

Université de Nantes
Master professionnel II : Ingénierie mathématique

Statistique Bayésienne et Simulation

Exercices

Anne PHILIPPE

`anne.philippe@math.univ-nantes.fr`

Le logiciel utilisé pendant les séances de TP est le logiciel libre R.

Quelques adresses :

<http://www.r-project.org/>

http://www.math.sciences.univ-nantes.fr/~philippe/R_freeware.html

5 décembre 2011

CHAPITRE 1

Pré requis

Exercice 1. Calcul de lois : la loi β

Soit $(a, b) \in \mathbb{R}_+ \times \mathbb{R}_+$.

Soit X et Y deux variables aléatoires *indépendantes*. On suppose que

- X est distribuée suivant la loi $\Gamma(a, 1)$
- Y est distribuée suivant la loi $\Gamma(b, 1)$

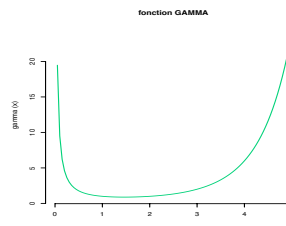
On rappelle que la loi $\Gamma(a, 1)$ admet pour densité

$$f(x) = \frac{1}{\Gamma(a)} e^{-x} x^{a-1} \mathbb{I}_{\mathbb{R}_+}(x).$$

La fonction Γ est définie sur $\mathbb{R}_+$ par

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt = (a-1)\Gamma(a-1).$$

On a en particulier pour tout $a \in \mathbb{N}$, $\Gamma(a+1) = a!$



- 1) Écrire la densité de la loi du couple (X, Y)
- 2) Donner la loi du couple $(V, W) = (X + Y, \frac{X}{X + Y})$.
- 3) Les variables aléatoires V et W sont-elles indépendantes? Préciser les lois marginales de V et W .
- 4) En déduire une expression de $B(a, b) := \int_0^1 x^{a-1} (1-x)^{b-1} dx$.

Exercice 2. Convergence

Soit $(\alpha, \beta) \in \mathbb{R}_+^*$. On dispose de deux échantillons indépendants $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ qui possèdent les propriétés suivantes

- $(X_1, \dots, X_n)$ sont iid suivant la loi exponentielle de paramètre α
- $(Y_1, \dots, Y_n)$ sont iid suivant la loi exponentielle de paramètre $\alpha\beta$.

La loi exponentielle de paramètre $\tau > 0$ a pour densité $f(x) = \tau e^{-\tau x} \mathbb{I}_{\mathbb{R}_+}(x)$. On souhaite estimer les paramètres α et β à partir des observations $(X_1, \dots, X_n, Y_1, \dots, Y_n)$.

- 1) Écrire la fonction de vraisemblance de l'échantillon $(X_1, \dots, X_n, Y_1, \dots, Y_n)$.
- 2) Calculer les estimateurs du maximum de vraisemblance des paramètres (α, β) . On les note $(\hat{\alpha}, \hat{\beta}_n)$.
- 3) On pose

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{et} \quad \bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$$

Montrer que le vecteur $Z_n := \begin{pmatrix} \bar{X}_n \\ \bar{Y}_n \end{pmatrix}$ converge presque sûrement vers $z = \begin{pmatrix} \alpha^{-1} \\ \alpha^{-1}\beta^{-1} \end{pmatrix}$.

- 4) En déduire que $(\hat{\alpha}_n, \hat{\beta}_n)$ est un estimateur convergent au sens de la convergence presque sûre.
- 5) Montrer que $\sqrt{n}(Z_n - z)$ converge en loi vers une variable gaussienne. Préciser la moyenne et la variance de la limite.
- 6) En déduire que

$$\sqrt{n} \left(\begin{pmatrix} \hat{\alpha}_n \\ \hat{\beta}_n \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right)$$

converge en loi vers une variable gaussienne. Préciser la moyenne et la variance de la limite.

Exercice 3. Conditionnement pour des lois discrètes

Soit (X_1, X_2) un couple de variables aléatoires indépendantes et de même loi définie par

$$\mathbf{P}(X_1 = k) = \alpha^k(1 - \alpha) \quad \forall k \in \mathbb{N}$$

où α est fixé dans $]0, 1[$. (On note $\tilde{G}(\alpha)$ cette loi)

On pose

$$N = \begin{cases} 1 & \text{si } X_1 \leq X_2 \\ 2 & \text{sinon} \end{cases}$$

- 1) Déterminer la loi de X_1 sachant $N = 1$, puis l'espérance conditionnelle.
- 2) Déterminer la loi de X_1 sachant $N = 2$, puis l'espérance conditionnelle.
- 3) Exprimer l'espérance de X_1 en fonction des espérances conditionnelles calculées aux questions précédentes
- 4) Calculer la loi et l'espérance de N conditionnellement à X_1 .

Exercice 4. Conditionnement dans le cas de lois continues

Les deux variables aléatoires X et Y sont positives, et la loi du vecteur (X, Y) est donnée par

– Y a une loi de densité

$$f(t) = \lambda^2 t e^{-\lambda t} \mathbb{I}_{\mathbb{R}^+}(t)$$

– La loi de X conditionnellement à Y est la loi uniforme sur $[0, Y]$.

- 1) Quelle est la loi du vecteur (X, Y) ?
- 2) Quelle est la loi de X ?
- 3) Montrer que la loi de Y conditionnellement à X a comme densité

$$\lambda e^{-\lambda(t-X)} \mathbb{I}_{[X, +\infty[}(t)$$

- 4) Calculer $\mathbb{E}(Y|X)$ et $\mathbb{E}Y$.

CHAPITRE 2

Statistique Bayésienne

Exercice 1. Modèle exponentiel – [avec R]

Récupérer le fichier de données

<http://www.math.sciences.univ-nantes.fr/~philippe/data/duree-de-vie.txt>

qui contient des durées de fonctionnement de 1000 ampoules.

On modélise ces données par des variables aléatoires $X_1, \dots, X_n$ iid suivant la loi exponentielle de paramètre $\theta \in \mathbb{R}_+^*$. On veut estimer le paramètre θ à l'aide d'une modélisation bayésienne.

Rappel sur le modèle exponentiel :

- Si la loi a priori du paramètre θ est la loi gamma $\Gamma(a, b)$ alors la loi a posteriori est aussi une loi gamma de paramètres $n + a, b + \sum_{i=1}^n X_i$.
- Loi $\Gamma(a, b)$: l'espérance est égale à a/b et la variance à a/b^2

- 1) L'information fournie a priori sur θ devrait être proche de $1/2$. On note τ la variance de la loi a priori. Proposer des paramètres pour la loi a priori.
- 2) On choisit $\tau = 1/2$. Superposer la densité de la loi a priori et les densités a posteriori pour différentes tailles d'échantillon n (par exemple $n \in \{2, 5, 10, 100, 500, 1000\}$)
- 3) Reprendre la question précédente pour différentes valeurs de $\tau = 1/100, 1/10, 10, 100$.
- 4) Proposer trois estimateurs de θ à partir de la loi a posteriori.
- 5) Pour les différentes valeurs de τ :
 - 5 - a) Calculer et représenter ces estimateurs en fonction de n le nombre d'observations.
 - 5 - b) Calculer et ajouter au graphique précédent l'estimateur du maximum de vraisemblance de θ
- 6) Interpréter les résultats obtenus.
- 7) L'information a priori sur θ est maintenant la suivante " θ est autour de 3". Reprendre les questions précédentes
- 8) Conclure

Exercice 2. Information a priori

Rappel

Loi beta $\beta(\alpha, \beta)$: l'espérance est égale à $\frac{\alpha}{\alpha+\beta}$ et la variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Trois personnes veulent estimer la proportion p d'étudiants qui ne résident pas sur le campus.

- A suppose que la loi a priori sur p est la loi discrète définie par

p_i	.1	.2	.3	.4	.5
$\pi(p_i)$	.5	.2	.2	.05	.05

- B suppose que la loi a priori sur p est la loi Beta de paramètres (3, 12)
- C suppose que la loi a priori sur p est Beta de paramètres (1, 4)

- 1) Calculer la moyenne et l'écart type de ces trois lois a priori.

- 2) Ont-ils la même information a priori sur la localisation du paramètre p ? Accordent ils la même confiance à l'information a priori obtenue?
- 3) Donner l'expression de la loi prédictive a priori $m(y)$ dans les trois situations. On veut prédire a priori y le nombre d'étudiants qui habitent hors du campus dans un échantillon de taille 12.
- 4) Superposer les courbes représentatives des densités des lois prédictives a priori.

Exercice 3. [avec R]

On veut estimer p la proportion des étudiants qui dorment plus de 8 heures par nuit. Les observations sur un échantillon de 27 étudiants sont :

11 étudiants dorment plus de 8 heures
16 étudiants dorment moins de 8 heures.

On envisage trois lois a priori différentes sur p :

A- La loi discrète définie dans la TABLE 1

i	b_i	$P(p = b_i)$
1	0.05	0.03
2	0.15	0.18
3	0.25	0.28
4	0.35	0.25
5	0.45	0.16
6	0.55	0.07
7	0.65	0.02
8	0.75	0.00
9	0.85	0.00
10	0.95	0.00

TABLE 1

B- Un mélange de loi uniforme (loi a priori de type histogramme) défini par

i	0	1	2	3	4	5	6	7	8	9
q_i	2	4	8	8	4	2	1	1	1	1

où

$$a_i = i/10$$

et

$$P(p \in [a_i, a_{i+1}]) = \frac{q_i}{\sum_{j=0}^9 q_j}$$

C- La loi Beta de paramètres $a = 3.4$ and $b = 7.4$.

- 1) Représenter graphiquement les trois lois a priori, calculer la moyenne et la variance de ces lois a priori
- 2) Pour les trois lois a priori proposées, calculer la loi a posteriori, sa moyenne et sa variance.
- 3) Représenter graphiquement les trois lois a posteriori.
- 4) Commenter les résultats obtenus

Exercice 4. Région de confiance bayésienne [avec R]

On considère un n échantillon iid suivant la loi exponentielle de paramètre θ inconnu. On suppose que la loi a priori est la loi $\Gamma(a, b)$. L'information a priori est " θ proche de 1".

- 1) Proposer une loi a priori sur θ de variance τ .
- 2) En utilisant les résultats du cours, programmer une fonction qui calcule l'intervalle de confiance bayésien de niveau 95%.
- 3) 3 - a) Simuler $N(= 100 \text{ ou } 500)$ échantillons de taille $n = 15$ suivant la loi exponentielle de paramètre 1
 3 - b) Calculer, puis tracer les intervalles de confiance bayésiens pour $\tau = .01; 1, 10$.
 3 - c) Estimer la longueur moyenne des intervalles.
 3 - d) A partir des N intervalles, estimer la probabilité fréquentiste des intervalles bayésiens pour les trois valeurs de τ
- 4) Reprendre la question 3) avec des échantillons simulés suivant la loi exponentielle de paramètre $1/4$ et $n = 15$
- 5) Reprendre la question 3 avec des échantillons simulés suivant la loi exponentielle de paramètre $1/4$ et $n = 100$

Exercice 5. Modèle Gaussien

On dispose de n observations $X_1, \dots, X_n$ iid suivant une loi gaussienne $\mathcal{N}(\theta, 1)$.

On choisit comme loi a priori sur θ la loi gaussienne $\mathcal{N}(0, \tau^{-2})$, $\tau > 0$

- 1) Montrer que la loi a posteriori est une loi Gaussienne

$$\mathcal{N}\left(\frac{\bar{X}_n}{1 + \tau^2/n}, \frac{1}{n + \tau^2}\right)$$

- 2) Montrer que les régions HPD de niveau $1 - \alpha (= .95)$ sont de la forme

$$\theta \in \left[\frac{\bar{X}_n}{1 + \tau^2/n} - \frac{u_{1-\alpha/2}}{\sqrt{n + \tau^2}}; \frac{\bar{X}_n}{1 + \tau^2/n} + \frac{u_{1-\alpha/2}}{\sqrt{n + \tau^2}} \right] = I^{HPD}(\tau, \bar{X}_n)$$

où u_α est le quantile d'ordre α de la loi gaussienne standard.

- 3) Montrer que

$$P_\theta(\theta \in I^{HPD}(\tau, \bar{X}_n)) = F\left(\frac{\theta\tau^2}{\sqrt{n}} + u_{1-\alpha/2}\sqrt{\frac{n + \tau^2}{n}}\right) - F\left(\frac{\theta\tau^2}{\sqrt{n}} - u_{1-\alpha/2}\sqrt{\frac{n + \tau^2}{n}}\right)$$

où F est la fonction de répartition de la loi gaussienne standard.

- 4) Quelle est la limite de cette probabilité quand $n \rightarrow \infty$. Commenter.
- 5) Quelle est la limite de cette probabilité quand $\tau \rightarrow 0$. Commenter.
- 6) Comment choisir la loi a priori pour que les régions HPD de niveau $1 - \alpha$ soient aussi des régions de confiance au sens classique de niveau $1 - \alpha$.

Exercice 6

Soit $\theta > 0$ un paramètre. Soit $\mathbf{X} = (X_1, \dots, X_n)$ des variables aléatoires indépendantes et identiquement distribuées suivant la loi uniforme sur $[\theta, 2\theta]$. On note $m_n = \min\{X_1, \dots, X_n\}$ et $M_n = \max\{X_1, \dots, X_n\}$.

- 1) Écrire la vraisemblance du modèle.

Indication : commencer par montrer que presque sûrement, $M_n \leq 2m_n$

On suppose que la loi a priori impropre est donnée par $\pi(\theta) \propto \mathbb{I}_{\mathbb{R}^+}(\theta)$.

- 2) Montrer que la loi a posteriori est bien définie.
- 3) Calculer la densité de la loi a postérieure.

- 4) Représenter graphiquement l'allure de densité de la loi a posteriori .
- 5) En déduire que les régions de confiance HPD pour la paramètre θ sont de la forme $[\frac{M_n}{2}, b_\alpha]$.
- 6) Donner l'expression b_α en fonction de $1 - \alpha$ le niveau de la région de confiance.
- 7) Calculer le niveau fréquentiste de la région HPD de niveau $1 - \alpha$

Exercice 7

Soient $X_1, \dots, X_n$ des variables aléatoires iid suivant la loi de densité f_θ , où $\theta \in \Theta$. Le paramètre θ est inconnu, on l'estime par une approche bayésienne.

On considère la fonction de coût

$$L(\theta, \delta) = e^{\delta - \theta} - (\delta - \theta) - 1.$$

- 1) Montrer que la fonction L définit bien une fonction de coût.
- 2) Montrer que

$$\delta^\pi(X_1, \dots, X_n) = -\log(\mathbb{E}(e^{-\theta} | X_1, \dots, X_n))$$

est un estimateur de Bayes pour la fonction de coût L .

Exercice 8. Calcul des lois a priori conjuguées

Donner une famille de lois conjuguées pour les modèles $\{P_\theta, \theta \in \Theta\}$ suivants

- 1) Poisson : $\{\mathcal{P}(\theta), \theta \in \mathbb{R}_+^*\}$

Indication : loi gamma

- 2) Gaussien de variance connue : $\{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$

Indication : loi gaussienne

On dit que X suit une loi inverse Gamma de paramètres $(a, b) \in \mathbb{R}_+^{*2}$, si X^{-1} suit une loi Gamma de paramètres (a, b) .

- 3) Montrer que la loi inverse Gamma de paramètres (a, b) admet pour densité

$$b^a \frac{1}{\Gamma(a)} e^{-b/x} x^{-a-1} \mathbb{I}_{\mathbb{R}_+^*}(x)$$

- 4) Gaussien de moyenne connue : $\{\mathcal{N}(\mu, \theta), \theta \in \mathbb{R}_+^*\}$

Indication : loi inverse gamma.

- 5) Gaussien : $\{\mathcal{N}(\theta_1, \theta_2), \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}_+^*\}$

Indication : Considérer la famille des lois de paramètres $(\lambda, \tau, a, b) \in \mathbb{R} \times \mathbb{R}_+^{*3}$ définies par

- la loi conditionnelle de θ_1 sachant θ_2 est la loi gaussienne de moyenne λ et de variance θ_2/τ
- la loi de θ_2 est la loi inverse Gamma de paramètres (a, b) .

Exercice 9. Calcul des lois de Jeffrey

Calculer la loi non informative de Jeffrey pour les modèles suivants :

- 1) Poisson : $\{\mathcal{P}(\theta), \theta \in \mathbb{R}^+\}$
- 2) Gaussien : $\{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$ dans les deux situations suivantes
 - a. $\theta = \mu$ et σ^2 est connu.
 - b. $\theta = \sigma$ et μ est connu.
 - c. les deux paramètres sont inconnus : $\theta = (\mu, \sigma)$.

Exercice 10

On dispose de n observations $N_1, \dots, N_n$. La i^{e} observation N_i est égale au nombre de pièces défectueuses dans le i^{e} lot. Un lot est constitué de 50 pièces. On veut estimer la probabilité p de produire une pièce défectueuse.

- 1) Décrire le modèle statistique.
- 2) On suppose que la loi a priori sur p est une loi beta de paramètre (a, b) . Quelle est la loi a posteriori ?
- 3) Le service qualité fournit les informations suivantes
 - la proportion de pièces défectueuses est en moyenne 0.15
 - la proportion de pièces $X_1, \dots, X_n$ défectueuses appartient à l'intervalle $[0.1, 0.2]$ avec une probabilité de 95%

Utiliser ces informations pour fixer les paramètres de la loi a priori conjuguée.

- 4) Calculer la loi a posteriori et l'estimateur de Bayes associé au coût quadratique.
- 5) **Situation non informative :** Calculer la loi a priori de Jeffrey pour ce modèle.
- 6) Calculer la loi a posteriori et l'estimateur de Bayes.

Simulation et estimateurs de Monte Carlo

Exercice 1. Algorithme SIR : choix de m en fonction de n

On cherche à simuler par la méthode SIR un échantillon de taille n suivant la loi Gamma $\Gamma(1.5, 1)$ à partir d'un échantillon de taille m suivant la loi exponentielle de paramètre 1

1. Programmer l'algorithme SIR.

Indication : utiliser la fonction `sample` pour l'étape de rééchantillonnage.

On cherche à calibrer m en fonction de n .

2. Pour évaluer les performances de cet algorithme, on teste l'ajustement à la loi de student par le test de Komogorov de niveau 5% pour $N = 500$ échantillons simulés de façon indépendante. Comparer la proportion d'échantillons rejetés avec l'erreur de première espèce 5% dans les situations suivantes

n	50	500	1000
-----	----	-----	------

et

m/n	2	5	10	50	100
-------	---	---	----	----	-----

3. Pourrait on remplacer l'algorithme SIR par l'algorithme d'acceptation rejet ?

Exercice 2. Importance Sampling / Échantillonnage Pondéré

On cherche à estimer l'intégrale

$$I = \int h(x)f(x) dx$$

où $h(x) = x^{10} \mathbb{I}_{\mathbb{R}^+}(x)$ et f est la densité de la loi gaussienne standard.

- 1) Représenter l'estimateur de Monte Carlo en fonction de n , ajouter la région de confiance construite à partir de N trajectoires indépendantes.

On cherche à améliorer l'estimateur de Monte Carlo à l'aide de la méthode d'"importance sampling" On utilise comme loi instrumentale, une loi gamma de paramètre $(11, b)$.

- 2) Représenter l'estimateur $I_n^{IS}(b)$ en fonction de n pour différentes valeurs de b , ajouter une région de confiance.
- 3) même question avec l'estimateur d' "importance sampling normalisé" $\hat{I}_n^{NIS}$
- 4) Comparer les résultats obtenus.

Exercice 3. Approximation des estimateurs de Bayes par les méthodes de Monte Carlo

Soit $X_1, \dots, X_n$ iid suivant une loi gaussienne $\mathcal{N}(\theta, 1)$. Le choix de la loi a priori sur θ est la loi de Cauchy

$$\pi(\theta) = \frac{1}{\pi} \frac{1}{1 + \theta^2}$$

- 1) Tester l'hypothèse : "la distribution des observations est gaussienne" sur les données suivantes

<http://www.math.sciences.univ-nantes.fr/~philippe/lecture/data-gauss>

puis tester l'hypothèse

H_0 La variance est égale à 1

- 2) Calculer la loi a posteriori.
- 3) Peut on calculer explicitement l'estimateur de Bayes de θ ?
- 4) Proposer et implémenter un algorithme d'acceptation rejet pour simuler suivant la loi a posteriori. (Indication : prendre la loi gaussienne de moyenne $\bar{X}_n$ et de variance $1/n$ comme loi instrumentale)
- 5) À partir d'échantillons simulés via l'algorithme d'acceptation rejet
 - Construire une approximation de la densité de la loi a posteriori
 - Donner une approximation de l'estimateur de Bayes. Estimer la précision de votre approximation.
 - Donner une approximation de la région de confiance bayésienne de niveau 95%.
- 6) En utilisant la même loi instrumentale que pour l'algorithme d'acceptation rejet, mettre en oeuvre un algorithme SIR pour générer des nombres aléatoires suivant la loi a posteriori
- 7) À partir d'échantillons simulés via l'algorithme SIR
 - Construire une approximation de la densité de la loi a posteriori
 - Donner une approximation de l'estimateur de Bayes. Estimer la précision de votre approximation.
 - Donner une approximation de la région de confiance bayésienne de niveau 95%.
- 8) A partir de l'échantillon simuler suivant la loi instrumentale, calculer une approximation de l'estimateur de Bayes par la méthode d'échantillonnage pondéré.
- 9) Comparer les résultats obtenus et conclure.

Exercice 4

On dispose de 240 observations suivant une loi de Poisson de paramètre λ qui ont été regroupées de la façon suivante :

valeurs observées j	0	1	2	$[3, +\infty[$	Total
nombre d'observations n_j	11	37	64	128	240

On suppose que la loi a priori sur λ est la loi a priori impropre $\pi(\lambda) \propto \lambda^{-1} \mathbb{I}_{\mathbb{R}^+}(\lambda)$.

- 1) Montrer que la loi a posteriori est bien définie et proportionnelle à

$$\pi(\lambda | n_0, \dots, n_3) \propto e^{-\lambda(n_0 + n_1 + n_2)} \lambda^{n_1 + 2n_2 - 1} [1 - e^{-\lambda}(1 + \lambda + \lambda^2/2)]^{n_3} \mathbb{I}_{\mathbb{R}^+}(\lambda)$$

- 2) Écrire une fonction R qui calcule la loi a posteriori à une constante multiplicative près.
- 3) Utiliser un algorithme SIR pour construire un générateur de nombres aléatoires suivant la loi a posteriori (Indication prendre une loi gamma $(n_1 + 2n_2, n_0 + n_1 + n_2)$ comme loi instrumentale)
- 4) À partir d'un échantillon simulé suivant la loi a posteriori, calculer une approximation de l'estimateur de Bayes de λ , puis une approximation de l'intervalle de confiance de niveau 95%.
- 5) Construire une deuxième approximation de l'estimateur de Bayes de λ par la méthode d'échantillonnage pondéré en prenant comme loi instrumentale, la loi gamma utilisée à la question 2.
- 6) Comparer les deux approximations de l'estimateur de Bayes de λ .

CHAPITRE 4

Algorithme MCMC : programmation avec JAGS

Exercice 1. Suite de l'exercice 3 du chapitre 3

Soit $X_1, \dots, X_n$ iid suivant une loi gaussienne $\mathcal{N}(\theta, 1)$. Le choix de la loi a priori sur θ est la loi de Cauchy

$$\pi(\theta) = \frac{1}{\pi} \frac{1}{1 + \theta^2}$$

- 1) Récupérer les données suivantes

`http://www.math.sciences.univ-nantes.fr/~philippe/lecture/data-gauss`

- 2) Écrire le modèle sous le format JAGS
- 3) Simuler une chaîne de Markov dont la loi stationnaire est la loi a posteriori
- 4) Représenter graphiquement la chaîne de Markov et l'évolution de la moyenne en fonction du nombre d'itérations.
- 5) Calculer une approximation de la loi a posteriori
- 6) Calculer une approximation de l'estimateur de bayes sous cout L^2 , puis L^1
- 7) Calculer une approximation de la région de confiance bayésienne à 95 %, puis de la région hpd.
- 8) Vérifier la convergence de l'algo MCMC en utilisant le critère de Gelman et Rubin.

Exercice 2. Regression non linéaire

La part chauffage dans la consommation d'électricité s'écrit de la forme

$$Conso_t = f_t(Temp_t) + \varepsilon_t$$

avec

$$f_t(Temp_t) = (\alpha - \beta 1_{we}(t)) (u - Temp_t)^+$$

où

- les données sont journalières,
- 1_{we} vaut 0 du lundi au vendredi et 1 les jours de week-end, samedi et dimanche,
- les (ε_t) sont des variables indépendantes identiquement distribuées gaussiennes centrées et de variance inconnue σ^2 .
- $(u - Temp_t)^+ = \max(0, u - Temp_t)$ vaut zéro lorsque la température dépasse le seuil u

Définition de la loi a priori : les variables sont indépendantes et les lois marginales vérifient

- Les lois a priori sur α et β ont une variance très grande : $\Gamma(0.04, 0.01)$ et $\Gamma(0.01, 0.01)$
- La loi sur σ^2 est la loi impropre $1/\sigma^2$
- La loi sur u est faiblement informative $\mathcal{N}(15, 1)$

- 1) Récupérer les données suivantes

`http://www.math.sciences.univ-nantes.fr/~philippe/lecture/edf`
`load("edf")`

- 2) À partir de quelques représentations graphiques, justifier le choix du modèle. (Indication : tracer la consommation en fonction de la température)
- 3) Écrire le modèle sous le format JAGS

- 4) Simuler une chaîne de Markov dont la loi stationnaire est la loi a posteriori
- 5) Calculer une approximation de la loi a posteriori
- 6) Calculer une approximation de l'estimateur de bayes sous coût L^2 , puis L^1
- 7) Calculer une approximation de la région de confiance bayésienne à 95 %, puis de la région hpd.
- 8) Vérifier la convergence de l'algo MCMC en utilisant le critère de Gelman et Rubin.
- 9) On suppose maintenant que l'on dispose d'une information a priori sur u : u suit une loi normale $\mathcal{N}(15, 1/10)$
Reprendre les questions précédentes et comparer les résultats.

CHAPITRE 5

Bootstrap

Exercice 1

Soit $X_1, \dots, X_n$ des variables aléatoires indépendantes et identiquement distribuées suivant la loi uniforme sur $[0, \theta]$. Dans cet exercice, on admet que $M_n = \max(X_1, \dots, X_n)$ vérifie les propriétés suivantes

- M_n est l'estimateur du maximum de vraisemblance de θ .
- M_n converge presque sûrement vers θ .

On veut approcher la loi de $M_n - \theta$ par la méthode du bootstrap. On note $(X_1^*, \dots, X_n^*)$ l'échantillon bootstrapé, conditionnellement à $(X_1, \dots, X_n)$, les variables $X_1^*, \dots, X_n^*$ sont iid suivant la loi uniforme sur $\{X_1, \dots, X_n\}$. On pose $M_n^* = \max(X_1^*, \dots, X_n^*)$

- 1) Calculer $G_n(x) = P(M_n - \theta \leq x)$ pour tout $x \in \mathbb{R}$.
- 2) Quel est l'estimateur par bootstrap de $G_n(x)$?
- 3) Montrer $\pi_n = P(M_n^* - M_n \geq 0 | X_1, \dots, X_n) = 1 - (1 - n^{-1})^n$. Calculer la limite de π_n quand $n \rightarrow \infty$.
- 4) Calculer la probabilité $P(M_n - \theta \geq 0)$.
- 5) En déduire le bootstrap ne fournit pas un estimateur consistant de G_n .

On considère maintenant le bootstrap paramétrique.

- 6) Quelle est la loi conditionnelle de l'échantillon bootstrapé $(X_1^{**}, \dots, X_n^{**})$ sachant $(X_1, \dots, X_n)$.
- 7) Montrer que l'estimateur de G_n par bootstrap paramétrique est

$$G_n^{**}(x) = \left\{ \frac{x + M_n}{M_n} \right\}^n \mathbb{I}_{[-M_n, 0[}(x) + \mathbb{I}_{\mathbb{R}+}(x)$$

- 8) Comparer G_n et G_n^{**} . Le bootstrap paramétrique fournit-il un estimateur consistant de G_n ?

Exercice 2. Modèle exponentiel

On dispose de $n = 20$ données simulées suivant une loi exponentielle $1/\theta$

$$f(x) = \frac{1}{\theta} e^{-x/\theta} \mathbb{I}_{\mathbb{R}+}(x)$$

avec $\theta = 2$. On note

- F la fonction de répartition
- F_n la fonction de répartition empirique.

On estime la moyenne $\theta = \theta(F) = \int x dF(x)$ par

$$T(\underline{X}) = \hat{\theta}_n = \bar{X}_n.$$

Partie A : Estimation du biais et de la variance

- 1) Mettre en oeuvre la méthode du bootstrap (simuler B échantillons suivant la loi F_n), puis estimer la variance et le biais de l'estimateur $\hat{\theta}_n$ à partir de l'échantillon bootstrapé $\hat{\theta}_n^{*(i)}$ pour $i = 1, \dots, B$. Représenter l'évolution des estimateurs de la variance et du biais de $\hat{\theta}_n$ en fonction de B et comparer avec les valeurs théoriques.
- 2) Reprendre la question précédente en utilisant le bootstrap paramétrique.

- 3) Comparer les deux méthodes. A partir de ces résultats, proposer un choix de B (que vous conserverez dans la suite de l'exercice).

Partie B : Estimation de la loi de $\hat{\theta}_n - \theta$

- 1) A partir des échantillons bootstrapés (obtenus par bootstrap non paramétrique et bootstrap paramétrique), représenter une estimation de la densité de la loi de $\hat{\theta}_n - \theta$.
- 2) Ajouter la densité théorique de $\hat{\theta}_n - \theta$

$$x \mapsto \gamma_{n,n/\theta}(x + \theta)$$

où $\gamma_{n,n/\theta}$ est la densité de la loi gamma de paramètres $(n, \frac{n}{\theta})$ (à justifier)

- 3) Comparer avec l'approximation gaussienne de moyenne 0 et variance $\frac{1}{n}\hat{\theta}_n^2$ (Justifier cette approximation)
- 4) Commenter.

Partie C : Intervalle de confiance de niveau 90 %

Dans cette partie on considère uniquement le bootstrap non paramétrique (basé sur F_n).

- 1) A partir de l'échantillon bootstrapé $\hat{\theta}_n^{*(i)}$ pour $i = 1, \dots, B$. Construire un intervalle de confiance pour le paramètre θ .

$$\text{On pose } S = \sqrt{n} \frac{\hat{\theta}_n - \theta}{\hat{\theta}_n}$$

- 2) Approcher les quantiles de S par le bootstrap. Puis en déduire un intervalle de confiance de niveau 90%
- 3) Approcher la loi de S par une loi gaussienne. Puis déduire un intervalle de confiance de niveau 90%
- 4) Le niveau des intervalles proposés est asymptotique. Pour évaluer le niveau exact, simuler $N = 500$ échantillons de tailles $n = 20$ et de paramètre $\theta = 2$. Pour chaque échantillon calculer les 3 intervalles de confiance, puis évaluer la proportion des intervalles qui contient la vraie valeur du paramètre.
- 5) Comparer et commenter les résultats obtenus. Que se passe-t'il quand n et θ varient ?

Partie D Données non identiquement distribuées

On modélise le temps T de fonctionnement d'un réseau électrique après un blackout par une loi exponentielle de paramètre $\frac{1}{\theta_n}$ où n est le nombre de lignes impactée par le blackout.

On observe les données associées à 20 blackouts : (n_i, T_i) pour $i = 1 \dots 20$

- les n_i sont déterministes, observés
- On suppose que les T_i sont indépendants

- 1) Récupérer les données suivantes

```
http://www.math.sciences.univ-nantes.fr/~philippe/lecture/blackouts.txt
#matrice à 2 colonnes 1er colonne n et 2nd colonne T
a = matrix(scan("blackouts.txt"),ncol=2, byrow=T)
```

- 2) Vérifier que l'estimateur du maximum de vraisemblance est égale à $\frac{1}{20} \sum_{i=1}^{20} \frac{T_i}{n_i}$
- 3) On veut construire un intervalle de confiance par bootstrap pour la paramètre θ . Proposer une procédure de bootstrap en utilisant les variables $R_i = \frac{T_i}{n_i}$ pour $i = 1 \dots 20$