Méthodes factorielles de représentation et de discrimination: ACP, AFC, ACM et CAH

Lise Vaudor

2016-08-29

Contents

1 Introduction aux méthodes factorielles

1.1 Qu'est-ce qu'une méthode factorielle de représentation?

Supposons que l'on s'intéresse à un jeu de données comprenant seulement 2 ou 3 variables numériques.

Il est dans ce cas facile de "visualiser" ce jeu de données, par exemple en regardant les relations entre les variables prises 2 à 2:

On peut essayer de caractériser notamment la corrélation entre les variables à travers le calcul du coefficient de corrélation:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}.\sqrt{\sum_{i=1}^{n}(y_i - \bar{x})^2}}$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Plus ce coefficient est proche de 1 (en valeur absolue), plus la corrélation entre les variables est forte. La corrélation peut être positive ou négative.

```
cor(USA$Illiteracy,USA$Murder)
```

```
[1] 0.7029752
```

```
cor(USA$Illiteracy,USA$HS.Grad)
```

```
[1] -0.6571886
```

```
cor(USA$Illiteracy,USA$Area)
```

```
[1] 0.07726113
```

Dès lors que l'on s'intéresse à un jeu de données comprenant beaucoup de variables ($N > 3$), il devient difficile de comprendre ce qui se passe dans le jeu de données... car **il est impossible a priori de "visualiser" le nuage de points en N dimensions**.

Une **méthode factorielle de représentation telle que l'ACP va permettre de "réduire" le nombre de dimensions nécessaire à la représentation du nuage de points**. De cette manière, le nuage de points pourra être représenté, classiquement en 1D, 2D ou 3D. Les axes servant à cette représentation, classiquement au nombre de 1, 2 ou 3, correspondent aux **composantes principales**, (ou **facteurs**) i.e. des combinaisons linéaires des variables d'origine.

Une méthode factorielle de représentation est ainsi utile lorsque l'on s'intéresse à un jeu de données comprenant de nombreuses variables. Elle permet d'explorer ce jeu de données, en fournissant des éléments de réponse aux questions suivantes:

- Quels sont les individus/échantillons/sites qui se ressemblent ?
- Quelles sont les variables importantes dans mon jeu de données (i.e. quelles sont celles qui apportent une information qui n'est pas redondante)?
- Comment les variables (notamment les variables importantes) sont-elles corrélées les unes aux autres ?

1.2 Comment décrire un nuage de points?

Les méthodes factorielles de représentation reposent en grande partie sur la notion d'**inertie** du nuage de points. L'inertie est la somme des distances au carré de chacun des points du nuage à leur centre. La figure suivante illustre cette notion pour le cas simple où il n'y a que 3 variables (x, y, et z). Grosso modo, **l'inertie correspond à l'étalement du nuage de points**. La notion d'inertie est elle même liées à celle de variance (il s'agit de la variance multipliée par le nombre d'observations, n).

L'inertie liée à une variable V_i est donc simplement:

$$I(V_i) = Var(V_i) * n$$

L'inertie totale pour un jeu de données correspond à la somme des inerties liées à chaque variable

$$I_{tot} = n\left(Var(V_1) + Var(V_2) + \ldots + Var(V_N)\right)$$

- Les méthodes factorielles transforment des nuages de points en n dimensions (jeu de données comprenant n variables) en des nuages de points en 1, 2 ou 3 dimensions (nouveau jeu de données comprenant 1, 2 ou 3 composantes principales).
- Cette transformation, même si elle permet de représenter le jeu de données, correspond forcément à une perte d'information par rapport au jeu de données initial. Néanmoins, elle est effectuée de manière à perdre le moins possible d'information (i.e. de telle sorte que l'inertie des premières composantes principales soit la plus forte possible).

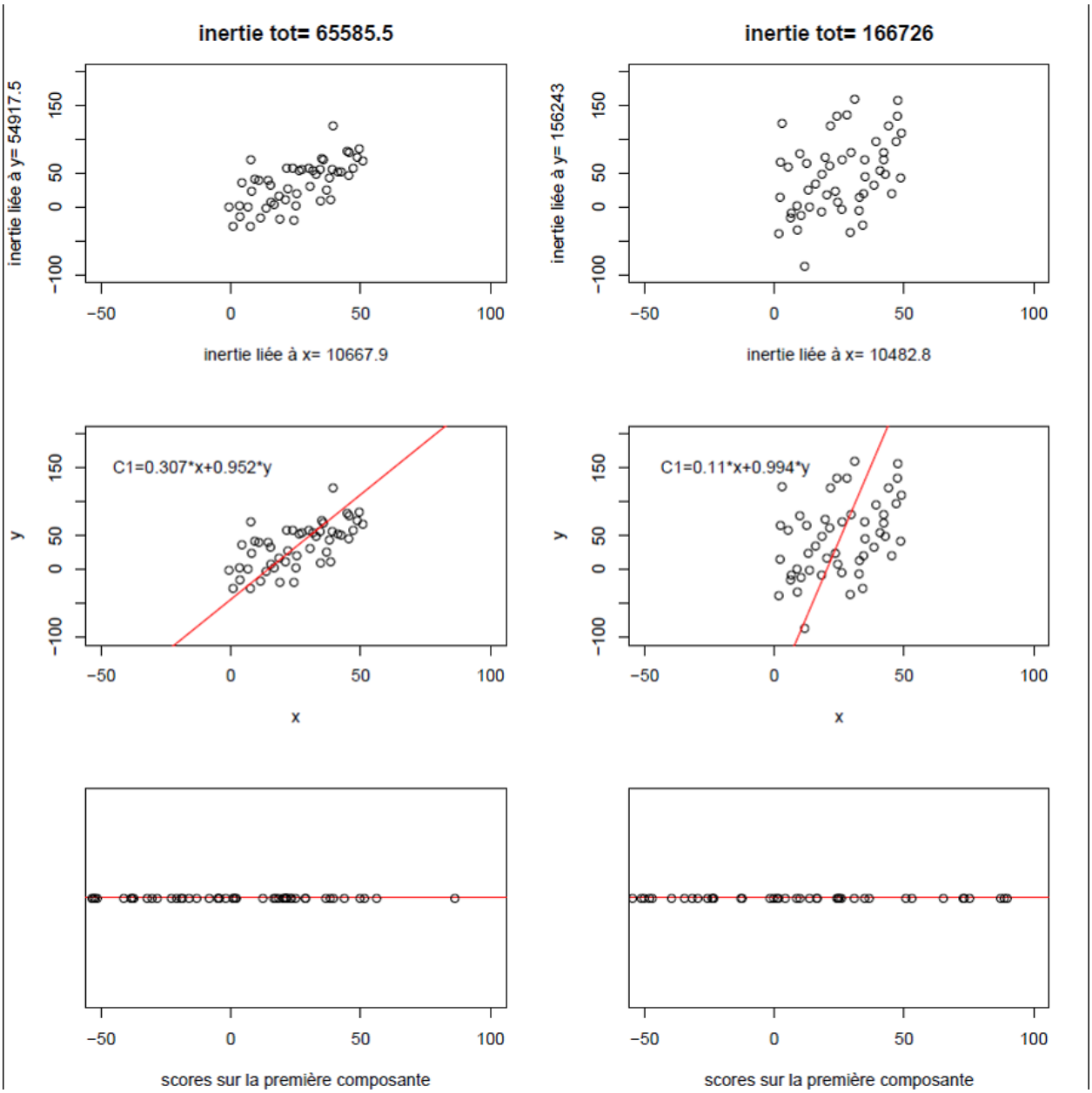


Figure 1: Illustration de la **notion d'inertie pour un nuage de points en 2D**: en réalisant une ACP sur un jeu de données comprenant 2 variables, on cherche à représenter de manière optimale le nuage de points dans un espace à une seule dimension, i.e. on projette les points sur une droite telle que la part d'inertie expliquée par cette projection soit maximale.

1.3 Quelques éléments de vocabulaire

Composantes principales ou **facteurs**

Les **composantes principales** C_1,C_2,C_3,... sont de **nouvelles variables formées par combinaison linéaire des anciennes variables**.

$$C_1 = \alpha_1 V_1 + \alpha_2 V_2 + \alpha_3 V_3 + \ldots + \alpha_n V_n C_2 = \beta_1 V_1 + \beta_2 V_2 + \beta_3 V_3 + \ldots + \beta_n V_n C_3 = \gamma_1 V_1 + \gamma_2 V_2 + \gamma_3 V_3 + \ldots + \gamma_n V_n$$

Inversement, on peut exprimer les variables en fonction des composantes principales:

$$V_1 = \lambda_1 C_1 + \lambda_2 C_2 + \lambda_3 C_3 + \ldots + \lambda_n C_n V_2 = \mu_1 C_1 + \mu_2 C_2 + \mu_3 C_3 + \ldots + \mu_n C_n V_3 = \nu_1 C_1 + \nu_2 C_2 + \nu_3 C_3 + \ldots + \nu_n C_n$$

Axes principaux

Les composantes principales définissent les axes d'un nouveau repère de dimension réduite (1, 2 ou 3D) dans lequel on va projeter les points. Les axes de ce nouveau repère sont dits **axes principaux**.

Scores

Les **scores** correspondent aux **coordonnées des individus ou des variables** dans l'espace défini par les axes principaux (i.e. les composantes principales sont des variables, et les scores sont les réalisations de ces variables).

Loadings

Les **loadings** correspondent aux coefficients $\alpha_1, \alpha_2, ..., \alpha_n$,$\beta_1$,$\beta_2$,...,$\beta_n$ et γ_1,γ_2,..., γ_n des combinaisons linéaires qui définissent les composantes principales en fonction des variables.

Eigenvalues/valeurs propres

Les valeurs propres renseignent sur la part de l'inertie totale prise en compte par chaque composante principale.

2 Un jeu de données comprenant de nombreuses variables quantitatives: l'ACP

L'ACP est l'exemple de base sur lequel nous allons nous appuyer pour comprendre le principe de fonctionnement des méthodes factorielles. Dans cette partie, nous reviendrons sur les éléments (notamment de vocabulaire) définis de manière théorique dans l'introduction.

On ne sait pas a priori ce qui se passe dans le jeu de données : quelles sont les variables importantes, quelles sont les variables qui vont nous apporter une information redondante par rapport à d'autres variables, quels sont les individus qui se ressemblent considérant l'ensemble des descripteurs, etc.

Réaliser une ACP va permettre d'explorer le jeu de données, en fournissant une réponse aux questions suivantes:

- Quels sont les **individus/échantillons/sites** qui se ressemblent?
- Quelles sont les **variables importantes** dans mon jeu de données (i.e. celles pour lesquelles il y a des différences importantes entre les individus)?
- Comment les variables (notamment les variables importantes) sont-elles **intercorrélées**?

2.1 Réaliser l'ACP: quels sont les éléments renvoyés par la fonction R?

Il existe plusieurs fonctions réalisant des ACP sous R (par exemple la fonction prcomp du package stats, la fonction `PCA` du package `FactoMineR`, ou encore la fonction `dudi.pca` du package `ade4`). Ces fonctions renvoient les mêmes éléments de réponse, mais sous des formes différentes, et ne sont pas associées aux mêmes fonctions graphiques. Par souci de simplicité, nous ne traiterons donc qu'une seule de ces fonctions, à savoir la fonction `PCA` du package `FactoMineR`.

```
# installation (si nécessaire)
# install.packages("FactoMineR")
require(FactoMineR)
```

On va réaliser l'ACP sur un jeu de données portant sur certaines **caractéristiques des différents états des USA dans les années 70**. Ce jeu de données fait partie des nombreux jeu de données d'exemple qui peuvent être chargés directement sur R:

```
data(state)
USA=as.data.frame(state.x77)
```

Pour ceux qui (comme moi) ne seraient pas hyper au point sur la géographie des USA une petite carte:

Figure 2: Carte des états des USA

Les variables de ce jeu de données sont

- **Population**: estimation de la population au 1er juillet 1975

- **Income**: revenu par tête (1974)
- **Illiteracy**: illetrisme (1970, pourcentage)
- **Life Exp**: espérance de vie (1969-1971, années)
- **Murder**: nombre d'homicides pour une population de 100 000 personnes (1976)
- **HS Grad**: pourcentage de diplômés de l'enseignement supérieur (1970)
- **Frost**: nombre moyen de jours ayant une température minimale inférieure à 0°C dans la capitale ou plus grande vile (1931-1960)
- **Area**: aire en miles au carré

Pour réaliser l'ACP, le code R est tout simple.

```
monACP=PCA(USA)
```

L'objet "monACP" renvoie un certain nombre d'éléments de réponse que l'on va vouloir interpréter. Les éléments les plus importants sont les suivants:

- `eig`, qui correspond aux eigenvalues (ou valeurs propres).
- `ind`, qui renvoie des informations relatives aux individus.
- `var`, qui renvoie des informations relatives aux variables.

Les éléments `ind` et `var` contiennent tous deux les sous-éléments suivants:

- `coord`
- `cos2`
- `contrib`

Nous allons voir dans les paragraphes suivant comment accéder à ces éléments et comment les interpréter.

2.1.1 Valeurs propres: `eig`

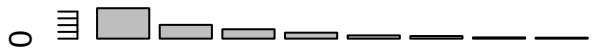
Dans le cas qui nous intéresse, on voit que les deux premiers axes expliquent 45% et 20%, respectivement, de l'inertie totale. Plus l'inertie expliquée par les 1,2 ou 3 premiers axes est importante, et plus l'ACP est "de qualité": cela signifie en effet que l'information peut être résumée efficacement par quelques axes. Evidemment, plus le nombre de variables dans le jeu de données est important, plus la part d'inertie expliquée par les deux ou trois premiers axes risque d'être réduite.

```
monACP$eig
```

```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1  3.5988956              44.986195                           44.98619
comp 2  1.6319192              20.398990                           65.38519
comp 3  1.1119412              13.899264                           79.28445
comp 4  0.7075042               8.843803                           88.12825
comp 5  0.3846417               4.808021                           92.93627
comp 6  0.3074617               3.843271                           96.77954
comp 7  0.1444488               1.805610                           98.58515
comp 8  0.1131877               1.414846                          100.00000
```

Pour représenter graphiquement ces valeurs propres, on peut faire appel à la fonction `barplot`:

```r
barplot(monACP$eig[,2])
```



2.1.2 Coordonnées: `ind$coord` et `var$coord`

Il s'agit des **coordonnées des individus (ou variables) dans le nouveau repère défini par les axes principaux**. Pour les individus, on parle aussi de **scores**.

```r
monACP$ind$coord[1:4,1:5]
```

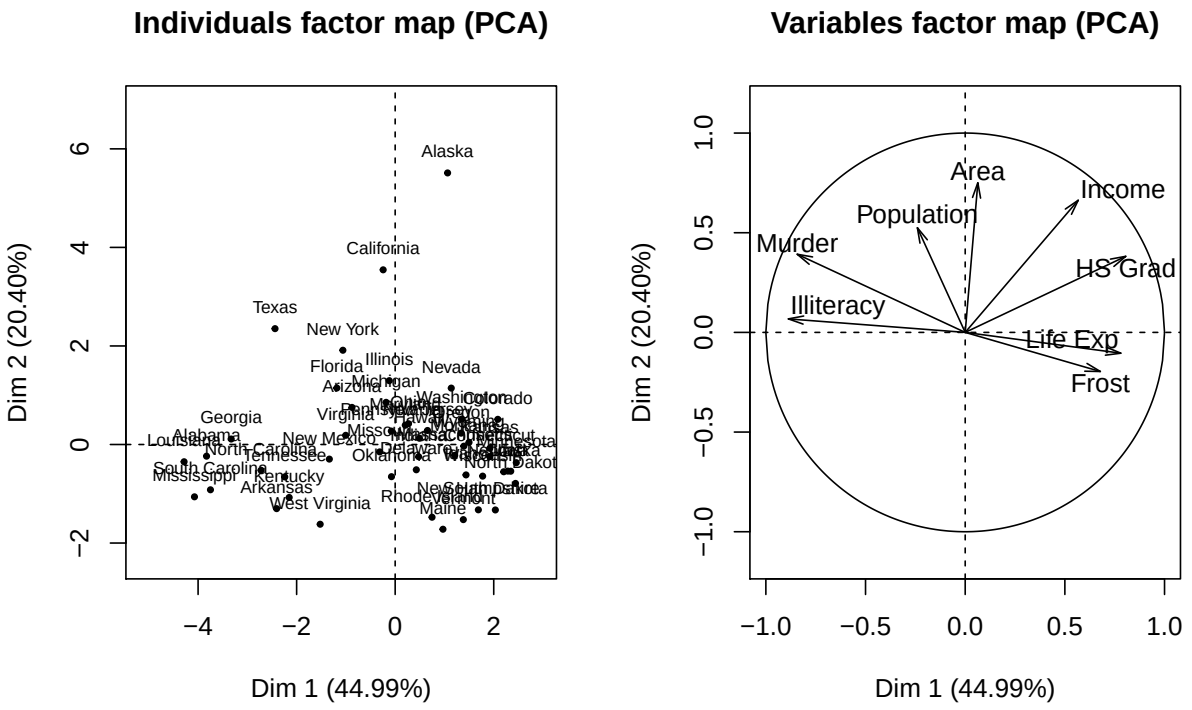
```
             Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
Alabama  -3.8283643 -0.2371626 0.23164558 -0.3871601 -0.2500637
Alaska    1.0638275  5.5115692 4.28364318 -0.5815183  0.1102403
Arizona  -0.8762354  0.7526258 0.07805313 -1.7362938 -0.5654377
Arkansas -2.4059587 -1.3014236 0.22505473 -0.6295345  0.6540497
```

```r
monACP$var$coord[1:4,1:5]
```

```
               Dim.1       Dim.2       Dim.3       Dim.4        Dim.5
Population -0.2398436  0.52487776 -0.69208615  0.34434757  0.251765858
Income      0.5669029  0.66297778 -0.10582738  0.07439531 -0.395428165
Illiteracy -0.8872037  0.06766573  0.07476148 -0.29677518  0.002186803
Life Exp    0.7808560 -0.10431289 -0.37954435 -0.37225450  0.202555453
```

Par défaut, lorsque l'on réalise l'ACP via l'appel à la fonction `PCA`, les graphiques montrant les coordonnées des individus et des variables dans le nouveau repère sont produits. On peut aussi les produire par l'appel à la fonction `plot`, de la manière suivante:

```r
layout(matrix(1:2,nrow=1))
plot(monACP, choix="ind", cex=0.7)
plot(monACP, choix="var")
```



Par défaut, les graphiques montrent le premier plan factoriel (c'est-à-dire les scores sur les deux premiers axes).

Le premier axe semble déterminé en grande partie par les variables Illiteracy, Murder, HS Grad, et Life Exp et Frost, tandis que le second axe semble lié aux variables Area et Population.

On peut interpréter deux variables proches (par exemple Murder et Illlliteracy) comme étant a priori corrélées positivement, tandis que deux variables opposées (par exemple Life Exp et Illiteracy) sont a priori corrélées négativement.

```
cor(USA$Murder, USA$Illiteracy)
```

```
[1] 0.7029752
```

```
cor(USA$"Life Exp", USA$Illiteracy)
```

```
[1] -0.5884779
```

A contrario, deux variables placées à peu près selon un angle droit (par exemple, Illiteracy et Area) sont a priori plutôt indépendantes l'une de l'autre.

```
cor(USA$Area, USA$Illiteracy)
```

```
[1] 0.07726113
```

En terme de scores, on voit que le long de l'axe 1, des états comme la Géorgie, l'Alabama, la Caroline du Sud, la Louisiane et le Mississipi se détachent. Ils doivent a priori être caractérisés par des valeurs hautes de

Illiteracy et Murder et basses de HS Grad, Life Exp et Frost. Le long de l'axe 2, ce sont des états comme l'Alaska, ou la Californie qui se détachent. Ils sont à priori caractérisés par des valeurs hautes de Population ou de Area.

Notez que cette représentation sur seulement deux axes montre une image approximative du jeu de données. Par exemple, d'après cette représentation, on pourrait supputer que l'Alaska a une valeur haute de Population, or il n'en est rien. En effet, à l'échelle du jeu de données, il y a certes une corrélation entre Population et Area. Mais si l'Alaska a une valeur de Area haute, sa population est en revanche très faible...

2.1.3 Contributions `ind$contrib` et `var$contrib` et cosinus `ind$cos2` et `var$cos2`

Au-delà des coordonnées des individus et variables dans l'espace défini par les composantes factorielles, on examine souvent des tableaux correspondant aux

- **contributions absolues** (ou plus simplement, "contributions") des variables et individus aux axes. Ces valeurs reflètent dans quelle mesure la direction d'un axe est déterminée par un individu ou une variable. Il s'agit du ratio entre l'inertie correspondant à l'individu ou à la variable et l'inertie totale portée par l'axe.
- **contributions relatives** (ou **cosinus**^{2}) des variables et individus aux axes. Ces valeurs reflètent la qualité de représentation de l'individu ou de la variable considérés par l'axe. Ces valeurs sont "visibles" sur le graphique comme correspondant au cosinus au carré de l'angle formé entre le vecteur reliant l'origine au point (représentant l'individu ou la variable) et l'axe en question.

Ainsi, en terme de contributions des variables aux axes,

```
monACP$var$contrib[,1:2]
```

```
               Dim.1      Dim.2
Population  1.5984061 16.8817585
Income      8.9299316 26.9339033
Illiteracy 21.8714450  0.2805685
Life Exp   16.9423099  0.6667720
Murder     19.7364030  9.4217898
HS Grad    18.0356857  8.9261493
Frost      12.7743655  2.3588074
Area        0.1114532 34.5302512
```

ce sont surtout Illiteracy, Murder, HS Grad, et Life Exp qui contribuent à l'axe 1, et Area, Population, et Income qui contribuent à l'axe 2.

En terme de qualité de représentation, ce sont les mêmes variables qui ressortent:

```
monACP$var$cos2[,1:2]
```

```
                Dim.1       Dim.2
Population 0.057524967 0.275496661
Income     0.321378913 0.439539542
Illiteracy 0.787130472 0.004578651
Life Exp   0.609736046 0.010881180
Murder     0.710292537 0.153755998
HS Grad    0.649085498 0.145667546
Frost      0.459736076 0.038493831
Area       0.004011086 0.563505803
```

2.1.4 variables supplémentaires

En plus des variables ayant un effet sur la production des composantes principales et axes associés (variables actives) on peut ajouter des variables supplémentaires et, notamment, les rendre visibles sur les graphiques.

Considérons par exemple la région des USA à laquelle les différents états appartiennent:

```
levels(state.region)
```

```
[1] "Northeast"     "South"         "North Central" "West"
```

On rajoute cette variable au jeu de données (pour donner le jeu de données USAr) et on fait tourner l'ACP sur ce jeu de données, en précisant le numéro de colonne où figure cette variable supplémentaire (ici, une variable qualitative). Ici, on va également traiter les variable Area et Population comme des variables supplémentaires.

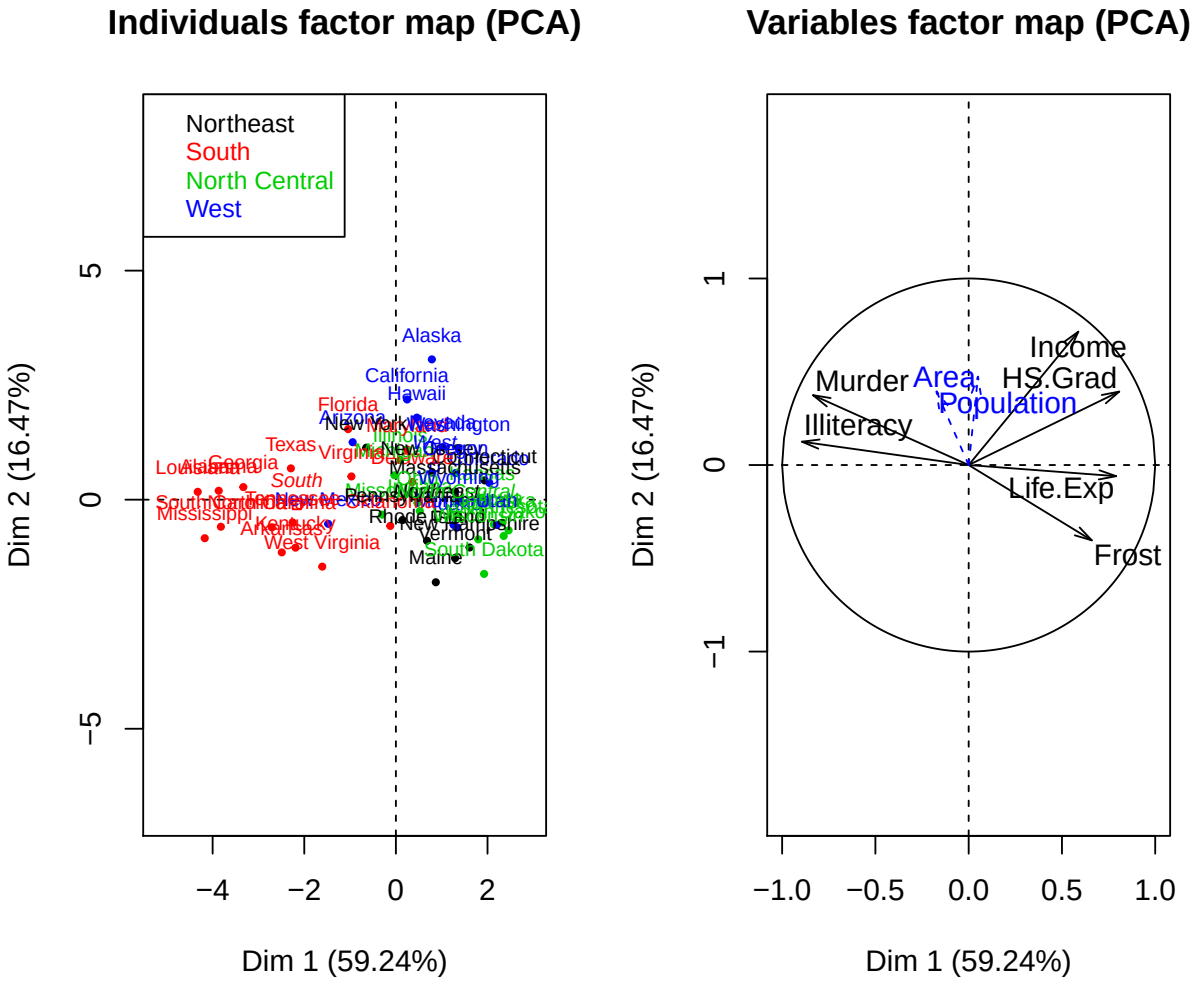
```
USAr=data.frame(region=state.region,USA)
print(colnames(USAr))
```

```
[1] "region"     "Population" "Income"     "Illiteracy" "Life.Exp"
[6] "Murder"     "HS.Grad"    "Frost"      "Area"
```

```
monACP=PCA(USAr, quali.sup=1, quanti.sup=c(2,9), graph=FALSE)
```

Il est alors possible d'afficher ces variables sur le graphique montrant les scores:

```
layout(matrix(1:2, nrow=1))
plot(monACP, choix="ind", cex=0.7, habillage="region")
plot(monACP, choix="var")
```



2.2 Examiner les résultats de l'ACP avec le package `explor`

Si les quelques fonctions examinées précedemment permettent d'afficher graphiquement les résultats de l'ACP, les graphiques produits ne sont ni toujours esthétiques (c'est en tout cas mon avis personnel) ni très flexibles (par exemple sur l'aspect ou le positionnement des étiquettes).

Le tout récent package `explor` permet de pallier ces défauts. Installez-le et testez la commande suivante:

```
# installation, si nécessaire:
# install.packages("explor")
layout(matrix(1:2), nrow=1)
require("explor") # chargement
explor(monACP)
```

Réaliser une ACP sur le jeu de données notes. Décrire les résultats:

- A-t-on réussi à bien décrire le jeu de données à travers les résultats de l'ACP?
- Quelles sont les variables qui pèsent le plus dans la définition des composantes principales?

- Comment les variables sont-elles inter-corrélées?

Remarque

Dans le package FactoMineR, les analyses sont réalisées, par défaut, sur des variables "réduites", i.e. transformées de manière à ce que toutes les variables aient la même variance. Cette opération consiste à donner le même poids (en terme d'inertie) aux différentes variables.

En effet, si une variable prend, par nature, des valeurs plus fortes ou plus variables que d'autres, elle va avoir tendance à peser davantage sur les résultats des analyses. Par exemple, dans le jeu de données "USA", la variable Area prend des valeurs particulièrement importantes: si l'on n'effectue pas de centrage et réducion du jeu de données, cette variables aura une poids énorme dans les résultats de l'analyse...

Vous pouvez vous en convaincre en exécutant la commande suivante:

```
PCA(USA, scale.unit=FALSE)
```

3 Autres exemples de méthodes factorielles: AFC et ACM

3.1 Un jeu de données comprenant deux variables qualitatives: l'AFC

L'AFC repose sur des principes similaires à ceux de l'ACP. Elle s'applique dans le cas où l'on considère **deux variables qualitatives**. On peut, par exemple, considérer des données classant les clients d'une banque selon leur catégorie socio-professionnelle (CSP) et leur âge:

```
dataAFC=read.table("../../datasets/dataAFC.csv",sep=";",header=T)
```

Pour comprendre ce genre de données on s'intéresse en premier lieu au **tableau de contingences** qui donne les effectifs observés en croisant les deux facteurs:

```
effectifs_observes=table(dataAFC$age,dataAFC$csp)
```

Considérons non plus les effectifs mais plutôt les proportions:

```
           agric artis cadsu emplo etudi inact inter ouvri retra somme_li
ai25         0.0   0.1   0.1   1.6   6.3   1.5   0.0   1.5   0.0     11.1
ai35         0.5   0.4   1.4   4.4   0.6   0.9   3.0   8.1   0.0     19.3
ai45         1.5   1.6   5.1   4.8   0.1   2.7   3.7   6.5   0.1     26.1
ai55         1.0   2.1   3.1   4.8   0.0   2.0   4.2   4.2   0.1     21.5
ai75         0.6   1.7   3.1   3.0   0.0   3.5   1.7   2.2   6.2     22.0
somme_col    3.6   5.9  12.8  18.6   7.0  10.6  12.6  22.5   6.4    100.0
```

A ce stade, on a "transformé" notre table de variables qualitatives en un tableau comprenant des nombres.

On ne peut pas interpréter directement ce tableau pour en déduire la force du lien entre deux modalités de facteurs. En effet, une forte valeur dans le tableau peut simplement s'expliquer par de fortes valeurs marginales (par exemple la proportion d'ouvriers âgés de moins de 45 ans est assez forte non pas parce qu'il y a un lien entre ces deux modalités, mais parce que les proportions marginales de personnes âgées de moins de 45 ans et la proportion d'ouvriers sont fortes: $ai45 * ouvri = 0.261 * 0.225 \approx 0.059$.

Il est donc utile de quantifier l'influence des proportions marginales sur les proportions observées. Cela revient à quantifier les valeurs que l'on observerait théoriquement si les deux facteurs, CSP et âge, étaient complètement indépendants.

Les effectifs dans chaque case seraient égaux à la multiplication des effectifs marginaux. Par exemple, le pourcentage d'agriculteurs âgés de moins de 25 ans serait égal au pourcentage global d'agriculteurs (3.6%) multiplié par le pourcentage global de personnes âgées de moins de 25 ans: $0.036 * 0.111 = 0.003996 \approx 0.4\%$

```
     agric artis cadsu emplo etudi inact inter ouvri retra
ai25   0.4   0.7   1.4   2.1   0.8   1.2   1.4   2.5   0.7
ai35   0.7   1.1   2.5   3.6   1.4   2.0   2.4   4.3   1.2
ai45   0.9   1.5   3.3   4.9   1.8   2.8   3.3   5.9   1.7
ai55   0.8   1.3   2.8   4.0   1.5   2.3   2.7   4.8   1.4
ai75   0.8   1.3   2.8   4.1   1.5   2.3   2.8   5.0   1.4
```

Soit, en termes d'effectifs:

```
           dataAFC$csp
dataAFC$age agric artis cadsu emplo etudi inact inter ouvri retra
       ai25  3.22  5.33 11.44 16.78  6.33  9.44 11.33 20.33  5.78
       ai35  5.59  9.24 19.84 29.08 10.98 16.37 19.64 35.24 10.01
       ai45  7.59 12.56 26.96 39.52 14.92 22.25 26.70 47.90 13.61
       ai55  6.23 10.31 22.13 32.44 12.24 18.26 21.91 39.31 11.17
       ai75  6.37 10.55 22.63 33.18 12.53 18.68 22.41 40.21 11.43
```

On s'intéresse à la différence entre les effectifs observés et les effectifs théoriques, divisée par les effectifs théoriques (ainsi, par exemple, observer 17 au lieu de 15 a moins de poids qu'observer 4 au lieu de 2).

On obtient ainsi une table qui quantifie la force de la corrélation entre les modalités des deux facteurs. La métrique

$$\frac{(Eff_{obs} - Eff_{theo})}{\sqrt{(Eff_{theo})}}$$

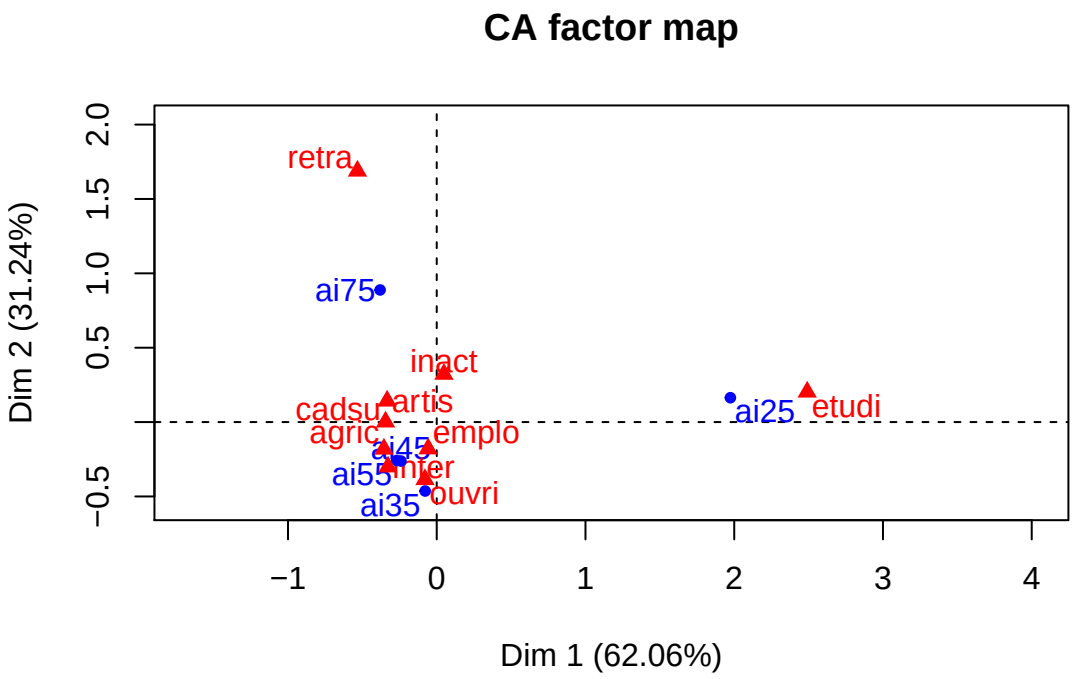
pour chacune des cases du tableau, correspond à la contribution de cette case à la distance dite **du** χ^2.

```
           dataAFC$csp
dataAFC$age agric artis cadsu emplo etudi inact inter ouvri retra
       ai25 -1.80 -1.88 -3.09 -0.92 17.75  0.83 -3.37 -1.85 -2.40
       ai35 -0.67 -2.05 -1.98  1.28 -1.80 -2.32  0.98  5.18 -3.16
       ai45  1.60  0.12  2.70 -0.08 -3.60 -0.05  0.64  0.74 -3.42
       ai55  0.71  2.08  0.61  1.15 -3.50 -0.53  2.58 -0.85 -3.04
       ai75 -0.54  1.06  0.50 -1.59 -3.54  2.16 -1.78 -3.50 11.41
```

On va dès lors pouvoir traiter ce tableau en réalisant le même type d'opérations ayant présidé au calcul de l'ACP. En effet, on va essayer de résumer en un graphique les liens existant entre les **colonnes** (i.e., non plus les variables comme dans le cas de l'ACP, mais les **modalités du premier facteur**), et les liens existant entre les **lignes** (i.e., non plus les individus comme dans le cas de l'ACP, mais les **modalités du second facteur**).

Pour cela, on applique la fonction `CA` au tableau de contingences qui nous intéresse:

```
table_contingences=table(dataAFC$age,dataAFC$csp)
monAFC=CA(table_contingences)
```



Remarque: vous pouvez ici aussi utiliser le package `explor`

Exercice 1

Interprétez les résultats de l'AFC. Quels sont les modalités les plus liées entre elles? Quelles sont les modalités de facteurs qui contribuent le plus à la construction des axes?

3.2 De nombreuses variables qualitatives: l'ACM (Analyse des Correspondances Multiples)

On peut généraliser le principe de l'AFC à plus de deux facteurs en réalisant une ACM. Pour ce faire, on (enfin, la fonction MCA) se base sur une table disjonctive qui "éclate" les variables qualitatives en autant de variables binaires:

```
tableau=data.frame(var=c("youpla","boum","tagada","tagada","tsoin","tsoin","boum","youpla"))
tableau
```

```
     var
1 youpla
2   boum
3 tagada
4 tagada
5  tsoin
6  tsoin
7   boum
8 youpla
```

```
tab.disjonctif(tableau)
```

```
  boum tagada tsoin youpla
1    0      0     0      1
2    1      0     0      0
3    0      1     0      0
4    0      1     0      0
5    0      0     1      0
6    0      0     1      0
7    1      0     0      0
8    0      0     0      1
```

On obtient ainsi un tableau numérique à partir de données qualitatives.

Examinons le jeu de données “tea”

Il comprend six questions relatives aux moments auxquels le thé est consommé (par exemple, “Buvez-vous du thé au petit-déjeuner?”, Réponse oui ou non):

1. **breakfast**
2. **tea.time**
3. **evening**

4. **lunch**
5. **dinner**
6. **always**

Il comprend six questions relatives aux endroits où le thé est consommé (par exemple, “Buvez-vous du thé à la maison?”, Réponse oui ou non):

7. **home**
8. **work**
9. **tearoom**
10. **friends**
11. **resto**
12. **pub**

Il comprend six questions relatives au type de thé consommé, et au mode de consommation:

13. **Tea**: What kind of tea do you drink the most (black tea, green tea, flavoured tea)?
14. **How**: How do you take your tea (nothing added, with lemon, with milk, other)?
15. **sugar**: Do you add sugar to your tea (yes, no)?
16. **how**: What kind of tea do you buy (tea bags, loose tea, both)?
17. **where**: Where do you buy your tea (in the supermarket, in specialist shops, both)?
18. **price**: What kind of tea do you buy (cheapest, supermarket brand, well-known brand, luxury, it varies, I don’t know)?

Il comprend six questions relatives aux caractéristiques et modes de vie des personnes interrogées:

19. **age** Age (quantitative)
20. **sex**: Female or male (F, M)?

21. **SPC**: Socio-professional category (employee, middle,non-worker, other, worker, senior, student, workman)
22. **Sport**: regular practice of sports (sportsman,
23. **age_Q**: age category (15-24, 25-34, 35,44, 45-59, +60) Not.sportsman)?
24. **frequency**: How often do you drink tea (more than twice a day, once a day, 3 to 6 times a week, once or twice per week)?

Enfin, il comprend une série de questions relatives à l'image du thé pour les personnes interrogées:

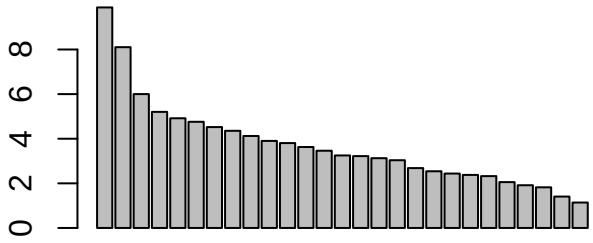
25. **escape.exoticism**: Do you consider tea to be exotic (yes, no)?
26. **spirituality**: Do you associate tea with spirituality (yes, no)?
27. **healthy**: Is tea good for your health (yes, no)?
28. **diuretic**: Is tea a diuretic (yes, no)?
29. **friendliness**: Do you associate tea with friendliness (yes, no)?
30. **iron.absorption**: Does tea stop the body from absorbing iron (yes, no)?
31. **feminine**: Is tea feminine (yes, no)?
32. **sophisticated**: Is tea refined (yes, no)?
33. **slimming**: Will tea help you to lose weight (yes, no)?
34. **exciting**: Is tea a stimulant (yes, no)?
35. **relaxing**: Is tea a relaxant (yes, no)?
36. **effect.on.health**: Does tea have any effect on your overall health (yes, no)?

Ce jeu de données contient donc beaucoup de variables! Ici on va réaliser une ACM en considérant les 18 premières variables du jeu de données comme variables actives et les 18 dernières comme des variables supplémentaires:

```
data(tea)
monACM=MCA(tea, quanti.sup=19, quali.sup=20:36, graph=FALSE)
```

Ici évidemment, la proportion d'inertie que l'on est à même d'expliquer est faible car il y a beaucoup de variables à prendre en compte...

```
barplot(monACM$eig[,2])
```

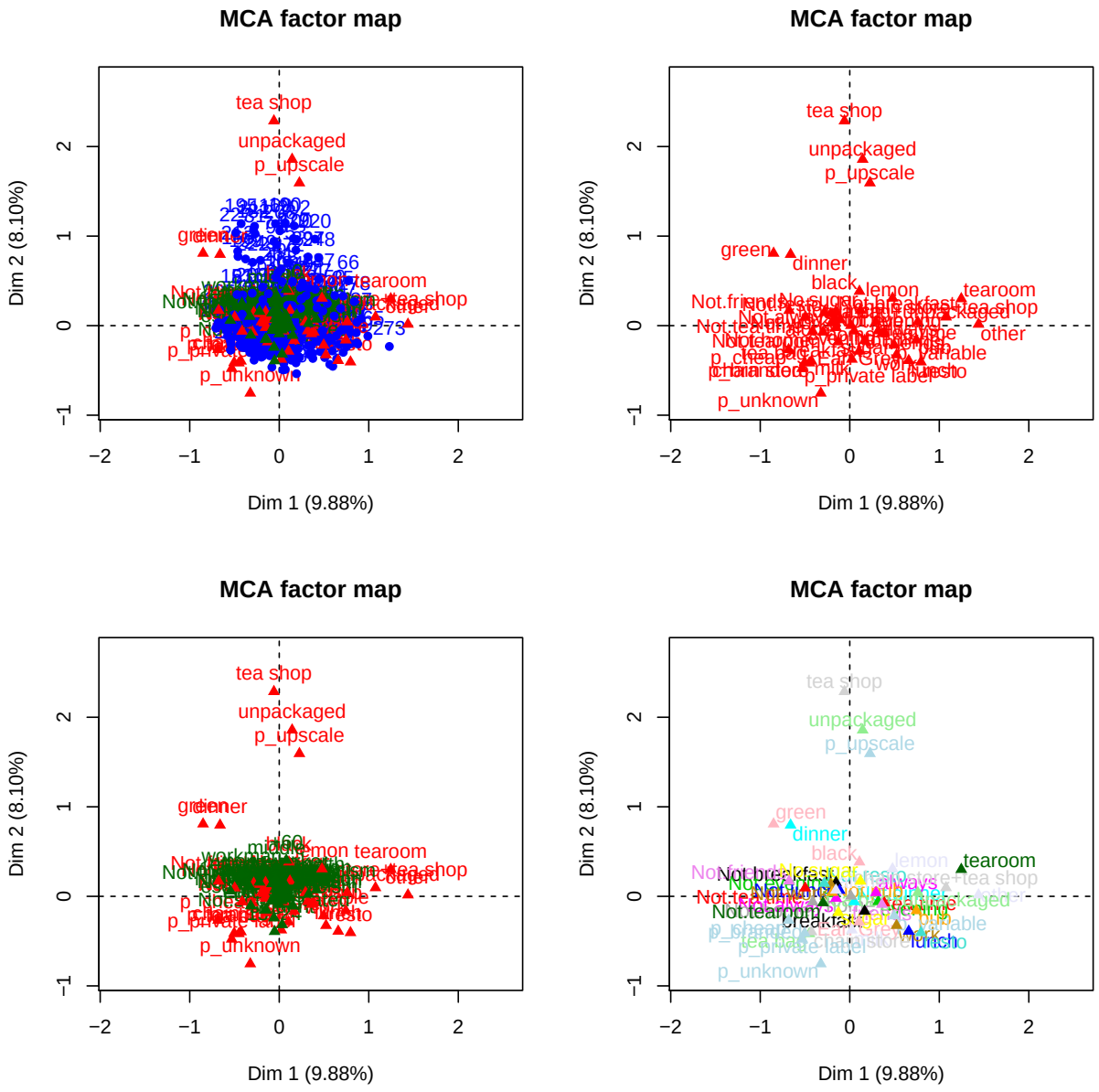


En considérant 2 axes au lieu des 27 dimensions de départ, on arrive tout de même à représenter environ 17% de l'inertie totale (ce qui n'est pas si mal car 2/27=7.4%).

Vous remarquerez qu'on a 27 dimensions pour 18 questions... En effet une question à k modalités correspond à $k-1$ dimensions...

Evidemment, avec un tel nombre de dimensions de départ, on obtient des graphiques quelque peu chargés, et il peut être intéressant de paramétrer un peu la fonction plot:

```
layout(matrix(1:4,nrow=2))
plot(monACM)
plot(monACM, invisible="ind")
plot(monACM, invisible=c("ind","quali.sup"))
plot(monACM, invisible=c("ind","quali.sup"), hab="quali")
```



Evidemment, il est aussi particulièrement intéressant ici d'utiliser le package `explor` car on a plus que jamais besoin de rendus graphiques un peu souples et interactifs...

```
explor(monACM)
```

4 Définir des classes à partir de (nombreuses) variables: la CAH

4.1 Principe

Les méthodes factorielles telles que l'ACP permettent de "représenter dans l'espace" un jeu de données. Pour comprendre la structure d'un jeu de données, il peut être tentant d'utiliser ces représentations dans l'espace pour **définir des classes**: on réalise alors une CAH (classification ascendante hiérarchique).

La CAH s'accompagne de la construction d'un arbre de classification. La figure ci-dessous illustre la manière dont la CAH se base sur les distances entre points, puis entre clusters, pour proposer une classification.

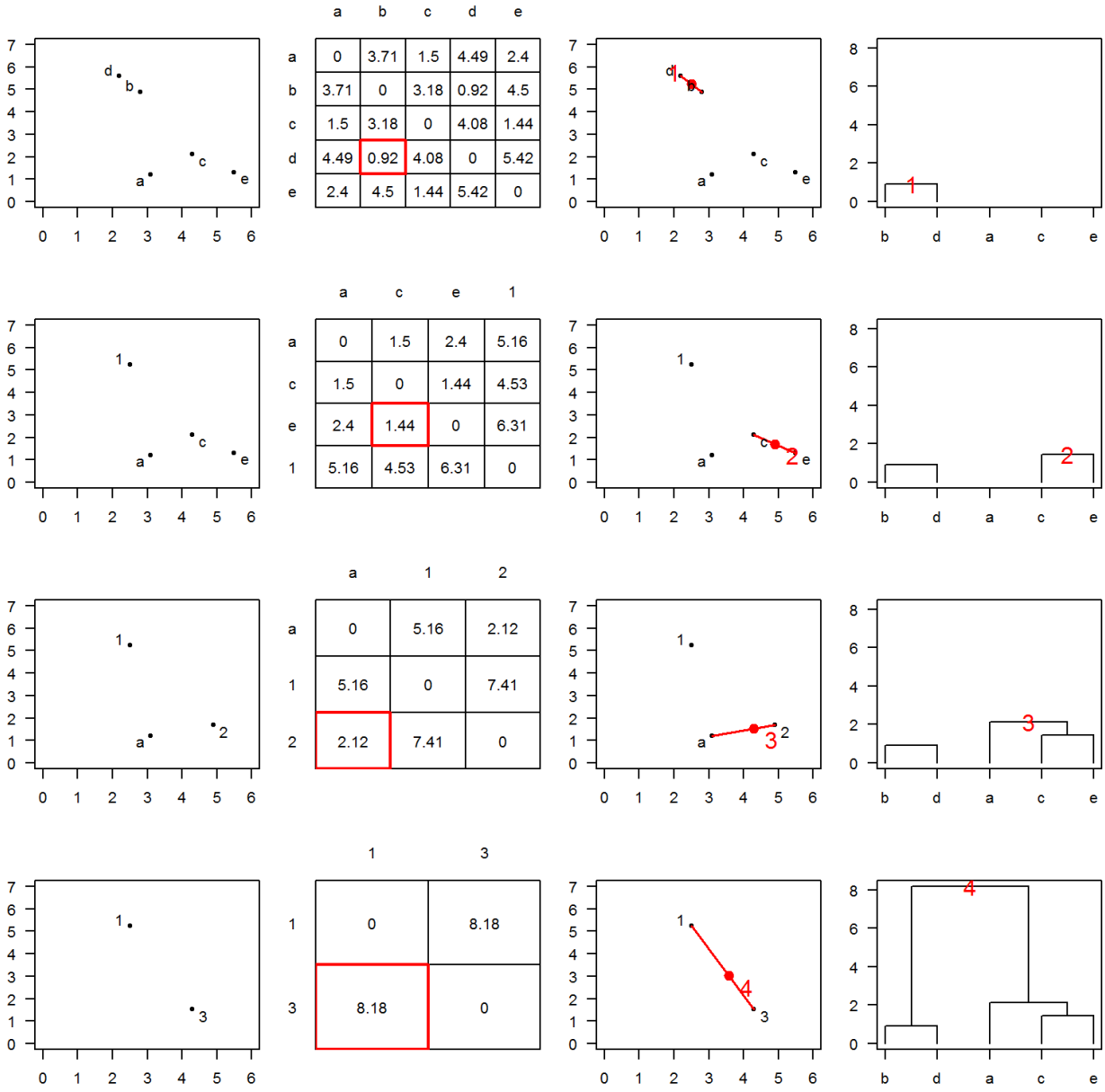
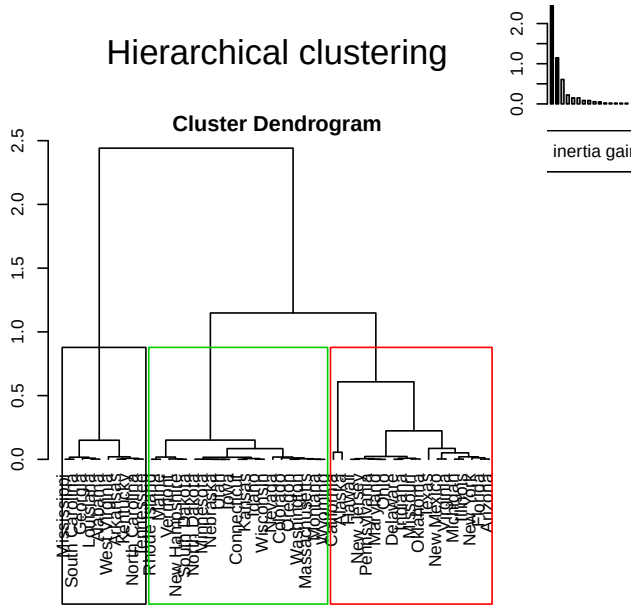


Figure 3: Algorithme de construction d'un arbre de classification: méthode de Ward,distance euclidienne

4.2 Mise en pratique: fonction `HCPC`

Ici, reprenons les données de la table USA, et réalisons à nouveau l'ACP en précisant le nombre de composantes principales à conserver. On peut réaliser une classification ascendante hiérarchique des différents états et afficher le résultat très simplement à l'aide de la fonction `HCPC`:

```
monACP=PCA(USA,ncp=2, graph=FALSE)
monHCPC=HCPC(monACP, nb.clust=3, graph=FALSE)
plot(monHCPC, choice="tree")
```




```
plot(monHCPC, choice="map")
```