TP1 – Maximum de vraisemblance

La FIGURE 1 montre n observations indépendantes que l'on considère comme une réalisation $(x_1, \ldots, x_n)$ d'un n-uplet $(X_1, \ldots, X_n)$ de variables aléatoires « iid » (indépendantes et identiquement distribuées). La loi des n variables X_i est soit $f_{\theta_1}(x)$ soit $f_{\theta_2}(x)$, de paramètres respectifs θ_1 et θ_2, qui se déduisent l'une de l'autre par translation. Bien sûr, ces données sont plus probablement issues de la densité $f_{\theta_1}(x)$ que de la densité $f_{\theta_2}(x)$.

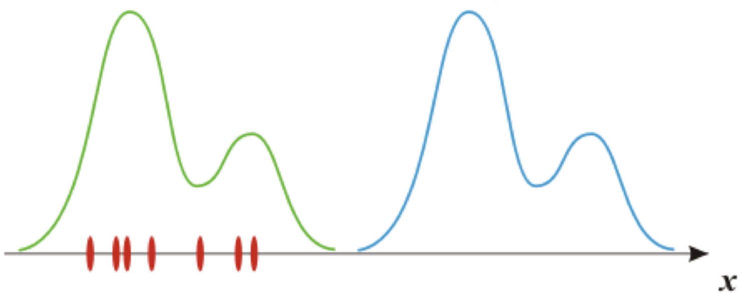

FIGURE 1 – Les n observations indépendantes (en rouge) d'un n-uplet de variables aléatoires correspondent plus probablement à la densité $f_{\theta_1}(x)$, en vert, qu'à la densité $f_{\theta_2}(x)$, en bleu, qui est une translatée de $f_{\theta_1}(x)$.

On peut formaliser cette intuition par la notion de *vraisemblance*, généralement notée L (pour *likelihood*). La vraisemblance $L_\theta(x_1, \ldots, x_n)$ est la loi du n-uplet $(X_1, \ldots, X_n)$, qui dépend de paramètres θ supposés connus :

$$L_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) \tag{1}$$

où f_θ est la densité de probabilité commune à toutes les variables indépendantes X_i (que l'on suppose continues).

Le but de ce TP est de montrer l'intérêt du *maximum de vraisemblance* pour l'estimation des paramètres. La loi qui semble le mieux « expliquer » les observations de la figure 1 est celle qui maximise leur vraisemblance $L_\theta(x_1, \ldots, x_n)$. On cherche ainsi la valeur θ^* de θ qui explique le mieux les observations $(x_1, \ldots, x_n)$.

Estimation des paramètres d'un cercle par maximum de vraisemblance

Lancez le script `donnees`, qui tire aléatoirement le centre C et le rayon R d'un cercle, ainsi que n points $P_i = (x_i, y_i)$ situés au voisinage de ce cercle. On souhaite estimer les paramètres (C^*, R^*) à partir des seuls P_i. En considérant l'écart $\epsilon(P_i) = d(P_i, C) - R$ entre le rayon R et la distance $d(P_i, C)$ du point P_i au centre C, il semble légitime de modéliser ces écarts par une *loi normale tronquée* d'écart-type σ :

$$f_{(C,R)}(P_i) = \begin{cases} K \exp\left\{-\frac{\epsilon(P_i)^2}{2\,\sigma^2}\right\} & \text{si } \epsilon(P_i) \geq -R \\ 0 & \text{sinon} \end{cases} \tag{2}$$

Comme les écarts $\epsilon(P_i)$ prennent leurs valeurs dans $[-R, +\infty[$ et non dans $\mathbb{R}$, le coefficient de normalisation K n'est pas exactement égal à $(\sigma\sqrt{2\pi})^{-1}$. Il est facile de vérifier que K dépend de R, mais pas de C.

Exercice 1 : estimation de la position du centre

Dans un premier temps, seul le centre C est estimé. Le rayon R est approché par la distance moyenne $\bar{R}$ des points P_i à leur centre de gravité G. Un produit étant plus difficile à maximiser qu'une somme, et la fonction logarithme étant strictement croissante, il est préférable de maximiser la *log-vraisemblance* $\ln L_{(C,\bar{R})}(P_1,\ldots,P_n)$:

$$C^* = \arg\max_{C\in\mathbb{R}^2}\left\{\ln\prod_{i=1}^n f_{(C,\bar{R})}(P_i)\right\} = \arg\min_{C\in\mathbb{R}^2}\sum_{i=1}^n\left[d(P_i,C)-\bar{R}\right]^2 \tag{3}$$

Dans un premier temps, complétez la fonction `G_et_R_moyen` pour déterminer le centre de gravité G des points P_i et le rayon moyen $\bar{R}$ comme la moyenne des distances $d(P_i,G)$. En suivant, et **sans boucle for**, complétez la fonction `estimation_C_uniforme`, appelée par le script `exercice_1`, censée résoudre le problème (3) par tirages aléatoires du centre C selon une loi uniforme (`rand`) dans un carré de centre G et de côté $2\bar{R}$.

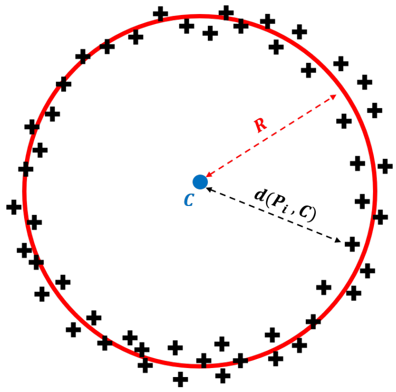

FIGURE 2 – Illustration de la distance $d(P_i,C)$ entre le centre C du cercle et un des n points P_i. Pour déterminer la position C^* du centre C, on minimise l'écart entre la distance $d(P_i,C)$ et le rayon R pour tous les points P_i.

Exercice 2 : estimation simultanée du centre et du rayon

On souhaite maintenant estimer simultanément C^* et R^*. L'estimation de R^* est un peu plus délicate, car le facteur de normalisation K de la loi (2) dépend de R. Au lieu de (3), on doit maintenant résoudre :

$$(C^*,R^*) = \arg\max_{(C,R)\,\in\,\mathbb{R}^2\times\mathbb{R}^+}\left\{\ln\prod_{i=1}^n f_{(C,R)}(P_i)\right\} = \arg\min_{(C,R)\,\in\,\mathbb{R}^2\times\mathbb{R}^+}\sum_{i=1}^n\left\{-\ln K+\frac{\left[d(P_i,C)-R\right]^2}{2\sigma^2}\right\} \tag{4}$$

Pour connaître la dépendance de K en R, écrivons la normalisation de la loi (2) en coordonnées polaires :

$$K\int_{\theta=0}^{2\pi}d\theta\int_{\rho=0}^{+\infty}\exp\left\{-\frac{(\rho-R)^2}{2\sigma^2}\right\}\rho\,d\rho = 1 \tag{5}$$

qui devient, avec le changement de variable $\tau=\rho-R$:

$$\int_{\tau=-R}^{+\infty}\exp\left\{-\frac{\tau^2}{2\sigma^2}\right\}\tau\,d\tau + R\int_{\tau=-R}^{+\infty}\exp\left\{-\frac{\tau^2}{2\sigma^2}\right\}d\tau = \frac{1}{K\,2\pi} \tag{6}$$

Dans (6), la première intégrale est facile à calculer, mais il n'existe pas d'expression analytique pour la seconde. En supposant $R\gg\sigma$, on peut néanmoins écrire l'approximation suivante (la borne rouge est inexacte) :

$$\sigma^2\exp\left\{-\frac{R^2}{2\sigma^2}\right\} + R\int_{\tau=-\infty}^{+\infty}\exp\left\{-\frac{\tau^2}{2\sigma^2}\right\}d\tau \approx \frac{1}{K\,2\pi} \tag{7}$$

Dans cette expression, on reconnaît l'intégrale de Gauss, donc :

$$\sigma^2\exp\left\{-\frac{R^2}{2\sigma^2}\right\} + R\,\sigma\sqrt{2\pi} \approx \frac{1}{K\,2\pi} \tag{8}$$

L'hypothèse $R \gg \sigma$ permet de négliger le premier terme du premier membre de (8), ce qui donne enfin :

$$K \approx \frac{1}{R\,\sigma\,(2\pi)^{3/2}} \tag{9}$$

La résolution du problème (4) revient donc à l'estimation approchée suivante :

$$(C^*, R^*) \approx \underset{(C,R)\in\mathbb{R}^2\times\mathbb{R}^+}{\arg\min} \sum_{i=1}^{n} \left\{ \ln R + \frac{[d(P_i,C)-R]^2}{2\sigma^2} \right\} \tag{10}$$

En utilisant à nouveau l'hypothèse $R \gg \sigma$, on voit que le premier terme de l'argument peut être négligé :

$$(C^*, R^*) \approx \underset{(C,R)\in\mathbb{R}^2\times\mathbb{R}^+}{\arg\min} \sum_{i=1}^{n} \left\{ [d(P_i,C)-R]^2 \right\} \tag{11}$$

Remarquez néanmoins qu'il aurait été impropre de déduire (11) de (3), puisque (11) est une approximation.

Complétez la fonction `estimation_C_et_R_uniforme`, appelée par le script `exercice_2`, censée résoudre le problème (11) par tirages aléatoires. Les moyennes des tirages aléatoires pour le centre C et pour le rayon R doivent être égales, respectivement, à G et à $\bar{R}$. Effectuez le même nombre n_{tests} de tirages pour C et pour R (ce dernier allant de $\bar{R}/2$ à $3\bar{R}/2$), puis testez chacun des couples (C_j, R_j), $j \in \{1, \ldots, n_{\text{tests}}\}$.
Remarque : Il n'est pas demandé de tester tous les R_j pour chaque C_j. Ici, pour un indice j donné, R_j et C_j sont testés ensemble une seule fois.

Exercice 3 : données partiellement occultées

On souhaite maintenant voir la robustesse de cette estimation si une partie des points P_i est manquante. Pour ce faire, complétez la fonction `occultation_donnees` appelée dans le script `donnees_occultees`. Cette fonction doit tirer aléatoirement deux angles θ_1 et θ_2 dans $[0, 2\pi[$, puis conserver seulement les points P_i d'angles polaires $\theta_i \in [\theta_1, \theta_2]$ si $\theta_1 \leq \theta_2$, et les points P_i d'angles polaires $\theta_i \in [0, \theta_2] \cup [\theta_1, 2\pi[$, si $\theta_1 > \theta_2$. Testez ensuite le script `exercice_3` pour visualiser l'estimation du cercle et du rayon avec seulement une partie des points P_i conservée.

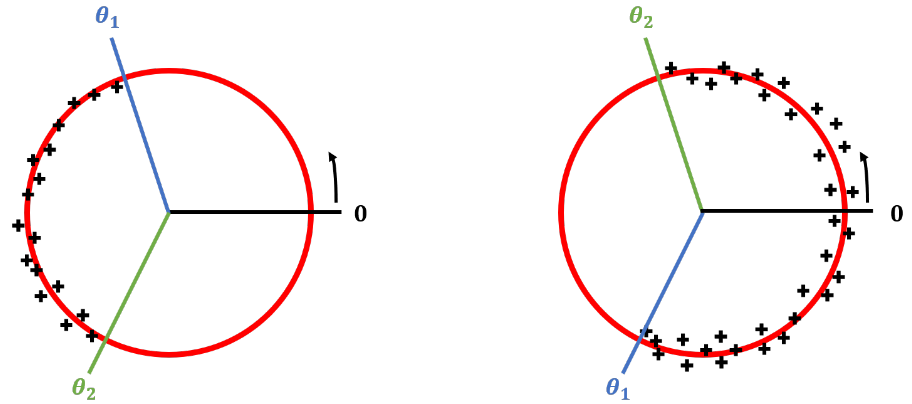

FIGURE 3 – Occultation des points P_i autour du cercle : cas où $\theta_1 \leq \theta_2$ à gauche, et $\theta_1 > \theta_2$ à droite.

Exercice 4 : modification des tirages aléatoires (facultatif)

Plutôt que des lois uniformes, il semble plus pertinent d'utiliser des lois normales pour les tirages aléatoires. Complétez la fonction `estimation_C_et_R_normale` dans laquelle vous traduirez cette idée en utilisant la fonction `randn` de Matlab (le `n` final indique qu'il s'agit d'une loi *normale*). Lancez ensuite le script `exercice_4` : vous devriez constater une amélioration dans les estimations par rapport à la loi uniforme, à nombre de tirages aléatoires équivalent.