

Scoring & Applications – Chapitre 1 – Méthodes géométriques

Contexte et objectif

On dispose de n individus décrits par d variables continues et répartis en G groupes. $\mathbf{X} = (x_{i,j}) \in \mathbb{R}^{n \times d}$ est la matrice des descripteurs : $x_{i,j}$ est la valeur de la variable j mesurée sur l'individu i . $\mathbf{Z} = (z_{i,g}) \in \{0,1\}^{n \times G}$ est la matrice des appartenances : $z_{i,g} = 1$ si l'individu i appartient au groupe g et $z_{i,g} = 0$ sinon. L'individu i est assimilé à la colonne i de $\mathbf{X}'$ que l'on note $\mathbf{x}_i \in \mathbb{R}^d$.

On souhaite s'appuyer sur les informations contenues dans $\mathbf{X}$ et $\mathbf{Z}$ pour attribuer à l'un des groupes un nouvel individu $\mathbf{x}_{n+1} \in \mathbb{R}^d$.

Statistiques usuelles
Statistiques par groupe

Le groupe g a pour effectif : $n_g = \sum_{i=1}^n z_{i,g}$, son centre est : $\bar{\mathbf{x}}_g = \sum_{i=1}^n z_{i,g} \mathbf{x}_i / n_g$ et sa matrice de variance : $\mathbf{V}_g = \sum_{i=1}^n z_{i,g} (\mathbf{x}_i - \bar{\mathbf{x}}_g)(\mathbf{x}_i - \bar{\mathbf{x}}_g)' / n_g$.

Statistiques globales

Le centre du nuage est : $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i / n$ et la matrice de variance : $\mathbf{V} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' / n$.

Des statistiques par groupe aux statistiques globales

Le centre du nuage peut aussi être calculé selon : $\bar{\mathbf{x}} = \sum_{g=1}^G n_g \bar{\mathbf{x}}_g / n$.

La matrice de variance intra groupes : $\mathbf{W} = \sum_{g=1}^G n_g \mathbf{V}_g / n$ traduit la dispersion des données dans les groupes.

La matrice de variance inter groupes : $\mathbf{B} = \sum_{g=1}^G n_g (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})' / n$ traduit la dispersion des groupes autour du centre du nuage.

La matrice de variance peut être obtenue par le théorème de König-Huygens : $\mathbf{V} = \mathbf{B} + \mathbf{W}$.

Analyse Factorielle Discriminante

Quelques prérequis mathématiques : <http://alexandre.lourme.free.fr/M2IREF/SCORING/CH0.pdf>

Description de la méthode

Une matrice $\mathbf{M} \in \mathbb{R}^{d \times d}$ symétrique définie positive définissant la métrique de $\mathbb{R}^d$, l'analyse factorielle discriminante (AFD) à un facteur se résume à quelques étapes simples.

◦ Conditionnement des données (centrage)

On note $\mathbf{1}$ le vecteur de $\mathbb{R}^n$ dont toutes les composantes valent 1.

(i) centrer le nuage de points : $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{1} \bar{\mathbf{x}}'$

(ii) recentrer le nouvel individu : $\mathbf{x}_{n+1} \leftarrow \mathbf{x}_{n+1} - \bar{\mathbf{x}}$

Dès lors, les statistiques de la section précédente sont calculées en prenant l'origine pour centre du nuage : $\bar{\mathbf{x}} = \mathbf{0}_{\mathbb{R}^d}$.

◦ Analyse spectrale de la matrice $\mathbf{V}^{-1} \mathbf{B}$

(iii) déterminer la plus grande des valeurs propres de $\mathbf{V}^{-1} \mathbf{B}$ et $\mathbf{u}$ un vecteur propre associé^a

(iv) normaliser $\mathbf{u}$ de sorte qu'il soit $\mathbf{M}^{-1}$ -normé : $\mathbf{u} \leftarrow \mathbf{u} / \sqrt{\mathbf{u}' \mathbf{M}^{-1} \mathbf{u}}$

(v) déterminer le vecteur directeur $\mathbf{M}$ -unitaire de l'axe factoriel : $\mathbf{a} = \mathbf{M}^{-1} \mathbf{u}$

◦ Allocation du nouvel individu $\mathbf{x}_{n+1}$

(vi) déterminer les abscisses du nuage projeté $\mathbf{M}$ -orthogonalement sur l'axe factoriel : $\mathbf{s} = \mathbf{X} \mathbf{u} \in \mathbb{R}^n$

(vii) déterminer le centre de chaque groupe sur l'axe factoriel : $\bar{s}_g = \sum_{i=1}^n z_{i,g} s_i / n_g$

(viii) attribuer $\mathbf{x}_{n+1}$ au groupe dont il est le plus proche sur l'axe factoriel : $\mathbf{x}_{n+1}$ est attribué au groupe g pour lequel le score $\mathbf{s}_g(\mathbf{x}_{n+1}) = |\mathbf{x}'_{n+1} \mathbf{u} - \bar{s}_g|$ est minimal

Interprétation de la méthode

Une fois le nuage centré, l'AFD détermine la direction (celle de $\mathbf{a}$) sur laquelle le nuage doit être projeté $\mathbf{M}$ -orthogonalement pour que la série résultante ait une variance inter groupes maximale.

Justification de la méthode

A. Lourme, Faculté d'économie, gestion & AES, Université de Bordeaux <http://alexandre.lourme.free.fr>

^asi le sous espace propre associé à la plus grande des valeurs propres de $\mathbf{V}^{-1} \mathbf{B}$ est de dimension strictement supérieure à 1, le problème d'optimisation sous jacent à l'AFD possède plusieurs solutions.

Notons $\mathbf{a}$ un vecteur M -normé ($\mathbf{a}'\mathbf{M}\mathbf{a} = 1$) de $\mathbb{R}^d$ et $\langle \mathbf{a} \rangle$ la droite vectorielle engendrée par $\mathbf{a}$. $(\mathbf{a}'\mathbf{M}\mathbf{x}_i)\mathbf{a}$ est le projeté M -orthogonal de $\mathbf{x}_i$ sur $\langle \mathbf{a} \rangle$, $s_i = \mathbf{a}'\mathbf{M}\mathbf{x}_i$ est son abscisse sur $\langle \mathbf{a} \rangle$ et la série $\mathbf{s} = (s_1, \dots, s_n)' = \mathbf{X}\mathbf{M}\mathbf{a} \in \mathbb{R}^n$ contient l'ensemble des abscisses des points du nuage projetés M -orthogonalement sur $\langle \mathbf{a} \rangle$. La série $\mathbf{s}$ hérite de la structure en G groupes des données de $\mathbf{X}$: si $\mathbf{x}_i$ est dans le groupe g du nuage, $s_i\mathbf{a}$ est dans le groupe g des points projetés sur $\langle \mathbf{a} \rangle$. La variance inter groupes de $\mathbf{s}$ est : $\mathbf{a}'\mathbf{M}\mathbf{B}\mathbf{M}\mathbf{a}$, la variance intra groupes : $\mathbf{a}'\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{a}$ et la variance totale : $\mathbf{a}'\mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a}$. On cherche $\mathbf{a}$ de sorte que le rapport de corrélation de $\mathbf{s}$: $\eta^2 = (\mathbf{a}'\mathbf{M}\mathbf{B}\mathbf{M}\mathbf{a})/(\mathbf{a}'\mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a})$ soit maximal ou, de façon équivalente, $\mathbf{u} = \mathbf{M}\mathbf{a}$ de sorte que $\eta^2(\mathbf{u}) = (\mathbf{u}'\mathbf{B}\mathbf{u})/(\mathbf{u}'\mathbf{V}\mathbf{u})$ soit maximal. On peut montrer^b que $\eta^2(\mathbf{u})$ est maximal lorsque $\mathbf{u}$ est un vecteur propre associé à λ , la plus grande des valeurs propre de $\mathbf{V}^{-1}\mathbf{B}$. Or, $\mathbf{u}$ est M^{-1} -normé puisque $\mathbf{a}$ est M -normé. Ainsi, $\mathbf{u}$ est un vecteur propre M^{-1} -normé associé à la plus grande des valeurs propres de $\mathbf{V}^{-1}\mathbf{B}$. $\mathbf{u}$ est le premier facteur discriminant ; il représente la forme linéaire à appliquer aux d variables initiales pour obtenir la série $\mathbf{s}$ la plus discriminante. $\langle \mathbf{a} \rangle$ est l'axe factoriel. λ est le rapport de corrélation maximal que l'on peut obtenir en projetant le nuage M -orthogonalement sur une droite vectorielle.

Un cas particulier

Lorsque les observations sont réparties en deux groupes ($G = 2$), la métrique de Mahalanobis ($\mathbf{M} = \mathbf{V}^{-1}$) permet d'obtenir le facteur discriminant et l'axe discriminant directement^c ; en effet, $\mathbf{a}$ est colinéaire à $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ et $\mathbf{u}$ est colinéaire à $\mathbf{V}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

Extension à plusieurs facteurs

L'analyse factorielle discriminante s'étend à p facteurs ($1 \leq p \leq \min\{G - 1, d\}$) en adoptant la métrique de Mahalanobis : $\mathbf{M} = \mathbf{V}^{-1}$.

$\mathbf{V}^{-1}\mathbf{B}$ est $\mathbf{V}$ -symétrique, donc diagonalisable dans une base $\mathbf{V}$ -orthonormée : $(\mathbf{u}_1, \dots, \mathbf{u}_d)$. En notant λ_i la valeur propre associée au vecteur propre $\mathbf{u}_i$ on peut supposer sans perte de généralité : $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$.

Rq. Les valeurs propres λ_i sont positives ou nulles : $\mathbf{V}^{-1/2}\mathbf{B}\mathbf{V}^{-1/2}$ est une matrice réelle, symétrique, positive ; elle est diagonalisable dans une base orthonormée et ses valeurs propres sont positives ou nulles ; elle possède les mêmes valeurs propres que $\mathbf{V}^{-1}\mathbf{B}$.

On note $\mathbf{U}$ la matrice $d \times p$ dont la colonne j est composée des coordonnées du j^{e} facteur discriminant : $\mathbf{u}_j$.

La matrice $\mathbf{S} = \mathbf{X}\mathbf{U} \in \mathbb{R}^{n \times p}$ est alors constituée des données factorielles : la ligne i de $\mathbf{S}$ représente l'individu i dans l'espace factoriel et la colonne j de $\mathbf{S}$ la j^{e} variable discriminante.

La matrice de variance des données factorielles : $\mathbf{U}'\mathbf{V}\mathbf{U} \in \mathbb{R}^{p \times p}$ est l'identité. La matrice de variance inter groupes : $\mathbf{U}'\mathbf{B}\mathbf{U} \in \mathbb{R}^{p \times p}$ est diagonale ; ses coefficients diagonaux sont : $\lambda_1, \dots, \lambda_p$. Donc : $\text{trace}(\mathbf{U}'\mathbf{B}\mathbf{U})/\text{trace}(\mathbf{U}'\mathbf{V}\mathbf{U}) = \sum_{j=1}^p \lambda_j/p$.

L'analyse factorielle discriminante à p facteurs en quelques étapes

Les étapes (i) et (ii) de conditionnement des données étant réalisées :

- déterminer la matrice $\mathbf{U} \in \mathbb{R}^{d \times p}$ formée en colonne des facteurs discriminants $\mathbf{u}_1, \dots, \mathbf{u}_p$
- calculer $\mathbf{S} = \mathbf{X}\mathbf{U} \in \mathbb{R}^{n \times p}$; la colonne j de $\mathbf{S}$ est la variable discriminante j ; elle contient les coordonnées des points du nuage projetés sur l'axe factoriel j
- déterminer le centre de chaque groupe dans l'espace factoriel : $\mathbf{U}'\bar{\mathbf{x}}_g$
- déterminer les coordonnées du nouvel individu dans l'espace factoriel : $\mathbf{U}'\mathbf{x}_{n+1}$
- affecter le nouvel individu au groupe dont le centre est le plus proche dans l'espace factoriel : $\mathbf{x}_{n+1}$ est attribué au groupe g pour lequel le score $\mathbf{s}_g(\mathbf{x}_{n+1}) = \|\mathbf{U}'\mathbf{x}_{n+1} - \mathbf{U}'\bar{\mathbf{x}}_g\|$ est minimal

Application

Le fichier `client`^d donne pour cent clients bancaires : la liquidité (cash), le flux (flow), l'épargne (saving), le niveau de consommation (consume) et la classe de risque (risk).

A quelle classe de risque l'Analyse Factorielle Discriminante affecte-t-elle un nouveau client avec les caractéristiques suivantes : liquidité = 6,2 ; flux = 2,9 ; épargne = 4,9 ; consommation = 1,7 ?

References

- [1] Gilbert Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

^bvoir [1], p.484

^csans avoir à déterminer le spectre de $\mathbf{V}^{-1}\mathbf{B}$

^ddisponible sous : <http://alexandre.lourme.free.fr/M2IREF/SCORING/client.csv>

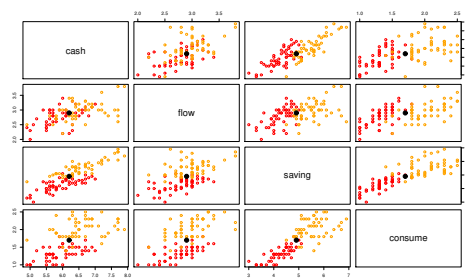


Fig. 1: Cent clients bancaires décrits par quatre variables continues (*cash*, *flow*, *saving*, *consume*) et répartis en deux classes de risque (0 en rouge, 1 en orange) ; à quelle classe de risque doit-on affecter le nouveau client (point noir) ?