ISAO - Méthodes MCMC

Magalie Fromont

ENSAI Deuxième année

2007-2008

Méthodes de Monte Carlo

Comment peut-on tirer profit des techniques de simulation de variables aléatoires ?

Définition

On appelle méthode de Monte Carlo toute méthode visant à calculer une valeur numérique par le biais de procédés aléatoires.

Méthodes de Monte Carlo

Problèmes typiques

1. **Intégration**

Calcul de $I = \int h(x)f(x)dx$.

Remarques fondamentales :

- $I = E_f[h(x)]$ si f est une densité de probabilité.
- $I = \int h(x)\frac{f(x)}{g(x)}g(x)dx = E_g\left[h(x)\frac{f(x)}{g(x)}\right]$ pour toute densité g dont le support contient celui de f.

Méthodes de Monte Carlo
Problèmes typiques

1. **Intégration**

Calcul de $I = \int h(x)f(x)dx$.

Remarques fondamentales :

- $I = E_f[h(x)]$ si f est une densité de probabilité.
- $I = \int h(x)\frac{f(x)}{g(x)}g(x)dx = E_g\left[h(x)\frac{f(x)}{g(x)}\right]$ pour toute densité g dont le support contient celui de f.

2. **Optimisation**

Calcul de $\max_{\{x\in\mathcal{X}\}} f(x)$ ou de $\mathrm{argmax}_{\{x\in\mathcal{X}\}} f(x)$.

Méthodes de Monte Carlo

Intégration

On cherche à évaluer $I = \int h(x)f(x)dx$.

Si f est une densité de probabilité, $I = E_f[h(x)]$.

Méthode de Monte-Carlo classique

Si $(x_1, \ldots, x_L)$ est la simulation d'un L-échantillon de la loi de densité f,

$$\hat{I}_L = \frac{1}{L}\sum_{l=1}^{L} h(x_l).$$

Justification : Loi forte des grands nombres + TCL si $\int h^2 f < \infty$.

Méthodes de Monte Carlo

Intégration

On cherche à évaluer $I = \int h(x)f(x)dx$.

Si f est une densité de probabilité, $I = E_f[h(x)]$.

Méthode de Monte-Carlo classique

Si $(x_1, \ldots, x_L)$ est la simulation d'un L-échantillon de la loi de densité f,

$$\hat{I}_L = \frac{1}{L}\sum_{l=1}^{L} h(x_l).$$

Justification : Loi forte des grands nombres + TCL si $\int h^2 f < \infty$.

Méthodes de Monte Carlo
Intégration

Si f n'est pas forcément une densité, pour toute densité g dont le support contient celui de f,

$$I = \int h(x)\frac{f(x)}{g(x)}g(x)dx = E_g\left[h(x)\frac{f(x)}{g(x)}\right].$$

Méthode de Monte-Carlo par échantillonnage pondéré

Si $(y_1, \ldots, y_L)$ est la simulation d'un L-échantillon de la loi de densité g :

$$\hat{I}_L = \frac{1}{L}\sum_{l=1}^{L} h(y_l)\frac{f(y_l)}{g(y_l)}.$$

Justification : Loi forte des grands nombres + TCL

Méthodes de Monte Carlo
Intégration

Si f n'est pas forcément une densité, pour toute densité g dont le support contient celui de f,

$$I = \int h(x)\frac{f(x)}{g(x)}g(x)dx = E_g\left[h(x)\frac{f(x)}{g(x)}\right].$$

Méthode de Monte-Carlo par échantillonnage pondéré

Si $(y_1, \ldots, y_L)$ est la simulation d'un L-échantillon de la loi de densité g :

$$\hat{I}_L = \frac{1}{L}\sum_{l=1}^{L} h(y_l)\frac{f(y_l)}{g(y_l)}.$$

Justification : Loi forte des grands nombres + TCL.

Méthodes de Monte Carlo

Intégration

Problème : Choix de g ?

$Var_g(\hat{I}_L) = Var_g\left(\frac{1}{L}\sum_{l=1}^{L} h(Y_l)\frac{f(Y_l)}{g(Y_l)}\right) < \infty \Leftrightarrow \int h^2\frac{f^2}{g} < \infty$

g doit avoir une queue de répartition plus lourde que celle de f.

Théorème (Rubinstein)

La densité g^* qui minimise la variance de $\hat{I}_L$ (pour tout L) est donnée par

$$g^*(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}} |h(x)|f(x)dx}.$$

Méthodes de Monte Carlo

Intégration

Problème : Choix de g ?

$Var_g(\hat{I}_L) = Var_g\left(\frac{1}{L}\sum_{l=1}^{L} h(Y_l)\frac{f(Y_l)}{g(Y_l)}\right) < \infty \Leftrightarrow \int h^2\frac{f^2}{g} < \infty$

g doit avoir une queue de répartition plus lourde que celle de f.

Théorème (Rubinstein)

La densité g^* qui minimise la variance de $\hat{I}_L$ (pour tout L) est donnée par

$$g^*(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}}|h(x)|f(x)dx}.$$

Si $h \geq 0$, $g^* = hf/I$, mais I est l'inconnue!

En pratique, on prend g telle que $Var_g(\hat{I}_L) < \infty$ et $|h|f/g \simeq cte$.

Méthodes de Monte Carlo
Intégration

Et si l'on connait g à une constante de normalisation près ?

Si $(y_1, \ldots, y_L)$ est la simulation d'un L-échantillon de la loi de densité g, remplacer $\hat{I}_L$ par

$$\sum_{l=1}^{L} h(y_l) \frac{f(y_l)}{g(y_l)} \Big/ \sum_{l=1}^{L} \frac{f(y_l)}{g(y_l)}$$

permet d'utiliser une densité g connue à une constante près.

Problème : l'optimalité de g^* ne se transmet pas à cet estimateur!

Méthodes de Monte Carlo

Optimisation

On cherche à calculer $\max_{\{x \in \mathcal{X}\}} f(x)$ **ou** $\operatorname{argmax}_{\{x \in \mathcal{X}\}} f(x)$.

Méthodes de Monte Carlo

Optimisation

On cherche à calculer $\max_{\{x\in\mathcal{X}\}} f(x)$ ou $\operatorname{argmax}_{\{x\in\mathcal{X}\}} f(x)$.

Deux méthodes de Monte Carlo simples :

- Si $\mathcal{X}$ est borné, si $(x_1,\dots,x_L)$ est la simulation d'un échantillon de $\mathcal{U}(\mathcal{X})$, on estime $\max_{\{x\in\mathcal{X}\}} f(x)$ par $\max\{f(x_l), l=1\dots L\}$.
  Si $\mathcal{X}$ n'est pas borné, changement de variables.

Méthodes de Monte Carlo
Optimisation

On cherche à calculer $\max_{\{x\in\mathcal{X}\}} f(x)$ ou $\operatorname{argmax}_{\{x\in\mathcal{X}\}} f(x)$.
Deux méthodes de Monte Carlo simples :

- Si $\mathcal{X}$ est borné, si $(x_1,\ldots,x_L)$ est la simulation d'un échantillon de $\mathcal{U}(\mathcal{X})$, on estime $\max_{\{x\in\mathcal{X}\}} f(x)$ par $\max\{f(x_l), l=1\ldots L\}$.
  Si $\mathcal{X}$ n'est pas borné, changement de variables.
- Si $f \geq 0$, estimer $\operatorname{argmax}_{\{x\in\mathcal{X}\}} f(x)$ revient à estimer le mode de la loi de densité $f/\int f$. Si $(x_1,\ldots,x_L)$ est la simulation d'un échantillon de cette loi, on prend comme estimateur le mode de l'histogramme de $(x_1,\ldots,x_L)$.
  Si $f \not\geq 0$, $g(x)=e^{f(x)}$ ou $g(x)=e^{f(x)}/(1+e^{f(x)})$.

Méthodes de Monte Carlo

Optimisation

On cherche à calculer $\max_{\{x\in\mathcal{X}\}} f(x)$ ou $\mathrm{argmax}_{\{x\in\mathcal{X}\}} f(x)$.

Deux méthodes de Monte Carlo simples :

- Si $\mathcal{X}$ est borné, si $(x_1,\ldots,x_L)$ est la simulation d'un échantillon de $\mathcal{U}(\mathcal{X})$, on estime $\max_{\{x\in\mathcal{X}\}} f(x)$ par $\max\{f(x_l), l=1\ldots L\}$.
  Si $\mathcal{X}$ n'est pas borné, changement de variables.
- Si $f \geq 0$, estimer $\mathrm{argmax}_{\{x\in\mathcal{X}\}} f(x)$ revient à estimer le mode de la loi de densité $f/\int f$. Si $(x_1,\ldots,x_L)$ est la simulation d'un échantillon de cette loi, on prend comme estimateur le mode de l'histogramme de $(x_1,\ldots,x_L)$.
  Si $f \not\geq 0$, $g(x) = e^{f(x)}$ ou $g(x) = e^{f(x)}/(1+e^{f(x)})$.

Problème pour la deuxième méthode : Calcul de la constante de normalisation $\int f$ ou $\int g$?

Méthodes de Monte Carlo
Optimisation

Méthodes de type Newton-Raphson :

- L'algorithme Monte Carlo Newton-Raphson MCNR (Mc Culloch 1997 et Tanner 1996).
  Approximations de MC des intégrales de score et des matrices hessiennes.
- L'algorithme Stochastic Approximation Newton-Raphson SANR (Robbins et Monroe 1951).

Méthodes de type Expectation-Maximisation (Dempster et al. 1977) pour la recherche d'un maximum de vraisemblance :

- L'algorithme Monte Carlo Expectation-Maximisation MCEM (Wei et Tanner 1990)
- L'algorithme Stochastic Approximation Expectation-Maximisation SAEM (Delyon et al. 1999)

Méthodes de Monte Carlo
vs. méthodes numériques

Intégration

Méthodes numériques

- Complexité numérique moins importante en petite dimension

Méthodes de Monte Carlo

- Complexité numérique moins importante en grande dimension
- Privilégie les zones importantes du support de la fonction à intégrer (zones où la fonction varie beaucoup ou zones où elle est grande en $|.|$)
- Extension à l'inférence statistique à moindre coût

Méthodes de Monte Carlo
vs. méthodes numériques

Optimisation

Méthodes numériques

- Tient compte de la régularité de f

Méthodes de Monte Carlo

- Pas d'hypothèse sur $\mathcal{X}$ et sur la nature de f
- Sort des bassins d'attraction des extrema locaux ($\rightarrow$ extrema globaux)
- Extension à l'inférence statistique à moindre coût

Méthodes de Monte Carlo
vs. méthodes numériques

Optimisation

Méthodes numériques

- Tient compte de la régularité de f

Méthodes de Monte Carlo

- Pas d'hypothèse sur $\mathcal{X}$ et sur la nature de f
- Sort des bassins d'attraction des extrema locaux ($\rightarrow$ extrema globaux)
- Extension à l'inférence statistique à moindre coût

L'idéal : une combinaison efficace des deux types de méthodes...

Méthodes de Monte Carlo et inférence statistique

Intégration et/ou optimisation

Intégration

- Estimation d'une espérance,
- Calcul de marginales,
- Estimation de la précision d'un estimateur,
- Analyse bayésienne,
- Modèles de mélange ou à données manquantes...

Optimisation

- Minimisation d'un critère,
- Estimation par maximum de vraisemblance,
- Analyse bayésienne,
- Modèles de mélange ou à données manquantes...

Méthodes de Monte Carlo et inférence statistique

L'approche bayésienne

Cadre paramétrique : Soit $X = (X_1, \ldots, X_n)$ un n-échantillon d'une loi de densité connue à un paramètre $\theta \in \Theta$ près.

L'approche bayésienne : on considère θ comme une v.a. suivant une loi de densité (dite a priori) $\pi(\theta)$.
La densité conditionnelle de X sachant θ est notée $f(x|\theta)$.
Le théorème de Bayes : loi a posteriori $\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_\Theta \pi(\theta)f(x|\theta)d\theta}$ i.e. $\pi(\theta|x) \propto \pi(\theta)f(x|\theta)$.

Méthodes de Monte Carlo et inférence statistique
L'approche bayésienne

Cadre paramétrique : Soit $X = (X_1, \ldots, X_n)$ un n-échantillon d'une loi de densité connue à un paramètre $\theta \in \Theta$ près.

L'approche bayésienne : on considère θ comme une v.a. suivant une loi de densité (dite a priori) $\pi(\theta)$.
La densité conditionnelle de X sachant θ est notée $f(x|\theta)$.
Le théorème de Bayes : loi a posteriori $\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_\Theta \pi(\theta)f(x|\theta)d\theta}$ i.e.
$\pi(\theta|x) \propto \pi(\theta)f(x|\theta)$.

Intérêt :

- la loi a priori permet d'intégrer des a priori physiques sur θ
- applications où cette approche s'impose (modèles de chaînes de Markov cachées, modèles de mélange, ruptures).

Méthodes de Monte Carlo et inférence statistique

L'approche bayésienne

Estimateurs bayésiens $T(X)$ de θ

1. Choix d'une fonction de coût : $L(\theta, T(X))$
2. Fonction de risque moyen :
   $R(T) = \int_{\mathcal{X}} \left(\int_{\Theta} L(\theta, T(x)) f(x|\theta)\pi(\theta) d\theta \right) dx.$
3. Estimateur bayésien $T^* = \text{argmin}_T R(T)$.
4. Estimateur bayésien généralisé :
   $T^*(x) = \text{argmin} \int_{\Theta} L(\theta, T(x))\pi(\theta|x) d\theta$ p.p.

Méthodes de Monte Carlo et inférence statistique

L'approche bayésienne

Fonctions de coût

1. $L(\theta, T(X)) = \delta^{\theta}_{T(X)} \Rightarrow T^*(x) = \mathrm{argmax}_{\theta \in \Theta} \pi(\theta|x)$
   Problème d'optimisation

2. $L(\theta, T(X)) = \|T(X) - \theta\|^2 \Rightarrow T^*(x) = \int_{\Theta} \theta\pi(\theta|x)d\theta.$
   Problème d'intégration

Méthodes de Monte Carlo et inférence statistique

L'approche bayésienne

Fonctions de coût

1. $L(\theta, T(X)) = \delta^{\theta}_{T(X)} \Rightarrow T^*(x) = \mathrm{argmax}_{\theta\in\Theta}\pi(\theta|x)$
   Problème d'optimisation

2. $L(\theta, T(X)) = \|T(X) - \theta\|^2 \Rightarrow T^*(x) = \int_\Theta \theta\pi(\theta|x)d\theta.$
   Problème d'intégration

Remarque : la loi a posteriori $\pi(\theta|x) \propto \pi(\theta)f(x|\theta)$, mais la constante de normalisation n'est pas souvent calculable...

Méthodes MCMC
Pourquoi et comment ?

Pourquoi ?

Les méthodes de Monte Carlo par Chaînes de Markov (Monte Carlo Markov Chains en anglais ou MCMC) sont utilisées lorsque la loi d'intérêt n'est pas simulable directement par les méthodes usuelles et/ou lorsque sa densité est connue à une constante de normalisation près.

Méthodes MCMC

Pourquoi et comment ?

Pourquoi ?

Les méthodes de Monte Carlo par Chaînes de Markov (Monte Carlo Markov Chains en anglais ou MCMC) sont utilisées lorsque la loi d'intérêt n'est pas simulable directement par les méthodes usuelles et/ou lorsque sa densité est connue à une constante de normalisation près.

Comment ?

Définition

Une méthode MCMC permet de simuler une chaîne de Markov $(X_n)_{n\geq 0}$ de noyau de transition P convergeant en un certain sens vers la loi d'intérêt π (ergodicité).

Méthodes MCMC

Pourquoi et comment ?

Théorème ergodique pour des chaînes de Markov homogènes
Sous certaines conditions (récurrence au sens de Harris et existence d'une loi invariante π par ex.), pour toute loi initiale μ_0 de X_0, la loi μ_n de X_n vérifie

$$\lim_{n \to +\infty} \|\mu_n - \pi\|_{TV} = 0,$$

et

$$\frac{1}{n} \sum_{k=0}^{n-1} h(X_k) \to E_\pi[h(X)] = \int h(x)\pi(x)dx \text{ p.s.}$$

Remarque : les $(X_n)_{n \geq 0}$ ne sont pas indépendantes, mais le théorème ergodique remplace la loi forte des grands nombres.
Egalement : Théorèmes ergodiques sous des conditions plus faibles, et pour les chaînes inhomogènes.

Méthodes MCMC

Batterie d'algorithmes

- **Simulation et intégration :** Metropolis-Hastings (1953-1970), échantillonnage de Gibbs (1990).
- **Optimisation :** Recuit simulé (1953).

Les algorithmes MCMC font appel, comme les méthodes d'acceptation-rejet et d'échantillonnage pondéré, à la simulation d'une loi instrumentale.

Cette loi instrumentale peut être caractérisée par un noyau de transition $q(.|.)$ ou par une loi conditionnelle.

Méthodes MCMC

Algorithme de Metropolis-Hastings

Initialisation : x_0.

A chaque étape $k \geq 0$:

1. Simuler une valeur y_k de $Y_k \sim q(.|x_k)$.
2. Simuler une valeur u_k de $U_k \sim \mathcal{U}([0,1])$.
3. Poser

$$x_{k+1} = \begin{cases} y_k & \text{si } u_k \leq \rho(x_k, y_k) \\ x_k & \text{sinon,} \end{cases}$$

où $\rho(x_k, y_k) = \min\left\{1, \frac{\pi(y_k)q(x_k|y_k)}{\pi(x_k)q(y_k|x_k)}\right\}$.

Méthodes MCMC

Algorithme de Metropolis-Hastings

Remarques

- L'algorithme de Metropolis-Hastings ne fait intervenir que les rapports $\frac{\pi(y_k)}{\pi(x_k)}$ et $\frac{q(y_k|x_k)}{q(x_k|y_k)}$
  $\Rightarrow$connaissance des constantes de normalisation inutile
- Les deux derniers points de l'étape k de l'algorithme peuvent se traduire : si $\pi(y_k)/q(y_k|x_k) \geq \pi(x_k)/q(x_k|y_k)$, on accepte y_k, sinon, on l'accepte quand même avec probabilité $\pi(y_k)q(x_k|y_k)/\pi(x_k)q(y_k|x_k)$.

Méthodes MCMC

Algorithme de Metropolis-Hastings indépendant

Cas particulier où $q(.|x) = g$ pour tout x

Initialisation : x_0.
A chaque étape $k \geq 0$:

1. Simuler une valeur y_k de $Y_k \sim g(.)$.
2. Simuler une valeur u_k de $U_k \sim \mathcal{U}([0,1])$.
3. Poser

$$x_{k+1} = \begin{cases} y_k & \text{si } u_k \leq \rho(x_k, y_k) \\ x_k & \text{sinon,} \end{cases}$$

où $\rho(x_k, y_k) = \min\left\{1, \frac{\pi(y_k)g(x_k)}{\pi(x_k)g(y_k)}\right\}$.

Méthodes MCMC

Algorithme de Metropolis-Hastings à marche aléatoire

Cas particulier où $q(y|x) = g(y-x)$

Initialisation : x_0.
A chaque étape $k \geq 0$:

1. Simuler une valeur e_k de $\varepsilon_k \sim g(.)$.
2. Simuler une valeur u_k de $U_k \sim \mathcal{U}([0,1])$.
3. Poser $y_k = x_k + e_k$.
4. Poser

$$x_{k+1} = \begin{cases} y_k & \text{si } u_k \leq \rho(x_k, y_k) \\ x_k & \text{sinon,} \end{cases}$$

où $\rho(x_k, y_k) = \min\left\{1, \frac{\pi(y_k)g(x_k - y_k)}{\pi(x_k)g(y_k - x_k)}\right\}$.

Méthodes MCMC

Algorithme de Metropolis-Hastings à marche aléatoire

Noyaux symétriques

Dans le cas d'un noyau symétrique (i.e. $g(t) = g(-t)$) :

$$\rho(x_k, y_k) = \min\left\{1, \frac{\pi(y_k)}{\pi(x_k)}\right\}$$

$\Rightarrow$ interprétation de l'algorithme

Exemples de noyaux symétriques :

- Pour une loi discrète : $g = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$,
- Pour une loi continue : g densité d'une loi normale centrée.

Méthodes MCMC

Algorithme du recuit simulé

Objectif : minimiser une fonction réelle f.

Idée : Metropolis-Hastings pour simuler la loi $\pi(x) \propto e^{-f(x)}$ + estimation du(des) mode(s).
En fait, on change la loi objectif à chaque étape k de l'algorithme :

$$\pi_k(x) = e^{-f(x)/T_k}.$$

La suite de paramètres $(T_k)_{k \geq 0}$ est une suite décroissante de *températures*.
En pratique, la température restera élévée dans les premières itérations pour pouvoir s'extraire des bassins des minima locaux, puis elle sera à décroissance lente jusqu'à tendre vers 0.

Méthodes MCMC

Algorithme du recuit simulé

Initialisation : x_0.

A chaque étape $k \geq 0$:

1. Simuler une valeur y_k de $Y_k \sim q(.|x_k)$.
2. Simuler une valeur u_k de $U_k \sim \mathcal{U}([0,1])$.
3. Poser
$$x_{k+1} = \begin{cases} y_k & \text{si } u_k \leq \rho(x_k, y_k) \\ x_k & \text{sinon,} \end{cases}$$
où $\rho(x_k, y_k) = \min\left\{1, \frac{e^{-f(y_k)/T_k} q(x_k|y_k)}{e^{-f(x_k)/T_k} q(y_k|x_k)}\right\}$.
4. Faire décroître la température $T_k \to T_{k+1}$.

Méthodes MCMC

Echantillonnage de Gibbs

L'échantillonnage de Gibbs : un algorithme de simulation d'une loi $\pi(x)$ telle que

- x admet une décomposition de la forme $x = (x_1, \ldots, x_n)$,
- les loi conditionnelles $\pi_i(.|(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n))$ sont simulables aisément.

Exemple : $(X, Y) \sim \mathcal{N}(0, \Sigma)$, avec $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

$X|Y = y \sim \mathcal{N}(\rho y, 1-\rho^2)$ et $Y|X = x \sim \mathcal{N}(\rho x, 1-\rho^2)$

Autre exemple : Mélange de gaussiennes (voir T.P.) $\hookrightarrow$ données complétées.

Principe de l'algorithme : Réactualisation "composante par composante".

Méthodes MCMC
Echantillonnage de Gibbs

Initialisation : $x_0 = (x_1^0, \ldots, x_n^0)$.
A chaque étape $k \geq 0$:

1. Simuler $x_1^{k+1} \sim \pi_1(.|x_2^k, \ldots, x_n^k)$
2. $\vdots$
3. Simuler $x_i^{k+1} \sim \pi_i(.|x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_{i+1}^k, \ldots, x_n^k)$
4. $\vdots$
5. Simuler $x_n^{k+1} \sim \pi_n(.|x_1^{k+1}, \ldots, x_{n-1}^{k+1})$

Méthodes MCMC

Convergence des algorithmes

On suppose le support de π connexe ou π-connexe.

- La chaîne de Markov générée par l'algorithme de Metropolis-Hastings est ergodique s'il existe $\varepsilon > 0$ et $\delta > 0$ tels que $|x - y| \leq \varepsilon \Rightarrow q(y|x) \geq \delta$.
- La chaîne générée par l'algorithme de Metropolis-Hastings indépendant est ergodique si g est strictement positive presque partout sur le support de π.
- La chaîne générée par l'algorithme de Metropolis-Hastings à marche aléatoire symétrique est ergodique si g est non nulle dans un voisinage de 0.

Méthodes MCMC
Convergence des algorithmes

- La chaîne générée par l'algorithme du recuit simulé est ergodique vers la loi uniforme sur les minima de f sous une condition (formelle) portant sur la température seule.
- La chaîne générée par l'algorithme de l'échantillonnage de Gibbs est ergodique s'il existe $\delta > 0$ tel que pour tout i, $\pi_i(x_i | x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n) > 0$ pour $\|x - x'\| < \delta$.

Méthodes MCMC

Accélération

A voir en T.P...