Examen 2009-2010

Centrale - Statistiques avancées (première partie) - deuxième session

6 mai 2010

A- Régression : une nouvelle observation diminue l'incertitude

On considère un modèle de régression linéaire simple :

$$Y_i = a + bx_i + \epsilon_i, \quad i \in \{1, \dots, n\},$$

où les variables ϵ_i sont indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$. On note

$$\bar{x}_n = \frac{x_1 + \cdots + x_n}{n} \quad et \quad S_n^2 = x_1^2 + \cdots + x_n^2.$$

1. *Rappeler les estimateurs du maximum de vraisemblance pour $\hat{a}_n$ et $\hat{b}_n$ (on ne demande pas de justifier les formules).*

$$\hat{b}_n = \frac{x_1 Y_1 + \cdots + x_n Y_n}{S_n^2 - n\bar{x}_n^2}, \quad \bar{a}_n = \frac{Y_1 + \cdots + Y_n}{n} - \bar{b}_n \bar{x}_n.$$

2. *Quelle est la loi de $\hat{b}_n$?*

$$\hat{b}_n \sim \mathcal{N}\left(b, \frac{\sigma^2}{S_n^2 - n\bar{x}_n^2}\right).$$

3. *Quel est le risque quadratique de $\hat{a}_n$ pour l'estimation de a ?*
   Comme $\hat{a}_n \sim \mathcal{N}\left(a, \frac{\sigma^2 S_n^2}{nS_n^2 - n^2\bar{x}_n^2}\right)$ est sans biais, son risque quadratique est égal à sa variance.

On ajoute une nouvelle observation (x_{n+1}, y_{n+1}). On définit donc

$$\bar{x}_{n+1} = \frac{x_1 + \cdots + x_n + x_{n+1}}{n+1}, \quad S_{n+1}^2 = x_1^2 + \cdots + x_n^2 + x_{n+1}^2$$

et on note $\hat{a}_{n+1}$ et $\hat{b}_{n+1}$ les estimateurs du maximum de vraisemblance respectifs de a et b avec l'échantillon $(x_1, Y_1), \dots, (x_{n+1}, Y_{n+1})$.

4. *Montrer que $\bar{x}_{n+1} = (n\bar{x}_n + x_{n+1})/(n+1)$.*
   évident.

5. *Montrer que le risque quadratique de $\hat{b}_{n+1}$ est toujours plus faible que celui de $\hat{b}_n$.*
   Quitte à reparamétrer par $r_i = x_i - \bar{x_n}$, ce qui ne change pas $\hat{b}_n$ et $\hat{b}_{n+1}$, on peut supposer que $\bar{x_n} = 0$. Il faut montrer que $\hat{b}_n$ a une variance plus grande que $\hat{b}_{n+1}$, c'est-à-dire que

$$\frac{1}{S_n^2} > \frac{1}{S_{n+1}^2 (n+1)\bar{x}_{n+1}^2}.$$

   Cela résulte aisément du fait que $\bar{x}_{n+1} = x_{n+1}/(n+1)$:

$$S_{n+1}^2 - (n+1)\bar{x}_{n+1}^2 = S_n^2 + x_{n+1}^2 - \frac{x_{n+1}^2}{n+1} > S_n^2.$$

6. *Cela est vrai-il aussi pour l'estimation de l'ordonnée à l'origine : le risque quadratique de $\hat{a}_{n+1}$ est-il toujours plus faible que celui de $\hat{a}_n$?*
   Au terme d'un calcul un peu plus long, on montre que c'est le cas.

B- Points de vue fréquentiste et bayésien

Pierre choisit un nombre entier strictement positif θ. Ensuite, il tire une suite de nombres aléatoires $X_1, \dots X_n$ indépendants de loi uniforme sur l'ensemble $\{1, \dots, \theta\}$, qu'il transmet à Jean. Jean cherche à estimer θ à partir des nombres fournis par Pierre.

7. *Pour Jean, quel est l'estimateur du maximum de vraisemblance $\hat{\theta}_{ML}$?*
   Soit $M_n = \max\{X_1, \dots, X_n\}$. La vraisemblance est $\ell(\theta) = 1/\theta^n \mathbb{1}_{\{M_n \leq \theta\}}$, elle est donc maximale en $\hat{\theta}_{ML} = M_n$.
8. *Montrer que $\mathbb{P}\left(\hat{\theta}_{ML} \neq \theta\right) \leq \exp(-n/\theta)$.*

$$\mathbb{P}(\hat{\theta}_{ML} \neq \theta) = \mathbb{P}(M_n < \theta) = \mathbb{P}\left(X_1 \leq \theta - 1, \dots, X_n \leq \theta - 1\right) = \left(1 - \frac{1}{\theta}\right)^n \leq \exp\left(-\frac{n}{\theta}\right),$$

   puisqu'il est bien connu que $\forall x > -1, \ln(1+x) \leq x$.
9. *Montrer que $\hat{\theta}_{ML}$ est consistant.*
   $\mathbb{P}(\hat{\theta}_{ML} \neq \theta) \leq \exp\left(-n/\theta\right)$ tend vers 0 quand n tend vers l'infini.

Jean apprend que Pierre a choisi θ au hasard : pour tout entier naturel non nul k la probabilité qu'il avait de choisir $\theta = k$ était de

$$\mathbb{P}(\theta = k) = \frac{c_a}{k^a}, \quad \textit{avec} \quad c_a = \left(\sum_{k=1}^{\infty} \frac{1}{k^a}\right)^{-1}.$$

10. *Déterminer la loi a posteriori, c'est-à-dire*

$$\Pi_n(\{k\}) = \mathbb{P}(\theta = k | X_1, \dots, X_n) = \sum_{x_1, \dots, x_n \in \mathbb{N}^*} \mathbb{P}(\theta = k | X_1 = x_1, \dots, X_n = x_n) \mathbb{1}_{\{X_1 = x_1, \dots, X_n = x_n\}}$$

    pour tout $k \geq 1$. La densité jointe pour $(\theta, X_1, \dots, X_n)$ au point $(k, x_1, \dots, x_n)$ est :

$$\frac{c_a}{k^a} \times \frac{1}{k^n} \mathbb{1}_{\{\max\{x_1, \dots, x_n\} \leq k\}},$$

    la loi a posteriori est donc

$$\Pi(\{k\}) = \frac{c'_{a, M_n}}{k^{a+n}} \mathbb{1}_{\{k \geq M_n\}},$$

    où c'_{a, M_n} est une constante de normalisation ne dépendant pas de k.
11. *Que proposez-vous comme estimateur bayésien $\hat{\theta}_B$?*
    Par exemple, le mode de la loi a posteriori : c'est $\hat{\theta}_{ML}$.
12. *Montrer que $\Pi_n(\{\theta\}) \to 1$ en probabilité quand n tend vers l'infini.*
    On peut d'abord montrer que $\Pi_n(\{1, \dots, \theta - 1\}) \to 0$ rapidement, puis on remarque que sur l'évènement $M_n = \theta$ on a

$$\Pi_n(\{\theta\}) = \frac{\theta^{-n-a}}{\theta^{-n-a} + S}, \quad \text{avec} \quad S = \sum_{j=1}^{\infty} (\theta + j)^{-n-a}.$$

    Or

$$\theta^{n+a} S = \sum_{j=1}^{\infty} \frac{1}{\left(1 + \frac{j}{\theta}\right)^{n+a}}$$

    tend vers 0 par les théorèmes classiques d'inversion de somme et de limite, ce qui donne le résultat.