Université de Poitiers
Département de Mathématiques

Statistique descriptive, 1er semestre, année univ. 2009-2010 **Fiche 7**

Analyse factorielle : analyse en composantes principales

Exercice 1

Projection d'un nuage de points, inertie

Dans $\mathbb{R}^p$ euclidien, on se donne un nuage $\mathcal{N}$ de n points $I_1; \ldots; I_n$ avec les poids respectifs $p_1; \ldots; p_n$. (on suppose les p_i positifs et de somme 1). Soit H un sous-espace vectoriel de $\mathbb{R}^p$ de dimension d contenant G le centre de gravité du nuage $\mathcal{N} = (I_i)_{1 \leq i \leq n}$. Pour $1 \leq i \leq n$ on note H_i le projeté orthogonal de I_i sur H et $\mathcal{N}_H$ le nuage projeté. Comme H contient G, on a $G_H = G$ où G_H désigne le centre de gravité du nuage $\mathcal{N}_H = (H_i)_{1 \leq i \leq n}$. On rappelle que l'inertie du nuage de points $\mathcal{N}$ est définie par

$$I(\mathcal{N}) = \sum_{i=1}^{n} p_i d_2(G, I_i)^2$$

et que l'inertie du nuage projeté sur H est

$$I_H(\mathcal{N}) = I(\mathcal{N}_H) = \sum_{i=1}^{n} p_i d_2(G_{\mathcal{N}}, H_i)^2$$

1) Montrer les relations

$$\sum_{i=1}^{n} \sum_{k=1}^{n} p_i p_k d_2(I_i, I_k)^2 = 2I(\mathcal{N})$$

et

$$\sum_{i=1}^{n} \sum_{k=1}^{n} p_i p_k d_2(H_i, H_k)^2 = 2I(\mathcal{N}_H)$$

2) On cherche un espace H de dimension d et contenant G qui maximise

$$\sum_{i=1}^{n} \sum_{k=1}^{n} p_i p_k d_2(H_i, H_k)^2 = 2I(\mathcal{N}_H).$$

Cela revient à dire que H restitue le plus fidèlement les distances. Montrer que maximiser l'inertie $I(\mathcal{N}_H)$ revient à minimiser l'erreur quadratique moyenne

$$I(\mathcal{N}) = \sum_{i=1}^{n} p_i d_2(I_i, H_i)^2.$$

3) Soit H et $\tilde{H}$ deux sous espaces orthogonaux de $\mathbb{R}^p$ contenant G. Montrer que

$$I_{H \oplus \tilde{H}}(\mathcal{N}) = I_H(\mathcal{N}) + I_{\tilde{H}}(\mathcal{N}).$$

4) La qualité de représentation du point I_i par sa projection H_i sur l'espace H est mesurée par

$$QLT_H(I_i) = \frac{GH_i^2}{GI_i^2} \in [0, 1].$$

Montrer que $QLT_H(I_i) = 1$ si et seulement si $I_i \in H$, puis que

$$QLT_{H\oplus\tilde{H}}(I_i) = QLT_H(I_i) + QLT_{\tilde{H}}(I_i)$$

où H et $\tilde{H}$ désignent deux sous-espaces orthogonaux contenant G. La qualité de représentation du nuage $\mathcal{N}$ est

$$QLT_H(\mathcal{N}) = \sum_{i=1}^{n} p_i QLT_H(I_i) \in [0,1].$$

Montrer que $QLT_H(\mathcal{N}) = 1$ si et seulement si N est porté par H, puis que

$$QLT_{H\oplus\tilde{H}}(\mathcal{N}) = QLT_H(\mathcal{N}) + QLT_{\tilde{H}}(\mathcal{N}).$$

Exercice 2

Exemple de calcul simple en dimension 2

On dispose du tableau x suivant représentant 2 variables mesurées sur 10 individus.

$$\begin{pmatrix} 0.5 & -0.1 & -0.5 & -0.3 & 0 & 1.6 & 2 & 2.4 & 0.5 & 2.7 \\ 0 & 1.2 & 0.5 & 0.1 & 2.5 & -0.7 & 2 & 1.2 & 3.5 & -0.9 \end{pmatrix}'$$

On calcule $\overline{x}_{.,1} = 0.88$, $\overline{x}_{.,2}0.94$, $s_1 = 1.128$, $s_2 = 1.346$, $\rho_{x_{.,1},x_{.,2}} = -0.237$. Le tableau Z des données centrées réduites est donné par

$$\begin{pmatrix} -0.34 & -0.87 & -1.22 & -1.05 & -0.78 & 0.64 & 0.99 & 1.35 & -0.34 & 1.61 \\ -0.7 & 0.19 & -0.33 & -0.62 & 1.16 & -1.22 & 0.79 & 0.19 & 1.9 & -1.37 \end{pmatrix}'$$

1) Tracer le nuage des individus correspondant aux données centrées-réduites.

2) Donner la matrice des corrélations associée aux deux variables.

3) Déterminer les axes factoriels du nuage des individus. Les représenter sur le nuage des individus.

4) Déterminer les composantes principales.

5) Pour chaque axe factoriel, déterminer la qualité de représentation et la contribution à la construction de l'axe de chaque individu.

6) Construire le cercle des corrélations.