A V A N T – P R O P O S

Cet ouvrage fournit des éléments de la théorie de la régression linéaire et de la théorie des axes factorielles avec quelques applications.

Il faut souligner qu'il s'agit d'un ouvrage à compléter – c'est pourquoi les suggestions de la part des lecteurs intéressés seront bienvenues.

La composition présente de l'ouvrage est répartie comme suit :
C. Costinescu – les chapitres 1 et 2, V. Petrehuş – le chapitre 3.

Les deux auteurs ont bénéficié du soutien de l'Agence Universitaire de la Francophonie par l'intermède du Projet de Coopération Scientifique interuniversitaire – 2005.

Cristian Costinescu et Viorel Petrehuş

TABLE DES MATIERES

LA REGRESSION SIMPLE

LA REGRESSION MULTIPLE

ANALYSE FACTORIELLE

LA REGRESSION SIMPLE

Etant donné deux variables aléatoires X et Y qui ne sont pas indépendantes, on peut admettre que le phénomène représenté par X puisse servir à prédire celui représenté par Y. On est conduit à la recherche d'une function f telle que $f(X)$ soit aussi proche que possible de Y dans un sens qui sera précisé.

Dans le cas théorique on va chercher une formule de prevision idéale au sens des moindres carrés, puis on va aborder le cas usual où les variables X et Y ne sont connues que par les valeurs d'un échantillon.

Dans ce qui suit X sera dit variable explicative ou prédicteur et Y sera dit variable expliquée ou critère.

1. Le modèle théorique. La régression linéaire

Soit $L^2(X)$ le sous-espace de L^2 formé par les variables aléatoires functions de X du type $f(X)$ et qui contient la droite D des variables aléatoires constantes. Alors l'espérance conditionnelle de Y sachant X, $E(Y/X)$, est considérée la projection orthogonale de Y sur $L^2(X)$; on sait que le minimum de l'expression

$$E[(Y - f(X))^2]$$

est atteint pour $f(X) = E(Y/X)$. On peut dire que $E(Y/X)$ est la meilleure approximation de Y par une function de X et il est alors immédiat, à cause de l'orthogonalité, que la différence $Y - E(Y/X)$ est non corrélée avec X.

En plus on peut interpréter le théorème de la variance totale comme le théorème de Pythagore appliqué au triangle rectangle de sommets Y, E (Y) et E (Y/X) – voir figure 1:

$\| Y - E(Y) \|^2 = V(Y)$,

$\| E(Y/X) - E(Y) \|^2 = V(E(Y/X))$ et

$\| Y - E(Y/X) \|^2 = E[(Y - E(Y/X))^2] = E(V(Y/X))$; on a donc:

$$V(Y) = V(E(Y/X)) + E(V(Y/X)) \qquad (1)$$

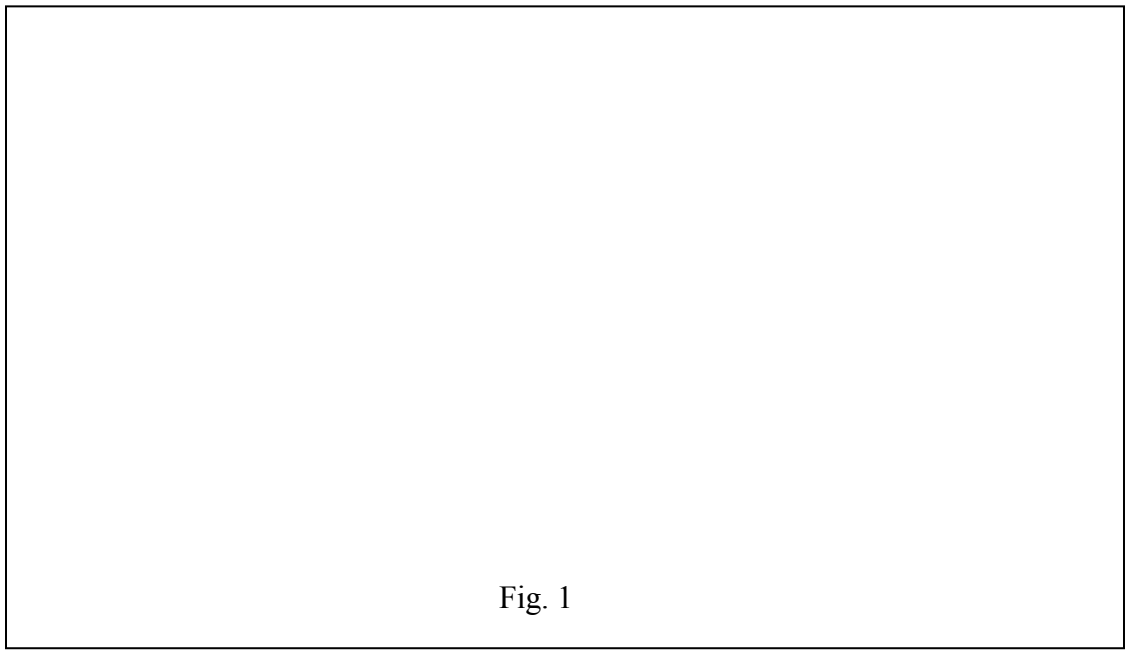

Fig. 1

On sait que le coefficient de corrélation linéaire $\rho = \frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}}$ est une mesure symétrique de dépendence, étant maximal dans le cas de la liaison linéaire.

En utilisant le théorème de la variance totale on peut introduire un autre rapport de corrélation $\eta (Y/X)$ tel que:

$$\eta^2(Y/X) = \frac{V(E(Y/X))}{V(Y)}$$

c. a. d. le quotient de la variance expliquée par la variance totale.

$\eta(Y/X)$ est non symétrique et il permet de "mesurer l'approximation" de Y par $E(Y/X)$.

$\eta^2(Y/X)$ est égal avec le cosinus carré de l'angle formé par $Y - E(Y)$ et l'espace $L^2(X)$; on a donc: $0 \leq \eta^2(Y/X) \leq 1$.

Si $\eta^2(Y/X) = 1$, tenant compte de la relation (1), on obtient $E(V(Y/X)) = 0$; alors on en déduit que $V(Y/X) = 0$ presque sûrement, c'est à dire qu'à X fixé la variance de Y est nulle, donc Y ne prend qu'une seule valeur.

En conclusion, le rapport de corrélation est maximal si Y est lié fonctionnellement à X, c. a. d. $\eta^2(Y/X) = 1$ implique $Y = f(X)$.

Si $\eta^2(Y/X) = 0$ on a $V(E(Y/X)) = 0$; alors $E(Y/X)$ est presque sûrement une constante et on dit que Y est non corrélé avec X. C'est en particulier le cas si X et Y sont indépendantes mais la réciproque est fausse !

En fait, $\eta^2(Y/X) = 0$ signifie seulement que $Y - E(Y)$ est orthogonal à l'espace $L^2(X)$.

D'autre part, $\rho^2 \leq \eta^2(Y/X)$ puisque ρ^2 est le cosinus carré de l'angle formé par $Y - E(Y)$ avec le sous-espace de dimension 2 de $L^2(X)$ engendré par X et la droite des constantes D.

Le cas $\rho^2 = \eta^2(Y/X)$ signifie que $E(Y/X)$ appartient à ce sous-espace de dimension 2 , donc que:

$$E(Y/X) = \alpha + \beta X$$

qui est le cas de la régression linéaire qu'on va étudier en détail ci-dessous.

Enfin, on rappelle que la fonction qui, pour une valeur x de la variable aléatoire X, associe $E(Y/X = x)$ est dite fonction de régression de Y en X.

Alors on peut considérer

$$Y = E(Y/X) + \varepsilon$$

où ε est un résidu aléatoire - pas toujours négligeable …

Le résidu ε a l'espérance nulle: $E(\varepsilon) = 0$ puisque $E(Y) = E(E(Y/X))$; de plus, tenant compte que ε est orthogonal à l'espace $L^2(X)$, il est non corrélé avec X et avec $E(Y/X)$.

La variance de ε, dite résiduelle, est de la forme:

$$V(\varepsilon) = [\, 1 - \eta^2 (Y/X) \,]\, V(Y) \qquad (2)$$

(voir les relations précédentes).

Dans la pratique, le cas de la régression linéaire :

$$E(Y/X) = \alpha + \beta X$$

est le plus important (il se produit en particulier si les variables aléatoires X et Y suivent une loi normale à deux dimensions).

En prenant l'espérance des deux membres de la relation

$$Y = \alpha + \beta X + \varepsilon$$

on obtient:

$$E(Y) = \alpha + \beta E(X)$$

car $E(\varepsilon) = 0$. Donc la droite de régression passe par le point de coordonnées $(E(X), E(Y))$ et alors on a :

$$Y - E(Y) = \beta (X - E(X)) + \varepsilon$$

Maintenant on multiplie par $X - E(X)$ les deux membres de la relation précédente et en prenant l'espérance il vient :

$$E[(X - E(X))(Y - E(Y))] = \beta E[(X - E(X))^2] + E[\varepsilon (X - E(X))]$$

soit

$$\text{cov}(X, Y) = \beta V(X) + \text{cov}(\varepsilon, X)$$

puisque l'espérance de ε est nulle.

Tenant compte à présent que le résidu ε est non corrélé avec la variable aléatoire X, il nous reste :

$$\beta = \frac{\text{cov}(X,Y)}{V(X)}$$

et alors l'équation de la droite de régression est de la forme:

$$E(Y/X) - E(Y) = \frac{\text{cov}(X,Y)}{V(X)} (X - E(X))$$

d'où on a :

$$Y = E(Y) + \frac{\text{cov}(X,Y)}{V(X)} (X - E(X)) + \varepsilon$$

En prenant la variance des deux membres de la dernière relation et tenant compte que le résidu ε est non corrélé avec X, il vient :

$$V(Y) = \rho^2 V(Y) + V(\varepsilon)$$

Vu la relation (2) on retrouve donc l'égalité $\rho^2 = \eta^2 (Y/X)$ si la régression est linéaire.

2. Estimation des paramètres par la méthode des moindres carrés

On va considérer n couples (x_i , y_i) d'observations indépendantes des variables aléatoires X et Y (c'est à dire un **n – échantillon**) et on suppose vraie l'hypothèse :

$$E (Y/ X) = \alpha + \beta X$$

La méthode utilisée s'applique encore si X n'est pas variable aléatoire, mais elle est connue à travers les valeurs d'un échantillon ; par exemple si Y est une grandeur mesurée à différents moments $x_1, x_2 , \dots , x_n$ – c. a. d. X est le temps. Il suffit alors de supposer que $y_i = \alpha + \beta x_i + \varepsilon_i$ où ε_i (i = 1, … , n) sont des réalisations indépendantes d'une variable ε d'espérance nulle et de variance constante σ^2 , quel que soit l'observation x_i . Dans ce cas on parle de **modèle linéaire.**

Pour ce qui suit on va estimer α , β et la variance du résidu ε par la méthode des moindres carrés, due à Gauss.

Puisque la méthode des moindres carrés ne dépende pas que des lois conditionnelles à X fixé, on peut aborder par les mêmes techniques la régression linéaire et le modèle linéaire.

En utilisant la méthode des moindres carrés sur le n – échantillon concerné, on va ajuster au nuage de points (x_i , y_i) une droite d'équation $y^* = a x + b$ tel que la somme des carrés

$$\sum_{i=1}^{n} (y_i - y_i^*)^2$$

soit minimale (voir figure 2).

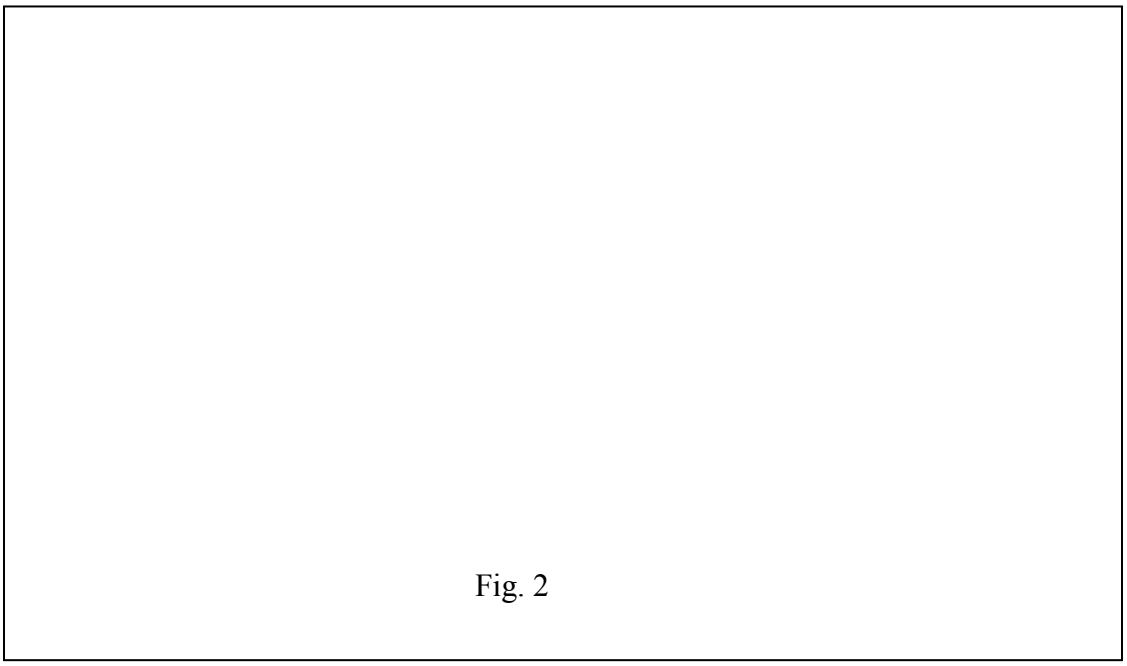

Fig. 2

Pour déterminer a et b on considère la fonction

$$F(a, b) = \sum_{i=1}^{n} (y_i - a - b x_i)^2$$

dont le minimum est atteint pour $\frac{\partial F}{\partial a} = 0 = \frac{\partial F}{\partial b}$; alors on obtient les équations :

$$\sum_{i=1}^{n} (y_i - a - b x_i) = 0$$

$$\sum_{i=1}^{n} x_i \, (y_i - a - b x_i) = 0 \qquad (3)$$

En divisant la première relation par n et en utilisant les notations clasiques pour les moyennes empiriques (dites « arithmétiques ») :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ , \quad \text{respectivement} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

il vient: $\bar{y} = a + b\,\bar{x}$.

En portant cette valeur de a dans la deuxième équation du système (3) on obtient :

$$\mathrm{b} = \frac{\sum_{i=1}^{n} y_i\, x_i - n\,\bar{y}\,\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^{\,2}} \ .$$

Par des calculs assez simples :

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} y_i\, x_i - \bar{y} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i\, x_i - n\,\bar{x}\,\bar{y} \quad (\text{puisque } \sum_{i=1}^{n} (y_i - \bar{y})\,\bar{x} = 0)$$

il vient finalement que

$$\mathrm{b} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\mathrm{cov}(x, y)}{s_x^2}$$

où cov (x, y) désigne la covariance empirique (dite « observée ») :

$$\mathrm{cov}\ (\mathrm{x},\ \mathrm{y}) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n}$$

et s_x est l'écart – type empirique de x, tel que

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Tenant compte de la définition du coefficient empirique de corrélation linéaire :

$$r = \frac{\mathrm{cov}(x, y)}{s_x \, s_y}$$

on obtient encore pour b :

$$\mathrm{b} = r \, \frac{s_y}{s_x}$$

d'où l'équation de la droite en question :

$$\mathrm{y}^* = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x}) .$$

En conclusion : la droite « des moindres carrés » passe par le centre de gravité du nuage des points et sa pente est l'analogue empirique de la pente de la droite de régression :

$$\rho \, \frac{\sigma_y}{\sigma_x} = \rho \, \frac{\sqrt{V(Y)}}{\sqrt{V(X)}} .$$

Il faut aussi souligner que le coefficient empirique de corrélation linéaire mesure exclusivement le caractère plus ou moins linéaire du nuage de points considérés.

Proposition 1. *a, b et y* sont des estimateurs sans biais de* α , β *et respectivement de* $\alpha + \beta x = E (Y/ X = x)$.

Démonstration. On note par

$$\mathrm{B} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

la variable aléatoire dont la réalisation est le paramètre b .

Puisque l'espérance de l'espérance conditionnelle est égale avec l'espérance de B, pour prouver que E (B) = β il suffit de montrer que $E_{x_i}(B)=\beta$, où $E_{x_i}(B)$ désigne l'espérance conditionnelle de B par rapport aux valeurs x_i des variables aléatoires X_i .

Il vient :

$$E_{x_i}(B) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, E_{x_i}(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Vu l'hypothèse de régression linéaire on a : $E_{x_i}(Y_i) = \alpha + \beta\, x_i$ et aussi $E_{x_i}(\bar{Y}) = \alpha + \beta\, \bar{x}$, d'où on obtient : $E_{x_i}(Y_i - \bar{Y}) = \beta\,(x_i - \bar{x})$.

Finalement il vient :

$$E_{x_i}(B) = \frac{\beta \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta$$

d'où E(B) = β .

Puisque a = $\bar{y}$ - b $\bar{x}$ est une réalisation de la variable aléatoire A = $\bar{Y} - B\bar{X}$, on obtient de la même manière :

$$E_{x_i}(A) = E_{x_i}(\bar{Y}) - \bar{x}\, E_{x_i}(B) = \alpha + \beta\, \bar{x} - \bar{x}\, \beta = \alpha$$

et alors on a: E (A) = α .

Car E (Y/ X = x) = α + β x il résulte que y* = a x + b est un estimateur sans biais de α + β x . □

Remarques 1. On peut prover de plus que la variable aléatoire B n'est pas corrélée avec $\bar{Y}$: tout d'abord on va simplifier l'expression de b

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

tenant compte que :

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i-\bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i-\bar{x}) \quad \text{et} \quad \sum_{i=1}^{n}(x_i-\bar{x})=0$$

Donc il vient

$$B = \frac{\sum_{i=1}^{n} Y_i\,(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

et alors la covariance conditionnelle des variables aléatoires B et $\bar{Y}$ aux valeurs x_i fixées, est de la forme :

$$\text{cov}(B, \bar{Y}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,\text{cov}(Y_i, \bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\ .$$

Or $\text{cov}(Y_i, \bar{Y}) = \text{cov}\left(Y_i, \frac{1}{n}\sum_{j=1}^{n} Y_j\right) = \frac{\sigma^2}{n}$ puisque les variables aléatoires Y_i et Y_j sont indépendantes pour tout $i \neq j$; finalement on obtient :

$$\text{cov}(B, \bar{Y}) = \frac{\sigma^2 \sum_{i=1}^{n}(x_i-\bar{x})}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} = 0$$

ce qui implique: B et $\bar{Y}$ sont non corrélées conditionnellement par rapport aux valeurs x_i.

2. Pourtant, le fait d'être sans biais n'est qu'une propriété mineure pour les estimateurs. Le résultat suivant (connu comme **le théorème Gauss – Markov**) donne la qualité des estimateurs obtenus : *a et b, parmi les estimateurs sans biais de α et β, sont de variance minimale.*

Proposition 2. *Pour les variances conditionnelles des variables aléatoires A et B par rapport aux valeurs x_i on a les formulas suivantes:*

$$V_{x_i}(B) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad ; \quad V_{x_i}(A) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Démonstration. Pour x_i fixés il vient $Y_i = \alpha + \beta x_i + \varepsilon$ et donc on a:

$$V(Y_i / X_i = x_i) = V(\varepsilon) = \sigma^2 .$$

Alors on obtient:

$$V_{x_i}(B) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 V_{x_i}(Y_i)}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} = \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} .$$

D'autre part $A = \bar{Y} - B\bar{X}$ d'où on a : $V_{x_i}(A) = V_{x_i}(\bar{Y}) + \bar{x}^2 V_{x_i}(B)$ et alors il résulte la formule de ci-dessus pour la variance conditionnelle de A par rapport aux valeurs x_i. □

Proposition 3. *La moyenne de l'écart résiduel est nulle et sa variance empirique (dite **résiduelle**) est égale à $(1 - r^2)\, s_y^2$.*

Démonstration. On va noter $e_i = y_i - y_i^*$; car $y_i^* = \bar{y} + b(x_i - \bar{x})$ il vient :

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \bar{y}) - b\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

et donc la moyenne de l'écart résiduel est nulle. Il résulte que les e_i ne sont pas des réalisations indépendantes d'une variable aléatoire.

La variance résiduelle est égale à $\frac{1}{n}\sum_{i=1}^{n} e_i^2$ et elle sera notée par $s^2_{y/x}$; il vient :

$$s^2_{y/x} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{b^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{2b}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) =$$

$$= s_y^2 + b^2\, s_x^2 - 2\, b\, \mathrm{cov}(x, y) = s_y^2 + r^2\, s_y^2 - 2\, r^2\, \frac{s_y}{s_x}\, s_x\, s_y = (1 - r^2)\, s_y^2$$

puisque $r = \frac{\mathrm{cov}(x,y)}{s_x\, s_y}$ et $b = r\, \frac{s_y}{s_x}$. □

Remarques 3. Pour évaluer $V(\varepsilon) = \sigma^2$ il faut utiliser la variance des e_i ; on peut montrer que

$$\frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n-2}$$

est un estimateur sans biais de σ^2 (voir plus loin …).

Il faut préciser que des nombreux modèles non linéaires se ramènent aux modèles linéaires par des transformations simples. Par exemple: le modèle $y = \alpha\, x^{\beta}$, très utile en économétrie (élasticité constante de y par rapport à x, où β est le coefficient d' élasticité), devient un modèle linéaire en passant au logarithme: $\ln y = \ln \alpha + \beta \ln x$, d'où on a

$$y' = \ln \alpha + \beta\, x', \quad \text{avec } y' = \ln y \text{ et } x' = \ln x$$

De même pour le modèle logistique souvent utilisé pour rendre compte des variations d'un taux de réponse y (compris entre 0 et 1) en fonction d'une « excitation » x :

$$y = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)};$$

il suffit de poser

$$y' = \ln \frac{y}{1-y} \qquad \text{avec } y' = \alpha + \beta\, x .$$

3. Applications et cas particuliers

a. Si on suppose que **le résidu ε suit une loi normale N (0; σ)** il résulte :

1) la variable aléatoire conditionnée Y / X = x suit une loi normale N (α + β x ; σ) puisque on travaille sous l'hypothèse E (Y/ X) = α + β X .

2) tenant compte que A, B et Y* sont des combinaisons linéaires des lois normales, elles suivent aussi (pour x_i fixés) des lois normales à savoir

$$A \in N\left(\alpha ; \sigma \sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right)$$

$$B \in N\left(\beta ; \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right)$$

$$Y^* \in N\left(\alpha+\beta x ; \sigma \sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right)$$

3) A et B sont des estimateurs de variance minimale de α et β, mais ils ne sont pas indépendants.

4) $\frac{\sum_{i=1}^{n}(y_i-y_i^*)^2}{\sigma^2}=\frac{n s_{y/x}^2}{\sigma^2}$ est une réalisation d'une variable qui suit une loi χ^2_{n-2} et qui est indépendante de A et B.

5) Puisque $\frac{(B-\beta)\sqrt{\sum_{i=1}^{n}(x_i - x)^2}}{\sigma} \in N(0;1)$ et $\frac{n s_{y/x}^2}{\sigma^2} \in \chi_{n-2}^2$ sont

indépendantes il résulte que

$$\frac{(B-\beta)\, s_x \sqrt{n-2}}{s_{y/x}}$$

suit une loi T_{n-2}, ce qui donne la possibilité d'obtenir des intervalles de confiance pour β.

Remarques.

- l'usage des lois normales de A et B suppose, théoriquement, σ connu ce qui en pratique n'est pas vrai.

- si le coefficient de corrélation linéaire ρ est nul on obtient $\beta = 0$, hypothèse dite de non – régression.

b. Pour effectuer des **tests dans le modèle linéaire** on utilise tout d'abord la décomposition :

$$y_i - \bar{y} = y_i - y_i^* + y_i^* - \bar{y}$$

et on suppose que le résidu ε suit une loi normale $N(0; \sigma)$.

On a immédiatement que

$$\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \bar{y}) = 0$$

et alors il résulte :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - y_i^*)^2 + \sum_{i=1}^{n}(y_i^* - \bar{y})^2$$

c'est-à-dire la somme des carrés totale est égale avec la somme des sommes des carrés résiduelle et expliquée.

Puisqu'on sait que :

$$\frac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{\sigma^2}$$

suit une loi χ^2_{n-2} (khi – carré d'ordre n-2), si l' hypothèse de non – régression linéaire

$$H_0 : \beta = 0$$

est vraie il résulte que :

$$\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2}$$

suit une loi khi – carré d'ordre n-1 et on a :

$$\frac{\sum_{i=1}^{n}(Y_i^* - \overline{Y})^2}{\sigma^2} = \frac{\sum_{i=1}^{n}B(X_i - \overline{X})^2}{\sigma^2} .$$

Car $\beta = 0$, on en déduit que $\frac{\sum_{i=1}^{n}(Y_i^* - \overline{Y})^2}{\sigma^2}$ suit suit une loi khi – carré d'ordre 1 puisque $\frac{\sum_{i=1}^{n}(B-\beta)^2(X_i - \overline{X})^2}{\sigma^2}$ suit suit une loi khi – carré d'ordre 1, étant le carré d'une variable normale standard N (0; 1).

Les variables aléatoires

$$\sum_{i=1}^{n}(Y_i - Y_i^*)^2 \text{ et } \sum_{i=1}^{n}(Y_i^* - \overline{Y})^2$$

sont indépendantes et tenant compte du fait que le carré d'une loi T_{n-2} est une loi Fisher – Snedecor F (1 ; n-2) il résulte que le quotient

$$\frac{\sum_{i=1}^{n}(Y_i^* - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i^* - Y_i)^2}(n-2)$$

suit une loi Fisher– Snedecor F (1 ; n-2) pour $\beta = 0$.

Alors on obtient immédiatement le test du caractère significatif de la régression, ce test étant d'ailleurs identique à celui du coefficient de corrélation linéaire :

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

c. Cas d'hétéroscédasticité : dans la pratique on a souvent $V (\varepsilon / X = x) = \sigma^2 x^2$ - c. a. d. l'écart type du résidu croît linéairement avec le prédicteur.

Les estimateurs obtenus par la méthode des moindres carrés sont sans biais mais ils ne sont pas de variance minimale. La vraisemblance des y_i est de la forme:

$$L (y_1, \ldots, y_n) = \frac{1}{(2n)^{n/2}\,\sigma^n \prod_{i=1}^{n} x_i} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{y_i - \alpha - \beta x_i}{x_i}\right)^2\right).$$

Alors les estimateurs de maximum de vraisemblance sont des estimateurs de variance minimale et il est évident que le problème en question est équivalent à une régression usuelle sur des dates transformées : si on note

$$y_i' = \frac{y_i}{x_i} \qquad x_i' = \frac{1}{x_i} \qquad \varepsilon_i' = \frac{\varepsilon_i}{x_i}$$

il vient :

$$y_i' = \beta + \alpha\, x_i' + \frac{\varepsilon_i}{x_i} = \beta + \alpha\, x_i' + \varepsilon_i'$$

avec V $(\varepsilon_i^{'}) = \sigma^2$. Donc c'est suffisant d'ajuster une droite au nuage des points de coordonnées ($x_i^{'} = \frac{1}{x_i}$, $y_i^{'} = \frac{y_i}{x_i}$).

On observe que la constante du modèle transformé est exactement la pente de la droite de régression du modèle initial et **réciproquement.**

LA REGRESSION MULTIPLE

Pour généraliser les notions du chapitre précédent la difficulté du sujet ne consiste pas tant de la complexité des calculs, mais plutôt de la distinction qui existe entre la régression multiple et le modèle linéaire – les hypothèses et les objectifs sont différentes.

1. Régression multiple entre variables aléatoires

On considère sur n individus k + 1 mesures représentées par des vecteurs de $\mathbf{R}^n$: y, x_1, x_2, … , x_k ; comme précédemment y est la variable expliquée ou critère et x_i sont les variables explicatives ou prédicteurs.

On suppose que les variables explicatives sont linéairement indépendantes, mais elles peuvent être corrélées, par exemple – c. a. d. elles ne sont pas supposées statistiquement indépendantes.

Pour exprimer y au moyen des x_i par une formule linéaire (la « recherche d'un ajustement linéaire ») on introduit la régression multiple :

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{11} + ... + \beta_k x_{1k} + \varepsilon_1 \\ \quad ..... \\ \quad ..... \quad ..... \quad ..... \\ y_n = \beta_0 + \beta_1 x_{n1} + ... + \beta_k x_{nk} + \varepsilon_n \end{cases} \qquad (1)$$

On va noter par X la matrice à n lignes, dont la première colonne est constituée par des unités et les autres k colonnes par les valeurs des variables explicatives x_1, …, x_k:

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

Elle sera dite **matrice modèle**.

Si on pose :

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}; \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_2 \\ \dots \\ \beta_k \end{pmatrix}; \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

alors le système (1) s'écrit sous la forme matricielle suivante :

$$\mathbf{y} = X\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2)$$

Remarque 1. La relation (2) caractérise aussi la régression simple.

Pour obtenir l'estimation du vecteur $\boldsymbol{\beta}$ on va utiliser la méthode des moindres carrés en minimisant l'expression :

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2 = (\mathbf{y} - X\,\boldsymbol{\beta})^t (\mathbf{y} - X\,\boldsymbol{\beta}) =$$

$$= \mathbf{y}\,\mathbf{y}^t - \boldsymbol{\beta}^t X^t \mathbf{y} - \mathbf{y}^t X\,\boldsymbol{\beta} + \boldsymbol{\beta}^t X^t X\,\boldsymbol{\beta} = \mathbf{y}\,\mathbf{y}^t - 2\,\boldsymbol{\beta}^t (X^t \mathbf{y}) + \boldsymbol{\beta}^t (X^t X)\,\boldsymbol{\beta}$$

puisque $\mathbf{y}^t X\,\boldsymbol{\beta} = \boldsymbol{\beta}^t (X^t \mathbf{y})$ est un scalaire.

Le minimum de S, noté par $\mathbf{b}$ et appelé vecteur des coefficients de régression**,** s'obtient de l'équation :

$$\frac{\partial S}{\partial \beta} = -2\, X^t \mathbf{y} + 2\, X^t X \boldsymbol{\beta} = 0 ;$$

alors il vient :

$$(X^t X)\, \mathbf{b} = X^t \mathbf{y} \qquad (3)$$

relation qui donne l'estimation du vecteur $\boldsymbol{\beta}$ par la méthode des moindres carrés.

Remarques 2. Si la matrice $X^t X$ est inversible, alors l'équation (3) a la solution unique :

$$\mathbf{b} = (X^t X)^{-1} X^t \mathbf{y} \qquad (4)$$

Si la matrice $X^t X$ est singulière on va résoudre l'équation (3) en utilisant l'inverse généralisée d'une matrice (voir l'annexe 1) ; bien que l'estimation $\mathbf{b}$ n'est pas unique, on obtient que $X\,\mathbf{b}$ est unique (voir également l'annexe 1).

3. Pour justifier théoriquement l'ajustement linéaire de y au moyen des x_i on va utiliser le modèle probabiliste suivant : on suppose que $y, x_1, x_2, \ldots, x_k$ forment un n – échantillon d'observations indépendantes de $k+1$ variables aléatoires $Y, X_1, X_2, \ldots, X_k$.

On sait que la meilleure approximation de Y par une fonction des X_i est donnée par l'espérance conditionnelle $E[\, Y / X_1, X_2, \ldots, X_k \,]$ et en utilisant l'hypothèse de régression **linéaire** multiple :

$$E[\, Y / X_1, X_2, \ldots, X_k \,] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \qquad (5)$$

on obtient le modèle suivant :

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

où ε est une variable aléatoire d'espérance nulle, non corrélée avec les X_i , dont la variance sera notée par σ^2.

Entre les réalisations des variables aléatoires $Y, X_1, X_2, \ldots, X_k$ et ε il existe une relation de type (2) déduite de l'hypothèse de régression linéaire multiple (5).

Il faut mentionner qu'en pratique les coefficients $\beta_0, \beta_1, \ldots, \beta_k$ et σ^2 ne sont pas connus ; alors il est nécessaire de les estimer le mieux possible.

La définition du résidu est tout à fait similaire que pour la régression simple

$$\mathbf{e} = \mathbf{y} - \mathbf{y}^*$$

où $\mathbf{e}$ et $\mathbf{y}^* = X\,\mathbf{b}$ sont les vecteurs:

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}, \qquad \mathbf{y}^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \ldots \\ y_n^* \end{pmatrix}$$

Tenant compte de (4) il vient que $\mathbf{y}^* = X\,(X^t X)^{-1} X^t \mathbf{y}$. Avec les notations $A = X\,(X^t X)^{-1} X^t$ et $M = I_n - A$ on obtient :

$$M\,X = X - X\,(X^t X)^{-1} X^t X = X - X = 0.$$

En utilisant maintenant la relation (2) il vient que :

$$\mathbf{e} = \mathbf{y} - A\,\mathbf{y} = M\,\mathbf{y} = M\,X\,\boldsymbol{\beta} + M\boldsymbol{\varepsilon} = M\boldsymbol{\varepsilon} \qquad (6)$$

et alors on a le résultat suivant

Proposition 1. *Le résidu* $\boldsymbol{e}$ *est orthogonal à* $\boldsymbol{y}^*$ *et à la matrice modèle* X *- voir la relation (2).*

Démonstration. Tenant compte des relations précédentes on a :

$$X^t\,\mathbf{e} = X^t M\boldsymbol{\varepsilon} = 0\boldsymbol{\varepsilon} = \mathbf{0}$$

où par $\mathbf{0}$ nous avons désigné le vecteur nul ; il vient aussi :

$$(\mathbf{y}^{*})^{t} \mathbf{e} = \mathbf{b}^{t} X^{t} \mathbf{e} = \mathbf{b}^{t} \mathbf{0} = 0 \qquad \square$$

Corollaire 2. *On a :*

$$\sum_{i=1}^{n} e_i = 0$$

Démonstration. Si on note par $\mathbf{1} = (1, \ldots, 1)^{t}$ la première colonne de la matrice modèle X (on dit que « β_0 est présent dans le modèle ») et en utilisant la proposition précédente, il vient :

$$\sum_{i=1}^{n} e_i = \mathbf{1}^{t} \mathbf{e} = 0 \qquad \square$$

Remarques 4. A l'aide de la proposition 1 on peut montrer que le minimum de l'expression S est atteint vraiment pour $\mathbf{b} = \boldsymbol{\beta}$; d'abord on a :

$$(\mathbf{b} - \boldsymbol{\beta})^{t} X^{t} (\mathbf{y} - X \mathbf{b}) = (\mathbf{y} - X \mathbf{b})^{t} X (\mathbf{b} - \boldsymbol{\beta}) = \mathbf{e}^{t} X (\mathbf{b} - \boldsymbol{\beta}) = 0$$

(puisque $(\mathbf{b} - \boldsymbol{\beta})^{t} X^{t} (\mathbf{y} - X \mathbf{b})$ est un scalaire). Alors il vient :

$$S = (\mathbf{y} - X \mathbf{b} + X \mathbf{b} - X \boldsymbol{\beta})^{t} (\mathbf{y} - X \mathbf{b} + X \mathbf{b} - X \boldsymbol{\beta}) =$$

$$= (\mathbf{y} - X \mathbf{b})^{t} (\mathbf{y} - X \mathbf{b}) + (\mathbf{b} - \boldsymbol{\beta})^{t} (X^{t} X) (\mathbf{b} - \boldsymbol{\beta}).$$

On observe que les expressions situées dans la dernière ligne sont des formes quadratiques normales – donc positives – et que la première forme quadratique ne dépend pas de $\boldsymbol{\beta}$. Donc il résulte que le minimum de l'expression S est atteint pour $\mathbf{b} = \boldsymbol{\beta}$. $\square$

5. Un simple calcul nous montre que A est idempotente :

$$X (X^{t} X)^{-1} X^{t} X (X^{t} X)^{-1} X^{t} = X (X^{t} X)^{-1} X^{t}$$

et donc M est aussi idempotente:

$$M^2 = (I_n - A) (I_n - A) = I_n - 2\,A + A^2 = I_n - 2\,A + A^2 = I_n - A = M$$

6. Si l'espace $\mathbf{R}^n$ est muni de la métrique **s** , la méthode des moindres carrés exige que $\| \mathbf{y} - \mathbf{y}^* \|^2$ soit minimale ; géométriquement, $\mathbf{y}^*$ est alors la projection **s** – orthogonale de **y** sur le sous – espace V engendré par $\mathbf{1, x_1, x_2, \dots , x_k}$.

On sait que l'opérateur de projection **s** – orthogonale sur V est donné par l'expression $X (X^t \mathbf{s} X)^{-1} X^t \mathbf{s}$; alors on a :

$$\mathbf{y}^* = X (X^t \mathbf{s} X)^{-1} X^t \mathbf{s}\, \mathbf{y} \quad \text{et} \quad \mathbf{b} = (X^t \mathbf{s} X)^{-1} X^t \mathbf{s}\, \mathbf{y}$$

En particulier, pour la métrique $\mathbf{s} = \frac{1}{n}$ Id on obtient la formule suivante pour le vecteur des coefficients de régression :

$$\mathbf{b} = (X^t X)^{-1} X^t \mathbf{y}$$

formule qui coïncide avec la relation (4) obtenue précédemment.

2. Le modèle linéaire général

En pratique on fixe d'habitude certaines valeurs des conditions expérimentales et on mesure plusieurs fois de suite un phénomène pour les mêmes valeurs des conditions expérimentales. On obtient donc un nuage de p vecteurs $\mathbf{y_1, y_2, \dots , y_p}$ dans $\mathbf{R}^n$ si on répète p fois l'expérience, les k variables explicatives $\mathbf{x_1, x_2, \dots , x_k}$ restant les mêmes.

Pour le modèle linéaire général il faut supposer que le centre de gravité du nuage des $\mathbf{y_1, y_2, \dots, y_p}$ soit situé dans le sous – espace V engendré par $\mathbf{1, x_1, x_2, \dots, x_k}$ (voir la remarque 5 du paragraphe précédent) : $\mathbf{h = X\,\beta}$.

Puisqu'en réalité on ne connaît pas (la plupart du temps) qu'un seul point du nuage, le but est d'approximer le mieux possible $\mathbf{h}$ à l'aide d'une seule observation $\mathbf{y}$.

Géométriquement, l'approximation $\mathbf{h}^*$ de $\mathbf{h}$ - obtenue à l'aide de $\mathbf{y}$, est exactement la projection orthogonale de $\mathbf{y}$ sur le sous – espace V, selon une certaine métrique $\mathbf{s}$; alors le problème est de trouver une métrique de telle sorte que $\mathbf{h}^*$ soit le plus proche possible de $\mathbf{h}$ - c. a. d. si on répétait la projection avec $\mathbf{y_1, y_2, \dots, y_p}$ les p approximations $\mathbf{h_1}^*, \dots, \mathbf{h_p}^*$ devraient être le plus concentrées possible autour de $\mathbf{h}$.

Remarques 1. On peut démontrer le résultat suivant (appelé **théorème de Gauss – Markov généralisé**):

Si C est la matrice de variance – covariance du nuage des $\boldsymbol{y_i}$ *, alors la metrique rendant l'inertie des* $\boldsymbol{h_i}^*$ *minimale est exactement* C^{-1} .

Pour une seule observation $\mathbf{y}$ on obtient donc :

$$\mathbf{h}^* = X\,(\,X^{\,t}\,C^{-1}\,X)^{-1}\,X^{\,t}\,C^{-1}\,\mathbf{y}$$

2. Le modèle probabiliste est la généralisation du cas précédent pour un grand nombre de répétitions. Si on considère que $\mathbf{y}$ est une réalisation d'un vecteur aléatoire d'espérance $\mathbf{X\,\beta}$ et de matrice variance – covariance C, alors dans ce qui suit on va noter un tel modèle par le triplet ($\mathbf{y}$; $\mathbf{X\,\beta}$; C).

Dans les deux cas : la régression multiple linéaire et le modèle linéaire général, on a la même formule : $\mathbf{y = X\,\beta + e}$, où $\mathbf{e}$ le vecteur aléatoire des résidus.

Cependant, les hypothèses sont différentes : dans le cas du modèle linéaire général $\mathbf{X}$ est un tableau de données et le vecteur des résidus a une matrice variance – covariance quelconque, alors qu'en régression multiple $\mathbf{X}$ est aléatoire et le vecteur $\mathbf{e}$ a pour matrice

variance – covariance $\sigma^2 I_n$ puisque l'hypothèse d'échantillonage suppose des observations indépendantes.

Les objectifs sont également différents : en régression multiple on cherche à approximer $\mathbf{y}$ le mieux possible ; dans le modèle linéaire général on estime l'effet moyen des variables explicatives.

Remarque 3. Il faut souligner que le terme « linéaire » s'applique en fait au vecteur $\boldsymbol{\beta}$ et non aux variables explicatives ; par exemple : la régression polynomiale

$$\beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$$

est un cas particulier du modèle linéaire général où on considère les k variables explicatives $X, X^2, \ldots, X^k$.

3. Conditions de Gauss – Markov. Estimations des paramètres du modèle $(y ; X \beta ; \sigma^2 Id)$

Pour que les estimateurs des paramètres possèdent des certaines propriétés statistiques (utiles dans les applications) nous allons utiliser les hypothèses suivantes, dites **conditions de Gauss – Markov** :

$$E (\varepsilon_i) = 0 \qquad (1)$$

$$E (\varepsilon_i^2) = \sigma^2 \qquad (2)$$

$$E (\varepsilon_i \varepsilon_j) = 0 \quad \text{si } i \neq j \qquad (3)$$

pour toutes $i, j = 1, \ldots, n$.

Matriciellement on écrit :

$$E (\boldsymbol{\varepsilon}) = \mathbf{0} , \quad E (\boldsymbol{\varepsilon}\, \boldsymbol{\varepsilon}^t) = \sigma^2 I_n$$

Remarques 1. Tenant compte de conditions Gauss – Markov on a :

$$E (\mathbf{y}) = X \boldsymbol{\beta}$$

et

$$\text{cov} (\mathbf{y}) = E ((\mathbf{y} - X \boldsymbol{\beta}) (\mathbf{y} - X \boldsymbol{\beta})^t) = E (\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^t) = \sigma^2 I_n .$$

2. En utilisant la relation (6) du premier paragraphe on obtient aussi :

$$E (\mathbf{e} \mathbf{e}^t) = M E (\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^t) M = \sigma^2 I_n$$

puisque M est idempotente (voir la remarque 5. 1.). Alors il vient :

$$V (e_i) = \sigma^2 m_{ii} = \sigma^2 (1 – a_{ii})$$

où par m_{ij} , respectivement a_{ij} on désigne le ij – élément de M, respectivement de A.

Remarque 3. Tenant compte que la variance est positive, de la relation précédente il résulte que $a_{ii} \leq 1$.

Soit $\mathbf{b} = (X^t X)^{-1} X^t \mathbf{y}$ l'estimation du vecteur $\boldsymbol{\beta}$ par la méthode des moindres carrés.

Théorème 1. *Sous les conditions de Gauss – Markov,* $\boldsymbol{b}$ *est un estimateur sans biais de* $\boldsymbol{\beta}$ *; en plus, la matrice variance – covariance de* $\boldsymbol{b}$ *est de la forme :*

$$cov (\boldsymbol{b}) = \sigma^2 (X^t X)^{-1}.$$

Démonstration. Puisque X est un opérateur linéaire et elle est constante, il vient :

$$E (\mathbf{b}) = (X^t X)^{-1} X^t E (\mathbf{y}).$$

D'autre part on a $E(\mathbf{y}) = X\boldsymbol{\beta}$ par l'hypothèse du modèle linéaire général et tenant compte de la condition (1) ; alors on obtient :

$$E(\mathbf{b}) = (X^t X)^{-1} X^t X \boldsymbol{\beta} = \boldsymbol{\beta}$$

c'est-à-dire la première affirmation du théorème.

Si on note $D = (X^t X)^{-1} X^t$, pour la matrice variance – covariance de $\mathbf{b}$ on a :

$$cov(\mathbf{b}) = D\, cov(\mathbf{y})\, D^t = (X^t X)^{-1} X^t\, cov(\mathbf{y})\, X (X^t X)^{-1}$$

et en utilisant maintenant la remarque 1 : $cov(\mathbf{y}) = \sigma^2 I_n$ on obtient la formule désirée

$$cov(\mathbf{b}) = \sigma^2 (X^t X)^{-1}. \qquad \square$$

Remarque 4. Pour prouver la première affirmation du théorème on a utilisé seulement la condition de Gauss – Markov (1) .

A présent nous allons chercher parmi les estimateurs sans biais de $\boldsymbol{\beta}$ celui de variance minimale. Soit $B\mathbf{y}$ un autre estimateur sans biais de $\boldsymbol{\beta}$ et on considère la différence de ces deux estimateurs $(X^t X)^{-1} X^t \mathbf{y} - B\mathbf{y}$; puisqu'ils sont sans biais on a

$$(X^t X)^{-1} X^t X \boldsymbol{\beta} = B X \boldsymbol{\beta}, \text{ pour tout } \boldsymbol{\beta}.$$

Alors on obtient $B X = I_{n+1}$ et si on note $B = (X^t X)^{-1} X^t + C$, on en déduit

$$C X = \mathbf{0} \qquad (4)$$

Pour la matrice variance – covariance de $B\mathbf{y}$ on a :

$$cov(B\mathbf{y}) = B\, cov(\mathbf{y})\, B^t = [(X^t X)^{-1} X^t + C]\, \sigma^2 I_n\, [(X^t X)^{-1} X^t + C]^t =$$

$$= \sigma^2 [(X^t X)^{-1} X^t X (X^t X)^{-1} + C X (X^t X)^{-1} + (X^t X)^{-1} X^t C^t + C C^t] =$$

$= \sigma^2 [(X^t X)^{-1} + C C^t]$

puisque $C X = \mathbf{0}$ (voir la relation 4). Alors il vient :

$$\text{cov} (B\mathbf{y}) = \text{cov} (\mathbf{b}) + \sigma^2 C C^t$$

et donc il résulte que chaque composante de $\mathbf{b}$ est un estimateur meilleur que $(B\mathbf{y})_i$. D'autre part, cov (B**y**) - cov (**b**) est semi – positive définie car les termes diagonaux de la matrice $C C^t$ sont positifs ou nuls.

Nous avons obtenu le résultat suivant appelé **théorème de Gauss – Markov** :

Théorème 2. ***b*** *est de tous estimateurs sans biais de* ***β****, de la forme B****y****, celui de variance minimale dans le sens précisé ci – dessus.*

Corollaire. *Si pour* $n \to \infty$ *on a* $tr [(X^t X)^{-1}] \to 0$, *alors* ***b*** *est un estimateur consistant de* ***β****.*

Démonstration. Vu la formule :

$$\text{cov} (\mathbf{b}) = \sigma^2 (X^t X)^{-1}$$

et tenant compte de l'hypothèse il vient que cov (**b**) $\to 0$ pour $n \to \infty$, c'est-à-dire que **b** est un estimateur consistant de **β**. □

Pour estimer σ^2 il existe le résultat :

Théorème 3. $s^2 = || \boldsymbol{y} - \boldsymbol{y}^* ||^2 / (n - k - 1) = || \boldsymbol{y} - X \boldsymbol{b} ||^2 / (n - k - 1)$ *est un estimateur sans biais et consistant de* σ^2.

Démonstration. Puisque $\mathbf{e} = \mathbf{y} - \mathbf{y}^* = \mathbf{y} - X \mathbf{b}$, en utilisant la relation 6.1. on a :

$$\| \mathbf{y} - X\,\mathbf{b} \|^2 = \mathbf{e}^t\,\mathbf{e} = \boldsymbol{\varepsilon}^t\, M^t\, M\, \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^t\, M\, \boldsymbol{\varepsilon}$$

car M est une matrice symétrique et idempotente (voir la remarque 5.1.). Alors on obtient :

$$\| \mathbf{y} - X\,\mathbf{b} \|^2 = \sum_{i=1}^{n} m_{ii}\, \varepsilon_i^2 + \sum_{i \neq j}^{n} m_{ij}\, \varepsilon_i\, \varepsilon_j \,. \qquad (5)$$

où par m_{ij} on désigne les éléments de la matrice $M = I_n - A$, avec $A = X\,(X^t X)^{-1} X^t$.

En prenant l'espérance dans les deux membres de la relation (5) on a:

$$E\,(\,\| \mathbf{y} - X\,\mathbf{b} \|^2\,) = \sum_{i=1}^{n} m_{ii}\, E(\varepsilon_i^2) + \sum_{i \neq j}^{n} m_{ij}\, E(\varepsilon_i\, \varepsilon_j)\,.$$

En utilisant à présent les conditions de Gauss - Markov (2) et (3) il vient :

$$E\,(\,\| \mathbf{y} - X\,\mathbf{b} \|^2\,) = \sum_{i=1}^{n} m_{ii}\, \sigma^2 = \sigma^2\, \mathrm{tr}\, M\,.$$

D'autre part, vu les propriétés de la trace d'une matrice on a :

$$\mathrm{tr}\, A = \mathrm{tr}\,[X\,(X^t X)^{-1} X^t\,] = \mathrm{tr}\,[X^t X\,(X^t X)^{-1}] = \mathrm{tr}\, I_{k+1} = k + 1$$

$$\mathrm{tr}\, M = \mathrm{tr}\, I_n - \mathrm{tr}\, A = n - k - 1$$

d'où la première affirmation du théorème:

$$E\,(\,\| \mathbf{y} - X\,\mathbf{b} \|^2\,) = \sigma^2\,(\,n - k - 1\,)$$

Pour la consistance on écrit $s^2 = \boldsymbol{\varepsilon}^t\, M\, \boldsymbol{\varepsilon} / (n - k - 1) = \boldsymbol{\varepsilon}^t\,(\,I_n - A)\,\boldsymbol{\varepsilon} / (n - k - 1)$; on va utiliser l'inégalité de Markov :

$$\text{Prob}\,(\,|\,X\,|\,\geq a) \leq \frac{E(|\,X\,|^{r}\,)}{a^{r}}$$

et premièrement il faut évaluer $E\,(\boldsymbol{\varepsilon}^t\,A\,\boldsymbol{\varepsilon}\,)$. Tenant compte que $\boldsymbol{\varepsilon}^t\,A\,\boldsymbol{\varepsilon}$ est un scalaire il vient que sa trace coïncide avec $\boldsymbol{\varepsilon}^t\,A\,\boldsymbol{\varepsilon}$ et alors on a :

$$E\,(\boldsymbol{\varepsilon}^t\,A\,\boldsymbol{\varepsilon}\,) = E\,(\,tr\,(\boldsymbol{\varepsilon}^t\,A\,\boldsymbol{\varepsilon}\,)) = E\,(\,tr\,(A\boldsymbol{\varepsilon}^t\,\boldsymbol{\varepsilon}\,)) = tr\,(\,A\,E\,(\boldsymbol{\varepsilon}^t\,\boldsymbol{\varepsilon}\,)) = \sigma^2\,tr\,A = \sigma^2\,(\,k+1).$$

Vu l'inégalité de Markov pour $r = 1$ on obtient :

$$\text{Prob}\,[(\,\boldsymbol{\varepsilon}^t\,A\,\boldsymbol{\varepsilon}\,)(\,n-k-1)^{-1} \geq \varepsilon\,] \leq \frac{\sigma^{2}\,(k+1)}{\varepsilon\,(n-k-1)} \to 0 \quad \text{pour } n \to \infty \qquad (6)$$

Maintenant on écrit $\boldsymbol{\varepsilon}^t\,\boldsymbol{\varepsilon}\,/\,(n-k-1) = (\,\boldsymbol{\varepsilon}^t\,\boldsymbol{\varepsilon}\,/\,n\,)\,\dfrac{n}{n-k-1}$ et en utilisant la loi de grands nombres on a :

$$(\,\boldsymbol{\varepsilon}^t\,\boldsymbol{\varepsilon}\,/\,n\,) \xrightarrow{prob} \sigma^{2}\;; \qquad (7)$$

puisque $\dfrac{n}{n-k-1} \to 1$ pour $n \to \infty$ il résulte finalement de relations (6) et (7) que s^2 est aussi un estimateur consistant de σ^2. □

Remarques 4. Si σ^2 n'est pas connu, alors un estimateur sans biais et consistant de cov (**b**) est donné par la formule

$$s^2\,(\,X^t\,X)^{-1}.$$

5. Géométriquement, $\mathbf{y}^* = X\,\mathbf{b}$ est la projection orthogonale de **y** sur le sous – espace V engendré par $\mathbf{1, x_1, x_2, \ldots, x_k}$ (voir la remarque 6.1.).

On sait que l'opérateur de projection orthogonale sur V est $A = X\,(\,X^t\,X)^{-1}\,X^t$ et alors le projecteur sur le complément orthogonal de V est exactement $I_n - A = M$; puisque

$\mathbf{y} - X\,\mathbf{b}$ est orthogonal à V, on observe immédiatement que $\mathbf{y} - X\,\mathbf{b}$ est égal à $M\,\mathbf{e}$. Ainsi on obtient une autre démonstration de la première affirmation du théorème 3.

Cas particulier. Si e_i suit une loi normale $N(0; \sigma)$ pour tout $i = 1, \ldots, n$, alors le vecteur aléatoire $\mathbf{y}$ est gaussien n – dimensionnel : $\mathbf{y} \in N_n(X\,\boldsymbol{\beta}, \sigma I_n)$ et sa densité est de la forme :

$$D(\mathbf{y}, \boldsymbol{\beta}, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp[-1/2\sigma^2 (\mathbf{y} - X\,\boldsymbol{\beta})^t (\mathbf{y} - X\,\boldsymbol{\beta})].$$

On peut prouver le résultat suivant :

Proposition 1. *Les estimateurs de maximum de vraisemblance de* $\boldsymbol{\beta}$ *et* σ^2 *sont* $\boldsymbol{b} = (X^t X)^{-1} X^t \mathbf{y}$, *respectivement* $\| \mathbf{y} - X\,\boldsymbol{b} \|^2 / n$, *le dernier étant biaisé.*

Puisque la propriété du maximum de vraisemblance ne donne pas des informations sur l'optimalité des estimateurs, on va déterminer des statistiques pour les paramètres inconnus $\boldsymbol{\beta}$ et σ^2 afin d'étudier l'efficacité de leurs estimateurs.

La densité de $\mathbf{y}$ s'écrit :

$$D(\mathbf{y}, \boldsymbol{\beta}, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp[-(1/2\sigma^2)(\mathbf{y}\,\mathbf{y}^t - 2\,\boldsymbol{\beta}^t X^t \mathbf{y} + \boldsymbol{\beta}^t X^t X\,\boldsymbol{\beta})];$$

en utilisant les notations suivantes :

$$\mathbf{a}(\boldsymbol{\beta}, \sigma) = \left(-\frac{1}{2\sigma^2}, \frac{\beta_0}{\sigma^2}, \frac{\beta_1}{\sigma^2}, \ldots, \frac{\beta_k}{\sigma^2} \right),$$

$$T(\mathbf{y}) = \begin{pmatrix} y^t y \\ X^t y \end{pmatrix} \quad \text{et} \quad C(\boldsymbol{\beta}, \sigma) = -(\boldsymbol{\beta}^t X^t X\,\boldsymbol{\beta}) / 2\sigma^2$$

il vient

$$D(\mathbf{y}, \boldsymbol{\beta}, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp[\mathbf{a}(\boldsymbol{\beta}, \sigma) T(\mathbf{y}) + C(\boldsymbol{\beta}, \sigma)].$$

Puisque le vecteur $\mathbf{y}$ (c. a. d. le domaine de définition de T) ne dépend pas de $\boldsymbol{\beta}$ ni de σ^2, et le rang de X est égal à $k+1$, il résulte que l'opérateur T est bijectif ; alors on sait que $T(\mathbf{y})$ est une statistique et donc les estimateurs concernés – étant des fonctions de T – sont des estimateurs sans biais de variance minimale de $\boldsymbol{\beta}$ et σ^2 .

Remarques 6. Le vecteur $\mathbf{b} = (X^t X)^{-1} X^t \mathbf{y}$ est gaussien :

$$\mathbf{b} \in N_{k+1}(\boldsymbol{\beta}, (X^t X)^{-1} \sigma^2)$$

puisqu'il est le transformé linéaire d'un vecteur gaussien.

7. Car le vecteur $\mathbf{e}$ est gaussien - les e_i suivent indépendamment des lois normales $N(0; \sigma)$ - il résulte que $\| \mathbf{e} \|^2 / \sigma^2$ suit une loi khi – carré d'ordre n. En utilisant maintenant le théorème de Pythagore dans le triangle rectangle de côtés $\mathbf{y}, \mathbf{y}^* = X\mathbf{b}, X\boldsymbol{\beta}$:

$$\| \mathbf{e} \|^2 = \| \mathbf{y} - X\mathbf{b} \|^2 + \| X\boldsymbol{\beta} - X\mathbf{b} \|^2$$

on obtient que $\| X\boldsymbol{\beta} - X\mathbf{b} \|^2 / \sigma^2$ suit une loi khi – carré d'ordre $k+1$ et $\| \mathbf{y} - X\mathbf{b} \|^2 / \sigma^2$ suit une loi khi – carré d'ordre $n - k - 1$.

Ainsi on peut déterminer des intervalles de confiance pour σ.

4. Tests dans le modèle linéaire

Tout d'abord on présente la liaison qui existe entre la somme des carrés résiduelle, la somme des carrés des observations et la somme des carrés des valeurs « ajustées » - c.a.d. issues d'un ajustement .

Proposition 1. *On a :*

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} (y_i^*)^2 = (\sum_{i=1}^{n} y_i^2 - n\bar{y}) - (\sum_{i=1}^{n} (y_i^*)^2 - n\bar{y})$$

Démonstration. Il vient successivement :

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (y_i - y_i^* + y_i^*)^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n} (y_i^*)^2 + 2 \sum_{i=1}^{n} e_i y_i^* ;$$

tenant compte à présent de la proposition 1.1. on obtient la première égalité de l'énoncé. La deuxième affirmation est immédiate. □

Corollaire 2. (voir aussi 1.3.b.) *Si* $\beta_0 \neq 0$ *on a:*

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i^* - \bar{y})^2$$

Démonstration. Vu le corollaire 1.2. il vient que $\sum_{i=1}^{n} e_i = 0$ et alors on a :

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i^*$$

c.a.d. : la moyenne des observations coïncide avec la moyenne des valeurs ajustées. En utilisant maintenant la deuxième affirmation de la proposition précédente on obtient successivement :

$$\sum_{i=1}^{n} e_i^2 = (\sum_{i=1}^{n} y_i^2 - n\bar{y}) - (\sum_{i=1}^{n}(y_i^*)^2 - n\bar{y}) = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i^* - \bar{y})^2 \quad \square$$

Car la somme des carrés résiduelle dépends de l'unité utilisée pour mesurer les observations y_i , on introduit **le coefficient de détermination**, à savoir :

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{pour } \beta_0 \neq 0 \qquad (1)$$

et

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} \quad \text{pour } \beta_0 = 0 \qquad (2)$$

Remarques 1. Il est évident qu'on ne peut pas comparer, en termes de R^2, les modèles à terme initial β_0 avec les modèles sans β_0 .

2. En utilisant le corollaire 2 il résulte de la formule (1) que $R^2 \in [0, 1]$; si $R^2 = 1$ il vient que $y_i = y_i^*$ pour tout $i = 1, \ldots, n$ et donc l'ajustement est optimal.

En conclusion, le coefficient de détermination est l'outil parfait pour l'ajustement de $\mathbf{y}$ par $\mathbf{y}^*$.

La racine carré du coefficient de détermination est exactement le coefficient de corrélation R entre les observations $y_1, \ldots, y_n$ et les valeurs $y_1^*, \ldots, y_n^*$ - c.a.d. la valeur maximale du coefficient de corrélation linéaire simple entre les composantes du vecteur $\mathbf{y}$ et les composantes de tout vecteur de la forme $X\,\mathbf{b}$.

Dans le cas du modèle qui possède un terme initial β_0 on définit :

$$R = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(y_i^* - \bar{y})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (y_i^* - \bar{y})^2}} \qquad (3)$$

On va prouver que son carré est le coefficient de détermination donné par la formule (1), exprimé en termes de variance expliquée ; en effet on a successivement :

$$\sum_{i=1}^{n} (y_i - \bar{y})(y_i^* - \bar{y}) = \sum_{i=1}^{n} (y_i - y_i^* + y_i^* - \bar{y})(y_i^* - \bar{y}) =$$

$$= \sum_{i=1}^{n} (y_i - y_i^*)(y_i^* - \bar{y}) + \sum_{i=1}^{n} (y_i^* - \bar{y})^2 = \sum_{i=1}^{n} e_i y_i^* - \bar{y} \sum_{i=1}^{n} e_i + \sum_{i=1}^{n} (y_i^* - \bar{y})^2 =$$

$$= \sum_{i=1}^{n} (y_i^* - \bar{y})^2$$

tenant compte de la proposition 1.1. et du corollaire 1.2.

Alors, pour $\beta_0 \neq 0$, il vient que

$$R^2 = \frac{\sum_{i=1}^{n} (y_i^* - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2};$$

en utilisant maintenant le corollaire 2 de ci-dessus, on obtient la formule (1).

Remarques 3. Car R est positif il résulte aussi que $R \in [0, 1]$.

4. Le coefficient de détermination R^2 est utile dans l'analyse de variance de la régression (pour tester la qualité de l'ajustement).

5. Dans le cas du modèle qui possède un terme initial β_0 , on utilise parfois le coefficient de détermination **ajusté**, à savoir :

$$R^2_a = 1 - \frac{(\sum_{i=1}^{n} e_i^{\,2})/(n-k-1)}{(\sum_{i=1}^{n}(y_i-\bar{y})^2)/(n-1)}.$$

On observe que le coefficient de détermination ajusté peut prendre des valeurs négatives dans un voisinage de 0.

6. Un calcul simple montre que l'estimateur de σ^2 (voir la proposition 1.3.) est égal à $\frac{n}{n-1}(1-R_a^2)\,s_y^2$ - on laisse la démonstration au soin du lecteur appliqué ! …

Dans le cas d'un modèle sans le terme initial β_0 on définit :

$$R = \frac{\sum_{i=1}^{n} y_i\, y_i^*}{\sqrt{\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} (y_i^*)^2}}$$

On va montrer que son carré est le coefficient de détermination donné par la formule (2) ; on a successivement :

$$\sum_{i=1}^{n} y_i\, y_i^* = \mathbf{y}^t\,\mathbf{y}^* = \mathbf{y}^t\, X\, \mathbf{b} = \mathbf{y}^t\, X\, (X^t X)^{-1}\, X^t\, \mathbf{y} =$$

$$= \mathbf{y}^t\, X\, (X^t X)^{-1}\, (X^t X)\, (X^t X)^{-1}\, X^t\, \mathbf{y} = \mathbf{b}^t\, X^t\, X\, \mathbf{b} = (\mathbf{y}^*)^t\, \mathbf{y}^* = \sum_{i=1}^{n} (y_i^*)^2 .$$

A présent on utilise la proposition 1 de ci-dessus, en obtenant pour le coefficient de détermination :

$$R^2 = \frac{\sum_{i=1}^{n} (y_i^*)^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2}$$

c'est-à-dire la formule (2) valable pour $\beta_0 = 0$.

Remarque 6. On observe que dans le cas d'un modèle avec $\beta_0 = 0$ on a aussi $R^2 \in [\,0, 1\,]$.

On peut démontrer que

$$\frac{\sum_{i=1}^{n} (y_i^* - \bar{y})^2}{\sigma^2}$$

suit une loi khi – carré d'ordre k pour $\beta_0 \neq 0, \beta_1 = \beta_2 = \ldots = \beta_k = 0$ car on sait que

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}$$ suit une loi khi – carré d'ordre n-k-1

pour tout vecteur $\boldsymbol{\beta}$.

Sous l'hypothèse de non-régression (c.a.d. $\beta_1 = \beta_2 = \ldots = \beta_k = 0$) il vient alors que

$$\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sigma^2}$$ suit une loi khi – carré d'ordre n-1

comme variance d'un échantillon de variables normales de mêmes lois.

En plus on a que

$$\frac{R^2}{1-R^2}\frac{n-k-1}{k} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n} e_i^{\,2}}\frac{n-k-1}{k}$$ suit une loi F (k; n-k-1)

pour $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ et β_0 quelconque.

Remarques 7. On retrouve alors, comme un cas particulier, la loi du coefficient de corrélation usuel si k = 1.

8. L'hypothèse de non-régression : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ correspond à l'annulation du coefficient de corrélation multiple théorique quand on considère la régression entre variables aléatoires.

ANALYSE FACTORIELLE

1. Vecteurs et valeurs propres

Pour chaque matrice $A \in M(n,n,R)$ on peut associer une application $f_A : R^n \to R^n$ qui sur une baisse $\{e_1, e_2, ..e_n\}$ agit de la manière suivante

$$f_A(\vec{e}_i) = \sum_{j=1}^{n} a_{j,i}\vec{e}_j$$

Dans les considérations suivantes $\{e_1, e_2, ..e_n\}$ c'est la base canonique de R^n. Si $\vec{x} = \sum_i x_i\vec{e}_i \in R^n$ alors $\vec{y} = f_A(\vec{x})$ s'écrit $\vec{y} = \sum_i y_i\vec{e}_i$ avec les coordonnées

$$\begin{pmatrix} y_1 \\ y_2 \\ y_2 \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & .. & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & . & a_{2,n} \\ a_{3,1} & a_{2,3} & .. & .. & a_{3,n} \\ .. & .. & .. & .. & .. \\ a_{n,1} & a_{n,2} & .. & .. & a_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ .. \\ x_n \end{pmatrix} \qquad (1)$$

Un vecteur $\vec{v} \in R^n$, $\vec{v} \neq \vec{0}$, s'appelle vecteur propre pour l'application linéaire f_A (ou pour la matrice A) s'il existe $\lambda \in R$ ainsi que

$$f_A(\vec{v}) = \lambda\vec{v} \qquad (2)$$

L'application f_A peut être considérée définie sur l'espace vectoriel complexe C^n avec des valeurs dans C^n par la même formule (1) dans la base canonique de C^n qui est dans le même temps la base canonique de R^n comme espace vectoriel réel. Pour $\vec{x} \in C^n$ les coordonnées $x_1, x_2, ..x_n$ et aussi les coordonnées $y_1, y_2, ..y_n$ du vecteur $\vec{y} = f_A(\vec{x})$ sont des nombres complexes. Les valeurs et les vecteurs propres sont définis dans le cas complexe par la même équation (2) avec la mention que $\vec{v} \neq \vec{0}$ et les coordonnées de $\vec{v}$ et la valeur propre λ sont généralement des nombres complexes.

Dans la suite l'application f_A sera considérée définie sur C^n a valeurs dans C^n. Si la matrice A est réelle alors le sous espace $R^n \subset C^n$ est transforme dans le sous espace R^n.

Si $\vec{v} = x_1\vec{e}_1 + ...x_n\vec{e}_n \in C^n$ alors l'équation (2) s'écrit à l'aide de (1)

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & .. & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & . & a_{2,n} \\ a_{3,1} & a_{2,3} & .. & .. & a_{3,n} \\ .. & .. & .. & .. & .. \\ a_{n,1} & a_{n,2} & .. & .. & a_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ .. \\ x_n \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ .. \\ x_n \end{pmatrix}$$

or

$$\sum_{j=1}^{1} a_{i,j}x_j = \lambda x_i$$

or

$$\begin{pmatrix} a_{1,1}-\lambda & a_{1,2} & a_{1,3} & .. & a_{1,n} \\ a_{2,1} & a_{2,2}-\lambda & a_{2,3} & . & a_{2,n} \\ a_{3,1} & a_{2,3} & .. & .. & a_{3,n} \\ .. & .. & .. & .. & .. \\ a_{n,1} & a_{n,2} & .. & .. & a_{n,n}-\lambda \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ .. \\ v_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ .. \\ 0 \end{pmatrix} \qquad (3)$$

Cet system admet une solution non identiquement nulle si et seulement si

$$\det \begin{pmatrix} a_{1,1}-\lambda & a_{1,2} & a_{1,3} & .. & a_{1,n} \\ a_{2,1} & a_{2,2}-\lambda & a_{2,3} & . & a_{2,n} \\ a_{3,1} & a_{2,3} & .. & .. & a_{3,n} \\ .. & .. & .. & .. & .. \\ a_{n,1} & a_{n,2} & .. & .. & a_{n,n}-\lambda \end{pmatrix} = 0 \qquad (4)$$

L'équation (4) s'appelle l'équation caractéristique de la matrice A et possède n racines réels ou complexes qui sont les valeurs propres de la matrice A. Les vecteurs propres se trouvent par la résolution du système (3). Un ensemble de vecteurs propres qui

correspondent à des valeurs propres distinctes est un ensemble indépendant. Tous les vecteurs propres qui correspondent à une valeur propre donnée, inclusivement le vecteur zéro forment un espace vectoriel appelé l'espace propre de la valeur propre donnée. La dimension de l'espace propre peut être plus grand que un. Dans ce cas la valeur propre c'est une racine d'ordre plus grand que un du polynôme caractéristique de la matrice.
Généralement si une valeur propre est racine d'ordre k du polynôme caractéristique alors il existe aux plus k vecteurs propres linéairement indépendants pour cette valeur propre. La dimension de l'espace correspondant propre ne dépasse pas k.

2. Matrices symétriques

Une matrice $A \in M(n,n,R)$ s'appelle matrice symétrique si $a_{i,j} = a_{j,i}$ pour chaque paire d'indices $1 \le i, j \le n$

Soit dans R^n ou dans C^n la forme bilinéaire

$$(\vec{x}, \vec{y}) = \sum_{i=1}^{n} x_i y_i \tag{5}$$

La condition de symétrie pour une matrice A c'est équivalente a

$$(f_A(\vec{x}), \vec{y}) = (\vec{x}, f_A(\vec{y})) \tag{6}$$

pour chaque $\vec{x}, \vec{y} \in R^n$. La condition s'écrit en coordonnées

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j} x_j y_i = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j} y_j x_i \tag{7}$$

Pour une matrice $A \in M(n,n,R)$ on peut associer une fonction bilinéaire $F_A : C^n \times C^n \to C$ or $F_A : R^n \times R^n \to R$ par la formule

$$F_A(\vec{x}, \vec{y}) = \sum_{i,j=1}^{n} a_{i,j} x_i y_j = (\vec{x}, f_A(\vec{y})) \tag{8}$$

Si la matrice A est symétrique alors la forme F est symétrique, c'est a dire $F_A(\vec{x}, \vec{y}) = F_A(\vec{y}, \vec{x})$, conformément a (7) et (8).

Remarquons que la formule (5) nous donne un produit scalaire sur R^n. Pour les matrices symétriques réelles nous avons le résultat suivant

Théorème 1. *Si* $A \in M(n,n,R)$ *est une matrice symétrique alors*

a) toutes les valeurs propres de A sont réelles

b) les vecteurs propres qui correspondent a des valeurs propres distinctes sont orthogonaux par rapport aux produit scalaire de R^n $(\vec{x},\vec{y}) = \sum_{i=1}^{n} x_i y_i$.

c) il existe une base de R^n *formée de n vecteurs propres orthogonaux.*

d) si $\lambda_1 \geq \lambda_2 \geq .. \geq \lambda_n$ *sont les valeurs propres alors*

$$\lambda_k = \max_{V \subset R^n,\ \dim V = k} \min_{x \in V} \frac{(f_A(\vec{x}),\vec{x})}{(\vec{x},\vec{x})} = \max_{V \subset R^n,\ \dim V = k} \left(\min_{x \in V} \frac{\sum_{i,j=1}^{n} a_{i,j} x_i x_j}{\sum_{i=1}^{n} x_i^2} \right) \qquad (9)$$

Demonstration. a) Soit $\lambda \in C$ une valeur propre et $\vec{0} \neq \vec{x} \in C^n$ le vecteur propre correspondant. Alors

$$\sum_{i,j=1}^{n} a_{i,j} x_i \overline{x}_j = \sum_{i=1}^{n} x_i \cdot \overline{\left(\sum_{j=1}^{n} a_{i,j} x_j \right)} = \sum_{i=1}^{n} x_i \overline{(\lambda x_i)} = \overline{\lambda} \sum_{i=1}^{n} x_i \overline{x}_i$$

La même somme peut s'écrire

$$\sum_{i,j=1}^{n} a_{i,j} x_i \overline{x}_j = \sum_{j=1}^{n} \overline{x}_j \left(\sum_{i=1}^{n} a_{j,i} x_i \right) = \sum_{j=1}^{n} \overline{x}_j \lambda x_j = \lambda \sum_{j=1}^{n} x_j \overline{x}_j$$

L'égalité de ces deux expressions nous donne $\lambda = \overline{\lambda}$, c'est à dire $\lambda \in R$. Comme conséquence les vecteurs propres qui correspondent a λ, solutions du système (3) peut être choisis avec les coordonnées réelles.

La démonstration peut être présentée indépendamment d'une base.

b) Si $\lambda_1 \neq \lambda_2$ sont des valeurs propres réelles de A et $\vec{v}_1 = (x_1, x_2, .. x_n)$, $\vec{v}_2 = (y_1, y_2, .. y_n)$ sont des vecteurs propres correspondants alors

$$\lambda_2(\vec{v}_1,\vec{v}_2) = (\vec{v}_1, \lambda_2 \vec{v}_2) = (\vec{v}_1, f_A(\vec{v}_2)) = (f_A(\vec{v}_1),\vec{v}_2) = (\lambda_1 \vec{v}_1, \vec{v}_2) = \lambda_1(\vec{v}_1,\vec{v}_2)$$

Parce que $\lambda_1 \neq \lambda_2$ il résulte $(\vec{v}_1, \vec{v}_2) = 0$.

c) Soit λ_1 la plus grande valeur propre et $\vec{v}_1$ le vecteur propre correspondant. Le sous espace $V \subset R^n$ des vecteurs de R^n orthogonaux a $\vec{v}_1$ c'est un espace invariant a l'application f_A parce que $(\vec{v}_1, \vec{v}) = 0$ implique

$$(\vec{v}_1, f_A(\vec{v})) = (f_A(\vec{v}_1), \vec{v}) = (\lambda_1 \vec{v}_1, \vec{v}) = \lambda_1 (\vec{v}_1, \vec{v}) = 0$$

Supposant par induction sur la dimension de l'espace que la restriction de f_A sur V (qui est aussi symétrique dans le sens de la formule (6)) a sur V (dim V=n-1), une base de n-1 vecteurs propres orthogonaux $\vec{v}_2, \vec{v}_3, .. \vec{v}_n$ alors sur R^n f_A aura la base de n vecteurs propres orthogonaux $\vec{v}_1, \vec{v}_2, .. \vec{v}_n$.

d) Soit $\vec{v}_1, \vec{v}_2, .. \vec{v}_n$ une base de R^n formée de n vecteurs propres normalises (c'est a dire $(\vec{v}_i, \vec{v}_i) = 1$, $(\vec{v}_i, \vec{v}_j) = 0$ pour $i \neq j$) de la matrice A et soit les valeurs propres en ordre décroissante $\lambda_1 \geq \lambda_2 \geq .. \geq \lambda_n$. Dans cette base $f_A(\vec{x}) = \lambda_1 x_1 \vec{v}_1 + .. \lambda_n x_n \vec{v}_n$ pour $\vec{x} = x_1 \vec{v}_1 + .. \lambda_n \vec{v}_n$. Pour le produit scalaire nous avons

$$(\vec{x}, \vec{y}) = \sum x_i y_j (\vec{v}_i, \vec{v}_j) = \sum x_i y_i$$

Pour un sous espace $V \subset R^n$, $\dim V = k$ et pour R^{n-k+1} le sous espace engendre par $\vec{v}_k, \vec{v}_{k+1}, .. \vec{v}_n$ nous avons $V \cap R^{n-k+1} \neq \{\vec{0}\}$ parce que la somme des dimensions dépasse n. Soit $\vec{0} \neq \vec{x} \in V \cap R^{n-k+1}$. Alors

$\vec{x} = b_k \vec{v}_k + b_{k+1} \vec{v}_{k+1} + .. b_n \vec{v}_n$ et nous avons

$$\frac{(f_A(\vec{x}), \vec{x})}{(\vec{x}, \vec{x})} = \frac{\lambda_k b_k^2 + .. \lambda_n b_n^2}{b_k^2 + .. b_n^2} \leq \frac{\lambda_k (b_k^2 + .. b_n^2)}{b_k^2 + .. b_n^2} = \lambda_k$$

Il résulte

$$\min_{x \in V} \frac{(f_A(\vec{x}), \vec{x})}{(\vec{x}, \vec{x})} \leq \lambda_k \qquad (10)$$

Mais si nous prenons $V' = R\vec{v}_1 \oplus R\vec{v}_2 \oplus .. \oplus R\vec{v}_k$ alors

$$\frac{(f_A(\vec{x}), \vec{x})}{(\vec{x}, \vec{x})} = \frac{\lambda_1 x_1^2 + \lambda_2 x_2^2 + .. + \lambda_k x_k^2}{x_1^2 + x_2^2 + .. + x_k^2} \geq \frac{\lambda_k (x_1^2 + x_2^2 + .. + x_k^2)}{(x_1^2 + x_2^2 + .. + x_k^2)} = \lambda_k$$

et pour ce sous espace nous avons

$$\min_{x \in V'} \frac{(f_A(\vec{x}), \vec{x})}{(\vec{x}, \vec{x})} = \lambda_k \qquad (11)$$

le minimum étant atteint pour $\vec{x} = \vec{v}_k$. Les formules (10) et (11) nous donnent

$$\max_{V \subset R^n,\ \dim V = k} \min_{x \in V} \frac{(f_a(\vec{x}), \vec{x})}{(\vec{x}, \vec{x})} = \lambda_k \ .$$

CQFD.

Nous appellerons $\lambda_1, \lambda_2, .., \lambda_n$ les valeurs propres de la matrice A, ou de l'application linéaire f_A ou de la forme bilinéaire F_A.

Remarque 1. *D'une manière analogue on peut démontrer*

$$\lambda_k = \min_{V \subset R^n,\ \dim V = n-k+1} \max_{x \in V} \frac{(f_A(\vec{x}), \vec{x})}{(\vec{x}, \vec{x})} \qquad (12)$$

Corollaire 1. *a) Si* $X \subset R^n$ *est un sous espace avec* $\dim(X) = m$ *alors la restriction de* F_A *sur* $X \times X$ *notée* $F_A^X : X \times X \to R$ *est symétrique et elle est donnée par la formule* $F_A^X(\vec{x}, \vec{y}) = (\vec{x}, f_A(\vec{y})) = (\vec{x}, Pr_X \circ f_A(\vec{y}))$ *ou* Pr_X *est la projection orthogonale sur* X.
b) L'application linéaire $f_A^X : X \to X$, $f_A^X(\vec{v}) = Pr_X(f_A(\vec{v}))$ *est une application symétrique par rapport aux produit scalaire, c'est a dire* $(f_A^X(\vec{x}), \vec{y}) = (\vec{x}, f_A^X(\vec{y}))$ *pour chaque* $\vec{x} \in X, \vec{y} \in X$.
c) Si $\lambda_1' \geq \lambda_2' \geq .. \geq \lambda_m'$ *sont les valeurs propres de l'application* f_A^X *(ou de la forme bilinéaire* F_A^X *) alors*

$$\lambda_1 \geq \lambda_1', \ \lambda_2 \geq \lambda_2', \ .., \ \lambda_m \geq \lambda_m'$$

Demonstration. a) et b) sont évidentes et c) résulte de (9) parce que dans la formule pour λ_k sont pris plusieurs sous espaces V de R^n que dans la formule correspondante pour λ'_k ou les sous espaces sont pris seulement de X.

Remarque 2. *D'une manière analogue de la formule (12) il résulte*

$$\lambda_n \leq \lambda_m', \ \lambda_{n-1} \leq \lambda_{m-1}, .., \ \lambda_{n-m+1} \leq \lambda_1'$$

Corollaire 2. *La somme* $\sum_{i=1}^{k} F_A(\vec{v}_i, \vec{v}_i) = \sum_{i=i}^{k} (\vec{v}_i, f_A(\vec{v}_i))$ *est maximum parmi toutes les k-uples des vecteurs* $\vec{v}_1, \vec{v}_2, .. \vec{v}_k$ *orthonormés de* R^n *si* $\vec{v}_1, \vec{v}_2, .. \vec{v}_k$ *sont les vecteurs*

propres des k plus grandes valeurs propres de f_A *et c'est minimum si* $\vec{v}_1, \vec{v}_2, ..\vec{v}_k$ *sont les vecteurs propres correspondant aux k plus petites valeurs propres de* f_A. *La valeur maximum est* $\sum_{i=1}^{k} \lambda_i$ *et la valeur minimum est* $\sum_{i=p-k+1}^{p} \lambda_i$.

Demonstration. Pour prouver l'affirmation sur le maximum remarquons que pour $X = Sp(\vec{v}_1, \vec{v}_2, ..\vec{v}_k)$ nous avons $\sum_{i=1}^{k}(\vec{v}_i, f_A(\vec{v}_i)) = \sum_{i=1}^{k}(\vec{v}_i, f_A^X(\vec{v}_i))$ et cette somme c'est la trace de l'application f_A^X et la trace ne dépend pas de la base choisie. Choisissant une base orthonormée des vecteurs propres de f_A^X nous obtenons de l'observation précédente

$$\sum_{i=1}^{k}(\vec{v}_i, f_A(\vec{v}_i)) = \sum_{i=1}^{k}(\vec{v}_i, f_A^X(\vec{v}_i)) = \sum_{i=1}^{k}\lambda'_i(\vec{v}_i, \vec{v}_i) = \sum_{i=1}^{k}\lambda'_i \le \sum_{i=1}^{k}\lambda_i$$

d'ou le résultat.

Pour le minimum l'affirmation peut être prouvée d'une manière analogue.

Remarque 3. *Si A est une matrice symétrique alors l'équation* $A\vec{x} = \lambda\vec{x}$ *nous donne par transposition* $\vec{x}^t A = \lambda\vec{x}^t$ *c'est a dire si* $\vec{x}$ *est un vecteur propre colonne (ou a gauche) de A alors* $\vec{x}^t$ *est un vecteur propre ligne (ou a droite) de A.*

3. Axes factorielles

Soit un ensemble de points de R^p dont les coordonnées forment une matrice M de $M(n, p, R)$

$$M = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & .. & x_{1,p} \\ x_{2,1} & x_{2,2} & x_{2,3} & . & x_{2,p} \\ x_{3,1} & x_{2,3} & .. & .. & x_{3,p} \\ .. & .. & .. & .. & .. \\ x_{n,1} & x_{n,2} & .. & .. & x_{n,p} \end{pmatrix} \quad (13)$$

Chaque ligne est un vecteur de R^p. Soit

$$\vec{m} = (m_1, m_2, ...m_p)$$

le centre des mass des lignes c'est à dire

$$m_j = \frac{\sum_{i=1}^{n} x_{i,j}}{n} \tag{14}$$

Soit $\{\vec{v}_1, \vec{v}_2, ...\vec{v}_p\}$ une base orthonormée de R^p. Alors les coordonnées des points $\vec{x}_i = (x_{i,1}, x_{i,2}, ...x_{i,p})$ peuvent être écrites

$$\vec{x}_i = (x_{i,1}, x_{i,2}, ...x_{i,p}) = \vec{m} + \sum_{k=1}^{p} t_{i,k}\vec{v}_k \tag{15}$$

pour i=1,2,..n. Soit $T = (t_{i,k})_{i=1,n;k=1,p}$ la matrice des coordonnées des points $\vec{x}_i - \vec{m}$ par rapport à la base $\{\vec{v}_1, \vec{v}_2, ...\vec{v}_p\}$.

La projection orthogonale de $\vec{x}_i$ sur le plan P_r de dimension r passant par $\vec{m}$ et parallèle aux vecteurs $\vec{v}_1, \vec{v}_2, ..\vec{v}_r$, $r \le p$ est

$$\vec{x}_i^{'} = \vec{m} + \sum_{k=1}^{r} t_{i,k}\vec{v}_k \tag{16}$$

Nous cherchons les vecteurs orthonormés $\vec{v}_1, \vec{v}_2, ..\vec{v}_r$ ainsi que l'expression $d^2 = \sum_{i=1}^{n} \| \vec{x}_i - \vec{x}_i^{'} \|^2$ soit minimum. Cette expression c'est la somme des carres des distances des points donnes au plan P_r. Nous avons le résultat suivant

Théorème 2. *Soit* M *une matrice de données (comme (13)) et* $\vec{m} = (m_1, m_2, ...m_p)$ *un point de* R^p *(pas nécessairement le centre des mass). Soit* M_1 *la matrice obtenue de M par le retranchement de* $\vec{m}$ *de chaque ligne. Alors une famille orthonormée* $\{\vec{v}_1, \vec{v}_2, ...\vec{v}_r\}$ *de vecteurs ligne ainsi que* $d^2 = \sum_{i=1}^{n} \| \vec{x}_i - \vec{x}_i^{'} \|^2$ *est minimum, ou* $\vec{x}_i^{'}$ *sont les projection de* $\vec{x}_i$ *sur le r plan qui passe par* $\vec{m}$ *et est parallèle aux vecteurs* $\{\vec{v}_1, \vec{v}_2, ...\vec{v}_r\}$, *c'est la famille des vecteurs propres ligne orthonormés de la matrice* $A = M_1^t \cdot M_1$ *correspondant aux premières valeurs propres de A en ordre décroissante. En plus nous avons* $d^2 = \sum_{k=r+1}^{p} \lambda_k$.

Demonstration. Soit $\vec{v}_k = (v_{k,1}, v_{k,2},...v_{k,p})$ pour k=1, 2,..p une complétion de la famille orthonormée a une base orthonormée et soit

$$V = \begin{pmatrix} v_{1,1} & v_{1,2} & v_{1,3} & .. & v_{1,p} \\ v_{2,1} & v_{2,2} & v_{2,3} & . & v_{2,p} \\ v_{3,1} & v_{2,3} & .. & .. & v_{3,p} \\ .. & .. & .. & .. & .. \\ v_{p,1} & v_{p,2} & .. & .. & v_{p,p} \end{pmatrix}$$

Alors $\{\vec{v}_1^t, \vec{v}_2^t,...\vec{v}_p^t\}$ est une base orthonormée des vecteurs colonne de R^n. Nous avons $V^t \cdot V = V \cdot V^t = Id$. Soit les coordonnées de $\vec{x}_i$ et $\vec{x}_i^{'}$ exprimées par (15) et (16) a l'aide de la matrice $T = (t_{i,k})_{i=1..n;\ k=1..p}$. Soit $(\vec{t}_k)^t = (t_{1,k}, t_{2,k},...t_{n,k})$. L'équation (15) s'écrit $M_1 = T \cdot V$ or $T = M_1 \cdot V^t$ or $\vec{t}_k = M_1 \cdot \vec{v}_k{}^t$ pour k=1,2,...p.

Nous avons

$$d^2 = \sum_{i=1}^{n} \| \vec{x}_i - \vec{x}_i^{'} \|^2 = \sum_{i=1}^{n} \left\| \sum_{k=r+1}^{p} t_{i,k} \vec{v}_k \right\|^2 = \sum_{i=1}^{n} \sum_{k=r+1}^{p} (t_{i,k})^2 = \sum_{k=r+1}^{p} \sum_{i=1}^{n} (t_{i,k})^2$$

$$= \sum_{k=r+1}^{p} (\vec{t}_k)^t \cdot \vec{t}_k = \sum_{k=r+1}^{p} \vec{v}_k \cdot M_1{}^t \cdot M_1 \cdot \vec{v}_k{}^t = \sum_{k=r+1}^{p} (\vec{v}_k \cdot A \cdot \vec{v}_k{}^t) = \sum_{k=r+1}^{p} (\vec{v}_k^t, A \cdot \vec{v}_k{}^t)$$

Conformément au corollaire 2 si les vecteurs $\vec{v}_{r+1}^t, \vec{v}_{r+2,...}^t \vec{v}_p^t$ sont les vecteurs propres colonne de A correspondant aux p-r les plus petites valeurs propres λ_{r+1} ,.. λ_p alors d^2 est minimum. Comme l'espace engendre par $\{\vec{v}_1^t, \vec{v}_2^t,...\vec{v}_r^t\}$ est orthogonal a l'espace engendre par $\{\vec{v}_{r+1}^t, \vec{v}_{r+2,...}^t \vec{v}_p^t\}$ nous pouvons prendre pour $\{\vec{v}_1^t, \vec{v}_2^t,...\vec{v}_r^t\}$ les vecteurs propres colonne de A correspondant aux valeurs propres $\lambda_1, \lambda_2..., \lambda_r$. Le corollaire 2 nous donne $d^2 = \sum_{k=r+1}^{p} \lambda_k$.

CQFD.

Dans la suite nous prenons pour $\vec{m}$ le centre des mass des points donnés par (13).

Les axes passant par $\vec{m}$ et parallèles aux vecteurs $\vec{v}_1, \vec{v}_2,...\vec{v}_p$ s'appelles les axes factorielles de l'ensemble donnée des points. Les coordonnées dans la base canonique des points projetés $\vec{x}_i^{'}$ forment une matrice M'. Par définition M_1 et $M_1^{'}$ s'obtient de M respectivement M' par retranchement de $\vec{m}$ dans chaque ligne. Nous avons

$$\| M - M' \|^2 = \sum_{i=1}^{n}\sum_{k=1}^{p}\left(x_{i,k} - x'_{i,k}\right)^2 = \sum_{i=1}^{n} \| \vec{x}_i - \vec{x}'_i \|^2 = \sum_{k=r+1}^{p} \lambda_k = \| M_1 - M'_1 \|^2$$

$$\left\|M_1\right\|^2 = \sum_{k=1}^{p}\sum_{i=1}^{n} x_{i,k}^2 = \mathrm{Trace}\left(M_1^t \cdot M_1\right) = \sum_{i=1}^{p}\lambda_i$$

Si le rapport

$$\frac{\sum_{k=1}^{r}\lambda_k}{\sum_{k=1}^{p}\lambda_k} = 1 - \frac{\sum_{k=r+1}^{p}\lambda_k}{\sum_{k=1}^{p}\lambda_k} = 1 - \frac{\| M_1 - M'_1 \|^2}{\| M_1 \|^2}$$

est assez proche de 1 alors $\frac{\|M_1 - M'_1\|^2}{\|M_1\|^2}$ est assez petite et nous pouvons considérer M'_1 une approximation bonne pour la matrice M_1 ou la même chose M' une bonne approximation de la matrice M. L'avantage c'est que les vecteurs ligne de la matrice M' peut être représentés par leurs coordonnées dans la basse des vecteurs propres $\vec{v}_1, \vec{v}_2, ...\vec{v}_p$, c'est a dire par la matrice $\left(t_{i,k}\right)_{i=1,n;k=1,r}$. A la place de np coefficients des données (13) nous avons besoin pour construire M' de p^2 coefficients pour la basse $\{\vec{v}_1, \vec{v}_2, ...\vec{v}_p\}$, nr coefficients $\left(t_{i,k}\right)_{i=1,n;k=1,r}$ et p coefficients $\left(m_i\right)_{i=1..n}$. Si le nombre n des données est assez grand par rapport au nombre p des caractéristiques de chaque point, et le rapport $\frac{\sum_{k=1}^{r}\lambda_k}{\sum_{k=1}^{p}\lambda_k}$ est assez proche de 1 pour r petit par rapport a p alors M' est une bonne approximation de M et le rapport entre les nombres des coefficients nécessaires pour représenter M' et le nombre des coefficients de M est aussi petit

$$\frac{p^2 + nr + p}{np} = \frac{p+1}{n} + \frac{r}{p}$$

ce que signifie une économie de mémoire.

Exemple 1. Une image bitmap blanc et noire c'est une matrice avec des valeurs entre 0 et 255. Nous pouvons appliquer la théorie des axes factorielle pour réduire le nombre des coefficients nécessaires pour retenir l'information contenue dans l'image. Les calculs suivants ont été effectues a l'aide de MATH CAD. Nous commençons par lire une image :

M := READBMP("e:\scoala\an6-2004-2005\dani.bmp")

M

Les indices des vecteurs et des matrices commencent a zéro dans mathcad. Ici nous avons

$$n := \text{rows}(M) \qquad p := \text{cols}(M)$$

$$n = 686 \qquad p = 500$$

L'indice des lignes varie de 0 à n-1 et l'indice des colonnes varie de 0 à p-1. Dans les calculs suivants nous nous retenons seulement

$$r := 50$$

axes factorielles de p=500.

La matrice M_1 et la matrice A sont obtenus par les instructions suivants

$$j := 0..\text{cols}(M) - 1$$

$$m_j := \text{mean}\left(M^{\langle j \rangle}\right)$$

$$i := 0..\text{rows}(M) - 1$$

$$M1_{i,j} := M_{i,j} - m_j$$

$$A := M1^T \cdot M1$$

Les premières et les dernières, par l'ordre de grandeur, des valeurs propres sont les suivantes

$$vp := eigenvals(A).$$

$vp =$

	0
0	329.881
1	332.189
2	336.233
3	323.797
4	316.748
5	310.989
6	304.48
7	302.807
8	349.915
9	345.069
10	356.955
11	359.15
12	288.32
13	283.034
14	281.704
15	271.008
16	266.016
17	260.739

$reverse(vp) =$

	0
0	$1.01 \cdot 10^9$
1	$1.966 \cdot 10^8$
2	$5.66 \cdot 10^7$
3	$3.463 \cdot 10^7$
4	$1.928 \cdot 10^7$
5	$1.734 \cdot 10^7$
6	$1.483 \cdot 10^7$
7	$9.992 \cdot 10^6$
8	$8.033 \cdot 10^6$
9	$7.505 \cdot 10^6$
10	$6.319 \cdot 10^6$
11	$5.529 \cdot 10^6$
12	$5.084 \cdot 10^6$
13	$4.257 \cdot 10^6$
14	$3.43 \cdot 10^6$
15	$3.372 \cdot 10^6$
16	$2.957 \cdot 10^6$
17	$2.416 \cdot 10^6$

On voit la grande différence entre les plus petites valeurs propres qui sont de l'ordre 300 et les plus grandes qui sont de l'ordre 10^9. Les valeurs propres doivent être toutes positives et vp en ordre croissante mais par la cause des erreurs numériques les plus petites valeurs propres ne sont toutes correctement rangées. Utilisant seulement r valeurs propres nous trouvons

$$S := \sum vp \qquad S = 1.457 \times 10^9$$

$$poz := p - r$$

$$poz = 450$$

$$S1 := \sum_{i=poz}^{p-1} vp_i$$

$$rap := \frac{S1}{S} \qquad rap = 0.99$$

Les vecteurs propres, ici vecteurs colonne, ranges dans l'ordre des valeurs propres sont obtenus par l'instruction

$$vects := eigenvecs(A).$$

Ici *vects* c'est la transpose de la matrice V de la théorie mais avec les vecteurs dans l'ordre

décroissante des valeurs propres. Parce que les valeurs propres sont rangées dans l'ordre décroissante nous extrayons les derniers r vecteurs propres de la matrice vects.

$$\text{vects1} := \text{submatrix}(\text{vects}, 0, p-1, \text{poz}, p-1)$$

Les coordonnées des points $x_i^{'}$ dans la base des vecteurs propres *vects1* dans le plan P_r sont données par

$$\text{coord} := M1 \cdot \text{vects1}$$

Ici *coord* c'est la matrice $T = \left(t_{i,k}\right)_{i=1,n;\,k=1,r}$ mais les vecteurs de la base de P_r sont ranges dans l'ordre croissant des valeurs propres. Cette matrice a un nombre d'éléments plus petits que M, précisément la fraction r/n du nombre de coefficients de M.

Maintenant nous construisons une approximation MA de M a l'aide des coefficients de T (*coord* dans le programme)

$$M1A := \left(\text{vects1} \cdot \text{coord}^T\right)^T$$

$$i := 0..\text{rows}(M1) - 1$$

$$j := 0..\text{cols}(M1) - 1$$

$$MA_{i,j} := M1A_{i,j} + m_j$$

La différence entre la matrice initiale M et la matrice MA construite à l'aide de T est (partiellement)

M – MA =

	0	1	2	3	4	5	6
0	-0.	-2.2	0.7	-2.5	-0.8	-0.3	-0.8
1	0.7	2.7	1.0	-1.8	-0.0	$1.185 \cdot 10$	-2.3
2	2.5	2.4	3.8	2.	0.1	-1.9	-2.3
3	4.1	3.1	2.5	0.	-0.7	-0.5	-0.8
4	3.8	2.7	4.0	3.1	0.	-2.2	-1.7
5	4.1	4.7	4.	2.6	1.3	1.8	1.
6	4.1	3.5	4.8	2.	2.0	2.3	0.6
7	2.	3.9	4.9	2.8	3.7	1.0	0.
8	3.2	4.4	4.3	4.5	3.7	-0.4	-1.1
9	4.3	3.6	3.8	4.0	3.1	-0.0	-1.6
10	2.7	4.0	4.2	2.6	1.9	1.4	0.9
11	2.8	3.0	2.3	1.2	1.5	0.	0.0
12	2.0	2.2	1.5	3.	1.6	0.9	0.1
13	3.5	2.1	0.7	5.1	-2		1.4
14	4.6	0.7	0.0	3.9	0.4	1.9	2.2
15	1.9	0.9	1.5	2.1	1.	1.8	-0.3

Pour voir mieux la différence représentons les images correspondantes a ces matrices

M

MA

On voit la différence entre l'image originale et l'image construite a l'aide de r=50 vecteurs propres de 500.

ANNEXE

Inverse généralisée d'une matrice

Pour les matrices singulières on va introduire la notion suivante

Définition. Une matrice B est dite **inverse généralisée** de A si elle remplit la relation $ABA = A$; on va noter l'inverse généralisée de A par A^- .

Exemples 1. Soit X une matrice avec m lignes et n colonnes, dont le rang $= n \leq m$; alors une inverse généralisée de X est donnée par la formule

$$X^- = (X^t X)^{-1} X^t$$

puisqu'on a : $X X^- X = X (X^t X)^{-1} X^t X = X$.

2. Pour une matrice idempotente A il vient $A^- = A$, car on a $AAA = AA = A$.

Remarques 1. Toutes les matrices possèdent des inverses généralisées ; pour les matrices non-singulières les inverses généralisées coïncident avec leures inverses car on a $AA^{-1}A = A$.

2. Si la matrice A est singulière, alors son inverse généralisée A^- n'est pas unique.

3. Si A^- est l'inverse généralisée de A : $A A^- A = A$, alors il vient

$$A^- A = A^- A A^- A$$

d'où il résulte que la matrice $A^- A$ est idempotente. D'une manière analogue on montre que $A A^-$ est idempotente.

4. Puisqu'on a : $\text{rang}(A) = \text{rang}(A A^- A) \leq \text{rang}(A^- A) \leq \text{rang}(A)$, il résulte que $\text{rang}(A) = \text{rang}(A A^-)$.

En général, on peut prouver le résultat suivant :

Proposition. *Soit A une matrice avec m lignes et n colonnes ; alors B est une inverse généralisée de A si et seulement si :*

$$BA \text{ est idempotente et } \text{rang}(BA) = \text{rang}(A)$$

ou

$$AB \text{ est idempotente et } \text{rang}(AB) = \text{rang}(A) .$$

Alors on démontre aisément les affirmations :

Corollaire. 1. *On a : $A (A^t A)^- A^t A = A$ et $A^t A (A^t A)^- A^t = A^t$ (c'est-à-dire $(A^t A)^- A^t$ est l'inverse généralisée de A, respectivement $A (A^t A)^-$ est l'inverse généralisée de A^t).*

2. La matrice $A (A^t A)^- A^t$ est symétrique, idempotente, de rang (A) et unique.

Démonstration. Pour la première partie on utilise la définition de l'inverse généralisée (à savoir $A^t A A^t A A^t A = A^t A$). Alors des simples calculs nous donnent les affirmations du point 2 ; par exemple :

$$\text{rang}(A) = \text{rang}[A (A^t A)^- A^t A] \leq \text{rang}[A (A^t A)^- A^t] \leq \text{rang}(A) .$$

B I B L I O G R A P H I E

1. **Armeanu I., Petrehus V**., *Probabilitati si statistică aplicate in biologie*, Editions MATRIX ROM, Bucarest, 2006.
2. **Costinescu C**., *Probabilités et statistique mathématique (recueil de problèmes)*, Editions CONSPRESS, Bucarest, 2003.
3. **Costinescu C., Popescu S.A., Mierlus-Mazilu I**., *Probabilitati si statistica tehnica (teorie si probleme)*, Editions CONSPRESS, Bucarest, 2005
4. **Gheorghe Mihoc, Virgil Craiu**, *Tratat de Statistică Matematică, vol I, II, Editura Academiei R.S.R*., 1976-1977
5. **Iosifescu, M., Mihoc, Gh., Theodorescu, R.,** *Teoria probabilitatilor şi statistică matematică*, Ed. Tehnică, Bucureşti, 1966.
6. **Monfort A**., *Cours de statistique mathématique*, Editions ECONOMICA, Paris, 1982.
7. **Petrehus V., Popescu S.A**., *Probabilitati si statistica (teorie, exemple, probleme)*, Editions de l'UTCB, Bucarest, 1997.
8. **Saporta G**., *Probabilités, analyse des données et statistique*, Editions Technip, Paris, 1990.
9. **Sen A., Srivastava M**., *Regression Analysis –Theory, Methods, and Applications*, Editions Springer- Verlag New- York Inc., 1990.
10. **Spiegel M.R**., *Probabilités et statistique. Cours et problèmes*, Série Schaum, McGraw – Hill Inc. New – York, 1981.
11. **Ventsel H**., *Théorie de probabilités*, Editions de Moscou, 1973.