Description statistique des données

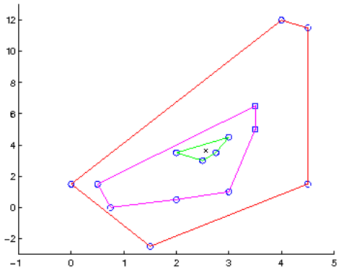
M8 – Chapitre 2

• **Variables :**	$X = \{a_1, \dots a_l\}$ $Y = \{b_1, \dots, b_k\}$
• **Échantillon :**	$S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ $(x_i, y_i) \in X \times Y$
• **Lignes :**	$n_{i\bullet} = \sum_{j=1}^{k} n_{ij}$
• **Colonnes :**	$n_{\bullet j} = \sum_{i=1}^{l} n_{ij}$
• **Total :**	$n_{\bullet\bullet} = \sum_{i=1}^{l}\sum_{j=1}^{k} n_{ij}$

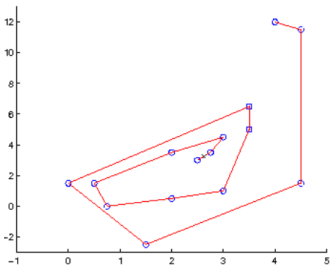
Tableau de contingence :

X \ Y	b_1	…	b_j	…	b_k	marge
a_1	n_{11}	…	n_{1j}	…	n_{1k}	$n_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
a_i	n_{i1}	…	n_{ij}	…	n_{ik}	$n_{i\bullet}$
⋮	⋮		⋮		⋮	⋮
a_l	n_{l1}	…	n_{lj}	…	n_{lk}	$n_{l\bullet}$
marge	$\boldsymbol{n_{\bullet 1}}$	…	$\boldsymbol{n_{\bullet j}}$	…	$\boldsymbol{n_{\bullet k}}$	$\boldsymbol{n_{\bullet\bullet} = n}$

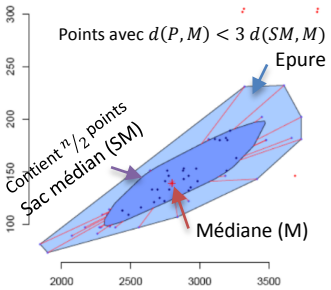
• **Fréquence :**	$f_{ij} = \frac{n_{ij}}{n} = \hat{\mathbb{P}}_{ij} = \hat{\mathbb{P}}(X = a_i \cap Y = b_j)$
• **Profil ligne :**	$l_{ij} = \frac{n_{ij}}{n_{i\bullet}} = \hat{\mathbb{P}}(Y = b_j \mid X = a_i)$
• **Profil colonne :**	$c_{ij} = \frac{n_{ij}}{n_{\bullet j}} = \hat{\mathbb{P}}(X = a_i \mid Y = b_j)$
• **Distribution marginale :**	$\frac{n_{i\bullet}}{n} = \hat{\mathbb{P}}_i \qquad \frac{n_{\bullet j}}{n} = \hat{\mathbb{P}}_j$
• **Moyenne :**	$\min_{(\bar{x},\bar{y})\in\mathbb{R}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + (y_i - \bar{y})^2$
• **Médiane :**	$\min_{(\overline{M_x},\overline{M_y})\in\mathbb{R}^2} \sum_{i=1}^{n} \lvert x_i - \overline{M_x}\rvert + \lvert y_i - \overline{M_y}\rvert$
• **Médiane euclidienne :**	$\min_{(\overline{M_x},\overline{M_y})\in\mathbb{R}^2} \sum_{i=1}^{n} \sqrt{\lvert x_i - \overline{M_x}\rvert^2 + \lvert y_i - \overline{M_y}\rvert^2}$
• **Information mutuelle : (X, Y discrètes)**	$I(X,Y) = \sum_{i\in\Omega_X}\sum_{j\in\Omega_Y} \hat{\mathbb{P}}_{ij} \log\left(\frac{\hat{\mathbb{P}}_{ij}}{\hat{\mathbb{P}}_i \hat{\mathbb{P}}_j}\right)$ X indep $Y \Rightarrow I(X,Y) = 0$
• **Dépendance : (X, Y quantitatives)**	$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})$ (Covariance) $s_X^2 = s_{XX} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $s_Y^2 = s_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ (Variance) $r = \frac{s_{XY}}{s_X s_Y}$ Corrélation $\in [-1;1]$ $r \approx 0 \Rightarrow x$ indep y $r \approx 1 \Rightarrow x \nearrow y \nearrow$ $r \approx -1 \Rightarrow x \nearrow y \searrow$
• **Dépendance : (X, Y qualitatives)** *Distance du χ^2*	$D_{\chi^2} = \sum_{i\in\Omega_X}\sum_{j\in\Omega_Y} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} = n\sum_{i\in\Omega_X}\sum_{j\in\Omega_Y} \frac{(\overbrace{\hat{\mathbb{P}}_{ij}}^{\text{eff obs}} - \overbrace{\hat{\mathbb{P}}_i\hat{\mathbb{P}}_j}^{\text{eff th}})^2}{\hat{\mathbb{P}}_i\hat{\mathbb{P}}_j}$
• **Dépendance : (X qual, Y quant)** *Coef. de détermination*	$S^2_{Y/X} = \frac{\sum_{j\in\Omega_X} n_j \left(\bar{y}_j - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$



Médiane de Tukey



Médiane de Jarvis



Sac médian

Description statistique des données

M8 – Chapitre 2

I. Matrice de covariance et matrice de corrélation

$X_{C_{i,j}} = X_{i,j} - \bar{X_j}$ Matrice centrée

$X_{R_{i,j}} = \frac{X_{i,j} - \bar{X_j}}{s_j}$ Matrice centrée réduite

$\boxed{\Sigma = \frac{{X_C}^{\mathrm{T}} X_C}{n}}$ Matrice de covariance

$\boxed{\rho = \frac{{X_R}^{\mathrm{T}} X_R}{n}}$ Matrice de corrélation

II. L'ACP

1. Le problème

On veut reconstruire au mieux X par le produit uv^{T}, c'est-à-dire minimiser :

$$\min_{u,v} J(u,v) = \min_{u,v} \left\| X - uv^{\mathrm{T}} \right\|_F^2 = \min_{u,v} -2(Xv)^{\mathrm{T}}u + \|u\|^2 + \|v\|^2 = \min_{u,v} \sum_i^n \sum_j^p \left(x_{ij} - u_i v_j\right)^2$$

2. La solution

$$\begin{cases} \nabla_u J(u) = -2Xv + 2\|v\|^2 u = 0 & \Leftrightarrow \quad Xv = \|v\|^2 u \\ \nabla_v J(v) = -2X^{\mathrm{T}}u + 2\|u\|^2 v = 0 & \Leftrightarrow \quad -X^{\mathrm{T}}u = \|u\|^2 v \end{cases} \Leftrightarrow X^{\mathrm{T}}Xv = \underbrace{\|u\|^2\|v\|^2}_{\lambda} v$$

La solution est un vecteur propre v de la matrice $X^{\mathrm{T}}X$. C'est celle qui a la plus grande valeur propre car à l'optimum on a $J(u,v) = -\lambda$.

3. Remarques sur les valeurs et vecteurs propres

- Les valeurs propres sont ordonnées : $\lambda_1 \geq \lambda_2$
- Les vecteurs propres sont normées : ${v_i}^{\mathrm{T}} v_i = 1$
- Les vecteurs propres sont orthogonaux : ${v_i}^{\mathrm{T}} v_j = 0$

4. L'ACP

$$[v, \lambda] = eig(X) \quad v : \text{vecteurs propres} \quad diag(\lambda) : \text{valeurs propres}$$

Les composantes principales : $u = Xv$

Corrélations entre variables observées normalisées Xn (centrées réduites) et les composantes principales u :

$$cor(X_R, u_i) = \frac{{X_R}^T u_i}{\sqrt{n}\|u_i\|} = \frac{\|u_i\|}{\sqrt{n}} v = \frac{\sqrt{\lambda_i}}{\sqrt{n}} v$$

Rôle des variables :

$$V_n = \sqrt{\frac{\lambda}{n}} V$$