Scroing et appli

Youming

2023-09-12

Un projet : fin septembre démonstration de la marque et EX5 T1 concerne exercie 2 du projet

CH1 : Une méthode géométrique : Analyse Factorielle Discriminante (AFD)

Nombre d'individus: n

Nombre de description : d

Nombre de goupe : G

Matrice de descripteur : $X_{i;j} \in R^{n,d}$

Valeur de variable mesurée sur individus : $X_{i;j}$

Matrice des appartenance : $Z(Z_{i;j}) \in \{0,1\}^{n \times G}$

$$\begin{cases} Z_{i;j} = 1 \Rightarrow Dans\ le\ groupe \\ Z_{i;j} = O \Rightarrow Pas\ dans\ le\ group \end{cases}$$

On note x_i (petit x) la colonne i de X' : la transposée de X ainsi $x_i = \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_{i,d} \end{pmatrix}$ resperente l'individu i

(coordonnés de l'individu i dans la base canonique)

$a = M^{-1}u$

NB : $u'M^{-1}u = 1$ $a'MM^{-1}Ma = 1$ $a'Ma = 1$ $|a|_n = 1$

La statistique par goupe :

Description	Equation
Nombre d'individu dans un group :	$n_g = \sum_{i=1}^{n} Z_{i,g}$
le centre de groups :	$g = \bar{x_q} = \sum Z_{i,g} \times X_i / n_g$
la matrice de covariance dans le groups h :	$V_g = \sum_{x=1}^{n} Z_{i,g}(X_i - \bar{X}_g)(X_i - \bar{X}_g)'/n_g$

La statistique globale| la statistique marginal :

Description	Equation
Le centre nuage :	$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$
La matrice ds covarience du nuage	$V = \sum_{i=1}^{n} \frac{(X_i - \bar{X}_g)(X_i - \bar{X}_g)'}{n}$

Lien entre les statistiques par la groupe et la statistique marginal :

Description	Equation
Centre de nuage	$\bar{X} = \sum_{g=1}^{G} \frac{n_g \bar{x}_g}{n}$
Matrice de variance intra :	$W = \sum_{g=1}^{G} \frac{n_g}{n} V_g$
Matrice de variance inter :	$B = \sum_{g=1}^{G} \frac{n_g}{n} (X_i - \bar{X}_g)(X_i - \bar{X}_g)'$
Et la matrice des covariance :	$V = B + W$

Analyse Factorielle Discriminante 主成分分析

1. Conditionnement des donnés

(i) Centre du nuage de point $X \leftarrow X -_1 \times \bar{X}'$,où $1 = (1;1;1....1)' \in R^n$ 用 1 矩阵取值 X 均值，然后相减求中心点

(ii) Centrage du nouvel individu $X_{n+1} \leftarrow X_{n+1} - \bar{X}$ 对新来的矩阵也求中心点。

2. Analyse spectrale de :$V^{-1}B$

$V^{-1}B$ 表示类别之间的差异与类别内部差异的比率。其中，V 代表总体的协方差矩阵，B 代表类别之间的协方差矩阵。最大的特征值对应的特征向量指示了最大的类间差异，求解 $V^{-1}B$ 的特征值和特征向量可以帮助我们找到最佳的投影方向，以最大化类间差异并最小化类内差异。

(iii) Trouver la plus grands de **valeur propres** λ de $V^{-1}B$ et $u \in R^d$ **vecteur popre** associée; 找到最大的特征值，特征向量。

(iv) Normaliser u de sort qu'il soit M^{-1} normée

$$u \leftarrow \frac{u}{\sqrt{u'M^{-1}u}}$$

u s'appel le facteur du discriminant

M 是对角线对称矩阵，$M \in R^{d\times d}$ symétrique défini positive.

Ps : $\sqrt{u'M^{-1}u}$; 马氏距离, 这里的 M 在 code 中是 V 的逆矩阵

(v) On détermine le vecteur $a = M^{-1}u$. 方向向量。

3. Allocation du nouvel individu x_{n+1}

$u'M^{-1}u$ 为何被称之为标准化?

$$\Rightarrow u \leftarrow \frac{u}{\sqrt{u'M^{-1}u}}$$

所以 $$u'M^{-1}u = \frac{u'}{\sqrt{u'M^{-1}u}} M^{-1} \frac{u}{\sqrt{u'M^{-1}u}} = \frac{u'M^{-1}u}{\left(\sqrt{u'M^{-1}u}\right)^2} = 1$$

不等于1都难，只能等于1

$$NB : u'M^{-1}u = 1 \Leftrightarrow a'MM^{-1}a = 1 \Leftrightarrow a'Ma = 1 \Leftrightarrow ||a||_M = 1$$

(vi) Déterminer **absisscce des pojetés** M-orthogonality des points du nuage sur **a** (droite vertorielle engendrée par a) : $s = Xu \in R$. 将 x 投影。

$$\begin{aligned} S_i &= x_i'u = x_i'Mai \\ &= a'Mx_i \end{aligned}$$

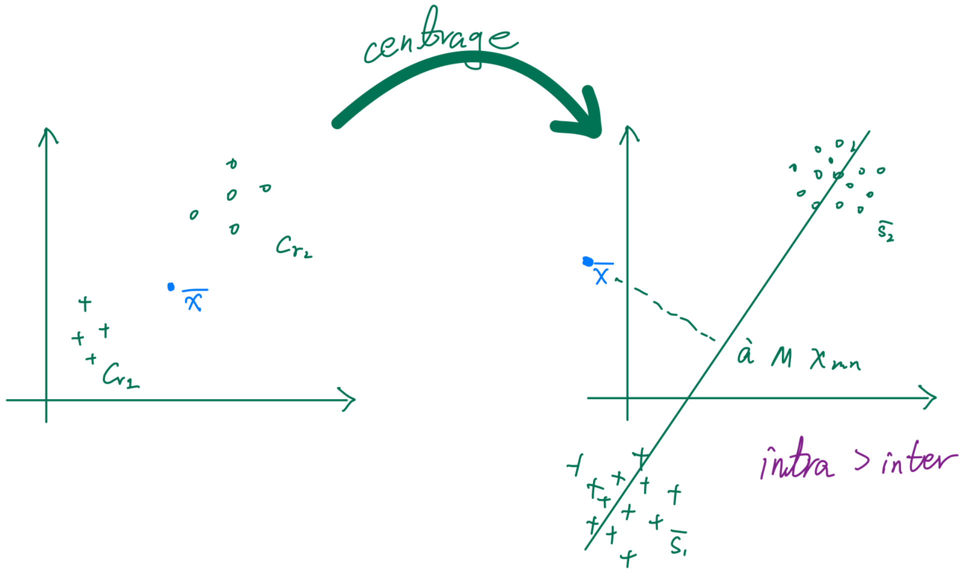
(vii) déterminer le centre de chaque groupe sur l'axe factoriel : $s = (s_1, s_2, ..., s_n)' \in R^m$

(vii) On détermine le centre des groups sur l'axe factoriel $\bar{s_g} = \sum_{i=1}^{n} z_{i,g}\frac{s_i}{n_g}$;

- z_i 是组内人数

(viii) On affecte x_{n+1} (nouveau entrant) au groue dont il est le plus proche sur l'axe factorielle

$$\begin{gathered} s_g(x_{mn}) = |u'X_{mn} - \bar{s_g}| \\ \downarrow \\ \textit{abcisse du projet d'orthogonality de } x_m n \textit{ sur } < a > \end{gathered}$$



Pourquoi l'axe d'orthorial est le meilleur ?
parce que la variance inter est très petit que la variance intra
PS : `<a>` = **axe factorielle**

Idée général : on se donne $a \in R^d$ R-normé

- On cherche `a` de sort que le rappoet de corrélation du nugae projeté sur `<a>` soit maximal
- On projet les points $X_i; i = 1, 2....n\ et\ X_{mn}$ sur l'axe factoriel
- On affecte x_{mn} au groupe dont il est **le plus proche sur l'axe factorielle**

$s_i = a'Mx_i$ **abcisse de projeté M orthogral** de x_i sur `<a>` et le projeté M-orthogral de X_i sur `<a>` et $a(a'Mx_i)$

S 世界的性质

La varience inter groupe de s est $a'MBMa$

La variance intra groupes : $a'MWMa$

La variance totale : $a'MVMa$

Le rapport de corrélation de s : $\eta^2(u) = \frac{(u'Bu)}{(u'Vu)}$

特殊情况

当只有数据集只被归类为两组，则 a 和 $(\bar{x}_1 - \bar{x}_2)$ 共线，u 和 $V^{-1}(\bar{x}_1 - \bar{x}_2)$ 共线。且 **马氏距离** $(M = V^{-1})$ 可以直接得到**判别因子（Discriminant Factor）**和 **判别轴（discriminant axes）**

多因子拓展

主成分分析 AFD 方法可以分析直到 $(1 < p < \min\{G-1, d\})$ 个因子。

1. 分析步骤如下

   (a) Centre du nuage de point $X \leftarrow X -_1 \times \bar{X}'$,où $1 = (1; 1; 1....1)' \in R^n$

   (b) Centrage du nouvel individu $X_{n+1} \leftarrow X_{n+1} - \bar{X}$

   (c) Former matrice U $R^{d \times p}$, U 由 $\{u_1, ..., u_j\}$ 判别因子组成

   (d) Calculer $A = M^{-1}U$, A 由 $\{a_1, ..., a_j\}$ 判别轴组成

   (e) $S = XU$ 将点用 U 投影进 S 世界

   (f) 找到 S 世界中每组在判别轴的中心点：$A'M\bar{X}$

   (g) 找到新个体在 S 世界判别轴的位置：$A'MX_{n+1}$

   (h) 判别新个体离哪个中心点近 $s_g(x_{n+1}) = \min \|A'MX - A'M\bar{X}\|$

Rappelle

$s = (s_1, s_2, ..., s_m)' =$ la série des abcisse des projetée de $x_i, ..., x_m = XMa$

le centre de s :

$$\begin{aligned}
\bar{s} &= \sum_{i=1}^{n} s_i/n \\
&= \sum_{i=1}^{n} a'Mx_i/n \\
&= a'M(\sum_{i=1}^{n} X_i/n) \\
&= a'M\bar{X} \\
&= 0
\end{aligned}$$

$$Car\ \bar{X} = 0$$

(étape 1, centre de nuage)

La somme de group g de s :

$$
\begin{aligned}
\bar{s} &= \sum_{i=1}^{n} z_{i,g} s_i / n_g \\
&= \sum_{i=1}^{n} z_{i,g} (a' M x_i) / n_g \\
&= a' M (\sum_{i=1}^{n} z_{i,g} X_i / n_g) \\
&= a' M \bar{X}_g
\end{aligned}
$$

La variance de s :

$$
\begin{aligned}
Var(s) &= \sum_{i=1}^{n} (S_i - \bar{S}_)^2 / n \\
&= \sum_{i=1}^{n} s\ddot{}2_i / n \\
&= \sum_{i=1}^{n} (a' M X_i)(X_i M a) / n \\
&= a' M (\sum_{i=1}^{n} x_i x_i' / n) M a \\
&= a' m (\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' / n) M a \\
&= a' M V M a
\end{aligned}
$$

La variance inter group de s :

$$
\begin{aligned}
Var_{Inter}(s) &= \sum_{g=1}^{G} (\bar{s_g} - \bar{s})^2 / n \\
&= \sum_{g=1}^{G} n_g \bar{S_g}^2 / n \\
&= \sum_{g=1}^{G} N_g (a' M \bar{X}_g \bar{X}_g' M a) / n \\
&= a' M (\sum_{g=1}^{G} n_g \bar{X}_g \bar{X}_g' / n) M a \\
&= a' M (\sum_{g=1}^{G} n_g (\bar{X}_g - \bar{X})(\bar{X}_g - \bar{X})') M a \\
&= a' M B M a
\end{aligned}
$$

On cherche a tem qie le rapport de corrélation de s soit maximal On maximise par rapport à `a` :

$$\frac{Var_{Inter}(s)}{var(s)} = \frac{a'MBMa}{a'MVMa} \qquad (*)$$

On poseui = Ma Optimiser (∗) par rapport à a recient à optimiser par rapport à `a` :

$$\frac{Var_{Inter}(s)}{var(s)} = \frac{a'MBMa}{a'MVMa} = \frac{u'Bu}{u'Vu}.......(*)$$

↓

"equation de Rayleigh"

Car $u = Ma$

-> la valeur maximale et obtenue quadratiquement un veteur prope de V^-1B associé à la plus grande des valeur propes

1h34'50 讲了 projet

> à démontréer comme EX 1 du projet Remark de projet : si G = 2 group et si la métrique choisi est celle de malalaibis ($M = V^{-1}$) alors l'axe estdirigé par $V^{-1}X_1 - \bar{X_2}$)

AFS à plusieur facteur on chercher à projeter le nuage de point D-orthogaralement sur un sev de R^d de domension p avec 1 p $\min(d_i G - 1)$

Rapelle :

$$B = \sum_{g=1}^{G} n_g \bar{x_g} \bar{x_g}'$$

$$\bar{x} = \sum_{g=1}^{G} n_g \bar{x_g}/n = 0$$

$$\bar{x_G} = -\sum_{g=1}^{G} \frac{n_g \bar{x_g}}{n n_G}$$

1h38 没听懂，什么 non null 很重要的样子

De façon ce que la somme des rappoets de correctation des nuages projetés sur la a< s facteurs b soit maximiser

- $V^{-1}B$ est V-symétrique , en effet

$$
\begin{aligned}
V(V^{-1}B) &= (V^{-1}B)'V \\
&= B'(V^{-1})'V \\
&= B'V'^{-1}V \\
&= B'V^{-1}V \\
&= B' \\
&= V(V'B)
\end{aligned}
$$

$V^{-1}B$ est donc diagonalisable dans une base $(u_1...u_d)$ qui est u-orthonervée <12> $V^{-1}B$ admet dans la valeur propre réelle $\lambda_1,\ \lambda_2\ ...\ \lambda_d$ avec λ_i = la valeur propore annoncée à u_i.

On peut supposée sur pert de générabilité : $\lambda_1 \geq lambda_2 \geq ... \geq \lambda_d$

- le d valeur proprs réelles de $V^{-1}B$ sont positive (non nulls)

Justification :

$$
\begin{aligned}
\lambda \in S_p(V^{-1}B) &\iff f_u \in R^d, \{0\} \\
tel\ que\ V^{-1}Bu = \lambda u \\
&\iff f_u \in R^d, \{0\}
\end{aligned}
$$

Tel que :

$$
\begin{aligned}
V^{\frac{1}{2}}RV^{-\frac{1}{2}}\underbrace{V^{\frac{1}{2}}u}_{=0} &= \lambda V^{\frac{1}{2}}u \\
&= \lambda \in Spect(V^{-\frac{1}{2}}RV^{-\frac{1}{2}})
\end{aligned}
$$

Donc le S_p qui s'appelle `Spect` $S_p(V^{-1/2}B) \subset R^+$

Ainsi, $\lambda_1 \geq lambda_2 \geq ... \geq \lambda_d \geq 0$

Le Vecteur $u_1, ..., u_d$ peivent être résidu M^{-1} normée sur chaque les v-orthogonal

$u_i \leftarrow \frac{u_i}{\sqrt{u_i' M^{-1} u_i}}$

ainsi chaque u_i est vecteur propre de $V^{-1}B$.

$V^{-1}B$ est V-symétrique $(V(V^{-1}B) = (V^{-1}B)'V$ à coefficients réels dont elle est diagonalisable dans une base $(u_i....u_d)$ de R^d qui sont M^{-1} unitaire pour chaque d'eux et deux à deux V-othogonale. Par ailleur, les valeurs prepres de $V^{-1}B$ sont réelles et positives $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d \geq 0$

Conditionnement de données

1. (i) et (ii) idem qu'en AFD 1 facteur

2. (iii) former la matrice $u = [u....u_p] \in R^{d\times p}$ dont la colonne $(i \in \{1....p\})$ est constituée des coordonné du vecteur u_i le vecteur u_i et le j^e facteur discriminant.

3. (iv) Calculé $A = M^{-1}U \in R^{d\times p}$. La colonne i de A est le vecteur a_j qui dirigele j^e axe factoriel. Puisque M_j est M^{-1} unitaire, a_j est M-normée.

4. (v) Calculer $S = XU = XMA \in R^{n\times p}$. La colonne $j, j \in \{1, ..., p\}$ de S se note $S^{[j]}$, elle est formée des cordonnées des projetée sur le j^e axe factoriel vect (a_j). $S^{[j]}$ est la j^e variable discriminant son rapport de corrélation est d_j.

Les variables $S^{[j]}, \quad j = 1, ..., p$ sont deux à deux non corrélés.

Calculer $S = XMA \in R^{n\times p}$. Elle est continuée des abcisse des projetés M-orthogonable des points du nuage sur `<a_i>`. le rapport de corrélation de s[1] est lendat_i . Le séris $s^{[i]}(i \in \{1...p\})$ sont non corrélés Si $i \neq j : cov(s^{[i]}, s^{[j]}) = \frac{1}{n}\sum_{k=1}^{n} S_k^{[i]} S_k^{[j]} = \frac{1}{n}(S^{[i]})'(S^{[j]}) = 0$

6. déterminer le centre de chaque group sur l'espace factoriel $A'M\bar{X_n} = U'\bar{X} \in R^p$
7. determiner le centre données de nouvel individu dans l'espace factoriel $A'M\bar{X_m}n = U'X_{mn}$
8. on affecté X_mn au groupe g pour meqiel le spre $S_g(X_{mn}) = ||A'MX_{mn} - A'M\bar{X}_g||$ est minimal.

Exercice 3

Anlyse Factorielle Discriminante à un facteur

```
client <- read.table(file = "http://alexandrelourme.free.fr/M2IREF/SCORING/client.csv" , sep=',',
```

Les données

```
X=as.matrix(client[,1:4]) # matrice des descripteurs
n=nrow(X) # taille de l'echantillon
d=ncol(X) # nombre de descripteurs/dimension de l'espace
G=length(unique(client[,5])) # nombre de groupes
Z=matrix(0,nrow=n,ncol=G) # tableau disjonctif complet/matrice des appartenances
for (i in 1:n){
  Z[i,1]=ifelse (client[i,5]==0,1,0)
  Z[i,2]=ifelse (client[i,5]==1,1,0)
}
```

Conditionnement des données

```
one=matrix(rep(1,n),nrow=n)
barx=t(colMeans(X)) # centre du nuage
```

[1] i

```
X <- X-one%*%barx # centrage du nuage
new=as.vector(c(6.2,2.9,4.9,1.7)) # nouveau client
new <- new - barx # recentrage du nouveau client
```

Effectifs des groupes

```
ng <- NULL
for (g in 1:G){
  ng[[g]]=sum(Z[,g]==1) # effectif du groupe g }
}
```

Centres des groupes

```
barxg <- NULL
for (g in 1:G){
  barxg[[g]]=colMeans(X[Z[,g]==1,]) # centres du groupe g
}
```

Matrices de variance des groupes

```
Vg <- NULL
for (g in 1:G){
  Vg[[g]]=var(X[Z[,g]==1,])*(ng[[g]]-1)/ng[[g]] # matrice de variance du groupe g
}
```

Matrice de variance intra groupes

```
W <- matrix(0,nrow=d,ncol=d) # matrice de variance intra groupes
for (g in 1:G){W=W+ng[[g]]/n*Vg[[g]]}
```

Matrice de variance inter groupes

```
B <- matrix(0,nrow=d,ncol=d) # matrice de variance inter groupes
for (g in 1:G){B=B+ng[[g]]/n*barxg[[g]]%*%t(barxg[[g]])}
```

Matrice de variance

```
V=B+W # théorème de Konig-Huygens
V=var(X)*(n-1)/n # calcul direct
```

Analyse spectrale de $V^{-1}B$

```r
# facteur discriminant
# M=diag(rep(1,d)) # choix de la métrique de Mahalanobis
M=solve(V) # choix de la métrique de Mahalanobis
EIG <- eigen(solve(V)%*%B)
lambda=EIG$values[1] # rapport de corrélation maximal d'une série obtenue par projection
u=as.vector(EIG$vectors[,1]) # facteur discriminant
u <- u/c(sqrt(t(u)%*%solve(M)%*%u)) # normalisation de u
```

Vecteur directeur M-unitaire de l'axe factoriel

```r
# axe discriminant
a=as.vector(solve(M)%*%u)
a=a/c(sqrt(t(a)%*%M%*%a))
```

Allocation du nouveau client

```r
s=X%*%u # variable discriminante
barvg <- NULL
for (g in 1:G){# centres des groupes sur la variable discriminantes
  barvg[[g]]=mean(s[Z[,g]==1,])
}
snew=new%*%u #abscisse du nouveau client sur l'axe factoriel
dist2group1=abs(snew-barvg[[1]]) # distance au groupe 1
dist2group2=abs(snew-barvg[[2]]) # distance au groupe 2
print(dist2group1)
```

```
##          [,1]
## [1,] 0.9251935
```

```r
##          [,1]
## [1,] 0.9251935
print(dist2group2)
```

```
##          [,1]
## [1,] 0.8455574
```

```r
##          [,1]
## [1,] 0.8455574
```

Le nouveau client est affecté au groupe 2 (codé 1 dans les données).

"非参数方法"（Méthode non paramétrique） 是统计学中的一种方法，用于分析数据和进行假设检验，而不依赖于特定的参数化概率分布。与参数方法不同，非参数方法不需要对数据的分布进行明确的假设，因此更具灵活性，适用于各种类型的数据分布和研究问题。

以下是非参数方法的一些关键特点和常见应用：

1. **无需分布假设**：非参数方法不要求研究人员提前假设数据服从特定的概率分布，如正态分布或泊松分布。这使得非参数方法在实际应用中更具通用性，因为真实世界的数据往往不容易用简单的分布来描述。
2. **基于排序或秩次的方法**：许多非参数方法基于数据的排序或秩次信息来进行分析。例如，Wilcoxon 符号秩检验和 Mann-Whitney U 检验用于比较两个样本的中位数，而不要求数据服从正态分布。
3. **典型应用**：非参数方法常用于以下情况：
   - 数据不满足正态分布假设。
   - 样本大小相对较小，不足以进行参数估计。
   - 研究问题要求更灵活的方法，而不是受限于特定的分布假设。
4. **常见的非参数方法**：一些常见的非参数方法包括：
   - 秩和检验（Wilcoxon 检验和 Mann-Whitney U 检验）。
   - 秩相关方法（Spearman 秩相关系数和 Kendall's ）。
   - 核密度估计。
   - 基于置换的方法（如置换检验）。
   - K-S 检验（Kolmogorov-Smirnov 检验）。

非参数方法的主要优点是它们通常更具**普适性**，它们可能需要更多的样本数据来达到相同的统计功效，并且在某些情况下可能不如参数方法那样精确。

> 描述统计：在统计学中，"valeur description" 可能指的是描述性统计量，例如均值、中位数、标准差等。这些统计量用于描述数据集的基本特征，以帮助理解数据的分布和性质