Analyse de données

Jacques Bailhache (jacques.bailhache@gmail.com)

August 9, 2020

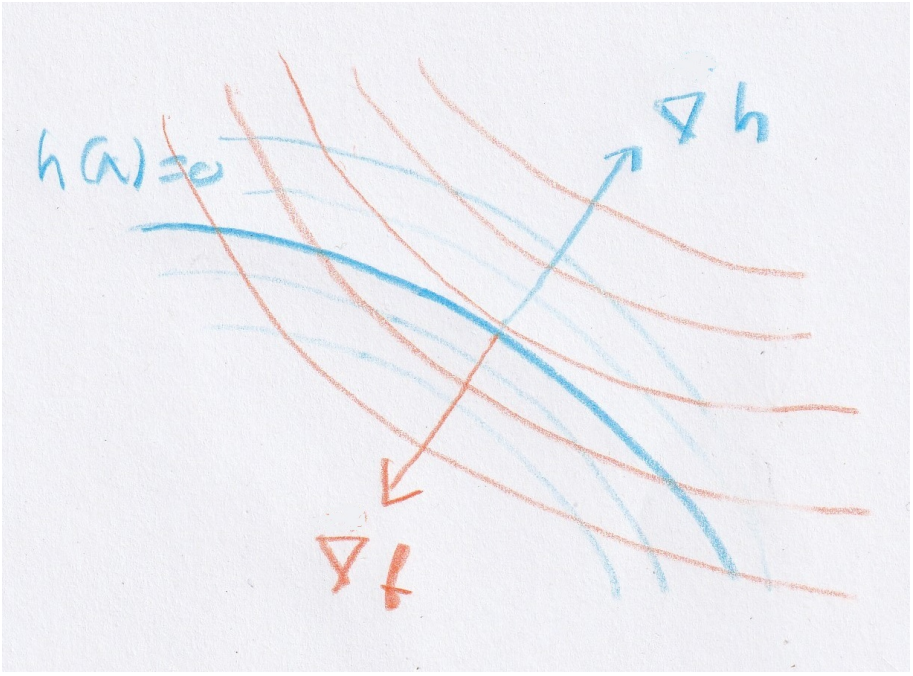
1 Prérequis

1.1 Multiplicateurs de Lagrange

1.1.1 Exemple dans un espace de dimension 2

Soit une fonction $h : \mathbb{R}^2 \to \mathbb{R}$.
L'équation h(x)=0 définit une courbe dans $\mathbb{R}^2$.
Soit une autre fonction $f : \mathbb{R}^2 \to \mathbb{R}$.
On cherche le point x de la courbe h(x)=0 qui minimise f, autrement dit x tel que :

- h(x)=0
- f(x) est minimal



On voit que x est le point où la courbe h(x)=0 est tangente à une courbe de niveau de f. Le gradient de f en ce point est perpendiculaire à la courbe h(x)=0. Le gradient de h étant également perpendiculaire à la courbe h(x)=0, les gradients de f et h sont alignés ou proportionnels, ce que l'on peut écrire sous la forme :

$$\nabla f(x) + \lambda \nabla h(x) = 0$$

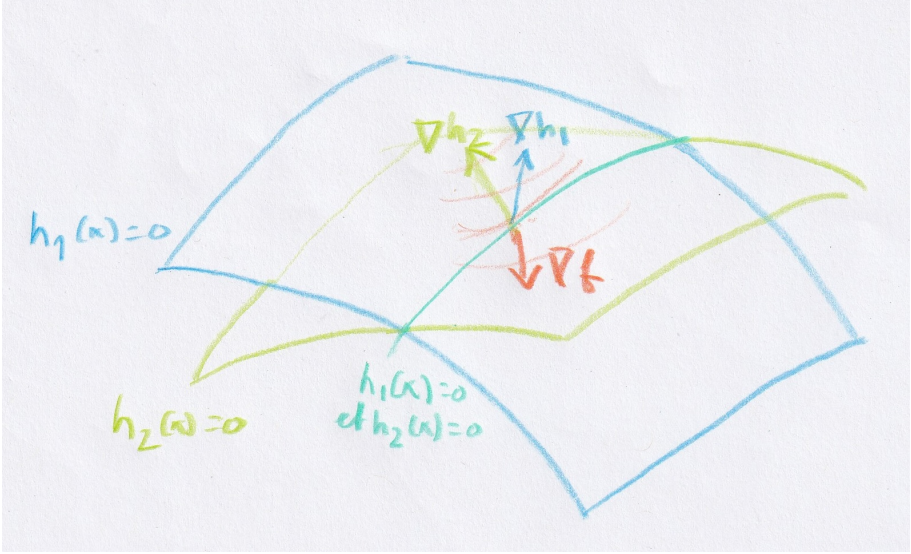
1.1.2 Exemple dans un espace de dimension 3

Soit deux fonction h_1 et $h_2 : \mathbb{R}^3 \to \mathbb{R}$.
L'équation $h_1(x) = 0$ définit une surface. L'équation $h_2(x) = 0$ définit une autre surface. La courbe C intersection de ces deux surfaces satisfait le système d'équations

- $h_1(x) = 0$
- $h_2(x) = 0$

On cherche le point x de cette courbe qui minimise une certaine fonction $f : \mathbb{R}^3 \to \mathbb{R}$, autrement dit x tel que :

- $h_1(x) = 0$
- $h_2(x) = 0$
- f(x) est minimal



Le point x est le point où la courbe C est tangente à une surface où f a la même valeur. On a :

$$\nabla f(x) + \lambda_1 \nabla h_1(x) + \lambda_2 \nabla h_2(x) = 0$$

1.1.3 Cas général

Dans $\mathbb{R}^n$, on considère p+1 fonctions $h_1, ..., h_p, f : \mathbb{R}^n \to \mathbb{R}$.
On cherche x tel que

- $h_1(x) = 0$
- ...
- $h_p(x) = 0$
- f(x) est minimal

Ce point x vérifie :

$$\nabla f(x) + \lambda_1 \nabla h_1(x) + \ldots + \lambda_p \nabla h_p(x) = 0$$

On définit le lagrangien L par :

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \ldots + \lambda_p h_p(x)$$

On a alors :

$$\nabla L(x, \lambda) = \frac{\partial L(x, \lambda)}{\partial x_i} = \nabla f(x) + \lambda_1 \nabla h_1(x) + \ldots + \lambda_p h_p(x)$$

La condition $\nabla f(x) + \lambda_1 \nabla h_1(x) + \ldots + \lambda_p \nabla h_p(x) = 0$ peut donc s'écrire :

$$\nabla L(x, \lambda) = 0$$

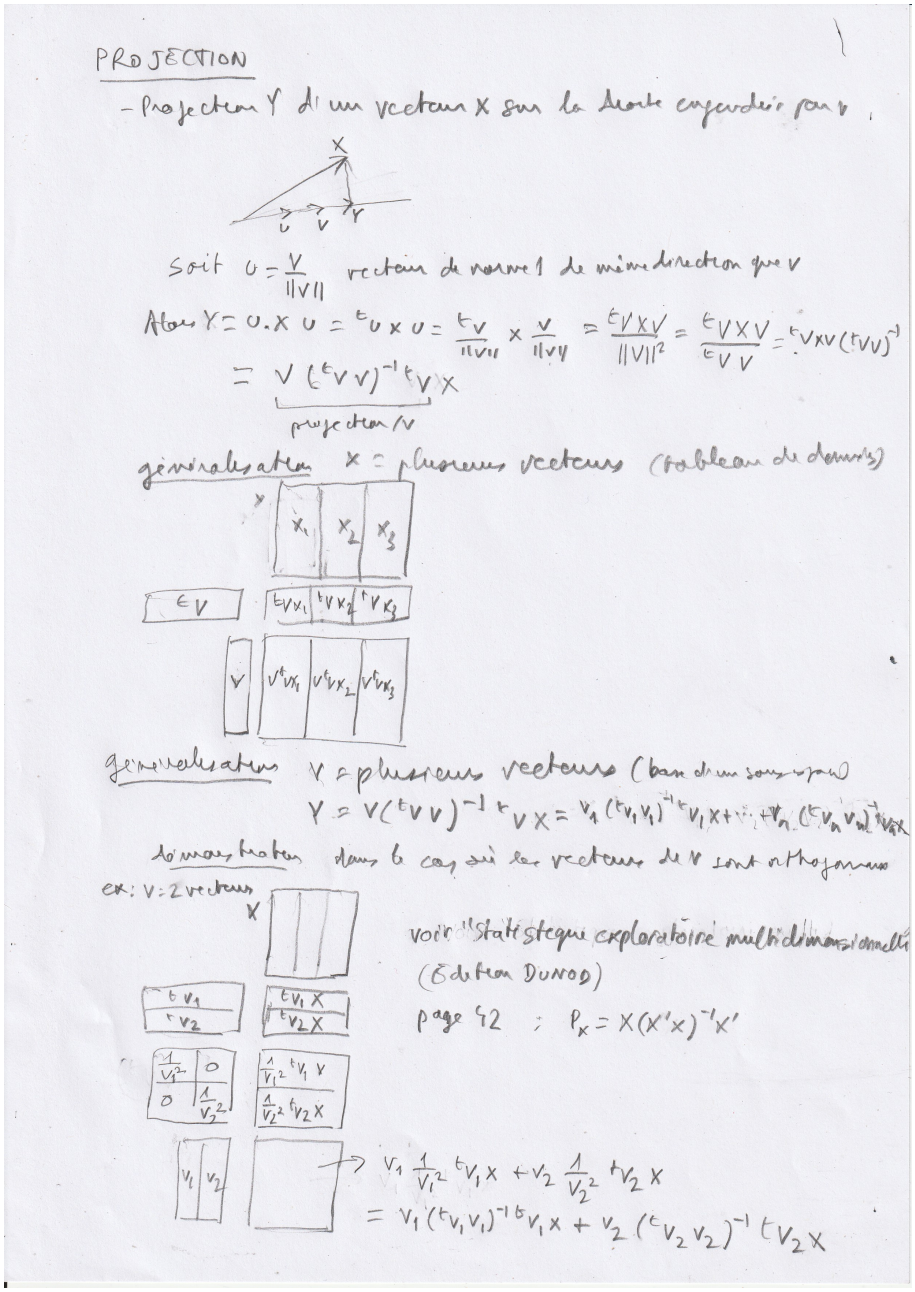
ou

$$\frac{\partial L(x,\lambda)}{\partial x_i} = 0$$

On peut également réécrire la condition $h_j(x) = 0$ sous la forme :

$$\frac{\partial L(x,\lambda)}{\partial \lambda_j} = 0$$

1.2 Projection



1.3 Valeurs et vecteurs propres d'une matrice

Calcul des valeurs et vecteurs propres par la méthode de la puissance itérée et de la déflation en J :

```
NB. Calcul des valeurs et vecteurs propres d'une matrice par la méthode de la puissance itérée et de la déflatio
NB. Référence :
```

```
NB.  http://www.bibmath.net/dico/index.php?action=affiche&quoi=./m/methodepuissance.html
NB.  http://log.chez.com/text/math/methodepuissance.pdf
NB.  http://www.normalesup.org/~pastre/meth-num/MN/9-val-pro/cours-valeurspropres.pdf
NB.  http://log.chez.com/text/math/cours-valeurspropres.pdf

and =: *.
or =: +.
not =: -.

transpose =: |:
extprod =: */
matprod =: +/ . *

avg =: 3 : 0
 (+/ y) % #y
)

puissance =: 3 : 0
 A =. y
 n =. #A
 u =. 1 , (n-1) $ 0
 i =. 0
 while. (i < 10000)
 do.
  i =. i + 1
  u1 =. u
  u =. A matprod u
  u =. u % (u matprod u)^0.5
  d =. u - u1
  if. (d matprod d) < 1e_30
  do. break.
  end.
 end.
 l =. avg (A matprod u) % u
 l; u; i
)

deflation =: 3 : 0
 A =. y
 NB. B = transpose A
 n =. #A
 l =. 0 $ 0
 u =. (n,0) $ 0
 for. A
 do.
  NB. echo 'A='; A
  lu1 =. puissance A
  NB. echo lu1
  l1 =. > 0 { lu1
  u1 =. > 1 { lu1
  l =. l , l1
  u =. u ,. u1
  lu2 =. puissance (transpose A)
  l2 =. > 0 { lu2
  u2 =. > 1 { lu2
  A =. A - l1 * u1 extprod u2 % u1 matprod u2
```

```
  NB. B =. B - l1 * u2 extprod u1 % u2 matprod u1
 end.
 l ; u
)
```

2 Analyse en composantes principales

On a un tableau de données X à n lignes et p colonnes. Chaque ligne représente un individu et chaque colonne une variable. Ce tableau représente un ensemble de points dans un espace de dimension p.
L'analyse en composantes principales consiste à projeter ces points sur un sous-espace de dimension $q < p$ tel que les projections des points soient le plus écartés possible.

2.1 Cas q = 1 : projection sur une droite

Soit u_1 un vecteur unitaire (${}^t u_1 u_1 = 1$) de cette droite.
Les projections de ces n points sur cette droits sont :

$$C = Xu_1$$

L'écartement de ces projections est mesuré par une valeur appelée inertie que l'on calcule par la formule :

$${}^t CC = {}^t u_1 {}^t XXu_1$$

Le problème consiste donc à maximiser ${}^t u_1 {}^t XXu_1$ sous la contrainte ${}^t u_1 u_1 = 1$.
On définit le lagrangien :

$$L(u_1, \lambda) = {}^t u_1 {}^t XXu_1 - \lambda_1({}^t u_1 u_1 - 1)$$

On a alors :

- $\frac{\partial L(u_1,\lambda_1)}{\partial u_1} = 0 \Rightarrow 2({}^t XXu_1 - \lambda_1 u_1) = 0 \Rightarrow {}^t XXu_1 = \lambda_1 u_1$
- $\frac{\partial L(u_1,\lambda_1)}{\partial \lambda_1} = 0 \Rightarrow {}^t u_1 u_1 - 1 = 0 \Rightarrow {}^t u_1 u_1 = 1$

u_1 est le premier vecteur propre de ${}^t XX$ et λ_1 la valeur propre correspondante.

2.2 Cas q = 2 : projection sur un plan

Le plan de projection optimal, c'est-à-dire celui qui maximise l'écartement des projections, contient nécessairement la droite optimale considérée dans la section précédente. Ce plan est donc défini par le vecteur u_1 déterminé précédemment, et par un autre vecteur u_2 que nous devons trouver.
On cherche u_2 tel que ${}^t u_2 {}^t XXu_2$ soit maximal, sous les contraintes ${}^t u_2 u_2 = 1$ et ${}^t u_2 u_1 = 0$.
Le lagrangien correspondant est :

$$L(u_2, \lambda_2, \mu) = {}^t u_2 {}^t XXu_2 - \lambda_2({}^t u_2 u_2 - 1) - \mu({}^t u_2 u_1)$$

En écrivant que la dérivée partielle de ce lagrangien respectivement par rapport à u_2, λ_2, μ est égale à 0, on obtient :

- ${}^t XXu_2 = \lambda_2 u_2$
- ${}^t u_2 u_2 = 1$
- ${}^t u_1 u_2 = 0$

u_2 est le deuxième vecteur propre de ${}^t XX$ associé à la deuxième valeur propre λ_2.

2.3 Cas général

On peut montrer que le sous-espace de dimension q pour lequel les projections des points représentés par X sont le plus écartés est défini par les q vecteurs propres de ${}^t XX$ associés aux q plus grandes valeurs propres.

2.4 Exemple d'ACP en J

```
NB. Analyse en composantes principales
NB. Exemple tiré de http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/ACP_L3.pdf
NB.                 http://log.chez.com/text/math/ACP_L3.pdf

transpose =: |:
matprod =: + / . *
id =: (= / ~) @ i.

diag =: 3 : 0
 y * id # y
)

NB. Tableau de données
X =: 1 3 $ 90 140 6.0
X =: X ,   60  85 5.9
X =: X ,   75 135 6.1
X =: X ,   70 145 5.8
X =: X ,   85 130 5.4
X =: X ,   70 145 5.0

n =: # X
p =: # 0 { X

M =: (+ / X) % # X                    NB. Moyenne des colonnes
Y =: X - (1 + 0 * i. # X) * / M       NB. Données centrées
E =: ((+ / Y ^ 2) % # Y) ^ 0.5        NB. Ecarts-types des colonnes
Z =: Y % (1 + 0 * i. # Y) * / E       NB. Données centrées-réduites

C =: ((transpose Y) matprod Y) % # Y  NB. Matrice des covariances
R =: ((transpose Z) matprod Z) % # Z  NB. Matrice des corrélations

LV =: deflation R                     NB. Valeurs et vecteurs propres
V =: > 1 { LV               NB. Vecteurs propres
NB. V =: V % (1 + 0 * i. # V) * / (+ / V ^ 2) ^ 0.5   NB. Vecteurs propres normés
F =: Z matprod V                      NB. Composantes principales

echo 'Composantes principales :'
echo F

S =: (Z matprod transpose Z) % # Z
KU =: deflation S
U =: > 1 { KU
NB. U =: U % (1 + 0 * i. # U) * / (+ / U ^ 2) ^ 0.5
A =: ((transpose Z) matprod U) % # Z

echo 'Coordonnées factorielles des variables :'
echo A

L =: diag (> 0 { LV) ^ 0.5
K =: diag (> 0 { KU) ^ 0.5

F1 =: U matprod K
A1 =: V matprod L
```

2.5 Références

- http://cours.polymtl.ca/geo/marcotte/glq3402/chapitre3.pdf
  ou http://log.chez.com/text/math/chapitre3.pdf
- http://www.arnaud.martin.free.fr/Doc/polyAD.pdf
  ou http://log.chez.com/text/math/polyAD.pdf
- http://www2.agroparistech.fr/IMG/pdf/AnalyseComposantesPrincipales-AgroParisTech.pdf
  ou http://log.chez.com/text/math/AnalyseComposantesPrincipales-AgroParisTech.pdf
- http://asi.insa-rouen.fr/enseignants/~gasso/public/Courses/DM/pca.pdf
  ou http://log.chez.com/text/math/pca.pdf
- https://www.lamsade.dauphine.fr/~atif/lib/exe/fetch.php?media=teaching:coursad_acp.pdf
  ou http://log.chez.com/text/math/coursad_acp.pdf
- http://hamrita.e-monsite.com/medias/files/chap2ad.pdf
  ou http://log.chez.com/text/math/chap2ad.pdf
- http://www.foad-mooc.auf.org/IMG/pdf/M03-5.pdf
  ou http://log.chez.com/text/math/M03-5.pdf
- https://www.iro.umontreal.ca/~vincentp/ift3395/cours/continuous_latent_variables_print.pdf
  ou http://log.chez.com/text/math/continuous_latent_variables_print.pdf
- http://www.math.tu-dresden.de/~gournay/SMV.pdf
  ou http://log.chez.com/text/math/SMV.pdf
- http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/d512ad5b22d73cc1c1257052003f1aed/da8b20974b2986
  $FILE/Me%CC%81moire%20TRIEU%20Thi%20Diep.pdf
  ou http://log.chez.com/text/math/Me%CC%81moire%20TRIEU%20Thi%20Diep.pdf

3 Analyse canonique

L'analyse canonique consiste, à partir de données avec deux groupes de variables, à trouver une combinaison linéaire des variables du premier groupe et une combinaison linéaire des variables du deuxième groupe qui soient les plus corrélées possibles.

3.1 Formulation mathématique

Les données sont représentées par deux tableaux X et Y correspondant aux deux groupes de variables. Les deux tableaux ont n lignes (nombre d'individus). Le tableau X a p colonnes (nombre de variables du premier groupe). Le tableau Y a q colonnes (nombre de variables du deuxième groupe).
Le problème consiste à trouver a et b qui rendent maximal ${}^t a {}^t XYb$ avec les contraintes ${}^t a {}^t XXa = 1$ et ${}^t b {}^t YYb = 1$.
Il s'agit de rendre maximal le lagrangien associé à ce problème :

$$ {}^t a {}^t XYb - \lambda({}^t a {}^t XXa - 1) - \mu({}^t b {}^t YYb - 1) $$

En écrivant que les dérivées partielles de ce lagrangien par rapport à a et b sont égales à 0, on obtient :

- ${}^t XYb - 2\lambda {}^t XXa = 0$
- ${}^t YXa - 2\mu {}^t YYb = 0$

En multipliant respectivement par ${}^t a$ et ${}^t b$ on obtient :

- ${}^t a {}^t XYb = 2\lambda$
- ${}^t b {}^t YXa = 2\mu$

On a donc $\lambda = \mu$. On pose $\beta = 2\lambda = 2\mu$. On a alors :

- ${}^tXYb = \beta {}^tXXa$
- ${}^tYXa = \beta {}^tYYb$

donc

$$ {}^tYX({}^tXX)^-1 {}^tXYb = \beta^2 \ {}^tYYb$$

Donc b est vecteur propre de

$$M = ({}^tYY)^{-1} \ {}^tYX({}^tXX)^{-1} \ {}^tXY$$

De même, a est vecteur propre de

$$N = ({}^tXX)^{-1} \ {}^tXY({}^tYY)^{-1} \ {}^tYX$$

et

$$a = \frac{1}{\beta}({}^tXX)^{-1} \ {}^tXYb \ ; \ b = \frac{1}{\beta}({}^tYY)^{-1} \ {}^tYXa$$

En multipliant respectivement par X ey Y on obtient :

- $Xa = \frac{1}{\beta}X({}^tXX)^{-1} \ {}^tXYb$
- $Yb = \frac{1}{\beta}Y({}^tYY)^{-1} \ {}^tYXa$

Dans ces formules apparaissent les opérateurs de projection orthogonale $P_X = X({}^tXX)^{-1} \ {}^tX$ et $P_Y = Y({}^tYY)^{-1} \ {}^tY$. Donc chacun des vecteurs Xa et Yb est colinéaire à la projection de l'autre.

3.2 Exemple d'analyse canonique en J

```
NB. Analyse canonique
NB. D'après http://iml.univ-mrs.fr/~reboul/canonique.pptx.pdf
NB.         http://log.chez.com/text/math/canonique.pptx.pdf

transpose =: |:      NB. Transposition de matrice
matprod =: + / . *   NB. Produit de matrices
inv =: %.            NB. Inverse
id =: (= / ~) @ i.   NB. Matrice identité

NB. Données

X =: 1 2 $ 100  100
X =: X ,   200  400
X =: X ,  _400 _200
X =: X ,   200 _300
X =: X ,  _100    0

Y =: 1 3 $ 200    0 _107
Y =: Y ,   600 _300  212
Y =: Y ,  _600 _200  233
Y =: Y ,  _200  200   92
Y =: Y ,     0  300 _430

NB. Réduction des données
X =: X % ((#X) # 1) */ (+/X^2)^0.5
Y =: Y % ((#Y) # 1) */ (+/Y^2)^0.5
```

```
VXX =: (transpose X) matprod X
VYY =: (transpose Y) matprod Y
VXY =: (transpose X) matprod Y
VYX =: (transpose Y) matprod X

RX =: (inv VXX) matprod VXY matprod (inv VYY) matprod VYX
RY =: (inv VYY) matprod VYX matprod (inv VXX) matprod VXY

LUX =: deflation RX
LUY =: deflation RY

UX =: > 1 { LUX
UY =: > 1 { LUY

echo 'Facteurs canoniques :'
echo ' '
echo UX
echo ' '
echo UY

NB. echo (transpose UX) matprod UX
NB. echo (transpose UY) matprod UY
```

3.3 Référence

Statistique exploratoire multidimensionnelle, DUNOD, 2.1 Analyse canonique

4 Analyse canonique généralisée

L'analyse canonique généralisée est une méthode d'analyse de tableaux de données de n lignes (individus) et p colonnes (variables) groupées en q groupes, que l'on peut écrire sous la forme :

$$X = (X_1, X_2, \ldots, X_k, \ldots, X_q)$$

Si q = 2, on retrouve l'analyse canonique classique.
Si chaque bloc X_k est un tableau disjonctif complet, on retrouve l'analyse des correspondances multiples.
Si chaque bloc n'est formé que d'une seule colonne, on retrouve l'analyse en composantes principales.

4.1 Formulation mathématique

Il s'agit de maximiser $\sum_{k=1}^{q} {}^t y X_k ({}^t X_k X_k)^{-1} {}^t X_k y$ avec la contrainte ${}^t y y = 1$.
Le vecteur y est le vecteur propre correspondant à la plus grande valeur propre de la matrice :

$$S = \sum_{k=1}^{q} X_k ({}^t X_k X_k)^{-1} {}^t X_k$$

4.2 Cas q = 2 : Analyse canonique classique

Pour q = 2, on a :

$$X_1 ({}^t X_1 X_1)^{-1} {}^t X_1 y + X_2 ({}^t X_2 X_2)^{-1} {}^t X_2 y = \lambda y$$

Posons $({}^t X_1 X_1)^{-1} {}^t X_1 y = a$ et $({}^t X_2 X_2)^{-1} {}^t X_2 y = b$.
On a alors

$$X_1 a + X_2 b = \lambda y$$

En multipliant à gauche par $({}^t X_1 X_1)^{-1}\ {}^t X_1$, on obtient :

$$({}^t X_1 X_1)^{-1}\ {}^t X_1 X_2 b = (\lambda - 1) a$$

De même, on a aussi

$$({}^t X_2 X_2)^{-1}\ {}^t X_2 X_1 a = (\lambda - 1) b$$

et par substitution

$$({}^t X_2 X_2)^{-1}\ {}^t X_2 X_1 ({}^t X_1 X_1)^{-1}\ {}^t X_1 X_2 b = (\lambda - 1)^2 b$$

On retrouve la matrice à diagonaliser de l'analyse canonique classique.

4.3 Cas où chaque bloc comprend une seule colonne : Analyse en composantes principales

Dans ce cas on a :

$$S = \sum_{k=1}^{q} x_k ({}^t x_k x_k)^{-1}\ {}^t x_k = \sum_{k=1}^{q} \frac{1}{n s_k^2} x_k\ {}^t x_k$$

avec

$$s_k^2 = \frac{1}{n}\ {}^t x_k x_k$$

Soit la matrice T dont la k-ième colonne vaut $t_k = \frac{1}{s_k} x_k$.
Alors on a $S = \frac{1}{n} T\ {}^t T$.
La relation $Sy = \lambda y$ s'écrit alors :

$$\frac{1}{n} T {}^t T y = \lambda y$$

En multipliant à gauche par ${}^t T$ on obtient :

$$\frac{1}{n}\ {}^t T T {}^t T y = \lambda {}^t T y$$

et en posant ${}^t T y = u$:

$$\frac{1}{n}\ {}^t T T u = \lambda u$$

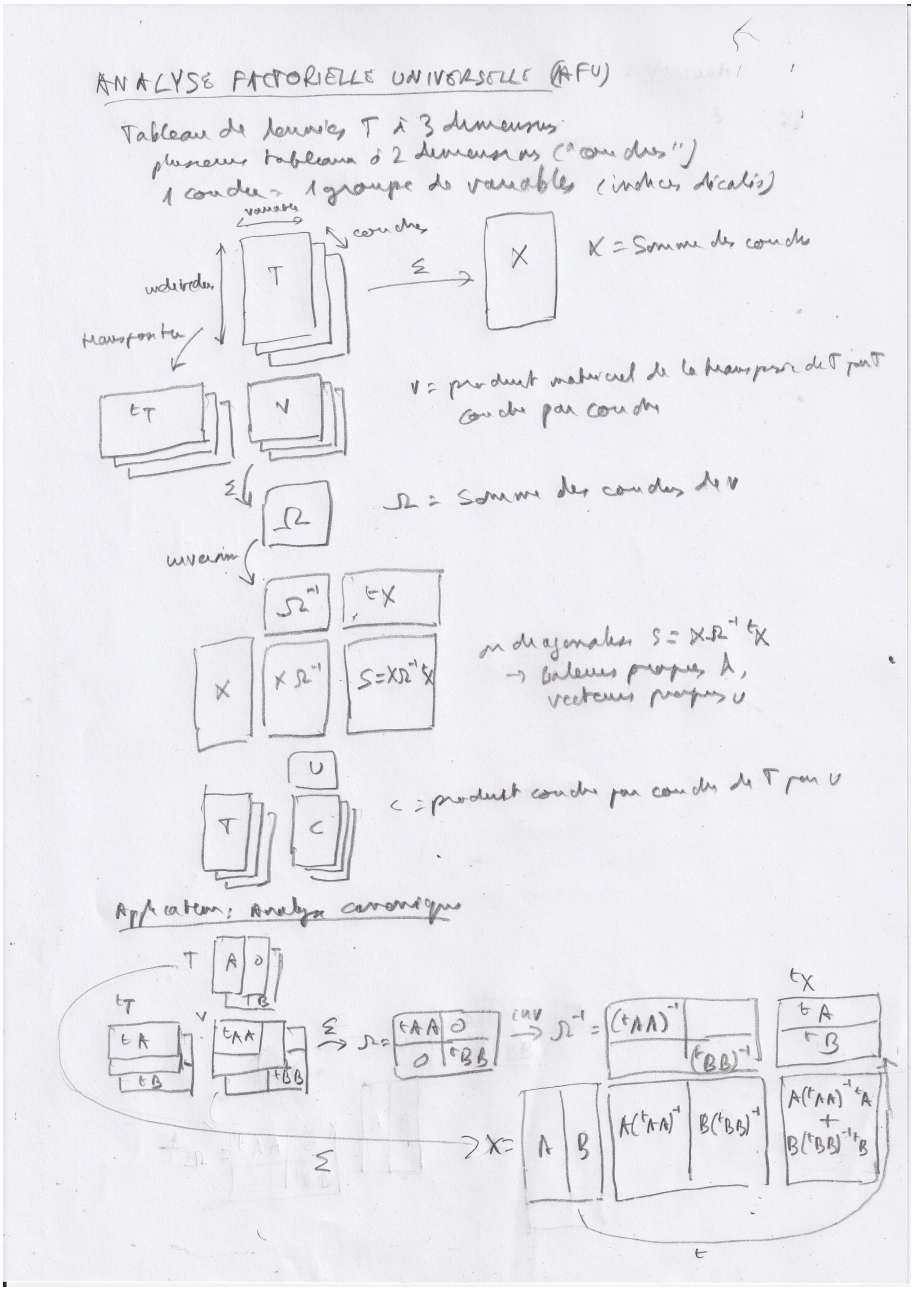
ou

$$C u = \lambda u$$

4.4 Référence

Statistique exploratoire multidimensionnelle, éditions DUNOD, 8.3.5 Analyse canonique généralisée

5 Analyse factorielle universelle



Exemple de programme J effectuant une ACP avec une AFU :

```
NB. Analyse factorielle universelle

transpose =: |:       NB. Transposition de matrices
matprod =: + / . *    NB. Produit de matrices
id =: (= / ~) @ i.    NB. Matrice identité

diag =: 3 : 0         NB. Matrice diagonale à partir d'un vecteur
 y * id #y
)

tmatprod =: 3 : 0
 (transpose y) matprod y
)
```

```
maptmatprod =: 3 : 0
 if. 0 = # y do. 0 0 0 $ 0
 else. (tmatprod {. y), maptmatprod }. y
 end.
)

mapprod =: 4 : 0
 if. 0 = # x do. 0 0 0 $ 0
 else. (({. x) matprod y) , (}. x) mapprod y
 end.
)

NB. Tableau de données
X =: 1 3 $ 90 140 6.0
X =: X ,   60  85 5.9
X =: X ,   75 135 6.1
X =: X ,   70 145 5.8
X =: X ,   85 130 5.4
X =: X ,   70 145 5.0

n =: # X
p =: # 0 { X

M =: (+ / X) % #X                    NB. Moyenne des colonnes
Y =: X - (1 + 0 * i. #X) * / M       NB. Données centrées
E =: ((+ / Y ^ 2) % #Y) ^ 0.5        NB. Ecarts-types des colonnes
Z =: Y % (1 + 0 * i. #Y) * / E       NB. Données centrées-réduites

NB. Tableau tridimensionnel pour l'AFU
T =: ((1 + 0 * i. # 0 { Z) * / Z) * 0 2 1 |: (id # 0 { Z) * / (1 + 0 * i. #Z)

Z1 =: + / T   NB. Tableau bidimensionnel somme des couches

V =: maptmatprod T
W =: + / V
S =: Z1 matprod (%. W) matprod (transpose Z1)
LU =: deflation S   NB. Valeurs et vecteurs propres de S
echo LU
U =: > 1 { LU
V =: (transpose Z1) matprod U
V =: V % (1 + 0 * i. #V) * / (+ / V ^ 2) ^ 0.5
C =: T mapprod V
F =: + / C
echo 'Composantes principales :'
echo F
```

Exemple de programme J effectuant une analyse canonique avec une AFU :

```
NB. Analyse canonique avec AFU
NB. Données d'après http://iml.univ-mrs.fr/~reboul/canonique.pptx.pdf
NB.                 http://log.chez.com/text/math/canonique.pptx.pdf

transpose =: |:      NB. Transposition de matrice
matprod =: + / . *   NB. Produit de matrices
extprod =: */
```

```
inv =: %.              NB. Inverse
id =: (= / ~) @ i.   NB. Matrice identité

diag =: 3 : 0        NB. Matrice diagonale à partir d'un vecteur
 y * id #y
)

tmatprod =: 3 : 0
 (transpose y) matprod y
)

maptmatprod =: 3 : 0
 if. 0 = # y do. 0 0 0 $ 0
 else. (tmatprod {. y), maptmatprod }. y
 end.
)

mapprod =: 4 : 0
 if. 0 = # x do. 0 0 0 $ 0
 else. (({. x) matprod y) , (}. x) mapprod y
 end.
)

NB. Données

X =: 1 2 $ 100  100
X =: X ,   200  400
X =: X ,  _400 _200
X =: X ,   200 _300
X =: X ,  _100    0

Y =: 1 3 $ 200    0 _107
Y =: Y ,   600 _300  212
Y =: Y ,  _600 _200  233
Y =: Y ,  _200  200   92
Y =: Y ,     0  300 _430

NB. Réduction des données
X =: X % ((#X) # 1) */ (+/X^2)^0.5
Y =: Y % ((#Y) # 1) */ (+/Y^2)^0.5

T =: (1 0 extprod (X ,. (0*Y))) + (0 1 extprod ((0*X) ,. Y))
echo 'Tableau de données de l''AFU :'
echo T

Z1 =: + / T   NB. Tableau bidimensionnel somme des couches

V =: maptmatprod T
W =: + / V
S =: Z1 matprod (%. W) matprod (transpose Z1)
LU =: deflation S   NB. Valeurs et vecteurs propres de S
echo 'Eléments propres :'
echo LU
U =: > 1 { LU

u1 =: 0 { transpose U
```

```
echo 'Premier vecteur propre :'
echo u1

a =: (inv (transpose X) matprod X) matprod (transpose X) matprod u1
b =: (inv (transpose Y) matprod Y) matprod (transpose Y) matprod u1

a =: a % (a matprod a)^0.5
b =: b % (b matprod b)^0.5

u2 =: 1 { transpose U

a2 =: (inv (transpose X) matprod X) matprod (transpose X) matprod u2
b2 =: (inv (transpose Y) matprod Y) matprod (transpose Y) matprod u2

a2 =: a2 % (a2 matprod a2)^0.5
b2 =: b2 % (b2 matprod b2)^0.5

echo 'Facteurs canoniques :'
echo a
echo a2
```

6 Analyses de tableaux de données avec deux groupes de variables et analyse comparative universelle

Il existe plusieurs méthode d'analyse de tableaux de données avec deux groupes de variables. Nous avons déjà vu l'analyse canonique. Il existe d'autres méthodes que l'on peut résumer dans le tableau ci-dessous :

Nom de l'analyse	Matrices à diagonaliser
Analyse canonique	$({}^tXX)^{-1}\ {}^tXY({}^tYY)^{-1}\ {}^tYX$
Analyse projetée	${}^tXY({}^tYY)^{-1}\ {}^tYX$
Analyse procrustéenne orthogonale	${}^tXY{}^tYX$
Analyse procrustéenne sans contrainte	$({}^tXX)^{-1}\ {}^tXY$

Pour plus d'informations, voir Statistique exploratoire multidimensionnelle, DUNOD, Chapitre 8 Analyse de données structurées.

On voit que pour toutes ces analyses, la matrice à diagonaliser est de la forme :

$$({}^tXX)^a\ ({}^tXY)^b\ ({}^tYY)^c\ ({}^tYX)^d$$

L'analyse comparative universelle généralise toutes ces méthodes. C'est une méthode paramétrée par 4 coefficients a, b, c, d, qui consiste à diagonaliser la matrice indiquée ci-dessus. Elle englobe les méthodes indiquées ci-dessus, ainsi que l'ACP avec a = 1 et b = c = d = 0 :

Nom de l'analyse	a	b	c	d	Matrices à diagonaliser
Analyse en composantes principales	1	0	0	0	${}^tXX$
Analyse canonique	-1	1	-1	1	$({}^tXX)^{-1}\ {}^tXY({}^tYY)^{-1}\ {}^tYX$
Analyse projetée	0	1	-1	1	${}^tXY({}^tYY)^{-1}\ {}^tYX$
Analyse procrustéenne orthogonale	0	1	0	1	${}^tXY{}^tYX$
Analyse procrustéenne sans contrainte	-1	1	0	0	$({}^tXX)^{-1}\ {}^tXY$