

TP1 – Maximum de vraisemblance

Notion de vraisemblance

La figure 1 montre n observations indépendantes que l'on considère comme une réalisation (x_1, \dots, x_n) d'un n -uplet (X_1, \dots, 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 ?

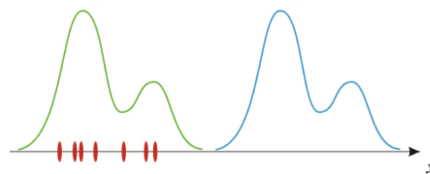


FIGURE 1 – Les n observations indépendantes (en rouge) d'un n -uplet de variables aléatoires correspondent plus probablement à la densité $f_{\theta_1}(x)$, en vert, qu'à la densité $f_{\theta_2}(x)$, en bleu, qui est une translatée de $f_{\theta_1}(x)$.

La réponse à cette question est donnée par la notion de *vraisemblance*, souvent notée L (pour *likelihood*). La vraisemblance $L_{\theta}(x_1, \dots, x_n)$ est la loi du n -uplet (X_1, \dots, X_n) , qui dépend de paramètres θ supposés connus :

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

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 $L_{\theta}(x_1, \dots, x_n)$. On trouve ainsi la valeur θ^* de θ qui explique le mieux les observations (x_1, \dots, x_n) .

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

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

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

Les écarts $\epsilon(P_i)$ prenant leurs valeurs dans $[-R_0, +\infty[$ 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_0 , mais pas de C_0 . Supposons dans un premier temps que le rayon réel R_0 soit connu, et que seule la position $C = (x_C, y_C)$ du centre soit inconnue. Cette 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* $\ln L_{(C, R_0)}(P_1, \dots, P_n)$:

$$(x_C^*, y_C^*) = \arg \max_{(x_C, y_C) \in \mathbb{R}^2} \ln \left\{ \prod_{i=1}^n f_{(C, R_0)}(P_i) \right\} = \arg \min_{(x_C, y_C) \in \mathbb{R}^2} \sum_{i=1}^n \left\{ [d(P_i, C) - R_0]^2 \right\} \quad (3)$$

Écrivez la fonction `estimation_1`, grâce à laquelle le script `exercice_1` est censé résoudre le problème (3) par tirages aléatoires selon deux lois uniformes (fonction `rand` de Matlab). Testez ce script en faisant varier les valeurs des paramètres (n , σ , nombre de tirages).

Exercice 2 : estimation simultanée du centre et du rayon

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 . Au lieu de (3), on doit maintenant résoudre :

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

Pour connaître la dépendance de K en R_0 , écrivons la normalisation de la loi (2) en coordonnées polaires :

$$K \int_{\theta=0}^{2\pi} d\theta \int_{\rho=0}^{+\infty} \exp \left\{ -\frac{(\rho - R_0)^2}{2\sigma^2} \right\} \rho d\rho = 1 \quad (5)$$

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

$$\int_{\tau=-R_0}^{+\infty} \exp \left\{ -\frac{\tau^2}{2\sigma^2} \right\} \tau d\tau + R_0 \int_{\tau=-R_0}^{+\infty} \exp \left\{ -\frac{\tau^2}{2\sigma^2} \right\} d\tau = \frac{1}{K 2\pi} \quad (6)$$

Dans (6), 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) :

$$\sigma^2 \exp \left\{ -\frac{R_0^2}{2\sigma^2} \right\} + R_0 \int_{\tau=-\infty}^{+\infty} \exp \left\{ -\frac{\tau^2}{2\sigma^2} \right\} d\tau \approx \frac{1}{K 2\pi} \quad (7)$$

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

$$\sigma^2 \exp \left\{ -\frac{R_0^2}{2\sigma^2} \right\} + R_0 \sigma \sqrt{2\pi} \approx \frac{1}{K 2\pi} \quad (8)$$

L'hypothèse $R_0 \gg \sigma$ permet de négliger le premier terme du premier membre de (8), ce qui donne enfin :

$$K \approx \frac{1}{R_0 \sigma (2\pi)^{3/2}} \quad (9)$$

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

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

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

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

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

Dupliquez `exercice_1` et `estimation_1`, sous les noms `exercice_2` et `estimation_2`, puis modifiez ces copies de manière à résoudre le problème (11) par tirages aléatoires selon trois lois uniformes distinctes.

Exercice 3 : données partiellement occultées

Faites une copie du script `exercice_2`, de nom `exercice_3`, où vous appellerez le script `donnees_occultees` au lieu du script `donnees`. Pour écrire `donnees_occultees`, faites une copie de `donnees`, que vous modifierez de manière à tirer aléatoirement deux angles θ_1 et θ_2 dans $[0, 2\pi[$, puis à 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 $\theta_i \in [0, \theta_2] \cup [\theta_1, 2\pi[$, dans le cas où $\theta_1 > \theta_2$.

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. Faites une copie du script `exercice_2`, de nom `exercice_4`, où vous appellerez le script `exercice_2` au lieu du script `donnees`, et où vous traduirez cette idée à l'aide de la fonction `randn` de Matlab (le `n` final indique qu'il s'agit d'une loi *normale*). Vous devriez constater une amélioration dans les estimations.