CHAPITRE 4

Analyse en composantes principales

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert – Hiver 2023

L'analyse en composantes principales est une technique permettant de réduire la taille d'un jeu de données en retenant le maximum d'information possible. Utilisée en régression, l'analyse en composantes principales permet d'éliminer les effets de la multicolinéarité. À la fin du chapitre, vous devriez être en mesure de :

- Effectuer une analyse en composantes principales ;
- Estimer le nombre de composantes principales requises ;
- Effectuer une régression sur les composantes principales comme variables explicatives ;
- Transformer les coefficients de régression de l'espace des composantes principales à l'espace des variables explicatives.

4.1 Rappel d'algèbre linéaire

Dans le cadre de ce chapitre, plusieurs concepts d'algèbres linéaires sont essentiels et je tenterai de les rappeler le plus clairement possible dans cette section.

4.1.1 Vecteurs propres et valeurs propres

Soit une matrice réelle carrée A de dimension m. Lorsque l'on multiplie la matrice A par un vecteur colonne $\boldsymbol{x}$ de dimension m, le vecteur change généralement de direction, à

l'exception de quelques directions particulières où la direction du vecteur demeure la même. Intéressons-nous à ces directions particulières. Soit $\boldsymbol{v}$ l'un de ces vecteurs particuliers, on a lors que :

$$A\boldsymbol{v} = \lambda\boldsymbol{v}. \tag{4.1}$$

Le vecteur $\boldsymbol{v}$ conserve sa direction mais sa longueur est modifiée par le facteur λ. Ces vecteurs particuliers sont appelés *vecteurs propres* et les facteurs correspondants sont appelés *valeurs propres*.

Exercice 1

Soit la matrice de projection

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

qui projette tout vecteur du plan (x, y) sur l'axe des x. Trouvez deux vecteurs orthogonaux $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$ pour lesquels l'équation (4.1) est satisfaite.

Lorsque la matrice A possède m colonnes linéairement indépendants, il existe alors m vecteurs propres indépendants $\boldsymbol{v}_1, \dots \boldsymbol{v}_m$ correspondant respectivement aux m valeurs propres $\lambda_1, \dots, \lambda_m$. Dans cette situation, n'importe quel vecteur $\boldsymbol{x}$ dans l'espace des colonnes peut s'écrire comme une combinaison linéaire des vecteurs propres :

$$\boldsymbol{x} = c_1\boldsymbol{v}_1 + \dots c_m\boldsymbol{v}_m.$$

La multiplication $A\boldsymbol{x}$ correspond donc à

$$\begin{aligned} A\boldsymbol{x} &= A\left(c_1\boldsymbol{v}_1 + \dots c_m\boldsymbol{v}_m\right) \\ &= c_1A\boldsymbol{v}_1 + \dots c_mA\boldsymbol{v}_m \\ &= c_1\lambda_1\boldsymbol{v}_1 + \dots c_m\lambda_m\boldsymbol{v}_m; \end{aligned}$$

ce qui est une nouvelle combinaison linéaire des vecteurs propres. Les vecteurs propres font donc partie de l'essence de la matrice A. Dans de nombreuses applications, les vecteurs propres correspondent d'ailleurs à des modes fondamentaux du système modélisé.

4.1.2 Calcul des valeurs propres et des vecteurs propres

L'équation (4.1) implique à la fois les valeurs propres λ inconnues et les vecteurs propres $\boldsymbol{v}$ inconnus. L'astuce pour calculer à la fois les valeurs propres et les valeurs propres à partir d'une seule équation consiste à la réécrire de la façon suivante :

$$(A - \lambda I_m)\boldsymbol{v} = \boldsymbol{0}, \tag{4.2}$$

où I_m correspond à la matrice identité de dimension m et $\boldsymbol{0}$ dénote le vecteur colonne nul de dimension appropriée. Cette dernière équation permet de constater que l'on cherche le

noyau de la matrice $(A - \lambda I_m)$: les valeurs propres λ correspondent donc aux valeurs qui rendent la matrice $(A - \lambda I_m)$ singulière. Autrement dit, on cherche les valeurs de λ telles que

$$|A - \lambda I_m| = 0,$$

où $|A|$ dénote le déterminant de la matrice A. Cette procédure permet d'identifier dans un premier temps les valeurs propres, puis dans un deuxième temps de calculer les vecteurs propres correspondants à l'aide de l'équation (4.2).

Exemple 1

Soit la matrice de projection de l'exercice précédent

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

On trouve les valeurs propres en résolvant l'équation suivante :

$$\begin{vmatrix} 1-\lambda & 0 \\ 0 & -\lambda \end{vmatrix} = 0$$

On obtient ainsi

$$\lambda^2 - \lambda = 0$$

qui donne deux solutions qui sont les valeurs propres : $\lambda_1 = 1$ et $\lambda_2 = 0$.
Le vecteur propre $\boldsymbol{v}_1$ correspondant à la valeur propre $\lambda_1 = 1$ s'obtient en résolvant l'équation (4.2) :

$$\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

On obtient que $-y = 0$. Donc tout vecteur orienté dans la direction $[1, 0]^\top$ est un vecteur propre. Choisissons $\boldsymbol{v}_1 = [1, 0]^\top$.
De la même façon, on peut montrer que le vecteur propre correspondant à $\lambda_2 = 0$ est $\boldsymbol{v}_2 = [0, 1]^\top$.

Pour une matrice réelle quelconque, rien n'indique que ses valeurs propres sont réelles comme l'illustre l'exercice suivant.

Exercice 2

Soit la matrice de rotation

$$Q = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Montrez que les valeurs propres sont $\lambda_1 = i$ et $\lambda_2 = -i$ où $i = \sqrt{-1}$.

4.1.3 Matrices symétriques

Les matrices symétriques possèdent des propriétés remarquables que nous exploiterons au fil du cours. Si la matrice S de dimension m est symétrique, *i.e.* $S^\top = S$, alors on a que

— Les m valeurs propres sont réelles ;
— Les m vecteurs propres peuvent être choisis orthogonaux.

Exemple 2

Soit la matrice symétrique

$$S = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

Cette matrice possède les valeurs propres réelles $\lambda_1 = 5$ et $\lambda_2 = 0$. Les vecteurs propres correspondants sont $\boldsymbol{v}_1 = [1, 2]^\top$ et $\boldsymbol{v}_2 = [-2, 1]^\top$ sont orthogonaux :

$$\boldsymbol{v}_1^\top \cdot \boldsymbol{v}_2 = \begin{bmatrix} 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 1 \end{bmatrix} = 0.$$

Le sens de l'expression *les vecteurs propres peuvent être choisis orthogonaux* est illustré par l'exemple suivant.

Exemple 3

Soit la matrice identité de dimension 2 :

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Cette matrice symétrique possède les valeurs propres réelles $\lambda_1 = 1$ et $\lambda_2 = 1$. Tous les vecteurs du plan sont des vecteurs propres. Il est alors possible d'en choisir deux orthogonaux. Par exemple $\boldsymbol{v}_1 = [1, 0]^\top$ et $\boldsymbol{v}_2 = [0, 1]^\top$.

Les m vecteurs propres orthogonaux d'une matrice symétrique S de dimension m peuvent être normalisés (divisés par leur norme respective) afin de les rendre **orthonormaux**. Les m vecteurs propres orthonormaux $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m$ sont respectivement associés aux valeurs propres $\lambda_1, \ldots, \lambda_m$.

Pour le vecteur propre $\boldsymbol{q}_i$ tel que $1 \leq i \leq m$, on a bien sûr que

$$S\boldsymbol{q}_i = \lambda_i \boldsymbol{q}_i.$$

Le système des m équations peut être écrit à l'aide de la notation matricielle. D'abord, définissons la matrice Q composée de la concaténation horizontale de tous les vecteurs

propres (vecteurs colonnes) :

$$Q = \begin{bmatrix} | & | & & | \\ \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_m \\ | & | & & | \end{bmatrix}.$$

Remarque. *Q est une base orthonormée. On a donc que*

- $Q^\top Q = I$;
- $Q^{-1} = Q^\top$.

Effectuons le produit matricielle SQ :

$$\begin{aligned} SQ &= \begin{bmatrix} | & | & & | \\ S\boldsymbol{q}_1 & S\boldsymbol{q}_2 & \cdots & S\boldsymbol{q}_m \\ | & | & & | \end{bmatrix} \\ &= \begin{bmatrix} | & | & & | \\ \lambda_1\boldsymbol{q}_1 & \lambda_2\boldsymbol{q}_2 & \cdots & \lambda_m\boldsymbol{q}_m \\ | & | & & | \end{bmatrix} \\ &= \begin{bmatrix} | & | & & | \\ \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_m \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} \\ &= Q\Lambda. \end{aligned}$$

De cette dernière égalité, on trouve que

$$S = Q\Lambda Q^{-1} = Q\Lambda Q^\top. \tag{4.3}$$

Ce résultat est central en algèbre linéaire. Il stipule qu'une matrice symétrique peut toujours s'exprimer sous cette forme. Une matrice symétrique est donc toujours **diagonalisable**. Le résultat est résumé par le théorème suivant.

Théorème 1 (Théorème spectral)**.** *Toute matrice symétrique S de dimension m peut s'exprimer sous la forme suivante*

$$S = Q\,\Lambda\,Q^\top,$$

où les colonnes de la matrice Q correspondent aux vecteurs propres orthonormaux de S et où Λ correspond à la matrice diagonale formée des m valeurs propres réelles.

Nous démontrerons en classe certaines parties de ce théorème si le temps le permet.

Remarque. *La diagonalisation et l'inversion d'une matrice correspondent à des concepts différents. Une matrice est diagonalisable si elle possède suffisamment de vecteurs propres indépendants. Elle est inversible si toutes ses valeurs propres sont non nulles. Une matrice peut être diagonalisable sans être inversible, c'était d'ailleurs le cas de la matrice de l'Exemple 2.*

4.1.4 Matrices définies positives

À la section précédente, nous avons vu qu'il existe toujours suffisamment de vecteurs propres indépendants pour diagonaliser une matrice symétrique, une propriété importante dans les applications. Dans cette section, nous verrons que si une matrice symétrique est en plus définie positive, alors elle possède des propriétés additionnelles très intéressantes qu'il sera possible d'exploiter dans les applications. Introduisons d'abord les matrices définies positives.

Définition 1. *La matrice symétrique S de dimension m est dite définie positive si pour n'importe quel vecteur $\boldsymbol{x} \in \mathbb{R}^m$ non nul, on a que*

$$\boldsymbol{x}^\top S \boldsymbol{x} > 0.$$

Exemple 4

La matrice

$$S = \begin{bmatrix} 2 & 4 \\ 4 & 9 \end{bmatrix}$$

est définie positive. Pour tout vecteur $\boldsymbol{x} = \begin{bmatrix} x & y \end{bmatrix}^\top$, on a en effet que

$$\begin{aligned} \boldsymbol{x}^\top S \boldsymbol{x} &= \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 4 & 9 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= 2x^2 + 8xy + 9y^2 = 2\left(x^2 + 4xy\right) + 9y^2 \\ &= 2\left(x^2 + 4xy + 2y^2 - 2y^2\right) + 9y^2 \\ &= 2\left(x^2 + 4xy + 2y^2\right) - 8y^2 + 9y^2 \\ &= 2\left(x + 2y\right)^2 + y^2. \end{aligned}$$

Il suffit de remarquer que la dernière ligne, étant une somme de carrés, ne peut qu'être positive pour tout vecteur non nul.

La proposition suivante permet de relier plusieurs concepts équivalents avec la définition de matrice définie positive.

Proposition 1. *Pour une matrice S définie positive de dimension m, les conditions suivantes sont équivalentes :*

(i) Pour tout $\boldsymbol{x} \in \mathbb{R}^m$ non nul, $\boldsymbol{x}^\top S \boldsymbol{x} > 0$.

(ii) Les m valeurs propres $\lambda_1, \ldots, \lambda_m$ sont positives.

(iii) La matrice S peut s'écrire comme le produit $A^\top A$ d'une matrice A dont les colonnes sont linéairement indépendantes.

La première propriété intéressante découle du deuxième point de la proposition précédente : les valeurs propres d'une matrice définie positive sont toutes positives. Supposons maintenant que λ_1 dénote la plus grande valeur propre, λ_2 à la deuxième plus grande valeur propre, et ainsi de suite. Les valeurs propres sont maintenant dénotées en ordre décroissant :

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$$

Si on reprend la diagonalisation de l'équation (4.3), on trouve que

$$S = Q\Lambda Q^\top = \boldsymbol{q}_1\lambda_1\boldsymbol{q}_1^\top + \ldots + \boldsymbol{q}_m\lambda_m\boldsymbol{q}_m^\top.$$

La propriété remarquable qui en découle est que toute troncature à droite de la somme correspond à une approximation de la matrice S. Par exemple, les $k \in \{1, \ldots, m\}$ premiers termes de la somme dénotés par S_k :

$$S_k = \boldsymbol{q}_1\lambda_1\boldsymbol{q}_1^\top + \ldots + \boldsymbol{q}_k\lambda_k\boldsymbol{q}_k^\top$$

correspondent à une approximation de S. Il est possible d'aller plus loin en stipulant que S_k est la meilleure approximation de rang k de la matrice S. Il s'agit du théorème de Eckart-Young.

Cette façon d'approximer la matrice S joue un rôle très important en pratique. Nous verrons un exemple lors du TD6 pour la reconnaissance faciale. Cette méthode est également utilisée pour la compression vidéo avec la norme de codage H.264.

4.1.5 Décomposition en valeurs singulières

On peut se demander si le théorème spectral est vraiment utile. Il s'applique en effet sur les matrices symétriques mais dans les applications, les matrices sont rarement carrées ! La décomposition en valeurs singulières est une méthode de diagonalisation permettant de remédier à ce problème. Soit la matrice rectangulaire A de dimension $(n \times p)$ et de rang r. Le rang de la matrice A est défini ici comme le nombre de vecteurs colonnes de A linéairement indépendants.

Puisque la matrice A est maintenant rectangulaire, deux ensembles de vecteurs, appelés **vecteurs singuliers**, sont nécessaires pour la *diagonalisation* :

$$A\boldsymbol{v}_1 = \gamma_1\boldsymbol{u}_1, \ \ldots, \ A\boldsymbol{v}_r = \gamma_r\boldsymbol{u}_r;$$

$$A\boldsymbol{v}_{r+1} = 0, \ \ldots, \ A\boldsymbol{v}_p = 0.$$

Par convention, on ordonne les valeurs singulières en ordre décroissant :

$$\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_r.$$

De façon matricielle, le système d'équations s'écrit de la façon suivante :

$$\begin{aligned} AV &= \begin{bmatrix} | & & | \\ A\boldsymbol{v}_1 & \dots & A\boldsymbol{v}_r \\ | & & | \end{bmatrix} \\ &= \begin{bmatrix} | & & | \\ \gamma_1\boldsymbol{u}_1 & \dots & \gamma_r\boldsymbol{u}_r \\ | & & | \end{bmatrix} \\ &= \begin{bmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_r \\ | & & | \end{bmatrix} \begin{bmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_r \end{bmatrix} \\ &= U\Gamma. \end{aligned}$$

Les colonnes de U sont orthogonales et de norme unitaire de même que les colonnes de V. La décomposition en valeurs singulières de la matrice A s'exprime donc ainsi :

$$A = U\Gamma V^\top,$$

avec

$$\dim(U) = (n \times r), \quad \dim(\Gamma) = (r \times r) \quad \text{et} \quad \dim(V) = (p \times r);$$

et où

$$U = \begin{bmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_r \\ | & & | \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_r \end{bmatrix} \quad \text{et} \quad V = \begin{bmatrix} | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_r \\ | & & | \end{bmatrix}.$$

Par conséquent, la matrice A s'écrit de la façon suivante :

$$A = \gamma_1\boldsymbol{u}_1\boldsymbol{v}_1^\top + \ldots + \gamma_r\boldsymbol{u}_r\boldsymbol{v}_r^\top.$$

Deux ensembles de vecteurs U et V sont nécessaires pour diagonaliser une matrice rectangulaire par opposition à un seul pour une matrice carré. L'utilité de la décomposition en valeurs singulières tient du théorème de Eckart-Young stipulant que l'approximation de A par les $k \in \{1, \dots, r\}$ premières valeurs singulières :

$$A_k = \gamma_1\boldsymbol{u}_1\boldsymbol{v}_1^\top + \ldots + \gamma_k\boldsymbol{u}_k\boldsymbol{v}_k^\top;$$

constitue la meilleure approximation de rang k de la matrice A.

Remarque. *Dans le cadre de ce cours, nous utilisons la* ***forme compacte*** *de la décomposition en valeurs singulières, c'est-à-dire que les valeurs singulières nulles sont ignorées.*

On doit maintenant calculer les matrices U, Γ et V de la décomposition en valeurs singulières. Pour ce faire, calculons d'abord le produit $A^\top A$. On sait que cette matrice est symétrique et définie positive (ou au moins semi-définie positive).

$$\begin{aligned} A^\top A &= \left(U\Gamma V^\top\right)^\top \left(U\Gamma V^\top\right) \\ &= \left(V\Gamma^\top U^\top\right)\left(U\Gamma V^\top\right) \\ &= V\Gamma^\top\Gamma V^\top \\ &= V \begin{bmatrix} \gamma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \gamma_r^2 \end{bmatrix} V^\top \end{aligned}$$

On reconnaît la diagonalisation usuelle de $A^\top A$. Par conséquent, $\gamma_1^2, \dots, \gamma_r^2$ correspondent aux valeurs propres de la matrice $A^\top A$. Les vecteurs $\boldsymbol{v}_1, \dots, \boldsymbol{v}_r$ correspondent aux vecteurs propres de $A^\top A$.

Calculons maintenant le produit $AA^\top$. Encore une fois, la matrice résultante est symétrique et au moins semi-définie positive.

$$\begin{aligned} AA^\top &= \left(U\Gamma V^\top\right)\left(U\Gamma V^\top\right)^\top \\ &= \left(U\Gamma V^\top\right)\left(V\Gamma^\top U^\top\right) \\ &= U\Gamma\Gamma^\top U^\top \\ &= U \begin{bmatrix} \gamma_1^2 & & \\ & \ddots & \\ & & \gamma_r^2 \end{bmatrix} U^\top \end{aligned}$$

On reconnaît la diagonalisation usuelle de $AA^\top$. Les $\gamma_1^2, \dots, \gamma_r^2$ correspondent aux valeurs propres de $AA^\top$. Les vecteurs $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ correspondent aux vecteurs propres de $AA^\top$.

Exemple 5

Décomposons en valeurs singulières la matrice suivante :

$$A = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}.$$

Calculons d'abord la matrice $A^\top A$:

$$A^\top A = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}.$$

Les valeurs propres de $A^\top A$ sont $\lambda_1 = 8$ et $\lambda_2 = 2$. Alors on a que

$$\Gamma = \begin{bmatrix} \sqrt{8} & 0 \\ 0 & \sqrt{2} \end{bmatrix}.$$

Les vecteurs propres de $A^\top A$ sont $\boldsymbol{v}_1 = [1, 1]^\top$ et $\boldsymbol{v}_2 = [-1, 1]^\top$. La matrice V correspond alors à

$$V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

On peut reprendre la procédure précédente avec la matrice $AA^\top$:

$$AA^\top = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}.$$

Les valeurs propres sont bien les mêmes. Les vecteurs propres normalisés de cette matrice sont les colonnes de la matrice suivante :

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

4.2 Analyse en composantes principales

L'analyse en composantes principales est une technique permettant de réduire la taille d'un jeu de données en limitant au maximum la perte d'information. L'analyse en composantes principales utilise la décomposition en valeurs singulières pour diagonaliser une matrice de données. Les composantes principales constituent les combinaisons linéaires des variables dont les poids sont donnés par les vecteur singuliers. Dans de multiples applications, l'analyse en composantes principales est utilisée comme méthode d'apprentissage non-supervisé.

Soit la matrice Z de dimension $(n \times p)$ contenant les données centrées des n observations et des p variables explicatives :

$$Z = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{bmatrix};$$

où

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \text{ pour } j \in \{1, \dots, p\}.$$

La matrice des covariances échantillonnales est donnée par

$$S = \frac{1}{n-1} Z^\top Z.$$

Les variances échantillonnales des variables se retouvent sur la diagonale de S et les covariances échantillonnales se retrouvent sur les éléments hors diagonales. La trace de la matrice S constitue la variance totale s^2 contenue dans la matrice des données :

$$s^2 = s_1^2 + \ldots + s_p^2;$$

où

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \text{ pour } j \in \{1, \ldots, p\}.$$

La première composante principale T_1 est définie comme la combinaison linéaire des variables qui explique la plus grande partie de la variance des données. On peut montrer que les poids de cette combinaison linéaire sont donnés par le vecteur propre correspondant à la plus grande valeur propre de la matrice S. La deuxième composante principale T_2 est définie comme la combinaison linéaire orthogonale à T_1 qui maximise la variance résiduelle. On peut montrer que les poids de cette combinaison linéaire sont donnés par le vecteur propre correspondant à la deuxième plus grande valeurs propre de la matrice S. De façon analogue, le j^e composante principale T_j est définie comme la combinaison linéaire orthogonale aux $j-1$ premières composantes principales qui maximise la variance résiduelle. On peut montrer que les poids de cette combinaison linéaire sont donnés par le vecteur propre correspondant à la j^e valeur propre la plus grande de la matrice S.

De façon générale, pour les matrices de grande taille, il est plus efficace et précis d'effectuer une décomposition en valeurs singulières de la matrice rectangulaire $Z = U\Gamma V^\top$ que d'effectuer un diagonalisation de la matrice symétrique S. Par conséquent, les poids des composantes principales sont données par les colonnes de la matrice V de la décomposition en valeurs singulières de Z :

$$\begin{aligned} \boldsymbol{t}_1 &= Z\boldsymbol{v}_1; \\ \boldsymbol{t}_2 &= Z\boldsymbol{v}_2; \\ &\vdots \\ \boldsymbol{t}_p &= Z\boldsymbol{v}_p. \end{aligned}$$

La matrice des données dans l'espace des composantes principales peut être représentée sous la forme suivante :

$$ZV = \begin{bmatrix} | & & | \\ \boldsymbol{t}_1 & \ldots & \boldsymbol{t}_p \\ | & & | \end{bmatrix} \tag{4.4}$$

Les composantes principales $\boldsymbol{t}_i$ et $\boldsymbol{t}_j$ sont orthogonales pour $i \neq j$. Cette propriété est utile dans le cadre de la régression en présence de multicolinéarité comme nous le verrons à la section suivante. Les composantes principales sont également utiles pour la compression de données. On peut montrer que la variance contenue dans les $k \in \{1, \ldots, p\}$ premières composantes principales est égale à :

$$\frac{1}{n-1}\left(\gamma_1^2 + \ldots + \gamma_k^2\right).$$

Si cette quantité est suffisamment élevée par rapport à la variabilité totale, alors les k premières composantes principales peuvent être utilisée pour approximer la matrice Z. Le nombre de colonnes peut donc passer de p à k en limitant la perte d'information. On peut également montrer que si toutes les composantes principales sont utilisées, toute la variance du jeu de données est recouvrée :

$$s^2 = \frac{1}{n-1}\left(\gamma_1^2 + \ldots + \gamma_p^2\right).$$

Il n'y a alors aucune perte d'information.

L'analyse en composantes principales présentée dans cette section est appropriée si l'échelle de mesure de toutes les variables est la même. Dans le cas contraire, qui se présente beaucoup plus souvent en pratique, il convient d'effectuer l'analyse en composantes principales sur la matrice de corrélation R plutôt que sur la matrice de covariance S. La matrice de corrélation s'obtient en standardisant d'abord la matrice des données :

$$Z = \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{s_1} & \frac{x_{12}-\bar{x}_2}{s_2} & \cdots & \frac{x_{1p}-\bar{x}_p}{s_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{s_1} & \frac{x_{n2}-\bar{x}_2}{s_2} & \cdots & \frac{x_{np}-\bar{x}_p}{s_p} \end{bmatrix}.$$

Puis, la matrice des corrélations peut être calculée par la produit matriciel suivant :

$$R = \frac{1}{n-1} Z^\top Z.$$

4.3 Régression sur composantes principales

La régression sur composantes principales s'effectue en utilisant les composantes principales comme variables explicatives en remplacement des variables explicatives originales. Puisque les composantes principales sont orthogonales, la régression sur composantes principales ne souffre pas de multicolinéarité même si la régression originale en souffrait.

Généralement, pour simplifier l'analyse en composantes principales et la régression qui s'ensuit, on évite d'ajouter une ordonnée à l'origine dans le modèle de régression. Pour ce faire, il suffit de centrer (ou standardiser) la matrice des variables explicatives, étape de toute

façon nécessaire pour l'analyse en composantes principales, et de centrer (ou standardiser) le vecteur des observations $\boldsymbol{y}$. Avec les variables explicatives et les observations centrées ou standardisées, l'ordonnée à l'origine n'est plus nécessaire dans le modèle de régression.

La régression sur composantes principales s'effectue en trois étapes :

1. Effectuer une analyse en composantes principales sur la matrice des variables explicatives Z centrées ou centrées et réduites. À cette étapes, vous pouvez choisir de prendre toutes les composantes principales ou seulement un sous-ensemble.
2. Estimer les coefficients de régression linéaire entre la variable d'intérêt Y et la matrice T composantes principales retenues.
3. Transformer les coefficients de régression obtenue à l'aide des vecteurs propres afin de pouvoir effectuer la régression avec les variables explicatives X directement.

À l'étape 1, on obtient la diagonalisation suivante pour la matrice des variables explicatives :

$$Z = U\Gamma V^\top.$$

Si on choisit de ne prendre qu'un sous-ensemble des composantes principales, on sélectionne les valeurs singulières et les vecteurs singuliers appropriés des matrice U, Γ et V. Les composantes principales sont obtenues avec l'équation 4.4 et on définit la matrice de structure de la façon suivante (l'ordonnée à l'origine n'est pas nécessaire puisque les variables ont été standardisées) :

$$T = ZV$$

À l'étape 2, les coefficients de régression sont estimés par la méthode des moindres carrés à l'instar de la régression linéaire étudiée au chapitre 2 :

$$\hat{Y} = T\hat{\boldsymbol{\eta}} \quad \Rightarrow \quad \hat{\boldsymbol{\eta}} = (T^\top T)^{-1} T^\top \boldsymbol{y}.$$

Les coefficients de régressions obtenus $\hat{\eta}$ sont dans l'espace des composantes principales. Si on veut les transformer pour utiliser les variables explicatives originales, il suffit de solutionner l'équation suivante :

$$ZV\hat{\boldsymbol{\eta}} = Z\hat{\boldsymbol{\beta}},$$

ce qui donne

$$\hat{\boldsymbol{\beta}} = V\hat{\boldsymbol{\eta}}.$$

4.4 Exercices

1. Soit la matrice de covariance
$$S = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}.$$
   (a) Quelles sont les valeurs propres de cette matrice et les vecteurs propres correspondants ?

(b) Quelle est la première composante principale.

(c) Quelle pourcentage de la variance la première composante principale explique-t-elle ?

2. Soit S la matrice de covariance du vecteur de variables aléatoires $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$. Soit $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ les valeurs propres de S. Dans laquelle des situations décrites ci-dessous l'analyse en composantes principales est le plus utile pour réduire la dimension du jeu de données :

   (i) Lorsque la variance empirique des valeurs propres $\frac{1}{p}\sum_{j=1}^{p}(\lambda_j - \bar{\lambda})^2$ est grande.

   (ii) Lorsque la variance empirique des valeurs propres $\frac{1}{p}\sum_{j=1}^{p}(\lambda_j - \bar{\lambda})^2$ est petite.

   Justifiez votre réponse.

3. Lors de l'analyse en composantes principales, expliquez dans vos propres mots pourquoi il est utile de standardiser les variables au préalable. Vous pouvez argumenter en indiquant ce qui pourrait se passer si les variables ne sont pas standardisées.

4. Si une valeur singulière est très très grande par rapport aux autres, expliquez ce que vous pouvez conclure. Par exemple, est-il possible de compresser le jeu de données en limitant la perte d'information ? Est-ce que le coefficient de détermination de la régression sera grand ? Y a-t-il présence de multicolinéarité ? Etc.

5. Soit la matrice

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

   a) Calculez les valeurs singulières de la matrice A. Donnez les étapes de votre calcul. Vérifiez qu'elles sont bien réelles et positives.

   b) Calculez la meilleure approximation de rang 1 de la matrice A.

6. Reprenez le jeu de données `bodyfat.csv` du TD2.

   (a) Effectuez une analyse en composante principale des variables explicatives. Quelle pourcentage de la variance sont expliquées par les composantes principales ?

   (b) Effectuez une régression linéaire entre le pourcentage de gras et les les composantes principales.

(c) Transformez vos coefficients de régression obtenus à partir des composantes principales à l'espace des variables explicatives standardisées.

(d) Estimez le pourcentage de matière grasse d'une patiente ayant une épaisseur de pli du triceps de 25 mm (`:Triceps`), d'une circonférence de cuisse de 45 mm (`:Thigh`) et une circonférence de bras de 30 mm (`:Midarm`).