DEA-AMSB 2003-2004 - Module 1
Analyse des données - Sur machine

Contrôle sur machine. Merci d'utiliser l'espace imparti pour vos réponses.

Le sujet de cet examen est disponible au format pdf à l'endroit habituel. On utilise la librairie ade4 dans R.

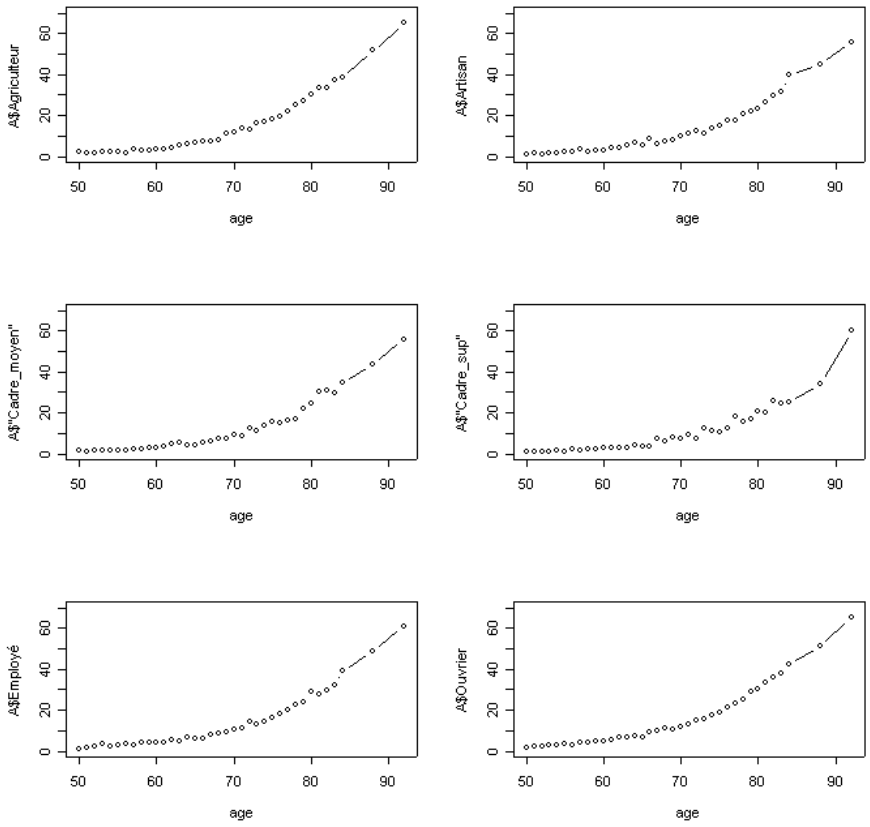
```
data(veuvage)
veuvage
$tab
   Agriculteur Artisan Cadre_sup Cadre_moyen Employé Ouvrier
1         2.06    1.00      1.25        1.37    1.13    1.83
2         1.42    1.72      1.19        1.15    1.85    2.25
...
36       51.60   44.94     33.91       43.64   48.89   51.29
37       65.03   55.87     59.89       55.95   61.12   65.59

$age
 [1] 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73
[25] 74 75 76 77 78 79 80 81 82 83 84 88 92
```

On trouve dans veuvage le pourcentage (x100) des hommes veufs pour un âge donné et dans une catégorie socio-professionnelle donnée (enquête de l'INSEE). Pour simplifier la suite, on extrait le vecteur age et on utilise les variables rangées par ordre alphabétique dans le tableau A :

```
cate = sort(names(veuvage$tab))
A = veuvage$tab[,cate]
age = veuvage$age
```

Question 1

Indiquer clairement quelles fonctions sont nécessaires pour reproduire la figure (en abscisse l'âge, en ordonnée le taux de veuvage, dans chaque fenêtre une catégorie socio-professionnelle)

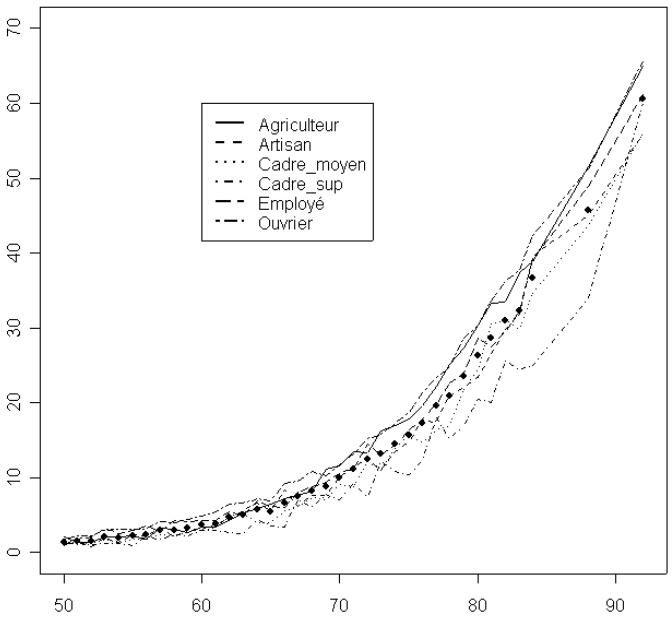


Question 2

On commence la figure par :

```
plot(age, seq(0,70,le=37),type="n",xlab="",ylab="")
for (k in 1:6) points(age,A[,k], pch=k, lty=k, type="l")
legend(60,60,leg=cate,lty=1:6,lwd=2)
```

Que fallait-il rajouter pour obtenir ce qui suit ?



Question 3

On considère les trois tableaux :

```
w11 = scalewt(A,scale=F)
w12 = sweep(A,2,apply(A,2,mean),"-")
w13 = dudi.pca(A,scale=F,scannf=F)$tab
```

Comparer ces trois objets.

Question 4

```
w21 = sweep(A,2,apply(A,2,mean),"-")
w22 = A-apply(A,1,mean)
```

Caractériser les points communs et les différences entre ces deux objets.

Question 5

On exécute deux analyses :

```
dudi1=dudi.pca(w21, scale=FALSE, scannf=FALSE, center=FALSE, nf=1)
dudi2=dudi.pca(w22, scale=FALSE, scannf=FALSE, center=FALSE, nf=1)
```

Donner pour chacune d'elles, la répartition de l'inertie entre les axes principaux.

Question 6

Justifier mathématiquement l'absence d'une valeur propre pour le second.

Question 7

Comparer le premier axe principal de chacune des deux analyses.

Question 8

Comparer la première composante principale de chacune des deux analyses.

Question 9

Représenter la première coordonnée factorielle des lignes de chacune d'entre elles en fonction du vecteur age. Comparer.

Question 10

Qu'a-t-on appris sur les données par ces observations ?

Question 11

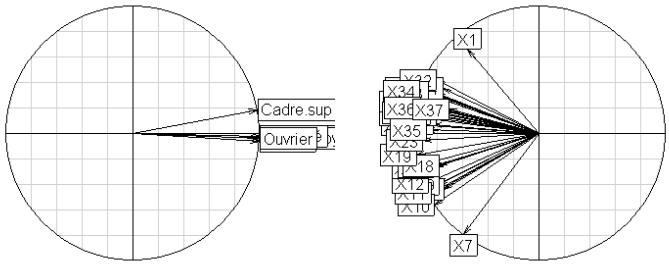
Ces deux ACP, à votre avis, disent la même chose, disent des choses contradictoires, disent des choses sans rapport ou ne disent rien d'intéressant ?

Question 12

Si l'une des deux analyses doit être qualifiée de stupide, pensez-vous qu'il s'agit de dudi1 ou de dudi2 ?

Question 13

Écrire trois lignes de R qui suffisent à reproduire la figure qui suit à partir de l'objet veuvage et qui n'a aucun intérêt statistique :



Question 14

Quand on fait

```
acpade4 = dudi.pca (A,scale=FALSE,scannf=FALSE,nf=6)
```

on obtient

```
acpade4$c1
                CS1      CS2      CS3      CS4      CS5      CS6
Agriculteur 0.4527 -0.09270  0.29534 -0.70901  0.0605  0.43917
Artisan     0.3907 -0.29233 -0.51642  0.32739  0.5223  0.33942
Cadre.moyen 0.3879 -0.14198  0.61238  0.60318 -0.2505  0.16659
Cadre.sup   0.3446  0.92259 -0.05458  0.09958  0.1298  0.01915
Employé     0.4121 -0.07014 -0.50248 -0.05735 -0.7491 -0.09106
Ouvrier     0.4510 -0.17234  0.12493 -0.11442  0.2876 -0.80963
```

Est-il possible de retrouver ce résultat exactement avec la fonction prcomp ?

Question 15

Est-il possible de retrouver ce résultat exactement avec la fonction princomp ?

Question 16

Comment faire l'ACP du tableau étudié centré par ligne et par colonne ? Quelle est alors la signification du tableau ?

Question 17

Dans l'ACP doublement centrée observe-t-on encore un élément de structure ?

Question 18

Exécuter les ordres qui suivent. Observer le résultat. Qu'apprend-t-on de nouveau ?

```
w1 = as.numeric(log(as.matrix(A)))
csp.fac=as.factor(rep(cate,rep(37,6)))
age.rep=rep(veuvage$age,6)
lm1 = lm (w1~age.rep+csp.fac)
coe1=coefficients(lm1)
alpha=coe1[2]
coe1 = c(coe1[1],(coe1[3:7]+coe1[1]))
par(mfrow=c(2,3))
par(mar=c(3,2,1,1))
for (k in 1:6) {
    z = log(A[,k])
    plot(age,z,ylab="",ylim=c(-0.5,4.5))
    abline(lm(z~age))
    abline (c(coe1[k],alpha),col="red",lwd=2)
    text(70,-0.25,cate[k],cex=2)
}
```

Question 19

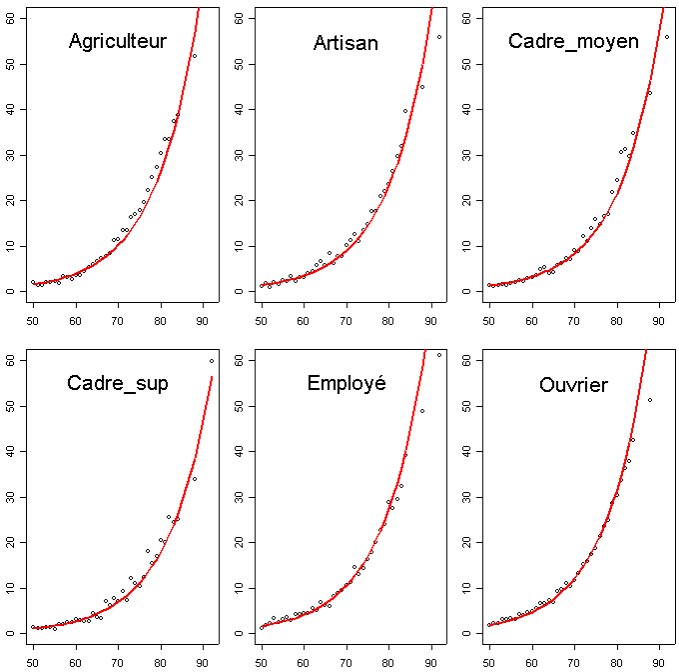
Que suffit-il alors d'ajouter pour avoir la figure ci-dessous.

Question 20

On achève cette réflexion par :

```
plot(dudi1$l1[,1],exp(veuvage$age*alpha))
abline(lm(I(exp(veuvage$age*alpha)~dudi1$l1[,1])))
```

Expliquer pourquoi une telle proximité et ce qu'elle implique.



DEA-AMSB 2003-2004 - Module 1
Solution

Question 1

```
par(mfrow=c(3,2))
plot(age,A$"Agriculteur",type="b",ylim=c(0,70))
plot(age,A$"Artisan",type="b",ylim=c(0,70))
plot(age,A$"Cadre_moyen",type="b",ylim=c(0,70))
plot(age,A$"Cadre_sup",type="b",ylim=c(0,70))
plot(age,A$"Employé",type="b",ylim=c(0,70))
plot(age,A$"Ouvrier",type="b",ylim=c(0,70))
```

Il y a quatre difficultés : le multifenêtrage 3-2, le passage des noms de variables avec un underscore, le type "b" pour trait+point, les bornes identiques dans les 6 fenêtres par ylim.

Question 2

Il faut rajouter la moyenne par ligne dans le tableau :

```
points(age,apply(A,1,mean),pch=20,cex=1.5)
```

Question 3

Ces trois tableaux sont identiques. Ils ont 37 lignes et 6 colonnes et sont centrés par variables-colonnes. Il s'agit du tableau ordinaire de l'ACP centrée.

```
max(w11-w12)
[1] 0
max(w11-w13)
[1] 0
max(w12-w13)
[1] 7.105e-15
```

Question 4

```
dim(w21)
[1] 37  6
 dim(w22)
[1] 37  6
apply(w21,1,sum)
      1       2       3       4       5       6       7       8       9
-72.751 -71.811 -72.441 -68.281 -69.511 -67.751 -66.681 -63.461 -63.211
...
apply(w21,2,sum)
Agriculteur     Artisan Cadre.moyen   Cadre.sup     Employé     Ouvrier
  3.375e-14   1.243e-14  -3.553e-15  -1.688e-14  -4.441e-15   9.948e-14
apply(w22,2,sum)
Agriculteur     Artisan Cadre.moyen   Cadre.sup     Employé     Ouvrier
      56.98      -19.42      -41.93     -113.72       23.36       94.75
 apply(w22,1,sum)
         1          2          3          4          5          6          7
-8.882e-16 -2.220e-16  9.992e-16 -6.661e-16  1.332e-15  0.000e+00 -2.220e-16
 ...
```

Ils ont même dimension. Le premier est centré par colonne. Le second est centré par ligne. Dans le premier on a enlevé, pour chaque colonne, la valeur moyenne pour la catégorie socio-professionnelle. Dans le second on a enlevé, pour chaque ligne, la valeur moyenne de la classe d'âge. Ils ont des significations radicalement différentes.

Question 5

```
dudi1$eig/sum(dudi1$eig)
[1] 0.9915268 0.0054054 0.0012269 0.0010372 0.0005719 0.0002319
Pratiquement un seul axe exprime la quasi totalité de la variabilité.
dudi2$eig/sum(dudi2$eig)
[1] 0.84216 0.08712 0.03817 0.01654 0.01601
C'est moins caricatural mais le dépouillement isolé du premier axe s'impose.
```

Question 6

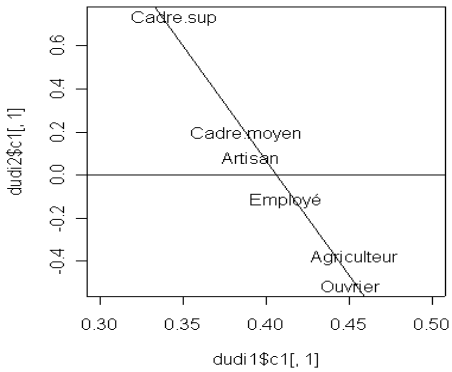
Un tableau 37-6 est de rang inférieur ou égal à 6. La somme des 6 colonnes est nulle par le centrage par ligne dans le second. Le rang est donc au plus égal à 5.

Question 7

```
dudi1$c1                  dudi2$c1
              CS1                       CS1
Agriculteur 0.4527        Agriculteur -0.38011
Artisan     0.3907        Artisan      0.08364
Cadre.moyen 0.3879        Cadre.moyen  0.19562
Cadre.sup   0.3446        Cadre.sup    0.73005
Employé     0.4121        Employé     -0.11542
Ouvrier     0.4510        Ouvrier     -0.51377
```

Différence : les composantes du premier sont toutes positives, les composantes du second sont de signe variable.

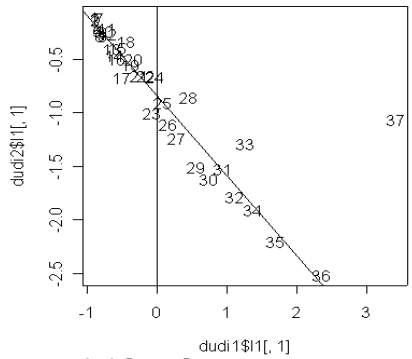
```
plot(dudi1$c1[,1],dudi2$c1[,1],type="n",xlim=c(0.3,0.5))
text(dudi1$c1[,1],dudi2$c1[,1],row.names(dudi1$c1))
abline(lm(dudi2$c1[,1]~dudi1$c1[,1]))
abline(h=0)
```



Ressemblance : l'ordre des valeurs des composantes est le même au signe près et reproduit la hiérarchie traditionnelle des classes sociales.

Question 8

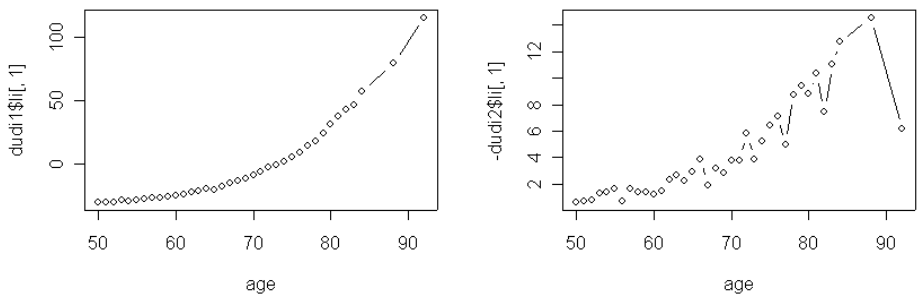
```
plot(dudi1$l1[,1],dudi2$l1[,1],type="n")
text(dudi1$l1[,1],dudi2$l1[,1],row.names(dudi1$l1))
abline(v=0)
abline(lm(dudi2$l1[-37,1]~dudi1$l1[-37,1]))
```



Différence : les composantes de la première sont de signe variable, les composantes de la seconde sont toutes négatives.
Ressemblance : l'ordre des valeurs des composantes est le même au signe près si on isole la dernière (92 ans).

Question 9

```
par(mfrow=c(1,2))
plot(age,dudi1$li[,1],type="b")
plot(age,-dudi2$li[,1],type="b")
```



On peut changer de signe une coordonnée sans changer sa nature (il suffit de changer le sens de l'axe). Ressemblance : une évolution continue croissante, deux différences : la première est régulière, la seconde plus chaotique ; la première est continue, la seconde présente un accident en fin de parcours.

Question 10

La première analyse est centrée par variable, les poids des variables sont tous du même signe, elle exprime un compromis. Elle rend compte de la différence entre les âges qui se reproduit dans chacune des CSP : le veuvage croît régulièrement avec l'âge toutes CSP confondues. D'où le signe des composantes, la régularité de la courbe, le taux d'inertie caricatural.
Cet effet est éliminé par le centrage par lignes dans la seconde analyse. L'axe principal a des composantes de signe varié : elle oppose les cadres supérieurs aux ouvriers - agriculteurs. Les premiers ont une courbe en dessous de la moyenne, les second au-dessus. L'inégalité sociale croît plus ou moins régulièrement avec l'âge mais vers 90 ans tous se retrouvent à la moyenne (60%).

Question 11

On pourrait croire, vues les ressemblances tant dans la répartition de l'inertie que dans l'ordre des composantes des éléments principaux que ces deux analyses disent la même chose. Il n'en est rien bien au contraire. La première analyse exprime l'inégalité issue de l'âge, la seconde l'inégalité sociale, les deux sont croissantes avec l'âge. La ressemblance technique entre les deux analyses est liée au mode de variation de la mesure en fonction de l'âge.

Question 12

La première enfonce une porte ouverte : les courbes des données en disent assez long. Toutes CSP confondues, le veuvage masculin passe de 3 à 60% entre 50 et 90 ans, chez les hommes survivants. Une ACP pour montrer cette évidence est une bêtise. La seconde qui montre que l'écart social se creuse régulièrement est plus utile. dudi2 est préférable à dudi1. Hors sujet : la fin du problème montre que *in fine* la coordonnée de **dudi1** est un modèle raffiné et que la coordonnée de **dudi2** est un effet parasite du problème non posé correctement.

Question 13

Il s'agit des deux ACP normées

```
par(mfrow=c(1,2))
s.corcircle(dudi.pca(A,scannf=FALSE)$co)
s.corcircle(dudi.pca(t(A),scannf=FALSE)$co)
```

Question 14

La réponse est OUI. Consulter la documentation.

```
?prcomp

acpprcomp = prcomp(A)

acpprcomp$rotation
                PC1       PC2       PC3       PC4      PC5       PC6
Agriculteur 0.4527 -0.09270  0.29534 -0.70901  0.0605  0.43917
Artisan     0.3907 -0.29233 -0.51642  0.32739  0.5223  0.33942
Cadre_moyen 0.3879 -0.14198  0.61238  0.60318 -0.2505  0.16659
Cadre_sup   0.3446  0.92259 -0.05458  0.09958  0.1298  0.01915
Employé     0.4121 -0.07014 -0.50248 -0.05735 -0.7491 -0.09106
Ouvrier     0.4510 -0.17234  0.12493 -0.11442  0.2876 -0.80963
On a bien exactement les mêmes résultats sur les axes principaux.
```

Question 15

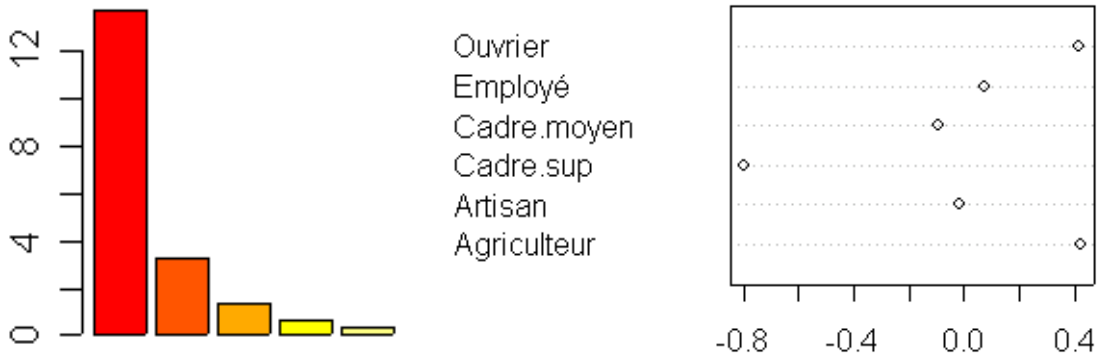
```
Consulter la documentation.
?princomp
acpprincomp=princomp(A)
acpprincomp$loadings

Loadings:
            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Agriculteur  0.453         -0.295  0.709         -0.439
Artisan      0.391  0.292  0.516 -0.327  0.522 -0.339
Cadre_moyen  0.388  0.142 -0.612 -0.603 -0.250 -0.167
Cadre_sup    0.345 -0.923                0.130
Employé      0.412         0.502        -0.749
Ouvrier      0.451  0.172 -0.125  0.114  0.288  0.810

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000

On a des valeurs voisines mais une édition qui laisse des blancs. On n'est pas sûr d'avoir
exactement les mêmes valeurs. Si on pense qu'il ne s'agit que d'un format d'édition, on peut
contourner la fonction générique :
unclass(acpprincomp$loadings)
            Comp.1   Comp.2   Comp.3   Comp.4  Comp.5   Comp.6
Agriculteur 0.4527  0.09270 -0.29534  0.70901  0.0605 -0.43917
Artisan     0.3907  0.29233  0.51642 -0.32739  0.5223 -0.33942
Cadre_moyen 0.3879  0.14198 -0.61238 -0.60318 -0.2505 -0.16659
Cadre_sup   0.3446 -0.92259  0.05458 -0.09958  0.1298 -0.01915
Employé     0.4121  0.07014  0.50248  0.05735 -0.7491  0.09106
Ouvrier     0.4510  0.17234 -0.12493  0.11442  0.2876  0.80963
La réponse est OUI, on a exactement les mêmes axes principaux.
```

Question 16

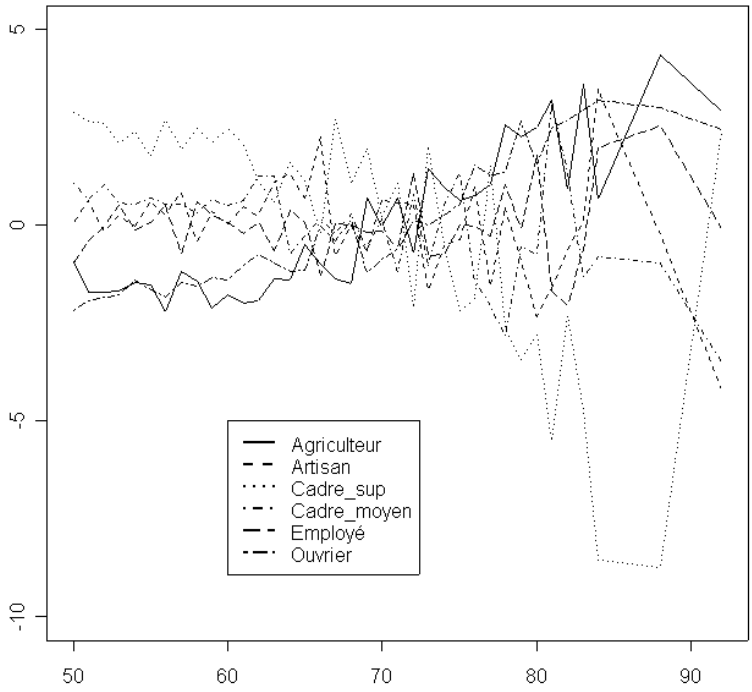
```
Pour avoir le  tableau doublement centré (par exemple) :
w1=scale(t(scale(t(A),scale=F)),scale=F)
dudi3 = dudi.pca(w1,scale=F,center=F)
Select the number of axes: 1
Le tableau contient les valeurs initiales corrigées de la moyenne par CSP et de la moyenne par
classe d'âge du type x[i,j]-x[i,.]-x[.,j]+x[.,.] le . désignant la moyenne sur l'indice associé.
```

Question 17

```
dotchart(dudi3$c1[,1],lab=row.names(dudi3$c1))
```



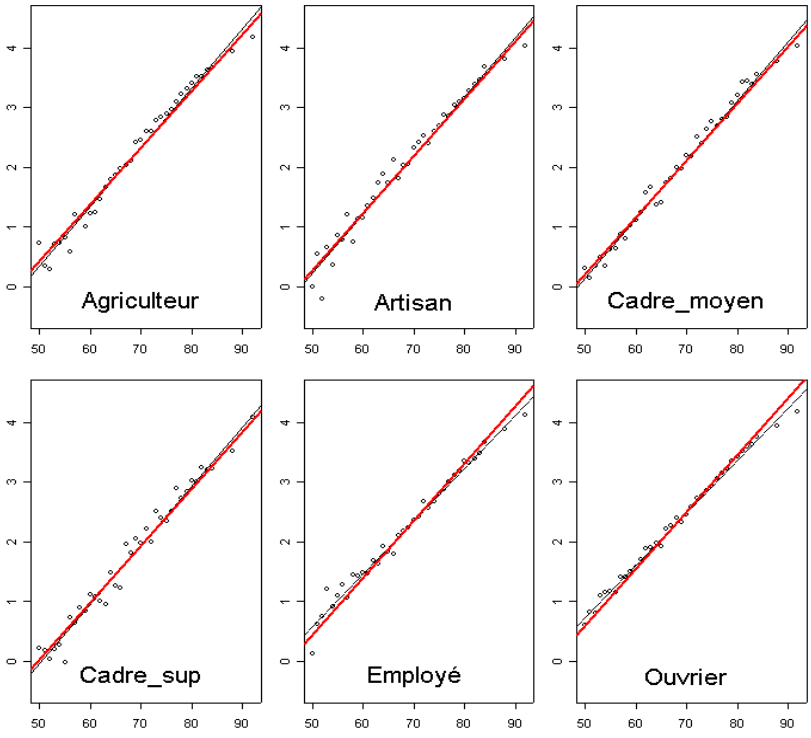

```
Il semble qu'un facteur soit interprétable.

plot(age, seq(-10,5,le=37),type="n",xlab="",ylab="")
for (k in 1:6) points(age,dudi3$tab[,k],pch=k,lty=k,type="l")
legend(60,-5,leg=cate,lty=1:6,lwd=2)
```



On retrouve des courbes de résidus structurés qui se coupent aux environs de 70 ans. Le modèle observations = effet âge + effet CSP a des résidus structurés par ce que l'effet n'est pas additif mais multiplicatif. Une décomposition en valeurs singulières directe (double centrage multiplicatif) éliminerait cet effet.

Question 18

On obtient :



Le logarithme du taux de veuvage est sensiblement une fonction affine de l'âge. De plus les droites d'ajustement par catégorie sont très voisines d'une droite commune à un effet additif près.

```
anova(lm1)
Analysis of Variance Table

Response: w1
           Df Sum Sq Mean Sq F value Pr(>F)
age         1  249.4   249.4   10266 <2e-16
csp         5    7.3     1.5      60 <2e-16
Residuals 215    5.2   0.024
```

L'intérêt est d'avoir linéarisé la relation, donc de pouvoir comparer les pentes des droites. Le modèle multiplicatif devient additif en log. L'ajustement à une droite de pente constante est presque parfait. Il n'y a donc dans les données qu'un modèle du type:

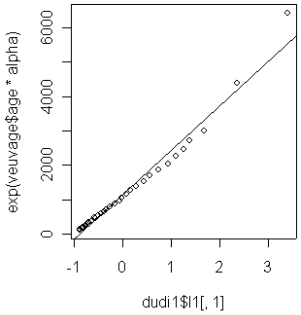
$$y = a_j \exp(bx)$$

Les courbes par CSP sont donc de même nature mais pas de même paramètre. Il n'y a pas amplification de l'inégalité sociale mais conséquence directe du changement de paramètres dans un modèle exponentiel.

Question 19

La modification est très simple :

```
for (k in 1:6) {
    z = A[,k]
    plot(age,z,ylab="",ylim=c(0,60))
    ymod=exp(coe1[k]+age*alpha)
    lines(age,ymod,col="red",lwd=2)
    text(70,55,cate[k],cex=2)
}
```



Question 20

On trouve en ordonnée le modèle commun à toutes les courbes, ajusté par un modèle linéaire sur le log. Ce modèle donne un très bon ajustement. En abscisse on a le meilleur modèle des courbes, toute catégorie, qui dérive directement des données sans a priori (principe de l'ACP de Pearson 1901). Ces deux modèles sont très voisins. Avec la famille des modèles exponentiels, on est très près de l'auto-modèle optimal. Mais pas exactement dessus. En fait la coordonnée montre que l'exponentielle est une approximation, les données montrent un processus légèrement moins accéléré. En fait on a traité de bêtise une analyse (question 12) qui montre

1) que les centrages habituels dans un sens comme dans l'autre comme dans les deux sont des erreurs,
2) qu'il convient de chercher un modèle externe simple,
3) que ce dernier, très accessible, ne convient pas totalement,
4) ou, pour simplifier, on peut se passer d'analyse en faisant simplement :

```
par(mfrow=c(2,3))
par(mar=c(3,2,1,1))
for (k in 1:6) {
  plot(age,log(A[,k]),pch=k,type="b",ylim=c(0,4.5))
  abline(lm(I(log(A[,k])~age)))
  text(70,4,cate[k],cex=1.5)
}
```