

Analyse en composantes principales (ACP) : application à la reconnaissance de visages

Pierre Gurdjos

March 13, 2008

1. Description multidimensionnelle de données numériques

Les données numériques

La donnée élémentaire

individu i = collection de p variables $\{x_1 \dots x_p \in \mathbb{R}\}$

Le « tableau des données »

observations de n individus \rightarrow matrice $X \in \mathcal{M}_{\mathbb{R}}(n, p)$

$$X = \begin{pmatrix} \vdots & & & & \\ \vdots & & & & \\ \cdots & \cdots & x_{ij} & \cdots & \cdots \\ \vdots & & & & \\ \vdots & & & & \end{pmatrix}$$

- ◆ ligne $\mathbf{a}_i^\top \in \mathbb{R}^p \rightarrow$ valeurs des p variables de l'individu i
- ◆ colonne $\mathbf{x}_j \in \mathbb{R}^n \rightarrow$ valeurs de la variable j prises pour les n individus

Espace des individus. Espace des variables

Espace des individus (X selon ses lignes)

- Individu : vecteur \mathbf{a}_i de l'e.v. $\mathcal{E} = \mathbb{R}^p$ dim appelé *espace des individus*.
- Dans \mathcal{E} , on étudie les distances entre individus.

Espace des variables (X selon ses colonnes)

- Variable = vecteur \mathbf{x}_j de l'e.v. $\mathcal{E}^* = \mathbb{R}^n$ appelé *espace des variables*.
- Dans \mathcal{E}^* , on étudie les angles entre variables.

Caractéristique d'ordre 1 : tendance centrale

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

■ Tendance centrale des variables

◆ Centre de gravité ou « point moyen »

→ vecteur $\mathbf{g} \in \mathbb{R}^p$ des moyennes arithmétiques des variables :

$$\mathbf{g} = \left(\dots, \bar{g}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \dots \right)^\top$$

Exemple 3

for powerdot - 5 / 15

Caractéristique d'ordre 2 : dispersion

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

■ Dispersion des variables

◆ vecteur des écarts-type :

$$\sigma = \left(\dots, \sigma_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{g}_j)^2}, \dots \right)^\top \in \mathbb{R}^p$$

◆ matrice diagonale (p, p) de réduction :

$$D_{1/\sigma} = (\text{diag } \sigma)^{-1} = \begin{pmatrix} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_j} & \\ & & & \ddots \\ & & & & \frac{1}{\sigma_p} \end{pmatrix}$$

Exemple 3

for powerdot - 6 / 15

Normalisation des données

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

■ Centrage du tableau de données :

$$X_C = \begin{pmatrix} & x_{1j} - \bar{g}_j \\ & \vdots \\ \dots & x_{ij} - \bar{g}_j & \dots & \dots \\ & \vdots \\ & x_{nj} - \bar{g}_j \end{pmatrix}$$

$$= X - \mathbf{1}_n \mathbf{g}^\top$$

Exemple 3

for powerdot - 7 / 15

Réduction du tableau de données

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

■ Tableau de données centrées et réduites :

$$X_{0,1} = \begin{pmatrix} \frac{x_{1j} - \bar{g}_j}{\sigma_j} \\ \vdots \\ \dots & \frac{x_{ij} - \bar{g}_j}{\sigma_j} & \dots & \dots \\ \vdots \\ \frac{x_{nj} - \bar{g}_j}{\sigma_j} \end{pmatrix}$$

$$= X_C D_{1/\sigma}$$

Exemple 3

for powerdot - 8 / 15

Matrice de covariance/corrélation

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

■ Matrice de covariance :

$$V = \begin{pmatrix} \cdots & \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{g}_i)(x_{kj} - \bar{g}_j) & \cdots \\ \vdots & & \vdots \\ \cdots & \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{g}_i)(x_{kj} - \bar{g}_j) & \cdots \end{pmatrix}$$
$$= \frac{1}{n-1} X_C^T X_C$$

à noter que

$$V_{ii} = \sigma_i^2$$

Exemple 3

for powerdot - 9 / 15

Matrice de covariance/corrélation

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

■ Matrice de corrélation :

$$R = \begin{pmatrix} \cdots & \frac{1}{n-1} \sum_{k=1}^n \frac{(x_{ki} - \bar{g}_i)(x_{kj} - \bar{g}_j)}{\sigma_i \sigma_j} & \cdots \\ \vdots & & \vdots \\ \cdots & \frac{1}{n-1} \sum_{k=1}^n \frac{(x_{ki} - \bar{g}_i)(x_{kj} - \bar{g}_j)}{\sigma_i \sigma_j} & \cdots \end{pmatrix}$$
$$= \frac{1}{n-1} X_{0,1}^T X_{0,1}$$

Corrélation : mesure de dépendance linéaire entre deux variables i et j

si $R_{ij} = 1$ alors $\mathbf{x}_j = a\mathbf{x}_i + b$

Exemple 3

for powerdot - 10 / 15

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

1. Analyse en composantes principales

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

Problème :

- Soit un ensemble de n individus décrits par p variables :
→ tableau de données *centré* $X \in \mathcal{M}(\mathbb{R}, n, p)$
- On cherche un sous-espace \mathcal{F}_k de \mathcal{E} , en général de dimension $k \ll p$, dans lequel on va représenter/approcher « au mieux » le nuage de points au sens d'un certain critère.

Exemple 3

for powerdot - 12 / 15

Analyse en composantes principales

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

Formulation du problème P_k : On cherche k vecteurs de base $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\} \subseteq \mathcal{E}$ engendrant \mathcal{F}_k i.e., on cherche

$$\mathbf{U} = (\mathbf{u}_1 \mid \mathbf{u}_2 \mid \dots \mid \mathbf{u}_k) \in \mathcal{M}(\mathbb{R}, p, k)$$

et un nouveau tableau de données représentant n individus et k variables

$$\mathbf{Y} = (\mathbf{y}_1 \mid \mathbf{y}_2 \mid \dots \mid \mathbf{y}_k) \in \mathcal{M}(\mathbb{R}, n, k)$$

tels que

$$\mathbf{X}_C = \mathbf{Y}\mathbf{U}^\top \quad (1)$$

avec, comme critère, que les n^{elles} variables $\mathbf{y}_j = (y_{1j}, y_{2j}, \dots, y_{nj})^\top$ soient :

- 2 à 2 non corrélées,
- de variance maximale.

Example 3

for powerdot – 13 / 15

Analyse en composantes principales

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

Théorème : Le sous-espace \mathcal{F}_k , de dimension k , solution du problème P_k , est engendré par les k vecteurs propres de la matrice de covariance $\mathbf{V} = \mathbf{X}_C^\top \mathbf{X}_C$ associés aux k plus grands vecteurs propres.

Soient $\{\lambda_j, \mathbf{u}_j\}$ les couples valeurs/vecteurs propres de $\mathbf{X}_C^\top \mathbf{X}_C$ avec $\|\mathbf{u}_j\| = 1$ et $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. Par définition,

$$\mathbf{X}_C^\top \mathbf{X}_C \mathbf{u}_j = \lambda_j \mathbf{u}_j \Leftrightarrow \mathbf{X}_C^\top \mathbf{X}_C \mathbf{U} = \mathbf{U} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix}$$

Propriété : $\mathbf{U} = (\mathbf{u}_1 \mid \mathbf{u}_2 \mid \dots \mid \mathbf{u}_k) \in \mathcal{M}(\mathbb{R}, p, k)$ alors $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$;
On déduit donc de (1) le nouveau tableau de données \mathbf{Y} :

$$\mathbf{X}_C = \mathbf{Y}\mathbf{U}^\top \Leftrightarrow \mathbf{X}_C \mathbf{U} = \mathbf{Y}.$$

Example 3

for powerdot – 14 / 15

Analyse en composantes principales

1. Description multidimensionnelle de données numériques 1. Analyse en composantes principales

- On appelle **composantes principales** les (valeurs des) nouvelles variables correspondant aux colonnes de

$$\mathbf{Y} = \mathbf{X}_C \mathbf{U}$$

$$= (\mathbf{y}_1 \mid \mathbf{y}_2 \mid \dots \mid \mathbf{y}_k) = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix} \in \mathcal{M}(\mathbb{R}, n, k).$$

- Les composantes principales pour l'individu i sont obtenues par projections orthogonales de \mathbf{a}_i sur $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$, respectivement :

$$y_{ij} = \langle \mathbf{u}_j \mid \mathbf{a}_i \rangle.$$

Example 3

for powerdot – 15 / 15