STA305 : Partie II
Calcul numérique pour l'analyse bayésienne

Boris Hejblum

ISPED M2 Biostatistique, Université de Bordeaux

Inserm BPH U1219 / Inria BSO, équipe SISTM

boris.hejblum@u-bordeaux.fr

https://borishejblum.science

Objectifs du cours

1. Comprendre les algorithmes d'échantillonnage et leur utilité
2. Comprendre le fonctionnement des algorithmes MCMC
3. Savoir utiliser le logiciel JAGS et en interpréter les sorties

Introduction

Une difficile estimation de la loi *a posteriori*

Intégration numérique – I

Applications réelles : $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$

$\Rightarrow$ la loi *a posteriori* conjointe des d paramètres

⚠ difficile à calculer :

- vraisemblance complexe
- constante d'intégration $f(\boldsymbol{y}) = \int_{\Theta^d} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$
- ...

Solution analytique rarement disponible

$\Rightarrow$ calcul numérique : intégrale de multiplicité d
– difficile lorsque d est grand (les problèmes numériques apparaissent dès que $d > 4$)

Intégration numérique – II

Des problèmes peuvent se poser même en dimension 1 !

Exemple :

Soit un échantillon $x_1,\dots,x_n$ *iid* selon une loi de Cauchy $\mathscr{C}(\theta,1)$ avec l'*a priori* $\pi(\theta)=\mathcal{N}(\mu,\sigma^2)$ (μ et σ connus)

$$\begin{aligned} p(\theta|x_1,\dots,x_n) &\propto f(x_1,\dots,x_n|\theta)\pi(\theta) \\ &\propto e^{-\frac{(\theta-\mu)^2}{2\sigma^2}} \prod_{i=1}^{n}(1+(x_i-\theta)^2)^{-1} \end{aligned}$$

⚠ La constante de normalisation ne peut se calculer analytiquement ⇒ impossible de donner une expression analytique pour cette loi *a posteriori*

Distributions marginales *a posteriori*

Objectif : tirer des conclusions à partir de cette distribution *a posteriori* conjointe

⇒ probabilité de toutes les valeurs possibles pour chaque paramètre (i.e. leurs distributions marginales, unidimensionnelles)

⚠ Reconstituer toute la densité a posteriori **numériquement** nécessite le calcul d'intégrales multidimensionnelles **pour chaque valeur possible du paramètre**

⇒ un calcul suffisamment précis de ces intégrales parait impossible

Distributions marginales *a posteriori*

Objectif : tirer des conclusions à partir de cette distribution *a posteriori* conjointe

⇒ probabilité de toutes les valeurs possibles pour chaque paramètre (i.e. leurs distributions marginales, unidimensionnelles)

⚠ Reconstituer toute la densité a posteriori **numériquement** nécessite le calcul d'intégrales multidimensionnelles **pour chaque valeur possible du paramètre**

⇒ un calcul suffisamment précis de ces intégrales parait impossible

Algorithmes basés sur des **simulations d'échantillonnage**
en particulier les méthodes de **Monte-Carlo par chaînes de Markov**
(*Markov chain Monte Carlo* – MCMC)

Solutions computationnelles

Théorème de Bayes $\Rightarrow$ loi *a posteriori*

Solutions computationnelles

Théorème de Bayes $\Rightarrow$ loi *a posteriori*

⚠ en pratique :

- une expression analytique rarement possible (cas bien particuliers)
- calcul de l'intégrale au dénominateur est souvent très difficile

Solutions computationnelles

Théorème de Bayes $\Rightarrow$ loi *a posteriori*

⚠ en pratique :

- une expression analytique rarement possible (cas bien particuliers)
- calcul de l'intégrale au dénominateur est souvent très difficile

Comment estimer la distribution *a posteriori* ?

$\Rightarrow$ générer un échantillon distribué selon la loi a posteriori

- **méthodes d'échantillonnage** directes
- méthodes de **Monte-Carlo par chaînes de Markov**

Méthode de Monte-Carlo

Monte-Carlo : von Neumann & Ulam
(*Los Alamos Scientific Laboratory* – 1955)

$\Rightarrow$ utiliser des nombres aléatoires pour estimer des quantités difficiles (ou impossible) à calculer analytiquement

Méthode de Monte-Carlo

Monte-Carlo : von Neumann & Ulam
(*Los Alamos Scientific Laboratory* – 1955)

⇒ utiliser des nombres aléatoires pour estimer des quantités difficiles (ou impossible) à calculer analytiquement

- **Loi des Grands Nombres**
- **échantillon** dit « **de Monte-Carlo** »

⇒ calculer divers fonctionnelles à partir de la distribution de l'échantillon

Exemple : On veut calculer $\mathbb{E}[f(X)] = \int f(x) p_X(x) dx$

Si $x_i \overset{iid}{\sim} p_X$, $\mathbb{E}[f(X)] = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$ (LGN)

⇒ si on sait échantillonner selon $p(x)$, on peut ainsi estimer $\mathbb{E}[f(X)]$ ...

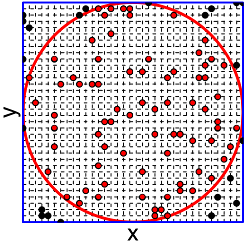
Méthode de Monte-Carlo : illustration

Estimation de π :

Méthode de Monte-Carlo : illustration

Estimation de π :

Une roulette de casino (à Monte-Carlo ?)



Un cible quadrillée en 36×36

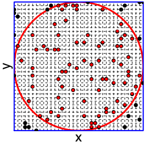
1. La probabilité d'être dans le cercle plutôt que dans le carré : $p_C = \frac{\pi R^2}{(2R)^2} = \frac{\pi}{4}$
2. n points $\{(x_{11}, x_{21}), \dots, (x_{1n}, x_{2n})\} = \{P_1, \dots, P_n\}$ dans le repère 36×36 (à l'aide de la roulette qui génère les coordonnées une à une)
3. Compter le nombre de points dans le cercle

$\Rightarrow$ Calculer le ratio (probabilité estimée d'être dans le cercle) : $\widehat{p}_C = \frac{\sum P_i \in cercle}{n}$

Méthode de Monte-Carlo : illustration

Estimation de π :

Une roulette de casino (à Monte-Carlo ?)



Un cible quadrillée en 36×36

Si $n = 1000$ et que l'on trouve 765 points dans le cercle : $\widehat{\pi} = 4 \times \frac{765}{1000} = 3,06$

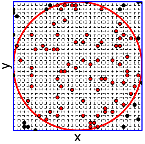
On peut améliorer l'estimation en augmentant :

- la résolution de la grille et aussi
- le nombre de points n : $\lim\limits_{n \to +\infty} \widehat{p}_C = p_C = \pi$ (LGV)

Méthode de Monte-Carlo : illustration

<u>Estimation de π :</u>

Une roulette de casino (à Monte-Carlo ?)



Un cible quadrillée en 36×36

Si $n = 1000$ et que l'on trouve 765 points dans le cercle : $\widehat{\pi} = 4 \times \frac{765}{1000} = 3,06$

On peut améliorer l'estimation en augmentant :

- la résolution de la grille et aussi
- le nombre de points n : $\lim_{n \to +\infty} \widehat{p}_C = p_C = \pi$ (LGV)

échantillon de Monte-Carlo $\Rightarrow$ calculer de nombreuses fonctionnelles
e.g. $\pi = 4$ fois la probabilité d'être dans le cercle