Analyse en composante principales

Choix du nombre de composantes

Véronique Tremblay

- Est-ce qu'il faut standardiser?
- Quel est l'effet des valeurs extrêmes?

En pratique (avec R)

En pratique, on utilisera la fonction `PCA` de la librairie `FactoMineR` (http://factominer.free.fr/index_fr.html).

```
library(FactoMineR)
# Faire l'ACP
climat_pca <- PCA(climat[6:13],
                  ncp = 8, #Pour conserver toutes
                           #les composantes principales
                  graph = FALSE)

# Extraire les composantes principales
cp <- climat_pca$ind$coord
```

Interprétation des poids

```
# Extraire 6 premiers les vecteurs propres
poids <- round(climat_pca$svd$V[,1:6],2)
```

tmax_mars	0.41	-0.33	-0.12	-0.29	0.09	0.02
tmax_juin	0.35	-0.48	-0.09	0.06	0.40	0.12
tmax_sept	0.46	-0.23	0.09	-0.12	-0.01	-0.01
tmax_dec	0.39	0.20	-0.09	-0.40	-0.69	-0.10
precip_mars	0.27	0.50	0.30	-0.21	0.50	-0.53
precip_juin	0.32	0.02	0.73	0.48	-0.21	0.23
precip_sept	0.33	0.13	-0.51	0.68	-0.10	-0.37
precip_dec	0.26	0.54	-0.27	-0.02	0.23	0.71

Interprétation des poids

Quelles villes auront une valeur élevée sur la première composante?

Lien vers le tableau interactif

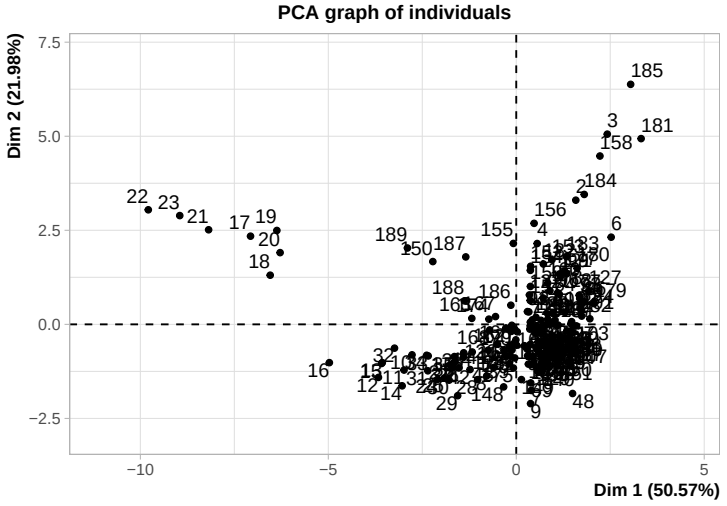
Analyse des observations

Visualisation des observations

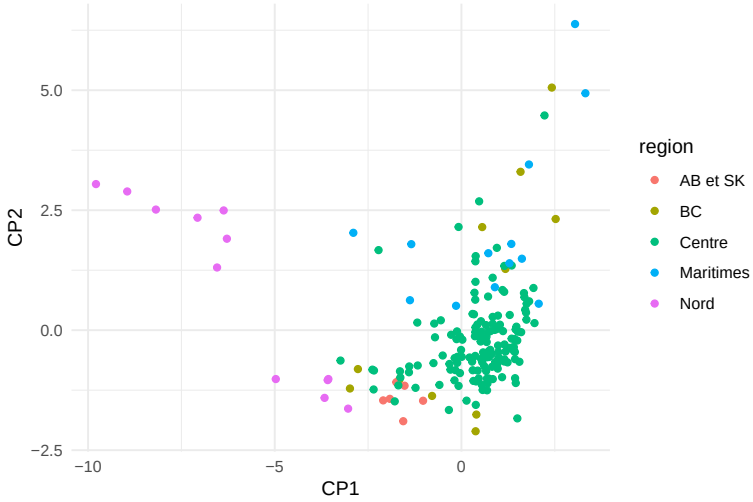
D'un point de vue géométrique, l'ACP projette les observations dans un sous-espace de dimensions inférieur.

Les composantes principales calculées précédemment sont simplement les coordonnées des observations sur les nouveaux axes (axes factoriels).

Visualisation des observations



Visualisation des observations



Lien vers la figure interactive

Qualité de la représentation d'une observation sur chaque composante

Indique à quel point l'individu est bien représenté par cet axe.

$$Q_{i,k}^{(obs)} = \frac{Y_{i,k}^2}{d_i^2}$$

$$d_i = \sqrt{\sum_1^k Y_{i,k}^2}$$

Contribution de chaque observation aux composantes

Proportion de la variabilité de la composante k provenant d'un individu donné i.

$$C_{i,k}^{(obs)} = \frac{Y_{i,k}^2/n}{\lambda_k}$$

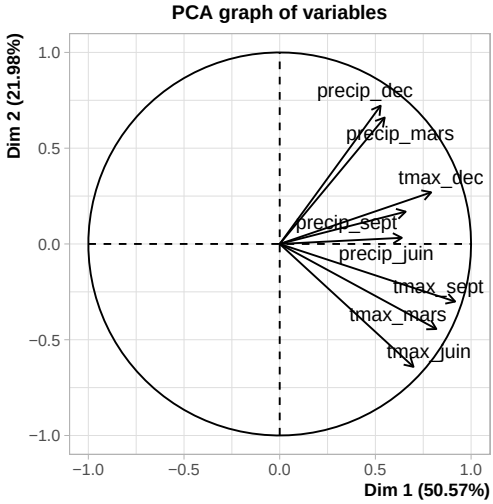
Analyse des variables

Visualisation des variables

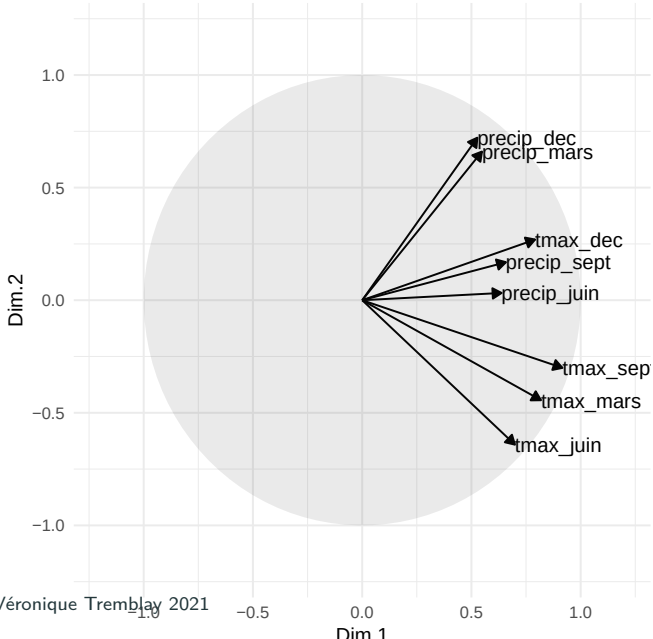
La coordonnée de la j^e variable sur l'axe k correspond à la corrélation entre cette variable et la k^e composante principale.

$$\mathbf{R} = cor(X, Y)$$

Visualisation des variables



Visualisation des variables



Qualité de la représentation des variables sur les composantes

Proportion de la variabilité d'une variable j expliquée par la composante k.

$$Q_{j,k}^{(var)} = r_{j,k}^2$$

Contribution des variables à chaque composante

$$C_{j,k}^{(var)} = \frac{r_{j,k}^2}{\lambda_k}$$

Ajout de variables et d'individus

Ajout de variables quantitatives

```
climat_pca <- PCA(climat[3:13],
                  quanti.sup = c(1,2),
                  quali.sup = c(3),
                  ncp = 8,
                  graph = FALSE)
```

Illustration des variables quantitatives supplémentaires

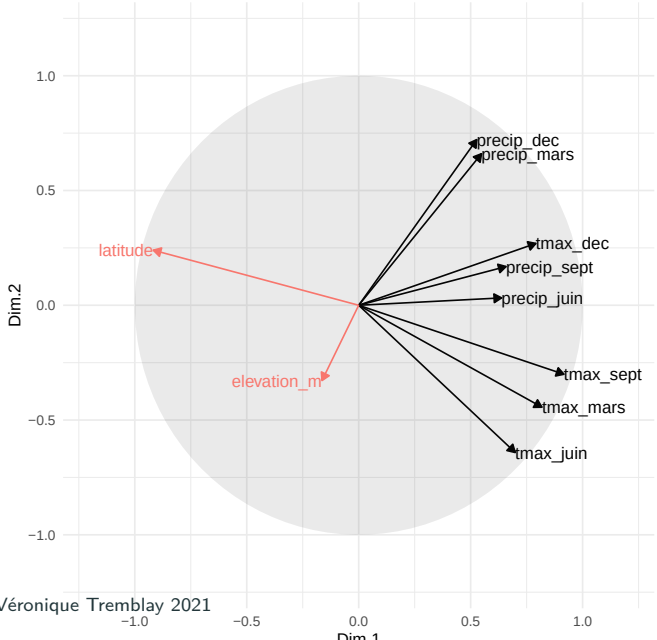


Illustration des variables qualitative supplémentaires

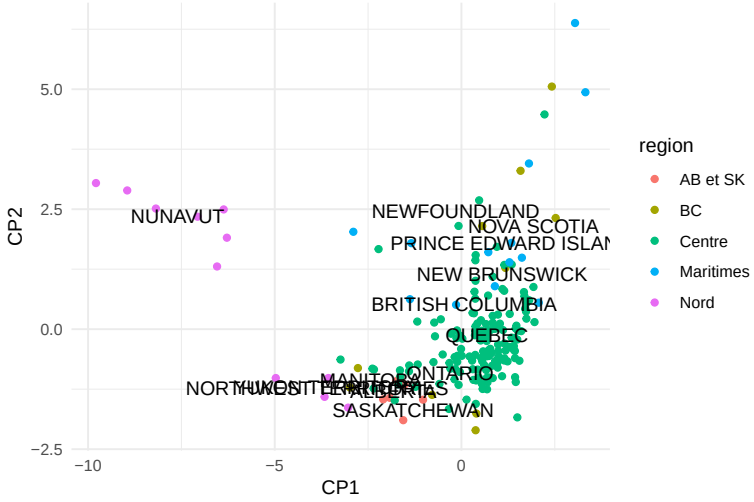


Figure interactive

Qualité des variables supplémentaires

	latitude	elevation_m
Dim.1	0.84	0.03
Dim.2	0.06	0.11
Dim.3	0.00	0.00
Dim.4	0.00	0.00
Dim.5	0.02	0.02
Dim.6	0.00	0.00
Dim.7	0.00	0.31
Dim.8	0.00	0.00

Choix du nombre de composantes

La règle des 80%

On garde les composantes qui représentent 80% de la variance.

	Variance	% Variance	% Cumulé
comp 1	4.05	50.57	50.57
comp 2	1.76	21.98	72.55
comp 3	0.77	9.57	82.12
comp 4	0.65	8.15	90.27
comp 5	0.38	4.71	94.98
comp 6	0.24	3.05	98.03
comp 7	0.11	1.33	99.36
comp 8	0.05	0.64	100.00

La règle de Kaiser

On garde les composantes qui correspondent aux $\lambda > 1$.

	Variance	% Variance	% Cumulé
comp 1	4.05	50.57	50.57
comp 2	1.76	21.98	72.55
comp 3	0.77	9.57	82.12
comp 4	0.65	8.15	90.27
comp 5	0.38	4.71	94.98
comp 6	0.24	3.05	98.03
comp 7	0.11	1.33	99.36
comp 8	0.05	0.64	100.00

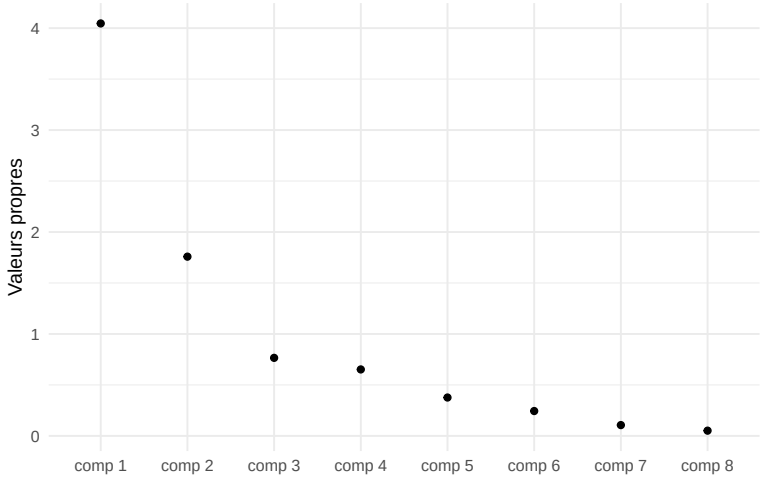
La règle de Joliffe

On garde les composantes qui correspondent aux $\lambda > 0.70$.

	Variance	% Variance	% Cumulé
comp 1	4.05	50.57	50.57
comp 2	1.76	21.98	72.55
comp 3	0.77	9.57	82.12
comp 4	0.65	8.15	90.27
comp 5	0.38	4.71	94.98
comp 6	0.24	3.05	98.03
comp 7	0.11	1.33	99.36
comp 8	0.05	0.64	100.00

La règle de Cattell

On conserve les composantes qui précèdent le pied de l'éboulis.



Dans un contexte prédictif

Les composantes de variance maximale ne sont pas nécessairement les plus importantes pour la prédiction.

Extensions de l'ACP et thèmes connexes

ACP avec rotation

- Orthogonale, obliques, varimax
  - Abdi, H. and Williams, L.J. (2010), Principal component analysis. WIREs Comp Stat, 2: 433-459. doi:10.1002/wics.101
  - ESL section 14.7
- Analyse factorielle (STT-7620-Modèles d'équations structurelles)

ACP avec noyau

Pour comprendre le principe de l'astuce du noyau:

Astuce du noyau

Pour plus de détails, voir ELS, section 14.5.4

Autres extensions de l'ACP

- Courbes et surfaces principales (ESL 14.5.2)
- ACP parcimonieuse (ESL 14.5.5)
- ACP probabiliste, probabiliste parcimonieuse, probabiliste parcimonieuse bayésienne,...
- Analyse des correspondances (prochain module)

Thèmes liés

- Autoencodeurs
- Uniform Manifold Approximation and Projection (UMAP)