TP1 – Droites de régression

Afin que vous n'ayez qu'un seul fichier à rendre pour ce TP, au lieu de créer un fichier pour chaque fonction que vous aurez à écrire, un fichier nommé `fonctions_TP1_stat` vous est fourni pour que vous complétiez les différentes fonctions qui y sont présentes.

Lancez le script `donnees` qui génère et affiche une droite ainsi que des points P_i autour de cette dernière, simulant du bruit sur les données. À partir de ces données, on va chercher à estimer les paramètres de la droite de quatre manières différentes qui sont :

- par le maximum de vraisemblance à partir de l'équation paramétrique,
- par les moindres carrés à partir de l'équation paramétrique,
- par le maximum de vraisemblance à partir de l'équation cartésienne normalisée,
- par les moindres carrés à partir de l'équation cartésienne normalisée.

Exercice 1 : estimation de $\mathcal{D}_{YX}$ par le maximum de vraisemblance

Si n points $P_i = (x_i, y_i)$ du plan se situent au voisinage d'une droite $\mathcal{D}$ d'*équation paramétrique* $y = ax + b$, il est légitime de modéliser les résidus $r_{(a,b)}(P_i) = y_i - a\,x_i - b$ par une loi normale centrée d'écart-type σ :

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

La *droite de régression de Y en X* d'un tel nuage de points, notée $\mathcal{D}_{YX}$, 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 :

$$(\sigma^*, a^*, b^*) = \underset{(\sigma,a,b)\in\mathbb{R}^+\times\mathbb{R}^2}{\arg\max} \left\{\ln \prod_{i=1}^{n} f_{(\sigma,a,b)}(P_i)\right\} = \underset{(\sigma,a,b)\in\mathbb{R}^+\times\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n}\left\{\ln\sigma + \frac{r_{(a,b)}(P_i)^2}{2\,\sigma^2}\right\} \tag{2}$$

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

$$(a^*, b^*) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} r_{(a,b)}(P_i)^2 = \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (y_i - a\,x_i - b)^2 \tag{3}$$

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 $\mathcal{D}_{YX}$ contient le centre de gravité G des points P_i. On peut donc calculer les coordonnées (x_G, y_G) de G, puis centrer les données. L'équation de $\mathcal{D}_{YX}$ devenant $y' = a^*x'$ après changement d'origine, et le problème se simplifie encore :

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

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

Dans un premier temps, complétez la fonction `centrage_des_donnees` qui retourne les coordonnées x_G et y_G du centre de gravité ainsi que les vecteurs centrés des données ($x'_i = x_i - x_G$ et $y'_i = y_i - y_G$). Afin d'effectuer les tirages aléatoires de ψ selon une loi uniforme, complétez également la fonction `tirages_aleatoires_uniformes` en vous basant sur celle du TP1 de Probabilités. Complétez ensuite la fonction `estimation_Dyx_MV`, appelée par le script `exercice_1`, permettant de résoudre le problème (4) correspondant au maximum de vraisemblance pour l'équation paramétrique.

Exercice 2 : estimation de $\mathcal{D}_{YX}$ par les moindres carrés

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

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

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 :

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

Minimiser $\mathcal{G}(a,b)$ revient donc à chercher une solution approchée du système linéaire $AX = B$, au sens des moindres carrés (voir cours d'Analyse de Données au second semestre). Le problème se résout en écrivant les *équations normales* $A^\top AX = A^\top B$, dont la solution s'écrit $X^* = (A^\top A)^{-1}A^\top B = A^+ B$, où $A^+ = (A^\top A)^{-1}A^\top$ est la *matrice pseudo-inverse* de A.

Complétez la fonction `estimation_Dyx_MC`, appelée par le script `exercice_2`, permettant de comparer cette méthode d'estimation de $\mathcal{D}_{YX}$ avec celle de l'exercice 1. En lançant plusieurs fois le script, observez ce qui se passe lorsque la droite réelle est quasi-verticale.

Exercice 3 : estimation de $\mathcal{D}_\perp$ par le maximum de vraisemblance

Une droite $\mathcal{D}$ du plan peut également être définie par son *équation cartésienne normalisée* $x\cos\theta + y\sin\theta = \rho$, où (ρ,θ) sont les coordonnées polaires de la projection orthogonale sur $\mathcal{D}$ de l'origine O du repère. Si l'on note (x_Q, y_Q) les coordonnées cartésiennes de ce point, appelé Q, alors la distance à l'origine de Q vaut $\rho = \sqrt{x_Q^2 + y_Q^2} \in \mathbb{R}^+$ et l'angle polaire $\theta = \arctan\left(\frac{y_Q}{x_Q}\right) \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

Dans le cas où la droite $\mathcal{D}$ passe par l'origine O, l'angle polaire θ de $Q = O$ n'est pas défini. L'équation cartésienne normalisée de $\mathcal{D}$ s'écrit alors $x\cos\theta + y\sin\theta = 0$, où θ est l'angle polaire d'un des vecteurs orthogonaux à $\mathcal{D}$, défini à π près.

Il semble légitime de modéliser les résidus $r_{(\theta,\rho)}(P_i) = x_i\cos\theta + y_i\sin\theta - \rho$ par une loi normale centrée :

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

La *droite de régression en distance orthogonale* du nuage de points, notée $\mathcal{D}_\perp$, est la droite ayant pour équation $x\cos\theta^* + y\sin\theta^* = \rho^*$, où θ^* et ρ^* sont les valeurs des paramètres θ et ρ qui maximisent la log-vraisemblance :

$$(\sigma^*,\theta^*,\rho^*) = \underset{(\sigma,\theta,\rho)\in\mathbb{R}^+\times\left]-\frac{\pi}{2},\frac{\pi}{2}\right]\times\mathbb{R}^+}{\arg\max}\left\{\ln\prod_{i=1}^{n} f_{(\sigma,\theta,\rho)}(P_i)\right\} = \underset{(\sigma,\theta,\rho)\in\mathbb{R}^+\times\left]-\frac{\pi}{2},\frac{\pi}{2}\right]\times\mathbb{R}^+}{\arg\min}\sum_{i=1}^{n}\left\{\ln\sigma + \frac{r_{(\theta,\rho)}(P_i)^2}{2\sigma^2}\right\} \tag{8}$$

En supposant σ fixé, et sachant que la droite de régression $\mathcal{D}_\perp$ contient elle aussi le centre de gravité G, la résolution du problème (7) est en tout point analogue à celle du problème (2). Par analogie avec (4) :

$$\theta^* = \underset{\theta\in\left]-\frac{\pi}{2},\frac{\pi}{2}\right]}{\arg\min}\sum_{i=1}^{n}(x_i'\cos\theta + y_i'\sin\theta)^2 \tag{9}$$

Complétez la fonction `estimation_Dorth_MV`, appelée par le script `exercice_3`, permettant de résoudre le problème (8) correspondant au maximum de vraisemblance pour l'équation cartésienne normalisée.

Exercice 4 : estimation de $\mathcal{D}_\perp$ par les moindres carrés

Le critère $\mathcal{I}(\theta) = \sum_{i=1}^{n} (x'_i \cos\theta + y'_i \sin\theta)^2$ à minimiser dans (8) s'appelle l'*inertie*. Il s'écrit également :

$$\mathcal{I}(\theta) = \|CY\|^2 \quad \text{, où } C = \begin{bmatrix} x'_1 & \cdots & x'_n \\ y'_1 & \cdots & y'_n \end{bmatrix}^\top \text{ et } Y = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \tag{10}$$

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 $\|Y\| = 1$. Ce nouveau problème se résout en introduisant le *lagrangien* $\mathcal{L}(Y, \lambda) = \|CY\|^2 + \lambda\,(1 - \|Y\|^2)$, où λ constitue un *multiplicateur de Lagrange*. La condition d'optimalité de $\mathcal{L}$ s'écrit :

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

Sachant que $C^\top C$ est symétrique réelle, cette matrice admet une base orthonormée de vecteurs propres. De plus, comme $C^\top C$ est *semi-définie positive*, ses valeurs propres sont positives ou nulles. Le minimiseur de $\mathcal{I}(\theta)$ dont la norme est égale à 1, noté Y^*, est donc le vecteur propre associé à la plus petite valeur propre de $C^\top C$. Il s'agit du vecteur perpendiculaire à celui correspondant à l'axe qui est le long des points (par analogie au vecteur normal définissant un plan dans l'espace). En effet, pour un vecteur propre Y_p de norme 1, associé à la valeur propre λ_p, on a :

$$\|CY_p\|^2 = Y_p^\top C^\top C Y_p = \lambda_p Y_p^\top Y_p = \lambda_p. \tag{12}$$

Écrivez la fonction `estimation_Dorth_MC`, appelée par le script `exercice_4`, permettant de comparer cette méthode d'estimation de $\mathcal{D}_\perp$ à celle de l'exercice 3. La fonction `eig` permet de calculer les valeurs propres et vecteurs propres de la matrice $C^\top C$. La valeur de l'angle θ s'obtient quant à elle avec la fonction `atan`. Observez l'évolution des résultats en fonction de n et de n_{tests}, et aussi dans le cas où la droite réelle est quasi-verticale.