Echantillonneurs de Monte Carlo par Chaines de Markov (MCMC) Application à l'Estimation Statistique Bayésienne

Paulo Gonçalves

CPE Lyon

7 janvier 2015

Objectifs

- Rappel sur l'inférence Bayésienne
- Méthodes de simulation de densité (inversion, rejection)
- MCMC : Techniques de Simulation Stochastique pour Inférence Statistique

Références Bibliographiques

- **Monte Carlo Statistical Methods**. Christian Robert, George Casella. Springer. 2004 (2eme édition)
- **Markov Chain Monte Carloin Practice**. Walter R. Gilks, Sylvia Richardson (Eds.) Chapman & Hall, 1996
- **Understanding Monte Carlo Markov Chain**. Gareth O. Roberts and Richard L. Tweedie. Springer Series in Statistics, 2004
- **Stochastic simulation : Algrithms and Analysis**. Soren Asmussen, Peter W. Glynn. Springer, 2000
- **A first course in Bayesian Statistical Methods**. P. Hoff. Springer, 2009
- **Non uniform Random variable generation**. Luc Devroye. Springer Verlag, 1986

- *Statistique bayésienne et algorithme MCMC*. Cours de M1 Inst. Math. Toulouse, J. Dupuis (LPS-UPS)
- *An introduction to MCMC sampling methods*. I. Ntzoufras, Dept. of Statistics, Athens University.
- *Introduction to Markov Chain Monte Carlo*. Charles Geyer, Dept. of Statistics, University of Minnesota.
- *Introduction aux méthodes de Monte Carlo*. S. Allassonnière, CMAP, Ecole Polytechnique.

Inférence Bayésienne (Introduction)

Notations

On dispose d'un échantillon d'observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ de taille n

Chaque observation x_i est une réalisation d'une variable aléatoire X_i

On fait l'hypothèse que les v.a. X_i sont Indépendantes et Identiquement Distribuées (i.i.d)

La densité de X_i dépend d'un paramètre θ (inconnu)

On appelle information a priori sur le paramètre θ, toute information disponible sur θ en dehors de celle apportée par les observations.

Il existe une incertitude sur l'information a priori sur θ (sinon on n'aurait pas à estimer le paramètre θ, il serait connu avec certitude)

C'est cet a priori sur le paramètre θ qui différentie l'inférence bayésienne des méthodes classiques d'estimation (optimisation).

Lois de probabilité

On modélise donc l'information a priori sur θ au travers d'une loi de probabilité, appelée loi a priori :

La densité a priori est notée $\pi(\theta)$ (θ est ici une v.a.)

Le modèle paramétrique bayésien nécessite aussi la connaissance de la loi des observations :

La loi conditionnelle de $\mathbf{X}$ sachant θ : $f(x|\theta)$ (θ est ici un paramètre !)

$\big(f(x|\theta)$ est aussi appelée *vraisemblance de* θ : $L(\theta; x)$ t.q. $\widehat{\theta}_{MV} := \text{argmax}_\theta\, L(\theta; x)\big)$

Enfin, dans un schéma bayésien, la loi la plus importante est la loi a posteriori :

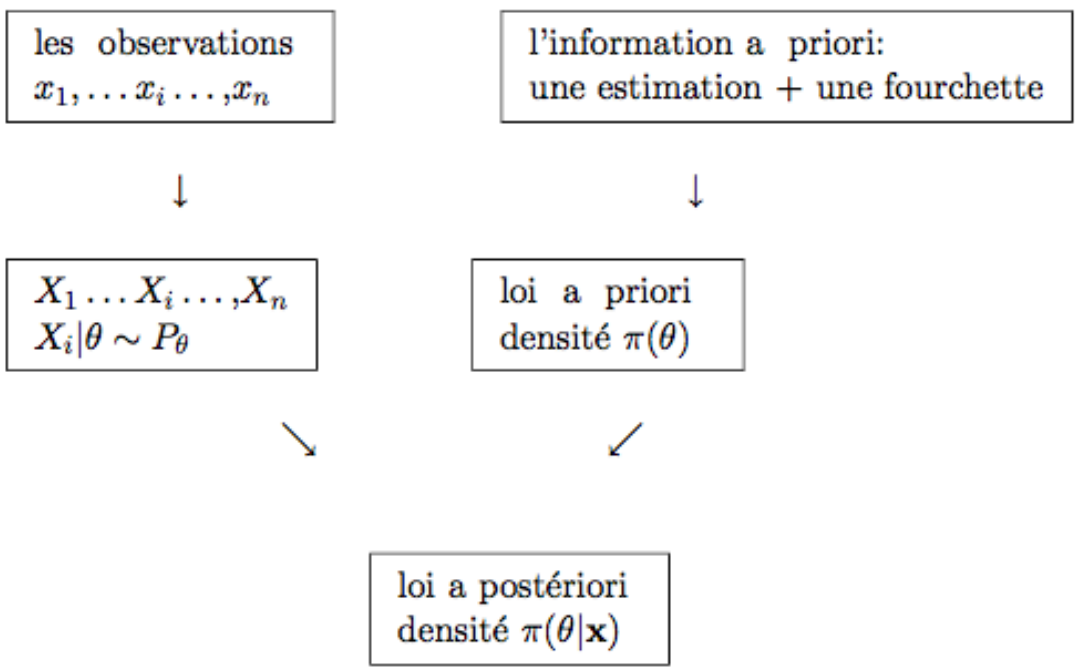
La densité a posteriori s'écrit $\pi(\theta|\mathbf{x}) = \dfrac{f(\mathbf{x}|\theta)\pi(\theta)}{f(\mathbf{x})} = \dfrac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_\theta f(\mathbf{x}|\theta)\pi(\theta)\, d\theta}$ (Bayes)

Principe

La loi a postériori s'interprète comme un résumé (en un sens probabiliste) de l'information disponible sur θ, une fois **x** observé.

L'approche bayésienne réalise l'actualisation de l'information a priori par l'observation **x**, au travers de $\pi(\theta|\mathbf{x})$

a priori $\rightarrow$ a postériori



Exemple d'un calcul de loi a postériori.

$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ t.q. : $X_i|\theta \sim \text{Bernoulli}(\theta)$, i.e. $\mathbb{P}\{X_i = x_i|\theta\} = \theta^{x_i}(1-\theta)^{1-x_i}$

et $\theta \sim \pi(\theta) = \text{Beta}(a,b) = \dfrac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{I}_{[0,1]}(\theta)$

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} P(X = x_i|\theta) = \theta^s(1-\theta)^{n-s} \text{ avec } s = \textstyle\sum_{i=1}^n x_i \quad (X_i \text{ i.i.d.})$$

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_\theta f(\mathbf{x}|\theta)\pi(\theta)\,d\theta} \quad \text{où} \quad \int_\theta f(\mathbf{x}|\theta)\pi(\theta)\,d\theta = \frac{B(\alpha = a+s, \beta = b+n-s)}{B(a,b)}$$

et donc $$\pi(\theta|\mathbf{x}) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\mathbb{I}_{[0,1]}(\theta)$$

Par conséquent : $$\theta|\mathbf{x} \sim \text{Beta}(a + \textstyle\sum_i x_i, b + n - \sum_i x_i)$$

Lorsque *loi a priori* et *loi a postériori* appartiennent à la même famille, la loi a priori est dite **conjuguée**

Estimation

Cas uni-dimensionnel.

On suppose que θ est un paramètre réel

Moralement, $\pi(\theta|\mathbf{x})$ est un résumé de l'information disponible sur θ une fois $\mathbf{x}$ observé.

Dans l'absolu, toute cette information est utile et devrait être communiquée à l'utilisateur (médecin, l'économiste, l'ingénieur. . .)

Mais si l'on souhaite disposer d'une estimation bayésienne de θ, on retient le plus souvent la moyenne de la loi a postériori :

$$\widehat{\theta}_B = \mathbb{E}[\theta|\mathbf{x}] = \int_\theta \theta\pi(\theta|\mathbf{x})d\theta = \frac{\int_\theta \theta f(\mathbf{x}|\theta)\pi(\theta)d\theta}{\int_\theta f(\mathbf{x}|\theta)\pi(\theta)d\theta}$$

On sait (cf. cours d'estimation statistique) que cet estimateur, aussi appelé estimateur en moyenne quadratique est celui qui minimise le risque conditionnel pour une fonction de coût quadratique.

D'autres fonctions de coût (e.g. valeur absolue, uniforme) donnent d'autres estimations bayésiennes (resp. médiane a postériori ou maximum a postériori)

Estimation (exemple)

Avec l'exemple précédent où

$$\theta|\mathbf{x} \sim \text{Beta}(a + \sum_i x_i, b + n - \sum_i x_i) = \text{Beta}(\alpha, \beta)$$

l'estimation $\widehat{\theta}_B = \mathbb{E}[\theta|\mathbf{x}]$ se déduit analytiquement de la moyenne d'une loi Beta :

$$\widehat{\theta}_B = \frac{\alpha}{\alpha + \beta} = \frac{a + \sum_i x_i}{a + b + n}$$

Comme on le verra, ce calcul n'est pas toujours possible ce qui justifie le recours aux méthodes de simulation d'une série de v.a. selon la loi a postériori (cible)...

Plus généralement, l'estimation bayésienne de $h(\theta) \in \mathbb{R}$ est par définition :

$$\mathbb{E}[h(\theta)|\mathbf{x}]$$

Estimation

Cas multi-dimensionnel.

Le paramètre θ est un vecteur à J composantes

$$\theta = (\theta_1, \theta_2, \ldots, \theta_J)$$

La moyenne a postériori $\mathbb{E}[\theta|\mathbf{x}]$ est égale au vecteur $(\mathbb{E}[\theta_1|\mathbf{x}], \mathbb{E}[\theta_2|\mathbf{x}], \ldots, \mathbb{E}[\theta_J|\mathbf{x}])$

où :

$$\mathbb{E}[\theta_j|\mathbf{x}] = \int_{\theta_j} \theta_j \pi(\theta_j|\mathbf{x})\, d\theta_j$$

et $\pi(\theta_j|\mathbf{x})$est obtenue en intégrant $\pi(\theta|\mathbf{x})$ sur toutes les composantes de θ autres que θ_j : Loi Marginale

Cette loi marginale peut encore formellement s'écrire selon la loi conditionnelle complète de θ_j (*full conditional density*) :

$$\pi_j(\theta_j) := \pi(\theta_j|\theta_{\setminus\{j\}}, \mathbf{x}) \propto p(\mathbf{x}|\theta_j)\pi(\theta_j|\theta_{\setminus\{j\}})$$

$\theta_{\setminus\{j\}} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_J)$ représente le vecteur des paramètres privé de θ_j

C'est cette relation qui sera utilisée dans l'échantillonneur de Gibbs

Propriété de l'estimateur bayésien

L'estimateur de Bayes est admissible : il existe une fonction de perte (risque conditionnel, cf. cours sur l'estimation statistique) pour laquelle il n'existe pas de "meilleur" estimateur que l'estimateur bayésien

L'estimateur de Bayes est biaisé... !

Et de plus, sous certaines hypothèses de régularité :

L'estimateur de Bayes est convergent en probabilité

$$\forall \varepsilon, \quad \lim_{n\to\infty} \mathbb{P}\{||\widehat{\theta}_B^{(n)} - \theta_0|| < \varepsilon\} = 0$$

La loi a postériori peut être asymptotiquement approchée par une loi normale

$$\pi(\theta|\mathbf{x}) \underset{n\to\infty}{\approx} \mathcal{N}(\mathbb{E}[\theta|\mathbf{x}], \mathrm{Var}[\theta|\mathbf{x}])$$

Modélisation de l'information a priori

En général, l'information disponible sur le paramètre θ ne permet pas de construire la loi a priori (e.g. intuition, expertise. . .)

Il existe alors des recours "empiriques" ou de "commodité calculatoire" permettant de rester dans le cadre bayésien. . .
. . . même en l'absence de toute information a priori sur le paramètre cherché θ !

Modélisation de l'information a priori

1. **Lois a priori conjuguées** La loi des observations $f(\mathbf{x}|\theta)$ étant connue. On se fixe un a priori $\pi(\theta)$ dans une famille de lois $\mathcal{F}$. Si la loi a postériori $\pi(\theta|\mathbf{x})$ appartient elle aussi à la famille $\mathcal{F}$, les lois sont dites conjuguées.
   Facilite le calcul explicite de l'estimation bayésienne (cf. exemple)
2. **Lois a priori non informatives** On ne dispose d'aucune information a priori sur le paramètre θ. Plusieurs possibilités :

   - On ne veut privilégier aucune valeur particulière de θ dans un intervalle de vraisemblable Θ : $\pi(\theta) = \mathcal{U}_\Theta(\theta)$

   - Utiliser la loi de Jeffreys :

$$\pi_J(\theta) \propto [I(\theta)]^{\frac{1}{2}} \mathbb{I}_\Theta(\theta), \quad \text{où } I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \log f(\mathbf{x}|\theta)}{\partial \theta} \right)^2 \right] \text{ est l'Info. de Fisher}$$

   L'information de Fisher mesure la Variance de $\widehat{\theta}_{MV}$ et s'interprète comme la quantité d'information apportée par l'observation $\mathbf{x}$ sur θ.

Modélisation de l'information a priori (exemple)

On reprend notre exemple où $X_i|\theta \sim$ Bernoulli(θ)

et $\theta \sim$ Beta(a, b) d'hyper-paramètres (a, b) inconnus.

Supposons qu'un expert nous fournit une information a priori sur θ sous la forme d'une estimation (plausible) θ^* et d'un intervalle vraisemblable $I^* = [\theta_{\min}, \theta_{\max}]$

On reparamètre la loi Beta selon : $\left\{ \begin{array}{l} a = \lambda\mu \\ b = \lambda(1-\mu) \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \mu = \frac{a}{a+b} = \mathbb{E}(\theta) \\ \lambda = a + b \end{array} \right.$

Avec cette reparamétrisation $\text{Var}(\theta) = \frac{\mu(1-\mu)}{1+\lambda}$ $\searrow$ quand $\lambda \nearrow$ (à μ fixé)

$\lambda \approx$ précision de l'a priori

Estimation loi a priori : $\left\{ \begin{array}{l} \mu = \mathbb{E}(\theta) \leftarrow \theta^* \\ \lambda \text{ t.q. } \int_{I^*} \text{Beta}\left(\theta\,;\, \lambda\theta^*, \lambda(1-\theta^*)\right)\, d\theta = 0.95 \text{ (Matlab)} \end{array} \right.$

Estimation bayésienne : $\widehat{\theta}_B = \mathbb{E}[\theta|\mathbf{x}] = \frac{\lambda}{\lambda+n}\mathbb{E}(\theta) + \frac{n}{\lambda+n}\overline{x} = \frac{\lambda}{\lambda+n}\mathbb{E}(\theta) + \frac{n}{\lambda+n}\widehat{\theta}_{MV}$

Simulation de Monte Carlo

Même avec une bonne information a priori, il arrive très souvent que le calcul explicite de l'estimation bayésienne $\mathbb{E}[\theta|\mathbf{x}]$ soit laborieux, voire impossible analytiquement.

Les méthodes (ordinaires) de simulation de Monte Carlo consistent à :

- tirer une série de variables aléatoires $(y_1, y_2, \ldots, y_n)$ **indépendantes** et **identiquement distribuées** selon la loi cible $\pi(\mathbf{y})$
- estimer l'espérance $\mu = \mathbb{E}_\pi[h(\mathbf{y})]$ par la moyenne empirique (**ERGODICITÉ**) :

$$\widehat{\mu_n} = \frac{1}{n}\sum_{i=1}^{n} h(y_i)$$

La théorie des grands nombres montre alors que : $\widehat{\mu_n} \xrightarrow{p.s.} \mu, \quad n \to \infty$

Et le théorème central limite, que

$$\sqrt{n}\,(\widehat{\mu_n} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2), \quad n \to \infty, \quad \text{avec} \quad \sigma^2 = \text{Var}\{h(\mathbf{y})\} < \infty$$

En général, σ est inconnu... mais l'Ecart-type de Monte Carlo : $\widehat{\sigma_n}^2 = \frac{1}{n}\sum_{i=1}^{n}(h(y_i) - \widehat{\mu_n})^2$ en est un estimateur convergent et le théorème de Slutsky dit que

$$\frac{\widehat{\mu_n} - \mu}{\widehat{\sigma_n}/n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Simulation selon une loi cible

L'estimation empirique de $\mathbb{E}[\theta|\mathbf{x}]$ par Monte Carlo
suppose que l'on sache simuler selon la loi a posteriori $\pi(\theta|\mathbf{x})$

Plusieurs cas de figure :

a) Il existe un générateur de nombres aléatoires selon la loi cible cherchée (e.g. *Statistics Toolbox* de Matlab©)

b) Il n'y a pas de générateur correspondant, mais la loi cible est associée à une fonction de répartition inversible (pour des v.a. à valeur dans $\mathbb{R}$ seulement)

c) La v.a. à simuler est à valeurs dans $\mathbb{R}^d$, ou la fonction de répartition de la v.a. n'est pas (facilement) inversible : Méthode d'acceptation - rejet

Méthode d'inversion

Fonction de Répartition

$$\begin{array}{rccll} F: & \mathbb{R} & \mapsto & [0,1] & \\ & x & \mapsto & F(x) & = \mathbb{P}(X \leq x) \\ & & & & = \mathbb{P}_X(]-\infty, x]) \end{array}$$

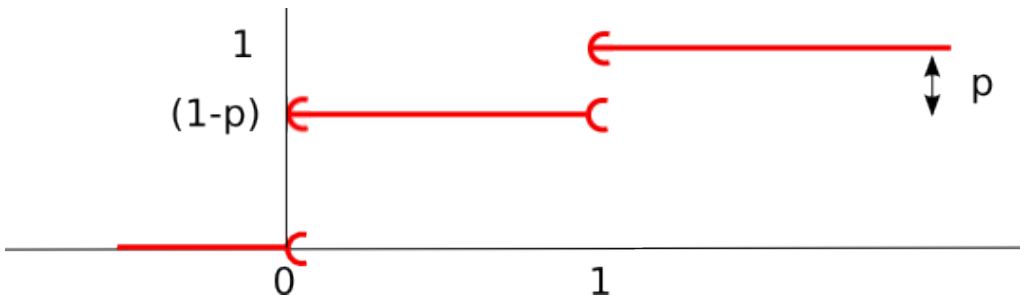
- F est croissante
- $\lim_{x\to-\infty} F(x) = 0$; $\lim_{x\to\infty} F(x) = 1$
- F est continue à droite
- F admet une limite à gauche
- $dF(x) = f(x)dx$ avec f la densité de x

Inverse généralisée

$$\begin{array}{rccl} F^{-1}: &]0,1[& \mapsto & \mathbb{R} \\ & u & \mapsto & \inf\{x \in \mathbb{R}, \text{ t.q. } F(x) \geq u\} \end{array}$$

- F^{-1} est croissante
- F^{-1} est continue à gauche
- $F(x) \geq u \Leftrightarrow x \geq F^{-1}(u)$
- $F(F^{-1}(u)) \geq u$ avec égalité si F continue

!!! F n'est pas nécessairement **bijective**. Ex. $F(x) = (1-p)\delta(x) + p\delta(x-1)$



Méthode d'inversion

Théorème d'inversion

(a) Soit $\mathcal{U}([0,1])$ la loi uniforme sur $[0,1]$ et $u \sim \mathcal{U}([0,1])$, alors :

$$X = F^{-1}(u) \sim F$$

(b) si $X \sim F$ et F est continue, alors $F(X) \sim \mathcal{U}([0,1])$

Méthode d'inversion

Algorithme d'inversion : pour simuler $X \sim F$

1- Simuler $U \sim \mathcal{U}([0,1])$

2- Définir la nouvelle v.a. $X = F^{-1}(U)$.

Pour calculer F^{-1} :

- Forme explicite
- Résolution numérique de $F(x) = u$
  - Bissection : on peut toujours encadrer la valeur de $x = F^{-1}(u)$ qu'on cherche par un intervalle $[a, b]$. Selon que $F\left(\frac{a+b}{2}\right)$ est supérieur ou inférieur à u, on ajuste l'intervalle $[a, b]$ au demi-intervalle à gauche ou à droite. On itère.
  - Newton-Raphson...
- Approximation algébrique de F^{-1}
  ex : pour $\mathcal{N}(0,1)$: $g(u) = (-\log u)^{1/2} + \frac{A[(-2\log u)^{1/2}]}{B[(-2\log u)^{1/2}]}$, où A et B sont 2 polynômes de degré 4.

Méthode d'acceptation - rejet

Soit X une v.a. de densité $f(X)$ sur $\mathbb{R}^d$.
Si $d > 1$, la méthode d'inversion ne permet pas de simuler X.

La méthode de simulation par acceptation - rejet s'appuie sur deux théorèmes simples.

Théorème 1

(a) Soient X de densité $f(X)$ sur $\mathbb{R}^d$, $U \sim \mathcal{U}([0,1])$, indépendant de X et $c > 0$, une constante.

Alors le couple $(X, c\,U\,f(X)) \sim \mathcal{U}(\mathcal{C})$ où $\mathcal{C} = \{(x,y) \text{ t.q. } 0 \leq y \leq c\,f(x)\}$

(b) Réciproquement, si $(X, Y) \sim \mathcal{U}(\mathcal{C})$ (avec $\mathcal{C}$ définie ci-dessus)

alors X a pour densité f

Algorithme pour générer une loi uniforme sur $\mathcal{C}$

- Tirer $X \sim \dfrac{f(x)}{\int f(x)}$ (on suppose que le choix de f permet de le faire, e.g. inversion)
- Tirer $U \sim \mathcal{U}([0,1])$ et **poser** $Y = c\,U\,f(X)$
- Retourner (X, Y) : Les v.a. (X, Y) sont uniformément distribuées sur le domaine $\mathcal{C}$.

Méthode d'acceptation - rejet

Théorème 2

Soit $(X_k)_{k\geq 0}$ une suite de v.a. i.i.d à valeur dans $\mathbb{R}^d$ de même loi que X.

Soit un ensemble A tel que $\mathbb{P}\{X \in A\} = p > 0$.

Soit $Y = X_{\tau_A}$ où $\tau_A = \inf\{k \geq 1 : X_k \in A\}$ (1er instant de X dans A)

$B \in \mathcal{X}$ ($\mathcal{X}$: $\sigma-$algèbre de l'espace probabilisé - en "gros" B est une partie de $\mathbb{X}$)

$$\text{Alors } \mathbb{P}(Y \in B) = \mathbb{P}(X \in B | X \in A)$$

Corollaire 2

Si $X \sim \mathcal{U}(\mathcal{C})$ où $\mathcal{C} \subset \mathbb{R}^d$. Soit $B \subset \mathcal{C}$

$$\text{Alors } Y = X_{\tau_B} \sim \mathcal{U}(B)$$

(puisque dans le cas uniforme $\mathbb{P}(Y \in B) = \mathbb{P}(X \in B | X \in A) = \mathbb{P}(X \in B)$)

Méthode d'acceptation - rejet

Algorithme du rejet

Soit g une densité t.q. $\exists c > 0$ t.q. $\forall x \in \mathbb{R}^d \quad f(x) \leq cg(x)$

Thm 1 : tirer une v.a. uniforme sur $\mathcal{C} = \{(x, y),\ 0 \leq y \leq cg(x)\}$

Thm 2 : Tant que X_k n'est pas dans $B = \{(x, y),\ 0 \leq y \leq f(x)\}$ on continue...

Corollaire et Thm 1(b) : Quand on est dans B, on sait que l'abscisse de $(X) \sim f$

La densité g est appelée densité instrumentale : son choix est crucial pour la rapidité de convergence de l'algorithme

Algorithme de Metropolis

Monte Carlo : simuler une séquence de v.a.

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}, \theta^{(t+1)}, \ldots, \theta^{(M)} \quad \sim \quad \pi(\theta|\mathbf{x})$$

$$\text{t.q.} \quad \frac{1}{M}\sum_{i=1}^{M} h(\theta^{(i)}) \quad \xrightarrow{p.s.} \quad \mathbb{E}\{h(\theta)|\mathbf{x}\}$$

Alternative : Construire dynamiquement la série $\{\theta^{(i)}\}_{i=1}^{M}$

$$\theta^{(t)} \rightarrow \theta^{(t+1)}$$

de telle sorte que :

- $\theta^{(t+1)}$ soit *vraisemblable* au sens de la loi a postériori $\pi(\theta|\mathbf{x})$
- et relativement à la réalisation courante $\theta^{(t)}$

Algorithme de Metropolis

Principe

- Tirer une réalisation candidate $\theta^{(\mathrm{cand})}$ selon une loi de proposition (*que l'on détaillera plus tard*)
- Former le rapport

$$\alpha := \min\left(1, \frac{\pi(\theta^{(\mathrm{cand})}|\mathbf{x})}{\pi(\theta^{(t)}|\mathbf{x})}\right)$$

- Si $\alpha = 1$, alors $\theta^{(t+1)} \leftarrow \theta^{(\mathrm{cand})}$, systématiquement
  Si $\alpha < 1$, $\theta^{(t+1)} \leftarrow \theta^{(\mathrm{cand})}$, avec probabilité α
  et $\theta^{(t+1)} \leftarrow \theta^{(t)}$, avec probabilité $(1-\alpha)$

Algorithme de Metropolis

Avantages

– $\alpha := \dfrac{f(\mathbf{x}|\theta^{(\text{cand})})\pi(\theta^{(\text{cand})})}{f(\mathbf{x}|\theta^{(t)})\pi(\theta^{(t)})}$ ne dépend plus de la loi marginale $p(\mathbf{x})$

Connaître la loi a postériori à une constante près suffit : $\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$, où (*vraisemblance* $*$ *prior*) est appelé *densité a postériori non normalisée*

– Si $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$

on peut simuler chaque série $\{\theta_j^{(i)}\}_{i=1}^M$ individuellement avec le rapport :

$$\frac{\pi(\theta_1, ..., \theta_j^{(\text{cand})}, ...\theta_J|\mathbf{x})}{\pi(\theta_1, ..., \theta_j^{(t)}, ...\theta_J|\mathbf{x})} = \frac{\pi(\theta_j^{(\text{cand})}|\theta_{\setminus\{j\}}, \mathbf{x})\pi(\theta_{\setminus\{j\}}|\mathbf{x})}{\pi(\theta_j^{(j)}|\theta_{\setminus\{j\}}, \mathbf{x})\pi(\theta_{\setminus\{j\}}|\mathbf{x})} = \frac{\pi_j(\theta_j^{(\text{cand})})}{\pi_j(\theta_j^{(t)})}$$

où les paramètres $\theta_{\setminus\{j\}}$ sont des constantes fixées à une valeur courante.

$$\alpha_j = \min\left(1, \frac{\pi(\theta_j^{(\text{cand})}|\theta_{\setminus\{j\}}, \mathbf{x})}{\pi(\theta_j^{(t)}|\theta_{\setminus\{j\}}, \mathbf{x})}\right) = \min\left(1, \frac{f(\mathbf{x}|\theta_j^{(\text{cand})})\pi(\theta_j^{(\text{cand})})}{f(\mathbf{x}|\theta_j^{(t)})\pi(\theta_j^{(t)})}\right) = \min\left(1, \frac{\pi_j(\theta_j^{(\text{cand})})}{\pi_j(\theta_j^{(t)})}\right)$$

– On peut simuler $\theta^{(\text{cand})}$ selon *"n'importe quelle loi de proposition"* ...

Metropolis : Choix de la loi de proposition

Idéalement, la loi de proposition pour le paramètre θ devrait s'approcher au plus près de la loi a postériori $\pi(\theta|\mathbf{x})\ldots$

$$\theta^{(\text{cand})} \sim q(\theta^{(\text{cand})}|\theta^{(t)})$$

L'algorithme de Metropolis se caractérise par une loi de proposition symétrique :

$$q(\theta^{(\text{cand})}|\theta^{(t)}) = q(\theta^{(t)}|\theta^{(\text{cand})})$$

Si de plus $q(\theta^{(\text{cand})}|\theta^{(t)}) = q(\theta^{(t)} - \theta^{(\text{cand})})$: **Marche Aléatoire**

où la probabilité d'atteindre $\theta^{(\text{cand})}$ depuis la valeur courante $\theta^{(t)}$ et probabilité d'y revenir depuis $\theta^{(\text{cand})}$ sont identiques

Exemples

⋆ $\theta^{(\text{cand})} \sim \theta^{(t)} + \sigma_q \mathcal{N}(0,1)$: σ_q fixe l'ampleur du saut entre $\theta^{(t)}$ et $\theta^{(\text{cand})}$

⋆ $\theta^{(\text{cand})} \sim \theta^{(t)} + \kappa\, \mathcal{U}_{[-1,1]}$: κ ——————— " ——————— .

⋆ Marche aléatoire de Langevin : $\theta^{(\text{cand})} \sim \theta^{(t)} + \sigma \left[\mathcal{N}(0,1) + 0.5 \nabla \log \pi(\theta^{(t)}|\mathbf{x})\right]$

Metropolis : Convergence

Le taux d'acceptation de la proposition $\theta^{(\text{cand})}$ simulée selon $q(\theta^{(\text{cand})}|\theta^{(t)})$ dépend :

- de la pertinence du choix de $q(a|b)$ relativement à $\pi(\theta|\mathbf{x})$
- de la distance entre $\theta^{(\text{cand})}$ et $\theta^{(t)}$, et donc de la valeur de σ_q (resp. de κ)

Une valeur de σ_q (resp. κ) trop faible entraine :

- un fort taux d'acceptation de la proposition $\theta^{(\text{cand})}$
- une exploration très lente de la loi a postériori $\pi(\theta|\mathbf{x})$
- une forte corrélation de la série $(\theta^{(1)}, ..., \theta^{(t)}, \theta^{(t+1)}, ...)$ car la marche aléatoire reste confinée à un sous espace de Θ

Une valeur de σ_q (resp. κ) trop forte entraine :

- un faible taux d'acceptation de la proposition $\theta^{(\text{cand})}$
- une forte corrélation de la série $(\theta^{(1)}, ..., \theta^{(t)}, \theta^{(t+1)}, ...)$ car la marche aléatoire reste longtemps bloquée sur la même valeur $\theta^{(t)}$

Heuristiques sur σ_q

- Utiliser un estimateur V_θ de la dispersion (variance) de θ et fixer $\sigma_q^2 = kV_\theta$ (typ. $k \in [2, 10]$)
- Une séquence $\{\theta^{(i)}\}_{i=1}^M$ *mélangeante* qui parcours bien la loi a postériori $\pi(\theta|\mathbf{x})$ correspond à un taux d'acceptation autour de 0.4 (*update* param. seul, 0.2 sinon)

Monte Carlo Markov Chain

- (Metropolis)
- **Metropolis-Hastings**
- **Echantillonneur de Gibbs**
- (Metropolis-Hastings within Gibbs)

Monte Carlo Markov Chain

Les marches aléatoires de l'algorithme de Metropolis sont des cas particuliers de
Chaînes de Markov à matrices de transition symétriques

Généralisation à l'**Algorithme de Metropolis-Hastings**

$$\theta^{(\text{cand})} \sim q(\theta^{(\text{cand})}|\theta^{(t)}) = Q(\theta^{(t)}, \theta^{(\text{cand})})$$

$Q(a,b) = \mathbb{P}\{x_{n+1} = b | x_n = a\}$: matrice de transition d'une chaîne de Markov $\{x_n\}_{n\in\mathbb{N}}$

On définit le nouveau rapport d'acceptation - rejet :

$$\alpha_{\text{M}-\text{H}}(\theta^{(\text{cand})}, \theta^{(t)}) := \min\left(1, \frac{\pi(\theta^{(\text{cand})}|\mathbf{x})}{\pi(\theta^{(t)}|\mathbf{x})} \frac{q(\theta^{(t)}|\theta^{(\text{cand})})}{q(\theta^{(\text{cand})}|\theta^{(t)})}\right)$$

Quelques rappels théoriques sur les Chaînes de Markov

Définition – Un processus aléatoire est une chaîne de Markov à valeurs dans $(E, \mathcal{E})$ ssi

$$\forall A \in \mathcal{E}, \quad \mathbb{P}\{x_{n+1} \in A | \sigma(x_1, x_2, ...x_n)\} = \mathbb{P}\{x_{n+1} \in A | x_n\}$$

Matrice / Noyau de transition –

Cas discret : $P : \mathbb{X} \times \mathbb{X}$ est une matrice de transition si $\forall (x, y) \in \mathbb{X}^2$

$$P(x, y) = \mathbb{P}\{x_{n+1} = y | x_n = x\}, \quad P(x, y) \geq 0 \quad \text{et} \quad \sum_{y \in \mathbb{X}} P(x, y) = 1$$

(en particulier $\forall x \in \mathcal{X}, \quad Q(y) = P(x, y)$ est un loi de probabilité sur $\mathbb{X}$)

Cas continu : $P : \mathbb{X} \times \mathbb{X}$ est un noyau de transition si

(i) $\forall x \in \mathbb{X}, \quad A \mapsto P(x, A)$ est une loi de probabilité sur $(\mathbb{X}, \mathcal{X})$

(ii) $\forall A \in \mathcal{X} \quad x \mapsto P(x, A)$ est une application mesurable

dans le cas continu, le noyau désigne également la densité conditionnelle $P(x, y)$ de la transition $P(x, \cdot)$:

$$P(\mathbf{x} \in A | x) = \int_A P(x, dy)$$

Quelques rappels théoriques sur les Chaînes de Markov

Composition (théorème de convolution Chapman-Kolmogorov) –

Soient (P_1, P_2) deux matrices (resp. 2 noyaux) de transition, On définit :

$$\begin{aligned} [P_1, P_2](x, y) &= \sum_{z \in \mathbb{X}} P_1(x, z) P_2(z, y) \\ [P_1, P_2](x, A) &= \int P_1(x, dz) P_2(z, A) \end{aligned}$$

et par récurrence

$$P^{n+m}(x, A) = \int P^n(x, dz) P^m(z, a)$$

Chaîne homogène – Si la matrice (noyau) de transition $P(x_n, x_{n+1})$ est indépendante de l'indice courant n, la chaîne est dite homogène

Stationnarité – Une chaîne $\mathbf{x}$ est stationnaire si $\forall h \in \mathbb{N}$, $\forall p \in \mathbb{N}$, $(x_{t_1}, ..., x_{t_p})$ et $(x_{t_1+h}, ..., x_{t_p+h})$ ont la même loi

Quelques rappels théoriques sur les Chaînes de Markov

Mesure invariante – Soit P un noyau de transition sur $(\mathbb{X}, \mathcal{X})$ et Π une mesure sur $(\mathbb{X}, \mathcal{X})$. Π est dite invariante pour P ssi

$$\Pi P = \Pi, \quad \text{i.e.} \quad \forall y \in \mathbb{X}, \ \sum_{x \in \mathbb{X}} \Pi(x) P(x, y) = \Pi(y)$$

$$\text{ou} \quad \forall A \in \mathcal{X}, \ \int_{\mathbb{X}} \Pi(dx) P(x, A) = \Pi(A)$$

Soit **x** une chaîne de Markov de noyau de transition P :

- Si **x** stat. et $x_0 \sim \mu$ (loi initiale) $\Rightarrow$ $x_1 \overset{\mathcal{L}}{=} x_0$ $\Rightarrow$ $\mu P = \mu$: mesure invariante
- Réciproquement, si $\mu P = \mu \Rightarrow \forall n, \ \mu P^n = \mu \Rightarrow x_n \sim \mu$: **x** est stationnaire

Ergodicité – Lorsque la mesure limite invariante Π est indépendante de la loi initiale μ, la chaîne est dite ergodique (propriété d'oubli des conditions initiales)

Temps de chauffe ("*Burn-in*") – C'est l'indice $n = T$ au-delà duquel la chaîne est entrée dans le régime stationnaire ("*steady state*") Π

$$x = (\underbrace{x_1, x_2, ..., x_{T-1}}_{\text{burn-in sequence}}, \underbrace{x_T, x_{T+1}, \ldots}_{\text{steady state}})$$

Quelques rappels théoriques sur les Chaînes de Markov

Réversibilité (CNS d'existence d'une loi invariante) – Si une mesure de probabilité μ vérifie la relation de réversibilité

$$\mu(i)P(i,j) = \mu(j)P(j,i)$$

ou (cas continu) $$\mu(dx)P(x,dy) = \mu(dy)P(y,dx)$$

(état d'équilibre du système : le flux de j vers i est égal au flux de i vers j)

alors μ est la mesure invariante pour P

Metropolis-Hastings

Soit une l**oi de probabilité** Π connue à une constante multiplicative près : peut-on construire un **noyau de transition** P dont Π est la **mesure de probabilité invariante** ?

OUI

Algorithme de Metropolis-Hastings (Hastings, 1970)

$$\theta^{(\text{cand})} \sim q(\theta^{(\text{cand})}|\theta^{(t)})$$

q : Loi de proposition (ou noyau de transition) et α : Rapport d'acceptation-rejet :

$$\alpha(\theta^{(t)}, \theta^{(\text{cand})}) := \min\left(1, \frac{\Pi(\theta^{(\text{cand})})}{\Pi(\theta^{(t)})} \frac{q(\theta^{(t)}|\theta^{(\text{cand})})}{q(\theta^{(\text{cand})}|\theta^{(t)})}\right)$$

- Avec probabilité $\alpha(\theta^{(t)}, \theta^{(\text{cand})})$ on pose $\theta^{(t+1)} = \theta^{(\text{cand})}$
- Et sinon, avec probabilité $1 - \alpha(\theta^{(t)}, \theta^{(\text{cand})})$ on pose $\theta^{(t+1)} = \theta^{(t)}$

Metropolis-Hastings

Noyau de transition – La chaîne générée a pour noyau de transition P dont la densité est donnée par deux cas :

- si on accepte $\theta^{(\mathrm{cand})}$ proposé : $p(\theta^{(t)}, \theta^{(t+1)}) = q(\theta^{(t+1)}|\theta^{(t)})\alpha(\theta^{(t)}, \theta^{(t+1)})$
- dans le cas de rejet : la probabilité de rester en $\theta^{(t)}$ s'écrit :

$$r(\theta^{(t)}) = p(\theta^{(t)}, \theta^{(t)}) = \int q(\theta^{(t+1)}|\theta^{(t)})[1 - \alpha(\theta^{(t)}, \theta^{(t+1)})]\,\mathrm{d}\theta^{(t+1)}$$

Et par conséquent, le noyau de transition s'écrit :

$$P(\theta^{(t)}, A) = \underbrace{\int_A \alpha(\theta^{(t)}, y) q(y|\theta^{(t)})\mathrm{d}y}_{\substack{\text{toutes les transitions acceptées} \\ \text{de } \theta^{(t)} \text{ vers } A \neq \{\theta^{(t)}\}}} + \underbrace{\delta_{\theta^{(t)}}(A) r(\theta^{(t)})}_{\substack{\text{tous les rejets qui maintiennent} \\ \text{l'état courant } \theta^{(t)} \ (A = \{\theta^{(t)}\})}}$$

Mesure invariante – On montre (cf. Ch. Robert & G.Casella) que Π vérifie la **relation de réversibilité** (équilibre des flux) :

$$\Pi(dx)P(x, dy) = \Pi(dy)P(y, dx)$$

et donc : Π **est la mesure invariante (cible) pour le noyau de transition** P

Metropolis-Hastings

≫ Run `[X,alpha] = MHexample({'exp',1}, {'norm',0,2},10,1000); plot(X)`

(in /Users/pgoncalv/M-files/MCMC/MHexample.m)

≫ run JAVA-applet : `http://www.lbreyer.com/classic.html`

and play with Target and Proposal Laws.

Echantillonneur de Gibbs

– Revenons au cas de l'estimation d'un paramètre multi-dimensionnel

$$\underline{\theta} = (\theta_1, \theta_2, ..., \theta_J)$$

– Alors que la loi a posteriori complète cible $\Pi(\theta) = \pi(\theta|\mathbf{x})$ peut être non-standard et difficile à échantillonner,

– la loi conditionnelle complète $\pi_j(\theta_j) := \pi(\theta_j|\theta_{\setminus\{j\}}, \mathbf{x})$ peut s'avérer être une loi cible de forme standard et facile à simuler

$$\theta_j^{(\mathrm{cand})} \sim \pi_j(\theta_j^{(t)})$$

– en itérant sur les indices $j = 1, ..., J$, la séquence multi-dimensionnelle

$$\underline{\theta}^{(1)}, \underline{\theta}^{(2)}, ..., \underline{\theta}^{(t)}, \underline{\theta}^{(t+1)}, ...\underline{\theta}^{(M)},$$

forme une **chaîne de Markov admissible**

Echantillonneur de Gibbs

Algorithme

$\underline{\theta}^{(0)} \leftarrow$ `init value`

`For t = 1 to M`

`Set` $\widetilde{\underline{\theta}} \leftarrow \underline{\theta}^{(t)}$

`For j = 1 : J`

`Draw` $\theta_j^{(t+1)} \sim \pi(\theta_j | \widetilde{\theta}_{\setminus\{j\}}, \mathbf{x})$

`Set` $\widetilde{\theta}_j \leftarrow \theta_j^{(t+1)}$

`End`

`Set` $\underline{\theta}^{(t+1)} \leftarrow \widetilde{\underline{\theta}}$

`End`

Remarques

- A l'indice j de l'itération t : $\widetilde{\underline{\theta}} = (\theta_1^{(t+1)}, ..., \theta_{j-1}^{(t+1)}, \theta_j^{(t)}, \theta_{j+1}^{(t)}, ..., \theta_J^{(t)})$
- Les propositions sont systématiquement acceptées (loi de proposition = loi cible)
- (les lois conditionnelles complètes sont connues et simulables . . .)

Echantillonneur de Gibbs

Propriétés

– Noyau de transition :

$$P(\underline{\theta}^{(t)}, \underline{\theta}^{(t+1)}) = \prod_{j=1}^{J} \pi_j(\theta_j^{(t+1)} | \theta_1^{(t+1)}, ..., \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, ..., \theta_J^{(t)})$$

– La loi cible (a posteriori complète) Π est la mesure invariante pour P

– Ce n'est pas un noyau réversible. L'ordre dans lequel on visite les composantes joue un rôle important.

Et si les lois conditionnelles complètes $\pi_j(\theta_j)$ne sont pas facilement simulables ? . . .

Metropolis-Hastings within Gibbs

Algorithme

$\underline{\theta}^{(0)} \leftarrow$ `init value`

`For t = 1 to M`

`Set` $\widetilde{\underline{\theta}} \leftarrow \underline{\theta}^{(t)}$

`For j = 1 : J`

`Draw` $\theta_j^{(t+1)} \sim \pi(\theta_j | \widetilde{\theta}_{\setminus\{j\}}, \mathbf{x})$

`Set` $\widetilde{\theta}_j \leftarrow \theta_j^{(t+1)}$

$\Rightarrow$

`Draw` $\theta_j^{(\text{cand})} \sim q(\theta_j^{(\text{cand})} | \theta_j^{(t)})$

`Compute`

$$\alpha = \min\left(1, \frac{\pi_j(\theta_j^{(\text{cand})})}{\pi_j(\theta_j^{(t)})} \frac{q(\theta_j^{(t)} | \theta_j^{(\text{cand})})}{q(\theta_j^{(\text{cand})} | \theta_j^{(t)})}\right)$$

`With probability` α: $\widetilde{\theta}_j \leftarrow \theta_j^{(\text{cand})}$

`With probability` $(1-\alpha)$: $\widetilde{\theta}_j \leftarrow \theta_j^{(t)}$

`End`

`Set` $\underline{\theta}^{(t+1)} \leftarrow \widetilde{\underline{\theta}}$

`End`