Université de Poitiers
Département de Mathématiques

Statistique descriptive, 1er semestre, année univ. 2009-2010 **Fiche 6**

Analyse factorielle : analyse en composantes principales

Exercice 1

Tableaux de données, résumés numériques et espaces associés

Soit $X = (x_{i,j})_{1\leq i\leq n;\ 1\leq j\leq p}$ un tableau numérique. On rappelle l'interprétation classique de ce tableau : p variables numériques sont mesurées sur n individus, la valeur $x_{i,j}$ représente la valeur de la j-ème variable pour le i-ème individus. Le i-ème vecteur ligne $x_{i,.} := (x_{i,j})_{1\leq j\leq p} \in \mathbb{R}^p$ est associé à l'individu i et $\mathbb{R}^p$ est l'espace des individus. Le j-ème vecteur colonne $x_{.,j} = (x_{i,j})_{1\leq i\leq j,\ n} \in \mathbb{R}^n$ est associé à la variable j et $\mathbb{R}^n$ est l'espace des variables. Dans le cas où les différents individus ont des poids p_i (avec les p_i positifs et de somme 1), on note $D = \mathrm{Diag}(p_1; \ldots; p_n)$. Lorsque rien n'est précisé, on prend $p_i = \frac{1}{n}$ et $D = \frac{1}{n} I_n$ où I_n désigne la matrice identité dans $\mathbb{R}^n$. Soit $\mathbb{1}_n$ le vecteur colonne de $\mathbb{R}^n$ dont toutes les entrées sont égales à 1.

1) Vérifier que $g = {}^tXD\mathbb{1}_n \in \mathbb{R}^p$ est l'individu moyen associé à la variable j de coordonnées $(\overline{x_{.,j}})_{1\leq j\leq p}$, puis que le tableau $Y = X - \mathbb{1}_n {}^tg = (I_n - \mathbb{1}_n {}^t\mathbb{1}_n D)X$ est le tableau des données centrées associé à X vérifant $y_{i,j} = x_{i,j} - \overline{x_{.,j}}$.

2) Vérifier que la matrice de variance-covariance V des p variables est donnée par

$$V = {}^tXDX - g{}^tg = Y'DY.$$

3) Notons s_j l'écart-type $s_j^2 = s_{\overline{x_{.,j}},\overline{x_{.,j}}}$ de la j-ème variable et $D_{1/s} = \mathrm{Diag}(()s_1^{-1}, \ldots, s_p^{-1})$. Montrer que le tableau des données centrées réduites Z est donné par $Z = YD_{1/s}$. Puis montrer que la matrice de corrélation est donnée par $R = D_{1/s}VD_{1/s} = {}^tZDZ$.

4) On suppose X centré ($X = Y$). Vérifier que l'inertie du nuage de point $\mathcal{N} = (x_{i,.})_{1\leq i\leq n}$ (dans l'espace des variables) définie par

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

est donnée par $I(\mathcal{N}) = \mathrm{Tr}(V)$. Que vaut $I(\mathcal{N})$ dans le cas d'un tableau centré réduit ($X = Z$) ?

5) L'espace des variables $\mathbb{R}^n$ est muni de la métrique associé à la matrice D. Vérifier que la moyenne d'une variable $c \in \mathbb{R}^n$ est alors donnée par le produit scalaire ${}^t\mathbb{1}_n Dc$ et la covariance de deux variables centrées c_1 et c_2 par le produit scalaire ${}^tc_1Dc_2$. Vérifier également que la norme d'une variable centrée est son écart type, et que le cosinus de l'angle entre deux variables centrées est égal à leur coefficient de corrélation.

6) Soit un vecteur unitaire u de $\mathbb{R}^p$ et Δ l'axe engendré. La liste des coordonnées c_i des individus sur cet axe forme une nouvelle variable artificielle $c \in \mathbb{R}^n$ donnée par

$$c = (\langle x_{i,.}; a\rangle)_{1\leq i\leq n}.$$

Vérifier que $c = Xu$, et donc que c est une combinaison linéaire des variables originelles $x_{.,j}$ $(1 \leq j \leq p)$, puis que la variance de c vaut $s_c^2 = {}^tuVu$.