1 Analyse en composantes principales

1.1 Matrice de variance-covariance

A partir des données multidimensionnelles, on étudie généralement les liaisons entre les variables observées: c'est ce qu'on appelle l'étude des corrélations. Soit X le tableau des données de dimensions $n \times p$ d'élément x_{ij} construit à partir de n individus définis par p variables.
On définit l'individu moyen par le vecteur de $\mathbb{R}^p$ par : $g = [\overline{x_1}, ...\overline{x_p}]$.
Soit maintenant X_C le tableau centré en g de dimensions $n \times p$ défini par : $X_{C_{ij}} = x_{ij} - \overline{x_j}$.

Différentes formes de nuages de points : forme allongée, forme parabolique, forme sphérique

Matrice de variance-Covariance Σ : On appelle *la matrice de variance-covariance*, la matrice symétrique de dimension $p \times p$ dont les éléments diagonaux représentent la variance des variables et les éléments hors diagonaux la covariance entre les variables :

$$\Sigma = \frac{1}{n} X_C^T X_C$$

La covariance de la variable j et l, notée Σ_{jl}, mesure la liaison/dépendance des paramètres :

$$\Sigma_{jl} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x_j})(x_{il} - \overline{x_l})$$

Corrélation entre les variables : A partir de cette matrice Σ, on définit aussi *la corrélation entre les variables X et Y*, indépendant des unités de mesure de X et de Y. Le coefficient de corrélation est symétrique :

$$-1 \leq Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \leq 1$$

- $Corr(X,Y) = 0$, les variables sont quasiment décorrélées c'est-à-dire étant donné X, on ne peut rien dire.prédire sur la valeur de Y.
- $Corr(X,Y) = 1$, dépendance linéaire positive de X et Y.
- $Corr(X,Y) = -1$, dépendance linéaire négative de X et Y.

1.2 Méthode

les veuteurs propres teur

Les composantes principales $C_1, .., C_q$ sont des nouvelles variables combinaison linéaire des variables d'origines $x_{.1}, ...x_{.p}$ telles que les C_k soient 2 à 2 non

corrélées, de variance maximale, d'importance décroissante.

Décomposition de la variance :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - g)^T(x_i - g)$$

Projection sur une droite : L'opérateur de projection orthogonale, noté π, sur une droite de vecteur directeur unitaire v s'écrit :

$$\Pi = vv^T$$

avec $v^T v = 1$. La variance des observations projetés s'écrit alors :

$$\begin{aligned}
\sigma_V^2 &= \frac{1}{n}\sum_{i=1}^{n}(\Pi(x_i - g))^T(\Pi(x_i - g)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\Pi(x_i - g))^T(\Pi(x_i - g)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(vv^T(x_i - g))^T(vv^T(x_i - g)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - g)^T vv^T(x_i - g)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(v^T(x_i - g)|v^T(x_i - g)\right) \\
&= \frac{1}{n}v^T\left(\sum_{i=1}^{n}(x_i - g)^T(x_i - g)\right)v \\
&= v^T\Sigma v
\end{aligned}$$

Recherche de la projection de variance maximale Observons que Σ est la matrice de variance-covariance. Cette matrice est symétrique définie positive. On doit donc maximiser cette variance des observations projetées:

$$\max_v v^T\Sigma v \text{ avec } v^T v = 1$$

Il s'agit d'un problème d'optimisation avec contrainte d'égalité. On introduit donc le Lagrangien :

$$\mathcal{L} = v^T\Sigma v + \lambda(1 - v^T v)$$

et on calcule les conditions nécessaires d'optimalité : $\partial_v\mathcal{L} = 0$. On obtient donc l'équation aux valeurs propres :

$$\Sigma v = \lambda v.$$

Comme la matrice Σ est symétrique définie positive, les valeurs propres sont réelles positives et les vecteurs propres peuvent être choisis orthonormés. Donc la solution est de projeter les données sur le vecteur propres associé à la plus grande valeur propres λ de Σ.

Recherche des projections de variance maximale orthogonales au premier axe : Afin de trouver le second axe de variance maximale, on résout :

$$\max_{v} v^T \Sigma v \text{ avec } v^T v = 1, v^T v_1 = 0.$$

Interprétation des vecteurs propres La somme des valeurs propres correspond à la variance totale:

$$Tr(\Sigma) = \sigma^2 = \sum_{i=1}^{p} \lambda_i$$

Chaque valeur propre mesure la part de variance expliquée par l'axe factoriel correspondant.

Choix de la dimension q

La qualité des estimations auxquelles conduit l'A.C.P. dépend, de façon évidente, du choix de q c'est-à-dire du nombre de composantes retenues pour reconstituer les données, ou encore de la dimension du sous-espace de représentations. De nombreux critères de choix pour q ont été proposés dans la littérature.
La "qualité globale" des représentations est mesurée par la part d'inertie expliquée :

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{i=1}^{p} \lambda_i}.$$