Examen 2009-2010

Centrale - Statistiques avancées (première partie)

18 février 2010

Un hopital a relevé chaque jour $i \in \{1, \dots, n\}$ le nombre y_i d'hospitalisations pour bronchiolite dans son service pédiatrique. Il souhaite éventuellement montrer un lien entre les cas de bronchiolite et la pollution atmosphérique : aussi s'est-il procuré la valeur x_i de la concentration d'azote dans l'atmosphère à midi chaque jour $i \in \{1, \dots, n\}$.

A- Estimation fréquentiste de la moyenne

On suppose dans un premier temps que les valeurs $y_1, \dots, y_n$ sont les réalisations de n variables aléatoires $Y_1, \dots, Y_n$ indépendantes identiquement distribuées selon une loi de Poisson de paramètre λ : pour tout $k \in \mathbb{N}$,

$$\mathbb{P}(Y_1 = k) = \exp(-\lambda)\frac{\lambda^k}{k!} \ .$$

1. *Calculer l'estimateur du maximum de vraisemblance de $\hat{\lambda}_{MV}$ pour λ.*
   En notant $S_n = Y_1 + \cdots + Y_n$, la log-vraisemblance s'écrit

   $$l(\theta) = \log \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{Y_i}}{Y_i!} = S_n \log(\lambda) - n\lambda + c(Y_1, \dots, Y_n),$$

   on vérifie aisément qu'elle est maximale pour $\lambda = \hat{\lambda}_{MV} = S_n/n$.
2. *Montrer que $\hat{\lambda}_{MV}$ est sans biais et consistant.*
   Comme moyenne empirique des observations, $\hat{\lambda}_{MV}$ a pour espérance et pour limite presque-sûre $\mathbb{E}[Y_1] = \lambda$.
3. *Proposer un intervalle de confiance asymptotique pour λ au risque $\alpha = 5\%$.*
   Le TCL donne : $(\hat{\lambda}_{ML} - \lambda)\sqrt{n/\lambda} \to \mathcal{N}(0, 1)$. On peut
   – soit en déduire directement un intervalle de confiance asymptotique en résolvant une équation de degré 2 ;
   – soit utiliser la méthode delta, en s'intéressant à $\sqrt{\hat{\lambda}_{MV}}$.

B- Estimation Bayésienne de la moyenne

On supposera dans cette partie que λ est une variable aléatoire de loi Gamma(a, b), dont la densité de probabilités sur $\mathbb{R}_+^*$ est donnée pour $a, b > 0$ par

$$\pi(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda).$$

On supposera en outre que les variables aléatoires $Y_1, \dots, Y_n$ sont, conditionnellement à λ, indépendantes identiquement distribuées selon une loi de Poisson de paramètre λ.

4. *(question bonus) Pourquoi a-t-on choisi la loi a priori dans la famille des lois Gamma ?*
   Les lois de Poisson forme une famille exponentielle, et les lois Gammas leur sont naturellement conjuguées

5. *Calculer l'espérance et la variance de la loi Gamma(a, b).*
   Avec deux calculs d'intégrales, on trouve une espérance de a/b et une variance de a/b^2.
6. *A priori, on s'attend à ce que λ soit plutôt situé entre 20 et 30. Quelles valeurs de a et b peut-on choisir ?*
   Si on veut une espérance de 25 et une variance de $5^2 = 25$, on peut prendre $a = 25$ et $b = 1$.
7. *Donner l'expression de la vraisemblance jointe de $\lambda, Y_1, \dots, Y_n$.*
   Cette vraisemblance jointe est le produit de la densité a priori de λ par la vraisemblance des observations sour λ, c'est-à-dire :

$$L(\lambda, Y_1, Y_n) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda)\prod_{i=1}^{n}\frac{\exp(-\lambda)\lambda^{Y_i}}{Y_i!} = \frac{b^a}{\Gamma(a)\prod_{i=1}^{n}Y_i!}\lambda^{a+S_n-1}\exp(-(b+n)\lambda).$$

8. *En déduire que la loi a posteriori de λ conditionnellement aux observations $Y_1, \dots, Y_n$ est la loi $\Gamma(a + Y_1 + \cdots + Y_n, b + n)$.*
   La loi a posteriori de λ a pour densité $L(\lambda, Y_1, \dots, Y_n)/\int_\mu L(\mu, Y_1, \dots, Y_n)d\mu$: comme c'est une densité de probabilités, on voit sans calcul que c'est celle de la loi Gamma$(a + S_n, b + n)$.
9. *Proposer un estimateur a posteriori $\hat{\lambda}_\pi$ de λ (par exemple, l'espérance de la loi a posteriori). Déterminer son biais, puis montrer qu'il est consistant.*
   L'espérance a posteriori vaut donc $\hat{\lambda}_\pi = (a+S_n)/b+n)$, ce qui n'est pas très différent $\hat{\lambda}_{ML} = S_n/n$, et la consistance du premier se déduit aisément de celle du deuxième.

C- Approximation gaussienne

On se donne dans cette partie une suite $N_1, N_2, \dots$ de variables indépendantes identiquement distribuées suivant une loi de Poisson de paramètre 1. On notera de plus pour tout entier m strictement positif $S_m = N_1 + \cdots + N_m$.

10. *Soit G_1 la fonction définie par $G_1(s) = \mathbb{E}[s^{N_1}]$. Montrer que G_1 est définie sur $\mathbb{R}$, et que pour tout $s \in \mathbb{R}$,*

$$G_1(s) = \exp(s - 1).$$

    Calcul immédiat :

$$G_1(s) = \sum_{k=0}^{\infty} s^k \frac{\exp(-1)1^k}{k!} = \exp(-1)\exp(s),$$

    pas de problème de sommabilité pous $s \in \mathbb{R}$.
11. *Calculer $G_m(s) = \mathbb{E}[s^{S_m}]$.*
    $G_m(s) = \mathbb{E}[s^{Y_1} \dots s^{Y_n}] = G_1(s)^n = \exp(n(s - 1))$
12. *Trouver une relation simple entre $\mathbb{P}(S_m = k)$ et $G_m^{(k)}(0)$ (la dérivée k-ième de G_m prise au point $s = 0$). On pourra commencer par regarder la valeur de $G_m(0)$, puis de $G'_m(0)$.*
    Il suffit de justifier qu'on peut dériver $\sum_{k=0}^{\infty}\mathbb{P}(S_m = k)s^k$ sous le signe somme, puis en $s = 0$ il ne reste que le terme constant.
13. *Montrer que S_m suit une loi de Poisson de paramètre m.*
    En appliquant la question précédente, on trouve $\mathbb{P}(S_m = k) = \exp(-m)m^k/k!$.
14. *En déduire que, quand m tend vers l'inifini,*

$$\mathbb{P}(S_m \le m + x\sqrt{m}) \to \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx.$$

    D'après le TCL, $(S_m - m)/\sqrt{m}$ converge en loi vers une gaussienne centrée réduite, d'où le résultat.

D- Régression linéaire

On cherche une relation simple entre le nombre d'hospitalisation pour bronchiolite et la concentration d'azote. On utilisera dans cette partie le modèle suivant :

$$Y_i = \mu + \alpha x_i + \sigma\epsilon_i,$$

et on fera comme si ϵ_i avait une distribution normale centrée réduite.

15. *(question bonus) Pourquoi les résultats précédents nous autorisent-ils à considérer un tel modèle de régression linéaire gaussienne, alors que nos données sont discrètes, si les valeurs y_i observées sont suffisamment grandes ?*

    Ces résultats montre que, vue de loin, la loi de Poisson est proche de la loi normale quand le paramètre est assez grand (cf. approximation binomiale-normale). Toutefois, en se plaçant dans le modèle gaussien standard on pert le lien qui existe entre l'espérance $\mu + \alpha x_i$ et la variance σ^2, supposée ici indépendante de x partout. Ca n'est qu'un modèle...

16. *Construire les estimateurs du maximum de vraisemblance de μ, α et σ.*

    C'est le modèle de régression linéaire simple habituel, cf cours.

17. *Proposer un intervalle de confiance de risque $\alpha = 5\%$ pour α (on indiquera quel quantile de quelle loi utiliser, sans chercher à préciser sa valeur).*

    On utilise le fait que $T = (\hat{\alpha} - \alpha)\sqrt{n\mathrm{Var}_n(x)/\sigma^2}$ suit une loi de Student à $n-2$ degrés de liberté.

18. *Comment tester si la pollution en azote a une influence sur le nombre d'hospitalisations ?*

    Il suffit de tester $\alpha = 0$ contre $\alpha \neq 0$ (ou contre $\alpha > 0$), à l'aide de T.

E- Régression Poissonnienne

Pour le cas où les nombres relevés ne sont pas très grands, l'hypothèse de normalité des ϵ_i n'est pas raisonnable. On cherchera plutôt à modéliser le nombre d'hospitalisations Y_i par une variable de Poisson dont le paramètre λ_i s'écrit :

$$\lambda_i = \exp(\eta + \gamma x_i)$$

pour une certaine valeur inconnue du couple de paramètres $\theta = (\eta, \gamma)$. On cherche à estimer θ.

19. *Pour tout $\theta \in \mathbb{R}^2$, calculer la log-vraisemblance des observations sous θ :*

$$l(\theta) = \log \prod_{i=1}^{n} p_{\lambda_i}(Y_i),$$

    où $p_\lambda(y)$ désigne la vraisemblance d'une observation y sous la loi de Poisson de paramètre λ.

$$l(\theta) = \log \prod_{i=1}^{n} p_{\lambda_i}(Y_i) = \sum_{i=1}^{n} -\exp(\eta + \gamma x_i) + Y_i(\eta + \gamma x_i).$$

20. *Montrer que la fonction l admet un unique maximum $\hat{\theta} = (\hat{\eta}, \hat{\gamma})$ sur $\mathbb{R}^2$, qui vérifie les équations :*

$$\sum_{i=1}^{n} \exp(\hat{\eta} + \hat{\gamma} x_i) = \sum_{i=1}^{n} Y_i$$
$$\sum_{i=1}^{n} \exp(\hat{\eta} + \hat{\gamma} x_i) x_i = \sum_{i=1}^{n} x_i Y_i$$

    La fonction l étant parfaitement régulière sur $\mathbb{R}^2$, un extremum annule son gradient. Le calcul de la matrice Hessienne montre que l'unique extremum est un maximum : comme celle-ci est définie négative, la fonction l est concave.

21. *(question bonus) Quelle méthode peut-on utiliser pour construire une approximation numérique de $\hat{\theta}$?*

    Pour ce problème de maximisation concave régulier on peut naturellement penser à une méthode de Newton-Raphson (on dispose de la matrice hessienne).

F- Modèle de mélange

En fait, une étude étrangère suggère que les hospitalisations sont beaucoup plus nombreuses les jours où l'usine d'incinération voisine fonctionne que les jours où elle est au repos, bien que celle-ci n'émette quasiment pas d'azote ; il semble donc plus réaliste de supposer que, quand le jour i est un jour d'incinération, le nombre d'hospitalisations suit une loi de Poisson de paramètre λ_1, alors que les autres jours il suit une loi de Poisson de paramètre λ_0. Le problème, c'est que l'hopital n'a pas accès au relevé des jours d'activité de l'usine.

On notera Z_i la variable aléatoire égale à 1 si le jour i est un jour d'incinération, et égale à 0 sinon. Les paramètres du problèmes sont la probabilité p que l'usine fonctionne un jour donné, et les intensités λ_0 et λ_1. On notera $\theta = (p, \lambda_0, \lambda_1) \in \Theta =]0,1[\times\mathbb{R}_+^* \times \mathbb{R}_+^*$, et P_θ la loi de probabilité correspondante. On supposera en outre que, sous P_θ, les variables Z_i sont indépendantes identiquement distribuées suivant une loi de Bernoulli de paramètre p.

22. *(question bonus) Dire à quel modèle de Markov caché très particulier la chaîne (Z_i, Y_i) correspond. Noter qu'il ne sera pas nécessaire d'utiliser ce fait pour les questions suivantes.*

    Les états cachés Z_i sont supposés i.i.d. : la chaîne de Markov correspondante a donc des lois de transitions qui sont toutes égales.

23. *Supposons, dans cette question seulement, que les valeurs de λ_0, λ_1 et p sont connues. Pour tout $i \in \{1, \dots, n\}$, calculer $P_\theta(Z_i = 1 | Y_i = y_i)$, puis pour tout $x \in \{0,1\}^n$ la quantité*

$$P_\theta\left(\bigcap_{i=1}^n \{Z_i = x_i\} | Y_1 = y_1, \dots, Y_n = y_n\right)$$

    en fonction de y_i, λ_0, λ_1 et p.

$$P_\theta(Z_i = 1, Y_i = y_i) = P_\theta(Z_i = 1) P_\theta(Y_i | Z_i = 1) = p\frac{\exp(-\lambda_1)\lambda_1^{y_i}}{y_i!} \ , \quad \text{donc}$$

$$P_\theta(Z_i = 1 | Y_i = y_i) = \frac{p\exp(-\lambda_1)\lambda_1^{y_i}}{(1-p)\exp(-\lambda_0)\lambda_0^{y_i} + p\exp(-\lambda_1)\lambda_1^{y_i}}, \quad \text{et}$$

$$P_\theta\left(\bigcap_{i=1}^n \{Z_i = x_i\} | Y_1 = y_1, \dots, Y_n = y_n\right) = \prod_{i=1}^n P_\theta(Z_i = 1 | Y_i = y_i) \ .$$

Pour estimer p, λ_0 et λ_1, on propose la procédure suivante :

Algorithm 1 - EM

1: choisir initialement $\theta^o = (p^o, \lambda_0^o, \lambda_1^o)$, avec

$$p^o = 1/2, \quad \lambda_0^o = \min\{y_i : 1 \le i \le n\}, \quad \lambda_1^o = \max\{y_i : 1 \le i \le n\}$$

2: **for** k = 1 .. 1000 **do**
3: pour $\theta \in \Theta$, définir $f^o(\theta) = \sum_{i=1}^n \sum_{x\in\{0,1\}} P_{\theta^o}(Z_i = x | Y_i = y_i) \log P_\theta(Z_i = x, Y_i = y_i)$
4: trouver θ^n tel que $f^o(\theta^n) = \max_{\theta\in\Theta} f^o(\theta)$
5: $\theta^o \leftarrow \theta^n$
6: **end for**
7: renvoyer θ^n comme estimateur de θ

Bien noter que dans θ^o et θ^n, les lettres o et n signifient "old" et "new" (ce sont sont pas des exposants !).

24. *Montrer que la fonction f^o (ligne 4) admet un unique maximum $\theta^n \in \Theta$, et donner sa valeur en fonction de n et des $P_{\theta^o}(Z_i = x | Y_i = y_i), x \in \{0,1\}, 1 \le i \le n$.*

    On peut écrire

$$\log P_\theta(Z_i = x, Y_i = y_i) = x \log p + (1-x)\log(1-p) - \lambda_x + y_i \log(\lambda_x) \log(y_i!) \ ,$$

et donc en notant $p_i(x) = P_{\theta^o}(Z_i = x | Y_i = y_i)$ on a :

$$\begin{aligned} f^o(\theta) &= \sum_{i=1}^{n} \sum_{x\in\{0,1\}} p_i(x)\Big(x\log p + (1-x)\log(1-p) - \lambda_x + y_i \log(\lambda_x) \log(y_i!)\Big) \\ &= N_0 \log(1-p) + N_1 \log p \\ &\quad - N_0\lambda_0 + \left(\sum_{i=1}^{n} p_i(0) y_i\right) \log\lambda_0 \\ &\quad - N_1\lambda_1 + \left(\sum_{i=1}^{n} p_i(1) y_i\right) \log\lambda_1 \\ &\quad - \sum_{i=1}^{n} \log(y_i!) \end{aligned}$$

si on note $N_0 = \sum_{i=1}^n p_i(0)$ et $N_1 = \sum_{i=1}^n p_i(1)$; on observera que $N_0 + N_1 = n$. pour maximiser f^o en fonction de θ, il suffit donc de maximiser séparément en p, en λ_0 et en λ_1. On trouve

$$p^n = \frac{N_1}{n}, \quad \lambda_0^n = \frac{\sum_{i=1}^n p_0(x) y_i}{N_0} \quad \text{et } \lambda_1^n = \frac{\sum_{i=1}^n p_1(x) y_i}{N_1} \ .$$

25. *Montrer qu'à chaque itération de l'algorithme on a*

$$P_{\theta^n}(Y_1 = y_1, \ldots, Y_n = y_n) \geq P_{\theta^o}(Y_1 = y_1, \ldots, Y_n = y_n).$$

C'est exactement la même preuve qu'en cours pour les HMM.

26. *En déduire un critère d'arrêt moins arbitraire pour la ligne 2.*
On peut choisir un critère du genre $\|\theta^n - \theta^o\| < \epsilon$, ou encore $l(\theta^n) - l(\theta^o) \leq \epsilon'$.

27. *La valeur renvoyée par l'algorithme est-elle une approximation de l'estimateur du maximum de vraisemblance ?*
L'algorithme converge vers un maximum local de la vraisemblance, mais rien ne garantit qu'il n'y en ait pas plusieurs.

28. *(question bonus) Pour cette dernière question seulement, on ne suppose plus que les variables Z_i sont indépendantes. On suppose que :*
– *si elle fonctionne le jour i, elle a une probabilité q de fonctionner le jour $i+1$;*
– *si elle est au repos le jour i elle a une probabilité $2q$ d'être encore au repos le lendemain.*
Proposer une procédure permettant d'estimer q, λ_0 et λ_1, en justifiant son fonctionnement.
Dans ce cas on a vraiment une HMM, il faut faire comme dans le cours...