

# Analyse en composantes principales

## Pourquoi changer de dimension ?

Travailler avec un grand nombre de variables peut poser plusieurs problèmes pratiques et théorique :

- Visualisation compliquée : il est impossible de représenter visuellement des données au-delà de 3 dimensions.
- Séparation des classes difficile : dans des problèmes de classification, la séparation entre les groupes peut être cachée dans une combinaison de variables plutôt que dans les variables prises individuellement.
- Coût computationnel élevé : des modèles complexes peuvent devenir difficiles à ajuster lorsque le nombre de variables est grand.
- Corrélations fortes : des variables redondantes rendent les modèles instables ou difficiles à interpréter.

La question naturelle à se poser est donc : peut-on réduire la dimension du jeu de données sans perdre trop d'information ?

Réduire la dimension, ce n'est pas simplement la suppression de variables. En effet, cela risquerait de faire disparaître de l'information pouvant être utile au modèle. Une meilleure approche consiste à construire de nouvelles variables, obtenues comme combinaisons linéaires des variables initiales, qui résument l'information essentielle du jeu de données. Une méthode possible pour cela est l'**Analyse en Composantes Principales** (ACP).

## Analyse en composantes principales

L'ACP est une méthode non-supervisée (sans variables à expliquer) permettant de réduire la dimension d'un jeu de données tout en conservant le plus d'information possible. Cette méthode est utilisée lorsque l'on dispose de  $n$  observations de  $p$  variables numériques continues avec  $p$  trop "grand" pour permettre une modélisation ou une visualisation efficace. La méthode a été introduite par dans Hotelling (1933).

### Applications courantes

1. Visualisation d'un jeu de données multidimensionnelles.
2. Réduction du nombre de variables de  $p$  à  $k \ll p$  pour faciliter la construction de modèle.
3. Compression d'images ou de signaux.
4. Exploration de données biologiques, textuelles ou environnementales.

### Exemples

1. Comparer des équipes de hockey sur la base de six statistiques de fin de saison.
2. Résumer la criminalité entre les provinces canadiennes sur la base des taux de sept types de crimes différents.
3. Compresser des images formées de  $1084 \times 1084$  pixels en quelques variables.
4. Identifier le nombre de variantes d'un type de tumeur à partir du degré d'expression de millions de gènes.

### Formulation mathématique

Soit un vecteur aléatoire composé de  $p$  variables  $X = (X_1, \dots, X_p)^\top$ , centré et ayant comme matrice de variance-covariance  $\Sigma$ . Notons  $\alpha_1 = (\alpha_{11}, \dots, \alpha_{1p})^\top$ , un vecteur de coefficients. On cherche une combinaison linéaire

$$Y_1 = \alpha_1^\top X = \sum_{k=1}^p \alpha_{1k} X_k,$$

telle que la variance de  $Y_1$  soit maximale. L'idée est simple : on désire combiner  $p$  variables en une seule, mais en "capturant" la plus grande partie possible de la variabilité.

Il faut d'abord ajouter une contrainte sur  $\alpha_1$ , puisque sinon on n'aurait qu'à prendre  $\alpha_{1k} = \pm\infty$  et on aurait  $\text{Var}(Y_1) = +\infty$  ce qui est définitivement maximal ! On contraint donc  $\alpha_1$  de sorte qu'il ait une norme égale à 1.

Cela revient à calculer :

$$\max_{\alpha_1^\top \alpha_1 = 1} \text{Var}(Y_1) = \max_{\alpha_1^\top \alpha_1 = 1} \alpha_1^\top \Sigma \alpha_1.$$

Ce problème se résout par les multiplicateurs de Lagrange. Il conduit à l'équation

$$\Sigma \alpha_1 = \lambda_1 \alpha_1,$$

où  $\lambda_1$  est la plus grande valeur propre de  $\Sigma$  et  $\alpha_1$  le vecteur propre associé.

On définit ainsi la première composante principale. On construit les suivantes en imposant des conditions d'orthogonalité (indépendance linéaire) avec les précédentes, ce qui revient à chercher les vecteurs propres suivants :

$$\Sigma \alpha_k = \lambda_k \alpha_k, \quad \text{avec} \quad \lambda_1 \geq \lambda_2 \geq \dots \lambda_p.$$

Les composantes principales sont donc données par

$$Y_k = \alpha_k^\top X, \quad \text{avec} \quad \alpha_k \text{ vecteur propre associé à } \lambda_k.$$

#### Remarque

Si  $\lambda_1 > \dots > \lambda_p$ , alors les composantes principales sont uniques, à un signe près.

#### Preuve

On cherche à calculer

$$\max_{\alpha_1^\top \alpha_1 = 1} \text{Var}(Y_1) = \max_{\alpha_1^\top \alpha_1 = 1} \alpha_1^\top \Sigma \alpha_1.$$

Le problème est donc de maximiser

$$F(\alpha_1) = \alpha_1^\top \Sigma \alpha_1, \quad \text{s.c.} \quad \alpha_1^\top \alpha_1 = 1.$$

On peut récrire ce problème à l'aide des multiplicateurs de Lagrange, soit maximiser

$$F(\alpha_1, \lambda) = \alpha_1^\top \Sigma \alpha_1 - \lambda(\alpha_1^\top \alpha_1 - 1),$$

où  $\lambda$  est un multiplicateur de Lagrange.

Pour solutionner ce problème, on dérive  $F$  par rapport à  $\alpha_1$  et à  $\lambda$ .

$$\begin{cases} \frac{\partial F(\alpha_1, \lambda)}{\partial \alpha_1} = 2\Sigma \alpha_1 - 2\lambda \alpha_1, \\ \frac{\partial F(\alpha_1, \lambda)}{\partial \lambda} = 1 - \alpha_1^\top \alpha_1. \end{cases}$$

Ensuite, on égalise à 0, ce qui donne :

$$\begin{cases} \Sigma \alpha_1 = \lambda \alpha_1 \\ \alpha_1^\top \alpha_1 = 1 \end{cases}.$$

La seconde équation est bien entendue notre contrainte. La première équation est celle qui nous intéresse. En utilisant cette équation et la définition des éléments propres, on déduit que

1.  $\alpha_1$  est un vecteur propre (normé) de  $\Sigma$ ;
2.  $\lambda$  est la valeur propre correspondante.

On a donc que

$$\text{Var}(Y_1) = \alpha_1^\top \Sigma \alpha_1 = \lambda \alpha_1^\top \alpha_1 = \lambda.$$

Puisque l'on veut maximiser cette quantité, on conclut que :

1.  $\lambda = \lambda_1$ , la plus grande valeur propre de  $\Sigma$  ;
2.  $\alpha_1$ , le vecteur propre normé correspondant.

On continue ensuite avec le calcul de la deuxième composante. Ici, on poursuit simultanément deux objectifs :

1. Conserver le maximum de variation présente dans  $X$  ;
2. Simplifier la structure de dépendance pour faciliter l'interprétation et assurer la stabilité numérique d'éventuelles méthodes qui utiliseront les composantes principales obtenues.

Étant donné  $Y_1$ , la deuxième composante principale  $Y_2 = \alpha_2^\top X$  est définie telle que

1.  $\text{Var}(Y_2) = \alpha_2^\top \Sigma \alpha_2$  est maximale ;
2.  $\alpha_2^\top \alpha_2 = 1$  ;
3.  $\text{Cov}(Y_1, Y_2) = 0$ .

Or, on a que

$$\text{Cov}(Y_1, Y_2) = \text{Cov}(\alpha_1^\top X, \alpha_2^\top X) = \alpha_1^\top \Sigma \alpha_2 = \alpha_2^\top \Sigma \alpha_1 = \lambda_1 \alpha_2^\top \alpha_1.$$

On cherche donc le vecteur  $\alpha_2$  qui maximise :

$$F(\alpha_2, \lambda, \kappa) = \alpha_2^\top \Sigma \alpha_2 - \lambda(\alpha_2^\top \alpha_2 - 1) - \kappa(\alpha_2^\top \alpha_1 - 0).$$

De même que pour la première composante, on dérive  $F$  par rapport à  $\alpha_2$ ,  $\lambda$  et  $\kappa$ .

$$\begin{cases} \frac{\partial F(\alpha_2, \lambda, \kappa)}{\partial \alpha_2} = 2\Sigma \alpha_2 - 2\lambda \alpha_2 - \kappa \alpha_1 \\ \frac{\partial F(\alpha_2, \lambda, \kappa)}{\partial \lambda} = 1 - \alpha_2^\top \alpha_2 \\ \frac{\partial F(\alpha_2, \lambda, \kappa)}{\partial \kappa} = -\alpha_2^\top \alpha_1 \end{cases}$$

En égalisant les équations à 0, on retrouve les deux équations des contraintes, ainsi que

$$2\Sigma \alpha_2 - 2\lambda \alpha_2 - \kappa \alpha_1 = 0.$$

En multipliant cette équation à gauche et à droite par  $\alpha_1^\top$ , on trouve

$$2\alpha_1^\top \Sigma \alpha_2 - 2\alpha_1^\top \lambda \alpha_2 - \kappa \alpha_1^\top \alpha_1 = 0.$$

Or  $\alpha_1^\top \Sigma = \lambda_1 \alpha_1^\top$ , et  $\alpha_1^\top \alpha_1 = 1$ , donc

$$2\alpha_1^\top \lambda \alpha_2 - 2\alpha_1^\top \lambda \alpha_2 - \kappa \alpha_1^\top \alpha_1 = 0 \implies -\kappa = 0.$$

En substituant ce résultat, on obtient

$$\Sigma \alpha_2 = \lambda \alpha_2.$$

et donc  $\lambda$  est une autre valeur propre de  $\Sigma$ . Puisque

$$\text{Var}(Y_2) = \alpha_2^\top \Sigma \alpha_2 = \alpha_2^\top \lambda \alpha_2 = \lambda,$$

on a que cette variance est maximale si  $\lambda = \lambda_2$ , la deuxième plus grande valeur propre de  $\Sigma$ , et conséquemment  $\alpha_2$  est le vecteur propre normé correspondant.

On peut généraliser ce résultat en utilisant des maximisations successives. On en conclut que

$$Y_k = \alpha_k^\top X,$$

où  $\alpha_k$  est le vecteur propre normé associé à  $\lambda_k$ , la  $k$ ème plus grande valeur propre de  $\Sigma$ .

Il est possible d'avoir une représentation plus compacte de l'ACP à l'aide de matrices. Soit  $A = (\alpha_1, \dots, \alpha_p) \in \mathbb{R}^{p \times p}$ , la matrice dont les colonnes sont les vecteurs propres. On a  $Y = AX$  et la covariance des composantes principales s'écrit

$$\text{Var}(Y) = A^\top \Sigma A.$$

#### Propriétés de $A$

1.  $A^\top A = AA^\top = I_p$ ;
2.  $A^\top = A^{-1}$ ;
3.  $\Sigma A = A\Lambda$ , où  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ ;
4.  $\text{Var}(Y) = A^\top \Sigma A = \Lambda \implies \text{Cov}(Y_k, Y_l) = 0$  si  $k \neq l$  et  $\text{Var}(Y_k) = \lambda_k \geq \text{Var}(Y_l) = \lambda_l$  si et seulement si  $k \geq l$ .

#### Preuves

1. Par construction, les colonnes de  $A$  sont orthogonales deux à deux et de norme 1 (cf. contraintes sur les vecteurs propres). La matrice  $A$  est donc une matrice orthogonale. Et donc  $A^\top A = AA^\top = I_p$ .
2. De même, comme  $A$  est orthogonal, on a  $A^\top = A^{-1}$ .
3. Le résultat est immédiat en faisant le produit de matrices.
4. On a  $\text{Var}(Y) = A^\top \Sigma A = A^\top A\Lambda = \Lambda$ . Comme  $\Lambda$  est une matrice diagonale,  $\text{Cov}(Y_k, Y_l) = 0$  si  $k \neq l$  (car pas sur la diagonale) et comme  $\lambda_1 \geq \dots \geq \lambda_p$ , on a  $\text{Var}(Y_1) \geq \dots \geq \text{Var}(Y_p)$ .

Une mesure globale de la variation présente dans les données est donnée par la trace de la matrice  $\Sigma$  :

$$\text{tr}(\Sigma) = \text{tr}(\Lambda) = \sum_{i=1}^p \lambda_i = \sum_{k=1}^p \text{Var}(Y_k).$$

La proportion de variation expliquée par la composante principale  $Y_k$  est donc donnée par le ratio entre la valeur propre  $k$  et la somme des valeurs propres :

$$\frac{\lambda_k}{\lambda_1 + \dots + \lambda_p} = \frac{\text{Var}(Y_k)}{\text{tr}(\Sigma)}.$$

De façon similaire, les  $m$  premières composantes expliquent

$$100\% \times \frac{\sum_{k=1}^m \lambda_k}{\sum_{k=1}^p \lambda_k} = 100\% \times \frac{\sum_{k=1}^m \text{Var}(Y_k)}{\sum_{k=1}^p \text{Var}(Y_k)}$$

de la variabilité dans les variables.

## Pratique de l'ACP

### Estimation de la matrice de variance/covariance

En pratique, la matrice  $\Sigma$  est inconnue. On doit donc l'estimer ! Pour cela, on a besoin d'un échantillon aléatoire  $X_1, \dots, X_n$  de réalisation de  $X$ . Un estimateur (sans biais) de  $\Sigma$  est donnée par

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}) (X_i - \widehat{X})^\top,$$

où  $\bar{X}$  est un estimateur de la moyenne de  $X$ .

La matrice  $\widehat{\Sigma}$  est symétrique à coefficients réels et donc elle est diagonalisable. Elle admet ainsi la représentation

$$\widehat{\Sigma} = \widehat{A} \widehat{\Lambda} \widehat{A}^\top,$$

où  $\widehat{A}$  est la matrice des estimateurs des vecteurs propres de  $\Sigma$  et  $\widehat{\Lambda}$  est la matrice des estimateurs des valeurs propres de  $\Sigma$ .

## Quelques remarques

### Sensibilité à l'échelle de $X_1, \dots, X_p$

Puisque l'on cherche une combinaison linéaire de  $X_1, \dots, X_p$  qui maximise la variance, une variable  $X_k$  ayant une grande variance aura un poids démesuré dans les composantes principales. On peut, par exemple, penser à la mesure de distance. En effet, exprimer une distance entre mètres plutôt qu'en kilomètres multiplierait la variance de cette variable par 1 million  $((10^3)^2)$ . Cette variable aurait donc un poids majeur dans toutes les composantes.

Ainsi, à moins que les variables aient des variances similaires, il est recommandé de standardiser les variables avant d'effectuer une ACP. Cela revient à effectuer une ACP sur la matrice des corrélations.

### Première étape dans une analyse prédictive

Il arrive que l'ACP soit effectuée parce que l'on veut prédire la valeur de variable  $Z$  à partir des valeurs de variables  $X_1, \dots, X_p$  mais que  $p$  est simplement trop grand. Dans ce cas, on applique l'ACP pour obtenir les  $k \ll p$  premières composantes principales  $Y_1, \dots, Y_k$  et on les utilise pour prédire  $Z$ .

Puisque les composantes principales retiennent la majeure partie de l'information contenue dans les variables originales, c'est généralement une façon raisonnable de faire.

### Rotation des axes

Puisque la matrice  $A$  est orthogonale, le produit matriciel  $Y = A^\top X$  représente une rotation d'axes. Les nouveaux axes correspondent aux directions orthogonales de variation maximale successives, en supposant que  $\lambda_1 > \dots > \lambda_p$ . Ainsi,  $Y_i = A^\top X_i$  donne les coordonnées de l'observation  $X_i$  dans le nouveau système d'axes. On appelle  $Y_{ik}$  le score de l'observation  $X_i$  sur l'axe principal  $k$  et se calcule comme

$$Y_{ik} = \alpha_k^\top X_i = \sum_{l=1}^p \alpha_{kl} X_{il}.$$

## Références

Hotelling, H. 1933. « Analysis of a Complex of Statistical Variables into Principal Components ». *Journal of Educational Psychology* 24 (6) : 417-41. <https://doi.org/10.1037/h0071325>.