TP2 – Maximum de vraisemblance

Notion de 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 $f_{\theta_2}(x)$.

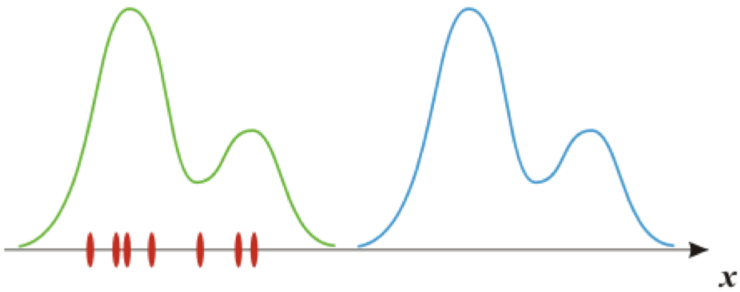

FIGURE 1 – Les n observations indépendantes (en rouge) d'un n-uplet de variables aléatoires correspondent plus probablement à la densité $f_{\theta_1}(x)$, en vert, qu'à la densité $f_{\theta_2}(x)$, en bleu, qui est une translatée de $f_{\theta_1}(x)$.

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_0 et le rayon R_0 d'un cercle $\mathscr{C}$, 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 un rayon donné R et la distance $d(P_i, C)$ du point P_i à un centre donné 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 avec le centre de gravité et le rayon moyen

Dans un premier temps, complétez la fonction `calcul_G_et_R_moyen` pour déterminer le centre de gravité $G = (x_G, y_G)^\top$ des points P_i et le rayon moyen $\bar{R}$ comme la moyenne des distances $d(P_i, G)$, comme illustré sur la Figure 2 ci-contre. Une fois fait, lancez le script `exercice_1` qui affiche le cercle de centre G et de rayon $\bar{R}$. Même si le tirage des points a été effectué de manière uniforme, on remarque cependant l'apparition d'un décalage entre le cercle initial $\mathscr{C}(C_0, R_0)$ et le cercle estimé $\mathscr{C}(G, \bar{R})$. C'est pourquoi dans la suite on va donner plus de latitude quant aux valeurs de C et R pour mieux estimer les paramètres du cercle.

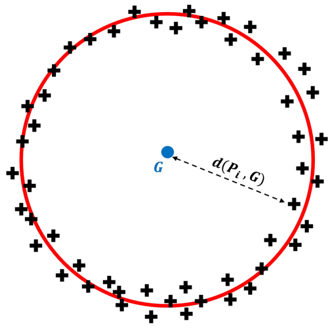

Figure 2 – Illustration de la distance $d(P_i, G)$ entre le centre de gravité G des n points P_i et un de ces points.

Exercice 2 : estimation de la position du centre par tirages aléatoires

Seul le centre C du cercle est estimé ici. 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* $\ln L_{(C,\bar{R})}(P_1, \ldots, P_n)$:

$$C^* = \underset{C \in \mathbb{R}^2}{\arg\max} \left\{ \ln \prod_{i=1}^{n} f_{(C,\bar{R})}(P_i) \right\} = \underset{C \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} \left\{ -\ln K + \frac{\left[d(P_i, C) - \bar{R}\right]^2}{2\sigma^2} \right\} \tag{3}$$

Comme K ne dépend pas de C, on obtient alors la minimisation suivante pour C :

$$C^* = \underset{C \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} \left[d(P_i, C) - \bar{R}\right]^2 \tag{4}$$

Dans un premier temps, complétez la fonction `tirages_aleatoires_uniformes` pour effectuer m tirages aléatoires de centres C suivant une loi uniforme (fonction `rand`) dont la moyenne est le centre de gravité G et la demi-étendue vaut $\bar{R}$ (i.e. les tirages sont effectués entre $G - \bar{R}$ et $G + \bar{R}$). En suivant, et **sans boucle for**, complétez la fonction `estimation_C` qui prend en entrée les n données bruitées (en colonnes), les m tirages de C (en lignes) ainsi que le rayon moyen $\bar{R}$, appelée par le script `exercice_2` pour résoudre le problème (4) pour le cercle $\mathscr{C}(C^*, \bar{R})$.

Exercice 3 : 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 :

$$(C^*, R^*) = \underset{(C,R)\,\in\,\mathbb{R}^2 \times \mathbb{R}^+}{\arg\max} \left\{ \ln \prod_{i=1}^{n} f_{(C,R)}(P_i) \right\} = \underset{(C,R)\,\in\,\mathbb{R}^2 \times \mathbb{R}^+}{\arg\min} \sum_{i=1}^{n} \left\{ -\ln K + \frac{\left[d(P_i, C) - R\right]^2}{2\sigma^2} \right\} \tag{5}$$

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

$$K \int_{\theta=0}^{2\pi} d\theta \int_{\rho=0}^{+\infty} \exp\left\{ -\frac{(\rho - R)^2}{2\sigma^2} \right\} \rho \, d\rho = 1 \tag{6}$$

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

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

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) :

$$\sigma^2 \exp\left\{-\frac{R^2}{2\sigma^2}\right\} + R \int_{\tau=-\infty}^{+\infty} \exp\left\{-\frac{\tau^2}{2\sigma^2}\right\} d\tau \approx \frac{1}{K\,2\pi} \tag{8}$$

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

$$\sigma^2 \exp\left\{-\frac{R^2}{2\sigma^2}\right\} + R\,\sigma\,\sqrt{2\pi} \approx \frac{1}{K\,2\pi} \tag{9}$$

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

$$K \approx \frac{1}{R\,\sigma\,(2\pi)^{3/2}} \tag{10}$$

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

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

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

$$(C^*, R^*) \approx \underset{(C,R)\in\mathbb{R}^2\times\mathbb{R}^+}{\arg\min} \sum_{i=1}^{n} [d(P_i,C)-R]^2 \tag{12}$$

Remarquez néanmoins qu'il aurait été impropre de déduire (12) de (4), puisque (12) est une approximation.

Dans un premier temps, dans la fonction `tirages_aleatoires_uniformes`, rajoutez les m tirages des rayons R suivant une loi uniforme, dont la moyenne est $\bar{R}$ et la demi-étendue vaut $\bar{R}/2$ (i.e. les tirages sont effectués entre $\bar{R}/2$ et $3\bar{R}/2$). Complétez ensuite la fonction `estimation_C_et_R` qui prend en entrée les n données bruitées ainsi que les m tirages de C et R, appelée par le script `exercice_3`, censée résoudre le problème (11) pour le cercle $\mathscr{C}(C^*, R^*)$.
Remarque : Ici, pour un indice $j \in [\![1, m]\!]$ donné, le rayon R_j et le centre C_j sont testés ensemble, l'indice du rayon devant rester le même que celui du centre (on ne teste pas R_j avec C_k si $j \neq k$).

Exercice 4 : estimation avec des données partiellement occultées

On souhaite tester 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`. À partir de deux angles θ_1 et θ_2 tirés aléatoirement dans $[0, 2\pi[$, cette fonction doit conserver seulement les points P_i d'angles polaires $\theta_i \in [\theta_1, \theta_2]$ si $\theta_1 \leq \theta_2$, ou les points P_i d'angles polaires $\theta_i \in [0, \theta_2] \cup [\theta_1, 2\pi[$ sinon. Testez ensuite le script `exercice_4` pour visualiser l'estimation du cercle et du rayon avec seulement les points P_i conservés.

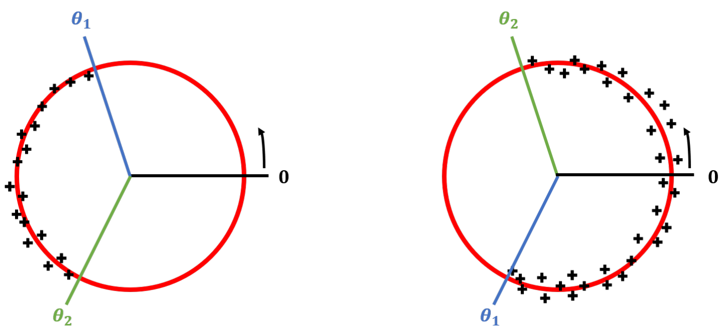

FIGURE 3 – Occultation des points P_i autour du cercle : cas où $\theta_1 \leq \theta_2$ à gauche, et $\theta_1 > \theta_2$ à droite.