéressant est celui de considérer la distance entre les valeurs de complexités de deux individus, celle-ci est petite dans le cas de 4 gaussiennes, légèrement plus grande dans le cas de 8 gaussiennes et dans le cas de 24 gaussiennes donc un lien existe entre le nombre de gaussienne et la variance intra classe des individus (on montrera ceci dans les résultats ci-dessous).

Clustering

On procède dans cette phase à la catégorisation des individus selon la complexité des signatures en ceci en utilisant deux méthodes : k-means et GMM. Chacune de ces deux méthodes vont être appliqués tout d'abord sur les individus afin de faire des classes d'individus ensuite sur les signature. Une vérification va être ensuite faite sur les classifications des signatures pour voir si toutes les signatures d'un même individu appartiennent à une même classe ou non.

Définition du clustering

Le clustering a pour but de classer les données, il est considéré comme une méthode statistique utile à l'analyse des données. Le principe consiste à la répartition des données en différents ensembles ayant des caractéristiques communes.

Classification des individus

On applique tout d'abord la méthode des k-means sur les individus et on choisit au préalable 3 classes.

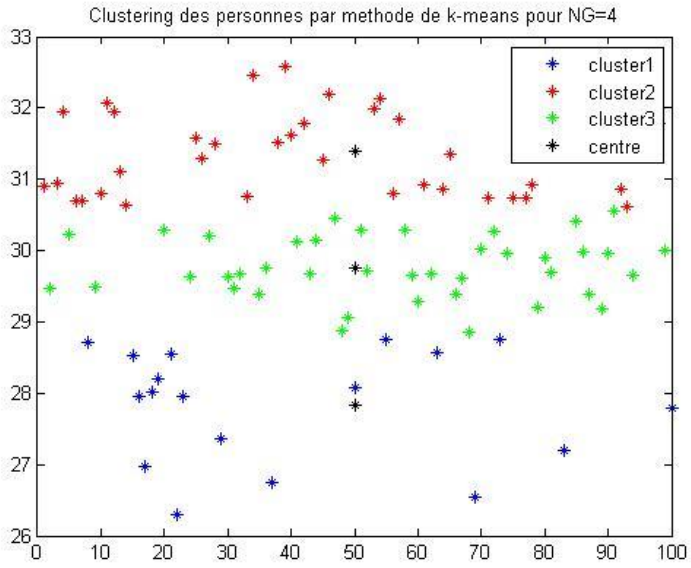


Figure 1 : k-means sur les individus dans le cas de 4 gaussiennes

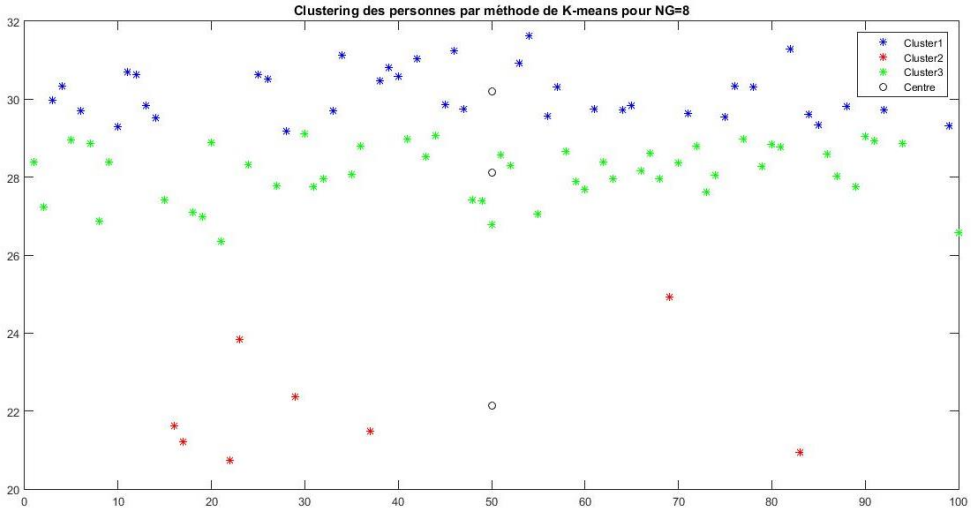


Figure 2 : k-means sur les individus dans le cas de 8 gaussiennes

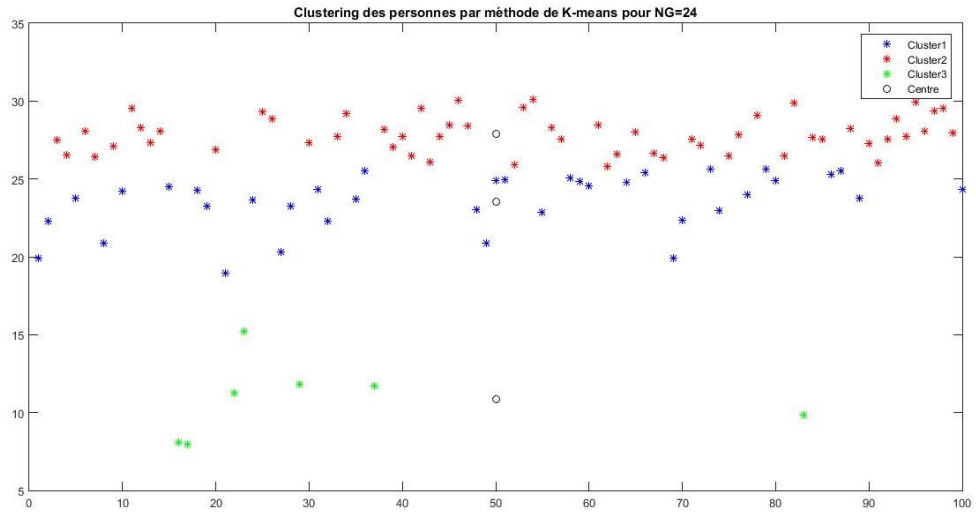


Figure 3 : k-means sur les individus dans le cas de 24 gaussiennes

D'après les deux dernières figures on remarque que les variances inter et intra classes se diffèrent : on remarque que plus on augmente le nombre de gaussiennes plus c'est facile de distinguer les surfaces de séparation entre les classes (variance inter classes augmente). De plus on remarque que l'augmentation du nombre de gaussiennes augmente la dispersion des individus dans chaque classe (variance intra classe augmente). Ceci peut être aussi vérifié par le tableau de variances suivantes :

NG	Complexité haute	Complexité moyenne	Complexité faible	Inter Classe
4	0,56	0,42	0,79	1,78
8	0,60	0,70	1,40	4,10
24	1,14	1,79	2,49	8,85

Tableau 1: Variances des classes dans le cas des k-means

Dans le tableau nous avons calculé les variances inter et intra classes selon le nombre de gaussiennes. Ce tableau montre l'augmentation de la variance inter classes en augmentant me nombre de gaussiennes. Donc un grand nombre de gaussiennes nous permet d'avoir une séparation des classes plus claire.

Cela vérifie ce qu'on a constaté avant dans les graphes.

A la fin de la classification des k-means, nous avons décidé d'appliquer la méthode de GMM sur les trois cas, et nous avons obtenu les résultats suivant :

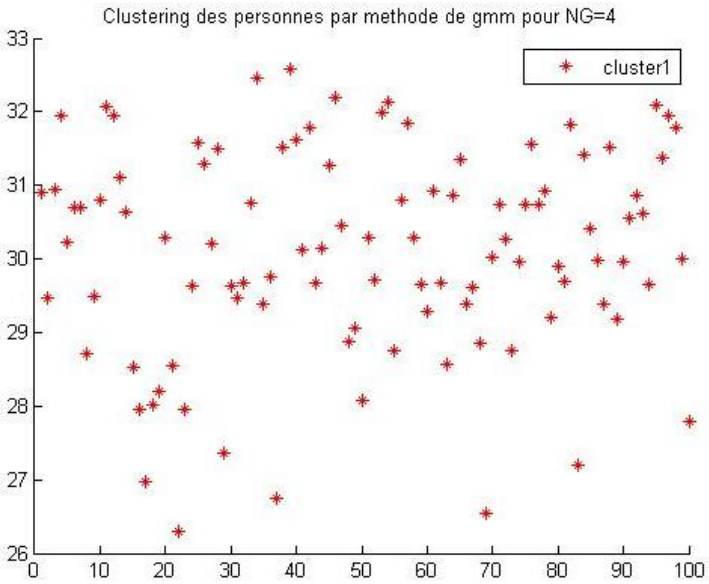


Figure 4: GMM sur les individus dans le cas de 4 gaussiennes

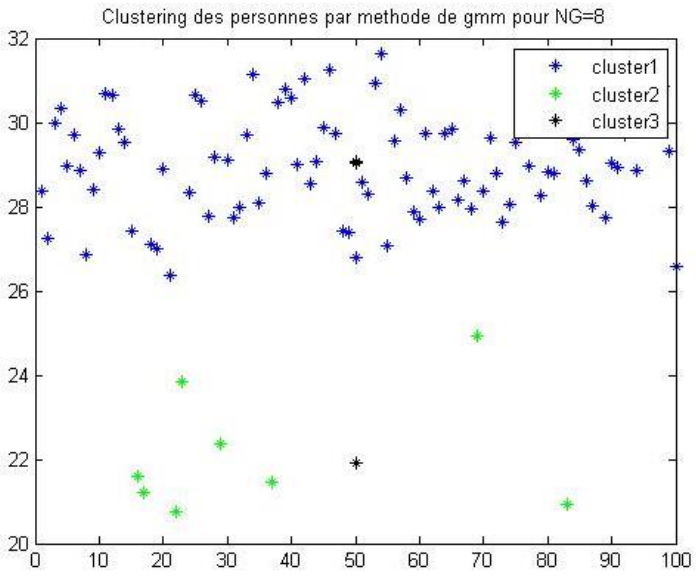


Figure 5: GMM sur les individus dans le cas de 8 gaussiennes

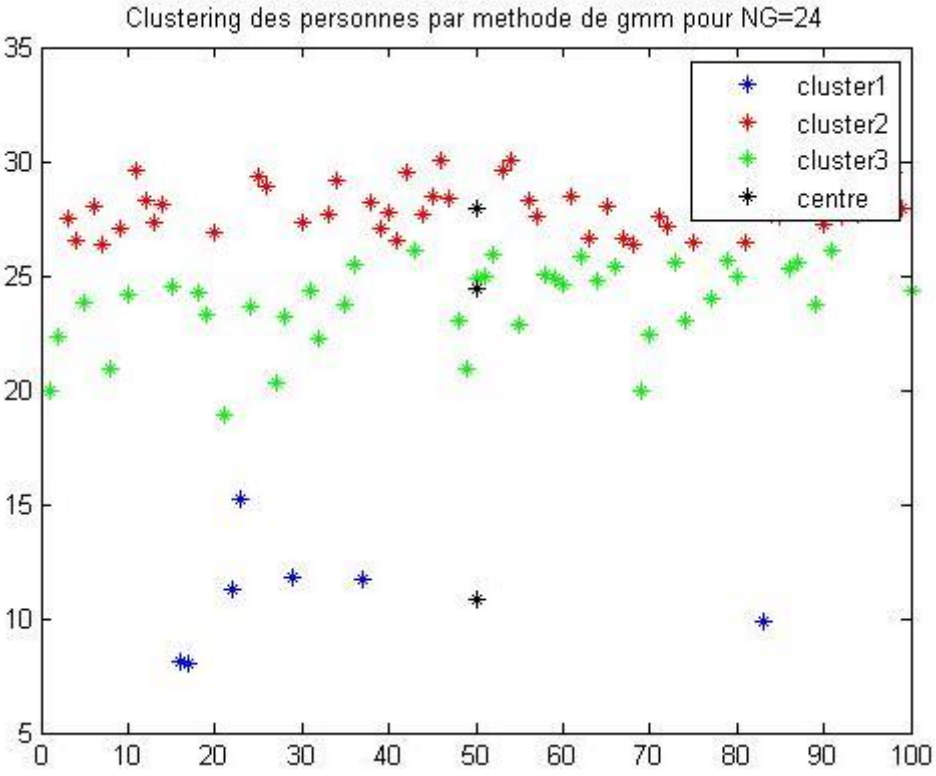


Figure 6: GMM sur les individus dans le cas de 24 gaussiennes

D'après ces trois graphes, nous remarquons que dans le cas des 4 gaussiennes, nous avons uniquement une classe, 2 classes dans le cas des 8 gaussiennes et 3 classes dans le cas des 24 gaussiennes. La conclusion est donc claire que le cas des 24

gaussiennes donne le meilleur résultat. Au tableau des variances montré ci-dessus, on ajoute les variances pour la méthode GMM, on obtient alors le tableau suivant :

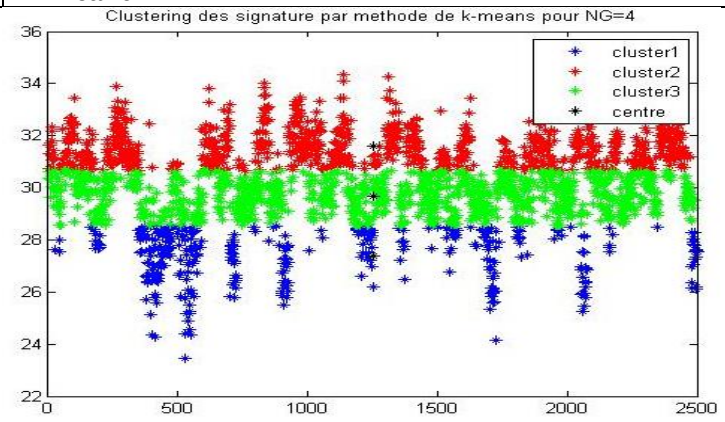
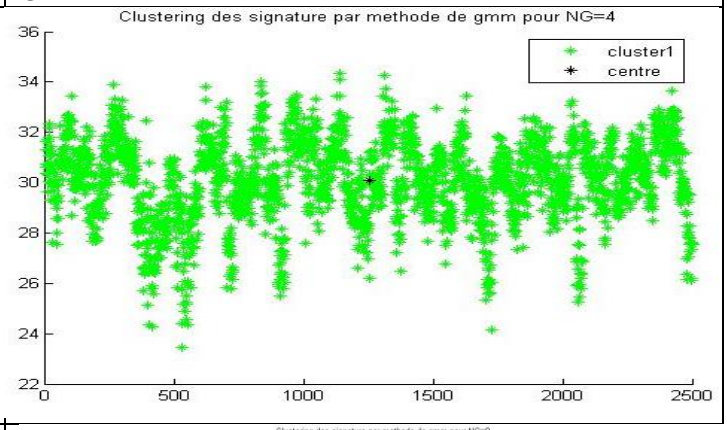
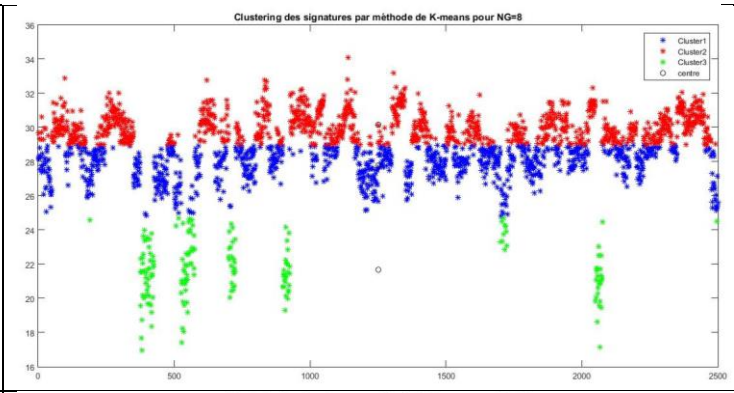
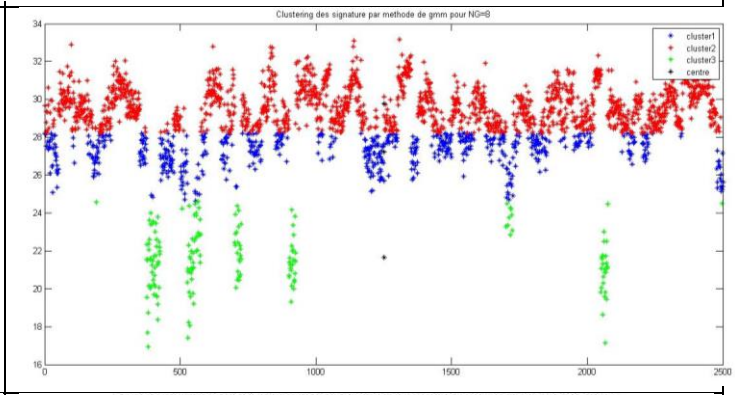
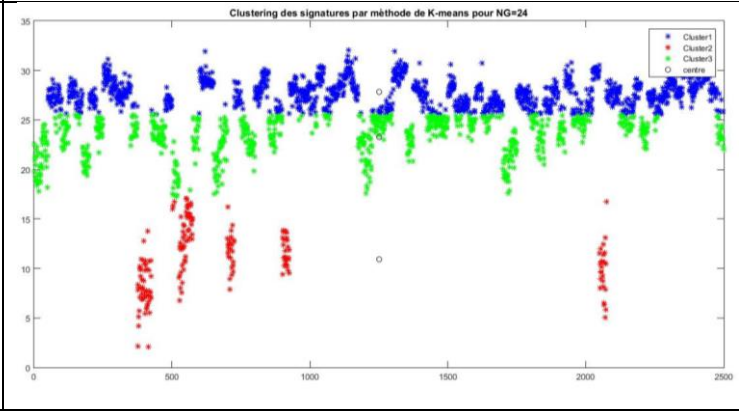
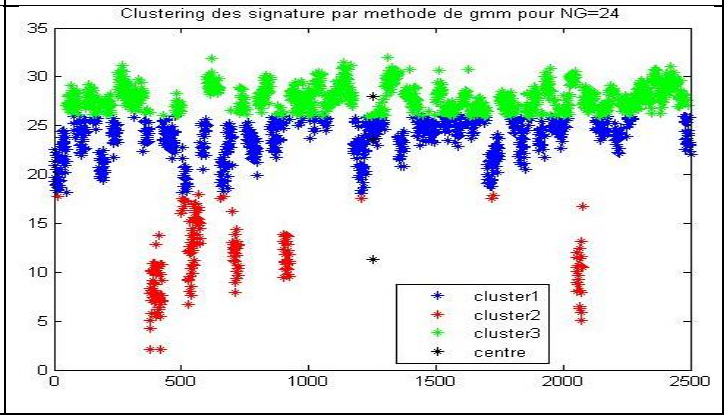
	NG	Complexité haute	Complexité moyenne	Complexité faible	Inter Classe
K-means	4	0,56	0,42	0,79	1,78
	8	0,60	0,70	1,40	4,10
	24	**1,14**	**1,79**	**2,49**	**8,85**
GMM	8	1,20	-	1,40	4,9
	24	**1**	**1,8**	**2,5**	**9**

Tableau 2: Variances des classes dans le cas des k-means et des GMM

Le résultat montré par le graphe est aussi validé par les calculs de variance, les plus grandes variances inter et intra classe sont obtenus pour le cas des 24 gaussiennes.

Classification des signatures

Nous avons dans un deuxième temps effectué des classifications sur les signatures afin de vérifier si les 25 signatures de chaque personne appartiennent à la même classe ou non. Nous avons appliquées les 2 méthodes de classifications sur les 3 cas de nombres de gaussiennes. On obtient les résultats suivants :

NG	K-means	GMM
4		
8		
24		

	NG	Complexité haute	Complexité moyenne	Complexité faible	Inter Classe
GMM	8	0,98	0,79	1,60	4,10
	24	**1,10**	**1,80**	**3,30**	**8,6**
K-means	4	0,69	0,57	0,98	2,10
	8	0,80	0,90	1,60	4,30
	24	**1,20**	**1,80**	**3,10**	**8,70**

Tableau 3: Variances des classes dans le cas des k-means et des GMM

Dans la classification des signatures, on remarque aussi que les variances inter-classes sont les plus en élevées dans le cas des 24 gaussiennes. Cela signifie que dans ce cas, la frontière de séparation des classes est plus claire et ceci se voit aussi sur les graphes surtout entre la classe des complexités faibles et celle des complexités moyenne. De plus les variances intra classes sont les plus élevés dans le cas des 24 gaussiennes et ceci est normal vu que l'augmentation du nombre de gaussiennes nous permet de détecter plus de détails dans la signature et donc chaque signature devient plus différente que les autres : les signatures seront donc plus éparpillés donc la variance intra augmente.

Ce qui nous intéresse le plus dans la classification des signatures est de voir si toutes les signatures d'une même personne appartiennent à la même classe. Nous avons fait le calcul du nombre de personnes dont toutes les signatures appartiennent à la même classe et ceci pour toutes les méthodes et cas présentés ci-dessus. On obtient le résultat suivant :

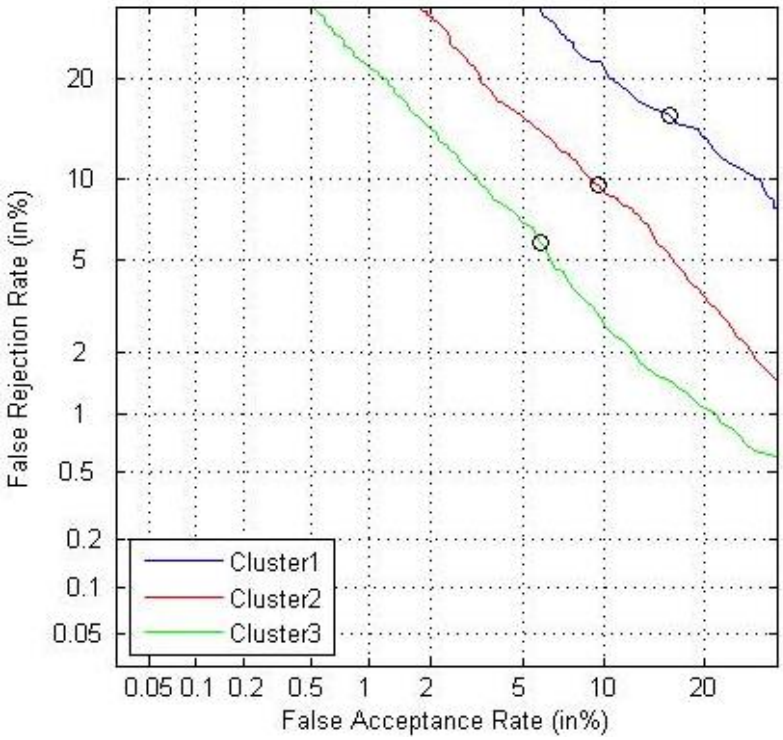
	NG	Nb individus dont signatures bien classées
K-means	4	9
	8	31
	24	**61**
GMM	8	46
	24	**59**

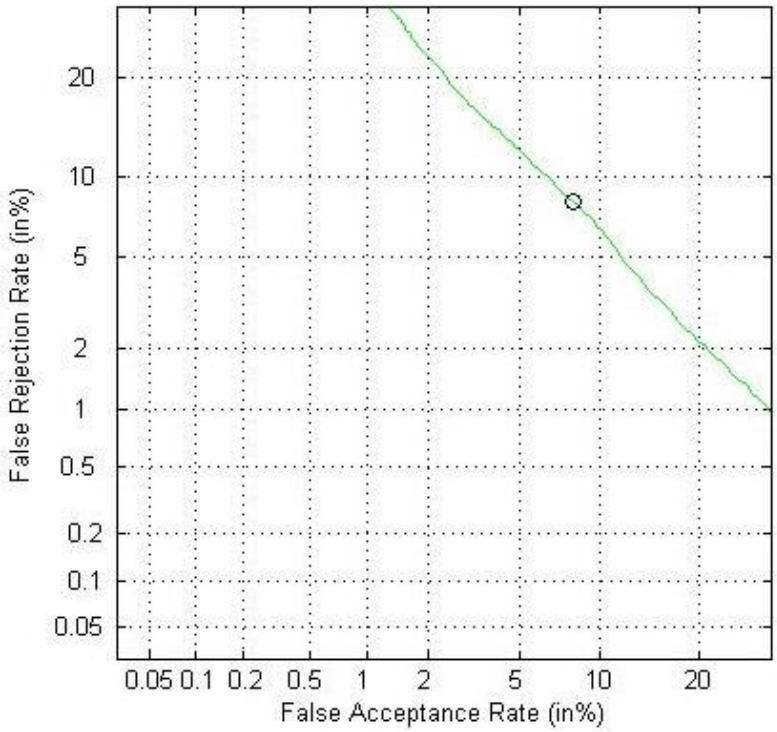
Tableau 4: Nombre d'individus dont les signatures sont bien classées

Le plus grand nombre de personnes dont les signatures appartiennent à la même classe est obtenu pour le cas des k-means appliquée dans le cas des 24 gaussiennes et très proche de celui obtenu pour le cas des GMM appliquée sur le même nombre de gaussiennes. Ce résultat est cohérent avec le résultat obtenu dans l'analyse des variances. On peut donc conclure que la méthode K-means dans le cas des 24 gaussiennes est la meilleure méthode de classification des signatures.

Performance

Nous nous sommes basés sur la courbe DET pour évaluer la performance des classifications. Nous avons tout d'abord tracé une courbe pour les 3 clusters ensemble puis une courbe pour chaque cluster. On obtient les résultats suivants :





D’après la première figure, on peut conclure que la courbe verte correspond à la classe des individus ayant la plus faible complexité. L’abscisse est donnée par une grande False Acceptation et l’ordonnée par une grande False Rejection. La courbe bleue correspond à la classe des individus ayant la plus grande complexité. L’abscisse est donnée par une petite False Acceptation et l’ordonnée par une petite False Rejection.

Le fait que la courbe du cluster des faibles complexités ait la meilleure performance s’explique par le fait que ce cluster est éloigné des deux autres clusters, donc les individus de ce cluster sont loin de la frontière de décision. Pour cela ils sont mieux classées que les individus des autres clusters.

La courbe rouge quant à elle prend en compte l’ensemble des classes des individus et présente la somme des trois courbes.

EER

Définition

L’EER (Taux d’erreur excessive) c’est la fréquence ou l’acceptante et la rejection sont égaux. La valeur de l’EER peut être obtenue facilement à partir du courbe DET. L’EER est une façon pour comparer la performance. En général, la courbe avec la plus petit EER a la meilleure performance.

Nous avons calculé le taux EER pour chacune des courbes DET de chacun des clusters et nous avons obtenu les résultats suivant :

	EER
Cluster 1 (complexité forte)	15.7
Cluster 2 (complexité moyenne)	9.4
Cluster 3 (complexité faible)	5.9

Ces résultats confirment l’analyse fait sur les graphes : le taux EER est le plus faible pour la courbe des faibles complexités et ceci puisque ce cluster est loin des autres donc pas trop de confusion avec les autres clusters. Pour les complexités fortes, le taux EER est le plus élevé et ceci s’explique par le fait que ce cluster est très proche du deuxième et donc le risque de confondre les classes est plus élevé.

Conclusion

La classification des personnes puis des signatures a été faites pour plusieurs cas afin de déterminer dans quels cas nous obtenons la meilleure classification. Dans un premier temps nous avons conclu que la classification des personnes est plus discriminante dans le cas des 24. Ensuite nous avons conclu que la méthode des k-means dans le cas des 24 gaussiennes est le meilleur