Statistique Mathématique 2 (MATH-F-309, Chapitre #4)

Thomas Verdebout

Université Libre de Bruxelles

2015/2016

Plan du cours

1. Vecteurs aléatoires.
2. Loi normale multivariée.
3. Inférence dans les modèles gaussiens.
4. Méthodes classiques de l'analyse multivariée.
5. Données directionelles.
6. Modèle linéaire.

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.1.1. Motivation, définition et calcul.

4.1.2. Propriétés et interprétation.

4.1.3. Version empirique.

4.1.4. Illustrations sur des données réelles.

4.1.5. Théorie asymptotique.

4.2. Analyse discriminante, classification.

Motivation, définition et calcul

Supposons qu'on a un p-vecteur aléatoire X. Nous aspirons à transformer X en un nouveau vecteur Y de dimension inférieure q (de préférence q est bien plus petit que p). Cette transformation devrait être opérée en minimisant la perte d'"information contenue" dans le vecteur original X.

$\Longrightarrow$ *réduction de la dimension de données*

En fin de compte nous espérons que Y soit plus facile à interpréter que X:

$\Longrightarrow$ *meilleure interprétation*

Motivation, définition et calcul

Exemple. Supposons qu'on ait sauvegardé pour une étude financière plusieurs centaines de prix d'actions sur base quotidienne (p.ex. S&P 500 contient 500 actions!) Il s'avère très difficile de tirer des conclusions statistiques à partir de vecteurs si larges. Comme bon nombre des prix d'actions sont fortement corrélés, on pourrait songer à en enlever une certaine quantité. Cependant, chaque prix peut contenir à lui seul des informations précieuses sur le marché, voilà pourquoi on ne peut pas les enlever de façon arbitraire!

Motivation, définition et calcul

Afin de "comprimer" le vecteur X nous allons utiliser une fonction H:

$$X \mapsto H(X) = Y, \quad Y \in \mathbb{R}^q,$$

où $q \leq p$. Souvent il est désirable que q soit beaucoup plus petit que p: $q \ll p$.

L'approche la plus simple consiste à choisir pour H une fonction linéaire, et alors $Y = HX$ avec $H \in \mathbb{R}^{q \times p}$.

Question: quel pourrait être un choix intelligent pour H?

Motivation, définition et calcul

L'ACP utilise la stratégie suivante: soit $H = (h_1, \ldots, h_q)'$.

(1) Commençons par $q = 1$ et choisissons h_1 tel que

$$\mathrm{Var}(Y_1) = \mathrm{Var}(h_1' X)$$

soit maximal sous la constrainte $\|h_1\| = 1$. (Sans contrainte ceci ne fait aucun sens!)

$$h_1 = \mathrm{argmax} \left\{ \mathrm{Var}(h' X) : h \in \mathbb{R}^p, \, \|h\| = 1 \right\}.$$

Remarquons que

$$\mathrm{Var}(h_1' X) = \mathrm{Var}(\|\langle h_1, X \rangle h_1\|).$$

En d'autres mots: projetons X sur le sous-espace 1-dimensionnel L_1 qui donne la plus grande (possible) variabilité pour la norme de la projection.

Motivation, définition et calcul

(2) Pour $q = 2$ choisissons h_1 comme avant et définissons

$$h_2 = \operatorname{argmax} \left\{ \operatorname{Var}(h'X) : h \in \mathbb{R}^p,\ \|h\| = 1, h \perp h_1 \right\}.$$

En d'autres mots: projetons X sur le sous-espace 1-dimensionnel L_2 qui donne la plus grande (possible) variabilité pour la norme de la projection, sous la contrainte que $L_2 \perp L_1$.

(3) Pour $q = 3$ choisissons h_1 et h_2 comme avant et définissons

$$h_3 = \operatorname{argmax} \left\{ \operatorname{Var}(h'X) : h \in \mathbb{R}^p,\ \|h\| = 1, h \perp h_1, h_2 \right\}.$$

(4) Continuons de la même manière pour $q = 4, \ldots, p$. Nous appelons $Y_i = h_i'X$ la *i-ème composante principale* de X.

Motivation, définition et calcul

Remarquons que les Y_i ne dépendent pas du nombre de composantes que nous voulons inclure. D'où le fait que nous pouvons toujours supposer que $H \in \mathbb{R}^{p \times p}$, $Y \in \mathbb{R}^p$ et que par après nous rejetons les composantes Y_i, $i > q$.

En fin de compte nous avons

$$Y = HX,$$

où H est une matrice orthogonale, et donc H effectue une rotation ou bien une réflection de nos données.

Une première question est de savoir comment implémenter H en pratique.

Motivation, définition et calcul

Puisque $\mathrm{Var}(v'X) = v'\Sigma v$, la détermination de H sera facile à l'aide des deux lemmes suivants.

Lemme

Soit Σ une matrice symétrique et définie positive, de valeurs propres $\lambda_1 > \lambda_2 > \ldots > \lambda_p (> 0)$ et soient $e_1, e_2, \ldots, e_p$ les vecteurs propres associés. Alors

$$\max \left\{ v'\Sigma v : \ v \in \mathbb{R}^p, \|v\| = 1 \right\} = \lambda_1,$$

et

$$\mathrm{argmax} \left\{ v'\Sigma v : \ v \in \mathbb{R}^p, \|v\| = 1 \right\} = e_1.$$

Remarque. Ceci reste vrai si nous demandons seulement que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Néanmoins le terme maximisant n'est alors plus unique.

Preuve.

Les vecteurs propres forment une base orthonormée de $\mathbb{R}^p$ et donc nous pouvons exprimer tout vecteur v comme $v = \sum_{i=1}^{p} \alpha_i e_i$. La condition $\|v\| = 1$ est équivalente par Pythagore à

$$1 = \|v\|^2 = \sum_{i=1}^{p} \|\alpha_i e_i\|^2 = \sum_{i=1}^{p} \alpha_i^2.$$

Remarquons que

$$\Sigma v = \Sigma \left(\sum_{i=1}^{p} \alpha_i e_i \right) = \sum_{i=1}^{p} \alpha_i \Sigma e_i = \sum_{i=1}^{p} \lambda_i \alpha_i e_i$$

et donc par l'orthonormalité des vecteurs propres il suit que

$$v' \Sigma v = \left(\sum_{i=1}^{p} \alpha_i e_i \right)' \sum_{i=1}^{p} \lambda_i \alpha_i e_i = \sum_{i=1}^{p} \lambda_i \alpha_i^2 \underbrace{e_i' e_i}_{=1}.$$

Clairement, par la contrainte $\sum_{i=1}^{p} \alpha_i^2 = 1$, cette somme est maximisée si $\alpha_1 = 1$ et $\alpha_i = 0$ pour $i = 2, \ldots, p$. Le maximum est λ_1 et le maximisant est e_1. (Discuter de l'unicité.) □

Lemme

Soit Σ une matrice symétrique et définie positive, de valeurs propres $\lambda_1 > \lambda_2 > \ldots > \lambda_p (> 0)$ et soient $e_1, e_2, \ldots, e_p$ les vecteurs propres associés. Alors pour $k = 2, \ldots, p$

$$\max \left\{ v' \Sigma v : \ v \in \mathbb{R}^p, \|v\| = 1, v \perp e_1, \ldots, e_{k-1} \right\} = \lambda_k,$$

et

$$\operatorname{argmax} \left\{ v' \Sigma v : \ v \in \mathbb{R}^p, \|v\| = 1, v \perp e_1, \ldots, e_{k-1} \right\} = e_k.$$

Preuve. La preuve est presque la même. Il nous suffit juste de remarquer que les conditions

$$v \perp e_1, \ldots, e_{k-1} \quad \text{et} \quad \|v\| = 1$$

impliquent que $v = \sum_{i=k}^{p} \alpha_i e_i$ avec $\sum_{i=k}^{p} \alpha_i^2 = 1$. □

Motivation, définition et calcul

Il est commun de centrer d'abord les données avant que les composantes principales ne soient implémentées. Cela mène à la définition suivante:

Définition

Soit X un p-vecteur aléatoire de moyenne μ et de matrice de variance-covariance définie positive Σ. Alors la transformation

$$X \mapsto Y = P_{tr}(X) = H(X - \mu)$$

est appelée transformation des composantes principales.

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.1.1. Motivation, définition et calcul.

4.1.2. Propriétés et interprétation.

4.1.3. Version empirique.

4.1.4. Illustrations sur des données réelles.

4.1.5. Théorie asymptotique.

4.2. Analyse discriminante, classification.

Proposition

Soit Σ une matrice de variance-covariance non-singulière d'un certain p-vecteur aléatoire X et supposons que $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ sont ses valeurs propres et que $O = (e_1, \ldots, e_p)$ est la matrice des vecteurs propres correspondants. Soit Y le vecteur des composantes principales de X. Alors

$$(i) \quad H = O',$$

$$(ii) \quad \mathrm{Var}(Y) = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p),$$

$$(iii) \quad \mathrm{tr}(\Lambda) = \mathrm{tr}(\Sigma) = E\|X - \mu\|^2,$$

$$(iv) \quad |\Lambda| = |\Sigma|.$$

Propriétés et interprétation

Preuve. Nous avons déjà démontré (i). Le théorème spectral nous donne

$$\mathrm{Var}(Y) = \mathrm{Var}(HX) = H\Sigma H' = HO\Lambda O' H' = \Lambda,$$

ce qui prouve (ii). De même,

$$\mathrm{tr}(\Sigma) = \mathrm{tr}(O\Lambda O') = \mathrm{tr}(O' O\Lambda) = \mathrm{tr}(\Lambda),$$

et

$$|\Sigma| = |O\Lambda O'| = |O||\Lambda||O'| = |\Lambda|,$$

ce qui établit (iii) et (iv). □

Propriétés et interprétation

Les composantes principales sont le plus facilement intelligibles dans le contexte d'un p-vecteur aléatoire $X \sim \mathcal{N}_p(\mu, \Sigma)$. Souvenez-vous que les contours de la densité de X sont donnés par

$$\{x | \, (x-\mu)' \Sigma^{-1} (x-\mu) = c^2\}.$$

Vous avez vu/verrez aux exercices que ces contours forment des ellipsoïdes de centre μ et d'axes principaux $e_1, \ldots, e_p$, où les e_i sont des vecteurs propres de Σ. Les rayons de ces axes sont $c\sqrt{\lambda_i}$.

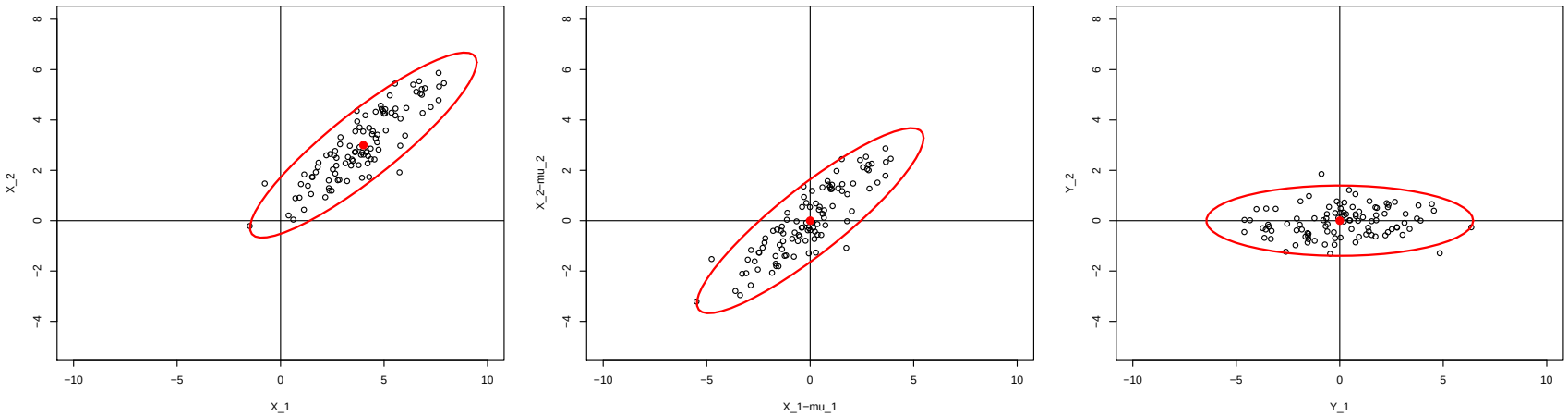
Les composantes principales $Y = H(X-\mu) \sim \mathcal{N}_p(0, \Lambda)$, et de ce fait les nouveaux contours correspondent à

$$\{y | \, y' \Lambda^{-1} y = c^2\}.$$

Nous avons donc des ellipsoïdes de centre 0, des axes principaux canoniques et les rayons sont les mêmes que précédemment.

Propriétés et interprétation

La figure ci-dessous montre la transformation des composantes principales appliquée à 100 données/observations issues d'une loi $\mathcal{N}_p(\mu, \Sigma)$ avec $\mu = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$ et $\Sigma = \begin{pmatrix} 5 & 3 \\ 3 & 2.25 \end{pmatrix}$.



Propriétés et interprétation

Jusqu'à présent nous n'avons pas encore retiré d'observations de notre jeu de données. Nous pouvons toujours réobtenir X tout entier à partir de Y:

$$X = H'Y + \mu. \tag{1}$$

Nous souhaitons remplacer le vecteur $Y = (Y_1, \ldots, Y_p)$ par un vecteur de plus petite dimension $Y^{(q)} = (Y_1, \ldots, Y_q)$, $q \leq p$.

Comme $H' = (e_1, \ldots, e_p)$ nous obtenons de (1) que

$$X - \mu = Y_1 e_1 + Y_2 e_2 + \cdots + Y_p e_p.$$

Si nous gardons seulement $Y_1, \ldots, Y_q$, nous pourrions utiliser l'approximation

$$X - \mu \approx Y_1 e_1 + Y_2 e_2 + \cdots + Y_q e_q.$$

Propriétés et interprétation

La proposition suivante montre que l'approximation

$$X - \mu \approx Y_1 e_1 + Y_2 e_2 + \cdots + Y_q e_q.$$

est "la meilleure" en termes de q vecteurs de base

Proposition

Soit $B = (b_1, \ldots, b_p)$ *une quelconque base orthonormale de* $\mathbb{R}^p$. *Alors*

$$X - \mu = W_1 b_1 + W_2 b_2 + \cdots + W_p b_p,$$

et pour tout $1 \leq q \leq p$

$$E \left\| (X - \mu) - \left(\sum_{i=1}^{q} W_i b_i \right) \right\|^2 \geq E \left\| (X - \mu) - \left(\sum_{i=1}^{q} Y_i e_i \right) \right\|^2 .$$

Propriétés et interprétation

Cette proposition montre que les composantes principales retiennent “le maximum" de l'information contenue dans le vecteur original X. Aucune autre base orthonormale jouit d'une meilleure puissance d'approximation que la fonction propre de Σ.

Notons que

$$E\left\|(X-\mu)-\left(\sum_{i=1}^{q} Y_i e_i\right)\right\|^2 = \lambda_{q+1}+\ldots+\lambda_p.$$

Le ratio

$$\frac{\lambda_{q+1}+\ldots+\lambda_p}{\lambda_1+\ldots+\lambda_p} = \frac{\lambda_{q+1}+\ldots+\lambda_p}{\mathrm{tr}(\Sigma)} \tag{2}$$

peut donc être vu comme une mesure de la quantité d'information expliquée par les q premières composantes principales. On dit que (2) est *la part de la variabilité de X expliquée par les q premières CP.*

Propriétés et interprétation

Preuve. Par simplicité, prenons $\mu = 0$ (sans perte de généralité).Tout d'abord, il est clair que X peut être réécrit sous la forme $\sum_{i=1}^{p} W_i b_i$. Cela suit du fait que B forme une base orthonormale. Remarquons aussi que $W_i = b_i' X$ (puisque la transformation des composantes principales nous dit que $W = B'X$).

On sait que

Maintenant, par le théorème de Pythagore, nous avons

$$E \left\| (X - \left(\sum_{i=1}^{q} W_i b_i \right) \right\|^2 = E \left\| \sum_{i=q+1}^{p} W_i b_i \right\|^2$$

$$= \sum_{i=q+1}^{p} E W_i^2 \underbrace{\|b_i\|^2}_{=1} = \sum_{i=q+1}^{p} b_i' \Sigma b_i = \mathrm{tr}(\tilde{B}' \Sigma \tilde{B}),$$

où $\tilde{B} = (b_{q+1}, \ldots, b_p)$ est une matrice $p \times (p - q)$ orthonormale.

Propriétés et interprétation

Nous cherchons donc une matrice B^* qui minimise $\mathrm{tr}(\tilde{B}'\Sigma\tilde{B})$ sur l'ensemble des matrices $(p-q)\times q$ orthonormales $\tilde{B}$. Comme les e_i forment une base de $\mathbb{R}^p$, nous avons que $\tilde{B} = HC$ où $C = (c_{jk})$ est $p \times (p-q)$. We have that

$$\mathrm{tr}(\tilde{B}'\Sigma\tilde{B}) = \mathrm{tr}(C'\Lambda C) = \sum_{j=1}^{p} \lambda_j \left(\sum_{k=1}^{p-q} c_{jk}^2 \right). \tag{3}$$

Nous avons que $C = \tilde{B}'H$ est telle que $CC' = \tilde{B}'HH'\tilde{B} = I_{p-q}$ (orthonormale). Puisque les valeurs propres sont supposées être bien ordonnées, la matrice orthonormale C qui minimise (3) est obtenue en prenant les vecteurs propres associés dans H associés aux plus petites valeurs propres.

□

Propriétés et interprétation

Exemple. Supposons que $X = (X_1, X_2, X_3)$ est de moyenne nulle et de matrice de variance-covariance

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Alors les valeurs propres sont

$$(\lambda_1, \lambda_2, \lambda_3) = (5.83, 2, 0.17)$$

avec comme vecteurs propres associés

$$e_1 = \begin{pmatrix} 0.383 \\ -0.924 \\ 0 \end{pmatrix} \quad e_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad e_3 = \begin{pmatrix} 0.923 \\ -0.383 \\ 0 \end{pmatrix}.$$

Propriétés et interprétation

Cela nous donne

$$Y_1 = 0.383X_1 - 0.923X_2$$
$$Y_2 = X_3$$
$$Y_3 = 0.924X_1 + 0.383X_2$$

Les coordonnées des vecteurs propres nous révèlent combien de poids est attribué à chaque X_i !

Dans cet exemple la proportion de la variance totale expliquée par les premières CP équivaut à $\lambda_1/\mathrm{tr}(\Sigma) = 5.83/8 = 0.73$, celle expliquée par les 2 premières CP correspond à $(\lambda_1 + \lambda_2)/\mathrm{tr}(\Sigma)$
$= (5.83 + 2)/8 = 0.98$.

Propriétés et interprétation

Une propriété défavorable de l'ACP est que cette méthode *n'est pas échelle-invariante*. Cela veut dire que si A est une matrice d'échelle et si P_{tr} représente la transformation des composantes principales, alors

$$P_{tr}(AX) \neq AP_{tr}(X).$$

Exemple. Soient $EX = 0$ et $\mathrm{Var}(X) = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$. Les valeurs propres sont $\lambda_1 = 100.16$ and $\lambda_2 = 0.84$ avec comme vecteurs propres associés

$$e_1 = \begin{pmatrix} 0.04 \\ 0.999 \end{pmatrix} \quad \text{et} \quad e_2 = \begin{pmatrix} 0.999 \\ -0.04 \end{pmatrix}.$$

Nous avons $\lambda_1/\mathrm{tr}(\Sigma) = 0.992$.

La grande variance de X_2 domine complètement la première composante principale obtenue à partir de Σ!

Propriétés et interprétation

Nous utilisons maintenant le même scénario, mais nous standardisons d'abord X en divisant ses composantes par leur dispersion. Cela veut dire que nous utilisons la matrice de corrélation

$$P = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

afin de déterminer les composantes principales. Les valeurs propres sont maintenant $\lambda_1 = 1.4$ and $\lambda_2 = 0.6$ avec comme vecteurs propres associés

$$e_1 = \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix} \quad \text{et} \quad e_2 = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}.$$

Nous avons $\lambda_1/\mathrm{tr}(P) = 0.7$.

Nous attribuons donc la même importance à X_1 et à X_2 lors du calcul de Y_1.

Propriétés et interprétation

Cet exemple montre que l'ACP doit être utilisée avec prudence.

L'ACP peut être utilisée au mieux si toutes les variables se trouvent à une échelle comparable.

Les variables devraient donc être proprement standardisées si elles sont mesurées sur des échelles fortement différentes.

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.1.1. Motivation, définition et calcul.

4.1.2. Propriétés et interprétation.

4.1.3. Version empirique.

4.1.4. Illustrations sur des données réelles.

4.1.5. Théorie asymptotique.

4.2. Analyse discriminante, classification.

Version empirique

Bien sûr, en pratique, nous ne connaissons ni μ ni Σ et par conséquent nous devons baser nos analyses sur leurs estimateurs

$$\bar{X} \quad \text{et} \quad S,$$

basés eux-mêmes sur l'échantillon $X_1, \ldots, X_n$.

Soient $\hat{H} = (\hat{e}_1, \ldots, \hat{e}_p)$ les vecteurs propres standardisés de S correspondant aux valeurs propres $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$. La transformation des composantes principales empiriques est donnée par

$$X_k \mapsto \hat{P}_{tr}(X_k) = \hat{Y}_k = \hat{H}(X_k - \bar{X}), \quad 1 \leq k \leq n.$$

Posant $\hat{Y}_k = (\hat{Y}_{k,1}, \ldots, \hat{Y}_{k,p})'$, les variables $\hat{Y}_{k,i}$ sont appelées les i-èmes scores des composantes principales.

Version empirique

Proposition

Soit S une matrice de variance-covariance non-singulière correspondant à un certain échantillon $X_1, \ldots, X_n$ de p-vecteurs. Supposons que $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_p$ soient ses valeurs propres et que $\hat{O} = (\hat{e}_1, \ldots, \hat{e}_p)$ soit la matrice des vecteurs propres correspondants. Soient $Y_1, \ldots, Y_n$ les composantes principales. Alors

$$(i) \quad \hat{H} = \hat{O}',$$

$$(ii) \quad S_Y = \hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p),$$

$$(iii) \quad \mathrm{tr}(\hat{\Lambda}) = \mathrm{tr}(S) = \frac{1}{n-1} \sum_{k=1}^{n} \|X_k - \bar{X}\|^2,$$

$$(iv) \quad |\hat{\Lambda}| = |S|.$$

Preuve. *Exercice.* □

Version empirique

Comme dans le cas de la population il est possible d'établir une propriété d'optimalité des fonctions propres.

Proposition

Soit $B = (b_1, \ldots, b_p)$ *une base orthonormale de* $\mathbb{R}^p$. *Alors*

$$X_k - \bar{X} = W_{k,1} b_1 + W_{k,2} b_2 + \cdots + W_{k,p} b_p,$$

et pour tout $1 \leq q \leq p$

$$\sum_{k=1}^{n} \left\| (X_k - \bar{X}) - \left(\sum_{i=1}^{q} W_{k,i} b_i \right) \right\|^2 \geq \sum_{k=1}^{n} \left\| (X_k - \bar{X}) - \left(\sum_{i=1}^{q} \hat{Y}_{k,i} \hat{e}_i \right) \right\|^2 .$$

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.1.1. Motivation, définition et calcul.

4.1.2. Propriétés et interprétation.

4.1.3. Version empirique.

4.1.4. Illustrations sur des données réelles.

4.1.5. Théorie asymptotique.

4.2. Analyse discriminante, classification.

Illustration sur des données réelles

Nous analysons les données de records nationaux de courses pour 54 pays en utilisant le logiciel R.

```
> track<-read.table("T8-6.dat")
> track

Country 100m 200m  400m  800m  1500m  5000m  10000m  Marathon
1 Arg 10.23 20.37 46.18  1.77   3.68  13.33   27.65   129.57
2 Aus  9.93 20.06 44.38  1.74   3.53  12.93   27.53   127.51
3 Aut 10.15 20.45 45.80  1.77   3.58  13.26   27.72   132.22
4 Bel 10.14 20.19 45.02  1.73   3.57  12.83   26.87   127.20
...
```

Afin d'obtenir un feeling pour les dépendances nous faisons une matrice de "scatterplot".

```
> plot(track[2:9])
```

Illustration sur des données réelles

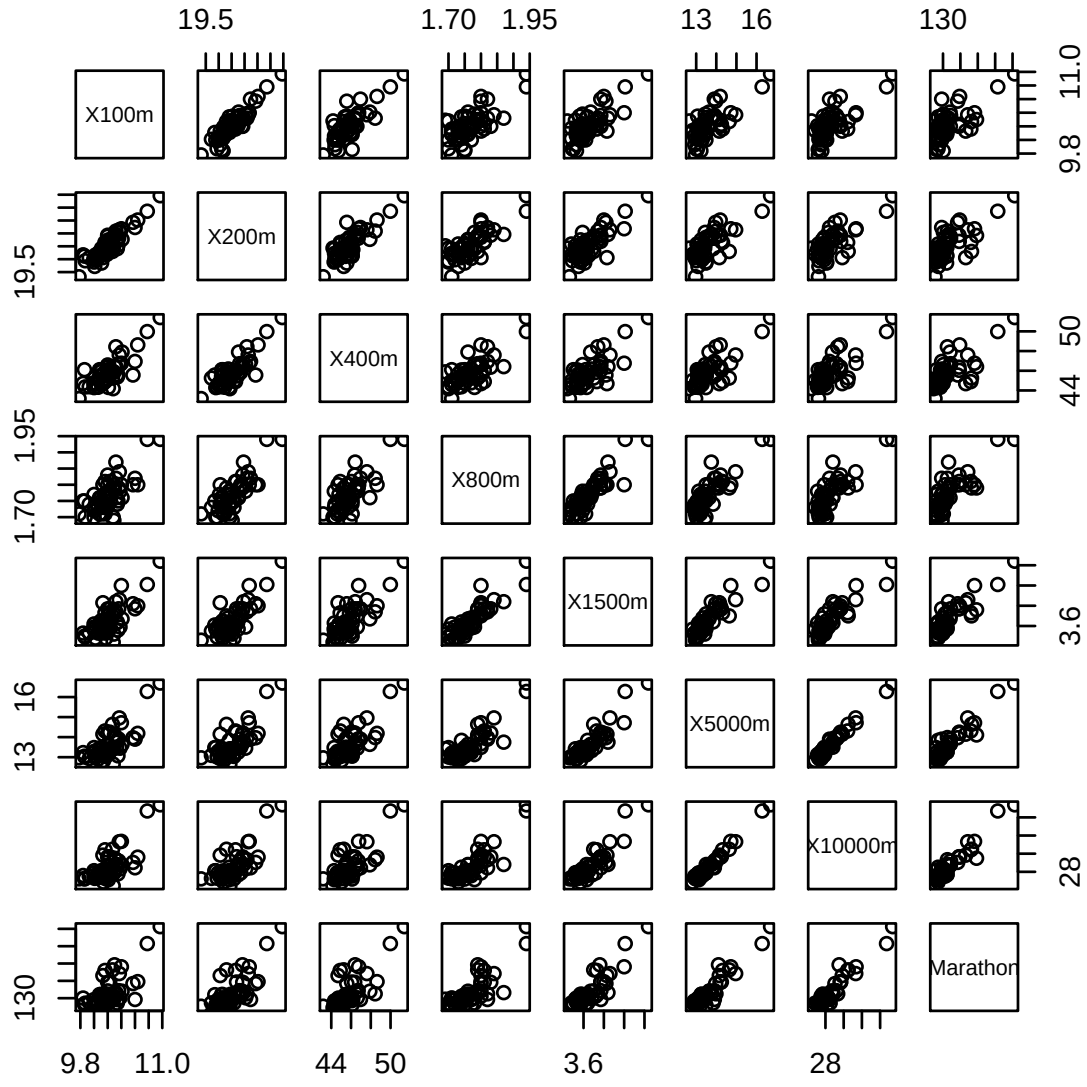


Illustration sur des données réelles

Nous commençons par une ACP des données originales

```
> pr=prcomp(track[2:9])
> pr
Standard deviations:
[1] 9.19278216 1.06833766 0.47829361 ...

Rotation:
                   PC1         PC2         PC3
X100m    -0.016521364  0.09994775  0.00805890 ...
X200m    -0.043603961  0.25282207  0.08082078 ...
X400m    -0.113581901  0.91633442  0.25345836 ...
X800m    -0.004643738  0.01405243 -0.01226878 ...
X1500m   -0.014583470  0.03076566 -0.03706409 ...
X5000m   -0.078593405  0.11653932 -0.37676524 ...
X10000m  -0.175189411  0.20893763 -0.87277505 ...
Marathon -0.973561706 -0.16745845  0.15475565 ...
```

Illustration sur des données réelles

Ces chiffres révèlent que les 2 premières CP expliquent quasiment toute la variance.

```
> summary(pr)
Importance of components:
                          PC1    PC2     PC3 ...
Standard deviation     9.193 1.0683 0.47829 ...
Proportion of Variance 0.983 0.0133 0.00266 ...
Cumulative Proportion  0.983 0.9961 0.99881 ...
```

Cependant, les CP ne sont pas faciles à interpréter. Une raison pourrait être que, avec la longueur de la course, l'échelle change dramatiquement. Il n'est donc pas surprenant que la contribution du marathon domine la valeur des premiers scores des CP.

Illustration sur des données réelles

Nous faisons à présent une ACP sur base de données standardisées.

```
> pr=prcomp(track[2:9],scale.=TRUE)
> pr
Standard deviations:
[1] 2.58907125 0.79900570 0.47699528 ...

Rotation:
                PC1         PC2          PC3
X100m    -0.3323877 -0.52939911 -0.343859303 ...
X200m    -0.3460511 -0.47039050  0.003786104 ...
X400m    -0.3391240 -0.34532929  0.067060507 ...
X800m    -0.3530134  0.08945523  0.782711152 ...
X1500m   -0.3659849  0.15365241  0.244270040 ...
X5000m   -0.3698204  0.29475985 -0.182863147 ...
X10000m  -0.3659489  0.33360619 -0.243980694 ...
Marathon -0.3542779  0.38656085 -0.334632969 ...
```

Illustration sur des données réelles

Ces chiffres révèlent que les 2 premières CP expliquent quasiment toute la variance.

```
> summary(pr)
Importance of components:
                          PC1    PC2    PC3
Standard deviation     2.589 0.7990 0.4770 ...
Proportion of Variance 0.838 0.0798 0.0284 ...
Cumulative Proportion  0.838 0.9177 0.9462 ...
```

Illustration sur des données réelles

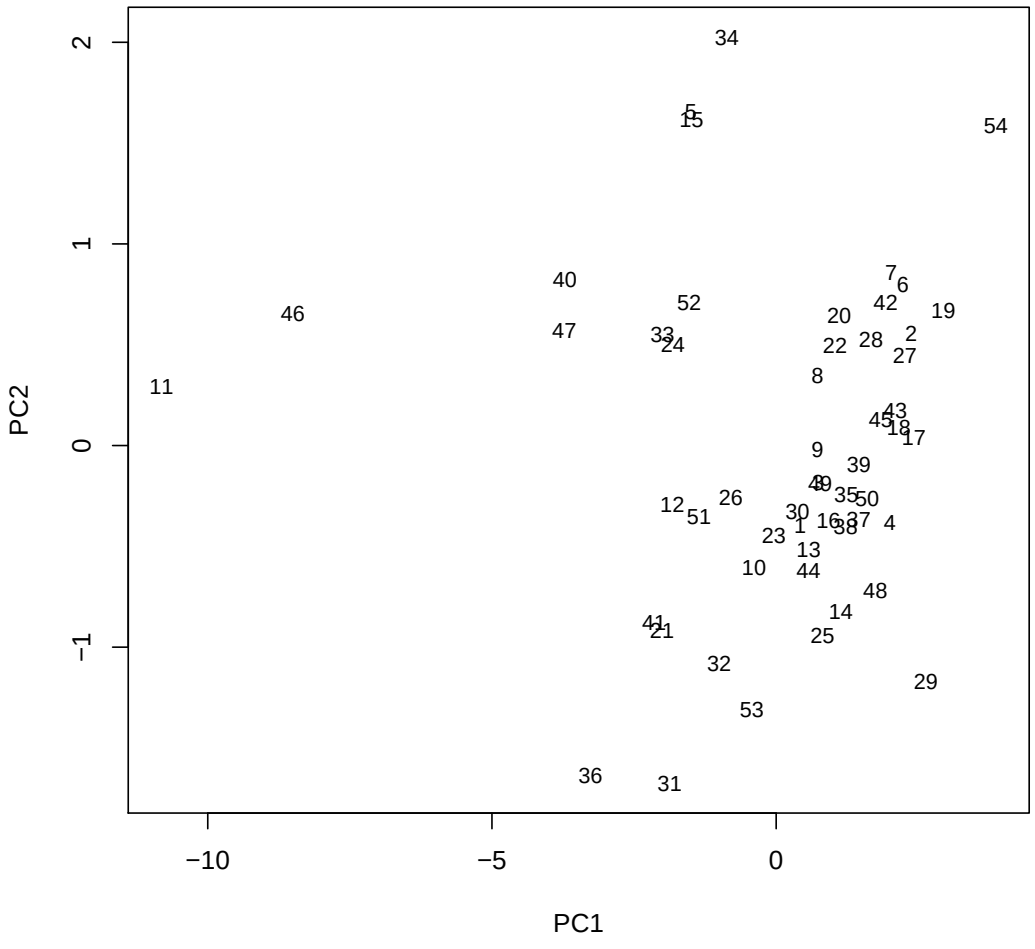
Interprétation.

La première CP est formée par une moyenne de toutes les disciplines. Si la première CP est petite/grande, la nation est mauvaise/bonne dans toutes les disciplines. La deuxième CP met en contraste les courses courtes contre les courses longues. Si cette deuxième CP est grande, cela veut dire qu'une nation est forte aux courses courtes, mais moins forte aux longues distances et vice-versa. La troisième CP met en contraste les distances moyennes contre les longues et courtes distances.

Illustration sur des données réelles

Un *bi-plot* peut s'avérer utile.

```
> library(sfsmisc)
> sco=princomp(track[2:9],cor=TRUE)$scores
> n.plot(sco[,1],sco[,2],cex=0.5,xlab="PC1",ylab="PC2")
```



54=USA
19=Great Britain
11=Cook Islands
46=Samoa
29=Kenyia
34=Mauritius
3=Austria
4=Belgium

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.1.1. Motivation, définition et calcul.

4.1.2. Propriétés et interprétation.

4.1.3. Version empirique.

4.1.4. Illustrations sur des données réelles.

4.1.5. Théorie asymptotique.

4.2. Analyse discriminante, classification.

Théorie asymptotique

Une question importante est de savoir si oui ou non les valeurs propres et vecteurs propres estimés, utilisés pour déterminer les CP, convergent vers les vrais valeurs propres et vecteurs propres.

De plus, il serait intéressant de voir si on peut obtenir la normalité asymptotique des estimateurs. Il n'est pas aisé de répondre à ces questions, et la plupart des résultats existants demandent un échantillon aléatoire issu d'une population normale.

Une hypothèse générale pour beaucoup de résultats est que

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p. \tag{4}$$

Ceci n'est pas très restrictif et assure essentiellement l'unicité de leurs estimateurs.

Théorème (Anderson (1963))

Soient $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, \Sigma)$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ *et* $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)'$. *Si* (4) *est satisfait, alors*

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \xrightarrow{\mathcal{D}} \mathcal{N}_p(0, 2\Lambda^2),$$

et

$$\sqrt{n}(\hat{e}_i - e_i) \xrightarrow{\mathcal{D}} \mathcal{N}_p(0, E_i), \quad 1 \leq i \leq p,$$

où

$$E_i = \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k e_k'.$$

Le théorème d'Anderson est assez difficile à montrer. Nous allons nous contenter de montrer la convergence, mais sous des hypothèses plus générales. Rappelons que la norme de Froebenius d'une matrice carrée A est donnée par

$$\|A\|_{\mathcal{F}}^2 = \mathrm{tr}(A^2) = \sum_{i=1}^{p} \|Ab_i\|^2,$$

où $b_1, \ldots, b_p$ est une base orthonormale. On que

$$\|A\|_{\mathcal{F}}^2 = \sum_{i,j=1}^{p} A_{ij}^2 \geq \|A\|_{\mathcal{L}}^2 = \sup_{x:\, \|x\|=1} \|Ax\|^2.$$

Théorème

Soit $\{X_i\}$ *un ensemble de p-vecteurs aléatoires iid avec* $E\|X_1\|^4 < \infty$ *et* $EX_1 = 0$. *Supposons que* (4) *soit satisfait. Soient* $(\hat{\lambda}_i, \hat{e}_i)$ *les estimateurs usuels et* $\hat{c}_i = \mathrm{sign}(\langle e_i, \hat{e}_i \rangle)$. *Alors*

$$\max_{1 \le i \le p} \limsup_{n \to \infty} \left\{ nE\|\hat{c}_i \hat{e}_i - e_i\|^2 + nE|\hat{\lambda}_i - \lambda_i|^2 \right\} < \infty.$$

Remarque. Quel est le rôle de $\hat{c}_i$?

Remarque. Le théorème implique que

$$\max_{1 \le i \le p} \{\|\hat{c}_i \hat{e}_i - e_i\|\} = O_P(1/\sqrt{n}),$$

et

$$\max_{1 \le i \le p} \left\{|\hat{\lambda}_i - \lambda_i|\right\} = O_P(1/\sqrt{n}).$$

Afin de simplifier la preuve, nous supposons dans ce qui suit que nos p-vecteurs sont de moyenne nulle et nous utilisons alors

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{k=1}^{n} X_k X_k'.$$

Le lemme suivant est crucial, le reste étant simplement de l'algèbre.

Lemme

Soit $\{X_i\}$ un ensemble de p-vecteurs aléatoires iid avec $E\|X_1\|^4 < \infty$ et $EX_1 = 0$. Alors $\forall n \geq 1$

$$nE\|\hat{\Sigma}_n - \Sigma\|_{\mathcal{F}}^2 = E\|X_1 X_1' - \Sigma\|_{\mathcal{F}}^2 < \infty.$$

Remarque. Il s'ensuit que $\|\hat{\Sigma}_n - \Sigma\|_{\mathcal{F}} = O_P(1/\sqrt{n})$.

Preuve. Pour une base orthonormale $e_1, \ldots, e_p$, nous avons

$$\begin{aligned}
\left\|\hat{\Sigma}_n - \Sigma\right\|_{\mathcal{F}}^2 &= \left\| \frac{1}{n} \sum_{k=1}^{n} (X_k X_k' - \Sigma) \right\|_{\mathcal{F}}^2 \\
&= \frac{1}{n^2} \sum_{i=1}^{p} \left\| \sum_{k=1}^{n} (X_k X_k' - \Sigma) e_i \right\|^2 \\
&= \frac{1}{n^2} \sum_{i=1}^{p} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \left\langle (X_k X_k' - \Sigma) e_i, (X_\ell X_\ell' - \Sigma) e_i \right\rangle .
\end{aligned}$$

L'indépendance des X_k implique que

$$E \left\|\hat{\Sigma}_n - \Sigma\right\|_{\mathcal{F}}^2 = \frac{1}{n^2} \sum_{i=1}^{p} \sum_{k=1}^{n} E \left\| (X_k X_k' - \Sigma) e_i \right\|^2 .$$

Comme les X_k sont identiquement distribués, il suit que

$$E\left\|\hat{\Sigma}_n - \Sigma\right\|_{\mathcal{F}}^2 = \frac{1}{n} E \sum_{i=1}^{p} \left\|(X_1 X_1' - \Sigma)e_i\right\|^2$$
$$= \frac{1}{n} E \left\|X_1 X_1' - \Sigma\right\|_{\mathcal{F}}^2 .$$

En prenant les propriétes de $\|\cdot\|_{\mathcal{F}}$, nous voyons que

$$E\left\|X_1 X_1' - \Sigma\right\|_{\mathcal{F}}^2 = \sum_{j=1}^{p}\sum_{i=1}^{p} E\left[(X_{1i}X_{1j} - \Sigma_{ij})^2\right]$$
$$\leq \sum_{j=1}^{p}\sum_{i=1}^{p} E X_{1i}^2 X_{1j}^2$$
$$\leq \sum_{j=1}^{p}\sum_{i=1}^{p} \sqrt{E(X_{1i}^4)E(X_{1j}^4)} < \infty. \qquad \square$$

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.2. Analyse discriminante, classification.

4.2.1. Considérations générales.

4.2.2. Classification pour deux populations normales.

4.2.3. Classification pour $m \geq 3$ populations.

4.2.4. Illustrations sur des données réelles.

Considérations générales

L'analyse discriminante et la classification ont pour objectifs de séparer en différents groupes des objets appartenant à diverses populations (par la construction de "discriminants", c'est-à-dire de quantités numériques qui différeront autant que possible d'une population à l'autre), et de définir des règles de classification qui permettront de "prédire" à quelle population appartient un nouvel objet.

Remarque: les procédures que nous allons construire (qui tentent bien entendu de minimiser les cas de misclassification) devront aussi prendre en compte les probabilités à priori qu'une nouvelle observation à classifier provienne de la population i ($i = 1, \ldots, m$), et des coûts de misclassification, qui peuvent ne pas être symétriques (exemple).

Considérations générales

Considérons deux populations π_1 et π_2, qui sont associées à des lois de probabilité (sur $\mathbb{R}^p$) absolument continues de densité f_1 et f_2, respectivement.

Ces deux populations diffèrent par leur position, leur dispersion, ou toute autre caractéristique.

Le problème que nous considérons est le suivant:

sur base de la valeur $x = (x_1, \ldots, x_p)'$ prise par un p-v.a. $X = (X_1, \ldots, X_p)'$ provenant de π_1 ou de π_2 (i.e., $X \sim f_1$ ou $X \sim f_2$), comment parier de manière raisonnable sur la population dont X est issu?

$\rightsquigarrow$ Une règle de classification consiste à donner une partition (R_1, R_2) de l'ensemble χ des valeurs possibles pour x, telle que

- X sera classifié en π_1 si $x \in R_1$ et
- X sera classifié en π_2 si $x \in R_2$.

Considérations générales

Si on se donne des probabilités à priori p_1, p_2 que X provienne respectivement de π_1 et π_2, la probabilité P de misclassification de X est donnée par

$$P = p_{2|1} \times p_1 + p_{1|2} \times p_2,$$

où

$$p_{i|j} = P[X \in R_i | X \in \pi_j] = \int_{R_i} f_j(x)\, dx.$$

Une procédure de classification (R_1, R_2) optimale veillera à minimiser P.

Considérations générales

Si en outre différents coûts de misclassification doivent être pris en compte (notons $c_{1|2}$ et $c_{2|1}$ respectivement les coûts si on classifie en π_1 un objet provenant de π_2 et si on classifie en π_2 un objet provenant de π_1), une procédure de classification (R_1, R_2) optimale devra cette fois minimiser le coût de misclassification moyen

$$\begin{aligned} E_{\text{coût}} &= E[\text{coût}|X \in \pi_1]\, p_1 + \mathrm{E}[\text{coût}|X \in \pi_2]\, p_2 \\ &= \left(0 \times p_{1|1} + c_{2|1}\, p_{2|1}\right) p_1 + \left(c_{1|2}\, p_{1|2} + 0 \times p_{2|2}\right) p_2 \\ &= c_{2|1}\, p_{2|1}\, p_1 + c_{1|2}\, p_{1|2}\, p_2. \end{aligned}$$

Considérations générales

Le résultat suivant décrit la procédure de classification (R_1, R_2) optimale dans ce contexte général:

Théorème: *la procédure de classification (R_1, R_2) qui minimise le coût de misclassification moyen est donnée par*

$$R_1 = \left\{ x \in \chi \;\middle|\; \frac{f_1(x)}{f_2(x)} \geq \frac{c_{1|2}\, p_2}{c_{2|1}\, p_1} \right\} \quad et \quad R_2 = \chi \backslash R_1.$$

Preuve. Le coût moyen $E_{\text{coût}}$ sera minimal si l'intégrande de

$$\begin{aligned} E_{\text{coût}} &= c_{2|1}\, p_{2|1}\, p_1 + c_{1|2}\, p_{1|2}\, p_2 \\ &= c_{2|1}\,(1 - p_{1|1})\, p_1 + c_{1|2}\, p_{1|2}\, p_2 \\ &= c_{2|1}\, p_1 + (c_{1|2}\, p_{1|2}\, p_2 - c_{2|1}\, p_{1|1}\, p_1) \\ &= c_{2|1}\, p_1 + \int_{R_1} (c_{1|2}\, f_2(x)\, p_2 - c_{2|1}\, f_1(x)\, p_1)\, dx \end{aligned}$$

est négative pour tout $x \in R_1$. □

Remarque. Il suffit de connaître les rapports $\frac{f_1}{f_2}$, $\frac{c_{1|2}}{c_{2|1}}$ et $\frac{p_2}{p_1}$ pour déterminer la règle de classification optimale.

Considérations générales

Quelques cas particuliers:

- Si les probabilités p_1, p_2 sont égales ou sont inconnues,

$$R_1 = \left\{ x \in \chi \,\middle|\, \frac{f_1(x)}{f_2(x)} \geq \frac{c_{1|2}}{c_{2|1}} \right\} \quad \text{et} \quad R_2 = \chi \backslash R_1.$$

- Si les coûts $c_{1|2}$, $c_{2|1}$ sont égaux ou non spécifiés,

$$R_1 = \left\{ x \in \chi \,\middle|\, \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \right\} \quad \text{et} \quad R_2 = \chi \backslash R_1.$$

- Si p_1, p_2, $c_{1|2}$, $c_{2|1}$ sont non spécifiés,

$$R_1 = \left\{ x \in \chi \,\middle|\, \frac{f_1(x)}{f_2(x)} \geq 1 \right\} \quad \text{et} \quad R_2 = \chi \backslash R_1.$$

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.2. Analyse discriminante, classification.

4.2.1. Considérations générales.

4.2.2. Classification pour deux populations normales.

4.2.3. Classification pour $m \geq 3$ populations.

4.2.4. Illustrations sur des données réelles.

Classification pour deux normales

Considérons d'abord le cas où $\pi_i = \mathcal{N}_p(\mu_i, \Sigma_i)$, $i = 1, 2$, avec $\Sigma_1 = \Sigma_2 (=: \Sigma)$.

$\rightsquigarrow$ **Proposition**: soit $a := \Sigma^{-1}(\mu_1 - \mu_2)$. Alors la procédure de classification optimale classifie x en π_1 si

$$a'x \geq \frac{1}{2} a'(\mu_1 + \mu_2) + \ln \left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1} \right]$$

et en π_2 sinon.

Preuve. □

Classification pour deux normales

Remarque. Si $\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1} = 1$, il faut donc classifier x en π_1 ssi

$$a'x \geq \frac{1}{2} a'(\mu_1 + \mu_2).$$

On voit facilement par la définition de a que

$$a'\mu_1 \geq \frac{1}{2} a'(\mu_1 + \mu_2) \quad \text{et} \quad a'\mu_2 \leq \frac{1}{2} a'(\mu_1 + \mu_2).$$

Il faut donc classifier x en π_1 si la projection de x sur l'axe a est plus proche de celle de μ_1 que de celle de μ_2.

Classification pour deux normales

De manière équivalente, il faut classifier x en π_1 ssi

$$2\left[a'x - \frac{1}{2}a'(\mu_1 + \mu_2)\right] = d^2_\Sigma(x, \mu_2) - d^2_\Sigma(x, \mu_1) \geq 0.$$

Dans le cas général, la règle de classification compare donc les projections de x, de μ_1 et de μ_2 sur l'axe a et classifie x en π_1, en fonction des coûts et des probabilités à priori, si

$$2\left[a'x - \frac{1}{2}a'(\mu_1 + \mu_2)\right] = d^2_\Sigma(x, \mu_2) - d^2_\Sigma(x, \mu_1) \geq 2\ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right].$$

Classification pour deux normales

La fonction $x \mapsto a'x - \frac{1}{2}a'(\mu_1 + \mu_2) - \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right]$ sur laquelle est fondée la règle de classification est appelée fonction discriminante linéaire de Fisher . On parlera *d'analyse discriminante linéaire.*

En pratique, μ_1, μ_2 et Σ sont inconnus et il faut les estimer sur base d'un échantillon ("training sample") d'observations indépendantes

$$X_1, \ldots, X_{m_1} \sim \mathcal{N}_p(\mu_1, \Sigma), \quad Y_1, \ldots, Y_{m_2} \sim \mathcal{N}_p(\mu_2, \Sigma).$$

En utilisant les notations dans la Section 3.1.4, la règle empirique consiste alors à classifier x en π_1 plutôt qu'en π_2 ssi

$$(\bar{X} - \bar{Y})' S_{\text{pool}}^{-1} x \geq \frac{1}{2}(\bar{X} - \bar{Y})' S_{\text{pool}}^{-1}(\bar{X} + \bar{Y}) + \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right].$$

Classification pour deux normales

Si on considére plutôt le cas où $\pi_i = \mathcal{N}_p(\mu_i, \Sigma_i)$, $i = 1, 2$, sans faire l'hypothèse d'égalité des covariances, on obtient alors (en procédant comme ci-dessus) le résultat suivant (exercice):

$\rightsquigarrow$ **Proposition**: soit

$$k := \ln\left[\frac{|\Sigma_1|}{|\Sigma_2|}\right] + (\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2).$$

Alors la procédure de classification optimale classifie x en π_1 si

$$-\frac{1}{2}x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x \geq \frac{k}{2} + \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right]$$

et en π_2 sinon.

Remarque. Pour des raisons évidentes, on parlera cette fois de règle de classification (et donc d'analyse discriminante) quadratique .

Classification pour deux normales

Si on dispose d'un échantillon d'observations indépendantes

$$X_1, \ldots, X_{m_1} \sim \mathcal{N}_p(\mu_1, \Sigma_1), \quad Y_1, \ldots, Y_{m_2} \sim \mathcal{N}_p(\mu_2, \Sigma_2),$$

la règle empirique classifie dans ce cas x en π_1 ssi

$$-\frac{1}{2}x'(S_x^{-1} - S_y^{-1})x + (\bar{X}'S_x^{-1} - \bar{Y}'S_y^{-1})x \geq \frac{\hat{k}}{2} + \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right],$$

où

$$\hat{k} := \ln\left[\frac{|S_x|}{|S_y|}\right] + (\bar{X}'S_x^{-1}\bar{X} - \bar{Y}'S_y^{-1}\bar{Y}).$$

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.2. Analyse discriminante, classification.

4.2.1. Considérations générales.

4.2.2. Classification pour deux populations normales.

4.2.3. Classification pour $m \geq 3$ populations.

4.2.4. Illustrations sur des données réelles.

Classification pour $m \geq 3$ populations

Nous considérons le cas de m populations π_i $(i = 1, \ldots, m)$, associées à des lois de probabilité (sur $\mathbb{R}^p$) absolument continues de densité f_i $(i = 1, \ldots, m)$.

Nous cherchons à déterminer une règle de classification qui vise à parier sur la population dont est issue la réalisation x d'un p-v.a. X (que l'on suppose provenir d'une des π_i).

Comme dans le cas à deux populations, une telle règle est complétement déterminée par une partition

$$(R_i,\ i = 1, \ldots, m)$$

de l'ensemble χ des valeurs possibles pour x, et consistera à classifier X en π_i ssi $x \in R_i$.

Classification pour $m \geq 3$ populations

Si on note

- $c_{i|j}$ le coût de classification en π_i d'un objet de π_j,
- p_j la probabilité à priori que X provienne de π_j et
- $p_{i|j} := P[X \in R_i | X \in \pi_j] = \int_{R_i} f_j(x)\, dx$ la probabilité conditionnelle qu'un objet soit classifié en π_i sachant qu'il provient de π_j,

une procédure de classification $(R_i,\ i = 1, \ldots, m)$ sera dite optimale si elle minimise le coût de misclassification moyen, qui s'écrit ici (avec $c_{j|j} = 0$)

$$E_{\text{coût}} = \sum_{j=1}^{m} \mathrm{E}[\text{coût} | X \in \pi_j]\, p_j = \sum_{j=1}^{m} \left[\sum_{i=1}^{m} c_{i|j}\, p_{i|j} \right] p_j.$$

Classification pour $m \geq 3$ populations

On peut alors montrer le résultat suivant:

Théorème: la procédure de classification qui minimise le coût de misclassification moyen consiste à classifier x en la population π_i pour laquelle

$$h_i(x) := \sum_{j=1}^{m} c_{i|j}\, p_j\, f_j(x)$$

est minimal.

Remarques:

- Ceci étend bien le résultat vu pour $m = 2$.
- Si les coûts $c_{i|j}$ sont égaux ou non spécifiés, la règle optimale classifie x en la population π_i telle que $p_i f_i(x) = \max_j\{p_j f_j(x)\}$.

Classification pour $m \geq 3$ populations

Preuve.

$$
\begin{aligned}
E_{\text{coût}} &= \sum_{j=1}^{m} \left[\sum_{i=1}^{m} c_{i|j}\, p_{i|j} \right] p_j \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} c_{i|j}\, p_j \int_{R_i} f_j(x)dx \\
&= \sum_{i=1}^{m} \int_{R_i} h_i(x)dx.
\end{aligned}
$$

□

Classification pour $m \geq 3$ normales

Considérons le cas où $\pi_i = \mathcal{N}_p(\mu_i, \Sigma_i)$, $i = 1, \ldots, m$, sans inclure de coûts de misclassification.

$\rightsquigarrow$ **Proposition**:

$$d_j(x) := -\frac{1}{2} \ln |\Sigma_j| - \frac{1}{2}(x - \mu_j)' \Sigma_j^{-1} (x - \mu_j) + \ln p_j.$$

Alors la procédure de classification optimale classifie x en π_i ssi $d_i(x) = \max_j \{d_j(x)\}$.

Preuve. Le théorème affirme qu'il est optimal de classifier x en π_i ssi $\ln(p_i f_i(x)) = \max_j \{\ln(p_j f_j(x))\}$. Or

$$\ln(p_j f_j(x)) = \ln p_j - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2}(x - \mu_j)' \Sigma_j^{-1} (x - \mu_j),$$

ce qui établit le résultat. □

Classification pour $m \geq 3$ normales

En pratique, on classifiera x en π_i ssi

$$\hat{d}_i(x) = \max_j \{\hat{d}_j(x)\},$$

où

$$\hat{d}_j(x) := -\frac{1}{2} \ln |S_j| - \frac{1}{2}(x - \bar{X}_j)' S_j^{-1}(x - \bar{X}_j) + \ln p_j.$$

Bien entendu, $\bar{X}_j$ et S_j désignent ici les estimateurs non biaisés usuels de μ_i et Σ_i calculés à partir de m échantillons indépendants

$$(X_{j1}, \ldots, X_{j,n_j}),$$

où $X_{j1}, \ldots, X_{j,n_j}$ sont i.i.d. $\mathcal{N}_p(\mu_j, \Sigma_j)$.

Classification pour $m \geq 3$ normales

Dans le cas particulier où $\Sigma_i = \Sigma$ pour tout $i = 1, \ldots, m$, la règle de classification revient à classifier x en π_i ssi $d_i(x) = \max_j\{d_j(x)\}$, où

$$d_j(x) = -\frac{1}{2}\ln|\Sigma| - \frac{1}{2}x'\Sigma^{-1}x + \mu_j'\Sigma^{-1}x - \frac{1}{2}\mu_j'\Sigma^{-1}\mu_j + \ln p_j,$$

ou de manière équivalente, à classifier x en π_i ssi $d_i(x) = \max_j\{d_j(x)\}$, où

$$d_j(x) := \mu_j'\Sigma^{-1}x - \frac{1}{2}\mu_j'\Sigma^{-1}\mu_j + \ln p_j.$$

Remarque. $d_j(x)$ peut être estimé en remplaçant μ_j par $\bar{X}_j$ et Σ par

$$S_{\text{pool}} := \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \ldots + (n_m - 1)S_m}{n_1 + n_2 + \ldots + n_m - m}.$$

Classification pour $m \geq 3$ normales

Au niveau population, cette règle de classification revient encore à classifier x en π_i ssi $d_i(x) = \max_j\{d_j(x)\}$, où

$$d_j(x) := -\frac{1}{2}x'\Sigma^{-1}x + \mu_j'\Sigma^{-1}x - \frac{1}{2}\mu_j'\Sigma^{-1}\mu_j + \ln p_j$$

$$= -\frac{1}{2}(x-\mu_j)'\Sigma^{-1}(x-\mu_j) + \ln p_j = -\frac{1}{2}d^2_\Sigma(x,\mu_j) + \ln p_j$$

(intuition dans le cas $p_j = 1/m$).

Pour l'analogue empirique, ces nouveaux $d_j(x)$ seront bien entendu estimés par

$$\hat{d}_j(x) := -\frac{1}{2}d^2_{S_{\text{pool}}}(x,\bar{X}_j) + \ln p_j.$$

4. Méthodes classiques de l'analyse multivariée.

4.1. Analyse en composantes principales.

4.2. Analyse discriminante, classification.

4.2.1. Considérations générales.

4.2.2. Classification pour deux populations normales.

4.2.3. Classification pour $m \geq 3$ populations.

4.2.4. Illustrations sur des données réelles.

Illustrations sur des données réelles

Classifier le saumon aléoute et canadien. Les pêcheurs de commerce aléoutes n'ont pas le droit d'attraper trop de saumons canadiens, et vice-versa. Comment faire pour classifier les saumons afin de sortir indemne de cette situation peu heureuse?

Les poissons ont un cycle de vie fort intéressant. Ils naissent dans des courants d'eau fraîche, nagent jusqu'à l'océan et après quelques années ils retournent à leur lieu de naissance afin d'y mourir paisiblement.

Comme ils sont récoltés alors qu'ils sont dans l'océan, on ne peut pas *a priori* décider à quel groupe ils appartiennent.

Cependant, ils jouissent d'anneaux de croissance qui sont notoirement connus pour être plus grands pour les poissons canadiens lors de leur croissance en eau fraîche.

Illustrations sur des données réelles

Alaskan		Canadien	
frais	marine	frais	marine
108	368	129	420
131	355	148	371
105	469	179	407
86	506	152	381
...			
...			
94	491	153	352
87	480	108	339

Illustrations sur des données réelles

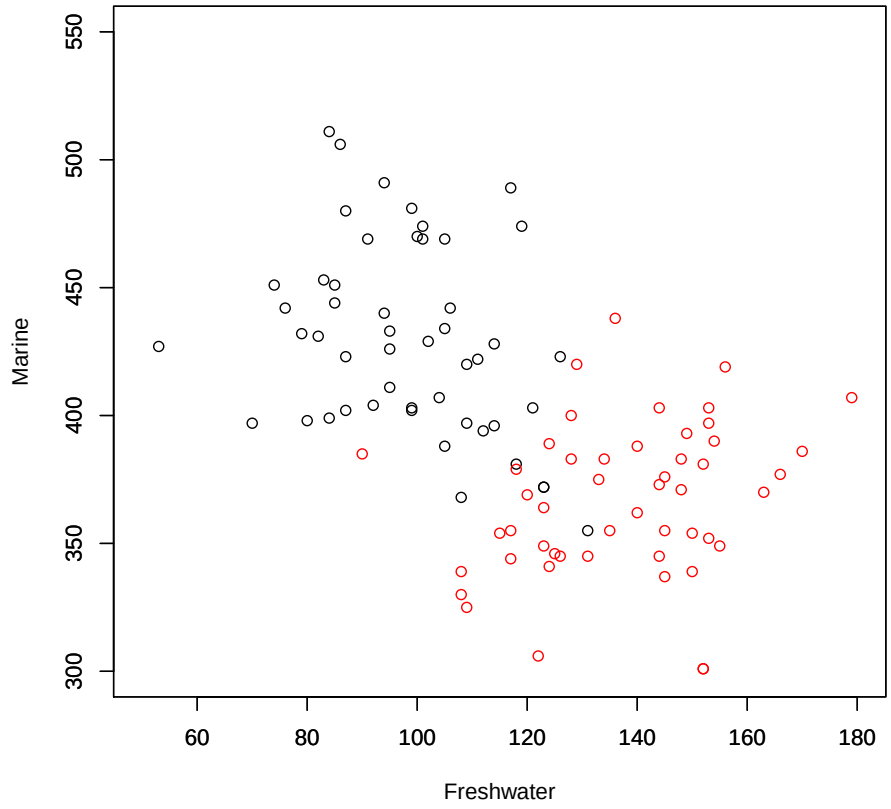


Figure: Alaskan fish (black) and Canadian fish (red) growthring diameters for a sample of 50.

Illustrations sur des données réelles

Nous avons pris les 30 premières données afin d'estimer S_x, S_y et μ_x, μ_y. Alors la règle de discrimination quadratique est appliquée aux 20 observations restantes. Dans ce cas particulier, nous n'avons pas fait de misclassification.

```
# estimate mean and covariance from the first 30 observations
Sx=var(sal[1:30,1:2])
Sy=var(sal[1:30,3:4])
k=log(det(Sx))-log(det(Sy))
+t(mx)%*%solve(Sx)%*%mx-t(my)%*%solve(Sy)%*%my
mx=c(mean(sal[1:30,1]),mean(sal[1:30,2]))
my=c(mean(sal[1:30,3]),mean(sal[1:30,4]))

# define discrimiant function
disc<-function(x){
if(-1/2*t(x)%*%(solve(Sx)-solve(Sy))%*%x
+(t(mx)%*%solve(Sx)-t(my)%*%solve(Sy))%*%x-k/2>0)  1 else 0}
```

Illustrations sur des données réelles

Afin d'obtenir une meilleure estimation de la qualité de notre procédure de classification, nous pouvons utiliser l'approche suivante connue sous le nom de *leave-one-out*.

- Prendre toutes les observations sauf une pour estimer le modèle (p.ex. S_x, S_y, ...)
- Classifier l'observation laissée de côté
- Répéter cette procédure pour chaque observation.

Nous pouvons alors estimer les probabilités de misclassification.

Illustrations sur des données réelles

En ayant recours à la discrimination linéaire, J&W ont appliqué cette procédure aux données saumoniennes.

		prédit	
		π_1	π_2
vrai	π_1	44	6
	π_2	1	49

Remarque. Il s'avère plus rare que des saumons canadiens (de naissance) sont misclassifiés comme aléoutes que vice-versa. La procédure n'est donc pas correcte dans un certain sens, et croire aveuglément en une certaine procédure est dangereux! Surtout pour les pêcheurs de saumons...