Introduction à la statistique

TP_ACP 1A ENSAE

Senghak ROU

2024-07-19

Introduction

Nous effectuons une analyse en composantes principales (ACP) sur un ensemble de données regroupant les indicateurs socio-économiques de 167 pays, en saisissant des paramètres clés tels que la mortalité infantile, les dépenses de santé, le revenu et le PIB par habitant. En utilisant l'ACP, nous visons à distiller la complexité des données en composantes principales qui révèlent les dimensions primaires de la variation entre les pays, mettant ainsi en lumière les modèles sous-jacents et informant les politiques et les stratégies de développement économique. Grâce à un examen détaillé du processus et des résultats de l'ACP, nous identifions les pays ayant des positions socio-économiques distinctes et discernons les indicateurs les plus critiques pour façonner le paysage économique mondial.

question 1

Retirer la première colonne de Data et renommer les lignes du jeu de données obtenu avec les noms des pays

```
df <- read.csv("Country-data.csv")

rownames(df) <- df$country
df <- df[, -1]

n <- nrow(df)
d <- ncol(df)
```

questions 2 & 3

La centralisation des données permet de supprimer les effets liés à l'échelle et aux unités de mesure. En réalité, si cette précaution n'est pas prise, les variables exprimées dans des unités plus grandes auraient un impact disproportionné sur les résultats par rapport à celles exprimées dans des unités plus petites. De même, la normalisation permet de réduire le risque que les variables avec des valeurs plus élevées dominent injustement l'analyse au détriment de celles ayant des valeurs moins élevées. Cela rend la comparaison des composantes principales plus facile, car elles sont des combinaisons linéaires des variables d'origine.

```
mean_col <- colMeans(df)
sd_col <- apply(df, MARGIN = 2, FUN = sd)

df_standard <- df %>%
                sweep(MARGIN = 2, STATS = mean_col, FUN = "-") %>%
                sweep(MARGIN = 2, STATS = sd_col, FUN = "/")
colMeans(df_standard)
```

```
##    child_mort       exports        health       imports        income
##  1.555642e-16 -3.234272e-16 -1.400078e-15  3.011563e-16 -7.445807e-17
##     inflation    life_expec     total_fer          gdpp
##  1.329608e-17  3.616535e-16  1.728491e-17  2.393295e-17
```

question 4

On calcule la matrice de variance-covariance $\hat{\Sigma} = \frac{1}{n-1} X^T X$. Cette matrice correspond à la matrice des corrélations car nos données ont été réduites.

```
hatSigma <- var(df_standard)
```

question 5

La diagonalisation consiste à déterminer les valeurs propres et vecteurs propres de la matrice hatsigma. Les valeurs propres sont toutes positives (logique car la matrice $\hat{\Sigma}$ est définie-positive).

```
spectre <- eigen(hatSigma)
print(spectre$values)
```

```
## [1] 4.13565658 1.54634631 1.17038330 0.99478456 0.66061903 0.22358112 0.11343874
## [8] 0.08831536 0.06687501
```

question 6

```
n <- nrow(df)
inertie <- (1/(n-1)) * sum(df_standard**2)
print(inertie)
```

```
## [1] 9
```

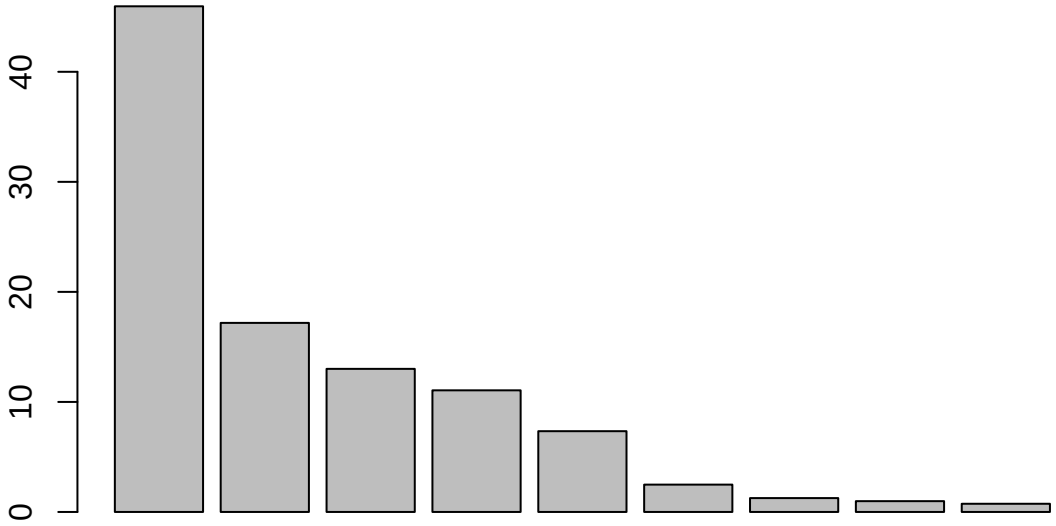
```
svp <- sum(spectre$values)
print(svp)
```

```
## [1] 9
```

Le résultat montre que l'inertie des variables et la somme des valeurs propres de hatSigma sont égales à 9. Cela confirme que l'ACP a été réalisée correctement et que les valeurs propres représentent bien la distribution de la variance totale entre les composantes principales.

question 7

```
pourcentage <- 100 * spectre$values / sum(spectre$values)
barplot(pourcentage)
```



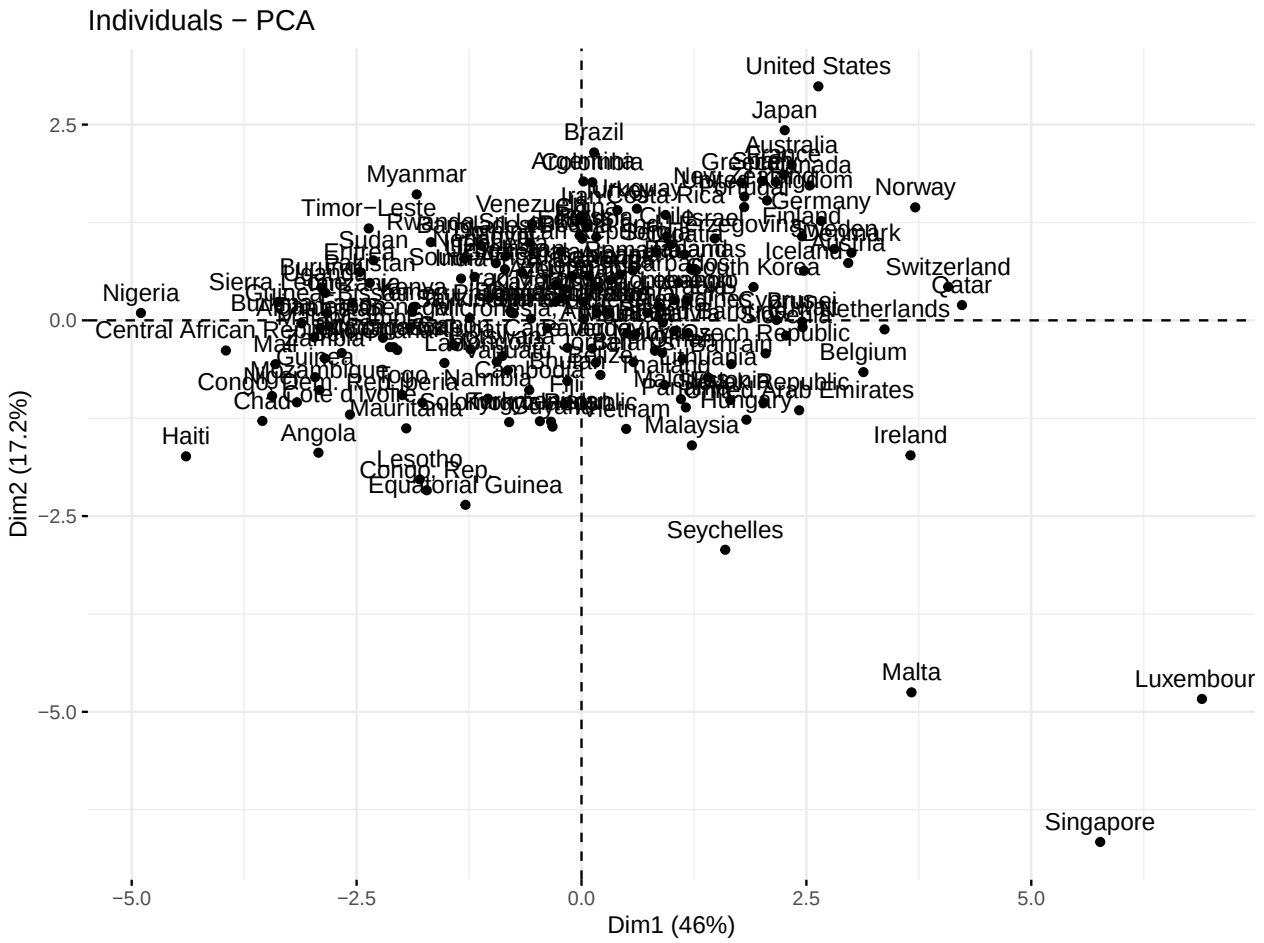
Le graphique demandé montre que la première composante principale (axe) explique 45,95 % de la variance totale, ce qui indique sa contribution significative à la capture de la variabilité de l'ensemble de données. La deuxième composante principale représente 17,18 % supplémentaires de la variance. Les contributions des

composantes suivantes, y compris la troisième, sont progressivement plus faibles, chacune expliquant moins de 17,18 % de la variance totale. Ce schéma suggère que la majorité de la variabilité de l'ensemble de données peut être attribuée aux quelques premières composantes, avec des rendements décroissants si l'on inclut des composantes supplémentaires.

question 8

Représentation des individus sur les deux premiers axes.

```
res.pca <- prcomp(df_standard, scale = TRUE)
fviz_pca_ind(res.pca)
```



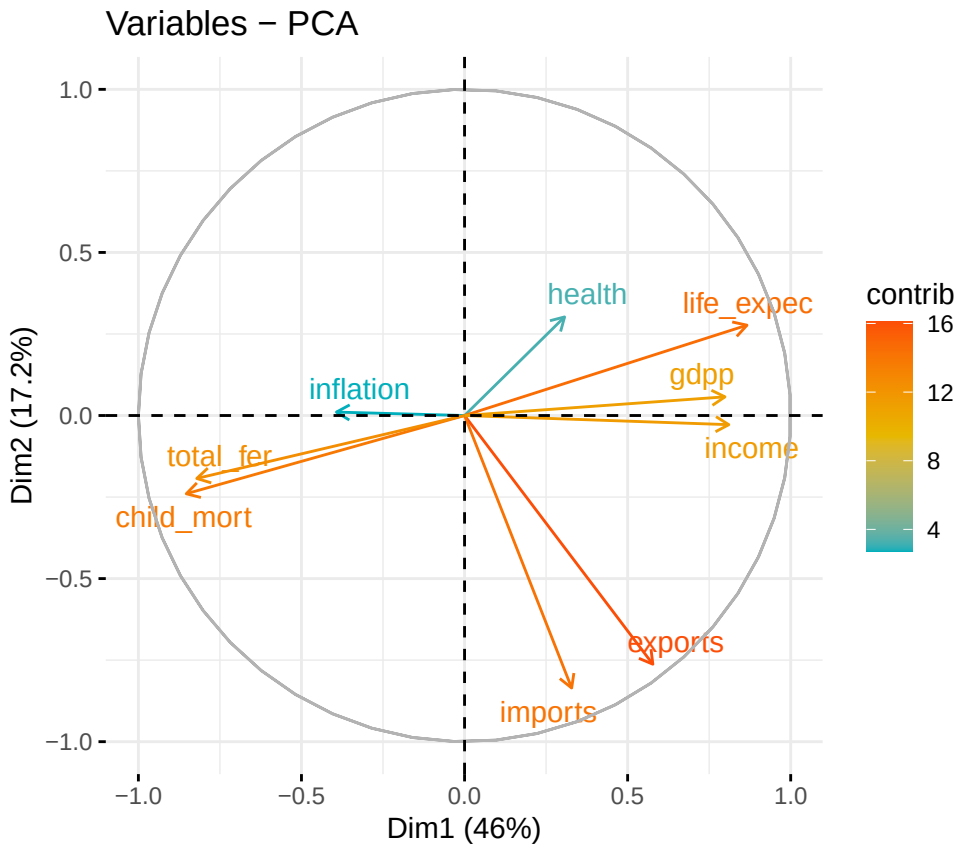
Le Luxembourg est positionné loin dans la direction positive de la première composante principale (Dim1), qui explique 46 % de la variance, ce qui le rend très représentatif des caractéristiques capturées par cette composante. Cela suggère que les indicateurs socio-économiques du Luxembourg diffèrent significativement de la moyenne mondiale. D'autre part, Singapour se situe à l'extrémité positive de la deuxième composante principale (Dim2), représentant 17,2 % de la variance. Cela indique que les caractéristiques socio-économiques uniques de Singapour sont distinctement capturées par la deuxième composante principale.

question 9

Représentation les variables sur le cercle des corrélations.

```
fviz_pca_var(res.pca,
             axes = c(1, 2),
             col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
```

```
             repel = TRUE)
```



Dans ce cercle de corrélations, les variables situées le plus loin sur l'axe horizontal (Dim1, qui explique 46 % de la variance) ont une corrélation plus forte avec la première composante principale. D'après le graphique, il semble que les exports, imports, income, et gdpp soient les variables qui se situent le plus loin sur l'axe horizontal, et donc celles qui expliquent le mieux la première composante principale.

question 10

Nous avons identifié le Luxembourg et Singapour comme des pays très représentatifs des première et deuxième composantes principales, respectivement, et en utilisant la représentation du cercle de corrélation, nous pouvons interpréter leurs positions par rapport aux variables socio-économiques: le Luxembourg s'est positionné loin dans la direction positive de la première composante principale (PC1), que nous avons identifiée comme étant associée à la prospérité économique (revenu élevé, PIB par habitant élevé). Cela suggère que le Luxembourg a un niveau élevé de performance économique par rapport aux autres pays de l'ensemble des données. Étant donné qu'il s'agit d'un centre financier avec un niveau de vie relativement élevé, cette position est conforme aux attentes.

Singapour se positionne loin dans la direction positive de la deuxième composante principale (PC2), qui semble être liée aux aspects du commerce (importations et exportations). Singapour est connu pour sa position stratégique et son importance en tant que plaque tournante mondiale du commerce et du transport, de sorte que sa position distincte sur la PC2 souligne son rôle dans le commerce international et ses solides activités liées au commerce.

Les deux pays sont "isolés" en ce sens qu'ils occupent des positions uniques sur le plan factoriel, ce qui indique que leurs profils socio-économiques sont sensiblement différents de ceux de la plupart des autres pays de l'ensemble des données. La richesse économique distincte du Luxembourg et les fortes activités commerciales de Singapour les distinguent dans l'espace multidimensionnel défini par l'ACP.

Conclusion

En conclusion, notre ACP a révélé des tendances socio-économiques distinctes d'un pays à l'autre, avec des exemples notables comme le Luxembourg et Singapour. Ces résultats soulignent l'utilité de l'ACP pour éclairer les politiques fondées sur des données et faire progresser l'analyse économique.