Apprentissage de données biologiques 2022-2023
Contrôle des connaissances (durée : 1h)

26 octobre 2022

Nom Prénom :

Tous les documents sont autorisés. La calculatrice est autorisée. L'usage du téléphone portable est interdit, même pour sa fonction calculatrice.

Exercice - Activité cérébrale et anxiété

Une équipe de chercheurs en neurosciences a mesuré l'activité cérébrale par électro-encéphalogramme (EEG) de 15 sujets à qui des images représentant un visage étaient présentées sur un écran. Chaque EEG mesure l'activité cérébrale dans l'intervalle débutant 200ms avant l'apparition de l'image sur l'écran et terminant 600ms après l'apparition de l'image (819 temps de mesure). Pour chaque sujet, on dispose de 8 mesures d'activité cérébrale, une pour chaque combinaison d'une image de visage neutre ou en colère (`emotion`), d'une durée courte ou longue d'exposition à l'image (`visibility`) et d'une position à droite ou à gauche sur l'écran (`direction`).

On donne ici un résumé des données (12 premières variables uniquement, les suivantes étant les 817 autres mesures d'activité cérébrale par EEG), disponibles dans le package R `permuco` :

```
summary(dta[,1:12])
```

```
       id     visibility    emotion   direction  laterality_id        age
 S01    : 8   16ms :60   angry  :60   right:60   Min.   : 25.0   Min.   :18.0
 S02    : 8   166ms:60   neutral:60   left :60   1st Qu.: 47.0   1st Qu.:20.0
 S03    : 8                                      Median : 62.0   Median :21.0
 S04    : 8                                      Mean   : 65.5   Mean   :21.5
 S05    : 8                                      3rd Qu.: 90.0   3rd Qu.:23.0
 S07    : 8                                      Max.   :100.0   Max.   :25.0
 (Other):72
    sex      STAIS_state    STAIS_trait       X.200             X.199
 male  :56   Min.   :21.0   Min.   :42.0   Min.   :-1.762   Min.   :-1.792
 female:64   1st Qu.:23.0   1st Qu.:43.0   1st Qu.:-0.672   1st Qu.:-0.651
             Median :25.0   Median :45.0   Median :-0.119   Median :-0.102
             Mean   :28.1   Mean   :46.4   Mean   : 0.193   Mean   : 0.195
             3rd Qu.:30.0   3rd Qu.:49.0   3rd Qu.: 0.319   3rd Qu.: 0.300
             Max.   :49.0   Max.   :55.0   Max.   : 7.582   Max.   : 7.549

    X.198
 Min.   :-1.8414
 1st Qu.:-0.6552
 Median :-0.0906
 Mean   : 0.1984
 3rd Qu.: 0.3115
 Max.   : 7.5160
```

Dans la suite, l'étude sert de support à l'analyse des relations entre le niveau général d'anxiété d'un patient (`STAIS_trait`) et l'activité cérébrale mesurée par EEG. Plus précisément, on cherche à répondre à deux questions :

- Peut-on prédire le score d'anxiété à partir des courbes d'activité cérébrale ?
- Si oui, le modèle d'association entre les courbes d'activités cérébrales et le score d'anxiété depend-il du genre du sujet (`sex`) ?

Question 1

Quelles sont les variables réponse et explicatives dans ces deux problématiques ? Donnez la nature (quantitative ou catégorielle) de ces variables et, si catégorielle, leurs modalités ?

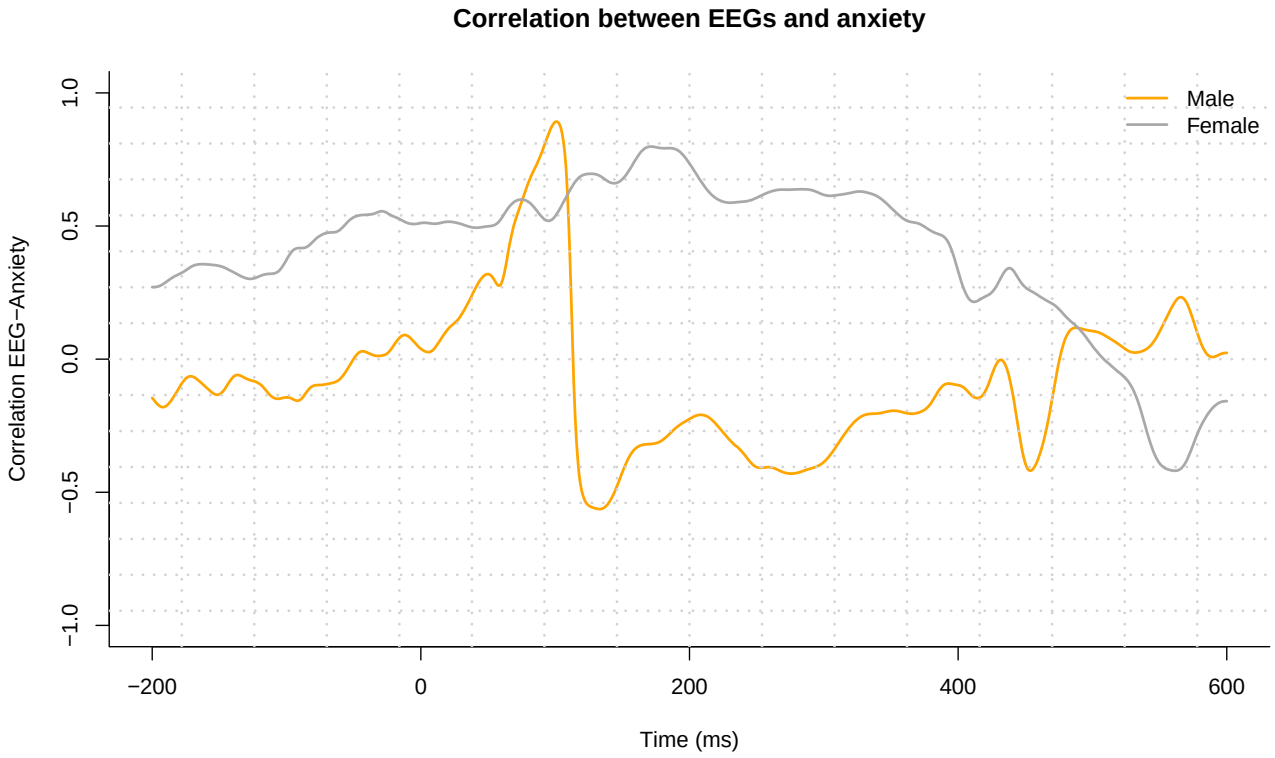
Réponse

Question 2

Donnez l'expression mathématique d'un modèle paramétrique linéaire permettant d'étudier l'association entre le score d'anxiété et l'activité cérébrale, avec potentiellement des paramètres d'association différents selon le genre du sujet. Combien ce modèle compte-t'il de paramètres ?

Réponse

Le graphique suivant représente les courbes de corrélations entre le score d'anxiété et les courbes d'activité cérébrale, et ce en distinguant les deux genres :



Question 3

Selon vous, le graphique précédent encourage-t'il à considérer que le modèle d'association entre l'anxiété et l'activité cérébrale est le même pour les sujets masculins et féminins ? (justifiez brièvement votre réponse).

Réponse

Afin de faire face à la grande dimension des données, on utilise dans la suite la méthode de la regression *Partial Least Squares* pour estimer le modèle d'association :

```
cvmod <- pls::plsr(STAIS_trait~.,data=dta[,-c(1:8)],
              validation="CV",ncomp=50)
```

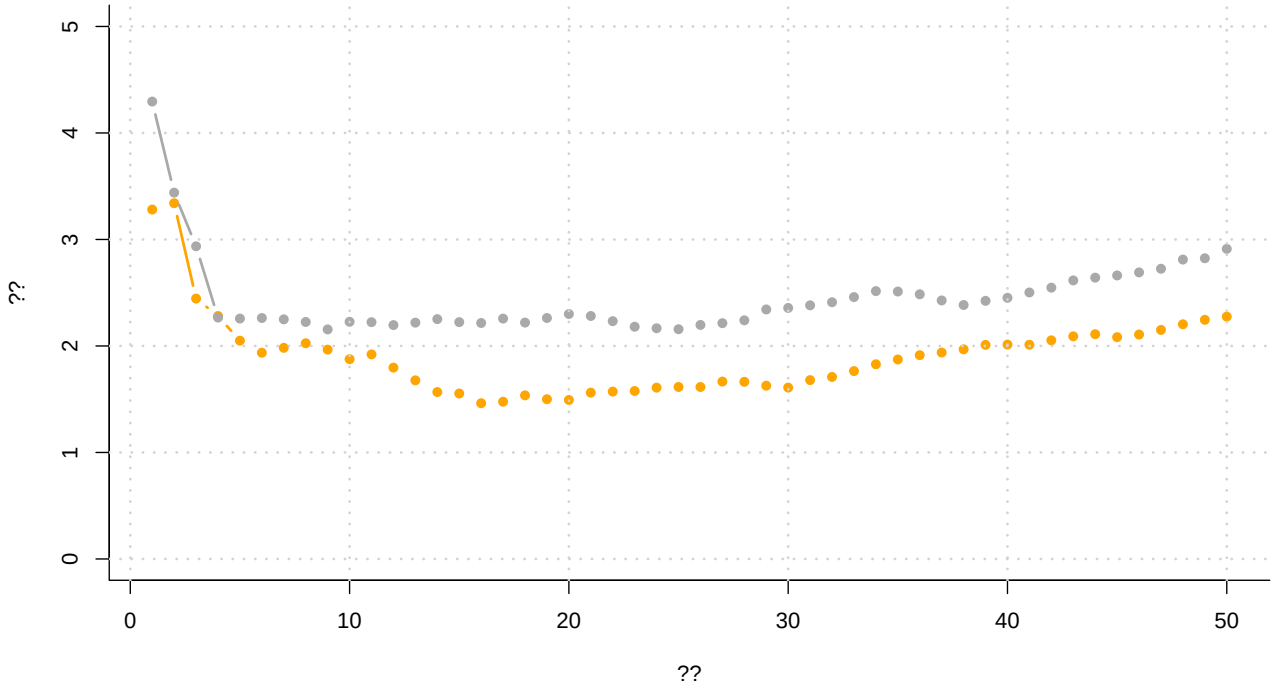
Question 4

Les modèles d'association estimés ci-dessus par la fonction `plsr` ont-ils des coefficients différents selon le genre du sujet ? (justifiez brièvement votre réponse)

Réponse

Le graphique ci-après permet de choisir le nombre de variables latentes à retenir dans la phase de réduction de la dimension de cette méthode :

```
cv_pred <- cvmod$validation$pred[,1,]
c_1 <- apply(cv_pred[dta$sex=="male",],MARGIN=2,
               FUN=function(pred,obs) {
                 sqrt(mean((pred-obs)^2)) },
               obs=dta$STAIS_trait[dta$sex=="male"])
c_2 <- apply(cv_pred[dta$sex=="female",],MARGIN=2,
               FUN=function(pred,obs) {
                 sqrt(mean((pred-obs)^2)) },
               obs=dta$STAIS_trait[dta$sex=="female"])
matplot(1:50,cbind(c_1,c_2),type="b",
     pch=16,col=c("orange","darkgray"),lwd=2,lty=1,
     bty="l",xlab="??",ylab="??",ylim=c(0,5))
grid(lwd=2)
```



Question 5

Dans le graphique précédent, quel nom donneriez-vous aux axes ? Que représentent les courbes grise et orange ? Enfin, vaut-il mieux choisir 10 ou 15 variables latentes ? (discutez brièvement ce choix)

Réponse

Si k désigne le nombre de variables latentes jugé optimal selon le graphique ci-dessus (dans la suite, $k = 10$ arbitrairement), on note L la matrice de dimension $120 \times k$, telle que L_{ij} contient la valeur de la $j$ème variable latente pour le $i$ème individu :

```
best_ncomp <- 10
L <- cvmod$scores[,1:best_ncomp]
colnames(L) <- paste("L",1:best_ncomp,sep="")
round(head(L),digits=2)
```

```
        L1    L2    L3     L4     L5     L6     L7     L8    L9    L10
27  -15.12  1.39  2.59   8.41 -22.00 -15.24 -31.62 -21.77  2.56  -3.64
96   60.33 -9.87  0.30 -19.13  -4.83  16.61  -9.68  -6.69  2.50 -21.03
165 -17.71  0.17  1.81  10.85 -16.71 -15.58 -37.15 -28.31  1.80  -3.38
234  58.69 -9.41 -3.61 -15.18  -2.60  19.95  -7.98  -1.57  1.11 -22.92
303 -16.26 -0.28  1.52  -6.16 -23.29 -15.04 -28.56 -14.65  0.14  -6.04
372  64.17 -6.38  1.54 -14.71   2.07  19.34  -5.36   5.99 -0.50 -19.91
```

Question 6

Selon quel critère, propre à la méthode d'estimation utilisée ci-dessus, peut-on affirmer que la variable L1 dans la matrice L prédit mieux l'anxiété que la variable L2 ?

Réponse

Question 7

Dans la méthode implémentée ci-dessus, donnez l'expression du modèle de prédiction de la variable à expliquer, à savoir l'anxiété, par toutes les variables latentes.

Réponse

On se demande ici si le modèle de prédiction de la question 5 doit être le même pour les deux genres ou s'il faut considérer un modèle différent pour chacun des deux genres :

```
anova(mod_0,mod_1)
```

```
Analysis of Variance Table

Model 1: Anxiety ~ L1 + L2 + L3 + L4 + L5 + L6 + L7 + L8 + L9 + L10
Model 2: Anxiety ~ sex * (L1 + L2 + L3 + L4 + L5 + L6 + L7 + L8 + L9 +
    L10)
  Res.Df   RSS Df Sum of Sq     F   Pr(>F)
1    109 279.9
2     98 167.3 11     112.5 5.992 2.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 8

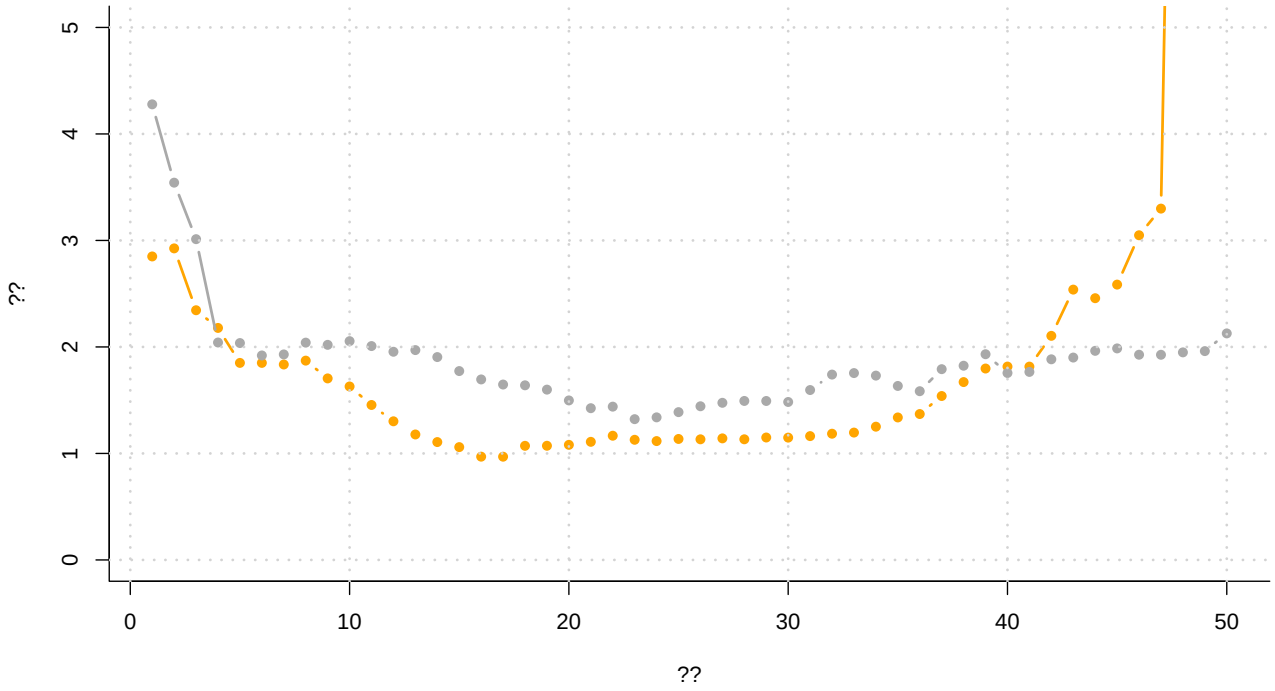
Dans le tableau ci-dessus, expliquez la valeur donnée dans la colonne `Df`. Diriez-vous que le modèle de prédiction de la question 5 doit être remplacé, on non, par un modèle dans lequel l'équation de prédiction dépend du genre du sujet ? (justifiez brièvement votre réponse)

Réponse

Dans la suite, on implémente dans une fonction `pls_z(xtrain,ytrain,ztrain,xtest,ztest,ncomp)` une nouvelle méthode de régression inspirée de la précédente mais dans laquelle l'étape de prédiction de `y` par `L` dépend aussi des modalités d'un facteur `z`. La fonction estime le modèle avec `ncomp` variables latentes sur des données d'apprentissage `(xtrain,ytrain,ztrain)` et utilise le modèle pour prédire la variable réponse sur des données test `(xtest,ztest)`. La fonction a deux sorties : `$pred` qui contient les valeurs prédites sans tenir compte de `z` et `$pred_z` qui contient les valeurs prédites en tenant compte de `z`.

```
segs <- cvsegments(N=nrow(dta),k=10)
cv_pred <- matrix(0,nrow=nrow(dta),ncol=50)
for (k in 1:10) {
   xtrain <- dta[-segs[[k]],-c(1:9)]
   xtest <- dta[segs[[k]],-c(1:9)]
   ytrain <- dta$STAIS_trait[-segs[[k]]]
   ztrain <- dta$sex[-segs[[k]]]
   ztest <- factor(dta$sex[segs[[k]]],levels=levels(ztrain))
   for (j in 1:50) {
      p <- pls_z(xtrain,ytrain,ztrain,xtest,ztest,ncomp=j)
      cv_pred[segs[[k]],j] <- p$pred_z
   }
}
c_1 <- apply(cv_pred[dta$sex=="male",],MARGIN=2,
               FUN=function(pred,obs) {
                 sqrt(mean((pred-obs)^2)) },
```

```
                obs=dta$STAIS_trait[dta$sex=="male"])
c_2 <- apply(cv_pred[dta$sex=="female",],MARGIN=2,
                FUN=function(pred,obs) {
                  sqrt(mean((pred-obs)^2)) },
                obs=dta$STAIS_trait[dta$sex=="female"])
matplot(1:50,cbind(c_1,c_2),type="b",
     pch=16,col=c("orange","darkgray"),lwd=2,lty=1,
     bty="l",xlab="??",
     ylab="??",ylim=c(0,5))
grid(lwd=2)
```



Question 9

D'après le graphique ci-dessus, proposez un nombre pertinent de variables latentes et expliquez en quoi la nouvelle méthode apporte des améliorations par rapport à la méthode classique.

Réponse