

## TP2 – Droites de régression

### Exercice 1 : estimation de $D_{YX}$ par maximum de vraisemblance

Lancez le script `donnees.m`, qui affiche  $n$  points  $P_i = (x_i, y_i)$  du plan censés être alignés. En utilisant l'équation paramétrique  $y = ax + b$  d'une droite et en notant  $r_{a,b}(P_i) = y_i - ax_i - b$  le résidu de cette équation pour le point  $P_i$ , il est légitime de modéliser ces résidus par une loi normale centrée :

$$f_{\sigma,a,b}(P_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{r_{a,b}(P_i)^2}{2\sigma^2}\right\} \quad (1)$$

La droite de régression de  $Y$  en  $X$  d'un tel nuage de points, notée  $D_{YX}$ , est la droite d'équation paramétrique  $y = a^*x + b^*$ , où  $a^*$  et  $b^*$  sont les valeurs des paramètres  $a$  et  $b$  qui maximisent la log-vraisemblance :

$$(\sigma^*, a^*, b^*) = \arg \max_{(\sigma,a,b) \in \mathbb{R}^+ \times \mathbb{R}^2} \left\{ \ln \left[ \prod_{i=1}^n f_{\sigma,a,b}(P_i) \right] \right\} = \arg \min_{(\sigma,a,b) \in \mathbb{R}^+ \times \mathbb{R}^2} \left\{ n \ln \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^n r_{a,b}(P_i)^2 \right\} \quad (2)$$

Si l'on suppose l'écart-type du bruit  $\sigma$  fixé, alors le problème se simplifie :

$$(a^*, b^*) = \arg \min_{(a,b) \in \mathbb{R}^2} \left\{ \sum_{i=1}^n r_{a,b}(P_i)^2 \right\} = \arg \min_{(a,b) \in \mathbb{R}^2} \left\{ \sum_{i=1}^n (y_i - ax_i - b)^2 \right\} \quad (3)$$

La résolution de (3) par tirages aléatoires n'est pas aussi simple qu'il y paraît, car d'une part, les inconnues  $a$  et  $b$  ne sont pas bornées, et d'autre part,  $a$  ne suit pas une loi uniforme. Néanmoins, il est facile de montrer que  $D_{YX}$  contient le centre de gravité  $G$ . On peut donc calculer les coordonnées  $(x_G, y_G)$  de  $G$ , puis centrer les données. L'équation de  $D_{YX}$  devenant  $y' = a^*x'$  après changement d'origine, le problème se simplifie encore :

$$a^* = \arg \min_{a \in \mathbb{R}} \left\{ \sum_{i=1}^n (y'_i - ax'_i)^2 \right\} = \tan \left[ \arg \min_{\psi \in ]-\frac{\pi}{2}, \frac{\pi}{2}[} \left\{ \sum_{i=1}^n (y'_i - \tan \psi x'_i)^2 \right\} \right] \quad (4)$$

Dans (4), la deuxième égalité vient de ce que le paramètre  $a$  d'une droite est égal à la tangente de son angle polaire  $\psi$ . La résolution de (4) peut être effectuée par tirages aléatoires de  $\psi$ , selon une loi uniforme sur  $] -\frac{\pi}{2}, \frac{\pi}{2}[$ .

Complétez le script `exercice_1.m` de façon à estimer  $\psi^*$ , puis les paramètres  $a^*$  et  $b^*$ , en suivant ce procédé.

### Exercice 2 : estimation de $D_{YX}$ par résolution d'un système linéaire

Le critère à minimiser dans (2) s'écrit  $\mathcal{F}(\sigma, a, b) = n \ln \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^n r_{a,b}(P_i)^2$ . Le problème (2) peut donc également être considéré comme un problème d'optimisation différentiable. En notant  $\mathcal{G}(a, b) = \sum_{i=1}^n r_{a,b}(P_i)^2$  :

$$\nabla \mathcal{F}(\sigma, a, b) = 0 \iff \begin{cases} \nabla_{\sigma} \mathcal{F}(\sigma, a, b) = 0 \\ \nabla_{a,b} \mathcal{F}(\sigma, a, b) = 0 \end{cases} \iff \begin{cases} \sigma^2 = \frac{1}{n} \sum_{i=1}^n r_{a,b}(P_i)^2 \\ \nabla \mathcal{G}(a, b) = 0 \end{cases} \quad (5)$$

La première de ces équations était prévisible, puisque c'est la définition même de la variance. Quant à la deuxième équation, elle correspond à l'optimalité du critère à minimiser dans (3). Or, ce critère s'écrit aussi :

$$\mathcal{G}(a, b) = \|AX - B\|^2, \text{ avec } A = \begin{bmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{bmatrix}^T, X = \begin{bmatrix} a \\ b \end{bmatrix} \text{ et } B = [y_1 \quad \cdots \quad y_n]^T \quad (6)$$

Minimiser  $\mathcal{G}(a, b)$  revient donc à chercher une solution approchée du système linéaire  $AX = B$ , au sens des moindres carrés ordinaires. Le problème se résout en écrivant les équations normales  $A^T AX = A^T B$ , dont la solution s'écrit  $X^* = (A^T A)^{-1} A^T B = A^+ B$ , où  $A^+ = (A^T A)^{-1} A^T$  est la matrice pseudo-inverse de  $A$  (`pinv`).

Complétez le script `exercice_2.m` de manière à comparer cette méthode d'estimation de  $D_{YX}$  avec celle de l'exercice 1. Observez l'évolution des résultats lorsque la valeur de `nb_tirages` diminue. Testez également différentes valeurs du nombre  $n \geq 2$  de données.

### Exercice 3 : estimation de $D_{\perp}$ par maximum de vraisemblance

Une droite  $D$  du plan peut également être définie par son *équation cartésienne normalisée*  $x \cos \theta + y \sin \theta = \rho$  :

- $\theta$  est l'angle polaire du vecteur  $\vec{v}$  orthogonal à  $D$ , de norme 1, tel que  $\theta \in ]0, \pi]$  (cf. figure 1).
- À partir d'un point  $P = (x, y)$  quelconque de  $D$ , on peut calculer le second paramètre  $\rho = x \cos \theta + y \sin \theta$ .

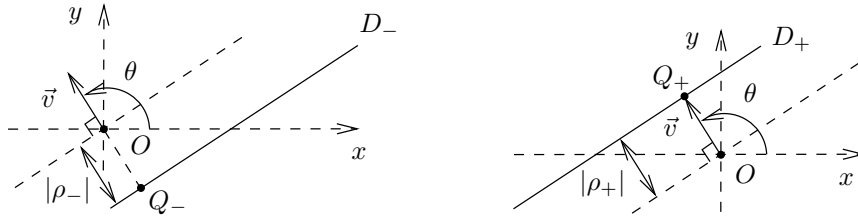


FIGURE 1 – Les droites  $D_-$  et  $D_+$  correspondent aux paramètres  $(\theta, \rho_-) = (130, -1)$  et  $(\theta, \rho_+) = (130, 1)$ .

Si l'on désigne par  $r_{\theta, \rho}(P_i) = x_i \cos \theta + y_i \sin \theta - \rho$  le résidu au point  $P_i$  de l'équation cartésienne normalisée, il semble à nouveau légitime de modéliser ces résidus par une loi normale centrée :

$$f_{\sigma, \theta, \rho}(P_i) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{r_{\theta, \rho}(P_i)^2}{2\sigma^2} \right\} \quad (7)$$

La *droite de régression en distance orthogonale* de ce nuage de points, notée  $D_{\perp}$ , est la droite d'équation  $x \cos \theta^* + y \sin \theta^* = \rho^*$ , où  $\theta^*$  et  $\rho^*$  sont les valeurs des paramètres  $\theta$  et  $\rho$  qui maximisent la log-vraisemblance :

$$(\sigma^*, \theta^*, \rho^*) = \arg \max_{(\sigma, \theta, \rho) \in \mathbb{R}^+ \times ]0, \pi] \times \mathbb{R}} \left\{ \ln \left[ \prod_{i=1}^n f_{\sigma, \theta, \rho}(P_i) \right] \right\} = \arg \min_{(\sigma, \theta, \rho) \in \mathbb{R}^+ \times ]0, \pi] \times \mathbb{R}} \left\{ n \ln \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^n r_{\theta, \rho}(P_i)^2 \right\} \quad (8)$$

En supposant  $\sigma$  fixé, et sachant que la droite de régression  $D_{\perp}$  contient elle aussi le centre de gravité  $G$ , la résolution du problème (8) est en tout point analogue à celle du problème (2). Par analogie avec (4) :

$$\theta^* = \arg \min_{\theta \in ]0, \pi]} \left\{ \sum_{i=1}^n (x'_i \cos \theta + y'_i \sin \theta)^2 \right\} \quad (9)$$

Complétez le script `exercice_3.m` de manière à estimer les paramètres  $\theta^*$  et  $\rho^*$  en suivant ce procédé.

### Exercice 4 : estimation de $D_{\perp}$ par résolution d'un système linéaire

Le critère  $\mathcal{I}(\theta) = \sum_{i=1}^n (x'_i \cos \theta + y'_i \sin \theta)^2$  à minimiser dans (9) s'appelle l'*inertie*. Il s'écrit également :

$$\mathcal{I}(\theta) = \|CY\|^2, \text{ avec } C = \begin{bmatrix} x'_1 & \cdots & x'_n \\ y'_1 & \cdots & y'_n \end{bmatrix}^T \text{ et } Y = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \quad (10)$$

Or, la solution approchée du système linéaire  $CY = O$ , au sens des moindres carrés ordinaires, vaut  $C^+O = O$ . Pour éviter cette solution, on impose la contrainte  $\|Y\| = 1$  (résolution approchée au sens des *moindres carrés totaux*). Ce nouveau problème se résout en introduisant le *lagrangien*  $\mathcal{L}(Y, \lambda) = \|CY\|^2 + \lambda(1 - \|Y\|^2)$ , où  $\lambda$  est un *multiplicateur de Lagrange*. La condition d'optimalité de  $\mathcal{L}$  s'écrit :

$$\nabla \mathcal{L}(Y, \lambda) = 0 \iff \begin{cases} \nabla_Y \mathcal{L}(Y, \lambda) = 0 \\ \nabla_{\lambda} \mathcal{L}(Y, \lambda) = 0 \end{cases} \iff \begin{cases} C^T CY = \lambda Y \\ \|Y\| = 1 \end{cases} \quad (11)$$

Sachant que  $C^T C$  est symétrique réelle, cette matrice admet une base orthonormée de vecteurs propres. De plus, comme  $C^T C$  est *semi-définie positive*, ses valeurs propres sont positives ou nulles. Le minimiseur de  $\mathcal{I}(\theta)$  de norme 1, noté  $Y^*$ , est donc un des deux vecteurs propres associés à la plus petite valeur propre de  $C^T C$ . En effet, pour un vecteur propre  $Y_p$  de norme 1, associé à la valeur propre  $\lambda_p$  :  $\|CY_p\|^2 = Y_p^T C^T CY_p = \lambda_p Y_p^T Y_p = \lambda_p$ .

Complétez le script `exercice_4.m` de manière à comparer cette méthode d'estimation de  $D_{\perp}$  avec celle de l'exercice 3. Observez l'évolution des résultats pour différentes valeurs de `nb_tirages` et de `n`.

Enfin, affichez sur une même figure les droites de régression  $D_{YX}$  et  $D_{\perp}$  : que constatez-vous ?