Université de Bordeaux - Master MIMSE - 2ème année

Rappels sur l'ACP avec métriques

Marie Chavent

http://www.math.u-bordeaux.fr/ machaven/

2014-2015

1 Les données

1. $\mathbf{Z}$ est une matrice $n \times p$ regroupant la mesure sur n individus de p variables **quantitatives**.
2. Chaque individu i est caractérisé par un vecteur $\mathbf{z}_i \in \mathbb{R}^p$ ($i$ème ligne de $\mathbf{Z}$). L'espace des individus $\mathbb{R}^p$ est muni d'une métrique $\mathbf{M}$ qui est souvent l'indentité $\mathbf{I}_p$.
3. Chaque variable j est caractérisée par un vecteur $\mathbf{z}^j \in \mathbb{R}^n$ ($j$ème colonne de $\mathbf{Z}$). L'espace des variables $\mathbb{R}^n$ est muni de la métrique $\mathbf{N} = \text{diag}(\ldots, p_i, \ldots)$ où p_i, est le poids de l'individu i. Généralement $p_i = \frac{1}{n}$.
4. Le tableau $\mathbf{Z}$ est généralement centré pour fixer le centre de gravité du nuage des individus à l'origine : pour chaque variable j on effectue $z^j \to z^j - \bar{z}_j \mathbf{1}_n$, où $\bar{z}_j = \sum_{i=1}^n p_i z_{ij}$.
5. Le tableau $\mathbf{Z}$ est souvent réduit ensuite pour éliminer le fait que les variables sont sur mesurées sur des échelles (de variation) différentes : pour chaque variable j on effectue $z^j \to \frac{z^j}{s_j}$, où $s_j^2 = \sum_{i=1}^n p_i (z_{ij} - \bar{z}_j)^2$ est la variance empirique de j.

2 Définitions

1. Le vecteur $\Psi^\alpha \in \mathbb{R}^n$ des coordonnées factorielles (des scores) des n individus sur l'axe α est le vecteur des projections orthogonales (pour le produit scalaire $\mathbf{M}$) des n individus sur l'axe engendré par le vecteur $\mathbf{v}_\alpha$ ($\mathbf{M}$-normé à 1) maximisant l'inertie du nuage projeté. On a :

$$\Psi^\alpha = \mathbf{Z}\mathbf{M}\mathbf{v}_\alpha$$

où $\mathbf{v}_\alpha$ est le vecteur propre de $\mathbf{Z}^t\mathbf{NZM}$ associé à la valeur propre λ_α.

Le vecteur Ψ^α s'appelle aussi αème composantes principales. Les composantes principales sont donc des nouvelles variables synthétiques (des combinaisons linéaires des variables initiales $\mathbf{z}^j$) qui vérifient :

— la variance de la composante principale Ψ^α est égale à la valeur propre λ_α : $Var(\Psi^\alpha) = \lambda_\alpha$.

— la corrélation linéaire entre deux composantes principales Ψ^α et $\Psi^{\alpha'}$ est nulle : $r(\Psi^\alpha, \Psi^{\alpha'}) = 0$.

En résumé, les k composantes principales $\Psi^1, \ldots, \Psi^k$ sont de nouvelles variables synthétiques (combinaisons linéaires des variables initiales) non corrélées entre elles, de variance maximale et les plus liées (en un certain sens) aux variables initiales $\mathbf{z}^1, ..., \mathbf{z}^p$.

2. Le vecteur $\Phi^\alpha \in \mathbb{R}^p$ des coordonnées factorielles (des loadings) des p variables sur l'axe α est le vecteur des projections orthogonales (pour le produit scalaire $\mathbf{N}$) des p variables sur l'axe engendré par le vecteur $\mathbf{u}_\alpha$ ($\mathbf{N}$-normé à 1) maximisant l'inertie du nuage projeté. On a :

$$\Phi^\alpha = \mathbf{Z}^t\mathbf{N}\mathbf{u}_\alpha$$

où $\mathbf{u}_\alpha$ est le vecteur propre de $\mathbf{ZMZ}^t\mathbf{N}$ associé à la valeur propre λ_α.

Le vecteur des loadings Φ^α contient les corrélations des p variables avec la αème composantes principales Ψ^α : $\phi_{j\alpha} = r(\mathbf{z}^j, \Psi^\alpha)$.

3 ACP avec métriques

L'ACP du triplet $(\mathbf{Z}, \mathbf{N}, \mathbf{M})$ revient à effectuer une décomposition en valeurs singulières généralisée (GSVD - Generalized Singular Value Decomposition) d'une matrice réelle $\mathbf{Z}$ de dimension $n \times p$ avec les métriques $\mathbf{N}$ sur $\mathbb{R}^n$ et $\mathbf{M}$ sur $\mathbb{R}^p$.

SVD généralisée. La GSVD de la matrice $\mathbf{Z}$ de rang r avec les métriques $\mathbf{N}$ et $\mathbf{M}$ donne la décomposition :

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t$$

avec

— $\mathbf{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ est la matrice diagonale des r valeurs singulières de $\mathbf{ZNZ}^t\mathbf{M}$ et $\mathbf{Z}^t\mathbf{NZM}$.

— $\mathbf{U}$ est la matrice de dimension $n \times r$ dont les colonnes sont les vecteurs propres de $\mathbf{ZMZ}^t\mathbf{N}$ et $\mathbf{U}^t\mathbf{NU} = \mathbb{I}_r$ (les vecteurs propres sont $\mathbf{N}$-orthonormés).

— $\mathbf{V}$ est la matrice de dimension $p \times r$ dont les colonnes sont les vecteurs propres de $\mathbf{Z}^t\mathbf{NZM}$ et $\mathbf{V}^t\mathbf{MV} = \mathbb{I}_r$ (les vecteurs propres sont $\mathbf{M}$-orthonormés).

En pratique, on peut obtenir la DVSG de $\mathbf{Z}$ en trouvant la DVS classique de $\tilde{\mathbf{Z}} = \mathbf{D}^{1/2}\mathbf{ZM}^{1/2}$, c'est à dire la DVSG avec les métriques $\mathbb{I}_n$ sur $\mathbb{R}^n$ et $\mathbb{I}_p$ sur $\mathbb{R}^p$. On trouve $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^t$ et :

$$\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}, \quad \mathbf{U} = \mathbf{D}^{-1/2}\tilde{\mathbf{U}}, \quad \mathbf{V} = \mathbf{M}^{-1/2}\tilde{\mathbf{V}}.$$

Coordonnées factorielles des individus et des variables

1. On note Ψ la matrice de dimension $n \times r$ des coordonnées des $\mathbf{M}$-projections des n individus (les lignes de $\mathbf{Z}$) sur les axes de vecteurs directeurs $\mathbf{v}^1, \ldots, \mathbf{v}^r$ (colonnes de $\mathbf{V}$). Cette matrice est aussi appellée matrice des coordonnées factorielles (scores) des individus. On a par définition que :

   $$\Psi = \mathbf{ZMV}$$

   On en déduit donc que :

   $$\Psi = \mathbf{U}\Lambda$$

   On en déduit également que $\mathbf{u}_\alpha = \Psi^\alpha/\sqrt{\lambda_\alpha}$ est la αème composante principale standardisée (divisée par son écart-type) et que $\mathbf{U}$ est la matrice des composantes principales standardisées.

2. On note Φ la matrice de dimension $p \times r$ des coordonnées des $\mathbf{N}$-projections des p variables (les colonnes de $\mathbf{Z}$) sur les axes de vecteurs directeurs $\mathbf{u}^1, \ldots, \mathbf{u}^r$ (colonnes de $\mathbf{U}$). Cette matrice est aussi appellée matrice des coordonnées factorielles (loadings) des variables. On a par définition que :

   $$\Phi = \mathbf{Z}^t\mathbf{MU}$$

   On en déduit donc que :

   $$\Phi = \mathbf{V}\Lambda$$