

Réduction de la dimension

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

On dispose de p variables explicatives X_1, \dots, X_p .

Comment réduire ce nombre de variables en conservant le plus d'information possible?

- maximiser la variabilité
- créer de nouvelles variables non corrélées les unes avec les autres.

Utilisé principalement pour les questionnaires

- Trouver automatiquement des regroupements de variables
- Créer des variables résumées (moyenne de variables)

Coefficient de corrélation linéaire

- Mesure la relation *linéaire* entre variables
- Valeur entre $-1 \leq r \leq 1$.
- Les points sont alignés (exactement) si $r = \pm 1$; le signe détermine l'orientation de la pente.

Datasaurus

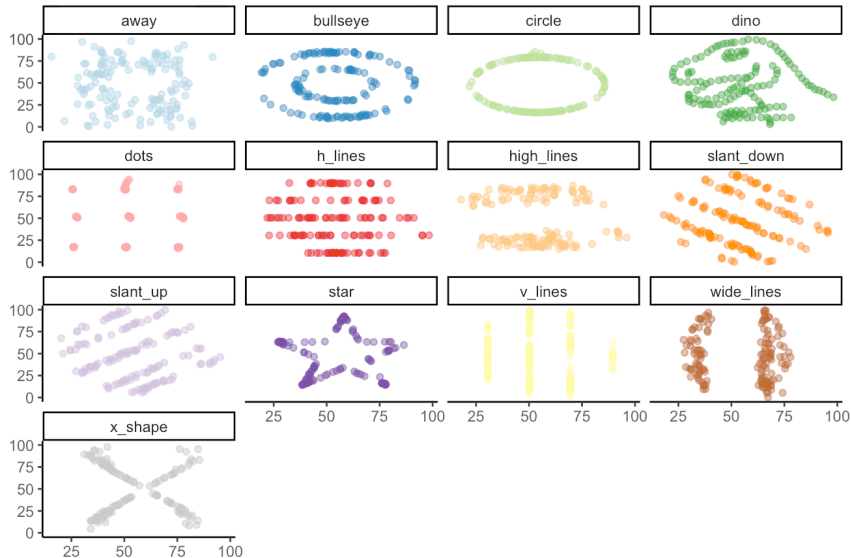


Figure 1: Datasaurus (13 jeux de données avec la même corrélation linéaire)

- **Questions** sur une étude dans un magasin (200 répondants).
- Réponses: échelles de Likert allant de pas important (1) à très important (5)

Pour vous, à quel point est-ce important

- | | |
|---|--|
| 1. que le magasin offre de bons prix tous les jours? | 7. que le magasin ait sa propre carte de crédit? |
| 2. que le magasin accepte les cartes de crédit majeures (Visa, Mastercard)? | 8. que le service soit rapide? |
| 3. que le magasin offre des produits de qualité? | 9. qu'il y ait une vaste sélection de produits? |
| 4. que les vendeurs connaissent bien les produits? | 10. que le magasin accepte le paiement par carte de débit? |
| 5. qu'il y ait des ventes spéciales régulièrement? | 11. que le personnel soit courtois? |
| 6. que les marques connues soient disponibles? | 12. que le magasin ait en stock les produits annoncés? |

Typiquement utilisé à des fins

- d'analyse exploratoire (visualisation) ou
- pour créer une variable réponse (par ex., le quotient intellectuel)

Réduire la dimension en préservant le plus de variabilité possible.

Comment manger le plus de krill?

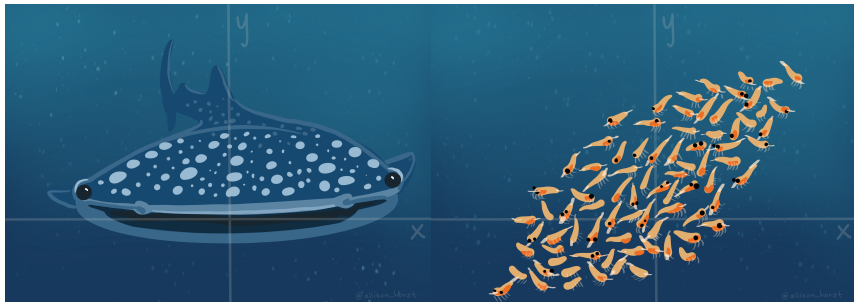


Figure 2: Illustrations d'Allison Horst (CC BY 4.0)

Composantes principales:

- Créer des nouvelles variables **non corrélées** appelées **composantes principales** et dénotées C_1, \dots, C_p .
- Les composantes principales sont des combinaisons linéaires des variables originales.

$$C_j = w_{j1}X_1 + w_{j2}X_2 + \dots + w_{jp}X_p, \quad (j = 1, \dots, p),$$

somme de poids fois variables explicatives

$$1 = w_{j1}^2 + \dots + w_{jp}^2$$

poids standardisés

$$\text{Va}(C_1) \geq \dots \geq \text{Va}(C_p)$$

et $\text{Cor}(C_i, C_j) = 0$ pour $i \neq j$.

- L'ensemble de k variables qui maximise la variance totale exprimée est C_1, \dots, C_k .
- Par construction, la variance des composantes principales est décroissante.

Table 1: Variance des premières composantes principales

C1	C2	C3	C4	C5	C6	C7	C8
2.43	2.00	1.94	1.30	0.74	0.69	0.57	0.54

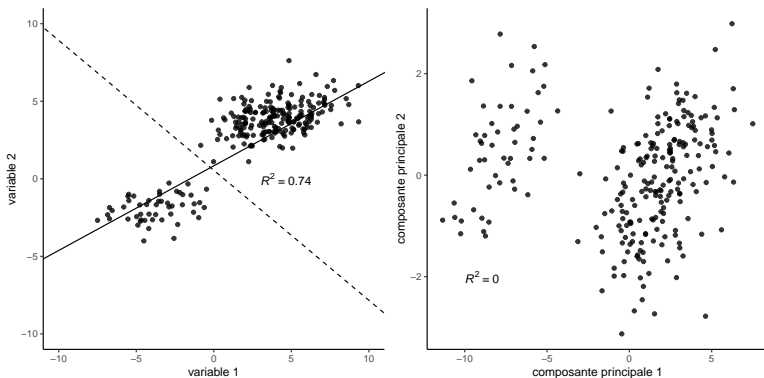


Figure 3: Nuage de points avant (gauche) et après (droite) analyse en composantes principales.

```
# Analyse en composantes principales  
# de la matrice de corrélation  
acp <- princomp(factor, cor = TRUE)  
loadings(acp) # chargements  
biplot(acp) # bigramme
```

Matrice de chargements

	C1	C2	C3	C4	C5	C6	C7	C8
x1	0.21			0.65	0.27	0.41	0.45	
x2						0.24		0.41
x3		0.36						
x4			0.25					0.25
x5				0.69				
x6		0.34			0.74			
x7								
x8			0.21					
x9		0.42					0.42	0.51
x10							0.44	
x11			0.35					
x12		0.36				0.63		

Les poids inférieurs à 0.2 en valeur absolue sont omis.

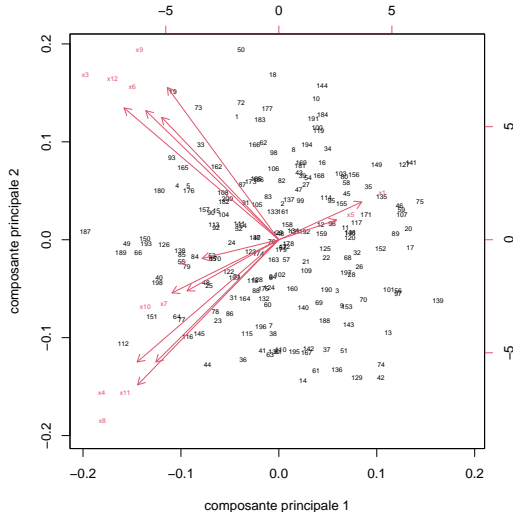


Figure 4: Bigramme (représentation sur les deux premières composantes principales)

- **critère des valeurs propres de Kaiser:** variances des composantes principales (valeurs propres) supérieures à 1.
- **critère du coude de Cattell:** diagramme d'éboulis (screeplot)

```
hecmulti::eboulis(acp)
```


Diagramme d'éboulis

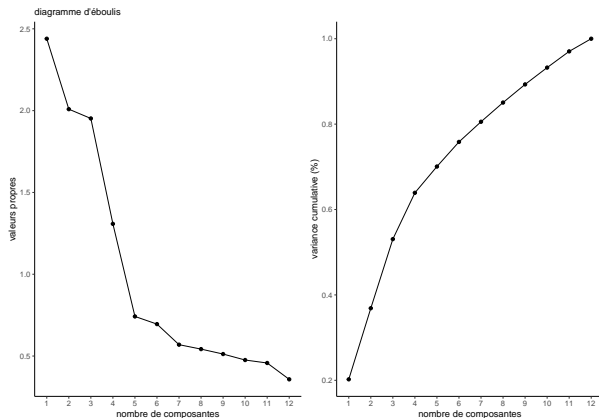


Figure 5: Diagramme d'éboulis (gauche) et variance cumulative (droite).

La variance totale des composantes principales est identique à celle des variables originales:

$$\text{Va}(C_1) + \dots + \text{Va}(C_p) = \text{Va}(X_1) + \dots + \text{Va}(X_p)$$

- Si on travaille avec la matrice de corrélation (variance unitaire), alors la variance totale est p .

La matrice de corrélation est la matrice de covariance des données **standardisées** (variance unitaire)

- la covariance accorde plus d'importance aux variables qui ont une variance élevée
- si les variables sont sur la même échelle, covariance et corrélation sont utilisables
- sinon, utiliser la matrice de corrélation *par défaut*.

- Toutes les variables sont nécessaires pour créer des composantes principales avec de nouvelles observations
- Le critère d'optimisation ne prend pas en compte une potentielle variable réponse (voir analyse canonique et moindres carrés partiels).

- Y a-t-il des groupements de variables?
- Est-ce que les variables faisant partie d'un groupement semblent mesurer certains aspects d'un facteur commun (non observé)?

De tels groupements peuvent être détectés (automatiquement) si plusieurs variables sont très corrélées entre elles.

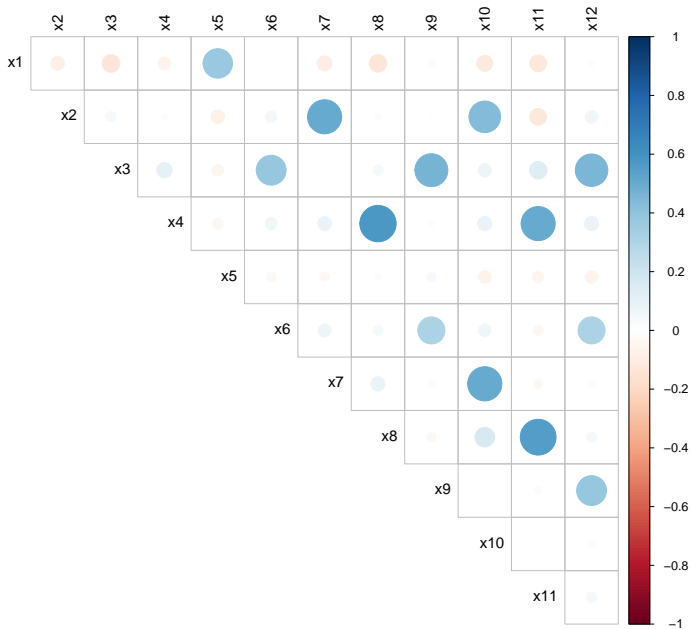


Figure 6: Corrélogramme des items du questionnaire

On possède n observations sur p variables et on s'intéresse à la matrice de covariance (corrélation) qui décrit la relation linéaire.

Avec np données, on cherche à estimer p paramètres de variances et $p(p - 1)/2$ corrélations.

On cherche un modèle plus **parcimonieux** pour expliquer la dépendances.

On suppose qu'il existe $m < p$ **facteurs** latents F_1, \dots, F_m qui suffisent à expliquer les variables explicatives

Les facteurs sont:

- des variables aléatoires non observables
- non corrélées entre elles
- standardisées de moyenne zéro et variance unitaire.

Soit X_1, \dots, X_p des variables explicatives *standardisées* (moyenne nulle, variance unitaire).

Cela revient à travailler avec la matrice de corrélation

$$X_j = \underbrace{\gamma_{j1}F_1 + \dots + \gamma_{jm}F_m}_{\text{combinaison linéaire pondérée des facteurs}} + \underbrace{\varepsilon_j}_{\text{aléa}}, \quad j = 1, \dots, p.$$

où ε_j est un aléa de variance ψ_j et de moyenne nulle.

$$X_j = \underbrace{\gamma_{j1}F_1 + \cdots + \gamma_{jm}F_m}_{\text{combinaison linéaire pondérée des facteurs}} + \underbrace{\varepsilon_j}_{\text{aléa}}, \quad j = 1, \dots, p.$$

Le chargement γ_{ij} mesure la corrélation entre X_i et F_j ,

$$\gamma_{ij} = \text{Cor}(X_i, F_j).$$

La proportion de la variance de X_i expliquée par les facteurs est $\gamma_{i1}^2 + \cdots + \gamma_{im}^2$.

Exemple d'estimation des chargements

Table 2: Estimés des chargements (pourcentage).

	F1	F2	F3	F4
x1				81
x2			-79	
x3		79		
x4	82			
x5				83
x6		66		
x7			-83	
x8	85			
x9		75		
x10			-79	
x11	82			
x12		73		

La corrélation entre les variables explicatives découlera de celle avec les facteurs

Le modèle factorielle donne une approximation de la corrélation.

Combien d'observations pour l'estimation?

Il faut un échantillon de taille conséquente.

- entre 5 et 20 fois p , le nombre de variables
- un nombre minimal de $n = 100$ à $n = 1000$ observations

Essentiellement des règles du pouce.

On transforme la solution pour garantir une solution **interprétable** puisque cette dernière n'est pas unique.

- La rotation *varimax* maximise la variance de la somme des carrés des chargements pour les facteurs.
- Donne des chargements dispersés (valeurs élevées positives ou négatives, d'autres presque nuls).

Garder m vecteurs propres et valeurs propres, puis effectuer une rotation (varimax)

- estimation toujours valide et rapide.
- sélection moins objective que maximum de vraisemblance (critère du coude ou de Kaiser)

```
library(hecmulti)
facto_cp <- factocp(factor,
                    nfact = "kaiser",
                    cor = TRUE)
# nfact: nombre de facteurs ("kaiser" par défaut)
# cor: matrice de corrélation? par défaut vrai
```

- Postulat de normalité des aléas et des facteurs.
- Nécessite une optimisation numérique délicate:
 - les problèmes de convergence sont fréquents
 - variances estimées parfois négatives!
 - appelés cas de **(quasi)-Heywood**.
- Méthodes de sélection plus informatives
 - critères d'information
 - tests d'hypothèse d'adéquation

Plus le nombre de facteurs m est grand, plus la corrélation modélisée se rapproche de la corrélation empirique.

Mais plus le nombre de paramètres est grand...

Valable uniquement pour les modèles ajustés par **maximum de vraisemblance**

$$\text{AIC} = -\text{ajustement} + 2 \times \text{nb param}$$

$$\text{BIC} = -\text{ajustement} + \ln(n) \times \text{nb param}$$

- Plus le critère d'information est petit, meilleur c'est
- Le BIC (critère Bayésien de Schwarz) pénalise davantage que le AIC (critère d'Akaike).

```
library(hecmulti)
ajustement_factanal(
  covmat = cor(factor), # matrice de corrélation
  factors = 1:5, # candidats pour nb de facteurs
  n.obs = nrow(factor)) # nombre d'observations
```

Table 3: Qualité de l'ajustement de modèles d'analyse factorielle (maximum de vraisemblance).

k	AIC	BIC	valeur-p	npar	heywood
1	2267.14	2346.30	$< 2e-16$	24	0
2	2137.87	2253.31	$< 2e-16$	35	0
3	2017.19	2165.61	0.09604	45	0
4	2002.56	2180.67	0.97262	54	1
5	2012.70	2217.19	0.97445	62	1

Le tableau inclut

- critères d'informations AIC et BIC
- valeur- p du test de rapport de vraisemblance comparant le modèle saturé (corrélations empiriques) et le modèle factoriel
- nombre de paramètres estimés
- indicateur pour les cas de (quasi)-Heywood

Les critères d'information suggèrent $m = 4$ facteurs (AIC) ou $m = 3$ (BIC)

- mais la solution à quatre facteurs n'est pas valide.
- le modèle à trois facteurs est préférable (simplification adéquate)

```
# Ajuster le modèle factoriel
# par maximum de vraisemblance
fa3 <- factanal(x = factor,
                factors = 3L)
# Imprimer les chargements en
# omettant les valeurs inférieures à 0.3
print(fa3$loadings,
      cutoff = 0.3)
```

Table 4: Estimés des chargements (pourcentage).

	F1	F2	F3
x1			
x2			67
x3		76	
x4	71		
x5			
x6		50	
x7			75
x8	79		
x9		63	
x10			67
x11	72		
x12		60	

Chargements de même signe et plus grands que 30%:

- Facteur 1: X_4 , X_8 et X_{11}
- Facteur 2: X_3 , X_6 , X_9 et X_{12}
- Facteur 3: X_2 , X_7 et X_{10}

Ces facteurs sont interprétables:

- F_1 : importance accordée au service.
- F_2 : importance accordée aux produits.
- F_3 : importance accordée à la facilité de paiement.

Si on ajuste le modèle à quatre facteurs, on obtient $\text{Cor}(X_1, F_4) = 0.99$ et $\text{Cor}(X_5, F_4) = 0.37$.

Cas de Heywood (trop de facteurs.)

Le facteur 4 représenterait le prix. On pourrait directement inclure X_1 .

- Créer de nouvelles variables selon les chargements
- moyenne équipondérée des variables explicatives fortement corrélées avec les facteurs

```
# Création des échelles
ech_service <- rowMeans(factor[,c("x4","x8","x11")])
ech_produit <- rowMeans(factor[,c("x3","x6","x9","x12")])
ech_paieement <- rowMeans(factor[,c("x2","x7","x10")])
ech_prix <- rowMeans(factor[,c("x1","x5")])
```

- En pratique, le coefficient α de Cronbach est fréquemment employé.
- Échelle fiable si $\alpha \geq 0.6$ (règle arbitraire)
- Plus α est élevé, plus les variables sont corrélées entre elles.

```
alphaC(factor[,c("x4","x8","x11")])  
alphaC(factor[,c("x3","x6","x9","x12")])  
alphaC(factor[,c("x2","x7","x10")])  
alphaC(factor[,c("x1","x5")])
```

Table 5: Coefficient alpha de Cronbach pour les quatre échelles formées.

service	produit	paiement	prix
0.781	0.718	0.727	0.546

La quatrième échelle (prix) n'est pas cohérente. On pourrait conserver la question X_1 plutôt.

- La corrélation mesure la force de la dépendance linéaire entre deux variables
 - plus elle est élevée, plus les points s'alignent.
- Si p grand et n petit, peu d'information disponible pour estimer de manière fiable les corrélations.

Une analyse en composante principales fait une décomposition en valeurs propres/vecteurs propres de la matrice de covariance ou de corrélation.

- Nouvelles variables sont orthogonales (corrélation nulle)
- Composantes principales en ordre décroissant de variance
- si on ne conserve que $k < p$ composantes principales, on maximise la variance expliquée.

- Choix du nombre de variables (diagramme d'éboulis, critère de Kaiser).
- Représentation graphique avec bigramme (directions des variables en fonction des deux premières composantes principales).

- L'analyse factorielle exploratoire fournit un modèle pour la matrice de corrélation
- Seules les variables numériques pour lesquelles on suspecte une dimension commune sont incluses dans l'analyse (questionnaires!)
- On doit avoir beaucoup d'observations (au moins 100, 10 fois plus que de variables) pour estimer le modèle.

On estime le modèle à l'aide de

- composantes principales
 - modèle toujours valide
 - moins coûteux en calcul
 - critères pour la sélection du nombre de facteurs arbitraires
- maximum de vraisemblance
 - optimisation numérique
 - solutions fréquemment problématique
 - critères d'information

Le nombre de facteurs retenu doit donner des regroupements logiques (facteur *wow*).

- La solution du problème n'est pas unique
 - on choisit celle qui permet de mieux séparer les variables.
 - par défaut, rotation varimax pour faciliter l'interprétation.
- L'interprétation se fait à partir des chargements (corrélation entre variables et facteurs).

- On crée des échelles en prenant la moyenne des variables qui ont un chargement élevés en lien avec un facteur donné (de même signe).
- Les échelles sont cohérentes si le α de Cronbach est supérieur à 0.6, faute de quoi elles sont rejetées.