ACP

Changer de dimension - pourquoi faire ?

Il y a plusieurs raisons de vouloir changer le dimension des données. Il est possible que la dimension des données soit trop importante pour avoir une visualisation intéressante, que la dimension actuelle ne permette pas une bonne séparation des classes dans les données, etc.

Analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode 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 a n observations de p variables continues avec p trop "grand" pour nos besoins.

Pourquoi l'ACP est-elle utilisée ?

- Visualisation d'un jeu de données ;
- Réduction du nombre de variables de p à $p' \ll p$ pour faciliter la construction de modèle. Exemples : analyse de texte, analyse de données génétique ;
- Effectuer une rotation d'axes pour simplifier la structure de corrélation ;
- Compression de données.

La méthode a été introduite par H. Hotelling dans (Hotelling 1933).

Visualiser, comprendre, modéliser, classifier des données sont toutes des tâches beacoup simples à accomplir si le nomnre de variables dans un jeu de données est faible. Si un jeu de données comprend un grand nombre de variables, une première question que l'on peut se poser est : "Est-il possible de réduire la dimension du problème sans trop perdre d'information ?". En omettant tout simplement des variables, on risque de perdre beaucoup d'information utile. Une meilleure solution consiste à trouver des combinaisons linéaires des variables en vue de conserver le maximum d'information sur le jeu de données.

Example

1. Comparer des équipes de hockey sur la base de six statistiques de fin de saison.
2. Comparer la criminalité entre états sur la base des taux de sept types de crimes différents.
3. Compresser des images formées de 1084×1084 pixels.
4. Identifier le nombre de variantes d'un type de tumeur à partir du degré d'expression de millions de gènes.

Les maths

Soit un vecteur aléatoire composé de p variables $X = \big(X_1, \dots, X_p\big)$ ayant comme matrice de variance Σ. On aimerait définir une première composante principale,

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

de sorte 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 verra qu'il est pratique de contraindre α_1 de sorte qu'il ait une norme égale à 1.

— Calcul de la première composante :

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

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 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

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

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.

— Notation matricielle :

Pour définir simultanément et de facon plus compacte les composantes principales, on pose

$$Y = AX,$$

où

$$A = (\alpha_1, \dots, \alpha_p) = \begin{pmatrix} \alpha_{1,1} & \alpha_{2,1} & \cdots & \alpha_{p,1} \\ \alpha_{1,1} & \alpha_{2,2} & \cdots & \alpha_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1,p} & \alpha_{2,p} & \cdots & \alpha_{p,p} \end{pmatrix}.$$

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 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$;
5. $\text{Var}(Y) = A^\top \Sigma A = \Lambda \implies \text{Cov}(Y_i, Y_j) = 0$ si $i \neq j$ et $\text{Var}(Y_i) = \lambda_i \geq \text{Var}(Y_j) = \lambda_j$ si et seulement si $i \leq j$.

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

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

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

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.