CLASSIFICATION

Le principe

Les méthodes de classification ont pour objectif le classement d'individus les uns par rapport aux autres. Une classification peut être réalisée à partir d'un tableau rectangulaire présentant en ligne les individus à classer, en colonne les variables observées. Elle peut aussi être réalisée à partir d'un tableau de distances entre individus.

On distingue essentiellement deux grands groupes de méthodes, les méthodes hiérarchiques — qui sont en général des **méthodes hiérarchiques ascendantes** — , et les **méthodes par partition**. Les méthodes hiérarchiques ascendantes consistent à agréger successivement les individus en classes, en calculant une nouvelle matrice de distances entre les classes à chaque agrégation. La classification obtenue est représentée sous la forme d'un dendogramme.
Les méthodes par partition sont des méthodes itératives, qui aboutissent à une partition des individus en classes. Elles nécessitent la connaissance a priori du nombre de classes. On peut citer la méthode des centres mobiles et la méthode des nuées dynamiques parmi les classifications par partition. Seule la méthode des centres mobiles est disponible dans Splus.

Une classification est le résultat d'une **succession de choix** :

- le choix de la **matrice de données** à utiliser. Quelles sont les variables à utiliser, faut-il les transformer, faut-il centrer-réduire la matrice ?

- le choix de la **distance** à utiliser : on peut citer entre autres la distance euclidienne pour les données continues, la distance du chi2 pour des données discrètes, différents types de distances basées sur des pourcentages de concordance pour des données binaires ...

- le choix de la **méthode de classification**.

- le choix de la **méthode d'agrégation pour une classification hiérarchique**. Choisir une méthode d'agrégation revient en fait à choisir la façon dont on calcule la distance entre deux classes : faut-il prendre la moyenne des distances entre toutes les paires d'individus (lien moyen), faut-il prendre la plus petite de ces distances (lien simple), ou la plus grande (lien complet) ?

- le choix du **nombre de classes pour une classification par partition**.

Un exemple

Résumé des principales commandes :
`> classons(scale(loup))`

Les données

Ces données sont extraites d'un tableau publié par M. Jambu, dans "Exploration informatique et statistique des données". Il s'agit de cranes de loups et de chiens pour lesquels différentes mesures ont été faites. On a mesuré entre autres, la longueur de la machoire, la largeur et la hauteur d'une molaire. Les variables sont notées LP, LM et LAM.

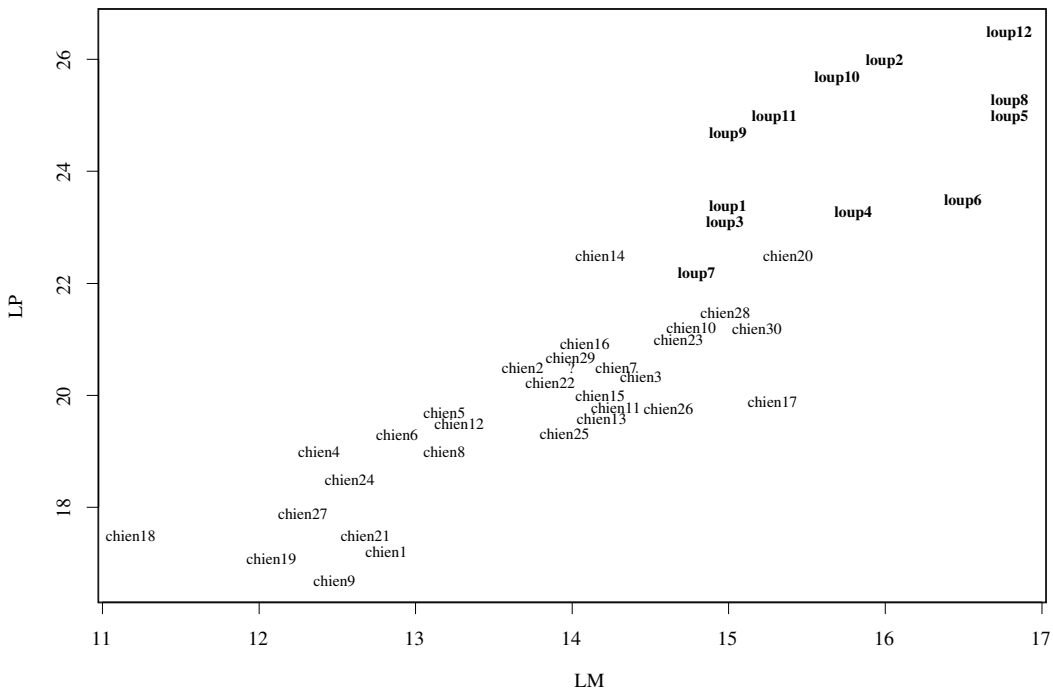
Ce jeu de données a été choisi pour l'illustration des méthodes de classification uniquement. Il a l'intérêt de ne comporter que 3 variables continues, et donc d'être facilement visualisable. Une classification de ces cranes présente peu d'intérêt en soi, puisqu'à l'exception de l'un d'entre eux, on sait s'il s'agit d'un crane de chien ou d'un crane de loup (on utiliserait plutôt une méthode d'analyse discriminante si l'on souhaitait déterminer les variables permettant de discriminer les loups et les chiens). Cependant, on peut toujours envisager d'utiliser la classification comme un outil de description des données.

Les lignes du tableau (les cranes) ont été permutés aléatoirement.

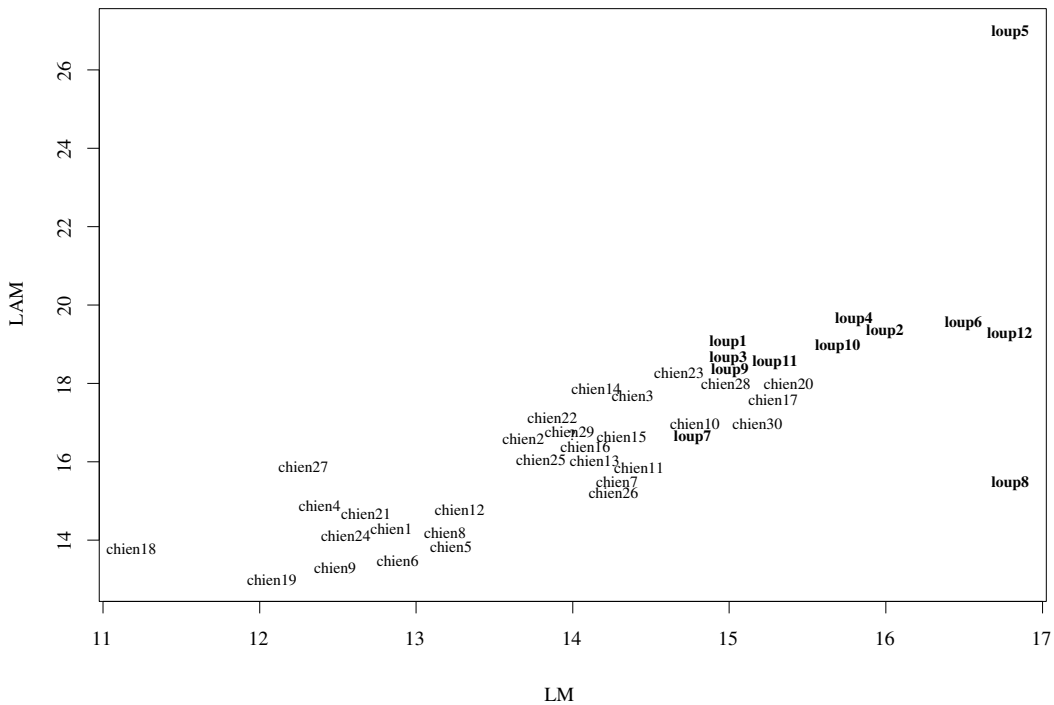
`> loup`

	LP	LM	LAM
loup1	23.4	15.0	19.1
chien1	17.3	12.8	14.3
chien2	20.5	13.7	16.6
loup2	26.0	16.0	19.4
chien3	20.5	14.4	17.7
chien4	19.0	12.4	14.9
chien5	19.7	13.2	14.0
chien6	19.3	12.9	13.5
chien7	20.5	14.3	15.5
loup3	23.3	15.0	18.7
loup4	23.3	15.8	19.7
?	20.5	14.0	16.7
chien8	19.0	13.2	14.2
chien9	16.7	12.5	13.3
chien10	21.3	14.8	17.0
loup5	25.0	16.8	27.0
chien11	19.8	14.3	15.9
chien12	19.5	13.3	14.8
chien13	19.7	14.2	16.0
loup6	23.5	16.5	19.6
chien14	22.5	14.2	17.8
chien15	20.0	14.2	16.5
chien16	20.8	14.1	16.4
chien17	19.9	15.3	17.6
chien18	17.5	11.2	13.8
chien19	17.1	12.1	13.0
chien20	22.5	15.4	18.0
loup7	22.2	14.8	17.0
chien21	17.5	12.7	14.7
loup8	25.3	16.8	15.5
loup9	24.7	15.0	18.5
chien22	20.3	13.9	17.0
chien23	21.0	14.7	18.3
loup10	25.7	15.7	19.0
loup11	25.0	15.3	18.6
chien24	18.5	12.6	14.2
chien25	19.6	14.0	16.4
chien26	19.8	14.3	15.6
chien27	17.9	12.3	15.9
loup12	26.5	16.8	19.3
chien28	21.4	15.0	18.0
chien29	20.7	14.0	16.8
chien30	21.2	15.2	17.0

```
> plot(loup[,"LM"],loup[,"LP"],type="n",xlab="LM",ylab="LP")
> text(loup[,"LM"],loup[,"LP"],dimnames(loup)[[1]],cex=.8)
```




```
plot(loup[,"LM"],loup[,"LAM"],type="n",xlab="LM",ylab="LAM")
text(loup[,"LM"],loup[,"LAM"],dimnames(loup)[[1]],cex=.8)
```



Mise en œuvre

On peut utiliser la fonction “dist” pour calculer une matrice de distances, la fonction “hclust” pour réaliser une classification hiérarchique ascendante, la fonction “kmeans” pour une classification par la méthode des centres mobiles.

On utilisera la fonction “classons”[1] pour classer ces cranes. La mise en œuvre et les résultats sont présentés plus loin[2], et commentés ci-dessous.

❶ On **centre et réduit** la matrice de données (“scale”) de manière à donner le même poids aux trois variables. On choisit la **distance euclidiennne**. Chacune des trois variables a la même contribution à la matrice de distances.

La **matrice des distances** entre toutes les paires de cranes est représentée graphiquement **sous forme de grisés** d’autant plus pales que la distance est faible. L’ordre des cranes étant aléatoire, on n’observe aucune structure particulière dans cette matrice. Cette représentation met en évidence le fait que le loup5 est très différent de tous les autres chiens et de tous les autres loups. Outre un LP et un LM parmi les plus élevés, il a un LAM très supèrieur aux autres. Le loup12 est assez différent de la plupart des autres. Les loups 10 et 11 se ressemblent, ils présentent les mêmes “profils de distances”. Les chiens 18 et 19 sont très semblables également ...
Cette représentation est très riche d’information. Cependant, un regroupement des individus voisins les uns des autres permettrait une meilleure interprétation. On aimerait par exemple regrouper les loups 12, 11, 10, 9 et 8.

❷ Une **classification hiérarchique ascendante** est réalisée. On choisit le **lien complet** comme méthode d’agrégation. Cela signifie que la distance entre deux classes sera définie comme la plus grande des distances entre toutes les paires d’individus des deux classes. Cette méthode a tendance à constituer des classes compactes, contrairement à la méthode du lien simple (la distance entre deux classes est la plus petite des distances entre les paires) avec laquelle on obtient en général des “phénomènes de chainage”.

La classification obtenue est représentée sous forme de **dendogramme**. Le graphique suivant présente les “indices de niveaux par ordre décroissant” : ce sont les distances correspondant aux nœuds de l’arbre, lorsqu’on le parcourt de la racine aux feuilles. C’est un outil d’aide au choix du nombre de classes. Il nous incite à choisir 2, 3 ou 4 classes.

Si on choisit 4 classes, une classe est constituée des chiens 4 à 19, une autre du loup5, une autre des loups 8 à 6, et la dernière est en majorité constituée de chiens, trois loups s’étant “glissés” parmi eux.

---

[1] fonction développée par Colette Roux (INRA Biométrie — Montpellier), pour l’aide à la mise en œuvre des méthodes de classification. Cette fonction n’est pas dans son état définitif aujourd’hui, elle est succeptible d’évoluer. On accèdera à la fonction par une commande “attach” ou par une commande “library”, selon le type d’installation.

[2] Tous les résultats obtenus à l’aide de la fonction “classons” sont encadrés.

L'ordre des individus (de gauche à droite par exemple) sur le dendogramme n'a aucun sens. Il faut imaginer ce dendogramme comme un mobile dans l'espace à trois dimensions. On peut présenter le même dendogramme après rotation de certaines classes et sous classes autour de leur "nœud racine".

❸ On est toujours dans le cadre de la **classification hiérarchique ascendante**, avec lien complet. La fonction demande à l'utilisateur de **choisir un nombre de classes**. Ce sont le dendogramme et la représentation des indices par niveaux qui doivent guider l'utilisateur dans ce choix, mais aussi des connaissances a priori sur les données. Les individus étant constitués de loups et de chiens, on pourrait être tenté de choisir deux classes. Le dendogramme et les indices de niveaux suggèrent 2, 3 ou 4 classes.

On choisit d'abord **2 classes**. La fonction représente tous les individus dans le plan formé par les variables LM et LP (elle choisit les deux variables qui contribuent le plus à la matrice de distance), et matérialise les 2 classes obtenues (classification hiérarchique ascendante) par leurs **enveloppes convexes**. Ces deux classes **expliquent 51% de l'inertie des données**.

On choisit successivement 3 puis 4 classes. Trois classes expliquent 65 % de l'inertie, alors que 4 classes en expliquent 84%.
On tente 5 classes, bien que rien ne nous incite à en choisir 5, le but est seulement de visualiser le gain d'inertie expliquée. Ce gain est très faible, on passe de 84 à 87 %. Cela nous conforte dans l'idée du choix de 3 ou 4 classes, éventuellement 2 si on souhaite résumer très fortement ces données.
Dans le cas de 3, 4 et 5 classes, l'une des classes est constituée d'un seul individu, c'est la classe qui apparait sous la forme d'un segment oblique.

Enfin, la **matrice de distances** est à nouveau représentée, les **individus** étant **réordonnés en fonction de l'ordre de présentation dans le dendogramme**. Cet ordre est un ordre parmi d'autres, il n'est pas unique, et dépend de l'ordre initial des individus dans le tableau de données. Cette représentation complète le dendogramme. Elle est beaucoup plus intéressante et beaucoup plus complète que le dendogramme, puisque toutes les distances sont représentées.

Avec une méthode de classification telle que celle qui est utiliséee ici (hiérarchique ascendante), on peut obtenir un dendogramme semblable à celui que l'on a obtenu, avec une distance très faible entre le chien 4 et le loup 5 par exemple, simplement parce que le chien 4 a d'abord été agrégé au chien 24, puis aux chiens 6 à 8, et que le loup 5 ne ressemble pas à ce groupe de chiens, bien qu'il ressemble au chien 4. La représentation en grisés de la matrice de distances, permet, elle, de visualiser la distance entre le chien 4 et le loup 5. Le fait que deux individus s'agrègent "haut" (près de la racine) dans le dendogramme ne prouve pas qu'ils soient très distants l'un de l'autre. C'est en ce sens que la représentation de la matrice de distances sous forme de grisés complète utilement le dendogramme.

Les 2, 3 ou 4 classes sont matérialisées par 2, 3 ou 4 **blocs diagonaux clairs** (distances faibles

entre individus). Les limites de classes ne sont pas toujours très nettes. Si on choisit 4 classes, l'examen visuel de la matrice des distances conduit à choisir les classes : [chien4 — chien19], [loup5], [loup8 — loup3], [chien7 — chien28], alors que le dendogramme définit les classes [chien4 — chien19], [loup5], [loup8 — loup6], [chien14 — chien28]. A l'examen de la matrice grisée, on est tenté de placer la limite entre les classes 3 et 4, entre loup3 et chien7, plutôt qu'entre loup6 et chien14. On est aussi tenté d'aménager la classification hiérarchique, en "faisant passer" le chien 14 (et peut-être le chien20) dans la classe 4.

Ceci est lié à un **inconvénient des méthodes hiérarchiques ascendantes qui ne remettent pas en cause les agrégations déjà réalisées**. L'utilisation des classifications non hiérarchiques est intéressante dans la mesure où, par des procédures itératives, elles consistent à échanger des individus d'une classe à l'autre, de manière à obtenir une homogénéité intra classe et une hétérogénéité inter classe les plus grandes possible.

❹ Une **classification non hiérarchique** par la **méthode des centres mobiles** est réalisée. **L'utilisateur choisit le nombre de classes**. On choisit **2 classes** dans un premier temps, puisqu'on sait que les cranes sont constitués de cranes de loups et de cranes de chiens. On espère, avec 2 classes et une méthode non hiérarchique, obtenir une inertie expliquée supèrieure à celle que l'on obtient avec le même nombre de classes et une classification hiérarchique.

La méthode des centres mobiles nécessite la **détermination d'autant de noyaux initiaux qu'il y a de classes**, chaque **noyau** étant composé d'**un seul individu.** La fonction **tire aléatoirement les noyaux** (ici deux cranes) parmi les individus, met en œuvre une **méthode itérative** qui conduit à la détermination d'une **partition** en 2 classes. La partition obtenue est représentée dans le plan formé par les variables ayant la plus forte contribution à la détermination de la matrice de distances, les noyaux initiaux sont matérialisés par des points noirs. Le pourcentage d'inertie expliquée est affiché. Une classification par la méthode des centres mobiles est ainsi réalisée. Plusieurs tirages aléatoires des noyaux initiaux sont réalisés, les classifications correspondantes sont réalisées et représentées.

Dans trois cas sur cinq, on obtient une partition avec une inertie expliquée égale à 58%. Dans les autres cas, la partition obtenue est celle qui correspond à une classification hiérarchique ascendante, le pourcentage d'inertie est infèrieur. La **partition** en 2 classes n'est **pas stable**, dans la mesure où le résultat dépend du choix des noyaux initiaux. Il existe deux partitions concurrentes.

La matrice des distances correspondant à la dernière classification non hiérarchique (58% d'inertie expliquée) est représentée. Elle fait apparaitre la partition en 2 classes, avec une classe assez homogène constituée de chiens uniquement, et une classe plus hétérogène constituée de chiens et de loups.

L'exécution de la fonction "classons" est terminée.

❺ On “relance” l’exécution de la fonction “classons”, en passant directement à la **classification non hiérarchique**. On choisit **3 classes**. La **partition** obtenue est **stable** (on obtient toujours la même partition, même avec un nombre de tirages aléatoires des noyaux bien supèrieur à celui qui est présenté). On explique 77% de l’inertie. On obtient une classe constituée de chiens uniquement, une autre classe constituée de chiens et d’un loup, et une dernière classe constituée de loups à l’exception d’un chien. A l’exception d’un chien et d’un loup mal classés, on obtient une classe de loups et deux classes de chiens qui se distinguent nettement l’une de l’autre.

On aurait pu obtenir une matrice de distances à l’allure plus ordonnée, en ordonnant avant l’utilisation de “classons”, les individus en fonction d’une classification hiérarchique :

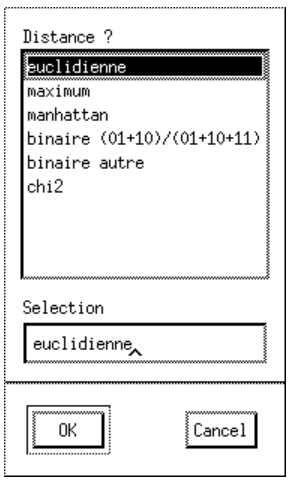
```
> loup1_loup[hclust(dist(scale(loup)))$order,]
> classons(scale(loup1))
```

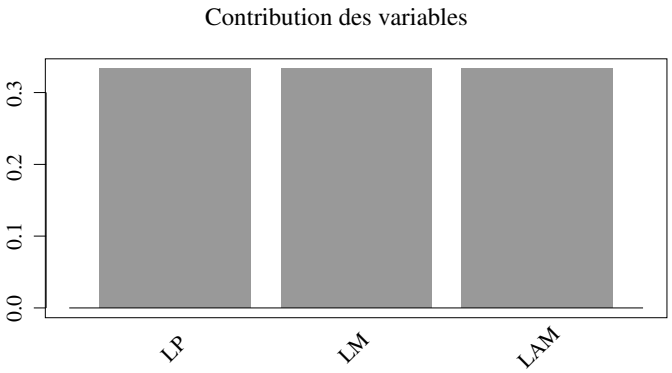
❻ On peut aussi ordonner les individus de chacune des trois classes en fonction d’une classification hiérarchique ascendante dans chacune des classes. Cela a l’avantage de faciliter un examen encore plus détaillé des distances entre certains individus, et l’inconvénient — dans ce cas — de rendre plus floue la limite entre les loups et la classe de chiens la plus ressemblante aux loups. L’œil aurait tendance à tracer la limite entre le chien 30 et le chien 7, alors que la méthode des centres mobiles la situe entre le loup 3 et le chien 14. Les chiens 14, 10, 30 et le loup 7 sont intermédiaires, mais ils sont classés avec les chiens, et la partition est stable.On pourrait remédier à cet inconvénient en choisissant un autre ordre des individus : loup5 ... loup3, chien18 ... chien24, chien14 ... chien28.

Des tentatives de classification non hiérarchique à 4 classes montrent que la partition obtenue n’est pas stable. On choisit donc une classification en 3 classes.

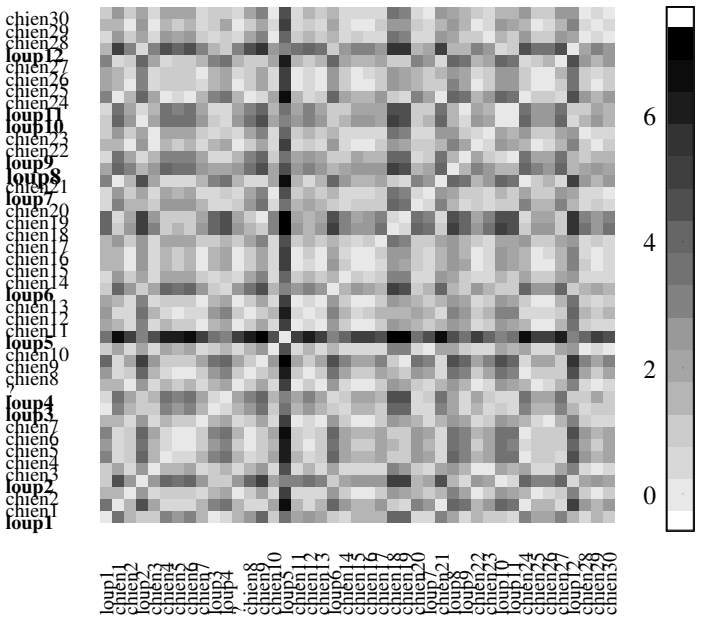
❶

```
> classons(scale(loup))
```

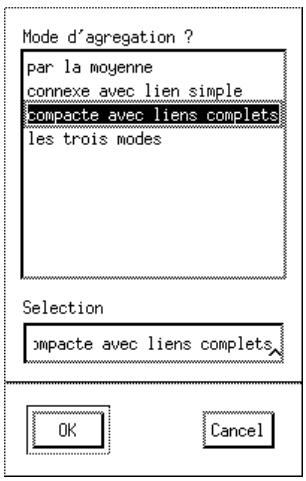




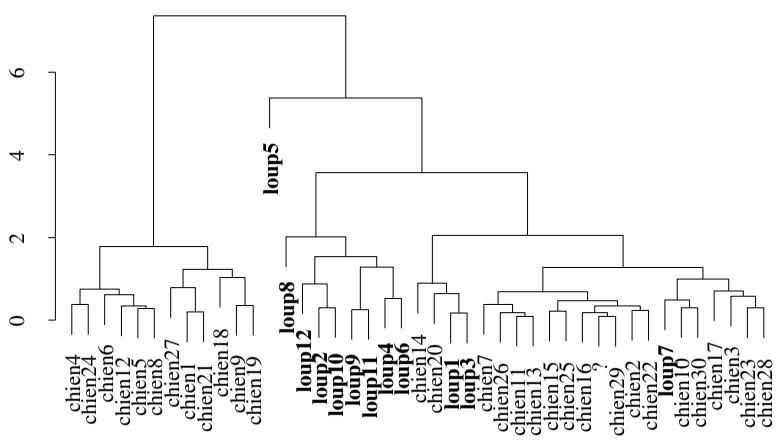
Matrice des distances (euclidienne)

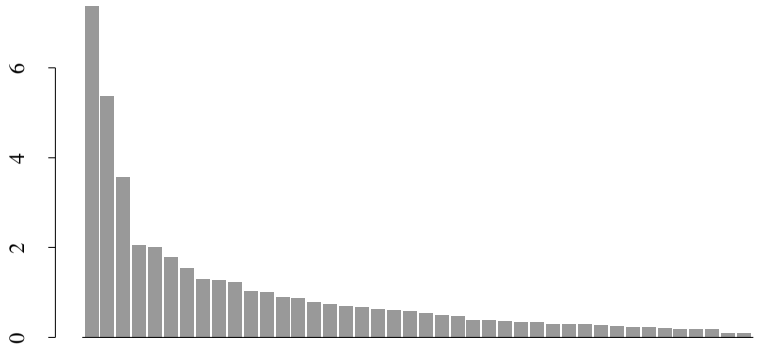


Classification hierarchique ascendante



Agregation compacte avec lien complet



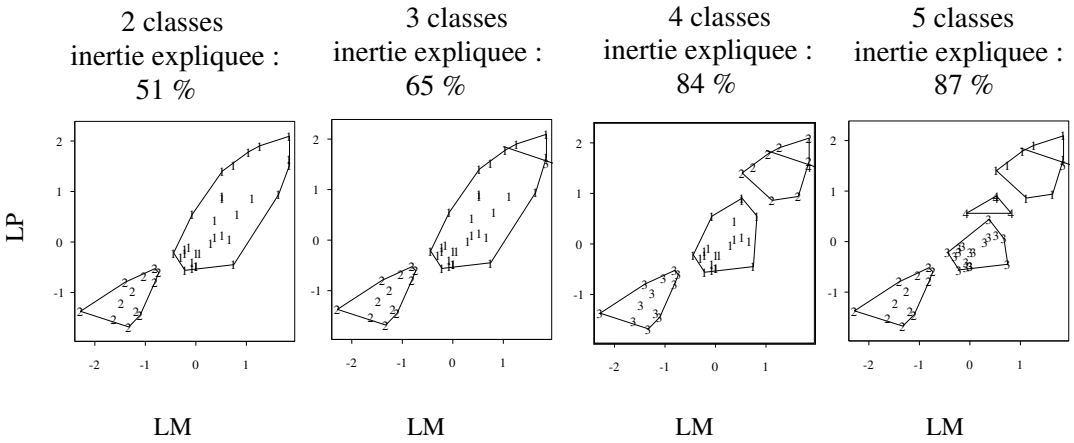


Indices de niveaux par ordre decroissant

❸

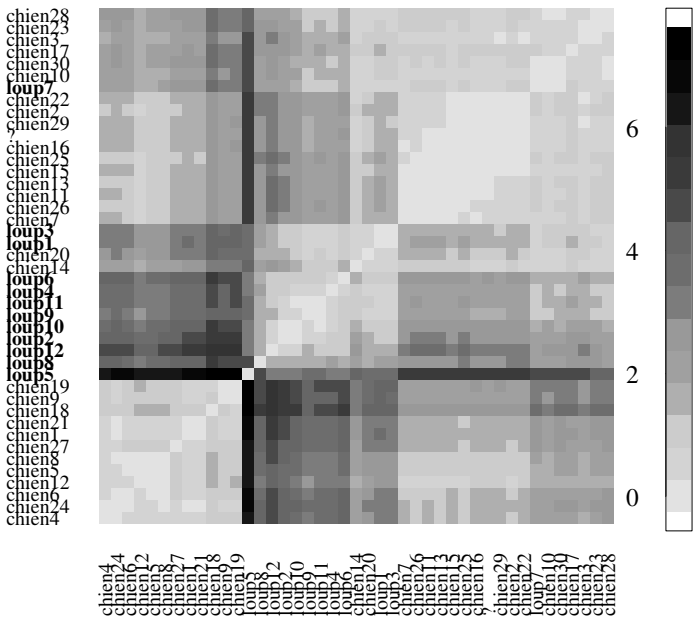
```
nombre de classes ? (ou RETURN) 2
nombre de classes ? (ou RETURN) 3
nombre de classes ? (ou RETURN) 4
nombre de classes ? (ou RETURN) 5
nombre de classes ? (ou RETURN)
```

Agregation compacte avec lien complet



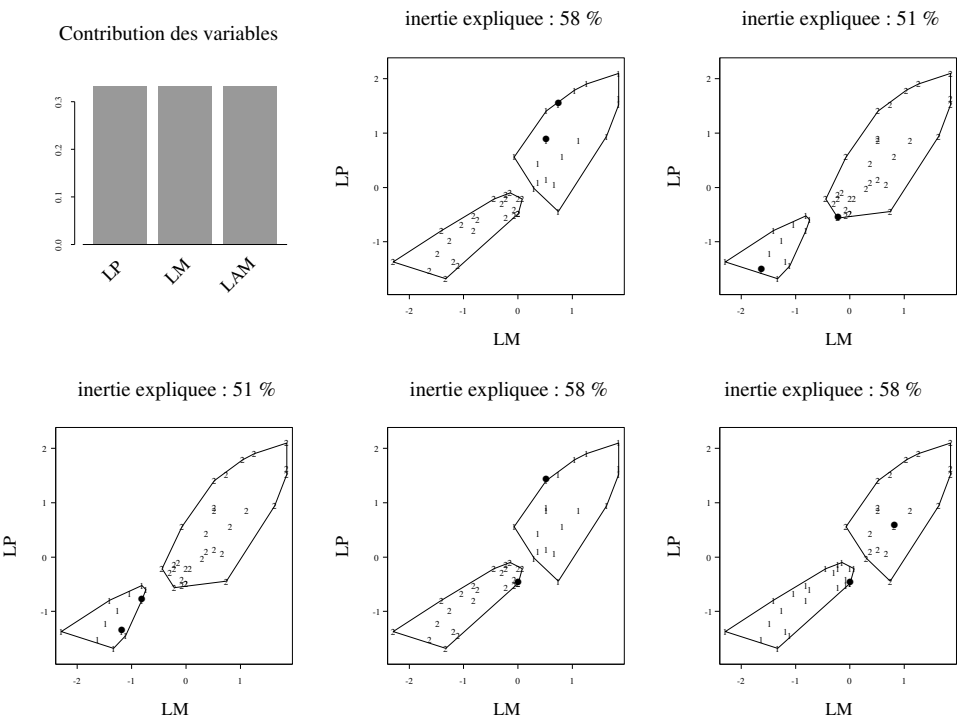
Matrice des distances (euclidienne)

individus reordonnes en fonction de la classification hierarchique



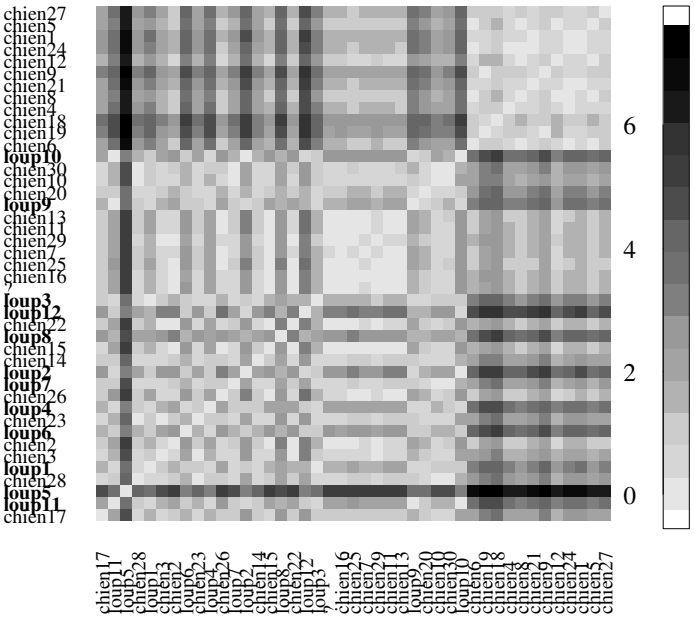
❹

```
Classification non-hierarchique
Nombre de classes ? (ou RETURN) 2
```



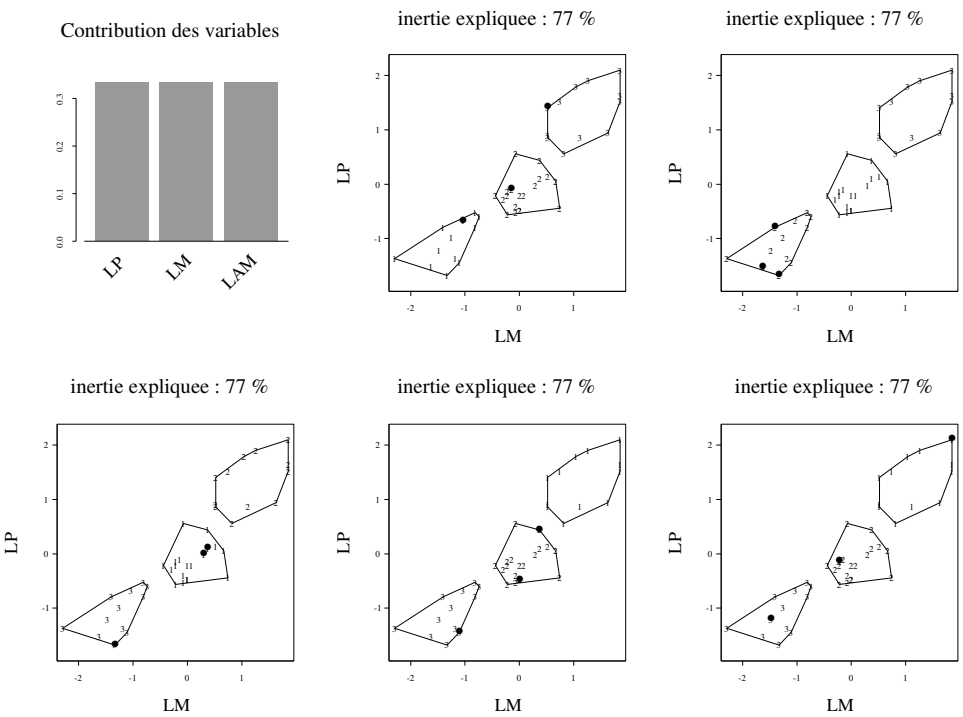
Matrice des distances (euclidienne)

individus reordonnes en fonction de la derniere classification non hierarchique



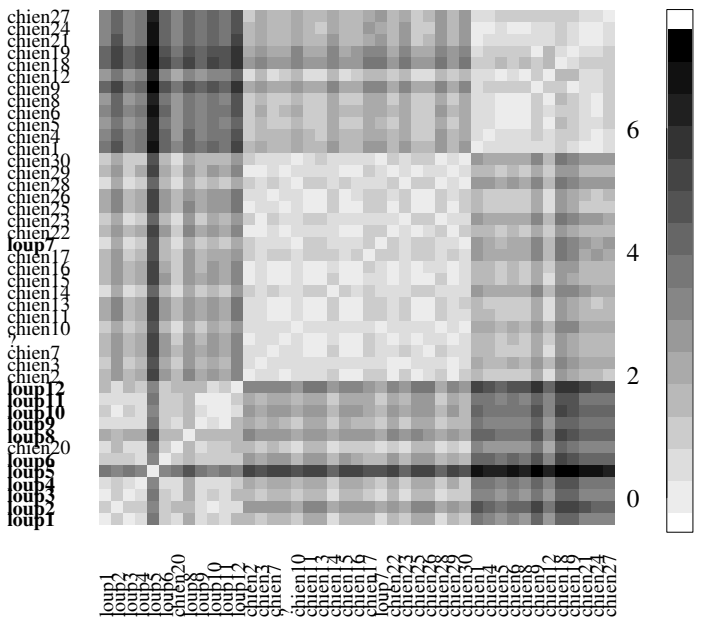
❺

```
> resultat_classons(scale(loup))
Classification non-hierarchique
Nombre de classes ? (ou RETURN) 3
```



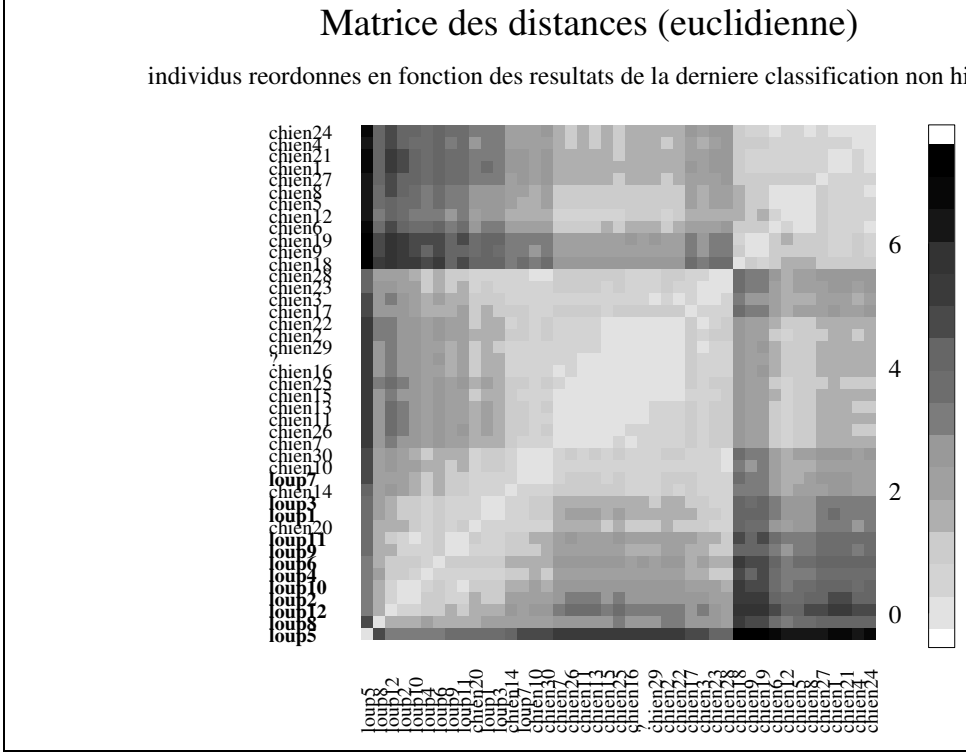
Matrice des distances (euclidienne)

individus reordonnes en fonction de la derniere classification non hierarchique



❻

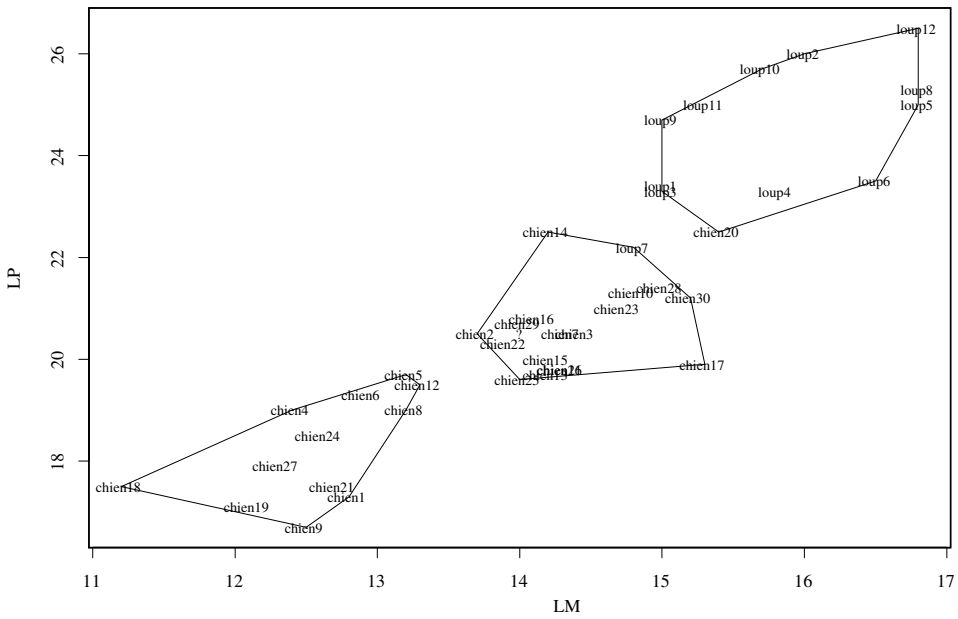
```
> loup2_NULL
> appartenance_matrix(resultat$app,ncol=11,byrow=F)[,11]
> for (i in 1:3)
  {sous.groupe_loup1[appartenance==i,]
   sous.groupe_sous.groupe[hclust(dist(scale(sous.groupe)))$order,]
   loup2_rbind(loup2,sous.groupe)
  }
> classons(scale(loup2))
Classification non-hierarchique
Nombre de classes ? (ou RETURN) 3
```

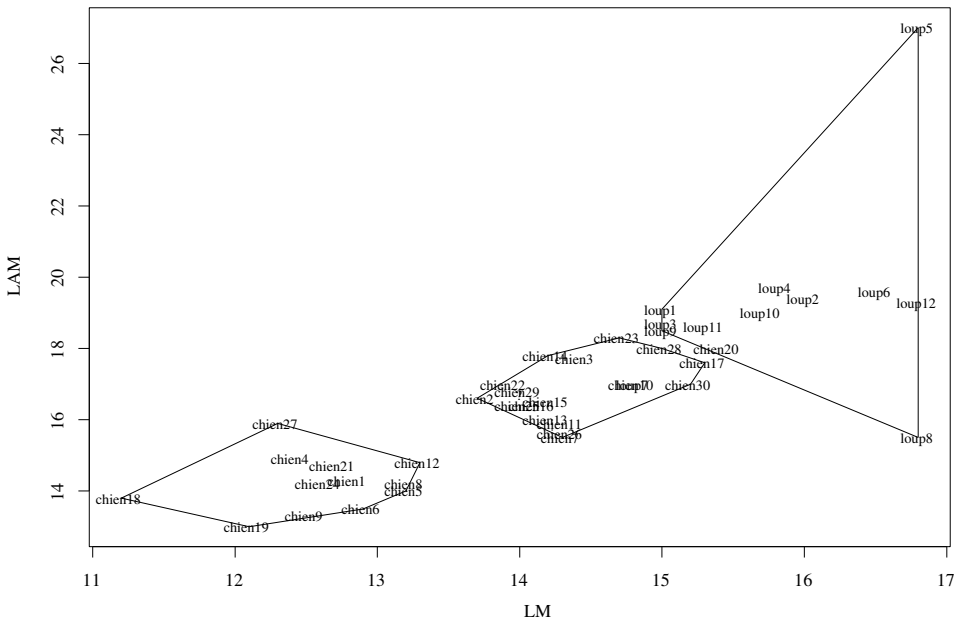


Enfin, on représente les enveloppes convexes des 3 classes obtenues par la méthode des centres mobiles.

```
> gr1_appartenance==1; gr2_appartenance==2; gr3_appartenance==3
> groupe1_loup[gr1,]; groupe2_loup[gr2,]; groupe3_loup[gr3,]
> hull1_chull(groupe1[,"LM"],groupe1[,"LP"])
> hull2_chull(groupe2[,"LM"],groupe2[,"LP"])
> hull3_chull(groupe3[,"LM"],groupe3[,"LP"])
> plot(loup1[,"LM"],loup1[,"LP"],type="n",xlab="LM",ylab="LP")
> text(loup1[,"LM"],loup1[,"LP"],dimnames(loup1)[[1]],cex=.8)
> polygon(groupe1[hull1,"LM"],groupe1[hull1,"LP"],density=0)
> polygon(groupe2[hull2,"LM"],groupe2[hull2,"LP"],density=0)
```

```
> polygon(groupe3[hull3,"LM"],groupe3[hull3,"LP"],density=0)
> hull1_chull(groupe1[,"LM"],groupe1[,"LAM"])
> hull2_chull(groupe2[,"LM"],groupe2[,"LAM"])
> hull3_chull(groupe3[,"LM"],groupe3[,"LAM"])
> plot(loup1[,"LM"],loup1[,"LAM"],type="n",xlab="LM",ylab="LAM")
> text(loup1[,"LM"],loup1[,"LAM"],dimnames(loup1)[[1]],cex=.8)
> polygon(groupe1[hull1,"LM"],groupe1[hull1,"LAM"],density=0)
> polygon(groupe2[hull2,"LM"],groupe2[hull2,"LAM"],density=0)
> polygon(groupe3[hull3,"LM"],groupe3[hull3,"LAM"],density=0)
```





Conclusion

La méthode des centres mobiles peut être utilisée pour améliorer une classification hiérarchique, lorsque les données peuvent effectivement être partitionnées en classes homogènes, bien séparées (concept de classifiabilité). Si c'est le cas, et si le nombre de classes est bien choisi, on obtient une partition stable. S'il n'existe pas de classes homogènes, bien séparées, une méthode de classification est quand-même envisageable, ne serait-ce que pour décrire et étudier les distances entre individus. Cependant, on ne pourra pas utiliser une méthode de classification par partition.

Dans le cas d'un tableau de données contenant à la fois des variables quantitatives et des variables qualitatives, on peut réaliser une AFC sur les données discrètes, une ACP sur les données continues, constituer un nouveau tableau dont les lignes sont les individus et les colonnes, les nouvelles coordonnées sur les axes factoriels de l'ACP et de l'AFC. On peut choisir de retenir les premiers axes factoriels seulement — l'AFC et l'ACP sont alors utilisées comme des filtres —, ou tous les axes factoriels. On utilise la distance euclidienne, et la méthode de classification de son choix.

Une autre solution dans le cas d'un tableau contenant à la fois des variables discrètes et des variables continues, consiste à coder en classes chaque variable, à éclater le tableau en tableau disjonctif complet[3]. On obtient un tableau de données binaires. On peut classer les individus en utilisant la distance du chi2.
On peut aussi réaliser une AFC à partir du tableau disjonctif complet, ne conserver que les premiers axes, récupérer les coordonnées, et travailler avec une distance euclidienne.

L'utilisation d'une AFC ou d'une ACP comme d'un filtre permet de supprimer "le bruit", et d'obtenir par la suite une classification plus stable.

Enfin, il existe de nombreux autres types de distances et de dissimilarités ainsi que de nombreuses autres méthodes de classification, dont on ne parlera pas ici.

[3] voir l'AFC pour le terme "tableau disjonctif complet".