# Introduction to Bayesian learning, some exam's answers.
# Year 2018-2019, master Datascience.

Anne Sabourin

January 15, 2019

**Solution: Exercise 1** Données exponentielles censurées.

1. On note $p$ les densités et $p_{a,\lambda}$ la densité d'une loi Gamma$(a,\lambda)$. La formule de bayes donne:

$$\begin{aligned} p(\theta|Y \geq 100 &\propto \mathbb{P}_\theta(Y \geq 100) p_{a,\lambda}(\theta) \\ &= e^{-100\theta} \frac{\lambda^a}{\Gamma(a)} \theta^{a-1} e^{-\lambda\theta} \\ &\propto \theta^{a-1} e^{-(100+\theta)\lambda} \\ &\propto p_{a,\lambda+100}(\theta) \end{aligned}$$

   La loi a posteriori sachant $Y \geq 100$ est donc une loi Gamma$(a, \lambda + 100)$.

2. On a directement d'après la question précédente en appliquant les propriétés des lois gamma données dans le formulaire

$$\mathbb{E}(\theta|Y \geq 100) = \frac{a}{\lambda + 100} \; ; \mathbb{V}\mathrm{ar}(\theta|Y \geq 100) = \frac{a}{(\lambda+100)^2}.$$

3. Si l'observation est $Y = 100$, la loi a posteriori est classiquement proportionnelle au produit des densités

$$\begin{aligned} p(\theta|Y \geq 100 &\propto p_\theta(100) p_{a,\lambda}(\theta) \\ &= \theta e^{-100\theta} \frac{\lambda^a}{\Gamma(a)} \theta^{a-1} e^{-\lambda\theta} \\ &\propto \theta^{a+1-1} e^{-(100+\theta)\lambda} \\ &\propto p_{a+1,\lambda+100}(\theta) \end{aligned}$$

   d'où

$$\begin{aligned} \mathbb{E}(\theta|Y = 100) &= \frac{a+1}{\lambda+100} > \mathbb{E}(\theta|Y \geq 100) \\ \mathbb{V}\mathrm{ar}(\theta|Y = 100) &= \frac{a+1}{(\lambda+100)^2} > \mathbb{V}\mathrm{ar}(\theta|Y \geq 100) \end{aligned}$$

**Solution: Exercise 2** Prior de Jeffreys.

1. Modèle de Bernoulli $X \sim Ber(p)$.
   On note $\theta = p \in [0,1]$. La log-vraisemblance poru une observation $x \in \{0,1\}$ s'écrit

$$\log p_\theta(x) = x \log \theta + (1-x)\log(1-\theta)$$

   d'où

$$(\partial_\theta \log p_\theta(X))^2 = \left(\frac{X}{\theta} - \frac{1-X}{1-\theta}\right)^2$$

   et

$$\begin{aligned} I(\theta) = \mathbb{E}_\theta[(\partial_\theta \log p_\theta(X))^2] &= \theta(\frac{1}{\theta^2} + 0) + (1-\theta)(0 + \frac{1}{(1-\theta)^2}). \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

   Le prior de Jeffreys est donc $\pi(\theta) \propto \mathbb{1}_{[0,1]}(\theta)\theta^{-1/2}(1-\theta)^{-1/2}$. On reconnaît une loi Beta$(1/2, 1/2)$. La loi a posteriori a pour densité

$$\begin{aligned} \pi(\theta|x_{1:n}) &\propto \theta^{-1/2}(1-\theta)^{-1/2}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \\ &= \theta^{-1/2+\sum x_i}(1-\theta)^{-1/2+n-\sum x_i} \end{aligned}$$

   On reconnait une loi Beta$(\sum_1^n x_i + 1/2, n - \sum x_i + 1/2)$. Le prior de Jeffreys est donc conjugué.

2. Modèle Gaussien $X \sim \mathcal{N}(\theta, 1)$. On a $\log p_\theta(X) = C^{te} - (X-\theta)^2/2$ d'où $\partial_\theta \log p_\theta(X) = X - \theta$. Ainsi $I(\theta) = \mathbb{E}_\theta(X-\theta)^2 = \mathbb{V}\mathrm{ar}(X) = 1$. L'information de Fisher est constante, donc le prior de Jeffreys est le prior impropre de densité constante par rapport à la mesure de Lebesgue. La loi a posteriori est donc proportionnelle à la vraisemblance

$$\begin{aligned} \pi(\theta|x_{1:n}) &\propto e^{-\frac{\sum(x_i-\theta)^2}{2}} \\ &\propto e^{-\frac{n\theta^2 - 2\theta\sum x_i}{2}} \\ &\propto e^{-\frac{n^2(\theta - \sum x_i/n)^2}{2}} \end{aligned}$$

   La loi a posteriori est donc la loi normale $\mathcal{N}(\mu = \frac{1}{n}\sum x_i, \sigma^2 = \frac{1}{n^2})$.

**Solution: Exercise 3** Approximation a usens de KL.

1. Par définition

$$\begin{aligned} KL(p||q_{\mu,\Lambda}) &= \int_{\mathbb{R}^d} p(x)\log\frac{p(x)}{q_{\mu,\Lambda}(x)}\,\mathrm{d}x \\ &= \mathbb{E}_p\{\log p(X)\} - \mathbb{E}_p\{\log[\frac{1}{\sqrt{(2\pi)^d \det \Lambda^{-1}}} e^{-\frac{1}{2}X-\mu)^\top \Lambda(X-\mu)}]\} \\ &= \mathbb{E}_p \log p(X) - \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det\Lambda - \frac{1}{2}\mathbb{E}_p[(X-\mu)^\top\Lambda(X-\mu)]. \end{aligned}$$

2. On cherche $(\mu^*, \Lambda^*) \in \operatorname{argmin} KL(p||q_{\mu,Lambda}) = \operatorname{argmin} F(\mu, \Lambda)$, avec

$$F(\mu, \lambda) = \log\det\Lambda - \mathbb{E}_p[(X-\mu)^\top \Lambda (X-\mu).$$

On annule les gradients par rapport à $\mu, \Lambda$:

$$\begin{aligned}\nabla_\mu F(\mu,\Lambda) &= -\nabla_\mu \mathbb{E}_p (X-\mu)^\top \Lambda (X-\mu) \\ &= 2\mathbb{E}_p \Lambda (X-\mu) \qquad\qquad = 2\Lambda \mathbb{E}_p (X-\mu).\end{aligned}$$

Ainsi $\nabla_\mu F(\mu,\Lambda) = 0 \iff \mu = \mathbb{E}_p(X)$.

De plus, en notant $S$ la matrice de covariance de $X$ sous la loi $p$,

$$\begin{aligned}\nabla_\Lambda F(\mu,\Lambda) &= \nabla_\Lambda \log\det\Lambda - \nabla_\Lambda \mathbb{E}_p[Trace[(X-\mu)(X-\mu)^\top\Lambda]] \\ &= \Lambda^{-1} - \nabla_\Lambda Trace(S\Lambda) \\ &= \Lambda^{-1} - S.\end{aligned}$$

Ainsi $\nabla_\Lambda F(\mu,\Lambda) = 0 \iff \Lambda = S^{-1}$.
Conclusion: la meilleure approximation de $p$ par une loi normale au sens de KL a même moyenne et matrice de covariance que $p$.

**Solution: Exercise 4** 1. Sachant $X_t = x$, on a $X_{t+1} = \rho x + \epsilon_t$ avec $\epsilon_t \sim \mathcal{N}(0,1)$ d'où la loi conditionnelle de $X_{t+1}$:

$$\mathcal{L}(X_{t+1}|X_t = x) = \mathcal{N}(\rho x, 1)$$

La densité $k(x,y)$ du noyau de transition est donc la densité d'une loi normale de moyenne $\rho x$ et de variance 1, soit

$$k(x,y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\rho x)^2}$$

2. On cherche les paramètres $(\mu, \sigma^2)$ d'une loi normale tels que si $X \sim \mathcal{N}(\mu,\sigma^2)$ on a $X \overset{\mathrm{d}}{=} \rho X + \epsilon_t$. Le membre de droite est une loi normale de moyenne $\rho\mu$ et de variance $\rho^2\sigma^2 + 1$. Ainsi on a

$$\begin{cases}\mu = \rho\mu \\ \sigma^2 = \rho^2\sigma^2 + 1\end{cases}$$

d'où $\mu = 0, \sigma^2 = \frac{1}{1-\rho^2}$.