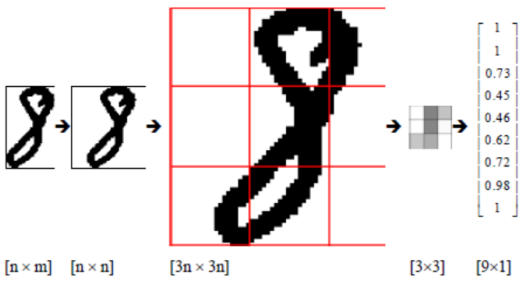
SYS800 Laboratoire 1
Représentation, extraction et sélection de caractéristiques

I. Prétraitement et préparation de données : extraction de caractéristiques

1) Pour extraire les caractéristiques des images de la base de données, nous allons utiliser le codage rétinien. Les images présentent des chiffres écris en noir (numériquement représenté par un 0) sur un fond blanc (numériquement représenté par un 1). Le codage rétinien va découper chaque image en différentes zones qui représenteront chacune une caractéristique de l'image : la proportion de 0 qu'elle contient.

Tout d'abord, pour que informations soit comparables, toutes les images seront rendues carré et les chiffres centrés dans leur image. Ensuite, la rétine découpera les images en n lignes et m colonnes de zones et calculera le pourcentage de pixel blanc dans chaque zone.

La figure suivante présente l'exemple d'une rétine 3x3 dont résulte un vecteur de 3*3 = 9 caractéristiques.



2) Dans un premier temps nous allons former des bases de données de caractéristiques pour différentes taille de rétine. Le but sera de commenter l'influence du nombre de caractéristiques sur les performances d'extraction de caractéristiques et la divergence entre les classes. Le tableau suivant inclut les tailles de rétine testée et le nombre de caractéristiques correspondant.

Rétine	2x2	2x5	5x5	10x10	15x10	20x10	15x15	20x15	20x20
Nb de caractéristiques	4	10	25	100	150	200	225	300	400

a) Pour calculer le taux de recouvrement, nous avons utilisé 3 métriques de similarité :

- La distance euclidienne
- La corrélation spectrale
- L'angle spectral

Nous avons aussi essayé d'utiliser la divergence d'information spectrale (SID), malheureusement les résultats n'étaient pas cohérent avec les trois autres métriques. Nous avons donc décidé de les écarter.

Pour chacun des éléments de la base de données, nous cherchons si l'élément le plus proche selon chaque métrique est de même classe (représente le même chiffre). Si ce n'est pas le cas, cette erreur représente un recouvrement des classes, le taux de recouvrement exprimé en % représente le taux d'erreurs sur toute la base de données. L'élément le plus proche est caractérisé par une courte distance euclidienne avec l'élément choisit, un faible angle spectral ou une grande corrélation spectrale.

Le tableau suivant regroupe les taux de recouvrement pour chaque rétine et chaque métrique :

Métrique	2x2	2x5	5x5	10x10	15x10	20x10	15x15	20x15	20x20
Distance	0.399	0.1473	0.0283	0.0202	0.0212	0.0212	0.0235	0.0235	0.0247
Corrélation	0.548	0.158	0.0297	0.019	0.0185	0.0187	0.0213	0.0207	0.0232
Angle	0.503	0.1475	0.03	0.0223	0.0233	0.024	0.0267	0.0272	0.0285

Nous remarquons que pour les rétines 2x2 et 2x5, le taux de recouvrement est très important et ne convient pas pour une classification efficace. Pour les autres rétines, le taux de recouvrement est inférieur à 3%, ce qui est plus satisfaisant, et globalement du même ordre de grandeur. On remarque qu'au sein du groupe de rétines testé, et bien que les rétines 15x10 et 20x10 donne de « meilleurs résultats » par la métrique de corrélation spectrale (0.05% et 0.03% de différence), la rétine 10x10 semble la mieux adapté à une application de classification. En effet elle présente deux des taux de recouvrement les plus faible et possède un nombre réduit de caractéristiques, 100, comparés aux autres rétines performantes. Nous considèrerons donc pour la suite cette rétine comme optimale et nous utiliserons la base de données qu'elle génère pour la suite de ce laboratoire.

b) Pour notre phase d'apprentissage, nous utiliseront donc les 100 caractéristiques, extraites par la rétine 10x10, de chaque image de la base de données d'apprentissage.

c) Afin d'évaluer la divergence entre les classes, soit notre capacité à les différencier deux à deux d'après les caractéristiques extraites, nous allons utiliser le critère de Fisher. Le critère de Fisher utilise les moyennes des classes ainsi que leur variance, c'est donc un moyen efficace d'évaluer le recouvrement de certaine classe entre elles. Le critère de Fisher définit la divergence de deux classes i et j pour une et une seule caractéristique avec les moyennes μ et les variances v de cette caractéristique au sein de chaque classe :

$$F_{ij} = \frac{|\mu_i - \mu_j|^2}{v_i^2 + v_j^2}$$

Finalement, nous calculons la moyenne des critères de Fisher de chaque caractéristique pour obtenir la divergence globale entre deux classes. La matrice suivante présente les divergences entre toutes les classes deux à deux. On remarque que cette matrice est symétrique, ce qui est logique puisque la divergence entre deux classes i et j est égale à la divergence entre les classes j et i. On note également que la diagonale est composée de 0 puisque la divergence entre une classe et elle-même est nulle.

	1	2	3	4	5	6	7	8	9	10
1	0	1.2723	0.3840	0.4814	0.7200	0.4001	0.4384	0.7474	0.5503	0.8173
2	1.2723	0	0.5253	0.6578	0.4936	0.3962	0.6323	0.4287	0.3584	0.4173
3	0.3840	0.5253	0	0.3506	0.4797	0.2889	0.2655	0.3577	0.2600	0.4877
4	0.4814	0.6578	0.3506	0	0.5222	0.3818	0.4833	0.3307	0.3122	0.3858
5	0.7200	0.4936	0.4797	0.5222	0	0.2555	0.5048	0.3404	0.2861	0.2913
6	0.4001	0.3962	0.2889	0.3818	0.2555	0	0.2946	0.3764	0.2025	0.3495
7	0.4384	0.6323	0.2655	0.4833	0.5048	0.2946	0	0.7000	0.3697	0.6414
8	0.7474	0.4287	0.3577	0.3307	0.3404	0.3764	0.7000	0	0.3115	0.1466
9	0.5503	0.3584	0.2600	0.3122	0.2861	0.2025	0.3697	0.3115	0	0.1821
10	0.8173	0.4173	0.4877	0.3858	0.2913	0.3495	0.6414	0.1466	0.1821	0

Figure 1; Matrice de divergence selon le critère de Fisher

On remarque d'après cette matrice que les classes qui divergent le plus sont :

- « 1 » et « 2 » avec $F_{12} = 1.27$
- « 1 » et « 0 » avec $F_{10} = 0.81$
- « 1 » et « 8 » avec $F_{18} = 0.74$

A l'inverse, les classes qui diverge le moins et qui sont donc plus difficile à différencier sont :

- « 0 » et « 9 » avec $F_{09} = 0.14$
- « 0 » et « 8 » avec $F_{08} = 0.18$
- « 6 » et « 9 » avec $F_{69} = 0.20$

On pouvait plus ou moins intuitivement prévoir ces résultats ou du moins nous pouvons les comprendre puisque par exemple, le chiffre « 1 » est celui qui en moyenne représente nettement moins de surface noire sur une image, il est donc facilement reconnaissable. On peut aussi noter la ressemblance entre des chiffres comme « 6 » et « 9 » que l'on retrouve ici sous forme de cette faible divergence.

Pour étudier la divergence globale, nous avons effectué la moyenne des différentes divergences entre les classes deux à deux (45 valeurs). Nous obtenons ainsi un critère de Fisher de0.44, en comparant avec les résultats précédents, on en conclut que la divergence globale est assez moyenne, sans être mauvaise. Nos caractéristiques sont donc à posteriori plutôt discriminantes.

II. Réduction de la dimension

Pour réduire la dimension de notre espace (10×10) nous pouvons utiliser deux méthodes :

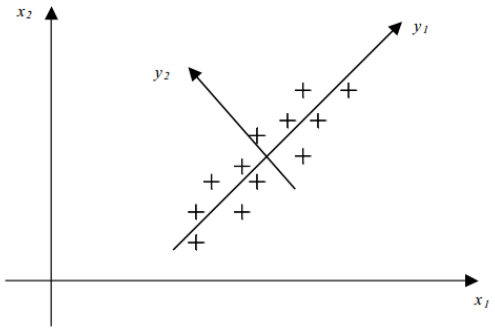
- Projection
- Sélection

La projection c'est garder les caractéristiques les plus précises. Quelles sont les dimensions qui donnent le plus d'information sur notre objet ? La sélection c'est plutôt choisir celles qui discriminent le mieux les caractéristiques. Quelles sont les meilleurs composantes parmi les n qui font le plus diverger les classes.

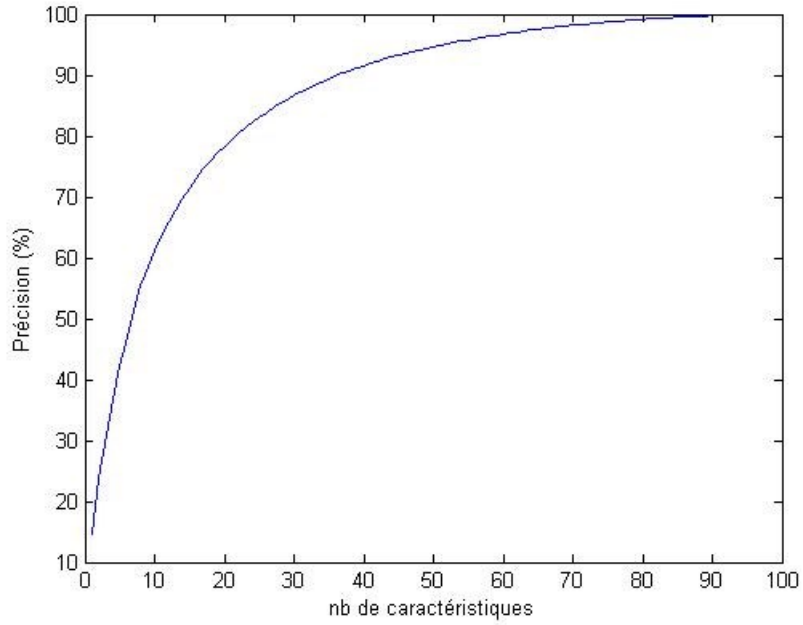
a) Projection des caractéristiques

Il existe beaucoup de méthodes de projection des informations d'un objet. Nous nous concentrerons sur une méthode qui est l'analyse par composante principale (ACP).
L'analyse en composantes principales est une méthode statistique qui consiste à extraire les vecteurs propres et les valeurs propres de la matrice de covariance, calculée à partir de l'ensemble de données d'apprentissage. L'intérêt de cette technique est de rechercher les axes principaux de l'ensemble des données qui ont la plus grande variation (valeur propres). En ayant cette information de vecteur propre on peut ainsi projeter nos données selon les axes fournissant le plus d'informations et composé de caractéristiques décorrélées.

Par exemple pour une classe donnée, on peut retrouver 2 vecteurs propres et on choisit l'axe qui possède le plus d'informations (la plus grande variance des données).



Nous avons donc calculé les différentes valeurs et vecteurs propres pour notre base de données totale. Voici un graphique représentant le pourcentage d'informations en fonction du nombre de caractéristiques gardées :



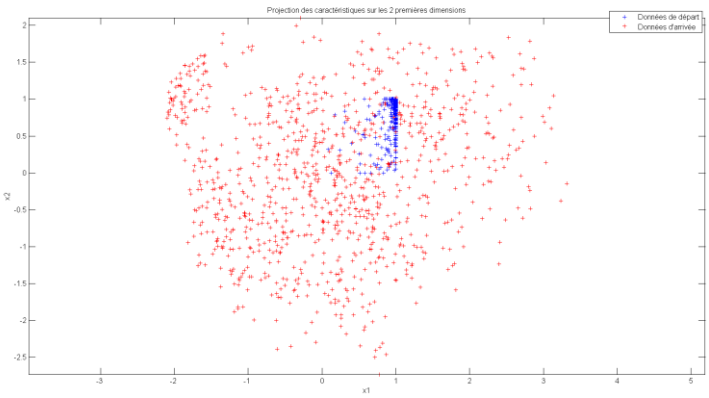
Nous nous fixons un objectif qui est d'obtenir :

$$\frac{\sum_1^m \lambda_i}{\sum_1^n \lambda_i} \geq 0.95$$

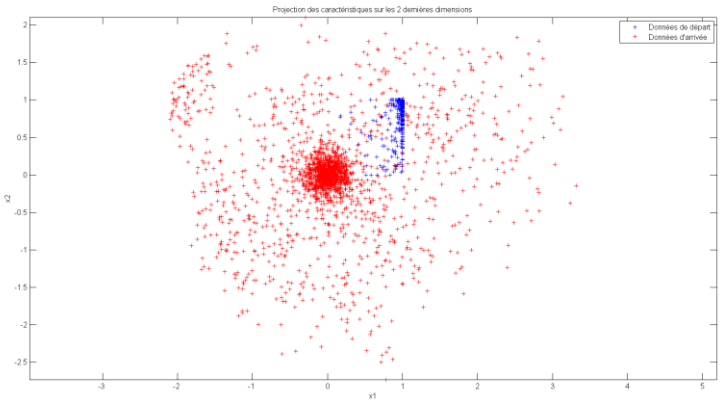
C'est-à-dire que la somme de nos valeurs propres choisies doit correspondre à un poids de 95% parmi toutes les valeurs propres. 95% est beaucoup utilisé en statistiques et est une règle générale pour le choix du taux de rejet à5%. En appliquant cette règle, nous avons décidé de garder les 52 meilleurs axes parmi les 100 disponibles. Il est nécessaire maintenant de projeter les données X de notre **base de test** dans notre nouvel espace défini par les vecteurs propres u_i précédemment choisis. Sans oublier bien évidemment de normaliser.

$$Y = (X - \mu) \times u_i \qquad \text{avec,} \qquad u_i = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \quad X = \begin{pmatrix} x_{1,1} & \dots & x_{1,d} \\ \vdots & \dots & \vdots \\ x_{n,1} & \dots & x_{n,d} \end{pmatrix}$$

Nous avons donc Y qui possède 52 caractéristiques. En guise de résultat, voici un exemple en prenant les 2 premières dimensions et en projetant les données dans les 2 meilleurs vecteurs propres. On voit que l'on a une répartition beaucoup plus variable que les données de départ en bleu.



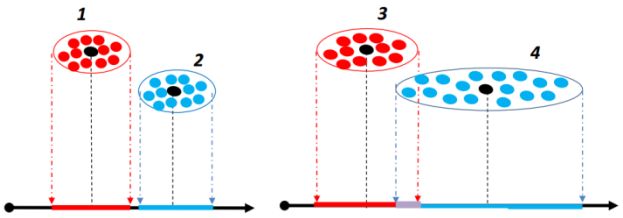
En projetant sur les 2 derniers vecteurs propres, on voit au contraire que la variabilité est moins élevée et les données sont beaucoup plus concentrées. On remarque donc que les 2 premiers vecteurs propres sont plus intéressants que les 51 et 52èmes (dans le cas où l'on en a choisi 52).



Par contre il faut faire attention à une chose, c'est qu'il ne faut pas généraliser le cas. Autant pour les 2 premières dimensions les résultats sont concluants, mais ça ne veut pas dire que l'on aura de bons résultats pour les autres. Il ne faut pas oublier que l'ACP fait perdre de l'information (dans notre cas 95% de l'information).

b) Sélection des caractéristiques

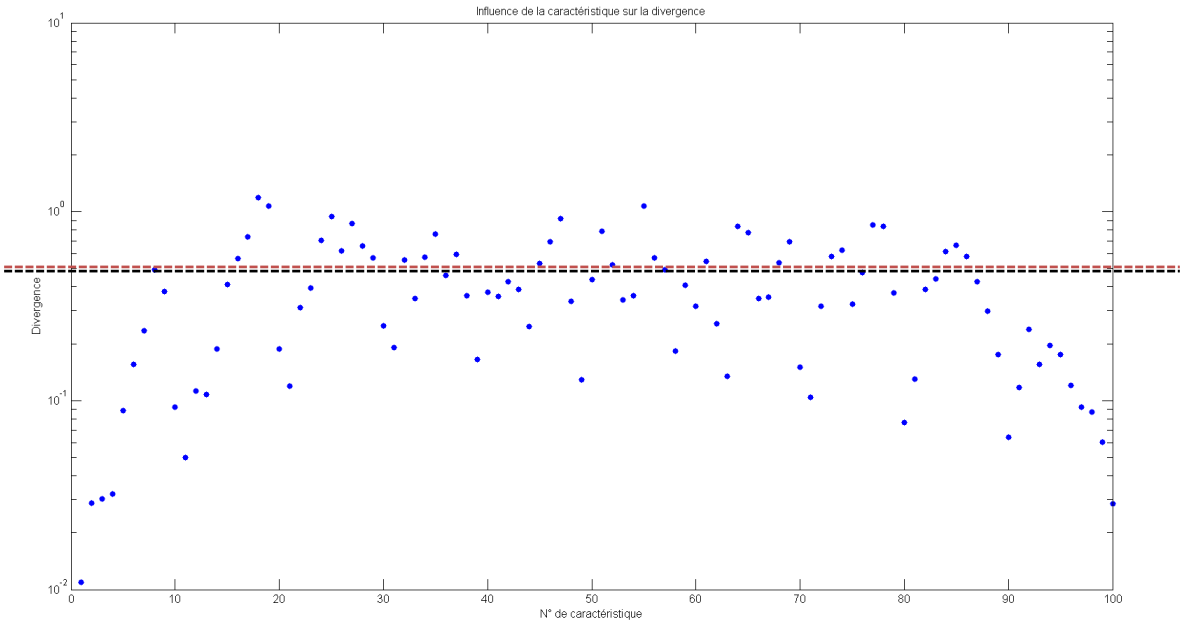
Une fois que l'on a notre information sur les axes contenant le plus d'informations, on peut sélectionner les meilleurs axes. Pour cela, on va calculer les divergences entre les classes, ainsi si la divergence est bonne il s'agit d'une bonne caractéristique.



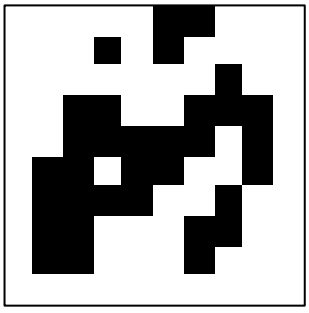
1 et 2 divergent bien contrairement à 3 et 4.

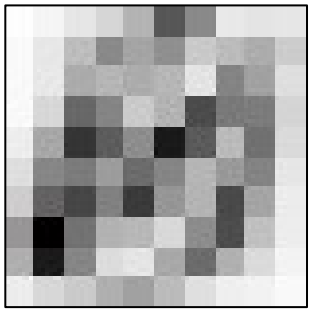
Pour calculer la divergence entre les classes, nous utiliserons le **critère de Fisher** qui a déjà été utilisé au préalable.

Pour sélectionner les dimensions, il est nécessaire de travailler sur chacune des dimensions. Ainsi sur *MATLAB* on a une matrice 3D avec 100 tableaux de divergence (un tableau pour chaque caractéristique). Pour obtenir l'information de la divergence entre les classes et ce pour chacune des caractéristiques, on effectue la moyenne de la matrice de divergence. On obtient ainsi les valeurs suivantes pour chaque dimension :



On considère que l'on a une bonne divergence à partir de 0.5 (au sens de Fisher). En effet à partir de0.5, on ne retrouve plus de recouvrement avec en plus une petite marge entre les 2 classes (pour être sûr). L'application de ce critère nous permet de sélectionner 33 caractéristiques parmi les 100 ce qui est raisonnable. Dans l'image de gauche ci-dessous, les régions de rétine coloriées en noir correspondent donc à celles qui sont les plus discriminantes après calcul. L'image de droite quant à elle donne le niveau d'importance de chaque région de l'image, plus le pixel est sombre, plus la région est importante.
On a donc une sélection plutôt logique au vu des images de la base de test. En effet on remarque entre autre que les côtés ne sont vraiment pas sélectifs car on a centré l'image et donc les pixels sont plutôt concentrés au milieu.





Conclusion

Alors que la sélection va permettre de mieux discriminer les données rentrantes, la projection permet, elle, d'obtenir plus d'informations en utilisant un nouvel espace pour nos données. Dans le premier cas, on a sélectionné 33 caractéristiques parmi les 100, dans la projection on en a sélectionné 52.
Avec la sélection, on travaille sur des zones connues puisque l'on ne change pas la base dans laquelle nous travaillons. La projection nous projette dans une nouvelle base qui n'a plus vraiment de sens physique (les caractéristiques ne sont plus les régions de la rétine). Par contre l'ACP permet d'obtenir une base contenant plus d'information que celle de départ. Malheureusement avec une mauvaise divergence entre les classes comme le montre le tableau suivant (pour la meilleure caractéristique, plus grande valeur propre) :

	1	2	3	4	5	6	7	8	9	10
1	0	0.4373	0.1434	0.1863	0.2402	0.1854	0.1767	0.2968	0.2208	0.2911
2	0.4373	0	0.2666	0.3583	0.2902	0.2920	0.3804	0.2716	0.2996	0.2894
3	0.1434	0.2666	0	0.1641	0.1586	0.1478	0.1349	0.1669	0.1182	0.1952
4	0.1863	0.3583	0.1641	0	0.2445	0.1545	0.2329	0.2000	0.1557	0.2365
5	0.2402	0.2902	0.1586	0.2445	0	0.1314	0.2097	0.2303	0.1899	0.1282
6	0.1854	0.2920	0.1478	0.1545	0.1314	0	0.1855	0.1849	0.1470	0.1387
7	0.1767	0.3804	0.1349	0.2329	0.2097	0.1855	0	0.3520	0.2303	0.3108
8	0.2968	0.2716	0.1669	0.2000	0.2303	0.1849	0.3520	0	0.1810	0.1843
9	0.2208	0.2996	0.1182	0.1557	0.1899	0.1470	0.2303	0.1810	0	0.1446
10	0.2911	0.2894	0.1952	0.2365	0.1282	0.1387	0.3108	0.1843	0.1446	0

Nous sommes donc dans le cas où les 2 méthodes se valent, il pourrait alors être intéressant de combiner les 2 techniques pour avoir les avantages de l'un et l'autre. Un dernier point à prendre en compte au niveau de la sélection et la projection des caractéristiques est la malédiction de la dimensionnalité (« *curse of dimensionnality* »). Elle met en garde le concepteur du système à utiliser un nombre limité de caractéristiques quand seulement un faible nombre d'échantillons d'apprentissage est disponible.
En général on dit que la taille de notre échantillon doit être proportionnelle à la dimension.

$$n = 5\sim10\ d$$

L'idéal étant de minimiser au maximum notre dimension est de maximiser la taille de l'échantillon. Dans notre cas, il faut donc une dimension $d \leq 1200$ ce qui est le cas pour les deux techniques que nous avons utilisées. Attention cependant au fait que le taux de recouvrement augmente avec le nombre de caractéristiques et il y a un compromis.