Analyse en composante principales

Mathematique

Véronique Tremblay

L'idée de l'ACP

On cherche une combinaison linéaire des variables qui maximise la variance.

$$Y = aX_1 + bX_2 + cX_3 + ...$$

p: le nombre de variables

n: le nombre d'observations

X: une matrice $n \times p$ qui contient les données

Σ: la matrice de variance-covariance de X

Y_k: la k^e composance principale

α_k: les coefficients permettant d'obtenir la k^e composante

$$\begin{aligned} Y_k &= \alpha_{k,1} X_1 + \alpha_{k,2} X_2 + \ldots \\ &= \alpha_k^\top X \end{aligned}$$

Trouver la première composante principale

On veut trouver une première composante principale qui maximise

$$\mathbb{V}\mathrm{ar}(Y_1) \quad = \quad \mathbb{V}\mathrm{ar}(\alpha_k^\top X)$$

Trouver la première composante principale

- Le problème est donc de maximiser

$$F(\alpha_1) = \alpha_1^\top \Sigma \alpha_1$$

- Sous la contrainte que $\alpha_1^\top \alpha_1 = 1$.

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

Trouver la première composante principale

On dérive, on pose la dérivée à zéro et on obtient:

$$2\Sigma\alpha_1 - 2\lambda\alpha_1 = 0$$

Qui est vrai seulement si:

$$\Sigma\alpha_1 = \lambda\alpha_1$$

Soit M une matrice carrée. Alors on dit que λ est une *valeur propre* de M s'il existe un vecteur $\mathbf{x} \neq \mathbf{0}$ tel que

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

Le vecteur $\mathbf{x}$ est appelé *vecteur propre* correspondant à la valeur propre λ et l'ensemble des nombres réels λ satisfaisant l'équation précédente est appelé *spectre* de la matrice M.

Trouver la première composante principale

$$\Sigma\alpha_1 = \lambda\alpha_1$$

On en déduit que:

1. α_1 est un vecteur propre (normé) de Σ;

2. λ est la valeur propre correspondante.

Trouver la première composante principale

Calculons la variance de Y_1.

$$\begin{aligned}\mathbb{V}\mathrm{ar}(Y_1) &= \alpha_1^\top \Sigma \alpha_1 \\ &= \lambda \alpha_1^\top \alpha_1 \\ &= \lambda.\end{aligned}$$

Trouver la première composante principale

On conclut que

1. $\lambda = \lambda_1$ est **la plus grande valeur propre de** Σ;

2. α_1 est le vecteur propre normé correspondant.

Trouver la deuxième composante principale

On poursuit simultanément deux objectifs:

1. conserver le **maximum de variation** préesente dans $\mathbf{X}$

2. simplifier la structure de dépendance pour - faciliter l'interprétation - assurer la stabilité numériques des méthodes qui utiliseront les composantes principales obtenues

Trouver la deuxième composante principale

Étant donné Y_1, la deuxième composante principale

$$Y_2 = \alpha_2^\top X$$

est définie telle que:

1. $\mathbb{V}\text{ar}(Y_2) = \alpha_2^\top \Sigma \alpha_2$ est maximale
2. $\alpha_2^\top \alpha_2 = 1$
3. $cov(Y_1, Y_2) = 0$

Trouver toutes les composantes principales

Quelques calculs permettent de conclure que:

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

où α_k est le vecteur propre normé associé à λ_k, la k^{e} plus grande valeur propre de Σ.

Écriture matricielle

$$\mathbf{Y} = X\mathbf{A}$$

$$\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{p1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{1p} & \alpha_{2p} & \cdots & \alpha_{pp} \end{pmatrix}$$

Rappel d'algèbre

- Si α est un vecteur propre de Σ correspondant à une valeur propre λ, alors $c\alpha$ sera également un vecteur propre de Σ correspondant à λ.
- Si Σ est symétrique et α_1 et α_2 sont des vecteurs propres correspondant à des valeurs propres différentes de Σ, alors α_1 et α_2 sont orthogonaux, i.e., $\alpha_1^\top \alpha_2 = 0$.
- Si Σ est symétrique, toutes ses valeurs propres sont réelles.
- Si Σ est définie non-négative [définie positive] alors toutes ses valeurs propres sont non-négatives [positives].