Estimation des paramètres d'une distribution Gamma

Préparation : avant la séance répondez aux questions 4, 5a, 5c, 5d

L'ESTIMATION de paramètres est une question cruciale pour le traitement de données expérimentales. A titre d'exemple, on s'intéresse ici à un « cas d'école », relativement simple concernant l'estimation des deux paramètres d'une distribution Gamma. L'objet de ce travail n'est pas une étude de cette distribution (moyenne, variance,... densité, fonction de répartition,... propriétés diverses,...) mais l'identification (l'estimation) de ses paramètres.

Fonction gamma – On rappelle la définition de la fonction Γ :

$$\Gamma(u) = \int_{\mathbb{R}_+} t^{u-1} \exp -t \, \mathrm{d}t \,, \quad \text{pour} \quad u \in \mathbb{R}_+ \,.$$

Cette fonction interviendra essentiellement au travers de son logarithme $\log \Gamma$ et la routine *Matlab* `gammaln` permet son calcul. Sa dérivée est la fonction di-Gamma notée ψ

$$\psi(u) = \frac{\mathrm{d}}{\mathrm{d}u} \log \Gamma(u) \,,$$

disponible par la routine *Matlab* `psi`. Cette même routine permet de calculer les dérivées de ψ. On fournit en plus la routine `InvPsi` qui permet d'inverser ψ. △

On s'intéresse à la densité Gamma paramétrée par $\alpha > 0$ et $\beta > 0$:

$$f_{X|A,B}(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma[\alpha]} \; x^{\alpha-1} \; \exp\left[-\beta x\right] \; \mathbb{1}_+(x) \,, \tag{1}$$

où $\mathbb{1}_+$ est l'indicatrice de $\mathbb{R}_+$. On rappelle que sa moyenne et sa variance valent respectivement : α/β et α/β^2. Lorsque $\alpha > 1$, son maximiseur est $(\alpha - 1)/\beta$.

On en observe N échantillons $x_1, x_2, \ldots, x_N$ indépendants et on les regroupe dans le vecteur $\boldsymbol{x} \in \mathbb{R}_+^N$. Pour alléger certaines écritures, on note

$$M_A(\boldsymbol{x}) = \frac{1}{N}\sum x_n \qquad \text{et} \quad V_A(\boldsymbol{x}) = \frac{1}{N}\sum (x_n - M_A(\boldsymbol{x}))^2$$

$$M_L(\boldsymbol{x}) = \frac{1}{N}\sum \log x_n \quad \text{et} \quad V_L(\boldsymbol{x}) = \frac{1}{N}\sum (\log x_n - M_L(\boldsymbol{x}))^2$$

les moyennes et variances empiriques des données et de leur logarithme.

1. *« **Théorie** » — En utilisant la routine `GammaPDF`, tracez la densité de probabilité pour diverses valeurs de (α, β). Attention, Matlab fonctionne à l'inverse pour le second paramètre : il faut lui passer $(\alpha, 1/\beta)$ pour lui causer de la densité définie par (1).*

2. *« **Simulations** » — Tirez les N réalisations de la variable en utilisant la routine* `GammaRND`*. Tracez-les sous la forme d'un histogramme que vous positionnerez en regard la densité.*

3. *Lien « **empirique – vrai** ».*

3a. *Sous Matlab, calculez les moments (moyenne et variance) empiriques et vérifiez qu'ils prennent la « bonne » valeur compte-tenu des vraies valeurs des paramètres α et β.*

3b. *A l'inverse, en approchant les deux moments vrais (fonction des vrais paramètres) par les deux moments empiriques (calculés à partir des données), proposez un estimateur empirique, simple et rapide, pour α et β. On parle d'estimation par une méthode de moments.*

4. *Écrivez la vraisemblance du couple (α, β) attachée aux données $\boldsymbol{x}$ ainsi que la log-vraisemblance que l'on notera $\mathrm{LV}_{\boldsymbol{x}}(\alpha, \beta)$.*

On se donne une densité *a priori* $\pi_{A,B}(\alpha, \beta)$ pour le couple (A, B) que l'on choisit uniforme sur le pavé $[0, \alpha_{\mathrm{M}}] \times [0, \beta_{\mathrm{M}}]$, où α_{M} et β_{M} sont des valeurs maximales prédéfinies pour les paramètres.

$$\pi_{A,B}(\alpha, \beta) = \begin{cases} (\alpha_{\mathrm{M}}\beta_{\mathrm{M}})^{-1} & \text{si } 0 < \alpha < \alpha_{\mathrm{M}} \text{ et } 0 < \beta < \beta_{\mathrm{M}} \\ 0 & \text{sinon} \end{cases}$$

On pourra prendre par exemple $\alpha_{\mathrm{M}} = 8$ et $\beta_{\mathrm{M}} = 5$. Concernant la vraie valeur des paramètres, commencera, par exemple, avec $\alpha^\star = 3$ et $\beta^\star = 2$.

5. *Travail sur la distribution A posteriori*

5a. *Déterminez la densité a posteriori $\pi_{A,B|\boldsymbol{X}}(\alpha, \beta|\boldsymbol{x})$ à un facteur multiplicatif près et surtout son logarithme, noté $\mathrm{LVP}_{\boldsymbol{x}}(\alpha, \beta)$, à un terme additif près.*

5b. *Tracez $\mathrm{LVP}_{\boldsymbol{x}}(\alpha, \beta)$ sur une grille de valeur du pavé $[0, \alpha_{\mathrm{M}}] \times [0, \beta_{\mathrm{M}}]$. La fonction* `meshgrid` *pourra être utile.*

5c. *Déduisez-en le logarithme de chacune des conditionnelles a posteriori $\pi_{A|\boldsymbol{X},B}(\alpha|\boldsymbol{x}, \beta)$ et $\pi_{B|\boldsymbol{X},A}(\beta|\boldsymbol{x}, \alpha)$ que l'on notera $\mathrm{LVP}_{\boldsymbol{x}}(\alpha|\beta)$ et $\mathrm{LVP}_{\boldsymbol{x}}(\beta|\alpha)$.*

5d. *Calculez le maximiseur de chaque conditionnelle a posteriori. Naturellement, le maximiseur par rapport à α sera fonction de β et inversement. On pourra faire intervenir l'inverse de la fonction ψ évoquée en introduction.*

On déduit de ce qui précède, un algorithme de calcul du maximiseur *a posteriori*. On montre qu'on peut maximiser la densité *a posteriori* par rapport à (α, β) en maximisant alternativement vis-à-vis de chacune des deux variables α et β, c'est-à-dire en maximisant à tour de rôle chaque conditionnelle $(\alpha|\boldsymbol{x}, \beta)$ et $(\beta|\boldsymbol{x}, \alpha)$. Une forme générique de l'algorithme d'optimisation est alors la suivante.

1. Initialisation : $\beta^{[1]} = 1$
2. Boucle pour $k = 2, \ldots, K$
   — Déterminez $\alpha^{[k]}$ comme maximiseur de $\mathrm{LVP}_{\boldsymbol{x}}(\alpha|\beta^{[k-1]})$
   — Déterminez $\beta^{[k]}$ comme maximiseur de $\mathrm{LVP}_{\boldsymbol{x}}(\beta|\alpha^{[k]})$
3. Fin

Dans le langage *Matlab*, le codage prend la forme suivante dont vous pouvez vous inspirer et qu'il vous appartient de compléter.

```
% Initialisation
BetaCourant = 1;

% Boucle d'optimisation variable par variable
while Delta ...

    % Mises A Jour Alpha
    AlphaCourant = ...
    GardeAlpha = [GardeAlpha AlphaCourant];
    GardeBeta = [GardeBeta BetaCourant];

    % Mises A Jour Beta
    BetaCourant = ...
    GardeAlpha = [GardeAlpha AlphaCourant];
    GardeBeta = [GardeBeta BetaCourant];

    % Variations
    Delta = ...

end
% Fin de boucle d'optimisation
```

6a. *Représentez les deux ensembles d'itérées produites $\alpha^{[k]}$ et $\beta^{[k]}$ pour $k = 1, \ldots, K$ en fonction de k sur deux sous figures d'une même figure.*

6b. *Représentez également, dans le plan, les couples d'itérées $(\alpha^{[k]}, \beta^{[k]})$ superposés à la représentation de la densité a posteriori de la question 5a.*

7. *Récupérez le fichier fourni `Mystere.mat` : il contient un jeu de données résultant de valeurs de paramètres α et β inconnues de vous. Déterminez les valeurs de ces paramètres.*