Méthodes de Simulation

JEAN-YVES TOURNERET

Institut de recherche en informatique de Toulouse (IRIT)

ENSEEIHT, Toulouse, France

Remerciements

- Christian Robert : pour ses excellents transparents
- Éric Moulines : pour ses excellents fichiers Latex
- Olivier Cappé : pour ses excellents conseils
- Nicolas Dobigeon : pour ses excellents programmes Matlab
- Florent Chatelain : pour ses excellents talents d'informaticien

Méthodes de simulation

- Partie 1 : Bayes et Simulation
- Partie 2 : Metropolis - Hastings
- Partie 3 : L'échantillonneur de Gibbs
- Partie 4 : Diagnostic de Convergence
- Partie 5 : Segmentation de signaux stationnaires par morceaux
- Pour finir : Livres, Sites Webs, Pages perso, ...

Cours 1 : Bayes et Simulation

- 1) Introduction : modèles statistiques
- 2) Maximum de vraisemblance
- 3) Méthodes Bayésiennes
- 4) Méthodes de base de simulation
- 5) Méthodes de Monte Carlo pour l'intégration
- 6) Méthodes numériques déterministes

Modèle Statistique (1)

Compromis entre

- un modèle compliqué proche de la réalité qui peut induire des méthodes d'estimation, de détection ou de classification non standards
- un modèle simple qui conduit à des hypothèses comme linéarité, Gaussianité, ... mais qui peut être trop éloigné du phénomène physique étudié

Avec le développement de la puissance de calcul, des méthodes comme les méthodes de simulation peuvent être envisagées plus facilement.

Modèle Statistique (2)

Parfois, on choisit un modèle simple mais la suppression de certaines informations rend le problème difficile :

- Modèles de censure

$$y_i = \min\{x_i, c\}$$

- Modèles de mélanges

$$y_i \sim p_1 f_1(x) + ... + p_k f_k(x)$$

- Modèles stationnaires par morceaux

$$y_i \sim f_k(x) \text{ si } i \in [t_k, t_{k+1}[$$

Maximum de Vraisemblance

- Définition : Pour un échantillon $\boldsymbol{x} = (x_1, ..., x_n)$ de densité $f(\boldsymbol{x}|\boldsymbol{\theta})$, la vraisemblance s'écrit :

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

- Propriétés asymptotiques : asymptotiquement sans biais, convergent et efficace
- Facile à comprendre et souvent facile à étudier

Mais pose problème pour de nombreux modèles statistiques

Exemple 1 : loi Gamma, α connu

$$f(x|\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}\mathbb{I}_{\mathbb{R}+}(x)$$

- Log-vraisemblance :

$$\ln L(\boldsymbol{x}|\alpha,\beta) = -n\ln\Gamma(\alpha) - n\alpha\ln\beta + (\alpha-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{\beta}\sum_{i=1}^{n}x_i$$

- Estimateur du maximum de vraisemblance de β :

$$\widehat{\beta} = \frac{1}{N\alpha}\sum_{i=1}^{n}x_i$$

Exemple 2 : loi Gamma, α inconnu

- Estimateur du maximum de vraisemblance

$$\frac{\partial}{\partial \alpha} \ln L(\boldsymbol{x}|\alpha, \beta) = 0$$
$$\frac{\partial}{\partial \beta} \ln L(\boldsymbol{x}|\alpha, \beta) = 0$$

Équations non-linéaires faisant intervenir la fonction digamma !!

Exemple 3 : loi de Student

$$f(x|\theta,p,\sigma) \propto \frac{1}{\sigma}\left(1+\frac{(x-\theta)^2}{p\sigma^2}\right)^{-\frac{p+1}{2}}$$

- Log-vraisemblance :

$$\ln L(\boldsymbol{x}|\theta,p,\sigma) = -\left(\frac{p+1}{2}\right)\ln\left(\sigma^{\frac{2n}{p+1}}\prod_{i=1}^{n}\left(1+\frac{(x_i-\theta)^2}{p\sigma^2}\right)\right)$$

possède n maxima locaux (p et σ^2 connus)
matlab: student

Modèles de censure

- Loi de Weibull

$$f(x|\alpha,\beta) = \alpha\beta x^{\alpha-1}\exp(-\beta x^{\alpha})\mathbb{I}_{\mathbb{R}+}(x)$$

- Données tronquées $z = \min(x,\omega)$

$$f(z|\alpha,\beta,\omega) = \alpha\beta z^{\alpha}e^{-\beta z^{\alpha}}\mathbb{I}_{]-\infty,\omega]}(z) + \left(\int_{\omega}^{\infty}\alpha\beta z^{\alpha}e^{-\beta z^{\alpha}}dz\right)\delta_{\omega}(z)$$

Modèles de mélange

- Définition

  Supposons que x_i suive la loi de densité $f_j(x_i)$ avec la probabilité p_j :

$$f(x_i) = p_1 f_1(x_i) + ... + p_k f_k(x_i)$$

- Vraisemblance

$$L(\boldsymbol{x}|\boldsymbol{\theta}, p) = \prod_{i=1}^{n} \left(p_1 f_1(x_i) + ... + p_k f_k(x_i)\right)$$

  comporte k^n termes. Donc les techniques classiques d'optimisation sont inappropriées à une telle fonction multimodale.

Méthodes Bayésiennes

- Vraisemblance

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

- Loi a priori sur $\boldsymbol{\theta}$

$$\pi(\boldsymbol{\theta})$$

- Loi a posteriori

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

où $f(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ est la loi marginale de $\boldsymbol{x}$.

Inférence Bayésienne

On rencontre deux types de problèmes avec les méthodes d'estimation Bayésiennes

$$E\left[C\left(\theta,\widehat{\theta}(x)\right)\right]=\int\left[\int C(\theta,\widehat{\theta}(x))f(\theta,\boldsymbol{x})d\boldsymbol{x}\right]d\theta$$

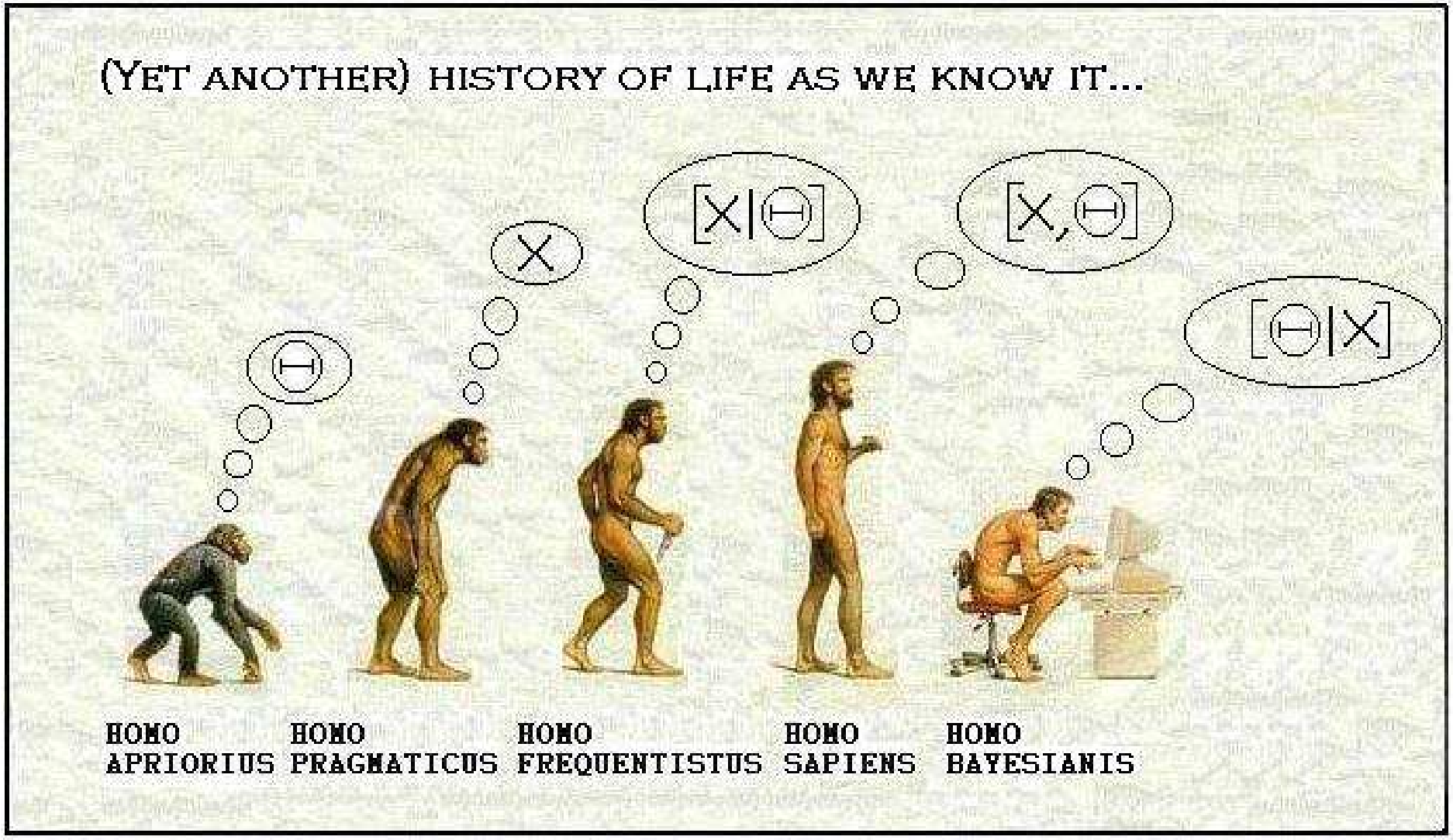
- Des problèmes d'optimisation (coût $0-1$) : estimateur du maximum a Posteriori

$$\widehat{\theta}_{\text{MAP}}(x)=\arg\max f(\theta|\boldsymbol{x})=\arg\max f(\boldsymbol{x}|\theta)\pi(\theta)$$

- Des problèmes d'intégration (coût quadratique) : estimateur MMSE

$$\widehat{\theta}_{\text{MMSE}}(x)=E[\theta|\boldsymbol{x}]=\int\theta f(\theta|\boldsymbol{x})d\theta$$

Méthodes Bayésiennes



Exemple 1 : le cas Gaussien

- Données

$$f(\boldsymbol{x}|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

- Loi a priori : $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

- Loi a posteriori : $\mu|\boldsymbol{x} \sim \mathcal{N}(\mu_n, \sigma_n^2)$

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\frac{1}{n}\sum_{i=1}^{n} x_i + \left(\frac{\sigma^2}{\sigma^2+n\sigma_0^2}\right)\mu_0$$

matlab: Bayes

Exemple 2 : Loi de Cauchy

- Données

$$f(\boldsymbol{x}|\mu,\sigma) = \prod_{i=1}^{n} \sigma^{-1}\left[1+\left(\frac{x_i-\mu}{\sigma}\right)^2\right]$$

- Loi a priori :

$$\pi(\mu,\sigma) = \sigma^{-1}$$

- Loi a posteriori de μ

$$f(\mu|\boldsymbol{x}) \propto \int_0^{\infty} \sigma^{-n-1} \prod_{i=1}^{n}\left[1+\left(\frac{x_i-\mu}{\sigma}\right)^2\right] d\sigma$$

Donc, pas d'expression explicite de cette loi a posteriori !

Lois conjuguées

- Définition : une loi a priori $\pi(\theta)$ est conjuguée si $f(\boldsymbol{x}|\theta)$ et $\pi(\theta)$ appartiennent à la même famille de lois.
- Cas Gaussien

$$f(\boldsymbol{x}|m,\sigma^2) \propto \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\right]$$

☞ loi conjuguée pour m : loi normale

$$m \sim \mathcal{N}(\mu,\beta^2)$$

☞ loi conjuguée pour σ^2 : loi inverse gamma

$$\pi(\sigma^2|\kappa,\gamma) \propto \frac{1}{(\sigma^2)^{\kappa+1}} \exp\left(-\frac{\gamma}{\sigma^2}\right)$$

Lois conjuguées

- Motivation : simplifie le calcul de la loi a posteriori
- Cas Particulier : lois impropres
  - ☞ $\pi(\theta) \propto$ Cste
  - ☞ Jeffreys prior $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$

Méthodes de simulation

- Générateur uniforme

  Pour une fonction de répartition F définie sur $\mathbb{R}$, on définit son inverse généralisée par

  $$F^{-1}(u) = \inf\{x; F(x) \geq u\}$$

  Alors, si U est uniforme sur $[0, 1]$, la variable aléatoire $F^{-1}(U)$ est de fonction de répartition F car

  $$P[F^{-1}(U) \leq x] = P[U \leq F(x)] = F(x)$$

- Cette méthode nécessite de connaître l'inverse généralisée de la fonction de répartition.

Méthodes de simulation

Certaines méthodes utilisent des propriétés spécifiques de la loi à simuler :

- loi Exponentielle

$$X = -\frac{1}{\lambda} \ln U, \quad U \sim \mathcal{U}([0,1])$$

la méthode de l'inverse généralisée donne $X = -\frac{1}{\lambda} \ln (1-U)$.

- Loi Gamma et Beta

$$Y = -b \sum_{j=1}^{a} \ln U_j \sim \mathcal{G}a(a,b), \quad a \in \mathbb{N}^*$$

$$Y = \frac{\sum_{j=1}^{a} \ln U_j}{\sum_{j=1}^{a+b} \ln U_j} \sim \mathcal{B}e(a,b), \quad a,b \in \mathbb{N}^*$$

Méthodes de simulation

- Méthode de Box Müller

  Si U_1 et U_2 sont deux variables indépendantes uniformes sur $[0,1]$, alors

$$Y_1 = \sqrt{-2\ln U_1}\cos(2\pi U_2)$$
$$Y_2 = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

  sont des variables iid distribuées suivant une loi $\mathcal{N}(0,1)$.

- Loi de Poisson

  Si $X_i \sim \mathcal{E}(\lambda)$ et $N \sim \mathcal{P}(\lambda)$ alors

$$P[N=k] = P\left[X_1 + ... + X_k \leq 1 < X_1 + ... + X_{k+1}\right]$$

Méthodes d'acceptation-rejet

- Beaucoup de lois sont difficiles à simuler directement avec les méthodes précédentes
- Il y a certaines applications où la loi à simuler f est connue à une constante multiplicative près (méthodes Bayésiennes)
- ☞ Une solution est de simuler à l'aide d'une loi de proposition g plus simple et d'utiliser un algorithme d'acceptation-rejet

Algorithme d'acceptation-rejet

Soit une loi d'intérêt de densité f et une loi de proposition de densité g telle que

$$f(x) \leq Mg(x)$$

sur le support de f. Alors, on peut simuler suivant f avec l'algorithme suivant

- 1) Générer $X \sim g$ et $U \sim \mathcal{U}([0,1])$
- 2) Accepter $Y = X$ si

$$U \leq \frac{f(X)}{Mg(X)}$$

- 3) Retourner en 1) si rejet

Probabilité d'acceptation

$$
\begin{aligned}
P[X \text{ accepté}] &= P\left[U \leq \frac{f(X)}{Mg(X)}\right] = E\left[\mathbb{I}_{\{U \leq \frac{f(X)}{Mg(X)}\}}\right] \\
&= E\left[E\left[\mathbb{I}_{\{U \leq \frac{f(X)}{Mg(X)}\}}\right] | X\right] \\
&= E\left[\frac{f(X)}{Mg(X)}\right] \\
&= \int \frac{f(x)}{Mg(x)} g(x) dx = \frac{1}{M}
\end{aligned}
$$

loi de X

$$
\begin{aligned}
P[X < x | X \text{ accepté}] &= \frac{P[X < x, X \text{ accepté}]}{1/M} \\
&= MP\left[X < x, U < \frac{f(X)}{Mg(X)}\right] \\
&= ME\left[\mathbb{I}_{\{X<x, U \leq \frac{f(X)}{Mg(X)}\}}\right] \\
&= ME\left[E\left[\mathbb{I}_{\{X<x, U \leq \frac{f(X)}{Mg(X)}\}}\right] | X\right] \\
&= ME\left[\mathbb{I}_{\{X<x\}} \frac{f(X)}{Mg(X)}\right] \\
&= \int_{-\infty}^{x} \frac{f(x)}{g(x)} g(x) dx = F(x)
\end{aligned}
$$

Remarques

- Cet algorithme permet de simuler une densité connue à une const. multiplicative près, e.g. $f(\theta|x) \propto f(x|\theta)\pi(\theta)$
- La probabilité d'acceptation est $1/M$ donc la valeur de M règle l'efficacité (vitesse) de l'algorithme
- Problème pour des densités à queues lourdes. Par exemple, on ne peut simuler une loi de Cauchy avec une loi de proposition normale (mais on peut faire l'inverse !)
- Utilisable pour un grand nombre de lois : $\mathcal{N}(0,1)$, $\mathcal{G}a(a,b)$, lois normales tronquées, ...

Exemple : Cauchy $\rightarrow$ Normale

- Loi cible

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$$

- Loi de proposition

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

- Choix de M

$$\frac{f(x)}{g(x)} = \sqrt{\frac{\pi}{2}}(1+x^2)e^{-x^2/2} \leq \sqrt{\frac{2\pi}{e}} = 1.52$$

valeur atteinte en ± 1. Proba d'acceptation $1/M \simeq 0.66$.

matlab: accept-reject pour différentes valeurs de M

Intégration par la méthode de Monte Carlo

On cherche à évaluer

$$E[h(\Theta)] = \int_{\mathcal{P}} h(\theta) f(\theta) d\theta,$$

où $\mathcal{P}$ est l'espace des paramètres, f est une densité connue et h est une fonction connue.

- Solution : générer un échantillon $(\theta_1, ..., \theta_n)$ distribué suivant f pour approcher cette intégrale :

$$E[h(\Theta)] \simeq \overline{h}_m = \frac{1}{m} \sum_{j=1}^{m} h(\theta_j),$$

- Justification : loi forte des grands nombres
- Erreur : $O\left(\frac{1}{\sqrt{n}}\right)$ (remember, curse of dimensionality!) Cste

Intervalles de confiance

- Variance :

$$v_m = \frac{1}{m^2} \sum_{j=1}^{m} [h(\theta_j) - \overline{h}_m]^2$$

- Loi asymptotique :

$$\frac{\overline{h}_m - E[h(\Theta)]}{\sqrt{v_m}} \sim \mathcal{N}(0,1)$$

☞ On peut déterminer des intervalles de confiance sur les paramètres inconnus !

Exemple : Fonction de répartition

- Définition :

$$F(\theta) = \int_{-\infty}^{\theta} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

- Approximation :

$$\widehat{F}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\theta_i < \theta},$$

où $(\theta_1, ..., \theta_n)$ est un échantillon généré avec l'algorithme de Box-Muller.

- Remarque : La variance de $\widehat{F}(\theta)$ est $\frac{F(\theta)[1-F(\theta)]}{n}$, e.g. $\frac{1}{4n}$ pour $\theta = 0$. Donc, pour avoir une précision de 10^{-4}, il faut un échantillon de taille $n = 200$ millions !

Échantillonnage d'importance

- Définition :

$$E[h(\Theta)] = \int_{\mathcal{P}} \left[h(\theta) \frac{f(\theta)}{g(\theta)} \right] g(\theta) d\theta,$$

  qui permet de simuler suivant g.

- Estimation : générer un échantillon $(\theta_1, ..., \theta_n)$ distribué suivant g pour approcher cette intégrale :

$$E[h(\Theta)] \simeq \frac{1}{m} \sum_{j=1}^{m} \frac{f(\theta_j)}{g(\theta_j)} h(\theta_j),$$

Choix de la loi de proposition

- Loi g simple à simuler
- Si le support de g contient celui de f, l'estimateur converge vers

$$\int_{\mathcal{P}} h(\theta) f(\theta) d\theta$$

- La variance de l'estimateur est finie si

$$E\left[h^2(\Theta)\frac{f(\Theta)}{g(\Theta)}\right] < \infty$$

- Éviter les lois de proposition telles que $\sup_{\theta\in\mathcal{P}} \frac{f(\theta)}{g(\theta)} = \infty$

  1) pb si le support de g n'est pas inclus dans celui de f,

  2) il existe une loi optimale minimisant la variance qui dépend de l'intégrale à calculer !

Exemple

Soit f la densité d'une loi de Student à ν degrés de liberté.
Calcul de

$$I = \int_a^\infty \theta^5 f(\theta) d\theta,$$

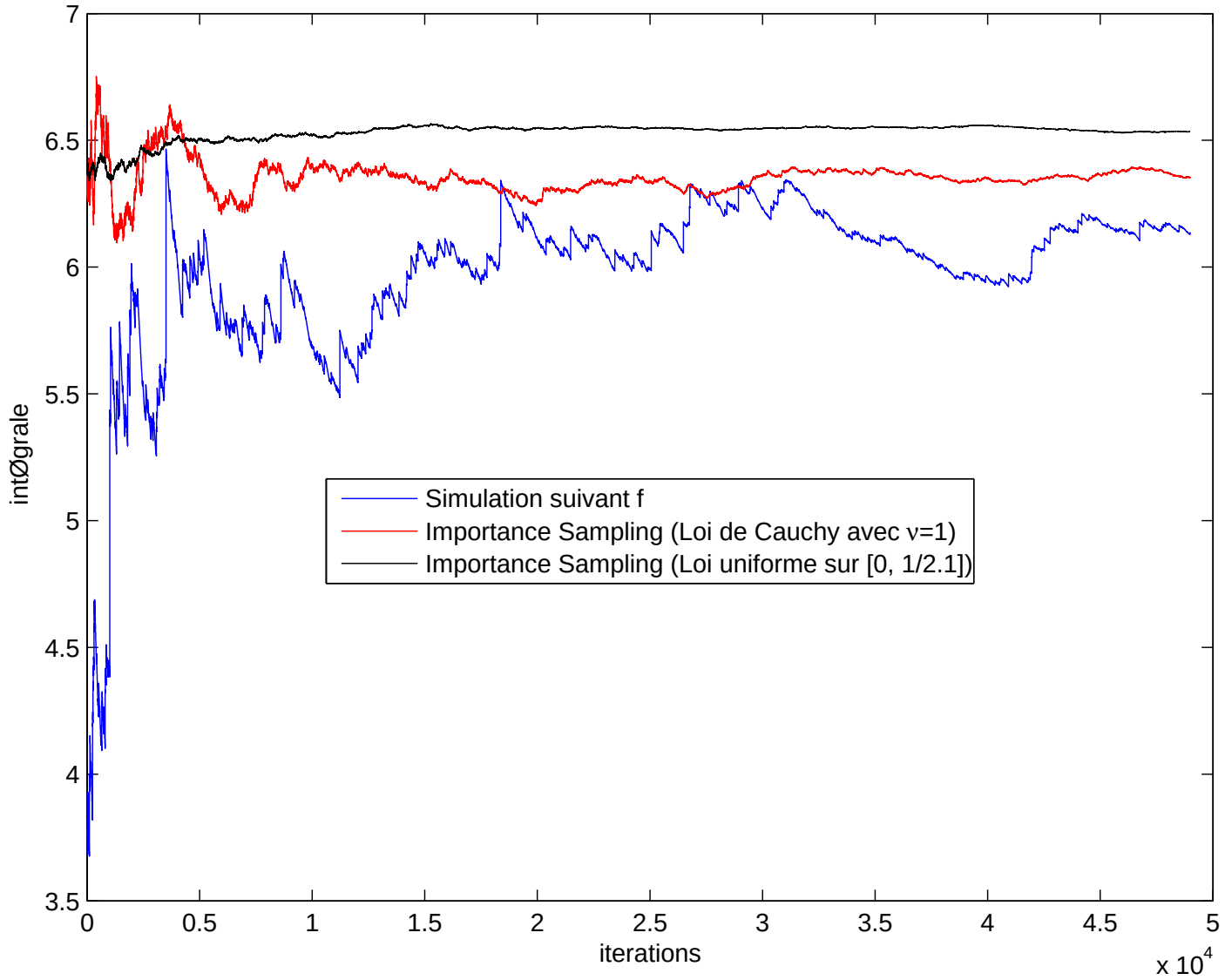
- Simulation suivant f
- Échantillonnage d'importance avec loi de Cauchy
- Un changement de variables $u = 1/\theta$ permet d'obtenir

$$I = \int_0^{\frac{1}{a}} a \frac{1}{au^7} f\left(\frac{1}{u}\right) du \simeq \frac{1}{a}\frac{1}{n}\sum_{i=1}^{n} \frac{1}{u_j^7} f\left(\frac{1}{u_j}\right),$$

où U suit une loi uniforme sur $[0, \frac{1}{a}]$.

matlab : integrale-student, $I = 6.54$, variance des estimées pour $n = 5000$

Exemple : $\nu = 12, a = 2.1$



Méthodes d'accélération

- Utiliser la corrélation pour diminuer la variance d'estimation. Soient deux échantillons $(\theta_1, ..., \theta_n)$ et $(\eta_1, ..., \eta_n)$ distribués suivant f. On a alors deux estimateurs non biaisés de $I = \int_{\mathbb{R}} h(\theta) f(\theta) d\theta$ définis par

$$\widehat{I}_1 = \frac{1}{n} \sum_{i=1}^{n} h(\theta_i), \quad \widehat{I}_2 = \frac{1}{n} \sum_{i=1}^{n} h(\eta_i)$$

- La variance de la moyenne de ces deux estimateurs est

$$\text{Var}\left(\frac{\widehat{I}_1 + \widehat{I}_2}{2} \right) = \frac{1}{4} \left(\text{Var}\widehat{I}_1 + \text{Var}\widehat{I}_2 \right) + \frac{1}{2} \text{Cov}(\widehat{I}_1, \widehat{I}_2)$$

☞ diminution de variance si la covariance est négative

Conditionnement - Rao-Blackwellization

- Espérances conditionnelles

$$E[h(\Theta)] = E\left[E[h(\Theta)|\Lambda]\right]$$

- Estimateurs

  Donc, si on sait calculer $g(\lambda) = E[h(\Theta)|\lambda]$, on en déduit deux estimateurs

$$\widehat{I}_1 = \frac{1}{n}\sum_{i=1}^{n} h(\Theta_i)$$

$$\widehat{I}_2 = \frac{1}{n}\sum_{i=1}^{n} g(\Lambda_i) = \frac{1}{n}\sum_{i=1}^{n} E[h(\Theta)|\Lambda_i]$$

Réduction de variance

Exemple

- Problème

$$I = \int_{-\infty}^{\infty} e^{-\theta^2} f(\theta) d\theta,$$

où f la densité d'une loi de Student à ν degrés de liberté.

- Estimateur usuel

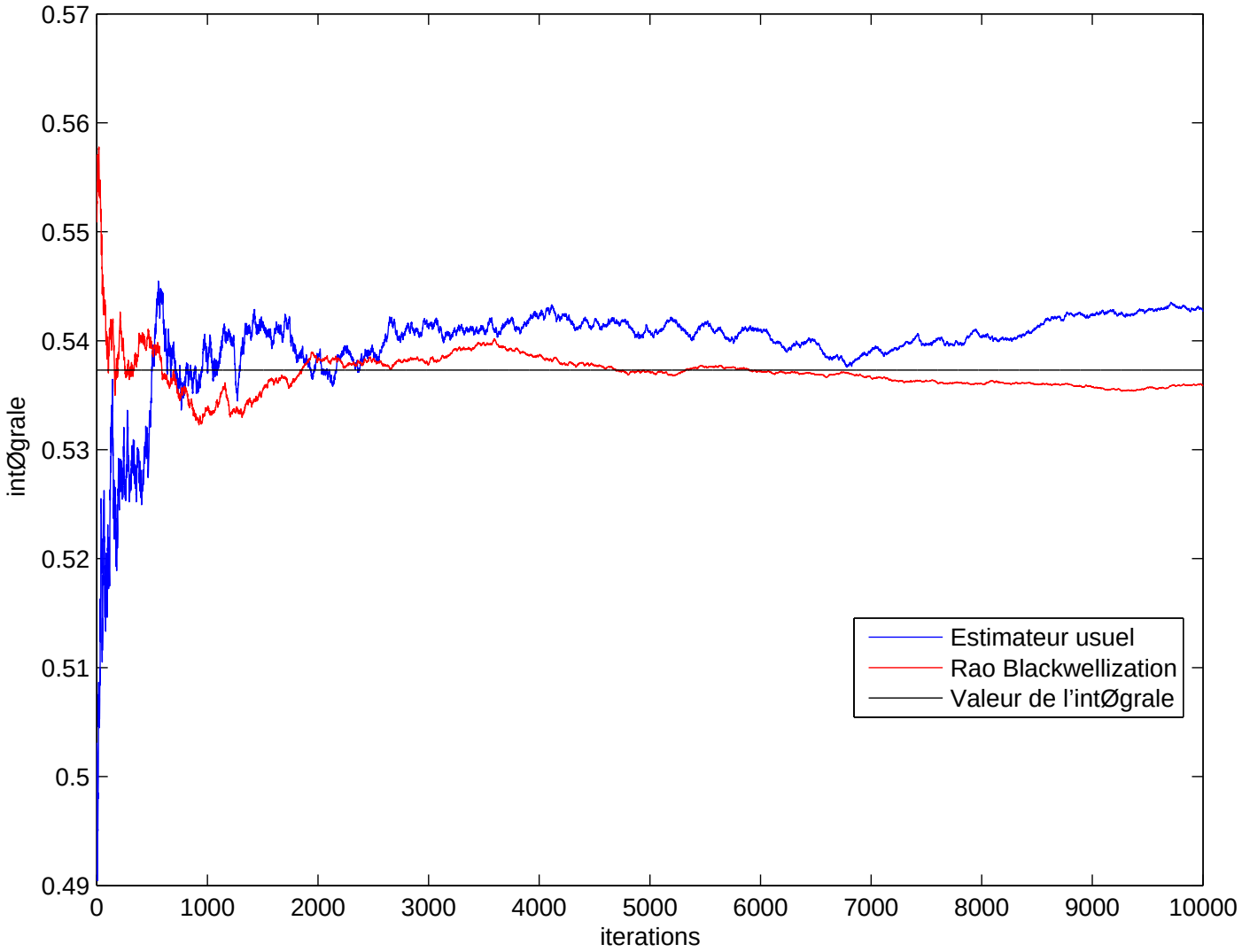
$$\widehat{I}_1 = \frac{1}{n} \sum_{i=1}^{n} e^{-\Theta_j^2}$$

- Réduction de variance $\Theta|\Lambda \sim \mathcal{N}(\mu, \sigma^2\Lambda)$ et $\Lambda^{-1} \sim \chi^2_\nu$

$$\widehat{I}_2 = \frac{1}{n} \sum_{i=1}^{n} E[e^{-\Theta^2}|\Lambda_i] = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\sigma^2\Lambda_j + 1}}$$

matlab : Raoblack, $I = 5373$

Exemple : $\nu = 4.6$, $\mu = 0$, $\sigma^2 = 1$



Méthodes déterministes d'optimisation

Pour résoudre une équation de la forme

$$f(\theta) = 0,$$

on peut utiliser des algorithmes comme l'algorithme de Newton-Raphson :

$$\theta_{n+1} = \theta_n + \left(\frac{\partial f}{\partial \theta}(\theta_n)\right)^{-1} f(\theta_n),$$

qui converge vers la solution $f(\theta) = 0$.

- convergence lente en $O(n^2)$ ou $O(n^3)$ alors que pour une méthode de simulation, on aura classiquement une convergence en $O(n)$!

Méthodes déterministes d'intégration

Pour calculer une intégrale de la forme

$$\int_a^b f(\theta)d\theta,$$

on peut utiliser des algorithmes basés sur les sommes de Riemann (méthode des trapèzes, méthode de Simpson, ...).

- On peut explorer des zones de faibles probabilités
- On a en général des problèmes pour des fonctions multi-modales.
- L'erreur est en $O\left(\frac{1}{n^{1/d}}\right)$, où d est la dimension de l'espace! (curse of dimensionality).

  Pour les méthodes de Monte-Carlo, on aura une erreur en $O\left(\frac{1}{\sqrt{n}}\right)$!