Méthodes de Monte Carlo

Travaux dirigés

Pierre Gloaguen

Contents

Références pour la simulation de loi sous `R`

`R` dispose d'un ensemble de fonctions pour générer les lois usuelles (multinomiale avec `sample`, loi uniforme avec `runif`, loi normale avec `rnorm`, etc ...).

En plus de l'aide de ces fonctions (`help(rnorm)`, par exemple), on pourra se referer à la partie 5 du polycopié de Christophe Chesneau.

1 Première implémentation

On cherche à évaluer la valeur de l'intégrale suivante:

$$I = \int_{\mathbb{R}^2} \cos^2(x) \sin^2(3y) \exp(-(x^2 + y^2)) \mathrm{d}x\mathrm{d}y$$

1. Ecrire un estimateur de Monte Carlo, noté $\hat{I}_M$ (où M est l'effort Monte Carlo) pour cette intégrale.
2. À l'aide du logiciel `R`, donnez une estimation de la valeur de cette intégrale pour un effort de Monte Carlo $M = 10000$. Pour simuler une loi normale sous `R`, vous utiliserez la fonction `rnorm` (voir `help(rnorm)`).
3. Quelle est la variance de $\hat{I}_1$? À l'aide des simulations obtenues précedemment, obtenez une estimation de cette variance. Servez vous de cette estimation pour calculer un intervalle de confiance asymptotique à 95% pour I.
4. Représentez graphiquement l'évolution de votre estimation en fonction de M ainsi que l'intervalle de confiance associé.

2 Aiguille de Buffon

Au XVIIIe siècle, naturaliste Georges Louis Leclerc de Buffon pose le problème suivant:

On considère un parquet avec une infinité de lattes de longueurs infinies, toutes de largeur 1. On considère ensuite l'expérience suivante: On jette une aiguille de longueur 1 en l'air, qui retombe ensuite sur le parquet. On cherche alors à calculer la probabilité que l'aiguille croise le bord d'une des lattes.

Le centre de l'aiguille tombant toujours entre deux lattes, on notera X la variable aléatoire correspondant à son ordonnée (on visualisera les lattes comme disposées "horizontalement"), comprise entre 0 et 1.

On notera θ l'angle formé par l'aiguille avec l'horizontale. θ est donc compris entre 0 et $\frac{\pi}{2}$.

On suppose que X et θ sont deux variables aléatoires indépendantes distribuées selon des lois uniformes sur $[0,1]$ et $[0, \frac{\pi}{2}]$ respectivement.

1. Montrer que la probabilité qu'une aiguille croise une latte dans ces conditions est de $\frac{2}{\pi}$.
2. Proposer un estimateur Monte Carlo de cette probabilité.
3. En déduire un estimateur de Monte Carlo de la valeur de π.
4. Donner un intervalle de confiance asymptotique à 95% pour cet estimateur.
5. Sur `R`, tracez, en fonction du nombres de simulation de Monte Carlo, l'estimation de π trouvée.
6. La vitesse d'approximation de π vous semble t'elle bonne?

3 Une comparaison avec l'intégration numérique

Cet exercice est une adaptation de l'exercice 1.2 de ce cours en ligne.

On se place dans l'hypercube unitaire de dimension d, autrement dit, l'espace $[0,1]^d$, pour $d \geq 2$.

Soit $0 < \varepsilon < 1/2$, on s'intéresse à évaluer le volume d'une sous région de cette hypercube, à savoir:

$$A_{\varepsilon,d} \cap B_{\varepsilon,d}$$

où

- $A_{\varepsilon,d}$ est l'ensemble des points du cube étant à une distance du bord plus petite que ε. Formellement:
$$A_{\varepsilon,d} = \left\{ x \in [0,1]^d, \min_{1 \leq j \leq d} \min(x_j, 1 - x_j) < \epsilon \right\}$$
- $B_{\varepsilon,d}$ est l'ensemble des points du cube étant à une distance de l'hyperplan $\left\{ x \in \in [0,1]^d, \sum_{j=1}^d x_j = \frac{d}{2} \right\}$ plus petite que ε. Formellement:
$$B_{\varepsilon,d} = \left\{ x \in [0,1]^d, \frac{1}{\sqrt{d}} \left| \sum_{j=1}^d (x_j - \frac{1}{2}) \right| < \epsilon \right\}$$

1. Justifier que le volume considéré grandit avec d. *On pourra justifier que le premier volume tende vers 1 quand $d \to \infty$ et que le second se stabilise vers une valeur finie. L'argument pour le premier volume est purement géométrique, l'argument pour le second peut se déduire du TCL.*
2. Ecrire le volume recherché sous forme d'une intégrale. En déduire un estimateur Monte Carlo de ce volume.
3. Donner une estimation de ce volume pour $\varepsilon = 0.1$ et $d = 2, 5, 10, 20$. Vous choisirez vous même l'effort de Monte Carlo, en justifiant ce choix. Donnez l'incertitude associée à votre estimation.
4. À l'aide de la fonction `hcubature` du package `cubature`, donnez une valeur du volume obtenue par approximation numérique pour les mêmes valeurs de d.
5. Comparez les résultats et commentez.

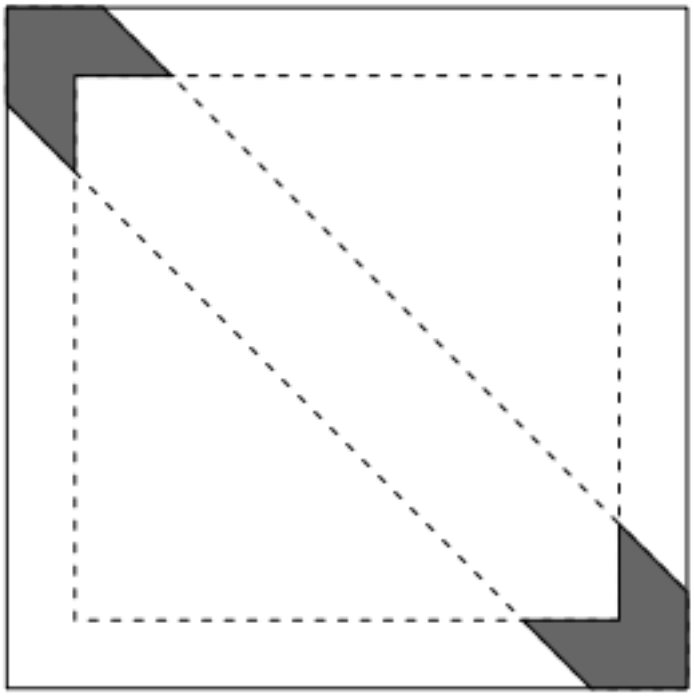

Figure 1: La surface grisée est la surface dont on cherche le volume (ici, $d = 2$). L'hyperplan définit dans le texte est ici donné par la droite $x_2 = 1 - x_1$.

4 Cas des évènements rares

On se propose d'étudier l'erreur relative de l'estimateur de Monte Carlo de la probabilité p d'un événement E $(0 < p \leq 1)$, en fonction de la valeur de p.

On se place dans le cas où pour estimer p, on simule n variables aléatoires indépendantes $X_1, \ldots, X_n$ de loi de Bernouilli de paramètre p.

L'estimateur de Monte Carlo de p est donné par

$$\hat{p} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

On s'intéresse à l'erreur relative de $\hat{p}$, à savoir la quantité:

$$\Delta_p = \frac{\hat{p} - p}{p}$$

1. Calculer la variance de Δ_p.
2. Pour $0 < \alpha < 1$, exprimer $\mathbb{P}(|\Delta_p| > \alpha)$ exactement en fonction de la loi d'une variable aléatoire binomiale de paramètres (n, p). Que pouvez vous conjecturer sur cette probabilité quand p devient petit?
3. En utilisant le théorème central limite, donner une expression asymptotique de cette probabilité basée sur la fonction de répartition de la loi normale centrée réduite.

4.1 Attention à la dimension!

4. On peut montrer que le volume d'une sphère de rayon 1 en dimension $d \geq 2$ est donné par la fonction:

$$V(d) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$$

où, pour $z > 0$

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

On se propose d'estimer la valeur de π en tirant, en dimension d, une U variable uniforme dans l'hypercube $[-1,1]^d$. On pose alors $X = \mathbf{1}_{\|U\|^2 \leq 1}$, sur un échantillon de taille $n = 10000$.

a. Quelle est la valeur de p, le paramètre de la loi de Bernouilli de X?
b. Donner alors l'estimateur de π en fonction de l'estimateur de p.
c. Discutez la qualité de l'estimateur quand d grandit. Vous pourrez vous aidez de R pour voir le comportement de la fonction V_d (en pourra utiliser la fonction `gamma` dans R).

5 Détection d'aggrégats dans une série temporelle

5.1 Présentation du problème

On s'intéresse à une série temporelle à valeurs dans $\mathbb{R}$. Ainsi, les données consistent en un vecteur $X_{1:n} = (X_1, \dots X_n)$ de valeurs ordonnées dans le temps.

La question est la suivante: *Existe-t-il une fenêtre temporelle de valeurs anormalement élevées?*.

Pour cela, on se propose de faire le test

- H_0: Les variables aléatoires $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées.
- H_1: Il existe une fenêtre temporelle où les valeurs de la série sont plus importantes.

Pour tester cette hypothèse, pour une série temporelle $X_{1:n} = (X_1, \dots, X_n)$, on va définir une statistique de test $T(X_{1:n})$.

Pour l'échantillon aléatoire $X_1, \dots X_n$, on note R_k le rang de X_k parmi les valeurs de l'échantillon (il est égal à 1 si X_k est la valeur la plus faible, à n si X_k est la valeur la plus élevée). Comme on considère des variables aléatoires continues, on considère dans la suite que deux rangs ne peuvent pas être égaux. **Vous remarquerez que l'hypothèse H0 ne fait pas d'hypothèse sur la distribution des valeurs observées, en effet, H0 fait une hypothèse sur la distribution jointe des rangs**.

1. Justifier que, sous H_0, la loi de R_k est une loi uniforme discrète sur $\{1, \dots, n\}$. Quelle est la loi de R_k sachant R_ℓ ($\ell \neq k$)?

Pour tout couple (i,j) tel que $1 \leq i \leq j \leq n$ on considère la variable aléatoire suivante:

$$S(i,j) = \sum_{k=i}^{j} R_k.$$

2. Que représente cette variable aléatoire? Dans quel cas prendra t'elle des grandes valeurs?
3. Montrer que, sous H_0, $m_{ij} := \mathbb{E}[S(i,j)] = \frac{1}{2}(n+1)(j-i+1)$ pour tout couple (i,j).
4. Calculer, sous H_0, $v_{ij} := \mathbb{V}[S(i,j)] = \frac{1}{12}(n+1)(j-i+1)(n-j+i-1)$ pour tout couple (i,j).

On définit maintenant la variable aléatoire centrée et réduite, pour tout couple d'entiers (i, j) tel que $1 \leq i \leq j \leq n$.

$$T(i,j) = \begin{cases} 0 & \text{si } i = 1 \text{ et } j = n \\ \frac{S(i,j) - m_{ij}}{\sqrt{v_{ij}}} & \text{sinon.} \end{cases}$$

Notre statistique de test $T_n(X_{1:n})$ sera donc donnée par

$$T_n(X_{1:n}) = \max_{1 \leq i \leq j \leq n} T(i,j). \tag{1}$$

5.2 Principe du test et prise de décision par méthode de Monte Carlo.

Le principe du test est le suivant: pour un échantillon observé $\mathbf{x}$ un risque α, on rejette H_0 si $T_n(\mathbf{x}) > t_{1-\alpha}$ où $t_{1-\alpha}$ est le quantile d'ordre α de la loi de $T_n(\mathbf{X})$. On concluera que la fenêtre temporelle pour laquelle la statistique est calculée (soit $(i_{\max}, j_{\max}) = \text{argmax}_{i,j}\ T(i,j)$) est anormalement loin de 0 sous H_0. On rejettera alors H_0 pour conclure a un aggrégat de valeurs élevées sur cette fenêtre.

5. La loi de T_n sous H_0 étant inconnue, on se propose d'approcher ses quantiles sous H_0 par méthode de Monte Carlo. Donner un algorithme simple de simulation de T_n sous H_0.

6. Proposer une méthode de Monte Carlo pour répondre à la question initiale à un risque α fixé, pour n'importe quelle série temporelle observée $x_{1:n}$.

5.3 Implémentation sous R pour les températures à Hobart, Tasmanie.

7. Ecrire une fonction `get_tn`, qui pour une série temporelle $x_{1:n}$ donnée, calcule $T_n(x_{1:n})$ et, si on le demande, renvoit les indices temporels de la fenêtre sur laquelle cette statistique est obtenue. Calculer cette statistique de test pour la série des températures à Hobart. On notera cette valeur t^*

8. Ecrire une fonction `get_h0_sample` qui, pour un entier n et un entier M permet d'obtenir M réalisations de T_n sous H_0.

9. Simuler un M échantillon de T_n sous H_0 pour une valeur de n correspondant à celles des données d'Hobart. Vous prendrez $M = 5000$. Représenter l'estimation obtenue de $\mathbb{P}(T_n > t^*)$ ainsi que son intervalle de confiance asymptotique à 95%.

10. Répondre à la question initiale sur les températures à Hobart

Echantillonnage préférentiel

Travaux dirigés

Pierre Gloaguen

Contents

1 Évènement rare

On veut estimer la probabilité p^* qu'une loi normale centrée réduite dépasse la valeur 3.

1. Pour un effort de Monte Carlo de taille M, proposer un estimateur de Monte Carlo pour p^* se basant sur la loi $\mathcal{N}(0, 1)$.
2. Implémenter cet estimateur sur `R` pour un effort de Monte Carlo de taille $M = 10000$
3. On se propose d'utiliser un échantillonnage préférentiel pour estimer cette probabilité. On utilisera comme loi d'échantillonnage une loi exponentielle translatée de 3, de paramètre λ notée $Y \sim t\mathcal{E}(3, \lambda)$, i.e., la variable aléatoire Y telle que $Y - 3 \sim \mathcal{E}(\lambda)$ Calculer les poids d'importance associés, et proposer un estimateur de p^*
4. Ecrire la variance de l'estimateur comme fonction de λ. Comment doit on choisir λ?
5. Donner un estimateur empirique de cette variance, et donner l'expression de l'intervalle de confiance à 95% pour l'estimateur de p^*.
6. Implémenter cet estimateur sur `R` avec $\lambda = 3.5$ et le comparer à celui de Monte Carlo.

2 Cas problématique

2.1 Premier cas jouet

Afin de montrer qu'un mauvais choix de loi d'échantillonnage peut augmenter drastiquement la variance de l'estimateur, on peut prendre un exemple très simple.

Supposons qu'on veuille estimer par méthode de Monte Carlo $\mathbb{E}[X]$ où $X \sim \mathcal{N}(0, 1)$. On se propose de le faire par méthode de Monte Carlo standard et par échantillonnage préférentiel en tirant dans une loi $Y \sim \mathcal{N}(\mu, 1)$ où μ est choisi par l'utilisateur.

1. Quelle est la variance de l'estimateur de Monte Carlo standard?
2. Quelle est la variance de l'estimateur par échantillonnage préférentiel?

2.2 Encore pire!

Soit X une variable aléatoire de densité $f_X(x) = 3x^{-4}\mathbf{1}_{x\geq 1}$ (on dit que X suit une loi de Pareto de paramètres (1, 3)).

1. Calculer $p = \mathbb{P}(X > 10)$
2. Supposons qu'on veuille estimer p par méthode de Monte Carlo. Proposer une méthode de simulation de X par méthode d'inversion. En déduire un estimateur de Monte Carlo standard.
3. Représenter graphiquement les performances empiriques de cet estimateur pour un effot Monte Carlo de $M = 10^4$.
4. On propose maintenant un estimateur par échantillonnage préférentiel comme dans l'exercice précédent. On choisit comme densité d'échantillonnage celle d'une loi exponentielle translatée de 10, de paramètre $\lambda = 1$. Donnez l'estimateur associé et mettez le en place sous R. Que constatez vous? Comment l'expliquez vous.
5. Que pouvez vous dire de la variance de cet estimateur?

3 Marche aléatoire du joueur optimiste

Le problème suivant peut être vu par le prisme d'un joueur dans un jeu à espérance négative, qui décide *a priori* de s'arrêter, soit quand il est ruiné, soit quand ses gains ont dépassé le jackpot.

Soit $X_0 = 0$ et $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi $\mathcal{N}(-\mu, 1)$ où μ est une constante strictement positive (gain pour une partie). On note $S_n = \sum_0^n X_n$ (somme des gains jusqu'à l'instant n). On se donne un réel $r < 0$ (représentant la ruine) et un réel $j > 0$ (représentant le jackpot). On considère le temps d'arrêt $T = \min\{n, S_n \leq r \text{ ou } S_n \geq j\}$. On souhaite estimer la probabilité de sortir vainqueur, soit $p^* = \mathbb{P}(S_T \geq j)$.

1. Proposer un algortihme de simulation de pour une trajectoire $(S_n)_{1\leq n\leq T}$.
2. Pour un effort de Monte Carlo de taille M, proposer un estimateur pour p^*.
3. Pour $\mu = 1, r = -50, j = 5$, implémenter cet estimateur, quelle est la valeur de $\hat{p}^*$ obtenue pour $M = 1000$? Représentez une trajectoire d'une estimation (i.e., la valeur de l'estimation en fonction de M)?
4. Pour une séquence observée $s_1, \ldots, s_T$ simulée grâce à la méthode de Monte Carlo ci dessus, donner la densité de l'échantillon, notée $f(s_{1:T})$.
5. On considère maintenant la marche aléatoire où les X_n sont tirés selon la loi $\mathcal{N}(\mu, 1)$ (n a toujours $X_0 = 0$ et $S_n = \sum_0^n X_n$). Pour une séquence $s_1, \ldots, s_T$ simulée ainsi on note $g(s_1, \ldots, s_T)$, la densité de l'échantillon, écrire le rapport $\frac{f(s_{1:T})}{g(s_{1:T})}$.
6. Proposer un estimateur par échantillonnage préférentiel pour estimer p^*, basé sur la simulation selon la densité g. L'estimateur sera noté $\hat{p}_M^{IS}$. De cet estimateur, vous déduirez que $p^* \leq \mathrm{e}^{-2\mu j}$.
7. Implémenter cet algorithme, tracez la valeur de l'estimation ainsi que son intervalle de confiance en fonction de M.

Simulations de variables aléatoires

Travaux dirigés

Pierre Gloaguen

La plupart des exercices de cette feuille nécessite la confection de programme en `R`.

Afin de garder trace de vos exercices, pensez à sauvegarder le script associé, voire à répondre à l'exercice dans un fichier *Rmarkdown* (extension `.Rmd`).

L'environnement `Rstudio` est plus que vivement conseillé pour programmer en `R`.

1 Générateurs pseudos aléatoires

1.1 Génération de loi uniforme

1. À l'aide du logiciel R, programmez un générateur à congruences pour la loi uniforme. Ce générateur prendra la forme d'une fonction prenant en argument:

- Un entier `n` donnant la taille de l'échantillon voulu.
- 4 entiers `a, m, c, x0` correspondant aux paramètres du générateurs vu en cours.

2. À l'aide de cette fonction, générer un échantillon de taille 10000 pour les valeurs

- `a = 41358`
- `m = 2^31 -1`
- `c = 0`

et la graine de votre choix. Refaites la même procédure avec

- `a = 3`
- `m = 2^31 -1`
- `c = 0`

et

- `a = 101`
- `m = 2311`
- `c = 0`

Vous stockerez chacun des échantillons obtenus

3. Pour chacun des échantillons obtenus, tracez l'histogramme empirique. Quels échantillons vous semblent tirés selon une loi uniforme $U[0, 1]$? En utilisant la fonction `ks.test`, effectuez un test de Kolmogorov-Smirnoff d'adéquation pour la loi uniforme. Que concluez vous sur la qualité des 3 générateurs?

4. Pour chacun des échantillons $(u_1, \dots, u_{10000})$ obtenus, tracez u_n en fonction de u_{n-1}. Que pouvez vous conclure sur la qualité des 3 générateurs?

2 Méthode d'inversion

2.1 Loi exponentielle

On rappelle qu'une variable aléatoire X est de loi exponentielle de paramètre $\lambda > 0$ si elle a pour fonction de densité $f_X(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{x \geq 0}$

1. En utilisant la méthode d'inversion, proposez un algorithme de simulation pour une variable aléatoire exponentielle.
2. Ecrire une fonction `R` mettant en oeuvre cette algorithme. Cette fonction prendra deux paramètres en entrée:

- `n` La taille de l'échantillon;
- `lambda` Le paramètre de la loi exponentielle

Vous testerez la qualité de votre fonction sur un échantillon de taille 10000, en comparant graphiquement l'histogramme empirique obtenu à la densité de la loi exponentielle correspondante.

2.2 Loi discrète

On considère une variable aléatoire discrète X à valeurs dans l'ensemble $\{1, \ldots, K\}$, dont la loi est définie par le vecteur de probabilité $(p_1, \ldots, p_K)$, i.e.:

$$\mathbb{P}(X = k) = p_k \tag{1}$$

$$\sum_{k=1}^{K} p_k = 1 \tag{2}$$

1. Pour tout $u \in]0, 1[$, écrire l'expression de l'inverse généralisée de la fonction de répartition de X.
2. En déduire un algorithme de simulation pour toute variable aléatoire discrète dans un ensemble fini.
3. Utilisez cet algorithme de simulation pour simuler un échantillon de taille 10000 loi binomiale de paramètres $n = 10$ et $p = 0.5$ avec `R`. Vous comparerez les fréquences obtenues avec les fréquences théoriques.

3 Méthode de rejet

3.1 Simulation d'une loi de Poisson pour $\lambda < 1$

1. On utilisant le résultat de l'exercice 2.2, proposez un algorithme pour simuler une variable aléatoire de Bernouilli de paramètre $p \in]0, 1[$.
2. En déduire un algorithme pour simuler une loi géométrique de paramètre p sur $\mathbb{N}$.
3. On souhaite obtenir un échantillon d'une loi de de Poisson de paramètre $\lambda \in]0, 1[$ par méthode d'acceptation rejet. On se propose d'utiliser comme loi de proposition la loi géométrique sur $\mathbb{N}$ de paramètre $1 - \lambda$. Définir l'algorithme de rejet correspondant.
4. Quelle est la probabilité d'acceptation dans l'algorithme de rejet?
5. Faites une fonction `R` permettant de générer une loi géométrique de paramètre p. Utiliser cette fonction dans une autre fonction `R` permettant de simuler selon une loi de Poisson de paramètre $p \in]0, 1[$. Simuler ainsi un échantillon de taille 10000. Comparer la distribution obtenue à celle de la vraie loi.

3.2 Loi uniforme sur le disque unité

1. À partir d'une variable aléatoire uniforme sur $[0,1]$, proposez une transformation pour simuler une loi uniforme sur $[-1,1]$.
2. Proposer une méthode d'acceptation rejet pour simuler, à partir de deux variables aléatoires indépendantes de loi uniforme sur $[-1,1]$, une variable aléatoire uniforme sur le disque unité.
3. Quelle est la probabilité d'acceptation de l'algorithme?
4. Ecrire une fonction `R` mettant en place la génération de variables aléatoires sur le disque unité. Dans cet algorithme, gardez en mémoire le nombre d'essais nécessaire avant chaque acceptation.
5. Générer un échantillon de taille 10000. Vérifiez graphiquement que ces points sont uniformément répartis sur le disque unité. Vérifiez également que le nombre d'essais moyens avant acceptation est en adéquation avec ce qui est attendu.

3.3 Proposition optimale

La loi normale tronquée de support $[b,+\infty[$ est définie par la densité f proportionnelle, pour tout réel x, à

$$f_1(x) = \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathbf{1}_{x\geq b}, \quad \text{avec } \mu>0,\ \sigma>0.$$

On propose de simuler suivant la loi de densité f par une méthode de rejet.

3.3.1 Méthode naïve

1. On note Φ la fonction de répartition de la loi normale centrée réduite. Montrer que f satisfait l'inégalité suivante pour tout réel x:
$$f(x) \leq \frac{1}{\sigma\sqrt{2\pi}\Phi(\frac{\mu-b}{\sigma})}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$
2. En déduire l'algorithme du rejet. Que peut-on dire du nombre d'essais moyen avant acceptation ?

3.3.2 Une distribution instrumentale alternative

On suppose que $b>\mu$. On considère la loi exponentielle translatée de b, $\tau\mathcal{E}(\lambda,b)$, de densité

$$g_\lambda(x) = \lambda e^{-\lambda(x-b)}\mathbf{1}_{x\geq b}, \quad x\in\mathbb{R}.$$

3. Montrer pour $x\geq b$ que
$$\frac{f_1(x)}{g_\lambda(x)} \leq \begin{cases} \frac{1}{\lambda}\exp\left\{\lambda(\mu-b)+\frac{(\lambda\sigma)^2}{2}\right\} & \text{si } \mu+\lambda\sigma^2>b, \\ \frac{1}{\lambda}\exp\left\{-\frac{(b-\mu)^2}{2\sigma^2}\right\} & \text{sinon.} \end{cases}$$
4. Proposer une méthode de simulation de la loi de densité f.
5. Calculer la valeur de λ^* telle que le temps moyen de calcul de la méthode proposée soit le plus petit possible.
6. En `R`, mettre en oeuvre les deux méthodes afin de constater empiriquement les différences.

4 Méthode de transformation

4.1 Simulation de lois Gaussiennes. Algorithme de Box-Muller

Soient X et Y deux variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$

1. Montrer que si U et V sont deux variables aléatoires indépendantes de loi $\mathcal{U}[0,1]$ alors le couple
$$\left(\sqrt{-2\ln(U)}\cos(2\pi V), \sqrt{-2\ln(U)}\sin(2\pi V)\right)$$
a la même loi que le couple (X,Y).
2. Ecrire une fonction `box_muller` permettant de simuler une loi $\mathcal{N}(0,1)$ en `R`. Vous comparez l'histogramme obtenu à la vrai densité de la loi.
3. En déduire, pour tout $\mu \in \mathbb{R}^{\not 2}$ et toute matrice 2×2 symmétrique semi-définie positive Σ une méthode pour simuler une variable aléatoire $Z \sim \mathcal{N}(\mu,\Sigma)$.

5 Autour de l'accepation rejet

5.1 Acceptation rejet étendu: Cas de deux fonctions positives.

Pour cette preuve, vous pourrez mimer les étapes de la preuve dans le cas usuel, détaillée dans le poly.

On se propose de montrer que pour que simuler selon une densité par algorithme d'acceptation rejet, il n'est nécessaire de connaître la densité qu'*à la constante de normalisation près.* Cette propriété est très utile dans le cas où le calcul de la constante de normalisation est coûteux, voir impossible (typiquement en statistiques Bayesiennes).

Plus formellement, soient $\tilde{f}$ une fonction positive et g une densité de probabilité, toutes deux définies sur $\mathbb{R}^d$ telles que:

- $0 < \int_{\mathbb{R}^d} \tilde{f}(x)\mathrm{d}x < \infty$. On note respectivement $I(\tilde{f})$ cette intégrale et
$$f(x) = \frac{\tilde{f}(x)}{I(\tilde{f})}$$
la densité associée à cette fonction positive.
- Il existe $M > 0$ tel que, pour tout réel x, $\tilde{f}(x) \leq Mg(x)$.

On note
$$\alpha(x) := \frac{\tilde{f}(x)}{Mg(x)}.$$
Soit $(U_m)_{m\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$. Soit $(Y_m)_{m\geq 1}$ une suite de variables aléatoires indépendantes et identiquement distributée, de densité donnée par g. On note T la variable aléatoire (à valeurs dans $\mathbb{N}^*$):
$$T = \inf\{m, \text{ tel que } U_m \leq \alpha(Y_m)\}.$$
.

1. Montrer la variable aléatoire $X := Y_T$ (T-ième valeur de la suite $(Y_m)_{m\geq 1}$) a pour densité f.
2. Donnez alors la loi de la variable aléatoire T. Quelle est l'espérance de T?
3. En déduire un estimateur de $I(\tilde{f})$ par méthode de Monte Carlo, obtenu uniquement à partir de l'algorithme d'acceptation rejet défini plus tôt.
4. Grâce au théorème central limite, donnez l'expression d'un intervalle de confiance asymptotique à 95% pour $I(\tilde{f})$, ne dépendant d'aucune quantité inconnue.

5.2 Recyclage dans l'acceptation rejet

Dans cette section, on se replace dans le cadre classique de l'acceptation rejet.

On se propose d'approcher une intégrale du type: $J = \mathbb{E}_f[\varphi(X)]$ où $f($ est la densité de la variable aléatoire X sur $\mathbb{R}^d$ selon laquelle on ne sait pas simuler, et φ est une fonction intégrable par rapport à cette densité.

À partir d'une densité $g(x)$ sur $\mathbb{R}^d$ selon laquelle on sait simuler, et telle que

$$\exists M > 0, \text{ tel que } \forall x \in \mathbb{R}^d, \ f(x) \leq Mg(x)$$

on obtient, par algorithme d'acceptation-rejet (pour un tel M fixé) un échantillon de variables aléatoires *i.i.d.* $X_1, \ldots, X_n$ de loi donnée par f.

Pour obtenir cet échantillon de taille n, on a simulé $N \geq n$ variables aléatoires indépendantes $Y_1, \ldots, Y_N$ de densité g. On note $Z_1, \ldots, Z_{N-n}$ l'échantillon i.i.d. de variables aléatoires ayant été rejetées dans l'algorithme d'acceptation rejet.

5. Donner l'expression de la densité de la variable aléatoire Z_1.

6. En déduire que
$$\hat{J}_N = \frac{1}{N}\left(\sum_{i=1}^{n} \varphi(X_i) + \sum_{j=1}^{N-n} \frac{(M-1)f(Z_j)}{Mg(Z_j) - f(Z_j)}\varphi(Z_j)\right)$$
est un estimateur sans biais de J. Quelle est l'intérêt de cette méthode selon vous?

5.3 Application

On reprend l'exemple vu en cours d'introduction à la statistique bayésienne. Vous reprendrez le même modèle ainsi que les mêmes données utilisées.

7. En utilisant le même prior que celui du cours, ainsi que le même loi de proposition g, implémentez l'algorithme d'acceptation rejet pour tirer selon le posterior. Les algorithmes efficaces seront valorisés.

8. Implémentez cette méthode, et tracez les densités empiriques des posteriors obtenus. Vous donnerez également une estimation de $\mathbb{E}[\theta|y_{1:n}]$ ainsi que l'intervalle de confiance associé.

9. À partir de cette méthode, et en utilisant les questions 3 et 4, donner une estimation ainsi qu'un intervalle de confiance pour à 95% de la quantité
$$\int_{\mathbb{R}^4} \pi(\theta)L(y_{1:n}\theta)\mathrm{d}\theta$$

10. Afin d'estimer de $\mathbb{E}[\theta|y_{1:n}]$, implémenter l'estimateur $\hat{J}_N$ de la question 6 (avec le même algorithme d'acceptation rejet que pour la question 8). Donnez un intervalle de confiance asymptotique pour l'estimation obtenue.

11. Comparez les deux estimateurs et commentez.

Inférence bayésienne et méthodes MCMC

Travaux dirigés

Pierre Gloaguen

1 Inférence bayésienne pour le modèle linéaire

Soit Y un vecteur d'observations de $\mathbb{R}^n$, β un vecteur de paramètres inconnus $\mathbb{R}^{p+1}$ (tel que $n > p+1$) et X une matrice $n \times (p+1)$ telle que la matrice X^TX soit inversible. On considère le modèle linéaire Gaussien:

$$Y = X\beta + E$$

où E est un vecteur Gaussien de loi $\mathcal{N}(0, \sigma^2 I_n)$.

1.1 Cas où σ^2 est connu

1. Dans le cas où σ^2 est connu, écrire la vraisemblance associée au modèle précédent. Montrer que cette vraisemblance est proportionelle, en tant que densité de probabilité pour le vecteur β, à la densité d'un loi $\mathcal{N}((X^TX)^{-1}X^TY, \sigma^2(X^TX)^{-1})$. En déduire la densité a posteriori sur β pour une inférence bayésienne effectuée avec un prior impropre.

1.2 Cas où σ^2 est inconnu

Dans ce cas, on pose comme loi a priori que le couple (β, σ^2) suit une loi normale inverse Gamma de paramètres $\mu \in \mathbb{R}^{p+1}$, V (une matrice de variance-covariance de taille $(p+1) \times (p+1)$, a et b (deux réels positifs).

Formellement:

$$\pi(\beta, \sigma^2 | \mu, \mathbf{V}, a, b) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{p+1}{2}} \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left(-\frac{b}{\sigma^2}\right) \exp\left(-\frac{(\beta-\mu)^T\mathbf{V}^{-1}(\beta-\mu)}{2\sigma^2}\right).$$

Remarque Cette modélisation est en fait assez naturelle, elle correspond au cas où σ^2 suit une loi inverse $\mathcal{G}amma(a, b))$ (usuelle pour les variances) et $\beta|\sigma^2 \sim \mathcal{N}(\mu, \sigma^2 V)$.

1. Montrer que la loi de $(\beta, \sigma^2)|Y, X$ suit également une loi Normale inverse Gamma dont vous préciserez les paramètres.
2. Interprétez les paramètres en terme "d'apprentissage bayésien", c'est à dire en distinguant le poids du prior et des données.

2 Modèle probit avec covariables

On reprend l'exemple vu en cours et dans l'exercice 5 du TD3 sur l'estimation de covariables corrélées à la présence d'oiseaux.

2.1 Notations et modèle

On note $y_1, \ldots, y_n$ les observations de présence (1 si on observe un oiseau, 0 sinon) sur les sites 1 à n.

On note x_{ij} la valeur de la j-ème ($1 \leq j \leq 3$) covariable sur le i-ème site.

On suppose que les $y_1, \ldots, y_n$ sont les réalisations de variables aléatoires $Y_1, \ldots, Y_n$ telles que

$Y_i \sim \mathcal{B}ern(p_i)$ où

$$p_i = \phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) = \phi(\mathbf{x}_i^T \theta)$$

où $\theta = (\beta_0, \ldots, \beta_3)^T$ et ϕ la fonction de répartition d'une $\mathcal{N}(0,1)$. L'objectif est d'estimer le vecteur θ dans un cadre bayésien.

2.1.1 Vraisemblance et posterior

1. Rappelez l'expression de la vraisemblance d'un paramètre θ pour vecteur d'observations $\mathbf{y}$ ainsi que l'expression du posterior associée à un prior $\mathcal{N}(0, 4I_4)$.

2.2 Algorithme de Metropolis Hastings

2. On se propose d'approcher la loi *a posteriori* en utilisant un algorithme MCMC. Plus précisemment, on se propose de générer une chaîne de Markov $(\theta_n)_{n\geq 0}$ dont l'unique loi stationnaire est le posterior défini plus haut. Pour cela, on utilisera un algorithme de Metropolis Hastings dont le noyau de transition est une marche aléatoire de loi normale $\mathcal{N}(0, \sigma^2 I_4)$ où I_4 est la matrice identité 4×4. Définir l'algorithme de Metropolis Hastings pour un jeu de données $\mathbf{y}$.
3. Le fichier `donnees_presence_complet.txt` contient les observations de 300 sites sur lesquels la présence d'oiseaux a été constatée, ainsi que différentes variables environnementales mesurées. Ecrire un programme `R` codant l'algorithme de Metropolis Hastings précédent pour ce jeu de données. Vous testerez plusieurs valeurs de σ^2 pour la variance de la marche aléatoire, et choisirez celle qui vous semble la meilleure.
4. Pour le σ^2 choisi quelle est la probabilité d'acceptation empirique?
5. Quelle est la valeur réalisée de l'estimateur Bayésien $\mathbb{E}[\theta|\mathbf{Y}]$?
6. Donner un intervalle de crédibilité pour chacun des paramètres.

3 Metropolis Hastings et loi de mélange

On s'intéresse à simuler à la loi d'un mélange de deux Gaussiennes:

$$f(x) = \frac{1}{2} f_Z(x-4) + \frac{1}{2} f_Z(x+4)$$

où f_Z est la densité d'une loi $\mathcal{N}(0,1)$.

1. À l'aide du logiciel `R`, tracez la densité de cette loi.
2. On se propose de construire un algorithme de Metropolis Hastings pour simuler une chaîne de Markov de loi stationnaire f à partir d'une marche aléatoire d'étendue uniforme. Plus précisemment, pour $\alpha \in \mathbb{R}_+^*$, à partir d'une position $X = x$, on définit la prochaine position $Y|X = x$ comme une variable aléatoire de densité

$$q_\alpha(x, y) = \mathbf{1}_{x-\alpha \leq y \leq x+\alpha}$$

   a. Ecrire un code `R` permettant de simuler selon cette loi.

b. Justifier que pour toute régions A et B de $\mathbb{R}$, on peut accéder de A à B en un nombre fini de pas.

3. Définir l'algorithme de Métropolis Hastings pour simuler une chaîne de Markov $(X_n)_{n\geq 0}$, de loi stationnaire f partir du noyau de Markov q_α et d'un point de départ x_0.
4. Implémentez cet algorithme en `R` pour $\alpha = 0.01$ à partir de $x_0 = 0$, pour $n = 5000$. Faites tourner cet algorithme plusieurs fois, que constatez vous?
5. Implémentez cet algorithme en `R` pour $\alpha = 5$ et à partir de $x_0 = -10$. Faites tourner cet algorithme plusieurs fois, que constatez vous?
6. Dans les deux cas, donnez la probabilité d'acceptation empirique de l'algorithme (ou le nombre d'essai moyen avant de faire un pas dans la chaîne).

4 Echantillonneur de Gibbs

Cet exercice est un exemple trivial d'implémentation (inutile ici!) de l'échantillonneur de Gibbs

On veut simuler par échantillonneur de Gibbs un échantillon de vecteur aléatoire (X, Y) distribué selon la loi $\mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

1. Donnez la loi conditionnelle de $Y|X$ et $X|Y$.
2. Implémentez un échantillonneur de Gibbs partant du point (10, 10) pour simuler selon la loi jointe de (X, Y).
3. *Burn-in:* Vérifiez empiriquement que l'algorithme converge vers la loi voulue. Quelle partie initiale de la chaîne simulée conseillez vous d'omettre?
4. *Thinning:* Regardez la fonction d'autocorrélation de l'échantillon simulé. Quelle proportion de l'échantillon préconisez vous de gardez en pratique poour avoir un échantillon qu'on pourra supposer indépendant?

5 Décryptage bayésien

```
[1] "CECI SERA LE DERNIER DEVOIR DU COURS."
```

5.1 Présentation du problème

On se place dans le cadre où on dispose d'un alphabet de taille finie, disons K. Chaque élément de l'alphabet est codé comme un nombre l'ensemble $\{1, \ldots, K\}$. On suppose qu'on dispose d'un message crypté de ce type:

```
[1] "NJNEFYJ OFXJFCJ 'EJ FCJAHE FCDFNHD YL"
```

L'objectif est de décrypter ce message, et retrouver le message original, à savoir:

```
[1] "CECI SERA LE DERNIER DEVOIR DU COURS."
```

en utilisant l'inférence bayésienne. Pour cela,

- On suppose que le message est issue d'une langue connue, disons le Français, dont on connaît certaines caractéristiques (décrites plus bas).
- On suppose que ce message est la transformation du vrai message par une permutation f_*^{-1} des éléments de l'alphabet. Ainsi, f_*^{-1} a envoyé chaque élément de la phrase initiale sur un autre élément de l'alphabet (deux éléments identiques dans la phrase de départ le sont encore dans la phrase d'arrivée).

On recherche donc la permutation f_* afin de décrypter le message. Cette inconnue vit donc dans un espace discret à $K!$ éléments.

- Pour un message X et permutation f donnée, la vraisemblance associée à f est donnée par

$$L(X|f) = \prod_{i,j=1}^{K} M(i,j)^{f_X(i,j)},$$

où

- $M(i,j)$ est le nombre de transitions $i \to j$ (pour chaque élément i et j de l'alphabet) observées *par ailleurs* dans la langue Française.
- $f_X(i,j)$ est le nombre de transitions de $i \to j$ dans la décryption du message X par f. Par exemple, pour si f est l'identité, notre message précédent présente 2 transitions "C" $\to$ "J", 0 transition "A" $\to$ "B", 3 transitions "J" $\to$ " ", 0 transition "J" $\to$ "C", etc...

On voit que cette fonction de vraisemblance est grande si la décryption de X par f présente une fréquence de transition conistante avec celle de la matrice M, connue dans la langue française.

5.2 Objectif

On veut obtenir une vision des décryptions possibles par inférence bayésienne. On suppose que la loi a priori de f est une loi uniforme sur l'ensemble des permutations possibles. Au vu d'un message X, on cherche à obtenir des échantillons tirés selon la loi a posteriori de f. Pour cela, vous implémenterez un algorithme de Metropolis Hastings dont la loi stationnaire sera donnée par cette loi a posteriori.

- C'est vous qui choisirez le(s) point(s) de départ de cet algorithme en justifiant votre démarche.
- De même, c'est vous qui choisirez le noyau de transition de l'algorithme de Metropolis Hastings utilisé.
- L'objectif est d'obtenir des tirages dans la loi a posteriori. Vous présenterez plusieurs de ces tirages et vous en servirez pour essayer de décrypter au mieux votre texte.
- La démarche devra être décrite clairement et reproductible (donc vous fournirez vos codes).

Afin de vous aider dans la démarche, vous pourrez la très complète référence: *Decrypting Classical Cipher Text Using Markov Chain Monte Carlo* de Chen et Rosenthal.

5.3 Détails techniques

L'alphabet considéré sera celui composé de toutes les lettres majuscules, sans accent, de la langue française, auxquelles s'ajouteront l'apostrophe "'", la virgule ",", le point "." et l'espace " ". On indexera ces éléments de 1 (la lettre"A") à 30 (l'espace" ").

```
# Un alphabet à 30 éléments
(my_alphabet <- c(LETTERS, "'", ",", ".", " "))
```

```
 [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N" "O" "P" "Q"
[18] "R" "S" "T" "U" "V" "W" "X" "Y" "Z" "'" "," "." " "
```

La matrice M transmise sera donc une matrice 30 par 30, indexée de la même manière. Ainsi, $M(5,29)$ comptera le nombre de transitions observées dans mon texte modèle (écrit en Français) entre le "E" et le ".".

Fatalement, l'exercice voudra que vous manipuliez des chaînes de caractères avec `R`. Vous pourrez vous aider du package `stringr` pour lequel il existe de nombreux tutoriels (dont celui ci).