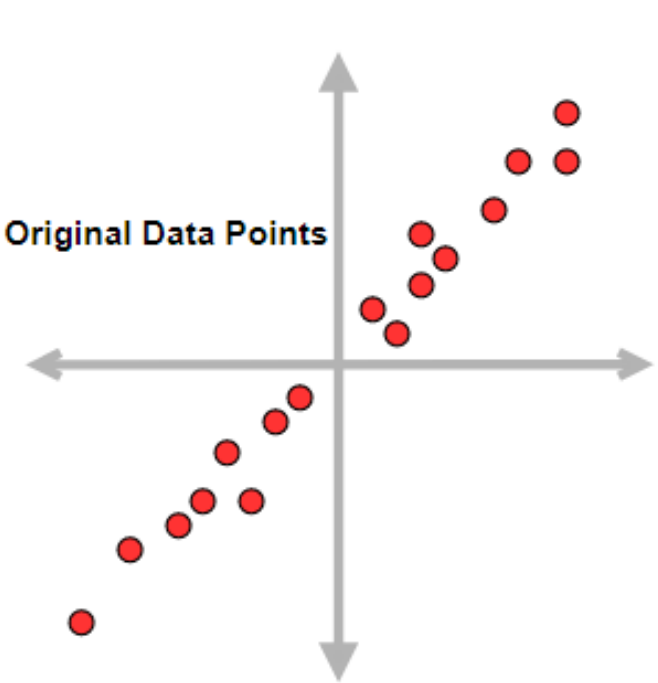
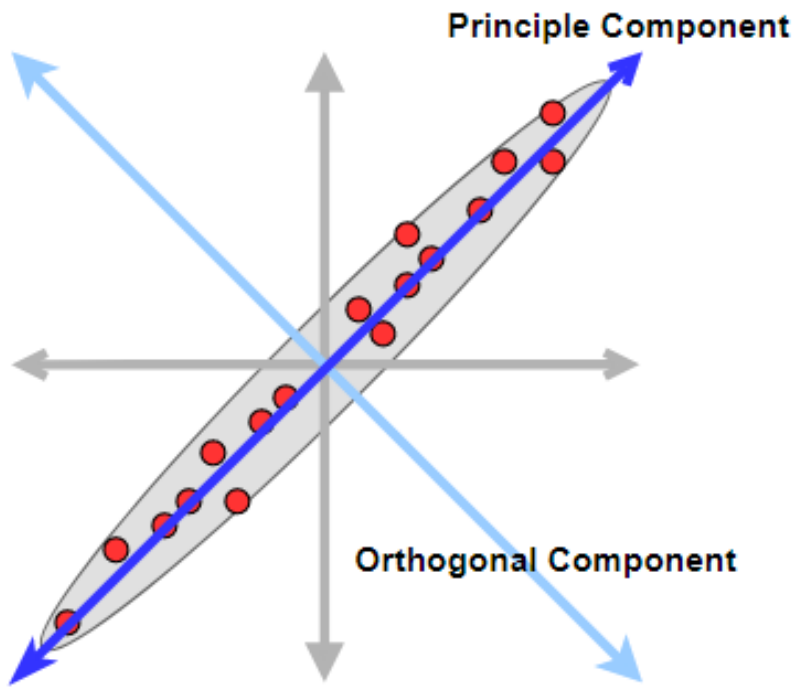
Cours - Informatique - Analyse en Composantes Principales

Gustave Cortal

Exemple introductif





Principe de l'ACP (1)

Soit p variables explicatives **quantitatives** : $X = (X_1, ..., X_p)$. L'Analyse en Composantes Principales (ACP) consiste à trouver q nouvelles variables non corrélées qui sont des combinaisons linéaires des p variables initiales avec $q < p$.

Soit les q nouvelles variables :

$$Z_1 = f_1(X_1, ..., X_p) \qquad ... \qquad Z_q = f_q(X_1, ..., X_p) \tag{1}$$

Les fonctions f sont linéaires et vont de $\mathbb{R}^p$ à $\mathbb{R}$.

Ces nouvelles variables sont appelées *composantes principales*.

Principe de l'ACP (2)

Si les fonctions f sont linéaires, alors on a :

$$Z_j = u_j^T X \tag{2}$$

où $u_j \in \mathbb{R}^p$ et $||u_j|| = 1$

Statistiquement, on recherche des axes orthogonaux u_j expliquant au mieux la variance des données.

Géométriquement, Z_j représentent les coordonnées de la projection de X sur l'axe dirigé par u_j.

Utilisation de l'ACP

L'ACP est utilisé pour :

- Décorréler les variables explicatives
- Représenter des données de haute dimension dans une plus petite dimension
- Visualiser les données dans deux ou trois dimensions seulement
- Réduire le nombre de paramètres des modèles d'apprentissage statistiques, en réduisant la dimension des données d'entrée
- Découvrir des dimensions latentes interprétables

Recherche des composantes principales (1)

Objectif : trouver des axes orthogonaux u_j qui maximisent la variance des données projetées.

Soit $X = (X_1, ..., X_p)$ un vecteur aléatoire ayant comme matrice de covariance V.

La première composante est $Z_1 = u_1^T X$, choisie de telle sorte que :

$$Var(Z_1) = max_{u_1} Var(u_1^T X) = max_{u_1} u_1^T V u_1 \tag{3}$$

soumis à la contrainte $u_1^T u_1 = 1$

Il faut résoudre un problème d'optimisation sous contrainte en utilisant le lagrangien :

$$L(u_1, \lambda) = u_1^T V u_1 - \lambda(u_1^T V u_1 - 1) \tag{4}$$

Recherche des composantes principales (2)

Pour obtenir l'axe u_1 qui maximise la variance, on dérive le lagrangien par rapport à u_1 :

$$\frac{\partial L}{\partial u_1} = 2Vu_1 - 2\lambda u_1 = 0 \Leftrightarrow Vu_1 = \lambda u_1 \tag{5}$$

L'axe u_1 est le vecteur propre de V avec une norme 1 et une valeur propre λ_1

On a $Var(Z_1) = Var(u_1^T X) = u_1^T V u_1 = \lambda_1$

On répète les mêmes opérations pour trouver le deuxième axe u_2, avec la contrainte supplémentaire qu'u_1 et u_2 soient orthogonaux : $u_1^T u_2 = 0$

Recherche des composantes principales (3)

En continuant le même raisonnement, on se retrouve avec p composantes principales $Z = (Z_1, ..., Z_p)$ ayant comme valeurs propres respectives $\lambda_1 > ... > \lambda_p$.

Pour résoudre notre problème, **il suffit de diagonaliser la matrice V**.

Rappel : V est matrice $p \times p$ symétrique définie positive, donc V admet p vecteurs propres orthogonaux ayant des valeurs propres réelles et positives

Variance expliquée par les composantes principales

La somme des valeurs propres est égale à la variance totale des variables initiales :

$$\sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var(X_j) \tag{6}$$

La proportion de la variance expliquée par les q premières composantes est :

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{7}$$

En pratique, on cherche les q premières composantes avec une variance expliquée proche de 1.

Comment obtenir la matrice V ?

En pratique, on calcule la matrice de variance empirique

$$V = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \tag{8}$$

Souvent, on centre les données par colonne et on divise par l'écart type, ce qui nous donne une matrice de corrélation. Les valeurs sont les coefficients de corrélation entre les variables, les valeurs sur la diagonale valent 1.

Aller plus loin

- Analyse des correspondances multiples (ACM) pour des individus décrits par des variables qualitatives
- Analyse factorielle des correspondances (AFC) pour des tableaux de contingences (deux variables qualitatives)
- Analyse factorielle de données mixtes (AFDM) pour des individus décrits par des variables qualitatives et quantitatives