UNIVERSITÉ DE NANTES

Master 2 Ingénierie Statistique

Statistique Bayésienne: Exercices

Ex 1. Soient X, Y deux variables aléatoires indépendantes de lois $\Gamma(a, 1)$ et $\Gamma(b, 1)$ respectivement, avec $a, b > 0$.

1. Ecrire la densité du couple (X, Y).
2. Calculer la loi du couple $(V, W) := \big(X + Y, X/(X + Y)\big)$.
3. Déterminer les lois marginales de V et W. Commenter.
4. En déduire une expression de $B(a, b) := \int_0^1 x^{a-1}(1-x)^{b-1}dx$.

Ex 2. Soit $X = (X_1, ..., X_n)$ un échantillon de va iid de loi de Poisson de paramètre $\lambda > 0$.

1. Générer les données sous $\lambda = 1$ et représenter en fonction de n l'évolution de la loi a posteriori en prenant comme a priori π une loi exponentielle de paramètre 1, une loi uniforme sur $[0, 1]$ ou une loi géométrique de paramètre 1.
2. Dans chaque cas, préciser la limite en probabilité (pour la convergence étroite) de la loi a posteriori $\pi(.|X)$.
3. Dans quels cas les hypothèses du théorème de Bernstein-von Mises sont-elles vérifiées?
4. Reprendre la question 1 pour $\lambda = 1.5$. Commenter.

Ex 3. Soient $X_1, ..., X_n$ iid de loi $\mathcal{N}(\theta, 1)$ avec comme a priori sur $\theta \in \mathbb{R}$ la loi $\mathcal{N}(0, 1)$.

1. Calculer la loi a posteriori.
2. Vérifier le théorème de Bernstein-von Mises dans ce cas.
3. Calculer les estimateurs de Bayes associés aux fonctions de pertes

$$L_1(\theta, \eta) = (\theta - \eta)^2 \quad \text{et} \quad L_2 = e^{-\theta^2}(\theta - \eta)^2.$$

Ex 4. Soit $X_1, ..., X_n$ iid de loi uniforme sur $[0, \theta]$. On choisit un a priori de la forme $\pi_\lambda(\theta) \propto \theta^{-\lambda}\mathbb{1}\{\theta > 0\}$ pour $\lambda \in \mathbb{R}$.

1. Déterminer pour quelles valeurs de λ l'a priori est valable en précisant si c'est une loi de probabilité ou une loi impropre.
2. Calculer la loi a posteriori $\pi_\lambda(.|X)$ lorsque celle-ci est bien définie.
3. Calculer la loi marginale f_X en fonction de λ.

Ex 5. Soit $X_1, ..., X_n$ iid de loi $\mathcal{N}(\theta, 1)$, on veut estimer $\theta \in \mathbb{R}$ avec comme critère le coût quadratique $L(\theta, \eta) = (\theta - \eta)^2$. On considère l'ensemble de règles de décision $\delta_a(x_1, ..., x_n) = \frac{a}{n}\sum_{i=1}^n x_i$ pour $a \geq 0$.

1. Montrer que δ_1 est préférable à δ_a pour tout $a > 1$.
2. Montrer que δ_a est admissible pour $a \in [0, 1]$.

Ex 6. Soient $X_1, ..., X_n$ iid de loi de Bernoulli de paramètre $\theta \in]0,1[$. Calculer la loi a posteriori (éventuellement à une constante multiplicative près) dans les cas suivants:

1. L'a priori sur θ est une loi discrète $\pi(\theta) = \sum_{j=1}^{k} \pi_j \delta_{t_j}(\theta)$ avec $t_1, ..., t_k \in]0,1[$ et $\sum_{j=1}^{k} \pi_j = 1$.
2. L'a priori est un mélange de lois uniformes de densité $\pi(\theta) = \sum_{j=1}^{k} \pi_j \frac{1}{t_{j+1} - tj} \mathbb{1}\{t_j < \theta < t_{j+1}\}$ avec $0 < t_1 < ... < t_{k+1} < 1$.
3. L'a priori est une loi beta $B(a, b)$, $a, b > 0$.

Ex 7. On veut estimer la proportion p de daltoniens dans une population. Sur un échantillon de 30 personnes issues de cette population, 5 sont diagnostiquées. On envisage trois lois a priori sur p:

i) la loi discrète $\mathbb{P}(p = j/10) = C/j$, $j = 1, ..., 9$ où C est une constante de normalisation.

ii) le mélange de lois uniformes $\pi(p) \propto 1/(\lfloor 10p \rfloor + 10)\mathbb{1}\{p \in]0,1[\}$ où $\lfloor . \rfloor$ désigne la partie entière.

iii) la loi continue de densité $\pi(p) \propto 1/(p+1)\mathbb{1}\{p \in]0,1[\}$

1. Représenter graphiquement ces trois lois a priori et calculer numériquement leurs moyennes et variances.
2. Superposer dans chaque cas la loi a posteriori calculée à partir de l'échantillon.
3. Calculer les moyennes et variances a posteriori. Commenter.

Ex 8. Soit $X = (X_1, ..., X_n)$ un échantillon iid de densité $f_\theta, \theta \in \Theta \subseteq \mathbb{R}$ et $\pi(.)$ une densité a priori sur θ telle que $\int |\theta| \pi(\theta) d\theta < \infty$.

1. Justifier que $\int \theta \pi(\theta|X) d\theta$ est bien défini presque sûrement.
2. Montrer que pour tout $\eta \in \Theta$,
$$\mathbb{E}\big(|\eta - \theta|\big|X\big) = \eta\big(2\mathbb{P}(\theta < \eta|X) - 1\big) + \mathbb{E}(\theta|X) - 2\int_{-\infty}^{\eta} \theta \pi(\theta|X) d\theta.$$
3. En déduire que la médiane a posteriori est l'estimateur Bayésien associé à la perte $\mathbb{L}^1$.

Ex 9. Simuler un échantillon iid $X = (X_1, ..., X_n)$ de taille $n = 1000$ de loi exponentielle $\mathcal{E}(1/2)$. On s'intéresse au comportement de la loi a posteriori dans le modèle exponentiel $\mathcal{M} = \{\mathcal{E}(\theta), \theta > 0\}$ lorsque la loi a priori sur θ est une loi gamma $\Gamma(a, b)$ avec des paramètres a et b à déterminer.

1. On suppose qu'une source d'information (extérieure) nous dit que θ est "vraisemblablement proche de 1/2". Proposer un ensemble de valeurs de (a, b) adapté dans ce cas.
2. On suppose que la fiabilité de l'information correspond à une variance a priori de $\tau = a/b^2 = 1$. En déduire la loi a priori choisie.
3. Superposer la densité a priori et les densités a posteriori construites à partir des k premières valeurs de l'échantillon pour $k = 2, 5, 10, 100, 500, 1000$. Commenter.
4. Reprendre la question précédente pour des variances a priori τ valant $0.01, 0.1, 10$ et 100, en gardant $1/2$ comme moyenne a priori. Commenter.
5. Proposer trois estimateurs de θ construits à partir de la loi a posteriori. On les notera $\widehat{\theta}_k^{(1)}, \widehat{\theta}_k^{(2)}, \widehat{\theta}_k^{(3)}$ où k représente la taille de l'échantillon.
6. Représenter graphiquement l'évolution de ces trois estimateurs et de l'estimateur du maximum de vraisemblance en faisant varier la taille de l'échantillon. Interpréter les résultats.
7. Reprendre l'exercice en partant de l'information initiale: "θ est vraisemblablement proche de 3".

Ex 10. Soit $N_1, ..., N_n$ les nombres de pièces défectueuses dans n lots de 50 pièces. On veut estimer la probabilité p qu'une nouvelle pièce soit défectueuse.

1. Décrire le modèle statistique.
2. On prend comme loi a priori sur p une loi beta $B(a, b)$. Calculer la loi a posteriori. Commenter.
3. Le service qualité nous apprend que "la proportion de pièces défectueuse dans un lot est probablement proche de 0.15" et "comprise entre 0.1 et 0.2 avec probabilité 95%". En déduire des valeurs de a, b adaptées.

Ex 11. Soit $X_1, ..., X_n$ iid de loi uniforme sur $[0, \theta]$ avec un a priori exponentiel $\pi_\lambda(\theta) = \lambda e^{-\lambda\theta}\mathbb{1}\{\theta > 0\}$, $\lambda > 0$. On considère un modèle hiérarchique en définissant un a priori sur λ, noté π. Calculer l'a priori sur θ correspondant pour π:

1. la loi exponentielle de paramètre 1.
2. la loi $\Gamma(2, 1)$.
3. la loi géométrique (sur $\mathbb{N}^*$) de paramètre $1/2$.

Ex 12. Montrer que les familles de lois sont conjuguées dans les modèles suivants:

1. Les lois gamma $\Gamma(a, b)$, $a, b > 0$ dans le modèle de Poisson $\mathcal{P}(\theta), \theta > 0$.
2. Les lois normales $\mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \lambda > 0$ dans le modèle Gaussien sur la moyenne $\mathcal{N}(\theta, 1), \theta \in \mathbb{R}$.
3. Les lois inverse-gamma $\Gamma^{-1}(a, b)$, $a, b > 0$ dans le modèle Gaussien sur la variance $\mathcal{N}(0, \theta), \theta > 0$.
4. Les a priori $\sigma^2 \sim \Gamma^{-1}(a, b)$ et $\theta \sim \mathcal{N}(c, \sigma^2/d)$ pour $a, b, c \in \mathbb{R}, d > 0$ dans le modèle Gaussien $\mathcal{N}(\theta, \sigma^2)$, $\theta \in \mathbb{R}$, $\sigma^2 > 0$.

Ex 13. Soit le modèle de Bernoulli $\mathcal{B}(p), p \in (0, 1)$.

1. Calculer l'information de Fisher $\mathcal{I}(p)$ pour un échantillon iid $X_1, ..., X_n$.
2. Montrer que l'a priori de Jeyffreys sur p est la loi beta $B(0.5, 0.5)$.
3. En déduire l'a priori de Jeyffreys sur $\theta := \arcsin(\sqrt{p}) \in]0, \pi/2[$.

Ex 14. Soit le modèle de Poisson $\mathcal{P}(\theta), \theta > 0$.

1. Déterminer l'a priori de Jeyffreys sur θ et vérifier que c'est une loi impropre.
2. En déduire l'a priori de Jeyffreys sur $\eta := e^{-\theta}$, $\eta \in]0, 1[$.

Ex 15. Déterminer l'a priori non-informatif de Jeyffreys dans le modèle Gaussien

1. sur la moyenne avec variance connue $\mathcal{N}(\theta, 1), \theta \in \mathbb{R}$.
2. sur la variance avec moyenne connue $\mathcal{N}(0, \sigma^2), \sigma^2 > 0$.
3. sur l'écart-type avec moyenne connue $\mathcal{N}(0, \sigma^2), \sigma > 0$.
4. sur le couple moyenne-variance dans le cas général $\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}, \sigma^2 > 0$.

Ex 16. Soit $X_1, ..., X_n$ iid de loi uniforme sur $[0, \theta]$ avec $\theta > 0$ et $M_n := \max\{X_1, ..., X_n\}$. On se place dans le modèle uniforme $\mathcal{M} = \{\mathcal{U}[0, \theta], \theta > 0\}$ avec l'a priori de Laplace $\pi(\theta) = \mathbb{1}\{\theta > 0\}$.

1. Ecrire la vraisemblance du modèle en fonction de M_n.
2. Montrer que l'a priori est impropre.
3. Calculer la loi a posteriori associée et la représenter graphiquement.
4. Déterminer la région HPD de niveau $1 - \alpha \in]0, 1[$, notée R_α^{HPD}.
5. Calculer la probabilité fréquentiste $\mathbb{P}_\theta\big(\theta \in R_\alpha^{HPD}\big)$.

Ex 17. Soit $X = (X_1, ..., X_n)$ un échantillon iid de loi normale $\mathcal{N}(\theta, 1)$. On prend comme a priori sur θ la loi normale $\mathcal{N}(0, 1/\tau)$, $\tau > 0$.

1. Montrer que la loi a posteriori est la loi normale $\mathcal{N}\Big(\frac{\overline{X}}{1+\tau/n}, \frac{1}{n+\tau}\Big)$ où $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.
2. Montrer que les régions HPD de niveau $1 - \alpha \in]0, 1[$ sont de la forme
$$R_\alpha^{HPD} = \Big[\frac{\overline{X}}{1+\tau/n} - \frac{q_{1-\alpha/2}}{\sqrt{n+\tau}}, \frac{\overline{X}}{1+\tau/n} + \frac{q_{1-\alpha/2}}{\sqrt{n+\tau}}\Big]$$
où q_α désigne le quantile d'ordre α de la loi normale standard.
3. Calculer la probabilité fréquentiste d'appartenance à la région HPD $\mathbb{P}_\theta\big(\theta \in R_\alpha^{HPD}\big)$ en fonction de α et θ (exprimer cette probabilité à l'aide de la fonction de répartition Φ de la loi normale standard).
4. Calculer la limite de cette probabilité quand n tend vers l'infini. Commenter.
5. Calculer la limite à n fixé quand $\tau \to 0$. Commenter.
6. Quel a priori choisir pour que les régions HPD correspondent à des intervalles de confiance fréquentistes classiques?

Ex 18. Simuler un échantillon $X_1, ..., X_n$ de taille $n = 50$ iid de loi exponentielle de paramètre $\theta = 2$. On prend comme a priori sur θ la loi $\Gamma(1, 1)$.

1. Calculer et représenter graphiquement tous les intervalles de crédibilité à 95% sur θ.
2. Rechercher numeriquement l'intervalle le plus court. A quoi correspond-il?
3. Calculer numériquement son niveau fréquentiste.
4. Refaire l'exercice pour $\theta = 10$ avec le même a priori. Commenter.

Ex 19. Sur les 48 derniers mois dans le désert d'Atacama, 6 ont connu au moins un jour de pluie. On modélise la présence de pluie un mois donné par une variable de Bernoulli X_i de paramètre $p \in (0, 1)$ indépendante du passé. On note $S = \sum_{i=1}^{48} X_i$, on observe donc ici $S = 6$. On fixe comme a priori sur p une loi beta $B(2, 10)$.

1. Calculer la loi a posteriori.
2. On veut prévoir le nombre T de mois pluvieux lors des N prochains mois. Déterminer la loi de T conditionnellement à (S, p).
3. En déduire la loi prédictive de T, c'est-à-dire la loi de T conditionnellement à S.
4. Donner la prévision Bayésienne $\widehat{T}$ de T pour la perte $\mathbb{L}^2$.
5. Comparer avec l'approche fréquentiste classique.
6. Superposer les bornes du plus court intervalle de prévision à $80\%, 95\%$ et 99%.

Ex 20. Soient $X_1, ..., X_n$ des v.a. telles que $X_1 \sim \mathcal{N}(0, 1)$ et pour $i = 2, ..., n$, la loi de X_i conditionnellement au passé $X_{i-1}, ..., X_1$ est la loi normale $\mathcal{N}(\theta X_{i-1}, 1)$ avec $\theta \in \mathbb{R}$ inconnu.

1. Ecrire la vraisemblance du modèle.
2. On choisit l'a priori Gaussien standard sur θ. Calculer la loi a posteriori.
3. Donner la loi d'une nouvelle valeur X_{n+1} conditionnellement à $(X_1, ..., X_n, \theta)$.
4. Simuler une trajectoire $X_1, ..., X_n, X_{n+1}$ pour $n = 50$ et $\theta = 1$.
5. Programmer une fonction qui, étant donné $X = (X_1, ..., X_n)$, simule un échantillon iid sous la loi prédictive de X_{n+1} sachant X.
6. A l'aide de cette fonction, représenter graphiquement une approximation de la densité de la loi prédictive.
7. Superposer au graphique la vraie valeur X_{n+1}, le prédicteur Bayésien pour la perte quadratique et le plus court intervalle de prévision de couverture 95%.

Ex 21. Soit $X = (X_1, ..., X_n)$ iid de loi de Poisson de paramètre θ. On considère l'a priori $\pi(\theta) \propto 1/(1+\theta)^2$.

1. Simuler un échantillon X pour $\theta = 3$ et $n = 100$.
2. Ecrire la loi marginale $f_X(X)$ sous la forme d'une espérance sous la loi a priori.
3. Construire une approximation par Monte-Carlo de $f_X(X)$ à partir d'un échantillon iid $\theta_1, ..., \theta_N$ de loi π.
4. Donner un intervalle de confiance à 95%.
5. Calculer l'information de Fisher $\mathcal{I}(\theta)$ (pour tout l'échantillon).
6. Calculer l'estimateur $\widehat{\theta}_{MAP}$ du maximum a posteriori.
7. Reprendre l'exercice en utilisant un échantillon $\theta_1, ..., \theta_N$ de loi $\mathcal{N}\big(\widehat{\theta}_{MAP}, 1/\mathcal{I}(\widehat{\theta}_{MAP})\big)$.

Ex 22. Soit $X = (X_1, ..., X_n)$ un échantillon iid de loi de Bernoulli $\mathcal{B}(p)$ et $S = \sum_{i=1}^n X_i$, on observe $S = 22$ dans un échantillon de taille $n = 50$. On cherche à estimer $\theta := \sqrt{\arcsin(p)}$ en prenant comme a priori sur p la loi beta $B(0.5, 0.5)$.

1. Ecrire $\delta(X)$ sous la forme d'une intégrale en p.
2. Construire une approximation $\widehat{\delta(X)}$ de $\delta(X)$ par Monte-Carlo à partir d'un échantillon simulé de taille $N = 10000$ sous la loi a posteriori.
3. Estimer la variance de l'approximation en se basant sur l'échantillon déjà simulé.
4. En déduire un intervalle de confiance asymptotique à 95% pour $\delta(X)$.
5. Représenter graphiquement l'approximation $\widehat{\delta(X)}$ et sa région de confiance en fonction de N.
6. Reprendre l'exercice en générant cette fois l'échantillon de Monte-Carlo sous la loi a priori (exprimer $\delta(X)$ comme une espérance sous cette loi).
7. Commenter.

Ex 23. Soit (X, Y) un couple de v.a. de densité sur $\mathbb{R}^2$ donnée par

$$f(x, y) = Ce^{-(x+y+2z+xy+yz)}\mathbb{1}\{x, y > 0\}$$

où C est une constante de normalisation.

1. Calculer la loi de X sachant Y, Z, de Y sachant X, Z et de Z sachant X, Y.
2. Générer une chaîne de Markov de loi invariante de densité f.
3. Représenter plusieurs trajectoires de la chaîne avec des points de départ différents.

Ex 24. On considère le modèle de régression linéaire simple. On observe

$$Y_i = a + bx_i + \epsilon_i \ , \ i = 1, ..., n$$

où les x_i sont déterministes et les ϵ_i sont iid de loi normale $\mathcal{N}(0, \sigma^2)$.

1. On suppose σ^2 connu. Calculer la loi a posteriori pour σ^2 connu et $\pi(a, b)$ est la densité d'un vecteur Gaussien standard de $\mathbb{R}^2$.
2. Que vaut l'estimateur du maximum a posteriori?
3. Proposer une interprétation Baysienne d'un estimateur du type

$$(\hat{a}, \hat{b}) = \arg\min_{a,b\in\mathbb{R}} \sum_{i=1}^n (Y_i - a - bx_i)^2 + \text{pen}(a, b)$$

où pen(.) est une pénalité quelconque.

4. On suppose maintenant σ^2 est inconnu. Donner l'a priori de Jeyffreys sur $\theta = (a, b, \sigma^2)$ puis la loi a posteriori associée.

5. Rappeler la forme des estimateurs de Bayes de a, b et σ^2 pour le coût quadratique. Quel problème rencontre-t-on ici si on veut les calculer explicitement?

6. Construire une approximation de ces estimateurs par un algorithme de type MCMC.

Ex 25. Régression logistique bayésienne

Ex 26. Soit $X = (X_1, ..., X_n)$ iid de loi de Poisson $\mathcal{P}(\theta)$. On choisit comme loi a priori sur θ la loi exponentielle de paramètre $\lambda > 0$ fixé.

1. Déterminer la loi a posteriori et l'estimateur de Bayes sous le coût quadratique.

2. Calculer numériquement le plus court intervalle de crédibilité de couverture 95%.

3. On définit une structure hiérarchique sur le modèle en prenant comme a priori sur λ une loi exponentielle de paramètre 1. Ecrire le DAG du modèle.

4. Sous JAGS, le modèle se définit dans un fichier à part (par exemple `model.R` dans le répertoire courant) comme suit

```
model{
  lambda~dexp(1)
  theta~dexp(lambda)
  for(i in 1:n){X[i]~dpois(theta)}}
```

   On génère maintenant une chaîne de Markov de loi invariante la loi a posteriori sur θ:

```
library(rjags)
set.seed(2048)
n<-10
X<-rpois(n,5)
J<-jags.model('model.R',data=list('X'=X,'n'=n),n.chains=1)
update(J,1000)
m=coda.samples(J,'theta',10000)
```

   Commenter ces lignes de codes et utiliser les fonctions `summary` et `plot` pour visualiser le résultat.

5. Comparer les résultats numériques avec les résultats théoriques sur la loi a posteriori et l'estimateur de Bayes.

6. En modifiant le paramètre `n.chains` de la fonction `jags.model`, générer simultanément 10 chaînes de Markov. Superposer les moyennes cumulées.

7. En utilisant les données déjà simulées, représenter graphiquement une approximation de la densité a posteriori, puis construire une estimation du plus court intervalle de crédibilité.

8. Comparer avec les résultats théoriques.

Ex 27. Dans un hôpital, on relevé la consommation mensuelle A_i d'un antibiotique et le pourcentage R_i de bactéries resistantes à l'antibiotique le mois suivant. On note a le taux de bactéries resistantes dans la population lorsque la consommation est nulle. On sait que la proportion de bactéries résistantes ne varie pas en-dessous d'un certain seuil s de consommation, et qu'elle croît avec la consommation quand ce seuil est dépassé. Des études ont montré que, sur 10 autres antibiotiques répertoriés, les taux de résistances dans la population à consommation nulle sont en moyenne de 0.05 avec un écart-type de 0.01, et, que le seuil de consommation à partir duquel des resistances se développent est toujours supérieur à 10 et est inférieur à 20 dans 95% des cas.

1. A partir des données `bacteries.csv`, proposer un modèle statistique pour décrire le taux de resistance à l'antibiotique en fonction de la consommation du mois précédent.

2. Proposer des lois a priori sur les paramètres et représenter les lois a posteriori, à l'aide de JAGS.

3. Conclure.