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 élévé : 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 H. Hotelling dans Hotelling (1933).

Applications courantes

1. Visualisation d'un jeu de données multidimensionnelles.
2. Réduction du nombre de variables de p à $p' \ll p$ 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 1084×1084 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 $X = \left(X_1, \ldots, X_p\right)^\top$, centré et ayant comme matrice de variance-covariance Σ. Notons $\alpha_1 = \left(\alpha_{11}, \ldots, \alpha_{1p}\right)^\top$, un vecteur de coefficients. On cherche une combinaison linéaire

$$Y_1 = \alpha_1^\top X = \sum_{i=1}^{p} \alpha_{1i} X_i,$$

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 α_1, puisque sinon on n'aurait qu'à prendre $\alpha_{1i} = \pm\infty$ et on aurait $\mathrm{Var}(Y_1) = +\infty$ ce qui est définitivement maximal ! On contraint donc α_1 de sorte qu'il ait une norme égale à 1.

Cela revient à calculer :

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

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

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

où λ_1 est la plus grande valeur propre de Σ et α_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 :

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

Les composantes principales sont donc données par

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

Preuve

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 λ est un multiplicateur de Lagrange.
Pour solutionner ce problème, on dérive F par rapport à α_1 et à λ.

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

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. α_1 est un vecteur propre (normé) de Σ ;
2. λ 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 Σ ;
2. α_1, le vecteur propre normé correspondant.

— Calcul de la deuxième composante :

On poursuit simultanément deux objectifs :

1. Conserver le maximum de variation présente dans X ;
2. Simplifier la structure de dépendancce 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. $\mathrm{Var}(Y_2) = \alpha_2^\top \Sigma \alpha_2$ est maximale ;
2. $\alpha_2^\top \alpha_2 = 1$;
3. $\mathrm{Cov}(Y_1, Y_2) = 0$.

On a que

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

On cherche donc le vecteur α_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 α_2, λ et κ.

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

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 $\lambda_1^\top \alpha_1 = 1$, 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 résulat, on obtient

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

et donc λ est une autre valeur propre de Σ. Puisque

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

on a que cette variance est maximale si $\lambda = \lambda_2$, la deuxième plus grande valeur propre de Σ, et conséquemment α_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 α_k est le vecteur propre normé associé à λ_k, la ke plus grande valeur propre de Σ.

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 collones sont les vecteurs propres. On a $Y = AX$ et la covariance des composantes principales s'écrit

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

Propriétés de A

1. Les colonnes de la matrice A sont les vecteurs propres de Σ ;
2. $A^\top A = AA^\top = I_p$;
3. $A^\top = A^{-1}$;
4. $\Sigma A = A\Lambda$, où $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$;
5. $\mathrm{Var}(Y) = A^\top \Sigma A = \Lambda \implies \mathrm{Cov}(Y_i, Y_j) = 0$ si $i \neq j$ et $\mathrm{Var}(Y_i) = \lambda_i \geq \mathrm{Var}(Y_j) = \lambda_j$ si et seulement si $i \leq j$.

Preuves

1.
2.
3.
4.

Une mesure globale de la variation presente dans les données est donnée par la trace de la matrice Σ :

$$\mathrm{tr}(\Sigma) = \mathrm{tr}(\Lambda) = \sum_{i=1}^{p} \lambda_i.$$

La proportion de variation expliquée par la composante principale Y_i est

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}.$$

Similairement, les m première composantes expliquent

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100\%.$$

de la variabilité dans les variables.

Pratique de l'ACP

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.