TP1 – Maximum de vraisemblance

Notion 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 $f_{\theta_2}(x)$, 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 $f_{\theta_2}(x)$. Comment formaliser cette intuition ?

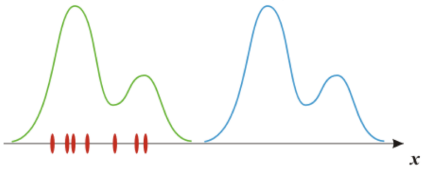

FIGURE 1 – Les n observations indépendantes (en rouge) d'un n-uplet de variables aléatoires correspondent plus probablement à la densité $f_{\theta_1}(x)$, en vert, qu'à la densité $f_{\theta_2}(x)$, en bleu, qui est une translatée de $f_{\theta_1}(x)$.

La réponse à cette question est donnée par la notion de *vraisemblance*, souvent 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).

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 trouve ainsi la valeur θ^* de θ qui explique le mieux les observations $(x_1, \ldots, x_n)$.

Exercice 1 : estimation du centre d'un cercle de rayon connu

Lancez le script `donnees.m`, qui affiche n points $P_i = (x_i, y_i)$ du plan se trouvant au voisinage d'un cercle de centre C_0 et de rayon R_0. Si l'on note $\epsilon(P_i) = d(P_i, C_0) - R_0$ l'écart entre la distance de P_i à C_0 et R_0, il semble légitime de modéliser ces écarts par une loi normale **tronquée** d'écart-type σ :

$$f_{(C_0,R_0)}(P_i) = \begin{cases} K \exp\left\{-\frac{\epsilon(P_i)^2}{2\,\sigma^2}\right\} & \text{si } \epsilon(P_i) \geq -R_0 \\ 0 \quad \text{sinon} \end{cases} \tag{2}$$

L'écart $\epsilon(P_i)$ prenant ses valeurs dans $[-R_0, +\infty[$ et non dans $\mathbb{R}$, le facteur de normalisation K n'est pas égal à $(\sigma\sqrt{2\pi})^{-1}$. Nous verrons par la suite que K est indépendant de C_0, mais pas de R_0. Dans un premier temps, nous supposons que le rayon R_0 est connu, et que seul le centre C est inconnu. Sa position peut être estimée en maximisant la vraisemblance. Comme un produit est plus difficile à maximiser qu'une somme, et que la fonction logarithme est strictement croissante, il est préférable de maximiser la *log-vraisemblance* $\ln L_{(C,R_0)}(P_1, \ldots, P_n)$:

$$(x_C^*, y_C^*) = \underset{(x_C,y_C)\in\mathbb{R}^2}{\arg\max} \left\{ \ln \left[\prod_{i=1}^{n} f_{(C,R_0)}(P_i) \right] \right\} = \underset{(x_C,y_C)\in\mathbb{R}^2}{\arg\min} \left\{ \sum_{i=1}^{n} \left[\sqrt{(x_i - x_C)^2 + (y_i - y_C)^2} - R_0 \right]^2 \right\} \tag{3}$$

Complétez le script `exercice_1.m`, qui est censé résoudre le problème (3) par tirages aléatoires selon deux lois uniformes (fonction `rand` de Matlab). Testez ce script en jouant sur les paramètres (n, σ, nombre de tirages).

Exercice 2 : estimation du centre et du rayon d'un cercle

L'estimation du rayon du cercle par maximum de vraisemblance est un peu plus délicate, car le facteur de normalisation K de la loi (2) dépend de R_0. La normalisation de cette loi s'écrit, en coordonnées polaires :

$$K\,2\pi \int_{\rho=0}^{+\infty} \exp\left\{-\frac{(\rho-R_0)^2}{2\sigma^2}\right\} \rho\, d\rho = 1 \tag{4}$$

qui devient, avec le changement de variable $\epsilon = \rho - R_0$:

$$\int_{\epsilon=-R_0}^{+\infty} \exp\left\{-\frac{\epsilon^2}{2\sigma^2}\right\} \epsilon\, d\epsilon + R_0 \int_{\epsilon=-R_0}^{+\infty} \exp\left\{-\frac{\epsilon^2}{2\sigma^2}\right\} d\epsilon = \frac{1}{K\,2\pi} \tag{5}$$

Dans (5), la première intégrale est facile à calculer, mais il n'existe pas d'expression analytique pour la seconde. En supposant $R_0 \gg \sigma$, on peut néanmoins écrire l'approximation suivante (la borne rouge est inexacte) :

$$\sigma^2 \exp\left\{-\frac{{R_0}^2}{2\sigma^2}\right\} + R_0 \int_{\epsilon=-\infty}^{+\infty} \exp\left\{-\frac{\epsilon^2}{2\sigma^2}\right\} d\epsilon \approx \frac{1}{K\,2\pi} \tag{6}$$

On reconnaît l'intégrale de Gauss, donc :

$$\sigma^2 \exp\left\{-\frac{{R_0}^2}{2\sigma^2}\right\} + R_0\,\sigma\,\sqrt{2\pi} \approx \frac{1}{K\,2\pi} \tag{7}$$

L'hypothèse $R_0 \gg \sigma$ permet de négliger le premier terme du premier membre de (7), ce qui donne enfin :

$$K \approx \frac{1}{R_0\,\sigma\,(2\pi)^{3/2}} \tag{8}$$

L'estimation simultanée du centre et du rayon s'écrit maintenant :

$$(x_C^*, y_C^*, R^*) = \underset{(x_C,y_C,R)\in\mathbb{R}^2\times\mathbb{R}^+}{\arg\max} \left\{ \ln\left[\prod_{i=1}^{n} K \exp\left\{-\frac{(d(P_i,C)-R)^2}{2\,\sigma^2}\right\}\right]\right\} \tag{9}$$

qui est donc effectivement plus délicate que le problème (3), puisqu'elle revient à l'estimation approchée :

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

En utilisant à nouveau l'hypothèse $R_0 \gg \sigma$, on voit que le premier terme de l'argument peut être négligé :

$$(x_C^*, y_C^*, R^*) \approx \underset{(x_C,y_C,R)\in\mathbb{R}^2\times\mathbb{R}^+}{\arg\min} \left\{ \sum_{i=1}^{n}\left[\sqrt{(x_i-x_C)^2+(y_i-y_C)^2} - R\right]^2\right\} \tag{11}$$

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

Effectuez une copie du script `exercice_1.m`, de nom `exercice_2.m`, que vous modifierez de manière à résoudre le problème (11) par tirages aléatoires selon trois lois uniformes. Testez ce script en jouant sur les paramètres (n, σ, nombre de tirages).

Exercice 3 : données partiellement occultées

Faites une copie du script `donnees.m`, de nom `donnees_occultees.m`, que vous modifierez de manière à produire les données au voisinage d'un arc de cercle d'angle variable, et non plus au voisinage d'un cercle entier. Vérifiez que les estimateurs précédents sont « robustes » à de telles données.

Exercice 4 : données aberrantes

Pour finir, faites une nouvelle copie du script `donnees.m`, de nom `donnees_aberrantes.m`, que vous modifierez de manière à introduire, parmi les données, une certaine proportion p de points tirés aléatoirement dans le pavé $[-\text{taille}, \text{taille}] \times [-\text{taille}, \text{taille}]$, qui constituent des *données aberrantes*. Observez jusqu'à quelle valeur de p les estimateurs précédents sont robustes à de telles données. Le remède à ce problème sera vu lors du TP3.