Chaînes de Markov en temps continu et génétique des populations.

Correction de la feuille 3 — 2014–2015

Nicolas Champagnat

Exercice 8

1. On a

$$\begin{aligned}\mathbb{P}_\theta(K_n=k) &= \mathbb{P}_\theta(K_n=k \mid \text{coal})\mathbb{P}(\text{coal}) + \mathbb{P}_\theta(K_n=k \mid \text{mut})\mathbb{P}(\text{mut})\\ &= \frac{n-1}{\theta+n-1}\mathbb{P}_\theta(K_{n-1}=k) + \frac{\theta}{\theta+n-1}\mathbb{P}_\theta(K_{n-1}=k-1).\end{aligned}$$

On va démontrer par récurrence sur n que $K_n = A_1 + \ldots + A_n$ en loi. On a $\mathbb{P}_\theta(K_1=1)=1=\mathbb{P}(A_1=1)$. Supposons que $\mathbb{P}_\theta(K_{n-1}=k)=\mathbb{P}_\theta(A_1+\ldots+A_{n-1}=k)$. Alors

$$\begin{aligned}&\mathbb{P}(A_1+\ldots+A_n=k)\\ &\quad= \mathbb{P}(A_1+\ldots+A_{n-1}=k)\mathbb{P}(A_n=0) + \mathbb{P}(A_1+\ldots+A_{n-1}=k-1)\mathbb{P}(A_n=1)\\ &\quad= \frac{n-1}{\theta+n-1}\mathbb{P}_\theta(K_{n-1}=k) + \frac{\theta}{\theta+n-1}\mathbb{P}_\theta(K_{n-1}=k-1) = \mathbb{P}_\theta(K_n=k).\end{aligned}$$

2. On trouve $\mathbb{E}_\theta(K_n) = \sum_{i=1}^n \frac{\theta}{\theta+i-1} \sim \theta\ln n$ et $\mathrm{Var}_\theta(K_n) = \sum_{i=1}^n \frac{\theta(i-1)}{(\theta+i-1)^2} \sim \theta\ln n$ quand $n\to+\infty$.

3. On obtient la convergence en probabilité de $K_n/\ln n$ vers θ avec l'inégalité de Tchebichev. Sa variance est équivalente à $1/\ln n$. Cela correspond à une vitesse de convergence en $1/\sqrt{\ln n}$, ce qui est *extrèmement lent*.

4. D'après la formule d'échantillonnage d'Ewens, si (Q_j) sont des v.a. indépendantes de Poisson de paramètres θ/j respectivement,

$$\begin{aligned}\mathbb{P}_\theta(B_1=b_1,\ldots,B_n=b_n \mid K_n=k) &= \mathbb{P}\left(Q_1=b_1,\ldots,Q_n=b_n \;\middle|\; \sum_{j=1}^n Q_j=k),\ \sum_{j=1}^n jQ_j=n\right)\\ &= \mathbb{E}\left[\mathbb{P}\left(Q_1=b_1,\ldots,Q_n=b_n \;\middle|\; \sum_{j=1}^n Q_j=k)\right) \;\middle|\; \sum_{j=1}^n jQ_j=n\right].\end{aligned}$$

Or $\sum Q_j$ est une v.a. de Poisson de paramètre $\theta\sum\frac{1}{j}$, donc

$$\mathbb{P}\left(Q_1=b_1,\ldots,Q_n=b_n \;\middle|\; \sum_{j=1}^n Q_j=k)\right) = \frac{\prod_{j=1}^n e^{-\theta/j}(\theta/j)^{k_j}/k_j!}{e^{-\theta\sum 1/j}(\theta\sum 1/j)^k/k!} = \frac{\prod_{j=1}^n \frac{1}{j^{k_j}k_j!}}{\frac{(\sum_{j=1}^k 1/j)^k}{k!}}.$$

Puisque le terme de droite ne dépend pas de θ, on a prouvé que $\mathbb{P}_\theta(B_1 = b_1, \ldots, B_n = b_n \mid K_n = k)$ ne dépend pas de θ.

5. D'après la question **1.**, $K_n = k$ ssi k des v.a. de Bernoulli $A_1, \ldots, A_n$ sont égales à 1 et les $n = k$ autres sont nulles. Soit $(\varepsilon_1, \ldots, \omega_n) \in \{0,1\}^n$.

$$\mathbb{P}(A_1 = \varepsilon_1, \ldots, A_n = \varepsilon_n) = \prod_{j=1}^{n} \frac{(i-1)^{1-\varepsilon_i}\theta^{\varepsilon_i}}{\theta + i - 1} = C(\varepsilon_1, \ldots, \omega_n) \frac{\theta^{\varepsilon_1 + \ldots + \varepsilon_n}}{\theta(\theta+1)\ldots(\theta+n-1)}.$$

On obtient la formule demandée en sommant sur tous les $(\varepsilon_1, \ldots, \varepsilon_n)$ tels que $\varepsilon_1 + \ldots + \varepsilon_n = k$. (On peut également montrer ce résultat directement à partir de la formule d'Ewens.)

6. On a donc

$$\frac{\partial}{\partial\theta} \ln L_n(\theta, k) = \frac{\partial}{\partial\theta} \left(k \ln\theta - \sum_{i=0}^{n-1} \ln(\kappa + i) \right) = \frac{1}{\theta} \left(k - \sum_{i=0}^{n-1} \frac{\theta}{\theta + i} \right).$$

L'estimateur du maximum de vraisemblance $\hat{\theta}$ satisfait donc

$$K_n = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i} \sim \hat{\theta} \ln n$$

lorsque $n \to +\infty$.

7. Le calcul donne

$$I_n(\theta) = \frac{1}{\theta^2} \mathbb{E}_\theta \left[\left(K_n - \sum_{i=0}^{n-1} \frac{\theta}{\theta + i} \right)^2 \right] = \frac{1}{\theta^2} \mathrm{Var}_\theta(K_n) \sim \frac{\ln n}{\theta}.$$

8. Puisque $\tilde{\theta}$ est sans biais,

$$0 = \mathbb{E}_\theta(\tilde{\theta} - \theta) = \sum_{k=1}^{n} [\tilde{\theta}(k) - \theta] L_n(\theta, k).$$

En dérivant cette relation par rapport à θ, on obtient

$$\sum_{k=1}^{n} [\tilde{\theta}(k) - \theta] L_n(\theta, k) \frac{\partial}{\partial\theta} \ln L_n(\theta, k) = 1.$$

L'inégalité de Cauchy-Schwartz implique alors

$$1 \leqslant \left(\sum_{k=1}^{n} [\tilde{\theta}(k) - \theta]^2 L_n(\theta, k) \right) \left(\sum_{k=1}^{n} \left[\frac{\partial \ln L_n(\theta, k)}{\partial\theta} \right]^2 L_n(\theta, k) \right) = \mathrm{Var}_\theta(\tilde{\theta})\, I_n(\theta).$$

9. Puisque $I_n(\theta) \sim \theta^{-1} \ln n$, il n'existe pas d'estimateur sans biais de θ, et donc pas d'estimateur de θ dont la variance est un $o(1/\ln n)$. Donc il n'existe pas d'estimateur convergeant plus vite que $\frac{K_n}{\ln n}$ (qui pourtant converge très lentement).