Examen_R

Sous la direction de M. Henri LAUDE

Arnaud Bruel YANKO

24/01/2021

Contents

I. Travail 1 : Cross Validation . . . 2
1. Les critères d'évaluation . . . 2
2. Le lien vers le document . . . 2
3. L'auteur de ce document qui fait l'objet de notre étude . . . 2
4. Synthèse du document . . . 2
5. Extrait commenté des parties du Rmarkdown . . . 2
6. Evaluation du travail suivant les 5 critères précisés . . . 4
7. Conclusion . . . 4
II. Travail 2 : Les facteurs . . . 6
1. Critères d'évaluation . . . 6
2. Lien vers le document commenté . . . 6
3. Auteurs du document commenté . . . 6
4. Synthèse du document . . . 6
5. Extrait commenté des parties de code . . . 6
6. Evaluation du travail suivant les 5 critères précités . . . 7
7. Conclusion . . . 8
III. Travail 3 : Package rSymPy . . . 9
1. Les critères d'évaluation . . . 9
2. Le lien vers le document . . . 9
3. L'auteur de ce document qui fait l'objet de notre étude . . . 9
4. Synthèse du document . . . 9
5. Extrait commenté des parties du Rmarkdown . . . 9
6. Evaluation du travail suivant les 5 critères précisés . . . 10
7. Conclusion . . . 11
IV. Travail 4 :le package caret . . . 12
1. Les critères d'évaluation . . . 12
2. Le lien vers le document . . . 12
3. L'auteur de ce document qui fait l'objet de notre étude . . . 12
4. Synthèse du document . . . 12
5. Extrait commenté des parties du Rmarkdown . . . 12
6. Evaluation du travail suivant les 5 critères précisés . . . 13
7. Conclusion . . . 13
V. Travail 5 : Travail 5 : FactoMineR . . . 15
1. Critères d'évaluation . . . 15
2. Lien vers le document commenté . . . 15
3. L'auteur de ce document qui fait l'objet de notre étude . . . 15
Synthèse du document . . . 15
6. Evaluation du travail suivant les 5 critères précités . . . 18
7. Conclusion . . . 19
VI. Travail 6 : Travail personnelle 6 : Interpolation et implémentation des données spatiales avec R 20

Petite confusion 21
VII. Travail 4 de Maths : Travail 4 : Marche aléatoire 21
1. Les critères d'évaluation 21
2. Le lien vers le document 21
3. L'uteur de ce document qui fait l'objet de notre étude 21
4. Synthèse du travail 21
5. Extrait commenté des parties du Rmarkdown 21
6. Evaluation du travail suivant les 5 critères précités 22
7. Conclusion 22

I. Travail 1 : Cross Validation

1. Les critères d'évaluation

1. Allure du Rmd à l'exécution
2. Qualité de la rédaction du dossier
3. Accessibilité, didactisme et pertinence du dossier
4. Portabilité et lisibilité du Rmarkdown
5. Qualité des applications permettant d'illustrer le package, Qualité du code LaTeX, fiabilité des codes

2. Le lien vers le document

Le lien suivant **ici**,nous permet d'arriver au GITHUB du travail dont le titre est cité ci-dessus.

3. L'auteur de ce document qui fait l'objet de notre étude

Le travail suivant est fruit des recherches de Marko ARSIC et rindra LUTZ étudiants du MSc Data managementà la Paris School of Business en partenariat avec l'Efrei de Paris.

4. Synthèse du document

Il est acquis qu'un modèle doit être évalué sur une base de test différente de celle utilisée pour l'apprentissage. Mais la performance est peut-être juste l'effet d'une aubaine et d'un découpage particulièrement avantageux. Pour être sûr que le modèle est robuste, on recommence plusieurs fois. On appelle cela la validation croisée ou cross validation.

Le principe dans son ensemnle est le suivant, on découpe la base de données en cinq segments de façon aléatoire. On en utilise 4 pour l'apprentissage et 1 pour tester. On recommander 5 fois. Si le modèle est robuste, les cinq de scores de test seront sensiblement égaux.

Une fois qu'un modèle est défini grâce aux reunions statistiques, la suite c'est de pouvoir valider la fiabilité du modèle d'où la nécessité de la Cross-validation.

Par ailleurs ils présentent les 3 principales méthodes de Cross-validation entre autre:

- LOOCV
- LKOCV
- K-fold cross validation

5. Extrait commenté des parties du Rmarkdown

Les auteurs ont commencé par charger les packages `tydyverse` et `caret` également(package dont nous avons étudié dans le document de Xueting YIN). Rappellons que ces packages sont nécéssaires pour la construction des statistiques et plus particulièrement les modèles prédictifs.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## Warning: package 'ggplot2' was built under R version 3.6.3
## Warning: package 'tibble' was built under R version 3.6.3
## Warning: package 'tidyr' was built under R version 3.6.3
## Warning: package 'readr' was built under R version 3.6.3
## Warning: package 'purrr' was built under R version 3.6.3
## Warning: package 'dplyr' was built under R version 3.6.3
## Warning: package 'stringr' was built under R version 3.6.3
## Warning: package 'forcats' was built under R version 3.6.3
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3
## Loading required package: lattice
##
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
##
##     lift
```

```
data("swiss")
sample_n(swiss, 3)
```

```
##              Fertility Agriculture Examination Education Catholic
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Orbe              57.4        54.1          20         6     4.20
##              Infant.Mortality
## Delemont                 22.2
## Franches-Mnt             20.2
## Orbe                     15.3
```

Allons sur l'exemple pratique:

Si on a compris le travail sur `caret` alors le code qui n'est pas étranger.

Cette fonction ci-dessous de `caret` permettra de fixer les paramètres du processus d'apprentissage, c'est ce qu'ils font via cette commande.

```
set.seed(123)
train.control <- trainControl(method = "cv", number = 10)
```

Une fois cela défini il faudrait entrainer le modèle, et c'est ce qui est fait par le bias de la commande ci-dessous.

```
model <- train(Fertility ~., data = swiss, method = "lm",
               trControl = train.control)
```

Et enfin on affiche le résultat via la commande:

```
print(model)
```

```
## Linear Regression
##
## 47 samples
##  5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 42, 42, 42, 42, 42, 44, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   7.424916  0.6922072  6.31218
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

6. Evaluation du travail suivant les 5 critères précisés

1. Allure du Rmd à l'exécution

Le code Rmd s'execute sans problème, l'ensemble du document est visuellement confortable . Tout le *chunk* s'exécute au lancement, il y a pas d'erreur d'exécution.

2. Qualité de la rédaction du dossier

La rédaction de ce document est propre, lisible et limpide, les explications sont accompagnées de bons exemples très explicites. Le sujet est relativement facilement pour la comprenhension de tous.

On a pas forcement besoin de connaissance poussée en mathématique pour y parvenir.

3. Accessibilité, didactisme et pertinence du dossier

La lecture de ce dossier est reste facile et digeste, le document est assez fourni à mon sens en explications et celui ci est accompagné d'un exemple très compréhensif.

Le sujet est pertinent et a vocation à servir de guide pour de futures travaux dans ce sens.

4. Qualité et lisibilité du RMarkdown

Le RMarkdown est bien écrit, lisible.

En dehors de quelques problème de mise en forme tout est nickel pour moi.

5. Qualité des applications permettant d'illustrer le package, Qualité du code LaTeX, fiabilité des codes

Les applications sont de bonne qualité, maitrisées. mais pour ma part les auteurs pouvaient aller unpeu plus loin pour donner plus d'originalité à leur remarquable travail, par exemple le calcul des prédictions et indicateurs de performance avec découpage en échantillon de Training et de test et calcul des prédictions et indicateurs de performance avec cross-validation.

7. Conclusion

Selon moi, il s'agit globalement d'un bon travail.

Les auteurs ont fait un travail remarquable à mon avis, le document est assez explite et la compréhension en découle de sa lecture.

Ce tutoriel a pu mettre en avant les fonctionnalités de la cross validation.

Pour ceux voulant aller plus loin dans ce sujet, la documentation est assez fournie en ligne sur le sujet, on y retrouve une panoplie de documents, de tutoriel et d'articles scientifiques autour du sujet.

II. Travail 2 : Les facteurs

1. Critères d'évaluation

1. Comportement du Rmd à l'exécution
2. Qualité de la rédaction du dossier
3. Accessibilité, didactisme et pertinence du dossier
4. Qualité et lisibilité du Rmarkdown
5. Qualité des applications permettant d'illustrer le package

2. Lien vers le document commenté

Le lien suivant **ici**, nous permet d'arriver au GITHUB du travail dont le titre est cité ci-dessus.

3. Auteurs du document commenté

Le document évalué dans le cadre de ce rendu est celui de Claire MAZZUCATO et Thuy AUFRERE toutes 2 étudiantes au MSc Data management à la Paris school of business en partenariat avec l'Efrei de Paris.

4. Synthèse du document

Le travail a pu but d'expliquer ceci avec exemple à l'appui l'utilisation et l'importance des facteurs sur *R*.

Les facteurs sont des vecteurs un peu particuliers, facilitant la manipulation de données qualitatives (qu'elles soient numériques ou caractères). En effet, en plus de stocker les différents éléments comme un vecteur classique, il stocke également l'ensemble des différentes modalités possibles dans un attribut accessible via la commande 'levels

Ils forment une classe d'objets et bénéficient de traitements particuliers lors de leur manipulation et lors de l'utilisation de certaines fonctions. Les facteurs peuvent être non ordonnés (homme, femme) ou ordonnés (niveaux de ski). c'est donc les auteurs expliquent dans ce tutoriel

5. Extrait commenté des parties de code

En entrer de jeux ils existent 3 des fonctions permettant à tout utilisateur de creer des facteurs, c'est ce dont les auteurs nous présentent ici.

- La fonction `factor`
- La fonction `as.factor`
- La fonction `ordered`

Les fonctions factor et as.factor

Ces deux fonctions sont très similaires dans leur utilisation. La première permet de créer un facteur en définissant directement les différents éléments du facteur, l'autre permet de transformer un autre objet en facteur. Dans tous les cas, ces deux fonctions permettent généralement de créer des facteurs non ordonnés.

La fonction `factor` permet de créer un facteur en définissant directement les différents éléments du facteur.

```
sexe <- factor(c("H", "H", "F", "H", "H", "F", "F", "F"))
```

Affichge : H, H, F, H, H, F, F et F.

A l'aide de cette deuxième fonction, on définit tout d'abord le vecteur **salto**.

```
salto <- c(1:5,5:1)
```

Celui-ci aura pour éléments les suites de chiffres suivantes :

- de 1 à 5 : 1, 2, 3, 4, 5
- de 5 à 1 : 5, 4, 3, 2, 1

La sortie de ce chunk donnera donc : 1, 2, 3, 4, 5, 5, 4, 3, 2, 1.

Puis on définit le facteur **salto.f** à partir du vecteur **salto** précédemment défini, tel que :

```
salto.f <- as.factor(salto)
salto.f
```

Attention :il existe tout de même au moins une différence si tu utilises factor à la place de as.factor. Ce dernier conserve toujours tous les niveaux d'un facteur alors que l'utilisation de factor supprime les niveaux vides (modalités non représentées). Il faut donc faire attention quand on fait des allers et venues entre la classe factor et d'autre classe

3. La fonction `ordered`

La 3^{eme} fonction `ordered` est la fonction de creation facteurs ordonnés.

```
niveau <- ordered(c("débutant","débutant","champion",
                    "champion","moyen","moyen","moyen",
                    "champion"),
levels=c("débutant","moyen","champion"))
niveau
```

Définition d'un facteur **niveau** avec les éléments suivants : débutant, débutant, champion, champion, moyen, moyen, moyen, champion.

Et par la suite la fonction `levels` permet de donner la liste des valeurs possibles : débutant $<$ moyen $<$ champion.

```
levels(sexe) <- c("Femme","Homme")
```

Le code précédent permet de renommer "F" et "H" respectivement en "Femme" et "Homme".

Il est également possible de modifier la valeur d'un facteur facilement par indexation comme pour un vecteur avec la fonction `levels`.

Par défaut, les niveaux d'un facteur nouvellement créés sont classés par ordre alphanumérique croissant ou selon l'ordre qui figure dans l'option `levels`. Cet ordre est utilisé chaque fois que le facteur est employé.

```
sexe <- factor(c("H", "H", "F", "H", "H", "F", "F", "F"))
```

Affichage :

–> Levels : F H

Le F arrivant avant le H dans l'alphabet latin.

Il est possible de modifier l'ordre des niveaux de ce facteur en utilisant la manipulation suivante à l'aide de la fonction `levels`:

```
sexe <- factor(sexe, levels = c("H", "F"))
```

Sortie : l'affichage des niveaux ne sera plus "F H" mais "H F".

6. Evaluation du travail suivant les 5 critères précités

1. Comportement du Rmd à l'exécution

Le code Rmd s'execute sans problème, faut juste penser à avoir le logo de PSB dans le même dossier que notre document. Tout les *chunk* s'exécutent au lancement.

2. Qualité de la rédaction du dossier

La rédaction de ce document est propre, lisible et limpide, et digeste pour tout lecteur, les auteurs ont fait du bon travail dans son ensemble, les explications sont accompagnées d'exemples claires.

3. Accessibilité, didactisme et pertinence du dossier

La lecture de ce dossier est reste facile et digeste

Les explications sont propres et accompagnées des bon exemples.

Les auteurs captivent les lecteurs par la qualité de leur expressions et de leur vocabulaire assez claire à mon sens. Le document a vocation à transmettre une connaissance et le but, pour ma part, est atteint.

Le sujet est pertinent et a vocation à servir de guide pour de futures travaux dans ce sens.

4. Qualité et lisibilité du RMarkdown

Le RMarkdown est bien écrit, lisible et aéré.

On rescent que du travail a été fait, le sujet est assez facile pour ma part car touche des choses que nous manipulons régulièrement.

5. Qualité des applications permettant d'illustrer le package

Les applications sont de bonnes qualités, maitrisées. On rescent que la recherche a été faite et bien faite.

7. Conclusion

Selon moi, il s'agit globalement d'un bon travail.

Les auteurs fournissent un dossier recherché, documenté et facile à lire.

Le document est acccompagné des reférences pour les curieux.

III. Travail 3 : Package rSymPy

1. Les critères d'évaluation

1. Allure du Rmd à l'exécution
2. Qualité de la rédaction du dossier
3. Accessibilité, didactisme et pertinence du dossier
4. Portabilité et lisibilité du Rmarkdown
5. Qualité des applications permettant d'illustrer le package, Qualité du code LaTeX, fiabilité des codes

2. Le lien vers le document

Le lien suivant **ici**, nous permet d'arriver au GITHUB du travail dont le titre est cité ci-dessus.

3. L'auteur de ce document qui fait l'objet de notre étude

Le travail suivant est fruit des recherches de Xueting YIN étudiant du Master grande école en Data management à la Paris School of Business.

4. Synthèse du document

Dans ce document, l'auteur nous explique et ceci en s'appuyant sur des exemples bien précis le fonctionnement de `rSimpy`.

Par ailleurs ce dernier nous explique que c'est un package pour calculer les opérations symboliques.

Ce ci étant accompagbé par des exemples je dirais précis, l'auteur termine par nous expliquer la différence entre `rSympy` et `Sympy`, nous interpellant ainsi à ne pas faire de confusion arbusive.

5. Extrait commenté des parties du Rmarkdown

Import de package

```
library(rSymPy)
```

```
## Warning: package 'rSymPy' was built under R version 3.6.3
```

```
## Loading required package: rJython
```

```
## Warning: package 'rJython' was built under R version 3.6.3
```

```
## Loading required package: rJava
```

```
## Warning: package 'rJava' was built under R version 3.6.3
```

```
## Loading required package: rjson
```

En entreé l'auteur commence par des définitions des variables

```
Q <- Var("Q")
n <- Var("n")
m <- Var("m")
g <- Var("g")
k <- Var("k")
Sf <- Var("Sf")
y <- Var("y")
```

La définition de ces variables nous permettent de définir ou de construire cette équation:

```
sympy("expr = n*Q*(2*y*sqrt(1+m**2))**(2/3) - k*(m*y**2)**(5/3)*sqrt(Sf)")
```

```
## [1] "Q*n - k*m*Sf**(1/2)*y**2"
```

Une équation qui pour ma part est plus que classique de part sa définition et ma simplicité.

L'application de la fonction `Sympy` à cette équation précédente nous donne le code suivant:

```
sympy("solve(expr.subs([(Q, 1.2), (n, 0.045), (m, 3.4), (Sf, 0.2), (g, 9.806), (k, 1)]), y)")
```

```
## [1] "[-0.188451640531767, 0.188451640531767]"
```

```
out <- sympy("solve(expr.subs([(Q, 1.2), (n, 0.045), (m, 3.4), (Sf, 0.2), (g, 9.806), (k, 1)]), y)")
out <- as.numeric(unlist(strsplit(gsub("\\[|\\]", "", out), ",")))
length(out)
```

```
## [1] 2
```

```
out
```

```
## [1] -0.1884516  0.1884516
```

Ce qui est tout à fait comprehensible à la sortie, c'est classique, rien à ajouter.

pour appuyer l'auteur un autre exemple assez gentil peut-être celui ci-dessous:

```
a1 <- Var("a1")
a2 <- Var("a2")
a3 <- Var("a3")
a4 <- Var("a4")

A <- Matrix(List(a1, a2), List(a3, a4))
#ici on définit ici l'inverse et le determinant de la matrice
Inv <- function(x) Sym("(", x, ").inv()")
Det <- function(x) Sym("(", x, ").det()")

A
```

```
## [1] "[a1, a2]\n[a3, a4]"
```

```
cat(A,"\n")
```

```
## ( Matrix( ( [ a1,a2 ] ),( [ a3,a4 ] ) ) )
```

L'inverse est donnée par:

```
Inv(A)
```

```
## [1] "[1/a1 + a2*a3/(a1**2*(a4 - a2*a3/a1)), -a2/(a1*(a4 - a2*a3/a1))]\n[            -a3/(a1*(a4 - a2
```

et le déterminant est donnée par:

```
Det(A)
```

```
## [1] "a1*a4 - a2*a3"
```

On peut ainsi donner une multitude d'exemples, libre à vous de le faire en fonction de vos attentes et vous convenances

6. Evaluation du travail suivant les 5 critères précisés

1. Allure du Rmd à l'exécution

Le code Rmd s'execute sans problème lors de la compilation, aucune erreur n'est signalée. Tout le *chunk* s'exécute au lancement.

2. Qualité de la rédaction du dossier

La rédaction de ce document est propre, lisible et limpide, les explications sont accompagnées d'exemples bonnes et pas très explicitées. Le sujet est relativement facilement pour la comprenhension de tous.

On a pas forcement besoin de connaissance poussée en mathématique pour y parvenir.

3. Accessibilité, didactisme et pertinence du dossier

La lecture de ce dossier est reste facile et digeste

Les explications sont propres et accompagnées des bon exemples.

4. Qualité et lisibilité du RMarkdown

Le RMarkdown est bien écrit, lisible.

En dehors de quelques problème de mise en forme tout est nickel pour moi.

5. Qualité des applications permettant d'illustrer le package, Qualité du code LaTeX, fiabilité des codes

Les applications sont de bonnes qualités, maitrisées. Cependant, elles auraient pu être davantage expliquées.

7. Conclusion

Selon moi, il s'agit globalement d'un bon travail.

L'auteur a fait un bon travail à mon avis, l'une des difficultés à ce sujet pour l'auteur je pense que c'est le manque d'informations assez digestes en ligne car on retrouve peux de travail sur le sujet.

J'ai appris de nouvelles choses que je ne connaissais, pourtant mon niveau est intéressant sur R, merci au travail de l'auteur.

IV. Travail 4 :le package caret

1. Les critères d'évaluation

1. Allure du Rmd à l'exécution
2. Qualité de la rédaction du dossier
3. Accessibilité, didactisme et pertinence du dossier
4. Portabilité et lisibilité du Rmarkdown
5. Qualité des applications permettant d'illustrer le package, Qualité du code LaTeX, fiabilité des codes

2. Le lien vers le document

Le lien suivant **ici**,nous permet d'arriver au GITHUB du travail dont le titre est cité ci-dessus.

3. L'auteur de ce document qui fait l'objet de notre étude

Le travail suivant est fruit des recherches de Xueting YIN étudiant du Master grande école en Data management à la Paris School of Business.

4. Synthèse du document

Le package "caret" (Classification And REgression Training) est une librairie pour R. Il couvre une large fraction de la pratique de l'analyse prédictive (classement et régression).

Un peu à la manière de [**scikit-learn**] (http://scikit-learn.org/stable/) pour Python, il intègre dans un ensemble cohérent les étapes clés de la modélisation : préparation des données, sélection, apprentissage, évaluation. La standardisation des prototypes des fonctions d'apprentissage et de prédiction notamment permet de simplifier notre code, facilitant les tâches d'optimisation et de comparaison des modèles.

C'est ce dont l'auteur de ce document produit tout au long de ce travail.

5. Extrait commenté des parties du Rmarkdown

Pour un début l'auteur va commencer par l'installation d'un certains nombres de packages.

```
library(caret)
```

```
library(lattice)
library(ggplot2)
```

L'auteur nous montre comment créer d'un tableau séparant les données en gardant l'écart-type global : createDataPartition

la fonction suivante represente la reproductibilité.

```
set.seed(2020)
```

Le travail ici est effectué sur la data `iris`.

Même si "caret" propose des techniques de rééchantillonnage pour l'évaluation des modèles, l'auteur ici subdivise les données en échantillons d'apprentissage (60%) et de test (40%). Ceci est possible grâce à la commande `createDataPartition()` de la librairie `caret`.

```
Essaie_1 <- createDataPartition(iris$Species, p = .6,
                                list = FALSE,
                                times = 3)
head(Essaie_1)
```

```
##      Resample1 Resample2 Resample3
## [1,]         1         2         1
## [2,]         2         4         2
## [3,]         3         7         4
## [4,]         4         8         7
## [5,]         6         9         8
## [6,]         7        13         9
```

```
Essaie_2 <- createDataPartition(iris$Species, p = .6,
                                list = TRUE,
                                times = 2)
head(Essaie_2)
```

```
## $Resample1
##  [1]   3   4   6   7  10  11  12  14  15  17  18  19  20  25  26  28  30  31  32
## [20]  33  36  37  38  42  43  44  45  47  48  50  52  55  57  58  60  64  65  67
## [39]  69  71  72  73  74  76  77  78  79  80  81  82  84  86  89  90  92  93  96
## [58]  97  98  99 101 105 106 108 110 112 113 115 116 118 119 120 121 122 123 124
## [77] 125 128 129 130 132 134 135 137 138 141 144 147 148 150
##
## $Resample2
##  [1]   1   3   4   6   7   9  10  11  12  14  15  17  19  20  21  24  26  27  28
## [20]  29  30  33  36  37  38  42  43  44  45  47  52  53  56  57  58  60  62  65
## [39]  69  70  72  76  77  80  81  82  83  85  87  89  90  91  92  93  94  95  96
## [58]  98  99 100 102 103 104 106 107 108 109 110 111 113 115 116 118 119 122 123
## [77] 125 126 127 128 129 134 136 138 139 141 143 145 146 150
```

6. Evaluation du travail suivant les 5 critères précisés

1. Allure du Rmd à l'exécution

Le code Rmd s'execute sans problème, l'ensemble du document est visuellement confortable . Tout est *chunk* s'exécute au lancement.

2. Qualité de la rédaction du dossier

La rédaction de ce document est propre, lisible et limpide, les explications sont accompagnées d'exemples explicites. Le sujet est relativement facile pour la comprenhension de tous.

3. Accessibilité, didactisme et pertinence du dossier

La lecture de ce dossier est reste facile et digeste

Les explications sont propres et accompagnées des bon exemples.

Le sujet est pertinent et a vocation à servir de guide pour de futures travaux dans ce sens.

4. Qualité et lisibilité du RMarkdown

Le RMarkdown est bien écrit, lisible il est de très bonne qualité.

5. Qualité des applications permettant d'illustrer le package, Qualité du code LaTeX, fiabilité des codes

Les applications sont de bonne qualité, maitrisées. Cependant, elles auraient pu être davantage expliquées.

7. Conclusion

Selon moi, il s'agit globalement d'un bon travail.

L'auteur a produit un bon travail à mon avis dans son ensemble.

Grâce à ce travail l'auteur de ce tutoriel a pu mettre en avant quelques fonctionnalités intéressantes du package `caret`. Le principal mérite de ce package est d'encapsuler dans des fonctions et procédures clés en main des processus types de la pratique du data mining.

V. Travail 5 : Travail 5 : FactoMineR

1. Critères d'évaluation

1. Comportement du Rmd à l'exécution
2. Qualité de la rédaction du dossier
3. Accessibilité, didactisme et pertinence du dossier
4. Qualité et lisibilité du Rmarkdown
5. Qualité des applications permettant d'illustrer le package

2. Lien vers le document commenté

Le lien suivant **ici**,nous permet d'arriver au GITHUB du travail dont le titre est cité ci-dessus.

3. L'auteur de ce document qui fait l'objet de notre étude

Le travail suivant est fruit des recherches de Rindra LUTZ et William ROBACHE tous 2 étudiants au MSc Data management à la Paris school of business en partenariat avec l'Efrei de Paris.

Synthèse du document

Commençons par rappeler que Il existe un certain nombre de packages R pour calculer les méthodes de composantes principales. Ces packages comprennent: `FactoMineR`, `ade4`, `stats`, `ca`, `MASS` et `ExPosition`.

Les auteurs dans ce travail utilisent le package `FactoMineR` pour l'analyse de composantes principales.

Mathématique le principe est le suivant : On cherche une représentation des n individus , dans un sous-espace F_k de $\mathbf{R}^p$ de dimension k (k petit 2, 3, ...; par exemple un plan).

Autrement dit, on cherche à définir k nouvelles variables combinaisons linéaires des p variables initiales qui feront perdre le moins d'information possible.

Rappelons qu'il est d'abord necessaire d'installer le package `FactoMineR` avant toutes manipulations de code, et c'est ce qui est fait par les auteurs:

```
install.packages("FactoMineR")
```

Une fois le packages installés les auteurs besoin de jeux de données pour pouvoir faire des analyses, et les données utilisées ici sont ceux de `USArrests`. Le jeu de données contient des statistiques sur les arrestations par 100 000 habitants pour assaut, meurtre et viol dans chacun des 50 États américains en 1973.

l'ACP ne supporte pas les données manquantes. Or, par défaut, la fonction `PCA()` de FactoMineR les substitue par la moyenne de la variable, dans le code suivant il applique cette fonction.

```
library(FactoMineR)
```

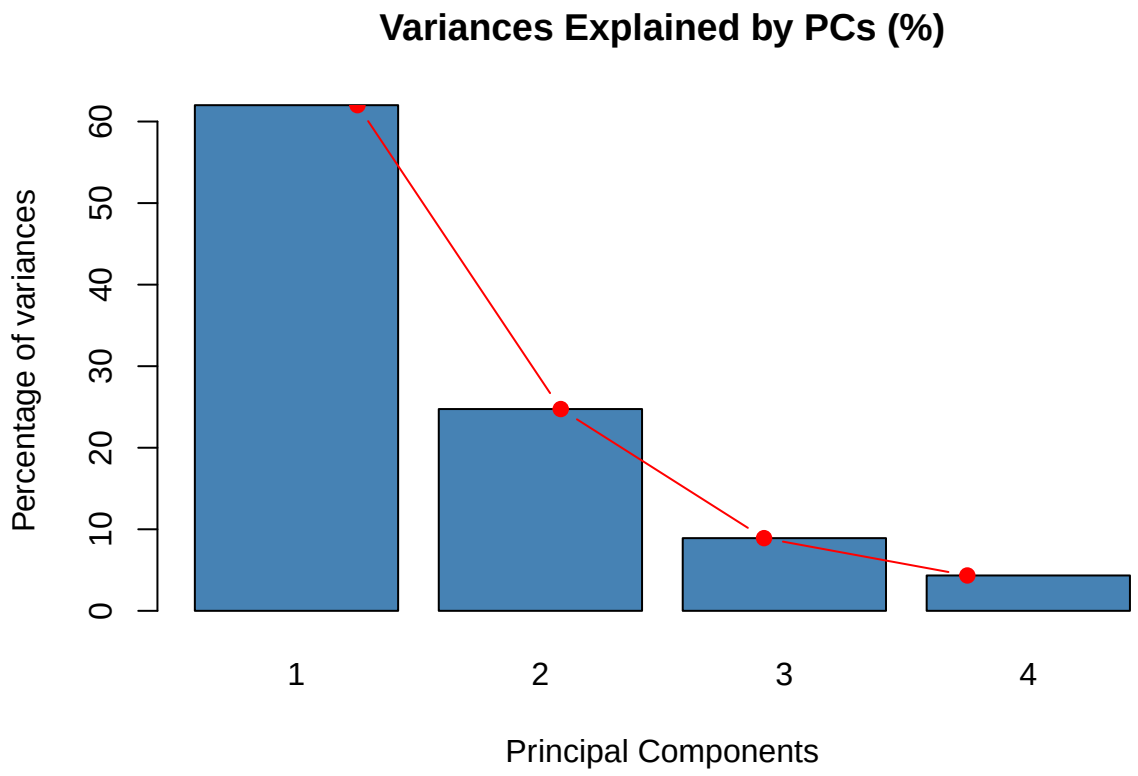
```
## Warning: package 'FactoMineR' was built under R version 3.6.3
```

```
data("USArrests")
res.pca <- PCA(USArrests, graph = FALSE)
```

Savoir comment sont constituées les données passe par exemple par la visualisation des valeurs propres.

```
eig.val <- res.pca$eig
barplot(eig.val[, 2],
        names.arg = 1:nrow(eig.val),
        main = "Variances Explained by PCs (%)",
        xlab = "Principal Components",
        ylab = "Percentage of variances",
        col ="steelblue")
```

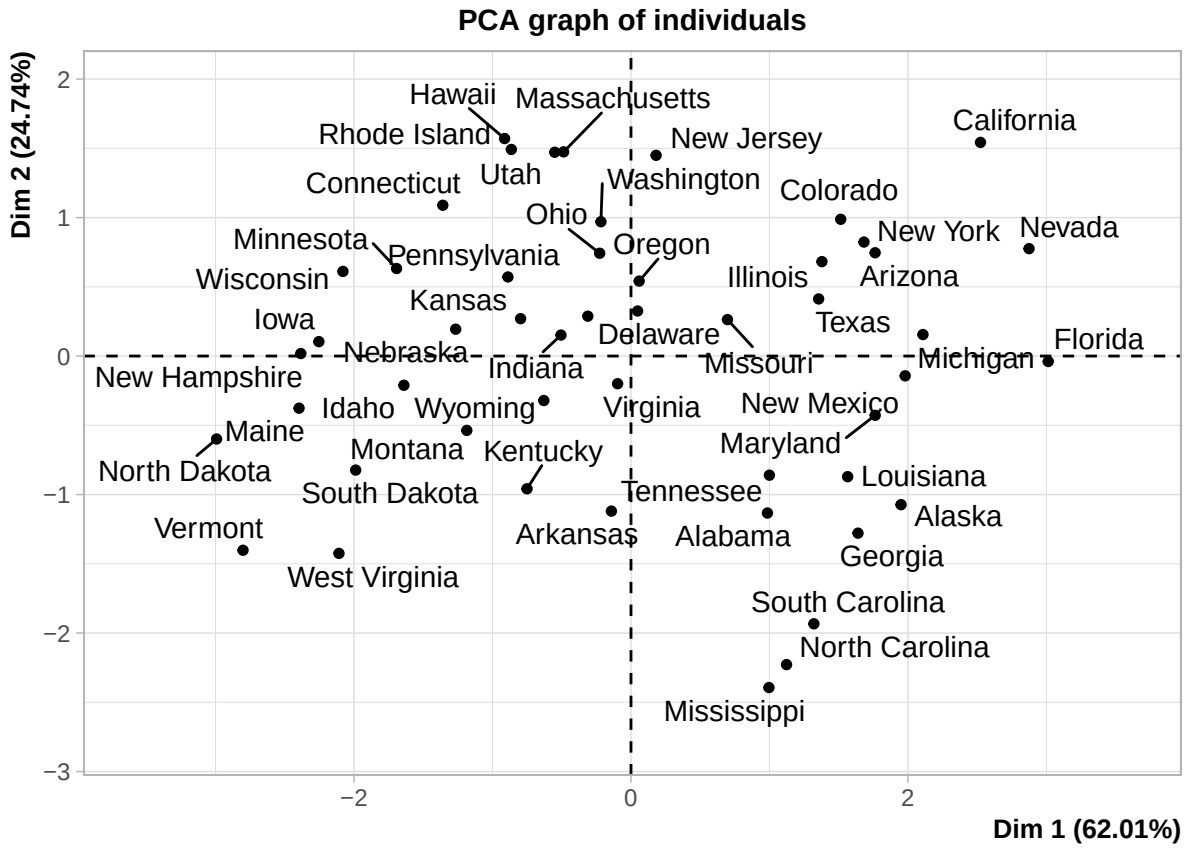
```
# Add connected line segments to the plot
lines(x = 1:nrow(eig.val), eig.val[, 2],
      type = "b", pch = 19, col = "red")
```



Et grâce à la commande suivante les auteurs nous permettent d'obtenir une repartition des individus.

```
plot(res.pca, choix = "ind", autoLab = "yes")
```

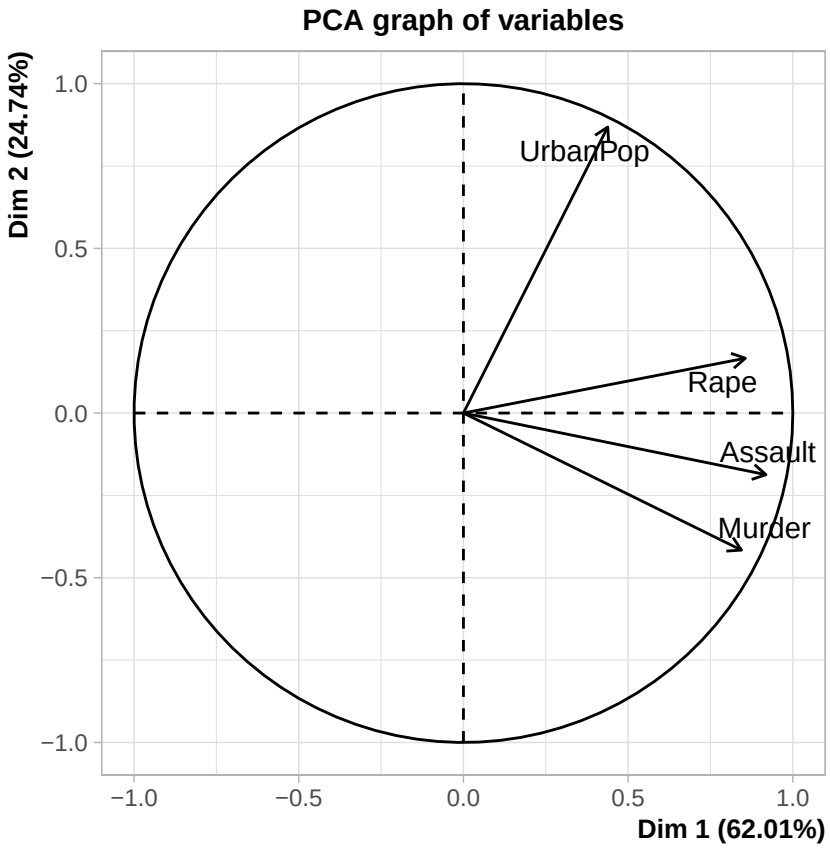
```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Par le suite ils nous permettent de visualiser la repartitionn des variables et tenant compte des valeurs manquantes qui ont été remplacées par la moyenne des autres variable via la fonction `PCA()`.

Ce qui est interessant c'est de voir que variables corrélées positivement sont du même côté du graphique, et les variables corrélées négativement sont sur des côtés opposés du graphique.

```
plot(res.pca, choix = "var", autoLab = "yes")
```



En conclusion, La contribution de certains points peuvent être très légèrement inférieures au seuil et conforter l'interprétation de l'axe que l'on aurait faite sans eux.

L'interprétation des nouvelles variables se fera à l'aide des individus et des variables contribuant le plus à l'axe avec la règle suivante : si une variable a une forte contribution positive à l'axe, les individus ayant une forte contribution positive à l'axe sont caractérisés par une valeur élevée de la variable.

De manière générale l'interprétation se fait dans la même manière une fois que les résulats ont été obtenus.

6. Evaluation du travail suivant les 5 critères précités

1. Comportement du Rmd à l'exécution

Le code Rmd s'execute sans problème lors de son éxécution, aucune erreur n'est affichée dans son ensemble.

2. Qualité de la rédaction du dossier

La rédaction de ce document est propre, lisible et limpide, et digeste pour tout lecteur, les auteurs ont fait du bon travail dans son ensembles , les explications sont accompagnées d'exemples claires.

3. Accessibilité, didactisme et pertinence du dossier

La lecture de ce dossier est reste facile et digeste

Les explications sont propres et accompagnées des bon exemples.

Le document a vocation à transmettre une connaissance et le but, pour ma part, est atteint.

Le sujet est pertinent et a vocation à servir de guide pour de futures travaux dans ce sens.

4. Qualité et lisibilité du RMarkdown

Le RMarkdown est bien écrit, lisible et aéré.

5. Qualité des applications permettant d'illustrer le package

Les applications sont de bonne qualité, maitrisées. L'exemple pris ici est asez bien expliqué par ces derniers.

Les details ressortent pleinement tout au long de la lecture.

7. Conclusion

Selon moi, il s'agit globalement d'un bon travail.

les auteurs fournissent un dossier recherché, documenté et facile à lire.

On y apprend des choses nouvelles. Personnellement, j'ai appris des choses que je connaissais certes, mais ma compréhension a été plus bonne.

Pour enrichir ce travail les auteurs pour nous parler des limites par exemple de l'ACP. Mais dans l'ensemle ce travail est très bien fait.

VI. Travail 6 : Travail personnelle 6 : Interpolation et implémentation des données spatiales avec R

Dans notre travail sur l'interpolation et implémentation des données spatiales avec R , il était question d'aborder l'aspect mathématique de l'interpolation des données spatiales et par ailleurs de faire un exemple d'implementation sur R de la dite interpolation avec le package `'DiceKriging'`.

Il en ressort que L'analyse spatiale est le processus de manipulation de l'information spatiale pour extraire de nouvelles informations à partir des données originales. Habituellement, l'analyse spatiale est réalisée avec un Système d'Information Géographique (SIG). Un SIG fournit généralement des outils d'analyses spatiales pour le calcul de statistiques sur les entités et la réalisation de géotraitements comme l'interpolation des données.

Dans ce travail nous avons définir ce qu'on entendait par données spatiales, nous avons présenté les différents modèles qui nous permettait de faire une pareille modélisation et nous avons terminé par faire des exemples d'applications concrèts.

Pour une auto critique pareil que celui de Mathématiques, je ne sais trop quoi dire, l'déal pour moi étant de laisser le choix aux autres de le faire.

Le travail demandé par l'enseignant a été fait dans son intégralité.

Je suis fier du travail et je laisse les autres jugés au mieux ce travail et les améliorations pourront etre apportées le cas échéant pourl' enrichir.

Petite confusion

J'ai inseré un travail de R dans celui de Maths par erreur, je me suis rendu compte après envoi, donc je le mets ici, merci pour votre bonne compréhension.

VII. Travail 4 de Maths : Travail 4 : Marche aléatoire

1. Les critères d'évaluation

1. Allure du Rmd à l'exécution
2. Qualité de la rédaction du dossier
3. Accessibilité, didactisme et pertinence du dossier
4. Portabilité et lisibilité du Rmarkdown
5. Qualité du code LaTeX, fiabilité des codes

2. Le lien vers le document

Le lien suivant **ici** nous permet d'arriver au GITHUB du travail dont le titre est cité ci-dessus.

3. L'uteur de ce document qui fait l'objet de notre étude

Le travail suivant est fruit des recherches de WILLIAM et MARKO tous étudiants du Master of science (MSc) à la Paris School of Business en partenariat avec l'Rfrei de Paris.

4. Synthèse du travail

La loi des grands nombres et le théorème central limite sont deux théorèmes clef de la théorie des probabilités. Ils montrent que la limite d'une somme de variables aléatoires indépendantes obéit à des lois simples qui permettent de prédire le comportement asymptotique.

Dans ce travail les auteurs travaillent autour de la notion de marche aléatoire et nous présente en quelques ligne comment le géant du numérique *google* utilise la marche aléatoire pour parcourir, identifier et classer les pages du réseau internet.

5. Extrait commenté des parties du Rmarkdown

Dans cette étude certains extraits sont interessants et méritent d'être commentés.

Commençons par une autre définition de cette notion qui est assez bien définit déja dans le document:

En effet une marche aléatoire est un modèle mathématique d'un système possédant une dynamique discrète composée d'une succession de pas aléatoires, ou effectués « au hasard ». On emploie également fréquemment les expressions marche au hasard, promenade aléatoire ou random walk.

L'exemple dans ce sens est comprensible normalement par tout le monde car il est assez claire je pense et surtout pour moi.

La suite de variables aléatoires $(X_n \in \mathbb{N})$ est appelée marche aléatoire sur le réseau $\mathbb{Z}^d$ si:

$X_n = X_0 + \sum_i Z_i$

et les déplacements successifs $(Z_n \in \mathbb{Z})$ sont indépendants et de même loi. Si chaque déplacement ne peut se faire que vers un de ses proches voisins, la marche aléatoire est dite simple. Si de plus les déplacements vers chacun des voisins immédiats se font avec la même probabilité $\frac{1}{2d}$, la marche est dite symétrique.

Partant de la définition précédente, on definit la marche aléatoire isotrope partant de 0, la suite (X_n de variables aléatoires suivante :

$$\begin{cases} X_0 = 0 & \text{n } = 0 \\ X_n = \sum_i X_i & n \geq 1. \end{cases}$$

Ensuite les auteurs parlent du *PageRank* qui est l'algorithme d'analyse des liens concourant au système de classement des pages Web utilisé par le moteur de recherche Google. Il mesure quantitativement la popularité d'une page web. Le *PageRank* n'est qu'un indicateur parmi d'autres dans l'algorithme qui permet de classer les pages du Web dans les résultats de recherche de Google.

Apres ce rappel ils definissent la notion de chaine de Markov qui est une notion très importante pour ceux qui ont eu à faire des statistiques et de la modélisation.

6. Evaluation du travail suivant les 5 critères précités

1. Allure du Rmd à l'exécution

Dans l'ensemble le code fonction bien à l'excution, tout est propre et ordonné.

2. Qualité de la rédaction du dossier

Le document est très bien redigé, on rescent que le travail a été produit et bien fait, explication sont claires et détaillées.

3. Accessibilité, didactisme et pertinence du dossier

Le documentest accesible, il est propre et pour ma part le travail fait est plus que pertinent et très utile pour toute personne debutante et en quête de connaissance sur ce sujet.

4. Portabilité et lisibilité du Rmarkdown

Le document reste accessible à tout le mmonde, vous pouvez le visualisez sur le Github de l'auteur. Le document est claire, lisible et propre.

5. Qualité du code LaTeX, fiabilité des codes

Le document est articulé, et sa mise en relation avec Rmarkdown est reussie, tous les codes /LATEX/ sont propres et très bien écris. Les codes /LATEX/ sont bien exécutés.

7. Conclusion

Pour conclure, je pense que les auteurs ont fait un bon travail de recherche tout au long de son travail, les expressions mathématiques présentent dans le document sont faciles à la comprendre, les définitions sont propres et les exemples qui les accompagnent le sont également à mon avis.

En guise d'application, à part le cas cités par les auteurs, les marches alétoires servent également en finance de marché pour le pricing et la valorisation des produits dérivés, la construction des martingales, par exemple.

Le sujet est très intessants et peut-être d'une grande aide pour les curieux.