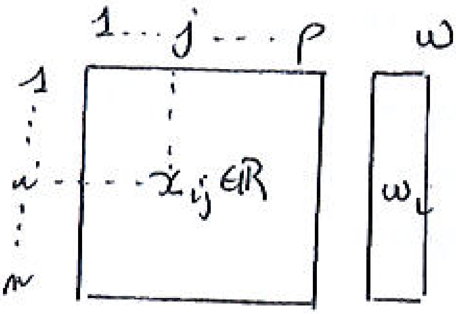
Notions de base pour l'analyse de données **quantitatives**

On se place dans le cadre d'un tableau de données numériques où n individus sont décrits sur p variables.



On notera :

* $X = (x_{ij})_{n \times p}$ la matrice des données "bruts"
  où x_{ij} = valeur du i-ème <u>individu</u> sur la j-ième <u>variable</u>

* $w_i > 0$: poids de l'individu i

* $x_i^t = (x_{i1}, \dots x_{ij}, \dots, x_{ip}) \in \mathbb{R}^p$ le i-ème vecteur ligne de X
  Il s'agit de la description du i-ème individu

* $x^j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{pmatrix} \in \mathbb{R}^n$ le j-ème vecteur colonne de X.
  Il s'agit de la description de la j-ème variable.

<u>Exemple</u> : On a mesuré la tension artérielle diastolique, systolique et le taux de cholestérol de 6 patients. Les résultats sont présentés dans le tableau ci-dessous :

Diast	Syst	Chol
90	140	6
60	85	5,9
75	135	6,1
70	145	5,8
85	130	5,4
70	145	5,0

$n =$

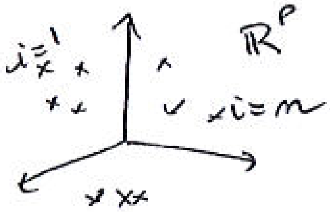
$p =$

$x_3^t =$

$x^2 =$

1) Nuage des n points-individus

Les n lignes de X définissent un nuage de n points de $\mathbb{R}^p$



En général, on pondère chaque individu i par $w_i > 0$.

En pratique,

$$\begin{cases} w_i = \frac{1}{n} \ \left(\text{parfois } \frac{1}{n-1}\right) & \text{en cas de tirage aléatoire} \\ w_i \neq \frac{1}{n} & \text{pour des échantillon redressés, des données regroupées....} \end{cases}$$

On notera :

$$\underset{n \times n}{\underset{\sim}{N}} = \operatorname{diag}(w_i) = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}$$

1.1) Centre de gravité du nuage des individus pondérés

On notera

$\bar{x} = (\bar{x}_1, \dots \bar{x}_j, \dots \bar{x}_p)$ ce centre de gravité

avec

$$\bar{x}_j = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \, x_{ij}$$

donc si $w_i = \frac{1}{n}$ ou $w_i = \frac{1}{n-1}$ $\forall i$,

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

1.2) Matrice centrée Y

Données brutes :

$$X = \begin{array}{c|ccc|} & 1 \dots & j & \dots p \\ \hline 1 & & \vdots & \\ \vdots & & & \\ i & \dots & x_{ij} & \\ \vdots & & & \\ n & & & \\ \hline \end{array}$$

$\bar{x} = \bar{x}_1, \bar{x}_j .. \bar{x}_p$

Données centrées :

$$\Rightarrow Y = \begin{array}{c|ccc|} & 1 \dots & j & \dots p \\ \hline 1 & & \vdots & \\ \vdots & & & \\ i & \dots & y_{ij} & \dots \\ \vdots & & & \\ n & & & \\ \hline \end{array} \quad \text{avec } y_{ij} = x_{ij} - \bar{x}_j$$

$\bar{y} = 0 \dots \dots 0$

$\Rightarrow$ Translation du nuage de points

Exemple : $\bar{x} =$

$Y =$

1.3) Variance empirique et matrice centrée-réduite Z

Variance empirique de la j-ème variable décrite par x^j :

$$\begin{cases} S_j^2 = \sum_{i=1}^{n} w_i \left(x_{ij} - \bar{x}_j\right)^2 & w_i = \frac{1}{n-1} \text{ : estimateur sans biais} \\ S_j = \text{écart-type} = \sqrt{S_j^2} & w_i = \frac{1}{n} \text{ : estimateur biaisé} \end{cases}$$

$\Rightarrow$ S_j^2 mesure la dispersion de la j-ème variable (colonne)

Données brutes : Données centrées-réduites

$$X = \begin{array}{c|ccc|} & 1 \cdots & j & \cdots p \\ 1 & & & \\ \vdots & & \vdots & \\ i & \cdots & x_{ij} & \\ \vdots & & & \\ n & & & \end{array} \quad \Rightarrow \quad Z = \begin{array}{c|ccc|} & 1 \cdots & j & \cdots p \\ 1 & & & \\ \vdots & & \vdots & \\ i & \cdots & z_{ij} & \\ \vdots & & & \\ n & & & \end{array} \quad \text{avec } z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$$

$S : (S_1 \cdots S_j \cdots S_p)$ $\qquad$ $\bar{z} : (0 \cdots 0)$

$\qquad\qquad\qquad\qquad$ $S : (1 \cdots 1)$

$\Rightarrow$ On divise chaque colonne de la matrice centrée par son écart-type

$\Rightarrow$ Variance $= 1$ dans toutes les directions du nuage centré-réduit.

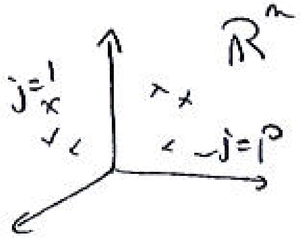
Exemple : $S^2 =$

$S =$

$Z =$

2) Nuage des p points-variables

Les p colonnes de X définissent un nuage de p points de $\mathbb{R}^n$



On notera :

$$\underbrace{M}_{p\times p} = \mathrm{diag}\left(1/s_j^2\right) = \begin{pmatrix} 1/s_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/s_p^2 \end{pmatrix}$$

remarque : $Z = Y M^{1/2}$

2.1) Matrice de variance-covariance V

- Covariance empirique entre j et j' :

$$s_{jj'} = \sum_{i=1}^{n} w_i \left(x_{ij} - \bar{x}_j\right)\left(x_{ij'} - \bar{x}_{j'}\right)$$

- Matrice de var-cov :

$$V = \left(s_{jj'}\right)_{p\times p} = Y^t N Y$$

2.2) Matrice des corrélations R

- Corrélation entre j et j' :

$$r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}} = \sum_{i=1}^{n} w_i \overbrace{\left(\frac{x_{ij}-\bar{x}_j}{s_j}\right)}^{z_{ij}} \left(\frac{x_{ij'}-\bar{x}_{j'}}{s_{j'}}\right) = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ij'}-\bar{x}_{j'}}{\sum_i (x_{ij}-\bar{x}_j)^2 \sum_i (x_{ij'}-\bar{x}}$$

indépendant du choix de w_i !

- Matrice des corrélations :

$$R = \left(r_{jj'}\right)_{p\times p} = Z^t N Z$$

Exemple : $V=$ $R=$

③ Métriques

3.1) Métrique sur l'espace $\mathbb{R}^p$ des points-individus

* . Rappel : Soit M une matrice $p\times p$, symétrique définie positive. Alors M définit sur $\mathbb{R}^p$:
  - un produit scalaire : $\langle x,y\rangle_M = x^t M y$
  - une norme : $\|x\|_M = \langle x,x\rangle_M^{1/2}$
  - une distance : $d_M(x,y) = \|x-y\|_M$
  - des angles $\cos\theta_M(x,y) = \dfrac{\langle x,y\rangle_M}{\|x\|_M \cdot \|y\|_M}$

De plus, pour M donné, on peut définir :

- une matrice A est M-symétrique si $(MA)^t = MA$
- deux vecteurs x et y sont M-orthogonaux si $\langle x,y\rangle_M = 0$
- un vecteur x est M-normé à 1 si $\|x\|_M = 1$.

* On munit l'espace $\mathbb{R}^n$ de la métrique M pour mesurer la distance entre deux individus i et i' :

$$d_M^2(x_i,x_{i'}) = (x_i - x_{i'})^t M (x_i - x_{i'})$$

Donc,

→ si $M=I$, $d_M^2(x_i,x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$ → distance Euclidienne simple

exemple : $d_I^2(1,2) =$

→ si $M=D_{1/s^2}$, $d_M^2(x_i,x_{i'}) = \sum_{j=1}^{p} \frac{1}{s_j^2}(x_{ij}-x_{i'j})^2$ → distance Euclidienne normalisée par l'inverse de la variance :

$$= \sum_{j=1}^{p} (z_{ij} - z_{i'j})^2 \rightarrow d_{D_{1/s^2}}(x_i,x_{i'}) = d_I(z_i,z_{i'})$$

exemple : $d_{D_{1/s^2}}^2(1,2) =$

3.2) Métrique sur l'espace $\mathbb{R}^n$ des points-variables

En général, on mesure la proximité entre deux variables j et j' par leur covariance empirique $s_{jj'}$, ou par leur corrélation $r_{jj'}$:

- On munit $\mathbb{R}^n$ de la métrique $N = \operatorname{diag}(w_i)$ et on a

$$* \quad s_{jj'} = \sum_i w_i \underbrace{(x_{ij} - \bar{x}_j)}_{y_{ij}} \underbrace{(x_{ij'} - \bar{x}_{j'})}_{y_{ij'}}$$

$$= (y^j)^t N y^{j'} = \langle y^j, y^{j'} \rangle_N$$

$\Rightarrow$ covariance entre x^j et $x^{j'}$ est le produit scalaire (associé à N) entre les variables centrées y^j et $y^{j'}$.

$$* \quad s_j^2 = \| y^j \|_N$$ $\Rightarrow$ variance de x^j est égale à la N-norme de la variable centrée y^j

$$* \quad r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}} = \frac{\langle y^j, y^{j'} \rangle_N}{\|y^j\|_N \, \|y^{j'}\|_N} = \cos \theta_N (y^j, y^{j'})$$

$$= \langle z^j, z^{j'} \rangle_N$$

$\Rightarrow$ • corrélation entre x^j et $x^{j'}$ est le cosinus de l'angle entre les variables centrées y^j et $y^{j'}$.

• corrélation entre x^j et $x^{j'}$ est le produit scalaire entre les variables centrées-réduites z^j et $z^{j'}$

4) Inertie du nuage des points-individus

L'inertie totale du nuage est la moyenne pondérée des carrés des distances des n points de $\mathbb{R}^p$ au centre de gravité $\bar{x}$:

$$I(X) = \sum_{i=1}^{n} w_i \, d_M^2 (x_i, \bar{x})$$

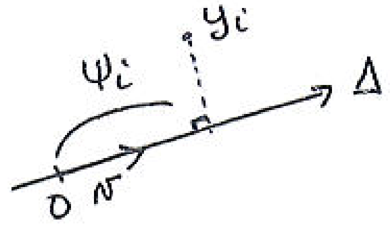
Montrer que si :

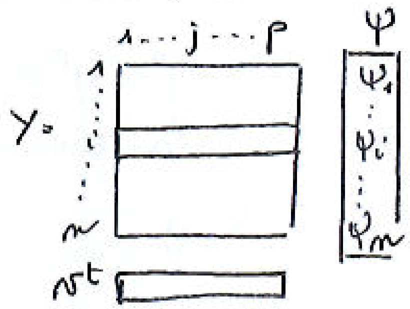
$M = I$, $\quad w_i = \frac{1}{n}$ (ou $\frac{1}{n-1}$) $\forall i$, alors $I(X) = s_1^2 + \dots + s_p^2$

$M = D_{1/s^2}$, $\quad$ ____________, alors $I(X) = p$

On considère le nuage des n points individus centrés $y_1 \cdots y_n \in \mathbb{R}^n$, pondérés par $\underset{n\times n}{N} = \mathrm{diag}(w_i)$ et munis de la métrique $\underset{p\times p}{M}$

- la projection M-orthogonale des n points individus <u>centrés</u> $y_1 \cdots y_n \in \mathbb{R}^p$, sur un axe Δ engendré par un vecteur unitaire v de M norme égale à 1 (i.e. $v^t M v = 1$) est un vecteur $\psi \in \mathbb{R}^n$





On a : $\psi_i = \langle y_i, v \rangle_M = y^t M v$ $\quad (b = Mv)^t$

$$\Rightarrow \psi = \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_i \\ \vdots \\ \psi_n \end{pmatrix} = \underset{n\times p}{Y} \cdot \underset{p\times p}{M} \cdot \underset{p\times 1}{v} = \underset{n\times p}{Y} \underset{p\times 1}{b} = \sum_{j=1}^{p} b_j y^j$$

$\Rightarrow$ ψ est une combinaison linéaire des colonnes de Y y^j avec les coefficients b_j

$\Rightarrow$ ψ est un "résumé" des colonnes de Y appelée <u>variable "synthétique"</u>

<u>Exemple</u> : $v_1 = \begin{pmatrix} 0,6407 \\ 0,720 \\ -0,265 \end{pmatrix}$ $\quad M = I_3 \quad w_i = \frac{1}{n}$

$\Rightarrow$ Vecteur ψ_1 des projections I-orthogonales des 6 points-individus <u>centrés-réduits</u> sur l'axe Δ_1 de vecteur directeur v_1 :

$$\psi_1 = Z v = 0,64 \begin{pmatrix} \\ \\ \end{pmatrix} + 0,720 \begin{pmatrix} \\ \\ \end{pmatrix} - 0,265 \begin{pmatrix} \\ \\ \end{pmatrix} =$$

$\|v_1\|^2 = 1$.

$v_2 = \begin{pmatrix} 0,445 \\ -0,065 \\ 0,894 \end{pmatrix}$ Vérifiez que v_2 est I-orthogonal à v_1 : $v_1^t v_2 = 0$

Calculer ψ_2^2 le vecteur des projections des 6 points-individus centrés-réduits sur l'axe Δ_2 de vecteur directeur v_2.

$\psi^2 = \begin{pmatrix} \\ \\ \\ \end{pmatrix}$

- Variance de la variable synthétique ψ :

→ Comme les colonnes $y^1 \cdots y^p$ de Z et les colonnes $z^1 \cdots z^p$ de Z sont centrées, la combinaison linéaire,

$$\psi = \sum_{j=1}^{p} b_j y^j$$

Exemple : $\overline{\psi_1} = \overline{\psi_2} = 0$

est centrée : $\overline{\psi} = 0$

TP : Reprendre l'exemple du cours dans R (en vous aidant du code se trouvant dans ~~le~~ fichier exemple-notions-bases.R) et dans Excel !