Statistique euclidienne
I Individus et variables

D. Chessel & A.B. Dufour

24 avril 2006

Introduction des principes euclidiens utilisés en statistique descriptive.

Table des matières

1 Moyennes et variances

Ce cours suppose qu'on a une idée des objectifs et des méthodes élémentaires de la statistique descriptive. L'analyse des données manipule des variables, c'est-à-dire des séries de n observations d'une quantité donnée. On commence par les variables quantitatives, celles dont les valeurs appartiennent à $\mathbb{R}$. Une variable $\mathbf{x} = (x_1, x_2, ..., x_n)$ est donc un vecteur de $\mathbb{R}^n$.

1.1 Définitions élémentaires

La moyenne naturelle de $\mathbf{x}$ est $m = \mathrm{m}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$. On note souvent $\mathrm{m}(\mathbf{x}) = \overline{\mathbf{x}}$.

La variance naturelle de $\mathbf{x}$ est $v = \mathrm{v}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mathrm{m}(\mathbf{x}))^2$. On peut utiliser la notation $\mathrm{v}(\mathbf{x}) = s^2_{\mathbf{x}}$.

L'écart-type naturel de $\mathbf{x}$ est $\mathrm{et}(\mathbf{x}) = \sqrt{\mathrm{v}(\mathbf{x})}$.

Ces paramètres caractérisent la position (moyenne) et la dispersion (variance) des mesures. La médiane, les quartiles sont d'autres paramètres de position, l'intervalle interquartile est un autre paramètre de dispersion. Ces paramètres décrivent la variable vue comme n points de $\mathbb{R}$.

Calculer la moyenne et la variance de $\mathbf{1}_n = (1, 1, ..., 1)$.

Calculer la moyenne et la variance de $\mathbf{x} = (1, 2, ..., n)$.

Montrer que la moyenne minimise sur $\mathbb{R}$ la quantité (inertie autour de h) : $\mathrm{iner}(\mathbf{x},h) = \frac{1}{n}\sum_{i=1}^{n} (x_i - h)^2$.

Comparer la variance et la quantité $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (x_i - x_j)^2$.

1.2 Pondérations

Une pondération des n individus porteurs d'une mesure est un vecteur de $\mathbb{R}^n$ dont toutes les composantes sont positives et dont la somme vaut 1. On notera une pondération :

$$\mathbf{p} = (p_1, p_2, ..., p_n) \text{ avec } \sum_{i=1}^{n} p_i = 1 \text{ et } 1 \leq i \leq n \Rightarrow p_i > 0$$

La moyenne pondérée de x est : $m_{\mathbf{p}} = \mathrm{m}_{\mathbf{p}}(\mathbf{x}) = \sum_{i=1}^{n} p_i x_i$.

La variance pondérée de x est : $v_{\mathbf{p}} = \mathrm{v}_{\mathbf{p}}(\mathbf{x}) = \sum_{i=1}^{n} p_i (x_i - \mathrm{m}_{\mathbf{p}}(x))^2$.

L'écart-type pondéré de $\mathbf{x}$ est $\mathrm{et}_{\mathbf{p}}(\mathbf{x}) = \sqrt{\mathrm{v}_{\mathbf{p}}(\mathbf{x})}$. La moyenne (resp. variance) naturelle est le cas particulier de la moyenne (resp. variance) pondérée pour la pondération uniforme définie par :

$$p_i = \frac{1}{n}.$$

Quand aucune ambiguïté n'est possible, on note simplement $\mathrm{m}_{\mathbf{p}}(\mathbf{x}) = \mathrm{m}(\mathbf{x})$, $\mathrm{v}_{\mathbf{p}}(\mathbf{x}) = \mathrm{v}(\mathbf{x})$ et $\mathrm{et}_{\mathbf{p}}(\mathbf{x}) = \mathrm{et}(\mathbf{x})$.

Montrer que la moyenne pondérée minimise sur $\mathbb{R}$ la quantité :

$$\mathrm{iner}_{\mathbf{p}}(\mathbf{x},h) = \sum_{i=1}^{n} p_i (x_i - h)^2$$

Comparer la variance pondérée et la quantité $\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \left(x_i - x_j\right)^2$

1.3 Produits scalaires

Un vecteur du $\mathbb{R}$-espace vectoriel $\mathbb{R}^s$ est un s-uple de nombres réels, soit $\mathbf{x} = (x_1, x_2, ..., x_s)$.

Étant donnés $\omega_1, \omega_2, ..., \omega_s$ s nombres strictement positifs on appelle ω-produit scalaire diagonal associé à $\omega = (\omega_i)_{1 \leq i \leq s}$ l'application qui à un couple $(\mathbf{x}, \mathbf{y})$de points de $\mathbb{R}^s$ associe le nombre réel :

$$\langle \mathbf{x}|\mathbf{y} \rangle_\omega = \sum_{i=1}^{s} \omega_i x_i y_i.$$

L'application ω-produit scalaire vérifie les propriétés :

- ⋆ PS1 $\langle \mathbf{x}|\mathbf{y} \rangle_\omega = \langle \mathbf{y}|\mathbf{x} \rangle_\omega$ pour tout x et y de $\mathbb{R}^s$,
- ⋆ PS2a $\langle \mathbf{x}|\mathbf{y} + \mathbf{z} \rangle_\omega = \langle \mathbf{x}|\mathbf{y} \rangle_\omega + \langle \mathbf{x}|\mathbf{z} \rangle_\omega$ pour tout $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ de $\mathbb{R}^s$,
- ⋆ PS2b $\langle \mathbf{x}|\alpha\mathbf{y} \rangle_\omega = \alpha \langle \mathbf{x}|\mathbf{y} \rangle_\omega$ pour tout $\mathbf{x}$ et $\mathbf{y}$ de $\mathbb{R}^s$ et pour tout α de $\mathbb{R}$,
- ⋆ PS3 $\langle \mathbf{x}|\mathbf{x} \rangle_\omega \geq 0$ pour tout $\mathbf{x}$ de $\mathbb{R}^s$,
- ⋆ PS4 $\langle \mathbf{x}|\mathbf{x} \rangle_\omega = 0 \Rightarrow x = 0 = (0, 0, ..., 0)$ pour tout $\mathbf{x}$ de $\mathbb{R}^s$.

Plus généralement, étant donnée une fonction de $\mathbb{R}^s \times \mathbb{R}^s$ dans $\mathbb{R}$ qui à un couple de points $(\mathbf{x}, \mathbf{y})$ associe un nombre réel noté indifféremment

$$(\mathbf{x}|\mathbf{y})_\Phi,\ \langle \mathbf{x}|\mathbf{y} \rangle_\Phi \text{ ou } \Phi(\mathbf{x}, \mathbf{y})$$

on dit que c'est un produit scalaire si elle vérifie les propriétés PS1, ..., PS4.

Les produits scalaires diagonaux sont des produits scalaires.

Les produits scalaires sont des fonctions *BSPND* (*B*ilinaires, *S*ymétriques, *P*ositives et *N*on *D*égénérés).

Quand on manipule un seul produit scalaire, si aucune confusion n'est possible, on note simplement $(\mathbf{x}|\mathbf{y})_\Phi$=$(\mathbf{x}|\mathbf{y})$. Quand on en manipule plusieurs, on les repère par une de leur matrice.

La matrice d'un produit scalaire dans une base donnée est le tableau :

$$\mathbf{S} = \text{Mat}\left(\Phi, \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_s\}\right) = [\Phi, \{\mathbf{v}\}] = \left[\langle \mathbf{v}_i|\mathbf{v}_j \rangle_\Phi\right]_{1 \leq i \leq s\ 1 \leq j \leq s}$$

Dans cette base, la bilinéarité permet de faire les calculs avec :

$$\mathbf{x} = \sum_{i=1}^{s} \xi_i \mathbf{v}_i \Longrightarrow [\mathbf{x}, \{\mathbf{v}\}]^t = [\xi_1, \xi_2, ..., \xi_s],$$
$$\mathbf{y} = \sum_{k=1}^{s} \psi_k \mathbf{v}_k \Longrightarrow [\mathbf{y}, \{\mathbf{v}\}]^t = [\psi_1, \psi_2, ..., \psi_s],$$
$$\langle \mathbf{x}|\mathbf{y} \rangle_\Phi = \sum_{i=1}^{s}\sum_{k=1}^{s} \xi_i \psi_k \langle \mathbf{v}_i|\mathbf{v}_k \rangle = [\mathbf{x}, \{\mathbf{v}\}]^t [\Phi, \{\mathbf{v}\}] [\mathbf{y}, \{\mathbf{v}\}].$$

Donner la matrice $\mathbf{R} = [\Phi, \{\mathbf{w}\}]$ en fonction de $\mathbf{S}$ et de la matrice de changement de base $\mathbf{H} = [\text{Id}, \{\mathbf{v}\}, \{\mathbf{w}\}]$ où Id est l'identité de $\mathbb{R}^s$ dans $\mathbb{R}^s$.

Soit dans $\mathbb{R}^3$ la fonctions h qui, aux vecteurs $\mathbf{v} = (x_1, x_2, x_3)$ et $\mathbf{w} = (y_1, y_2, y_3)$ associe : $\text{h}(\mathbf{v}, \mathbf{w}) = [x_1, x_2, x_3]\, \mathbf{A} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$ A quelles conditions h est-elle un produit scalaire ?

Justifier l'appellation de produit scalaire diagonal.

1.4 Longueur, angle et distance

Un produit scalaire permet de mesurer la longueur d'un vecteur et l'angle de deux vecteurs. La Φ-norme associée à un produit scalaire est définie par :

$$\|\mathbf{x}\|_\Phi = \sqrt{\langle \mathbf{x}|\mathbf{x}\rangle_\Phi}.$$

On a $\|\alpha\mathbf{x}\|_\Phi = |\alpha| \, \|\mathbf{x}\|_\Phi$. Quand aucune confusion n'est possible, on note simplement $\|\mathbf{x}\|_\Phi = \|\mathbf{x}\|$.

Deux vecteurs de $\mathbb{R}^s$ sont Φ-orthogonaux si et seulement si $\langle \mathbf{x}|\mathbf{y}\rangle_\Phi = 0$.

Théorème de Pythagore. Montrer que :

$$\langle \mathbf{x}|\mathbf{y}\rangle_\Phi = 0 \Leftrightarrow \|\mathbf{x}+\mathbf{y}\|_\Phi^2 = \|\mathbf{x}\|_\Phi^2 + \|\mathbf{y}\|_\Phi^2 \Leftrightarrow \|\mathbf{x}+\mathbf{y}\| = \|\mathbf{x}-\mathbf{y}\|$$

Projection sur un vecteur Si $\mathbf{x}$ et $\mathbf{y}$ sont deux vecteurs de $\mathbb{R}^s$ et si $\mathbf{x}$ est non nul, il existe un unique vecteur $\mathbf{z}$ de $\mathbb{R}^s$ proportionnel à $\mathbf{x}$ tel que $\mathbf{y}-\mathbf{z}$ soit orthogonal à $\mathbf{x}$. On dit que $\mathbf{z}$ est le projeté Φ-orthogonal de $\mathbf{y}$ sur $\mathbf{x}$.

Il vaut :

$$\mathbf{z} = \frac{\langle \mathbf{x}|\mathbf{y}\rangle_\Phi}{\langle \mathbf{x}|\mathbf{x}\rangle_\Phi}\mathbf{x}.$$

Le vecteur $\mathbf{w}$ de $\mathbb{R}^s$ proportionnel à $\mathbf{x}$ qui minimise $\|\mathbf{y}-\mathbf{w}\|^2$ est $\mathbf{z}$.

Donner les coordonnées du pied de la perpendiculaire abaissée de A (1,3) sur la droite $y = x/2$.

On a toujours $|\langle \mathbf{x}|\mathbf{y}\rangle_\Phi| \leq \|\mathbf{x}\|_\Phi \, \|\mathbf{y}\|_\Phi$ (*Cauchy-Schwartz*) Si $\mathbf{x}$ et $\mathbf{y}$ sont 2 points de $\mathbb{R}^s$, la Φ-mesure de l'angle de $\mathbf{x}$ et $\mathbf{y}$ est notée

$$\mathrm{A}_\Phi(\mathbf{x},\mathbf{y}) = a$$

Elle est définie par :

$$0 \leq a \leq \pi \text{ et } \cos a = \frac{\langle \mathbf{x}|\mathbf{y}\rangle_\Phi}{\|\mathbf{x}\|_\Phi \|\mathbf{y}\|_\Phi}.$$

On a toujours $\|\mathbf{x}+\mathbf{y}\|_\Phi \leq \|\mathbf{x}\|_\Phi + \|\mathbf{y}\|_\Phi$ (Inégalité triangulaire).

La Φ-distance de deux vecteurs est définie par :

$$\mathrm{d}_\Phi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_\Phi = \|\mathbf{y}-\mathbf{x}\|_\Phi.$$

On sait donc mesurer les angles et les distances entre points de $\mathbb{R}^s$ au sens d'un produit scalaire donné.

1.5 Définitions euclidiennes

Dans $\mathbb{R}^s$, on notera $\{\mathbf{e}\}_s = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_s\}$ la base canonique. On utilisera :

- ⋆ le produit scalaire canonique $\langle \mathbf{x}|\mathbf{y}\rangle_c = \sum_{i=1}^{s} x_i y_i = (\mathbf{x}|\mathbf{y})_{\mathbf{I}_s}$ où $\mathbf{I}_s$ est la matrice identité ;
- ⋆ le produit scalaire uniforme $\langle \mathbf{x}|\mathbf{y}\rangle_u = \frac{1}{s}\sum_{i=1}^{s} x_i y_i = (\mathbf{x}|\mathbf{y})_{\mathbf{U}_s}$ avec $\mathbf{U}_s = \frac{1}{s}\mathbf{I}_s$;
- ⋆ le produit scalaire associé à une pondération $\mathbf{p} = (p_1, p_2, ..., p_s)$

$$\langle \mathbf{x}|\mathbf{y}\rangle_{\mathbf{p}} = \sum_{i=1}^{s} p_i x_i y_i = (\mathbf{x}|\mathbf{y})_{\mathbf{D}_s} \text{ avec } \mathbf{D}_s = \mathrm{Diag}(p_1, p_2, ..., p_s)$$

Dans la plupart des cas, dès qu'il s'agit de variables, on se contentera de noter $(\mathbf{x}|\mathbf{y})_{\mathbf{D}}$ un produit scalaire associé à une pondération, le contexte rendant implicite la dimension et la nature de cette pondération qui, par défaut, est uniforme.

La moyenne naturelle devient $m = \mathrm{m}(\mathbf{x}) = \overline{\mathbf{x}} = \langle \mathbf{x}|\mathbf{1}_n \rangle_u$. La moyenne pondérée de x est : $m_{\mathbf{p}} = \mathrm{m}_{\mathbf{p}}(\mathbf{x}) = \overline{\mathbf{x}}_{\mathbf{p}} = \langle \mathbf{x}|\mathbf{1}_n \rangle_{\mathbf{p}}$.

Dans la plupart des cas, dès qu'il s'agit de variables, on se contentera de parler de moyenne et de noter $\mathrm{m}(\mathbf{x})$ la moyenne associée à une pondération, le contexte rendant implicite la dimension et la nature de cette pondération qui, par défaut, est uniforme. Le calcul de la moyenne est donc associé, dans tous les cas, à une projection euclidienne (figure 1).

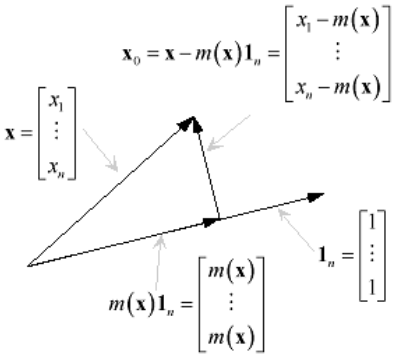


FIG. 1 – Le centrage est une projection euclidienne.

On utilisera toujours la notation $\mathrm{Proj}_A^{\mathbf{\Phi}}$ pour désigner la projection orthogonale au sens du **Φ-**produit scalaire sur un sous-espace A.

La variance est alors simplement : $v = \mathrm{v}(x) = \|\mathbf{x} - \overline{\mathbf{x}}\mathbf{1}_n\|^2 = \|\mathbf{x}_0\|^2$

Du théorème de Pythagore il vient :

$$\mathrm{v}(x) = \|\mathbf{x} - \overline{\mathbf{x}}.\mathbf{1}_n\|^2 = \|\mathbf{x}\|^2 - \overline{\mathbf{x}}^2$$

La variance est la moyenne des carrés moins le carré de la moyenne. On a là un exemple des plus cocasses du lien entre mathématique et pratique. La formule est célèbre. Elle servait, au temps des calculs manuels, à calculer la variance à l'aide d'une table de carrés. Elle a souvent été donnée pour une définition. Elle est devenue un ennui majeur dans le calcul numérique,à cause des erreurs d'arrondis. Elle reste la base théorique du système des carrés des écarts.

On note $\mathbf{x}_0 = \mathbf{x} - \overline{\mathbf{x}}\mathbf{1}_n$ et on l'appelle la variable centrée associée à $\mathbf{x}$. La décomposition s'écrit :

$$\mathbf{x} = \mathbf{x}_0 + \mathrm{m}(\mathbf{x})\mathbf{1}_n \Leftrightarrow \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 - \mathrm{m}(\mathbf{x}) \\ \vdots \\ x_i - \mathrm{m}(\mathbf{x}) \\ \vdots \\ x_n - \mathrm{m}(\mathbf{x}) \end{bmatrix} + \begin{bmatrix} \mathrm{m}(\mathbf{x}) \\ \vdots \\ \mathrm{m}(\mathbf{x}) \\ \vdots \\ \mathrm{m}(\mathbf{x}) \end{bmatrix}$$

C'est la plus simple des décompositions aux moindres carrés qui se retrouvera à tout moment.

2 Covariance et corrélation

Deux variables forment un tableau accompagné de la pondération commune des individus :

$$\mathbf{X} = \left[\begin{array}{cc} \mathbf{x} & \mathbf{y} \end{array} \right] = \left[\begin{array}{cc} x_1 & y_1 \\ \vdots & \vdots \\ x_i & y_i \\ \vdots & \vdots \\ x_n & y_n \end{array} \right] \left[\begin{array}{c} p_1 \\ \vdots \\ p_i \\ \vdots \\ p_n \end{array} \right]$$

2.1 Deux figures duales

Le tableau centré contient les variables centrées :

$$\mathbf{X}_0 = \left[\begin{array}{cc} \mathbf{x}_0 & \mathbf{y}_0 \end{array} \right] = \left[\begin{array}{cc} x_1 - \mathrm{m}(\mathbf{x}) & y_1 - \mathrm{m}(\mathbf{y}) \\ \vdots & \vdots \\ x_i - \mathrm{m}(\mathbf{x}) & y_i - \mathrm{m}(\mathbf{y}) \\ \vdots & \vdots \\ x_n - \mathrm{m}(\mathbf{x}) & y_n - \mathrm{m}(\mathbf{y}) \end{array} \right]$$

On a soit n points de $\mathbb{R}^2$, soit 2 points de $\mathbb{R}^n$. Dans le premier cas, le centrage est un changement de repère ; deux variables y sont vues comme n points de $\mathbb{R}^2$ (figure 2, à gauche). Dans le second, c'est une double projection ; deux variables y sont vues comme 2 points de $\mathbb{R}^n$ (figure 2, à droite).

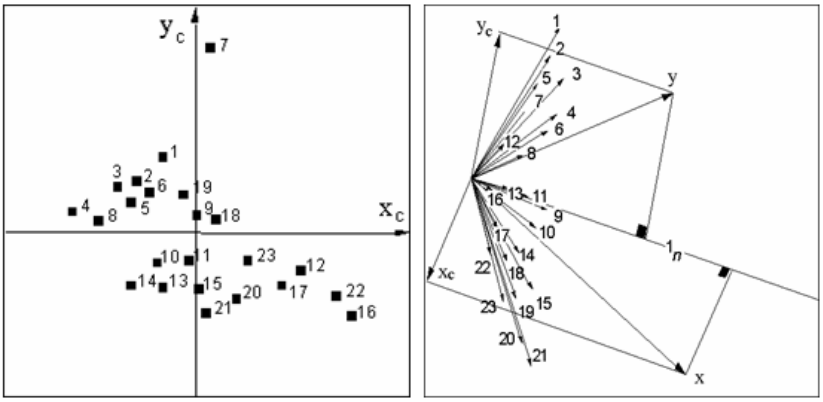


FIG. 2 – Les deux points de vue de la statistique linéaire. A gauche, celui des données permet de voir les modèles. A droite, celui de la théorie permet de comprendre les calculs. Celui de droite s'étend en dimension quelconque.

2.2 Les droites de régressions

La régression est la recherche d'un prédicteur linéaire.

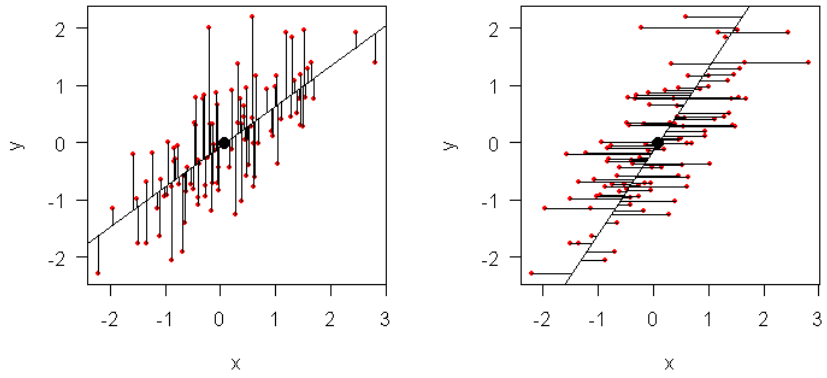


FIG. 3 – Définition des droites de régression.A gauche, la droite de prédiction de $\mathbf{y}$ par $\mathbf{x}$, à gauche, droite de la prédiction de $\mathbf{x}$ par $\mathbf{y}$. Dans les deux cas la somme des carrés des écarts données-modèles est minimum.

Il y a deux problèmes. On cherche une droite $y = ax + b$ qui minimise :

$$E(a,b) = \sum_{i=1}^{n} p_i M_i P_i^2 = \sum_{i=1}^{n} p_i (y_i - ax_i - b)^2$$

La solution analytique (utiliser les dérivées partielles) est :

$$b = \mathrm{m}(\mathbf{y}) - a\mathrm{m}(\mathbf{x})$$

$$a = \frac{\sum_{i=1}^{n} p_i (x_i - \mathrm{m}(\mathbf{x}))(y_i - \mathrm{m}(\mathbf{y}))}{\sum_{i=1}^{n} p_i (x_i - \mathrm{m}(\mathbf{x}))^2}$$

La solution algébrique utilise une droite qui passe par le centre de gravité. Chercher une droite $Y = aX$ qui minimise :

$$E(a) = \sum_{i=1}^{n} p_i M_i P_i^2 = \sum_{i=1}^{n} p_i (Y_i - aX_i) = \|\mathbf{y}_0 - a\mathbf{x}_0\|_{\mathbf{p}}^2$$

On trouve :

$$a = \frac{(\mathbf{x}_0|\mathbf{y}_0)_{\mathbf{D}}}{\|\mathbf{x}_0\|_{\mathbf{D}}^2} = \frac{\sum_{i=1}^{n} p_i (x_i - \mathrm{m}(\mathbf{x}))(y_i - \mathrm{m}(\mathbf{y}))}{\sum_{i=1}^{n} p_i (x_i - \mathrm{m}(\mathbf{x}))^2} = \frac{\mathrm{cov}_{\mathbf{p}}(\mathbf{x},\mathbf{y})}{\mathrm{var}_{\mathbf{p}}(\mathbf{x})}$$

On appelle covariance de deux variables le produit scalaire des variables centrées. La covariance naturelle est :

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

.

La corrélation est :

$$\mathrm{cor}_{\mathbf{p}}(x,y) = \frac{\mathrm{cov}_{\mathbf{p}}(\mathbf{x},\mathbf{y})}{\sqrt{\mathrm{var}_{\mathbf{p}}(\mathbf{x})}\sqrt{\mathrm{var}_{\mathbf{p}}(\mathbf{y})}} = \frac{\mathrm{cov}_{\mathbf{p}}(\mathbf{x},\mathbf{y})}{\mathrm{et}(\mathbf{x})\,\mathrm{et}(\mathbf{y})} = \frac{(\mathbf{x}_0|\mathbf{y}_0)_{\mathbf{D}}}{\|\mathbf{x}_0\|_{\mathbf{D}}\,\|\mathbf{y}_0\|_{\mathbf{D}}} = \cos(A(x_0,y_0))$$

Fondamentalement, la variance de la variable prédite se décompose en deux composantes additives (théorème de Pythagore), respectivement l'erreur de prédiction et la variance expliquée. Le carré de corrélation est dit pourcentage de variance expliquée.

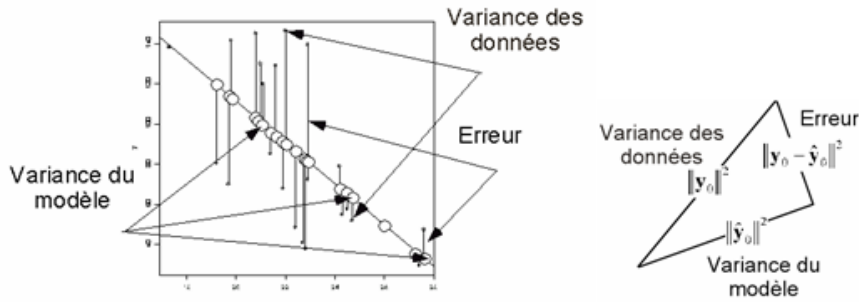


FIG. 4 – Décomposition de la variance dans une régression linéaire simple

2.3 Exercice : Les jurys

On considère un jury de sélection formé de L juges sélectionnant chacun R produits parmi C. Le résultat de la sélection est consigné dans un tableau $\mathbf{X} = [x_{ij}]$ où x_{ij} vaut 1 si le juge i sélectionne le produit j et 0 sinon.

a) Soit a_{ik} le nombre des sélections communes aux juges i et k. Donner le coefficient de corrélation linéaire r_{ik} entre les juges i et k, en fonction de a_{ik}, R et C.

b) On peut définir sur l'ensemble des C produits une statistique S en posant $S_j = \sum_{i=1}^{L} x_{ij}$. Quelle signification donner à S_j ? Calculer la moyenne des valeurs de S_j en fonction de L, C et R. Exprimer la variance v des valeurs S_j en fonction de L, C, R et la somme pour $i \neq k$ des a_{ik}.

c) Exprimer en fonction de L, C, R et v la moyenne Z des coefficients de corrélation linéaire r_{ik} sur l'ensemble des $L(L-1)/2$ paires de juges.

d) Vérifier le résultat précédent dans le cas où les L jugements concordent.

On considère un jury de classement de L juges qui ont à classer C produits pour un critère donné. Chacun des juges classe les produits par ordre de préférence (sans ex æquo) et attribue à chacun un entier compris entre 1 et C : b_{ij} est le rang proposé par le juge i pour le produit j. Le résultat de la dégustation est consigné dans un tableau de L lignes et C colonnes où chaque ligne est une permutation des entiers $1, ..., C$. On note m_i et v_i la moyenne et la variance des rangs attribués par le juge i. De plus S_j est la somme des rangs associés au produit j et on pose $t_{ik} = \sum_{j=1}^{C} b_{ij} b_{kj}$.

a) Calculer m_i et v_i.

b) Calculer le coefficient de corrélation linéaire r_{ik} entre les juges i et k (coefficient de Spearman) en fonction de C et t_{ik}

c) Calculer la moyenne m et la variance v de la série des S_j en fonction de L, C et la somme pour$i \neq k$ des t_{ik}.

d) Exprimer en fonction de L, C, v la moyenne z des coefficients de Spearman pour tous les couples (i, k) de juges.

e) Vérifier le résultat obtenu lorsque tous les jugements concordent.

f) Application : au premier tour d'un concours (Mâcon 1981) 3 juges ont classé 7 vins de la manière suivante :

	1	2	3	4	5	6	7
1	5	2	3	4	1	6	7
2	7	1	3	6	2	5	4
3	2	1	4	7	3	6	5

Calculer le coefficient de corrélation moyen entre les juges. Sachant que seuls 3 vins sont retenus pour le tour suivant on peut réduire les données aux choix effectué par chacun des juges soit :

	1	2	3	4	5	6	7
1	0	1	1	0	1	0	0
2	0	1	1	0	1	0	0
3	1	1	0	0	1	0	0

Calculer le nouveau coefficient de corrélation moyen. Commenter alors les trois méthodes de sélection pour le tour suivant :

- ⋆ Seront sélectionnés les produits ayant obtenu la meilleure somme des rangs ;
- ⋆ Seront sélectionnés les produits ayant obtenu le plus grand nombre de sélections par un juge ;
- ⋆ Sera d'abord sélectionné le produit ayant obtenu le plus grand nombre de places de premier, puis en cas dex æquo le plus grand nombre de places de second, puis en cas dex æquo le plus grand nombre de places de troisième, puis...

On peut s'intéresser alors aux coefficients de concordance.

g) Pour quelle situation la variance v sera-t-elle maximale ? Quelle est alors la variance v_{max} ?

h) On pose $K = v/v_{max}$. K est appelé coefficient de concordance de Kendall. Exprimer z en fonction de K et K en fonction de z. Justifier le nom donné à K.

i) On pose $u_{ik} = \sum_{j=1}^{C} (b_{ij} - b_{kj})^2$. Quelle relation existe-t-il entre r_{ik} et u_{ik} ? j) Application : 9 juges classent 6 vins de consommation courante pour la finesse de l'arôme :

	A	B	C	D	E	F
1	6	5	4	2	1	3
2	6	5	4	3	1	2
3	1	4	5	6	2	3
4	4	3	2	1	5	6
5	4	5	3	2	1	6
6	1	5	6	3	4	2
7	6	5	2	3	1	4
8	5	4	3	2	1	6
9	5	3	1	2	4	6

et l'harmonie générale (à droite) :

	A	B	C	D	E	F
1	6	3	2	1	4	5
2	5	6	4	2	1	3
3	3	6	1	2	5	4
4	5	3	1	2	4	6
5	6	5	4	2	1	3
6	4	2	1	5	3	6
7	6	5	2	3	1	4
8	5	4	3	2	1	6
9	5	2	1	3	4	6

Pour chacun des deux tableaux, calculer v, z, K. Pour quel critère le jury est-il le plus homogène ? Commentaires.

Pour en savoir plus : Tomassone R. & Flanzy C. (1977) Présentation synthétique de diverses méthodes d'analyse de données fournies par un jury de dégustateurs. *Annales de Technologie Agricole*, 26, 373-418.

3 Variables qualitatives

3.1 Projections et systèmes orthogonaux

Si $\{\mathbf{v}_1, ..., \mathbf{v}_s\}$ est une base de $\mathcal{E} = \mathbb{R}^s$, on dit qu'elle est Φ-orthogonale si $i \neq j \Rightarrow \langle \mathbf{v}_i | \mathbf{v}_j \rangle_\Phi = 0$. On dit qu'elle est Φ-normale si $1 \leq i \leq s \Rightarrow \|v_i\|_\Phi = 1$.

Tout système $\{\mathbf{v}_1, ..., \mathbf{v}_r\}$ de r vecteurs non nuls et orthogonaux deux à deux est une base orthogonale du sous-espace qu'ils engendrent.

Soit $\{\mathbf{v}_1, ..., \mathbf{v}_r\}$ un ensemble de r vecteurs non nuls orthogonaux deux à deux de $\mathbb{R}^s$, $\mathcal{F} = \text{sev}(\mathbf{v}_1, ..., \mathbf{v}_r)$ le sous espace qu'ils engendrent et $\mathbf{w}$ un vecteur de $\mathcal{E}$. Soient les r nombres réels définis par :

$$c_k = \frac{\langle \mathbf{w} | \mathbf{v}_k \rangle}{\langle \mathbf{v}_k | \mathbf{v}_k \rangle}$$

Pour tout ensemble de r nombres réels $a_1, ..., a_r$ on a :

$$\left\| \mathbf{w} - \sum_{k=1}^{r} c_k \mathbf{v}_k \right\| \leq \left\| \mathbf{w} - \sum_{k=1}^{r} a_k \mathbf{v}_k \right\|$$

Le vecteur défini par $\mathbf{w}_0 = \sum_{k=1}^{r} c_k \mathbf{v}_k$ est l'unique vecteur de $\mathcal{E}$ tel que :

$$1 \leq k \leq s \Rightarrow \langle \mathbf{w} - \mathbf{w_0} | \mathbf{v}_k \rangle = 0$$

On l'appelle projeté orthogonal de $\mathbf{w}$ sur $\mathcal{F}$ et on le note $\text{Proj}_{\mathcal{F}}^{\Phi}(\mathbf{w})$. L'intérêt de la proposition devient manifeste dés qu'on sait trouver une base orthogonale dans un sous-espace donné.

C'est l'objet du théorème de Gram-Schmidt : $\mathcal{E} = \mathbb{R}^n$. $\mathcal{F}$ est un sous-espace vectoriel de $\mathcal{E}$. $\{\mathbf{v}_1, ..., \mathbf{v}_r\}$ est une base de $\mathcal{F}$ Φ-orthogonale. Si $\mathcal{F} \subset \mathcal{E}$, il existe $\{\mathbf{w}_{r+1}, ..., \mathbf{w}_n\}$ dans $\mathcal{E}$ tels que

$$\{\mathbf{v}_1, ..., \mathbf{v}_r, \mathbf{w}_{r+1}, ..., \mathbf{w}_n\}$$

est une base orthogonale de $\mathcal{E}$. En bref, on peut compléter une base orthogonale en obtenant une base orthogonale.

On en déduit que, étant donnée une base $\{\mathbf{v}_1, ..., \mathbf{v}_r\}$, on peut trouver une base orthogonale $\{\mathbf{w}_1, ..., \mathbf{w}_r\}$ telle que

$$1 \leq k \leq r \Rightarrow \text{sev}(\mathbf{v}_1, ..., \mathbf{v}_k) = \text{sev}(\mathbf{w}_1, ..., \mathbf{w}_k).$$

.

Le principe est simple et induit une procédure dite de Gram-Schmidt d'orthogonalisation d'une base :

$$\begin{aligned}
&\text{pas 1} : \mathbf{w}_1 = \mathbf{v}_1 \\
&\text{pas 2} : \mathbf{w}_2 = \mathbf{v}_2 - \frac{\langle \mathbf{v}_2 | \mathbf{w}_1 \rangle}{\langle \mathbf{w}_1 | \mathbf{w}_1 \rangle} \mathbf{w}_1 \\
&\text{pas 3} : \mathbf{w}_3 = \mathbf{v}_3 - \frac{\langle \mathbf{v}_3 | \mathbf{w}_1 \rangle}{\langle \mathbf{w}_1 | \mathbf{w}_1 \rangle} \mathbf{w}_1 - \frac{\langle \mathbf{v}_3 | \mathbf{w}_2 \rangle}{\langle \mathbf{w}_2 | \mathbf{w}_2 \rangle} \mathbf{w}_2 \\
&... \\
&\text{pas r} : \mathbf{w}_r = \mathbf{v}_r - \frac{\langle \mathbf{v}_r | \mathbf{w}_1 \rangle}{\langle \mathbf{w}_1 | \mathbf{w}_1 \rangle} \mathbf{w}_1 - \frac{\langle \mathbf{v}_r | \mathbf{w}_2 \rangle}{\langle \mathbf{w}_2 | \mathbf{w}_2 \rangle} \mathbf{w}_2 - ... - \frac{\langle \mathbf{v}_r | \mathbf{w}_{r-1} \rangle}{\langle \mathbf{w}_r | \mathbf{w}_{r-1} \rangle} \mathbf{w}_{r-1}
\end{aligned}$$

On appelle complémentaire orthogonal d'un sous espace l'ensemble des vecteurs orthogonaux à tout vecteur de ce sous-espace et on note $\mathcal{G} = \mathcal{F}^{\perp_\Phi} = \mathcal{F}^\perp$. Un sous-espace et son orthogonal forment une somme directe orthogonale :

$$\mathcal{F} + \mathcal{G} = \mathcal{F} \oplus \mathcal{G} = \mathcal{F} \overset{\perp}{\oplus} \mathcal{G}$$

Le projecteur sur une somme directe orthogonale est simplement

$$\text{Proj}^\Phi_{\mathcal{F} \overset{\perp}{\oplus} \mathcal{G}} = \text{Proj}^\Phi_{\mathcal{F}} + \text{Proj}^\Phi_{\mathcal{G}}$$

Noter encore le théorème des trois perpendiculaires : Si $\mathcal{F}$ est un sous-espace de $\mathcal{E}$ et $\mathcal{G}$ un sous-espaces de $\mathcal{F}$, alors :

$$\text{Proj}^\Phi_{\mathcal{G}} \circ \text{Proj}^\Phi_{\mathcal{F}} = \text{Proj}^\Phi_{\mathcal{G}}.$$

Il convient de retenir ce qui précède que si $\mathcal{E} = \mathbb{R}^n$ est muni d'un produit scalaire et $\mathcal{F}$ est une sous-espace vectoriel de $\mathcal{E}$, alors :

a) $\mathcal{F}$ possède une base $\mathbf{v}_1$, ..., $\mathbf{v}_r$ et r est la dimension de $\mathcal{F}$

b) $\mathcal{F}$ possède une base orthogonale $\mathbf{w}_1$, ..., $\mathbf{w}_r$ de r vecteurs qu'on peut trouver avec la procédure de Gram-Schmidt

c) Tout vecteur $\mathbf{x}$ de $\mathcal{E}$ se décompose de manière unique sous la forme

$$\mathbf{x} = \mathbf{z} + (\mathbf{x} - \mathbf{z}) \text{ où } z \in \mathcal{F} \text{ et } (\mathbf{x} - \mathbf{z}) \in \mathcal{F}^{\perp}$$

d) on peut calculer le vecteur $\mathbf{z}$ par :

$$\mathbf{z} = \sum_{k=1}^{r} \frac{\langle \mathbf{x} | \mathbf{w}_k \rangle}{\langle \mathbf{w}_k | \mathbf{w}_k \rangle} \mathbf{w}_k$$

e) le vecteur $\mathbf{z}$ est dit projeté orthogonal de $\mathbf{x}$. Il minimise dans $\mathcal{F}$ la quantité $\|\mathbf{x} - \mathbf{w}\|$. Cette quantité est appelée distance de $\mathbf{x}$ au sous-espace vectoriel $\mathcal{F}$ (minimum de la distance à un quelconque vecteur de $\mathcal{F}$).

f) L'angle de $\mathbf{x}$ et $\mathbf{z}$ défini par $a \in \left[0, \frac{\pi}{2}\right]$ et $\cos a = \frac{\langle \mathbf{x} | \mathbf{z} \rangle}{\|\mathbf{x}\| \|\mathbf{z}\|}$ est appelé angle du vecteur $\mathbf{x}$ avec le sous-espace $\mathcal{F}$. On notera que :

$$\langle \mathbf{x} | \mathbf{z} \rangle = \langle \mathbf{x} - \mathbf{z} + \mathbf{z} | \mathbf{z} \rangle = \langle \mathbf{x} - \mathbf{z} | \mathbf{z} \rangle + \|\mathbf{z}\|^2 \Rightarrow \cos a = \frac{\|\mathbf{z}\|}{\|\mathbf{x}\|}$$

On applique ces notions au traitement élémentaire des variables qualitatives.

3.2 Sous-espace des indicatrices

Une variable qualitative prend ses valeurs dans un ensemble fini de valeurs appelées modalités de la variable.

$$\begin{array}{c|c|c|c|c|ccccccc}
1 & \texttt{bleu} & 1 & 2 & 1 & 0 & 1 & \cdots & 0 & \cdots & 0 \\
2 & \texttt{vert} & 2 & 1 & 2 & 1 & 0 & \cdots & 0 & \cdots & 0 \\
3 & \texttt{rouge} & 3 & m & 3 & 0 & 0 & \cdots & 0 & \cdots & 1 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
i & \texttt{bleu} & i & 2 & i & 0 & 1 & \cdots & 0 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
n & \texttt{blanc} & n & k & n & 0 & 0 & \cdots & 1 & \cdots & 0
\end{array}$$

Les modalités sont repérées par une étiquette ou un numéro d'étiquette. m est le nombre de modalités. Les porteurs de l'étiquette k forment la classe k. La variable qui prend la valeur 1 pour les élments de la classe k et 0 pour les autres est appelée l'indicatrice de la classe k et notée $\mathbf{I}_k$. Le sous-espace vectoriel engendré par les vecteurs $\{\mathbf{I}_1, \cdots, \mathbf{I}_k, \cdots, \mathbf{I}_m\}$ est l'ensemble des $\mathbf{x}$ du type $\mathbf{x} = \sum_{k=1}^{m} \alpha_k \mathbf{I}_k$, soit l'ensemble des variables qui sont *constantes par classe*. Les variables qualitatives définissent donc un sous-espace vectoriel, alors que les variables quantitatives définissent un vecteur.

On peut supposer qu'on utilise une pondération $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ des individus. Les indicatrices des classes sont alors orthogonales pour le produit scalaire associé à :

$$\mathbf{D}_n = Diag\,(p_1, p_2, \cdots, p_n).$$

On note n_k le nombre de porteur de la modalité k et $n = \sum_{k=1}^{m} n_k$. Le poids de la classe k est alors défini par la somme des poids des individus de la classe k et noté

$$p_k^+ = \sum_{i/\mathbf{I}_k(i)=1} p_i.$$

Quand on veut étudier le lien entre une variable qualitative et une variable quantitative, on note q_i le numéro de la modalité associée à l'individu i :

$$q_i = k \Leftrightarrow \mathbf{I}_k\,(i) = 1.\ \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} q_1 \\ \vdots \\ q_i \\ \vdots \\ q_n \end{bmatrix} \quad \begin{bmatrix} p_1 \\ \vdots \\ p_i \\ \vdots \\ p_n \end{bmatrix}$$

On transforme traditionnellement l'information en un tableau d'analyse de variance, c'est à dire qu'on réécrit les valeurs prises par $\mathbf{x}$ par classe de valeurs prises par $\mathbf{q}$:

$$\begin{array}{ccccc}
Classe\ 1 & \cdots & Classe\ k & \cdots & Classe\ m \\
x_{11} & & x_{k1} & & x_{m1} \\
x_{12} & & x_{k2} & & x_{m2} \\
\vdots & & \vdots & & \vdots \\
x_{1n_1} & & x_{kn_k} & & x_{mn_m}
\end{array}$$

Le nombre de valeurs varie en général d'une classe à l'autre. A chaque classe on attribue un poids et une moyenne conditionnelle de x sachant k :

$$\mathrm{mc}_{/k}(\mathbf{x}) = \sum_{i/q_i=k} \frac{p_i}{p_k^+} x_i$$

et une variance conditionnelle de x sachant k :

$$\mathrm{vc}_{/k}(\mathbf{x}) = \sum_{i/q_i=k} \frac{p_i}{p_k^+} \left(x_i - \mathrm{mc}_{/k}(\mathbf{x})\right)^2$$

La moyenne de $\mathbf{x}$ est la moyenne des moyennes conditionnelles en utilisant le poids des classes, ce qui conduit à envisager la variance des moyennes conditionnelles et la moyenne des variances conditionnelles. La première vaut (b est utilisé pour *between*) :

$$\mathrm{b}(\mathbf{x}) = \sum_{k=1}^{m} p_k^+ \left(\mathrm{mc}_{/k}(\mathbf{x}) - \mathrm{m}_{\mathbf{p}}(\mathbf{x})\right)^2$$

Cette quantité s'appelle la variance inter-classe. La seconde vaut (w est utilisé pour *within*) :

$$\mathrm{w}(\mathbf{x}) = \sum_{k=1}^{m} p_k^+ \mathrm{vc}_{/k}(\mathbf{x})$$

Elle s'appelle la variance intra-classe. On démontre alors l'équation d'analyse de la variance (totale = inter + intra) :

$$\mathrm{var}_{\mathbf{p}}(\mathbf{x}) = \mathrm{b}(\mathbf{x}) + \mathrm{w}(\mathbf{x})$$

Le pourcentage de variance formée par la variance inter-classe est appelé rapport de corrélation des variables $\mathbf{x}$ et $\mathbf{q}$ et est noté traditionnellement :

$$\eta^2_{\mathbf{xq}} = \frac{b(\mathbf{x})}{\mathrm{var}_{\mathbf{p}}(\mathbf{x})} = 1 - \frac{w(\mathbf{x})}{\mathrm{var}_{\mathbf{p}}(\mathbf{x})}$$

Ces notions sont plus simples à exprimer en terme algébrique, bien que leur utilité soit plus claire dans la présentation élémentaire. La variable $\mathbf{x}$ est un vecteur de $\mathcal{E} = \mathbb{R}^n$. La variable centrée $\mathbf{x}_0$ a pour carré de norme la variance $\mathrm{var}_{\mathbf{p}}(\mathbf{x})$. Les indicatrices des classes de la variable $\mathbf{q}$ forme un sous-espace $\mathcal{F}$ de dimension m. Les indicatrices de classes étant orthogonales, on cherche leur norme :

$$\|\mathbf{I}_k\|^2 = p_k^+$$

On projette alors $\mathbf{x}_0$ sur $\mathcal{F}$:

$$\begin{gathered}
\langle \mathbf{I}_j, \mathbf{I}_k \rangle = 0 \\
c_k = \frac{\langle \mathbf{x}, \mathbf{I}_k \rangle}{\langle \mathbf{I}_k, \mathbf{I}_k \rangle} = \mathrm{mc}_{/k}(\mathbf{x}) \\
\|\mathrm{P}_{\mathcal{F}}(\mathbf{x}_0)\|^2 = \sum_{i=1}^{n} p_i \left(\mathrm{mc}(\mathbf{x})_{/u_i} - \mathrm{m}(\mathbf{x})\right)^2 = \\
\sum_{k=1}^{m} \sum_{i/q(i)=k} p_i \left(\mathrm{mc}_{/k}(\mathbf{x}) - \mathrm{m}(\mathbf{x})\right)^2 = \mathrm{b}(\mathbf{x})
\end{gathered}$$

L'équation d'analyse de la variance est donc encore une conséquence du théorème de Pythagore.

3.3 Exercices

a) Une variable **q** qualitative prend ses valeurs dans {A,B,C,D} :

i	1	2	3	4	5	6	7	8	9	10	11
q_i	A	A	A	B	B	B	C	C	C	D	D
x_i	1	1	1	2	2	2	3	3	3	4	4

La variable **x** a pour valeurs les numéros de modalités de la variable **q**. Une variable **y** prend les valeurs :

i	1	2	3	4	5	6	7	8	9	10	11
y_i	1	2	3	1	3	5	3	5	7	6	8

Comparer le carré du coefficient de corrélation de **x** et **q** et le rapport de corrélation de **y** et **q**. Généraliser le résultat.

b) Une variable quantitative devient qualitative par la mise en classe. Une variable qualitative devient quantitative par le codage numérique des modalités. Les deux opérations sont cohérentes si on affecte à chaque classe son centre. Par exemple, on a relevé les notes de 50 étudiants. **x** est la note de mathématiques et **y** celle d'informatique. On obtient le tableau de contingence : Donner les

	y	*[5,7[*	*[7,9[*	*[9,11[*	*[11,13[*	*[13,15[*
x		*6*	*8*	*10*	*12*	*14*
[4,6[	*5*	0	5	0	0	1
[6,8[	*7*	2	2	2	1	2
[8,10[	*9*	1	4	6	0	2
[10,12[	*11*	3	1	3	5	3
[12,14[	*13*	0	0	0	4	4

moyennes et les variances marginales.

Calculer les effectifs, les moyennes et les variances conditionnelles.

Dessiner le nuage et les courbes de régression. Calculer les rapports de corrélation.

Déterminer les droites de régression. Calculer le coefficient de corrélation.

Commenter les décompositions de la variance.

c) **x** et **y** sont deux variables qui prennent les valeurs 0 ou 1. On peut les considérer soit comme variables quantitatives, soit comme variables qualitatives (non/oui). Pour n individus on recueille a observations du type (0,0), b observations du type (0,1), c observations du type (1,0) et d observations du type (1,1). La statistique est alors donnée par la table de contingence :

	$y=0$	$y=1$	total
$x=0$	a	b	$a+b$
$x=1$	c	d	$c+d$
total	$a+c$	$b+d$	n

Donner en fonction de a, b, c, d et n les moyennes et variances marginales, les moyennes et variances conditionnelles, la covariance, le carré du coefficient de corrélation linéaire et les rapports de corrélation.

4 Régression multiple

4.1 Position du problème

On considère p variables dites explicatives (ou prédictives) $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^p$ et une variable à prédire $\mathbf{y}$. Les variables prédictives forment un tableau $\mathbf{X}$ à n lignes et p colonnes de terme général x_{ij}. La pondération des individus (lignes) est consignée dans la matrice diagonale $\mathbf{D}$.

$$\mathbf{D} = \text{Diag}\,(p_1, \cdots, p_n)$$

Les moyennes des variables explicatives forment un vecteur :

$$\mathbf{m} = \mathbf{X}^t\mathbf{D}\mathbf{1}_n = \begin{bmatrix} m_1 \\ \vdots \\ m_p \end{bmatrix}$$

Les variables explicatives centrées forment le tableau $\mathbf{X}_0$ à n lignes et p colonnes de terme général $x_{ij} - m_j$. Les variances et covariances entre explicatives sont directement calculées dans la matrice de variances-covariances :

$$\mathbf{C} = [c_{jk}]_{1\leq j\leq p,\ 1\leq k\leq p} = \mathbf{X}_0^t\mathbf{D}\mathbf{X}_0.$$

Les écarts-types des variables explicatives sont donc les quantités $s_j = \sqrt{c_{jj}}$. On note $\mathbf{S}$ la matrice diagonale $\mathbf{S} = \text{Diag}\,(s_1, \cdots, s_p)$.

Les variables explicatives normalisées forment le tableau $\mathbf{X}_*$ à n lignes et p colonnes de terme général $x^*_{ij} = \frac{x_{ij}-m_j}{s_j}$. Algébriquement :

$$\mathbf{X}_* = \mathbf{X}_0\mathbf{S}^{-\frac{1}{2}}.$$

Les corrélations entre explicatives sont directement calculées dans la matrice de corrélation :

$$\mathbf{R} = [r_{jk}]_{1\leq j\leq p,\ 1\leq k\leq p} = \mathbf{X}_*^t\mathbf{D}\mathbf{X}_* = \mathbf{S}^{-\frac{1}{2}}\mathbf{C}\mathbf{S}^{-\frac{1}{2}}.$$

La j^{ème} colonne de $\mathbf{X}$ est $\mathbf{X}^j$, la j^{ème} colonne de $\mathbf{X_0}$ est $\mathbf{X}_0^j$, la j^{ème} colonne de $\mathbf{X}_*$ est $\mathbf{X}_*^j$.

Rappelons que $\left\|\mathbf{X}_0^j\right\|^2 = c_{jj} = \text{v}\left(\mathbf{X}^j\right)$ et $\left\|\mathbf{X}_*^j\right\|^2 = 1 = \text{v}\left(\mathbf{X}_*^j\right)$. Les variables normalisées sont des vecteurs $\mathbf{D}$-unitaires.

La variable à expliquée (ou à prédire ou dépendante) a pour moyenne $\text{m}\,(\mathbf{y})$, pour variance $\text{v}\,(\mathbf{y})$. La variable dépendante centrée est $\mathbf{y}_0 = \mathbf{y} - \text{m}\,(\mathbf{y})\,\mathbf{1}_n$. La variable dépendante normalisée est $\mathbf{y}_* = \frac{\mathbf{y}-\text{m}(\mathbf{y})\mathbf{1}_n}{\sqrt{\text{v}(\mathbf{y})}}$.

Faire la régression de $\mathbf{y}$ sur les variables $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^p$, c'est chercher à prédire l'observation y_i à l'aide d'un modèle du type

$$\widehat{y_i} = \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots \alpha_p x_{ip} + \beta = y_i + e_i$$

$\widehat{y_i}$ est la prédiction de y_i et e_i est l'erreur de prédiction. On cherchera à minimiser l'erreur totale, soit la quantité :

$$\text{E}(\alpha_1, \cdots, \alpha_p, \beta) = \sum_{i=1}^n p_i\,(\widehat{y_i} - y_i)^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2_{\mathbf{D}}$$

Ce critère est dit *des moindres carrés*. La solution est unique et de nature algébrique.

4.2 Projection sur un sous-espace

$E = \mathbb{R}^n$. Soient F un sous-espace de E, $\{\mathbf{f}\} = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_r\}$ une base quelconque de F, $\{\mathbf{v}\} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$ une base quelconque de E, $\mathbf{W}$ la matrice d'un produit scalaire Φ dans la base $\{\mathbf{v}\}$ et $\mathbf{F}$ la matrice qui contient en colonne les composantes des vecteurs $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_r$ dans la base $\{\mathbf{v}\}$. $\Pi = \text{Proj}_F^{\Phi}$ est le projecteur Φ-orthogonal sur F. Alors la matrice de Π dans la base $\{\mathbf{v}\}$ est :

$$[\Pi, \{\mathbf{v}\}] = \mathbf{F}\left(\mathbf{F}^t\mathbf{W}\mathbf{F}\right)^{-1}\mathbf{F}^t\mathbf{W}$$

Utiliser la formule ci-dessus pour projeter un vecteur sur un sous-espace dont on connaît une base orthogonale ou orthonormée. Préciser quand le sous-espace est de dimension 1 et quand il est engendré par les indicatrices d'une variable qualitative.

4.3 Variables sans redondance

Le critère des moindres carrés $\min\left(\|\mathbf{y}-\widehat{\mathbf{y}}\|^2_{\mathbf{D}}\right)$ est satisfait quand $\widehat{\mathbf{y}}$ est le projeté orthogonal au sens de $\mathbf{D}$ sur le sous espace engendré par l'ensemble des vecteurs $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^p$ et $\mathbf{1}_n$ puisque ce dernier assure le minimum de .

$$\left\|\mathbf{y}-\left(\alpha_1\mathbf{x}^1 + \alpha_2\mathbf{x}^2 + \cdots \alpha_p\mathbf{x}^p + \beta\mathbf{1}_n\right)\right\|^2_{\mathbf{D}}$$

Le vecteur projeté $\widehat{\mathbf{y}}$ existe et est unique. Les coefficients α_j et β ne sont par contre définis de manière unique que si les vecteurs $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^p$ et $\mathbf{1}_n$ sont indépendants. On dit dans ce cas que *les variables sont sans redondance*.

Les variables sont sans redondance si et seulement si $\mathbf{x}_0^1, \mathbf{x}_0^2, \cdots, \mathbf{x}_0^p$ et $\mathbf{1}_n$ sont indépendants. Dans ce cas :

$$\text{sev}\left(\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^p, \mathbf{1}_n\right) = \text{sev}\left(\mathbf{x}_0^1, \mathbf{x}_0^2, \cdots, \mathbf{x}_0^p, \mathbf{1}_n\right) = \text{sev}\left(\mathbf{x}_0^1, \mathbf{x}_0^2, \cdots, \mathbf{x}_0^p\right) \overset{\perp}{\oplus} \text{sev}\left(\mathbf{1}_n\right)$$

Les vecteurs $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^p$ et $\mathbf{1}_n$ sont indépendants si et seulement si les vecteurs $\mathbf{x}_0^1, \mathbf{x}_0^2, \cdots, \mathbf{x}_0^p$ le sont.

Donc, pour qu'un ensemble de variables explicatives soit sans redondance il faut et il suffit que la matrice des corrélations ou des covariances associée soit inversible.

4.4 Procédure de la régression multiple

La solution est obtenue dans le cas des variables sans redondance par :

$$\widehat{\mathbf{y}} = \text{Proj}_{\text{sev}\left(\mathbf{x}_0^1, \mathbf{x}_0^2, \cdots, \mathbf{x}_0^p\right)}(\mathbf{y}) + \text{Proj}_{\text{sev}(\mathbf{1}_n)}(\mathbf{y}) = \text{Proj}_1(\mathbf{y}) + \text{Proj}_2(\mathbf{y})$$
$$\widehat{\mathbf{y}} = \text{Proj}_1(\mathbf{y}_0) + \text{Proj}_2(\mathbf{y}_0) + \text{Proj}_1(\text{m}(\mathbf{y})\mathbf{1}_n) + \text{Proj}_2(\text{m}(\mathbf{y})\mathbf{1}_n)$$

Soit dans la base canonique :

$$\widehat{\mathbf{y}} = \mathbf{X}_0\left(\mathbf{X}_0^t\mathbf{D}\mathbf{X}_\mathbf{0}\right)^{-1}\mathbf{X}_0^t\mathbf{D}\mathbf{y}_\mathbf{0} + \text{m}(\mathbf{y})\mathbf{1}_n$$

On note l'apparition du vecteur :

$$\mathbf{d} = \mathbf{X}_0^t \mathbf{D} \mathbf{y_0} = \begin{bmatrix} \text{cov}\left(\mathbf{x}^1, \mathbf{y}\right) \\ \vdots \\ \text{cov}\left(\mathbf{x}^p, \mathbf{y}\right) \end{bmatrix} \quad \mathbf{a} = \left(\mathbf{X}_0^t \mathbf{D} \mathbf{X_0}\right)^{-1} \mathbf{X}_0^t \mathbf{D} \mathbf{y_0} = \mathbf{C}^{-1}\mathbf{d} = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$$

La solution s'écrit :

$$\widehat{\mathbf{y}} = \mathbf{X}_0 \mathbf{a} + \text{m}\left(\mathbf{y}\right) \mathbf{1}_n = a_1 \mathbf{x}_0^1 + \cdots + a_p \mathbf{x}_0^p + \text{m}\left(\mathbf{y}\right) \mathbf{1}_n$$
$$\widehat{\mathbf{y}} = a_1 \mathbf{x}^1 + \cdots + a_p \mathbf{x}^p + \left(\text{m}\left(\mathbf{y}\right) - a_1 \text{m}\left(\mathbf{x}^1\right) - \cdots - a_p \text{m}\left(\mathbf{x}^p\right)\right) \mathbf{1}_n$$
$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} = \mathbf{C}^{-1}\mathbf{d} \text{ et } \beta = \text{m}\left(\mathbf{y}\right) - a_1 \text{m}\left(\mathbf{x}^1\right) - \cdots - a_p \text{m}\left(\mathbf{x}^p\right)$$

Préciser cette solution quand il n'y a qu'une variable explicative (régression simple).

D'où la structure d'un programme de régression multiple :

1. Calcul des moyennes des variables explicatives et centrage du tableau $\mathbf{X}$
2. Calcul de la matrice des covariances des variables explicatives $\mathbf{C}$
3. Inversion de $\mathbf{C}$
4. Calcul de la moyenne de la variable à expliquer
5. Calcul du vecteur des covariances de la variable à expliquer et des variables explicatives $\mathbf{d}$
6. Calcul des coefficients de régression $\mathbf{a} = \mathbf{C}^{-1}\mathbf{d}$
7. Calcul de l'ordonnée à l'origine $\beta = \text{m}\left(\mathbf{y}\right) - a_1 \text{m}\left(\mathbf{x}^1\right) - \cdots - a_p \text{m}\left(\mathbf{x}^p\right)$
8. Calcul des valeurs prédites $\widehat{y_i} = \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots \alpha_p x_{ip} + \beta$
9. Calcul des résidus $e_i = y_i - \widehat{y_i}$
10. Calcul des prédictions pour des valeurs supplémentaires $\widehat{y_s} = \alpha_1 x_{s1} + \alpha_2 x_{s2} + \cdots \alpha_p x_{sp} + \beta$

Cette procédure est souvent utilisée pour estimer des valeurs manquantes.

4.5 Exercices

a) On considère les variables à $n = 4$ valeurs $\mathbf{x} = (-3, 0, 1, 2)$ et $\mathbf{y} = (0, 2, 2, 4)$.

Donner la valeur des paramètres qui minimise l'erreur :

$$E\left(f\right) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2$$

pour les fonctions

$$f_0(x) = a, f_1(x) = ax, f_2(x) = ax + b, f_3(x) = ax^2 + bx + c$$

Tracer le nuage de points et les courbes estimées. Préciser la variation de l'erreur d'un modèle à l'autre.

b) Autres exemples n'autorisant aucune approximation numérique : $\mathbf{x} = (0, 1, 2, 3)$ et $\mathbf{y} = (1, 3, 3, 9)$. $\mathbf{x} = (-2, -1, 0, 1, 2)$ et $\mathbf{y} = (-3, 1, 4, 1, -3)$. $\mathbf{x} = (-1, 0, 1, 2, 3)$ et $\mathbf{y} = (-2.5, 1.25, 1, -1.5, -4)$.

c) Un tableau d'observations sur une variable expérimentale a n lignes et p colonnes. Il est noté $\mathbf{X} = [x_{ij}]$. Un modèle de ce tableau est une matrice $\mathbf{M} = [m_{ij}]$. L'erreur associée au modèle est $E(\mathbf{M}) = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - m_{ij})^2$. Donner la solution qui minimise $E(\mathbf{M})$ pour les modèles du type $m_{ij}^1 = \alpha$, $m_{ij}^2 = \beta_i$, $m_{ij}^3 = \gamma_j$, $m_{ij}^4 = \beta_i + \gamma_j$, $m_{ij}^5 = \alpha + \beta_i + \gamma_j$. Pour le dernier modèle on impose les contraintes $\sum_{i=1}^{n} \beta_i = 0$ et $\sum_{j=1}^{p} \gamma_j = 0$. Expliciter quand et pourquoi les valeurs estimées sont uniques.

5 Représentation d'objets à trois dimensions

5.1 Repères dans $\mathbb{R}^3$

Représentation cartésienne et image euclidienne.

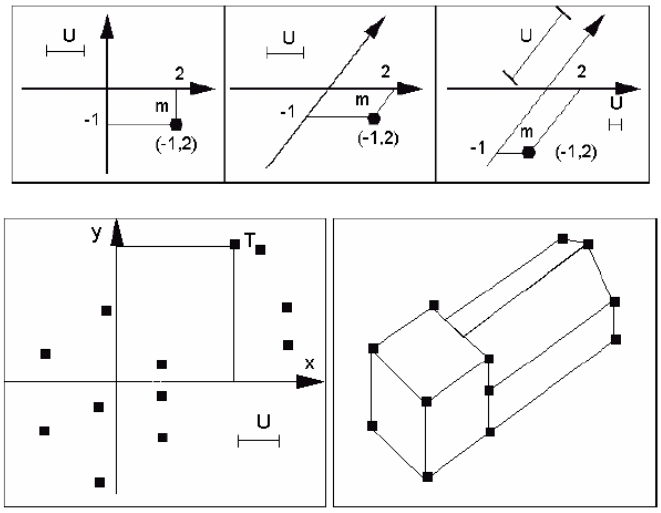


FIG. 5 – Entre la projection euclidienne de quelques points et la perception d'un objet, il y a l'interprétation. Il s'agit de donner un sens expérimental à un objet abstrait. C'est l'essentiel de la démarche statistique.

Soit trois variables donnant la proportion de la production en trois classes de qualité d'un grand centre de production d'huile d'olives (**x** extra, **y** moyenne, **z** lampante) :

	1972	1973	1974	1975	1976	1977	1978	1979	1980
i	1	2	3	4	5	6	7	8	9
x	70	54	49	43	28	18	26	25	12
y	25	27	33	28	36	18	32	13	6
z	5	19	18	29	36	64	42	62	82

Les données sont des points de $\mathbb{R}^3$.

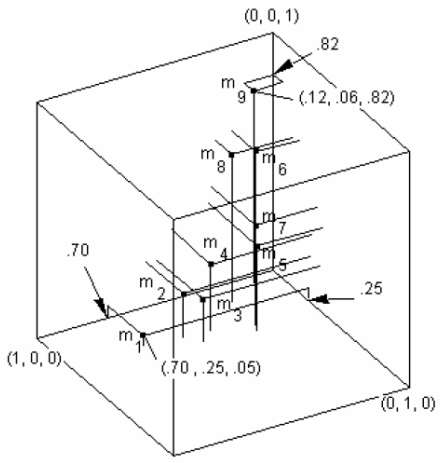
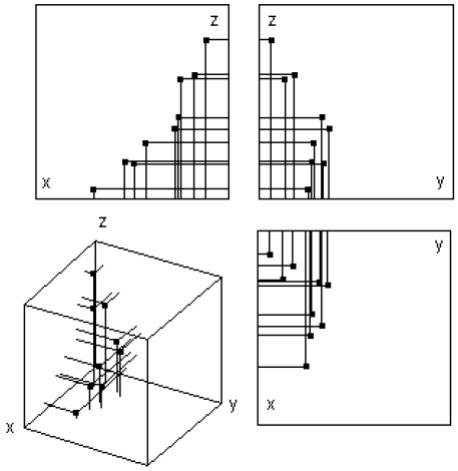


FIG. 6 – n points de $\mathbb{R}^3$: la figure utilise pour son exécution ce qu'elle essaye d'illustrer : la projection sur un plan (feuille de papier) dans $\mathbb{R}^3$.

Les nuages bivariés sont les projections sur les plans définis par la base canonique.



Définition d'une base associée à deux angles :

$$\mathbf{H} = [\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}, \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}] = \begin{bmatrix} \cos a \cos b & -\sin a & -\cos a \sin b \\ \sin a \cos b & \cos a & -\sin a \sin b \\ \sin b & 0 & \cos b \end{bmatrix}$$

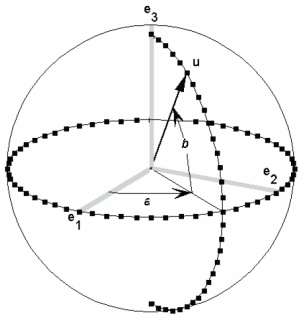


FIG. 7 – Définition d'une direction de projection par deux angles. Le plan perpendiculaire à cette direction recevra les projections.

Les coordonnées d'un vecteur dans une base orthonormée sont les produits scalaires avec les éléments de cette base. Ici on utilise la métrique canonique.

$$[\mathbf{m}, \{\mathbf{u}, \mathbf{v}, \mathbf{w}\}] = \mathbf{H}^t [\mathbf{m}, \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}]$$

Pour $a = 30^\circ$ et $b = 60^\circ$, par exemple :

$$\begin{bmatrix} 0.4330 & 0.25 & 0.8660 \\ -0.5 & 0.8660 & 0 \\ -0.75 & -0.4330 & 0.5 \end{bmatrix} \begin{bmatrix} 0.70 \\ 0.25 \\ 0.05 \end{bmatrix} = \begin{bmatrix} 0.4089 \\ -0.1335 \\ -0.6083 \end{bmatrix}$$

En utilisant les deux dernières coordonnées :

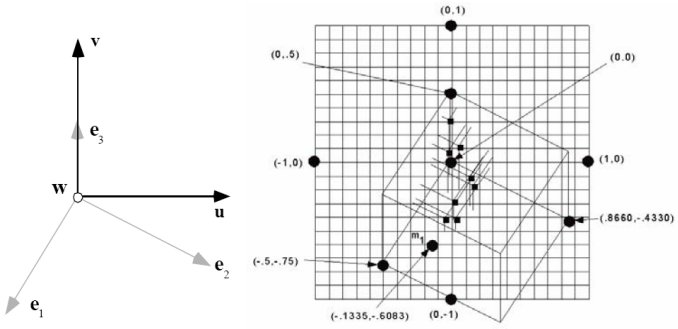
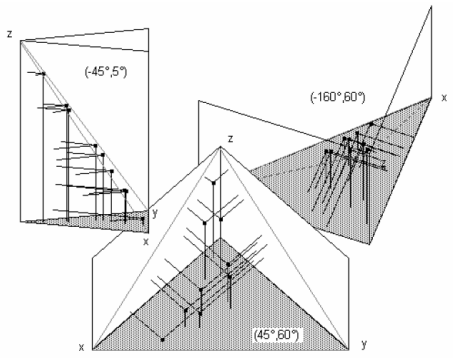


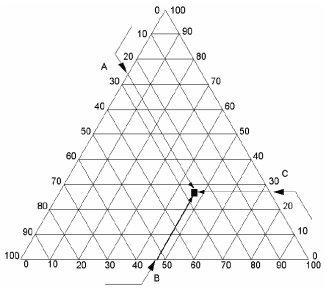
FIG. 8 – Passage du problème géométrique à sa mise en place numérique.

Ce qu'on voit sur une représentation euclidienne varie fortement d'une base à l'autre :



5.2 Représentation triangulaire

Mode de représentation traditionnel des données de fréquences à trois catégories :



C'est une image euclidienne.

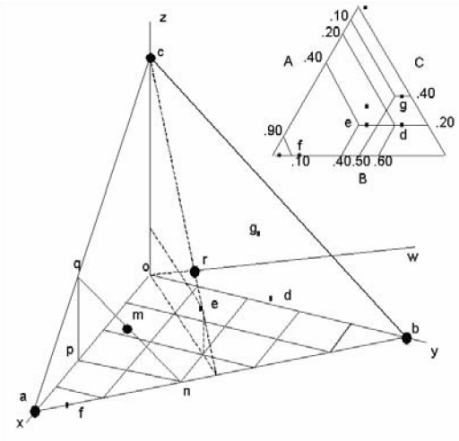


FIG. 9 – Lien entre le triangle de la représentation traditionnelle et la base canonique de $\mathbb{R}^3$.

Ceci conduit à une approche géométrique, une vision mécanique et une définition numérique de la représentation triangulaire :

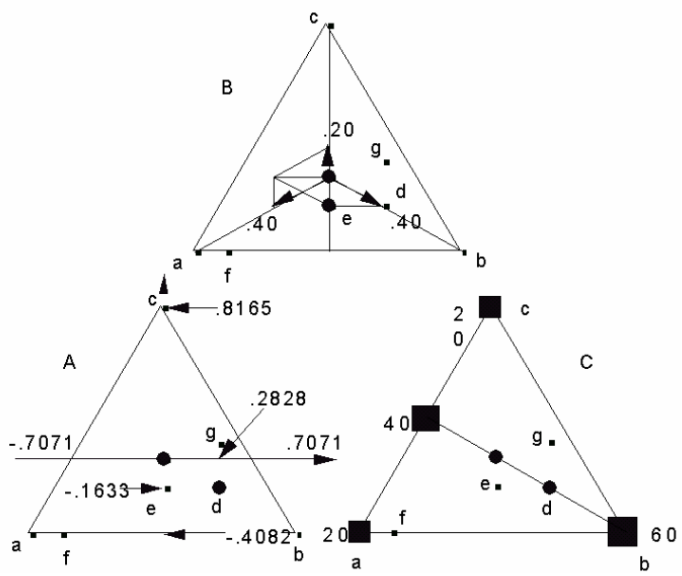
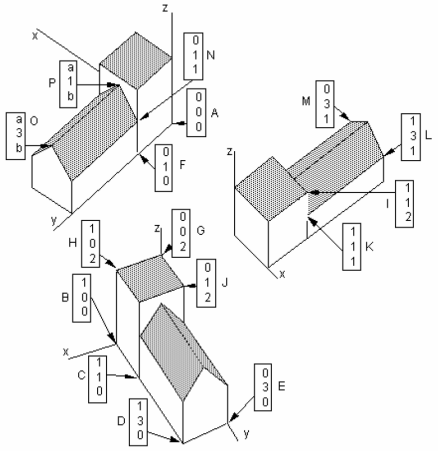


FIG. 10 – L'analyse en composante principale étend les schémas de la représentation triangulaire en dimension quelconque.

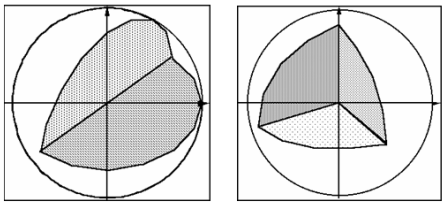
5.3 Exercices

a) Soit le cube défini par les 8 points A$(0,0,0)$, B$(0,1,0)$, C$(1,1,0)$, D$(1,0,0)$, E$(0,0,1)$, F$(0,1,1)$, G$(1,1,1)$ et H$(1,0,1)$. Représenter cet objet vu dans les directions $(30^\circ, 60^\circ)$ et $(60^\circ, 30^\circ)$

b) Représenter cet objet dans la directions de votre choix ($a = \frac{1}{2}$, $b = \sqrt{3}$).



c) S est la sphère de centre $(0,0,0)$ et de rayon unité. Représenter cet objet auquel on a enlevé tous les points (x,y,z) tels que $y > 0$ et $z > 0$ vu dans la direction $(45^\circ, 45^\circ)$.Représenter la même sphère de laquelle on a enlevé les points(x,y,z) tels que $x > 0$, $y > 0$ et $z > 0$ vue dans la direction $(60^\circ, 30^\circ)$.



d) Soit l'objet défini par les 8 points A$(2,0,0)$, B$(3,0,0)$, C$(3,1,0)$, D$(2,1,0)$, E$(2,0,1)$, F$(3,0,1)$, G$(3,1,1)$ et H$(2,1,1)$.

Représenter cet objet vu dans la direction $(60^\circ, 60^\circ)$. Sur le même graphique représenter la base associée à la direction $(30^\circ, 0^\circ)$ et la base canonique. Sur le même graphique représenter la projection de l'objet associée à la direction $(30^\circ, 0^\circ)$