Analyse des données (formulaire)

Analyse factorielle en composantes principales

La technique utilisée consiste à projeter l'ensemble des individus sur un espace de dimension *q* (inférieur à *p*) en minimisant la déformation lors de la projection. En général *q* est égal à 2.

Généralités

- *nombre d'individus* : *n* (indice *i*)
- *nombre de variables* : *p* (indice *j*)
- *tableau de données* : $X = \left(x_i^j\right)_{n\times p}$
- *vecteur individu* : $\underline{x}_i$ (correspond à la ligne *i* du tableau de donnée)
- *vecteur variable* : $\underline{x}^j$ (correspond à la colonne *j* du tableau de donnée)
- *moyenne* : $\bar{x}^j = \sum_i p_i x_i^j$ avec $p_i = \frac{1}{n}$ le poids
- *variance* : $\mathrm{var}\left(\underline{x}^j\right) = \sum_i p_i \left(x_i^j - \bar{x}^j\right)^2$
- *écart-type* : $\sigma^j = \sqrt{\mathrm{var}\left(\underline{x}^j\right)}$
- *covariance* : $\mathrm{cov}\left(\underline{x}^j, \underline{x}^{j'}\right) = \sum_i p_i \left(x_i^j - \bar{x}^j\right)\left(x_i^{j'} - \bar{x}^{j'}\right)$
- *corrélation* : $\mathrm{cor}\left(\underline{x}^j, \underline{x}^{j'}\right) = \frac{\mathrm{cov}\left(\underline{x}^j, \underline{x}^{j'}\right)}{\sigma^j \sigma^{j'}}$
- *centre de gravité* : $\underline{G} = \sum_i p_i \underline{x}_i$
- *inertie du nuage* : $I_N = \sum_i p_i d^2\left(\underline{x}_i, \underline{G}\right)$
- *distance du $\aleph^2$ (pour tableaux de données quantitatives)* : $d^2\left(\underline{x}_i, \underline{x}_{i'}\right) = \sum_j \frac{1}{x_\bullet^j}\left(\frac{x_i^j}{x_i^\bullet} - \frac{x_{i'}^j}{x_{i'}^\bullet}\right)$

Les grandes étapes du calcul

1. calcul de $\tilde{X}$, le tableau de données centré par rapport aux variables.
2. calcul de la matrice des variances - covariances : $V = \frac{1}{n}\tilde{X}^t\tilde{X}$ de dimension *p*.
3. calcul de la matrice des corrélations : $S = M^{1/2}VM^{1/2}$ avec $M = diag\left(\frac{1}{\sigma^{j^2}}\right)$.

4. calcul des valeurs propres λ_k ($\lambda_1 > \cdots > \lambda_q$) de *S* et des vecteurs propres $\underline{v}_k$.

5. calcul des axes principaux d'inertie : $\underline{u}_k = M^{-1/2}\underline{v}_k$.

6. calcul de la $k^{ème}$ composante principale $\underline{y}^k$ qui donne les (nouvelles) coordonnées des *n* individus sur $[\underline{u}_k]$: $\underline{y}_i^k = \langle \underline{x}_i, \underline{u}_k \rangle = \underline{x}_i^t . M . \underline{u}_k$.

Description des individus et des variables

- calcul de *l'inertie expliquée* : $IE_{[\underline{u}_k]} = \frac{\lambda_k}{trace(VM)}$, $IE_{[\underline{u}_k, \underline{u}_{k'}]} = \frac{\lambda_k + \lambda_{k'}}{trace(VM)}$. La représentation est d'autant meilleure que l'inertie expliquée est proche de 0 (cas limite où il n'y a pas de déformation).

- **description des individus**

- Les individus sont projetés sur la plan $[\underline{u}_k, \underline{u}_{k'}]$.

- calcul du *cos²* (qualité ponctuelle de la représentation) : La qualité ponctuelle par rapport à l'axe $[\underline{u}_k]$ est $e_i^k = \frac{\left(y_i^k\right)^2}{\|x_i\|^2}$. Plus cette quantité est proche de 1, meilleure sera la représentation du point par rapport à cet axe. La qualité ponctuelle par rapport au plan $[\underline{u}_k, \underline{u}_{k'}]$ est $e_i^k + e_i^{k'}$.

- calcul de la contribution absolue des points par rapport aux axes :Contribution de l'individu $\underline{x}_i$ à l'axe $[\underline{u}_k]$: $\alpha_i^k = \frac{p_i \left(y_i^k\right)^2}{\lambda_k} \leq 1$.

- **description des variables**

- Les variables sont projetés sur le plan $[\underline{y}^k, \underline{y}^{k'}]$.

- calcul des coordonnées de la variable $\underline{x}^j$ sur l'axe $[\underline{u}_k]$: $\langle \underline{x}^j, \underline{y}^k \rangle = \frac{\sqrt{\lambda_k}}{\sigma^j} \underline{u}_k^j$. Les variables $\underline{x}^j$ vont se projeter sur le cercle des corrélations.

Méthode

1. On choisit le plan de projection $[\underline{u}_1, \underline{u}_2]$, ayant la plus forte inertie expliquée.

2. On repère dans le tableau sur les individus les contributions élevées (réparties entre des coordonnées négatives et positives) pour chacun de ces axes. On identifie 4 groupes apposées sur le plan $[\underline{u}_1, \underline{u}_2]$. On marque la dépendance d'un axe par une flèche.

3. On repère dans le tableau sur les individus les corrélations absolues les plus élevées et l'on forme 4 groupes sur le plan $[\underline{y}_1, \underline{y}_2]$. Il faut penser à dessiner le cercle des corrélations (qui peut être déformé).

4. *Interprétation* : On décrit les 4 groupes individus - variables mis en évidence par la méthode.

Analyse factorielle des correspondances

Cette méthode traite les tableaux de contingences. Elle utilise deux analyses en composantes principales : l'une sur le nuage des individus, l'autre sur le nuage des variables.

Méthode

1. On choisit le plan de projection $[\underline{u}_1, \underline{u}_2]$, ayant la plus forte inertie expliquée.

2. On peut appeler les axes soit par les variables, soit par les individus. Il convient de donner une signification à cette axe, simplement en le nommant : *"échelle de…"*.

3. Si l'on choisit d'appeler les axes par les *variables*, on retiendra d'une part les *variables* ayant une contribution absolue élevée et d'autre part les *individus* ayant une contribution relative élevée (cos^2). Et inversement, si l'on choisit d'appeler les axes par les individus.

4. On forme ainsi 4 groupes principaux pour les variables et 4 groupes principaux pour les individus que l'on met en relation pour l'interprétation.

Classification automatique

Méthode