L'Analyse en Composantes Principales

Table des matières

1 Introduction

X étant un tableau de p variables numériques (en colonnes) décrivant n individus (en lignes), nous nous proposons de rechercher une représentation des n individus $\underline{e_1}$, $\underline{e_2}$,..., $\underline{e_n}$ dans un sous-espace de l'espace initial. Autrement dit, nous cherchons à définir k nouvelles variables, combinaison des p de l'espace initial, qui feraient perdre

le moins « d'information possible ». Ces k variables seront appelées « composantes principales » et les axes qu'elles déterminent « axes principaux ».

2 Notations

- $\underline{e_1}, \ldots, \underline{e_n}$: vecteurs « individus » dans l'espace initial
- $\underline{f_1}, \ldots, \underline{f_n}$: vecteurs « individus » dans l'espace de projection
- $\underline{x_1}, \ldots, \underline{x_p}$: vecteurs « variables »
- $p_1, \ldots, p_n$: poids associés à chaque individu
- $\underline{g}$: centre de gravité du nuage de points dans l'espace initial
- $I_{\underline{g}}$: inertie totale du nuage de points

Dans ce qui suit nous considérons que les variables sont centrées (la moyenne par colonne est égale à 0)

3 Définitions

Une métrique est une matrice permettant de définir un produit scalaire et donc des **distances** entre individus ou entre variables.

La métrique que l'on utilise de manière naturelle pour mesurer les proximités entre variables est celle définie par la matrice D_p qui est la métrique de la covariance quand les variables sont centrées :

$$\begin{aligned} COV(\underline{x_i}; \underline{x_j}) &= \underline{x_i}^t D_p \underline{x_j} \\ VAR(\underline{x_i}) &= \underline{x_i}^t D_p \underline{x_i} \end{aligned}$$

D_p est diagonale et chacun des éléments de la diagonale est égal à $\frac{1}{N}$.

Pour mesurer des distances entre individus, il n'y a pas de choix aussi naturel que D_p pour les variables et on prendra une matrice M, symétrique définie positive :

$$d^2(\underline{e_i}; \underline{e_j}) = (\underline{e_i} - \underline{e_j})^t M (\underline{e_i} - \underline{e_j})$$

les métriques les plus couramment utilisées sont :

- $M = I$
- $M = V^{-1}$: métrique de Mahalanobis
- $M = D_{1/\sigma^2}$ où D_{1/σ^2} désigne la matrice diagonale des inverses des variances de p variables

4 Projections sur un sous-espace

On va chercher un sous-espace de l'espace initial tel que :

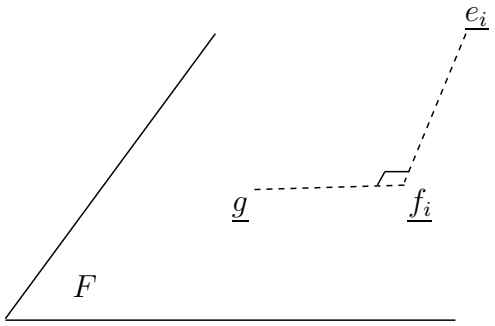
$$\sum_{i=1}^{n} p_i ||\underline{e_i} - \underline{f_i}||^2,$$

soit minimal.

Or d'après le théorème de Pythagore, minimiser l'expression ci-dessus, revient à maximiser

$$\sum_{i=1}^{n} p_i ||\underline{f_i} - \underline{g}||^2,$$

car on a :



$$||\underline{e_i} - \underline{g}||^2 = ||\underline{e_i} - \underline{f_i}||^2 + ||\underline{f_i} - \underline{g}||^2$$

Il est donc clair ici que **trouver les valeurs de $\underline{f_i}$ les plus proches de celles de $\underline{e_i}$ dans un nouvel espace, revient à maximiser la dispersion (ou inertie totale) des $\underline{f_i}$.**

L'inertie totale est définie comme la somme des distances de chaque individu au centre de gravité $\underline{g}$. Dans l'espace initial, on a donc :

$$I_{\underline{g}}^{init} = \sum_{i=1}^{n} = p_i \underline{e_i}^t M \underline{e_i}$$

puisque $\underline{g} = 0$ (les variables sont centrées). En utilisant l'écriture matricielle, on montre facilement que :

$$I_{\underline{g}}^{init} = Tr(MV)$$

où $V = X^t D_p X$, est la matrice de variance-covariance entre variables.
Dans l'espace des projetés, on calcule l'inertie du nuage projeté, $I_{\underline{g}}^{proj}$:

$$I_{\underline{g}}^{proj} = Tr(MVP)$$

où P est la matrice permettant de projeter le nuage de l'espace initial vers celui de l'espace des individus projetés.

5 Axes principaux

Nous cherchons la droite maximisant l'inertie du nuage projeté sur cette droite. Soit $\underline{a}$, un vecteur de cette droite, on a :

$$P = \underline{a}(\underline{a}^t M \underline{a})^{-1} \underline{a}^t M$$

Le but est donc de trouver $\underline{a}$ maximisant $Tr(VMP)$.
On montre que $\underline{a}$ est le vecteur propre de la matrice VM associé à la plus grande valeur propre.
De manière plus générale, le sous-espace des projetés de dimension k est engendré par les k vecteurs propres de VM associés aux k plus grandes valeurs propres. On appelle axes principaux d'inertie les vecteurs propres de VM normés à 1. Il y en a p.

6 Facteurs principaux

À l'axe principal $\underline{a}$ est associé le facteur principal $\underline{u} = M\underline{a}$. En partant du fait que $\underline{a}$ est déterminé par le vecteur propre de MV correspondant, on montre que $\underline{u}$ est déterminé par les vecteurs propres de VM.

7 Composantes principales

Ce sont les variables $\underline{c_i}$ définies par les facteurs principaux :

$$\underline{c_i} = X\underline{u_i}$$

$\underline{c_i}$ est le vecteur renfermant les coordonnées des projections des individus sur l'axe défini par $\underline{a_i}$. Ce sont donc les combinaisons linéaires de $\underline{x_1}$,..., $\underline{x_p}$ de variances maximales. On montre que la variance de $\underline{c_i}$ est égale à la valeur propre λ_i correspondante. Ces composantes sont orthogonales, et donc non corrélées entre elles.

8 Application

Récupérer le fichier `http://www.lirmm.fr/~guindon/dess/cocons.txt`. Les 3 variables étudiées ici sont la longeur, la plus petite largeur et la plus grande largeur de cocons de vers à soies. Les individus sont donc ici en lignes et les variables, en colonnes.

Questions :

– construire le tableau des données centrées

Solution :

```
> for(i in 1:3) d[,i] <- d[,i]-mean(d[,i])
> d
         X      Y   Z
1    13.92  -1.92  -1
2   -21.08 -19.92 -20
3     7.92  10.08   6
4    27.92  -6.92 -11
5    -2.08  -4.92  -3
6   -23.08  -9.92  -4
7    28.92  17.08  21
8     8.92  14.08  12
9     2.92  -5.92  -7
10   12.92   1.08   5
11    2.92  -2.92  -3
12  -14.08 -17.92  -9
13  -28.08 -19.92 -16
14    5.92  35.08  37
15    7.92  15.08  20
```

```
16   0.92  -5.92  -8
17  26.92  21.08  20
18 -16.08  17.08  20
19  42.92  -8.92  -7
20 -19.08   3.08  11
21  -6.08  -5.92 -11
22 -17.08   9.08   7
23 -15.08 -15.92 -24
24 -16.08 -10.92 -16
25 -13.08  -4.92 -19
```

– calculer la matrice de variance-covariance des données centrées (utiliser %*% pour le produit matriciel)
Solution :

```
> t(d)%*%mv%*%d
           X         Y      Z
  X 335.1136  98.8864  98.56
  Y  98.8864 190.3136 197.60
  Z  98.5600 197.6000 230.16
```

– selon vous, doit-on ici réduire les variables pour effectuer l'ACP ?

Solution : les variances des trois variables sont du même ordre de grandeur : on ne réduit pas les variables.

– calculer les valeurs propres et vecteurs propres de la matrice de covariance (utiliser la fonction `eigen`)

Solution :

```
eigen(t(d)%*%mv%*%d);
$values
[1] 516.23337 227.90979  11.44405

$vectors
            Z          Y           X
  X 0.6101958 -0.7918123 -0.02634923
  Y 0.5378277  0.3895879  0.74763805
  Z 0.5817237  0.4703770 -0.66358351
```

Chaque valeur propre correspond à la variance des projetés sur la composante principale correspondante. La somme des valeurs propres est égale à la variance totale dans l'espace des projetés. Cette variance totale est égale à celle de l'espace initial.

– calculer les coordonnées des projetés.

Solution :

```
> z <- d%*%eig$vectors
> z
```

	Z	Y	X
1	6.879573	-12.24041332	-1.13866287
2	-35.210929	-0.47672709	-1.06583792
3	13.744396	0.47815449	3.34600454
4	6.915940	-29.97749517	1.39009269
5	-5.660491	-1.68093386	-1.63282226
6	-21.745465	12.52880834	-4.15409509
7	39.049157	-6.36713424	-1.92761562
8	19.996245	4.06695569	2.32870646
9	-5.474234	-7.91109129	0.14212755
10	11.373203	-7.45757539	-2.85090054
11	-1.533856	-4.86081963	-0.26929234
12	-23.464942	-0.06609081	-7.05442503
13	-37.155405	6.94746706	-3.53572731
14	44.003130	26.38316334	1.51856549
15	24.577666	9.01137172	-2.20597431
16	-7.276349	-6.79784362	0.85840952
17	39.398353	-3.69553488	1.77921853
18	11.008621	28.79404331	-0.07831661
19	17.320117	-40.75234797	-3.15475593
20	-3.587067	21.48185662	-4.49395001
21	-13.292891	-2.66628827	3.03360467
22	-1.466604	20.35425161	2.59351383
23	-31.725338	-5.55075723	4.42095288
24	-24.992606	0.95201048	2.87682430
25	-21.680224	-0.49702987	9.27435540

– représenter graphiquement ces valeurs dans l'espace bi-dimensionel engendré par les deux premier vecteurs propres (Attention aux échelles !)

Il est possible de calculer les corrélations entre les composantes principales et les variable d'origine. Les éléments de la matrice de corrélations sont donnés par :

$$r_{i,l} = \frac{u_{i,l} * \sqrt{\lambda_l}}{\sigma_i}$$

où i et l sont les indices associés aux variables dans l'espace initial et l'espace des projetés respectivement.

La matrice d'importance permet de mesurer le degré d'expression de chaque variable initiale sur chaque facteur principal. Chaque élément i, l de cette matrice est égal à $r_{i,l}^2$. Pour nos valeurs on obtient :

	[,1]	[,2]	[,3]
[1,]	0.573049	0.426409	0.000000

```
[2,] 0.784996 0.179776 0.033124
[3,] 0.760384 0.219024 0.021904
```

avec, en ligne les variables initiales et en colonnes, les facteurs principaux.

Question : interprétation des résultats

Solution : La variable X s'exprime à 57.3% sur le premier axe principal et à 42.6% sur le deuxième. Y s'exprime à 78.5% sur le premier axe principal...