Chapitre 4

Analyse en Composantes Principales (ACP)

Méthode factorielle, ou de type R (en anglais). A pour but de réduire le nombre de variables en perdant le moins d'information possible, c'est à dire en gardant le maximum de la variabilité totale.

Pratiquement, cela revient à projeter les données des variables pour les individus sur un espace de dimension inférieure en maximisant la variabilité totale des nouvelles variables. On impose que l'espace sur lequel on projète soit orthogonal (pour ne pas avoir une vision déformée des données).

4.1 Etape 1 : Changement de repère

Soit X la matrice des données. Pour plus de visibilité, on considère la matrice des données centrées $X-\overline{X}$. Le $i^{\text{ème}}$ vecteur ligne $(X-\overline{X})_i^t$ représente les données de toutes les variables pour le $i^{\text{ème}}$ individu. Pour simplifier les notations, on écrit $\mathbf{x}^t = (X-\overline{X})_i^t$.

• Représentation graphique du $i^{\text{ème}}$ individu

On peut représenter $\mathbf{x}^t$ par un point de $\mathbb{R}^p$. Alors,

- chacun des axes de $\mathbb{R}^p$ représente une des p variables,
- les coordonnées de $\mathbf{x}^t$ sont les données des p variables pour le $i^{\text{ème}}$ individu.

• Nouveau repère

Soient $\mathbf{q}_1, \cdots, \mathbf{q}_p$, p vecteurs de $\mathbb{R}^p$, unitaires et deux à deux orthogonaux. On considère les p droites passant par l'origine, de vecteur directeur $\mathbf{q}_1, \cdots, \mathbf{q}_p$ respectivement. Alors

ces droites définissent un nouveau repère. Chacun des axes représente une nouvelle variable, qui est combinaison linéaire des anciennes variables.

• Changement de repère pour le $i^{\text{ème}}$ individu

On souhaite exprimer les données du $i^{\text{ème}}$ individu dans ce nouveau repère. Autrement dit, on cherche à déterminer les nouvelles coordonnées du $i^{\text{ème}}$ individu. Pour $j = 1, \cdots, p$, la coordonnée sur l'axe $\mathbf{q}_j$ est la coordonnée de la projection orthogonale de $\mathbf{x}$ sur la droite passant par l'origine et de vecteur directeur $\mathbf{q}_j$. Elle est donnée par (voir Chapitre 2) :

$$< \mathbf{x}, \mathbf{q}_j >= \mathbf{x}^t \mathbf{q}_j.$$

Ainsi les coordonnées des données du $i^{\text{ème}}$ individu dans ce nouveau repère sont répertoriées dans le vecteur ligne :

$$(\mathbf{x}^t \mathbf{q}_1 \ \cdots \ \mathbf{x}^t \mathbf{q}_p) = \mathbf{x}^t Q = (X - \overline{X})_i^t Q,$$

où Q est la matrice de taille $p \times p$, dont les colonnes sont les vecteurs $\mathbf{q}_1, \cdots, \mathbf{q}_p$. Cette matrice est *orthonormale*, *i.e.* ses vecteurs colonnes sont unitaires et deux à deux orthogonaux.

• Changement de repère pour tous les individus

On souhaite faire ceci pour les données de tous les individus $(X - \overline{X})_1^t, \cdots, (X - \overline{X})_n^t$. Les coordonnées dans le nouveau repère sont répertoriées dans la matrice :

$$Y = (X - \overline{X})Q. \tag{4.1}$$

En effet, la $i^{\text{ème}}$ ligne de Y est $(X - \bar{X})_i^t Q$, qui représente les coordonnées dans le nouveau repère des données du $i^{\text{ème}}$ individu.

4.2 Etape 2 : Choix du nouveau repère

Le but est de trouver un nouveau repère $\mathbf{q}_1, \cdots, \mathbf{q}_p$, tel que la quantité d'information expliquée par $\mathbf{q}_1$ soit maximale, puis celle expliquée par $\mathbf{q}_2$, etc... On peut ainsi se limiter à ne garder que les 2-3 premiers axes. Afin de réaliser ce programme, il faut d'abord choisir une mesure de la quantité d'information expliquée par un axe, puis déterminer le repère qui optimise ces critères.

4.2.1 Mesure de la quantité d'information

La variance des données centrées $(X - \overline{X})_{(j)}$ de la $j^{\text{ème}}$ variable représente la dispersion des données autour de leur moyenne. Plus la variance est grande, plus les données de cette variable sont dispersées, et plus la quantité d'information apportée est importante.

La quantité d'information contenue dans les données $(X - \overline{X})$ est donc la somme des variances des données de toutes les variables, c'est à dire la *variabilité totale* des données $(X - \overline{X})$, définie à la Section 3.4.1 :

$$\sum_{j=1}^{p} \mathrm{Var}((X - \overline{X})_{(j)}) = \mathrm{Tr}(V(X - \overline{X})) = \mathrm{Tr}(V(X)).$$

La dernière égalité vient du fait que $V(X - \overline{X}) = V(X)$. Etudions maintenant la variabilité totale des données Y, qui sont la projection des données $X - \overline{X}$ dans le nouveau repère défini par la matrice orthonormale Q. Soit $V(Y)$ la matrice de covariance correspondante, alors :

Lemme 3

1. $V(Y) = Q^t V(X) Q$

2. *La variabilité totale des données Y est la même que celle des données $X - \overline{X}$.*

Preuve:

$$\begin{aligned} V(Y) &= \frac{1}{n}(Y - \overline{Y})^t (Y - \overline{Y}) \\ &= \frac{1}{n} Y^t Y, \text{ (car } \overline{Y} \text{ est la matrice nulle)} \\ &= \frac{1}{n}((X - \overline{X})Q)^t (X - \overline{X})Q \text{ (par (4.1))} \\ &= \frac{1}{n} Q^t (X - \overline{X})^t (X - \overline{X}) Q \text{ (propriété de la transposée)} \\ &= Q^t V(X) Q. \end{aligned}$$

Ainsi, la variabilité totale des nouvelles données Y est :

$$\begin{aligned} \mathrm{Tr}(V(Y)) &= \mathrm{Tr}(Q^t V(X) Q) = \mathrm{Tr}(Q^t Q V(X)), \text{ (propriété de la trace)} \\ &= \mathrm{Tr}(V(X)) \text{ (car } Q^t Q = \mathrm{Id}, \text{ étant donné que la matrice } Q \text{ est orthonormale).} \end{aligned}$$

□

4.2.2 Choix du nouveau repère

Etant donné que la variabilité totale des données projetées dans le nouveau repère est la même que celle des données d'origine $X - \overline{X}$, on souhaite déterminer Q de sorte que

la part de la variabilité totale expliquée par les données $Y_{(1)}$ de la nouvelle variable $\mathbf{q}_1$ soit maximale, puis celle expliquée par les données $Y_{(2)}$ de la nouvelle variable $\mathbf{q}_2$, etc... Autrement dit, on souhaite résoudre le problème d'optimisation suivant :

$$\boxed{\begin{array}{c}\text{Trouver une matrice orthonormale } Q \text{ telle que} \\ \text{Var}(Y_{(1)}) \text{ soit maximale, puis Var}(Y_{(2)}), \text{ etc...}\end{array}} \tag{4.2}$$

Avant d'énoncer le Théorème donnant la matrice Q optimale, nous avons besoin de nouvelles notions d'algèbre linéaire.

• Théorème spectral pour les matrices symétriques

Soit A une matrice de taille $p \times p$. Un vecteur $\mathbf{x}$ de $\mathbb{R}^p$ s'appelle un *vecteur propre* de la matrice A, s'il existe un nombre λ tel que :

$$A\mathbf{x} = \lambda\mathbf{x}.$$

Le nombre λ s'appelle la *valeur propre* associée au vecteur propre $\mathbf{x}$.

Une matrice carrée $A = (a_{ij})$ est dite *symétrique*, ssi $a_{ij} = a_{ji}$ pour tout i, j.

Théorème 4 (spectral pour les matrices symétriques) *Si A est une matrice symétrique de taille $p \times p$, alors il existe une base orthonormale de $\mathbb{R}^p$ formée de vecteurs propres de A. De plus, chacune des valeurs propres associée est réelle.*

Autrement dit, il exite une matrice orthonormale Q telle que :

$$Q^t A Q = D$$

et D est la matrice diagonale formée des valeurs propres de A.

• Théorème fondamental de l'ACP

Soit $X - \overline{X}$ la matrice des données centrées, et soit $V(X)$ la matrice de covariance associée (qui est symétrique par définition). On note $\lambda_1 \geq \cdots \geq \lambda_p$ les valeurs propres de la matrice $V(X)$. Soit Q la matrice orthonormale correspondant à la matrice $V(X)$, donnée par le Théorème 4, telle que le premier vecteur corresponde à la plus grande valeur propre, etc...Alors, le théorème fondamental de l'ACP est :

Théorème 5 *La matrice orthonormale qui résoud le problème d'optimisation* (4.2) *est la matrice Q décrite ci-dessus. De plus, on a :*

1. $\text{Var}(Y_{(j)}) = \lambda_j$,

2. $\mathrm{Cov}(Y_{(i)}, Y_{(j)}) = 0$, *quand* $i \neq j$,
3. $\mathrm{Var}(Y_{(1)}) \geq \cdots \geq \mathrm{Var}(Y_{(p)})$,

Les colonnes $\mathbf{q}_1, \cdots, \mathbf{q}_p$ de la matrice Q décrivent les nouvelles variables, appelées les *composantes principales*.

Preuve:
On a :

$$\begin{aligned} V(Y) &= Q^t V(X) Q, \quad \text{(par le Lemme 3)} \\ &= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & & \lambda_{p-1} & 0 \\ 0 & \cdots & 0 & \lambda_p \end{pmatrix}, \quad \text{(par le Théorème 4)} \end{aligned}$$

Ainsi,

$$\begin{aligned} \mathrm{Var}(Y_{(j)}) &= (V(Y))_{jj} = (Q^t V Q)_{jj} = \lambda_j \\ \mathrm{Cov}(Y_{(i)}, Y_{(j)}) &= (V(Y))_{ij} = (Q^t V Q)_{ij} = 0. \end{aligned}$$

Ceci démontre 1 et 2. Le point 3 découle du fait que l'on a ordonné les valeurs propres en ordre décroissant.

Le dernier point non-trivial à vérifier est l'optimalité. C'est à dire que pour toute autre matrice orthonormale choisie, la variance des données de la première variable serait plus petite que λ_1, etc... Même si ce n'est pas très difficile, nous choisissons de ne pas démontrer cette partie ici. □

4.3 Conséquences

Voici deux conséquences importantes du résultat que nous avons établi dans la section précédente.

• Restriction du nombre de variables

Le but de l'ACP est de restreindre le nombre de variables. Nous avons déterminé ci-dessus des nouvelles variables $\mathbf{q}_1, \cdots, \mathbf{q}_p$, les *composantes principales*, qui sont optimales. La part de la variabilité totale expliquée par les données $Y_{(1)}, \cdots, Y_{(k)}$ des k

premières nouvelles variables ($k \leq p$), est :

$$\frac{\mathrm{Var}(Y_{(1)}) + \cdots + \mathrm{Var}(Y_{(k)})}{\mathrm{Var}(Y_{(1)}) + \cdots + \mathrm{Var}(Y_{(p)})} = \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}.$$

Dans la pratique, on calcule cette quantité pour $k = 2$ ou 3. En multipliant par 100, ceci donne le pourcentage de la variabilité totale expliquée par les données des 2 ou 3 premières nouvelles variables. Si ce pourcentage est raisonable, on choisira de se restreindre aux 2 ou 3 premiers axes. La notion de raisonable est discutable. Lors du TP, vous choisirez 30%, ce qui est faible (vous perdez 70% de l'information), il faut donc être vigilant lors de l'analyse des résultats.

• Corrélation entre les données des anciennes et des nouvelles variables

Etant donné que les nouvelles variables sont dans un sens "artificielles", on souhaite comprendre la corrélation entre les données $(X - \overline{X})_{(j)}$ de la $j^{\text{ème}}$ ancienne variable et celle $Y_{(k)}$ de la $k^{\text{ème}}$ nouvelle variable. La matrice de covariance $V(X, Y)$ de $X - \overline{X}$ et Y est donnée par :

$$\begin{aligned} V(X,Y) &= \frac{1}{n}(X - \overline{X})^t(Y - \overline{Y}) \\ &= \frac{1}{n}(X - \overline{X})^t(Y - \overline{Y}), \text{ (car } \overline{Y} \text{ est la matrice nulle)} \\ &= \frac{1}{n}(X - \overline{X})^t(X - \overline{X})Q, \text{ (par définition de la matrice } Y) \\ &= Q(Q^t V(X) Q) \text{ (car } Q^t Q = \mathrm{Id}) \\ &= QD, \text{ (par le Théorème spectral, où } D \text{ est la matrice des valeurs propres).} \end{aligned}$$

Ainsi :

$$\mathrm{Cov}(X_{(j)}, Y_{(k)}) = V(X,Y)_{jk} = q_{jk}\lambda_k.$$

De plus, $\mathrm{Var}(X_{(j)}) = (V(X))_{jj} = v_{jj}$, et $\mathrm{Var}(Y_{(k)}) = \lambda_k$. Ainsi la correlation entre $X_{(j)}$ et $Y_{(k)}$ est donnée par :

$$r(X_{(j)}, Y_{(k)}) = \frac{\lambda_k q_{jk}}{\sqrt{\lambda_k v_{jj}}} = \frac{\sqrt{\lambda_k} q_{jk}}{\sqrt{v_{jj}}}.$$

C'est la quantité des données $(X - \overline{X})_{(j)}$ de la $j^{\text{ème}}$ ancienne variable "expliquée" par les données $Y_{(k)}$ de la $k^{\text{ème}}$ nouvelle variable.

Attention : Le raisonnement ci-dessus n'est valable que si la dépendance entre les données des variables est linéaire (voir la Section 3.4.2 sur les corrélations). En effet, dire qu'une corrélation forte (faible) est équivalente à une dépendance forte (faible) entre les données, n'est vrai que si on sait à priori que la dépendance entre les données est linéaire. Ceci est donc à tester sur les données avant d'effectuer une ACP. Si la dépendance entre les données n'est pas linéaire, on peut effectuer une transformation des données de sorte que ce soit vrai (log, exponentielle, racine, ...).

4.4 En pratique

En pratique, on utilise souvent les données centrées réduites. Ainsi,

1. La matrice des données est la matrice Z.
2. La matrice de covariance est la matrice de corrélation $R(X)$. En effet :

$$\begin{aligned}\mathrm{Cov}(Z_{(i)}, Z_{(j)}) &= \mathrm{Cov}\left(\frac{X_{(i)} - \overline{X_{(i)}}}{\sigma_{(i)}}, \frac{X_{(j)} - \overline{X_{(j)}}}{\sigma_{(j)}}\right), \\ &= \frac{\mathrm{Cov}(X_{(i)} - \overline{X_{(i)}}, X_{(j)} - \overline{X_{(j)}})}{\sigma_{(i)}\sigma_{(j)}}, \\ &= \frac{\mathrm{Cov}(X_{(i)}, X_{(j)})}{\sigma_{(i)}\sigma_{(j)}}, \\ &= r(X_{(i)}, X_{(j)}).\end{aligned}$$

3. La matrice Q est la matrice orthogonale correspondant à la matrice $R(X)$, donnée par le Théorème spectral pour les matrices symétriques.
4. $\lambda_1 \geq \cdots \geq \lambda_p$ sont les valeurs propres de la matrice de corrélation $R(X)$.
5. La corrélation entre $Z_{(j)}$ et $Y_{(k)}$ est :

$$r(Z_{(j)}, Y_{(k)}) = \sqrt{\lambda_k} q_{jk},$$

car les coefficients diagonaux de la matrice de covariance (qui est la matrice de corrélation) sont égaux à 1.