Travail d'Etude et de Recherche

Présenté par :

Soufiane Fadel

Master 1 : Sciences Mathématiques et Applications

Inférence bayésienne approchée Du Big Data

soutenu le : 05 Juin 2018

Encadré par :

Mr PUDLO PIERRE Professeur de l'Enseignement Supérieur à l'AMU

Année universitaire : 2017/2018

Remerciements

Je tiens tout d'abord à adresser mes remerciements les plus chaleureux à Monsieur Pierre Pudlo pour l'intérêt qu'il a porté à mon projet et pour nous avoir transmis sa grande passion pour la recherche. Son enthousiasme m'a poussé à me surpasser et sa présence attentive couplée à sa bienveillante disponibilité m'a été d'une très grande aide. J'ai découvert, grâce à lui c'est quoi un Data Scientist dont je n'avais pas d'idée professionnelle très précise. Aujourd'hui de je me sens plus professionnel, j'ai appris à faire de la vraie recherche, particulièrement j'ai commencé d'acquérir les techniques et les réflexes d'un programmeur et surtout la méthodologie d'attaquer un problème de programmation, la chose qui est vraiment très rare d'apprendre... Je tiens donc à lui témoigner mon estime, ma considération et ma reconnaissance.

Mes remerciements vont également à tous nos professeurs enseignants en Master "mathématiques et applications", notamment Madame Florence Hubert et Monsieur Oleg Lepski qui nous ont initiés respectivement à la modélisation mathématique et à la Statistique, modules, qui nous ont été particulièrement utile dans ce projet.

Je tiens également à témoigner toute ma reconnaissance à tous ceux qui ont contribué de près ou de loin à la réalisation de ce projet.

Finalement, nous dédie ce travail à mes chers parents, qui par leur soutien, leur sacrifice et leur amour inconditionnel, me poussent à me surpasser et dépasser mes limites.

"People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world".

Pedro Domingos.

Résumé

Les méthodes Monte Carlo par chaîne de Markov (MCMC) sont souvent jugées trop lourdes en calcul pour être utiles aux applications Big Data. Dans les scénarios particuliers où les données sont supposées indépendantes, différentes approches pour étendre l'algorithme de Metropolis-Hastings dans un contexte d'inférence bayésienne ont été récemment proposées dans l'apprentissage automatique et les statistiques computationnelles. Ce rapport introduit un cadre pour accélérer l'inférence bayésienne menée dans la présence de grands ensembles de données. Nous concevons une chaîne de Markov dont le noyau de transition utilise une fraction inconnue de taille fixe des données disponibles qui est rafraîchie aléatoirement tout au long de l'algorithme. Inspiré par la littérature sur le *calcul bayésien approximatif (ABC)*, le processus de sous-échantillonnage informé est guidé par la fidélité aux données observées, mesurée par des *statistiques sommaires.* L'algorithme résultant, sous-échantillonnage-informé-MCMC (ISS-MCMC : *Informed Sub-Sampling MCMC*) , est une approche générique et flexible contrairement aux méthodologies évolutives existantes,il préserve la simplicité de l'algorithme de Metropolis-Hastings, Même si l'exactitude est perdue, c'est-à-dire que la distribution de la chaîne se *rapproche* de la cible, nous étudions et quantifions théoriquement ce biais et montrons sur un ensemble varié d'excellentes performances lorsque le budget de calcul est limité.

Table des matières

Chapitre 1

INTRODUCTION

Effectuer des inférences sur de grands ensembles de données est un aspect majeur du défi du Big Data. Les modèles statistiques, et les méthodes bayésiennes en particulier, requièrent souvent des algorithmes de chaînes de Markov Monte Carlo (MCMC) pour faire des inférences, mais l'exécution des MCMC sur de tels ensembles de données est souvent beaucoup trop complexe pour être utile. Par conséquent, les méthodes MCMC telles que l'algorithme de Metropolis-Hastings (chapitre 3) ne peuvent pas être considérées pour un temps d'exécution raisonnable.

Dans ce rapport, nous proposons le MCMC par échantillonnage-informé , une nouvelle méthodologie qui vise à tirer le meilleur parti d'une ressource de calcul disponible pour un temps de calcul donné, tout en préservant la simplicité de l'échantillonneur Metropolis-Hastings standard. Étant donné un ensemble de données observées $(Y_1, \therefore, Y_N)$, une distribution a priori spécifiée p et une fonction de vraisemblance f, des paramètres à estimer $\theta \in \Theta$ du modèle procède par l'exploration de la distribution postérieure π défini sur $(\Theta, \mathcal{B}\Theta)$ (chapitre 2 est consacré à l'inférence bayésienne) par

$$\pi(d\theta|Y_1, ..., Y_N) \propto f(Y_1, ..., Y_N|\theta)p(d\theta).$$

L'espace d'état Θ est étendu avec un vecteur n-dimensionnel d'entiers uniques $U_k = \{1, 2....N\}$ qui permet identifier un sous-ensemble des données utilisées par le noyau de transition de Markov à la k-ième itération de l'algorithme, où $n \ll N$ est défini en fonction du budget de calcul disponible. Le point central de notre approche est le fait que chaque sous-ensemble est pondéré en fonction d'une mesure de similarité par rapport à l'ensemble des données par des statistiques résumées. La variable de sous-ensemble est rafraîchie aléatoirement à chaque itération selon la mesure de similarité. Le noyau de transition de la chaîne de Markov utilise seulement une fraction $\frac{n}{N}$ des données disponibles qui est par construction maintenue constante tout au long de l'algorithme. De plus, contrairement à la plupart des articles existant, notre méthode peut être appliquée à pratiquement n'importe quel modèle (impliquant des données i.i.d. ou non), car elle ne nécessite aucune hypothèse sur la fonction de vraisemblance ni sur la distribution a priori. Dans le cas particulier où les données sont i.i.d réalisations à partir d'un modèle exponentielle, nous montrons que, lorsque les statistiques récapitulatives sont définies comme les

statistiques exhaustives, cela donne une approximation optimale, dans le sens de minimiser une limite supérieure de *la divergence de Kullback-Leibler (KL)* entre π et la cible marginale de notre méthode. Dans le cas général, nous montrons que la définition des statistiques récapitulatives en tant qu'estimateur du maximum de vraisemblance permet de lier l'erreur d'approximation (en distance L_1) de notre algorithme.

L'objectif de ce projet sera de donner une perspective sur les méthodes MCMC, nous présenterons l'algorithme*MCMC par sous échantillonnage informé* qui permet dans des conditions vérifiables d'approcher π grâce à une approximation modulable de l'algorithme M-H où le budget de calcul de chaque itération est fixé (par la taille du sous-ensemble n). Pour ce faire, il est nécessaire de choisir les sous-ensembles en fonction d'une mesure de similarité par rapport à l'ensemble de données complet et non uniformément au hasard.

- ♣ Dans le chapitre 2, nous présenterons rapidement l'inférence bayésienne.
- ♣ Le chapitre 3 présente une revue général les chaînes de Markov à espace d'état continue.
- ♣ Le chapitre 4 est une étude profonde de l'échantillonneur Metropolis-Hastings.
- ♣ Dans le chapitre 5 nous fournissons des résultats théoriques concernant les modèles de familles exponentielles, que nous illustrons à travers un exemple du modèle probit.
- ♣ Le chapitre 6 permet de justifier nos motivations à l'appui de la méthodologie générale du sous-échantillonnage informé.
- ♣ Dans le chapite 7 nous étudions le noyau de transition de notre algorithme et montrons qu'il donne une chaîne de Markov ciblant, marginalement, une approximation de π et que l'erreur d'approximation est quantifiée dont nous fournissons des justifications théoriques pour la mise en place des paramètres de réglage du sous-échantillonnage informé, y compris le choix de la statistique récapitulative.
- ♣ Finalement, nous terminerons notre travail par le chapitre 8 qui présentera une illustration ...

Chapitre 2

L'INFÉRENCE BAYÉSIENNE

2.1 Introduction

Les algorithmes Monte Carlo par chaîne de Markov (MCMC) - tels que l'algorithme Metropolis Hastings et l'échantillonneur de Gibbs sont devenus extrêmement populaires dans la statistique, comme un moyen d'échantillonner à partir des distributions de probabilité de grandes dimensions. De plus l'existence d'algorithmes MCMC a transformé l'inférence bayésienne, en permettant aux statisticiens d'échantillonner à partir des distributions a posteriori des modèles statistiques compliqués. En plus de leur importance pour les applications statistiques et autres, ces algorithmes soulèvent également de nombreuses questions liées à la théorie des probabilités et aux chaînes de Markov. En particulier, les algorithmes MCMC impliquent des chaînes de Markov $\{X_n\}$ ayant une distribution stationnaire (compliquée) $\pi(.)$.

2.2 Principes

Cependant que les algorithmes MCMC sont utilisés dans de nombreux domaines (physique statistique, informatique), leur application la plus répandue est l'inférence statistique bayésienne.

2.2.1 Notations et définitions

L'ensemble des observations est noté $\mathbf{x}$. Ici $\mathbf{x} = (x_1, ...x_i, ..., x_n)$; autrement dit, on dispose d'un échantillon de taille n. Le cadre statistique de ce cours étant celui de la statistique inférentielle, les observations x_i sont donc considérées comme des réalisations de variables aléatoires, notées X_i.

- ♣ On entend par ***information a priori*** sur le paramètre θ toute information disponible sur θ en dehors de celle apportée par les observations
- ♣ L'information a priori sur θ est entachée d'incertitude (si ce n'était pas le cas, le paramètre θ serait connu avec certitude et on n'aurait pas à l'estimer !). Il est naturel de modéliser cette information a priori au travers d'une loi de probabilité, appelée ***loi a priori***. Sa densité est notée $\pi(\theta)$.

♣ Le modèle statistique paramétrique bayésien consiste en la donnée d'une loi a priori et de la loi des observations. On appelle ***loi des observations ou bien la vraisemblance*** , la loi conditionnelle de X sachant θ. Sa densité est notée $f(x|\theta)$,que la variable aléatoire X soit discrète ou continue. Si X est discrète, $f(x|\theta)$ représente $Pr(X = x|\theta)$. Par exemple, l'hypothèse que, les v.a. X_i sont indépendantes, sachant θ, on a :

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

Indiquons maintenant les autres lois de probabilité qui interviennent en statistique bayésienne.

♣ **La loi a posteriori** : C'est la loi conditionnelle de θ sachant $\mathbf{x}$. Sa densité est notée $\pi(\theta|x)$. En vertu de la formule de Bayes, on a :

$$\pi(\theta|x) = \frac{f(\theta|x)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}$$

♣ **La loi du couple** (θ, X). Sa densité est notée $h(\theta, x)$.On a donc :

$$h(\theta, x) = f(x|\theta)\pi(\theta)$$

♣ **La loi marginale de** X, (*appelé aussi l'évidence*). Sa densité est notée $m(x)$. On a donc : $m(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$.

2.2.2 La philosophie de l'approche bayésienne

Alors que la statistique classique repose sur la loi des observations, la statistique bayésienne repose sur la loi a posteriori. La loi a posteriori peut s'interpréter comme un résumé (en un sens probabiliste) de l'information disponible sur θ, une fois $\mathbf{x}$ observé. L'approche bayésienne réalise en quelque sorte l'actualisation de l'information a priori par l'observation $\mathbf{x}$, a travers de $\pi(\theta|x)$.

Le schéma ci-dessous résume la démarche bayésienne dans le cadre de la statistique paramétrique inférentielle. Il fait également apparaitre, la modélisation stochastique des x_i comme étant des réalisations de variables aléatoires X_i (cette modélisation est caractéristique de la statistique inférentielle), ainsi que la modélisation stochastique de l'information a priori disponible sur le paramètre θ, a travers de la loi a priori.

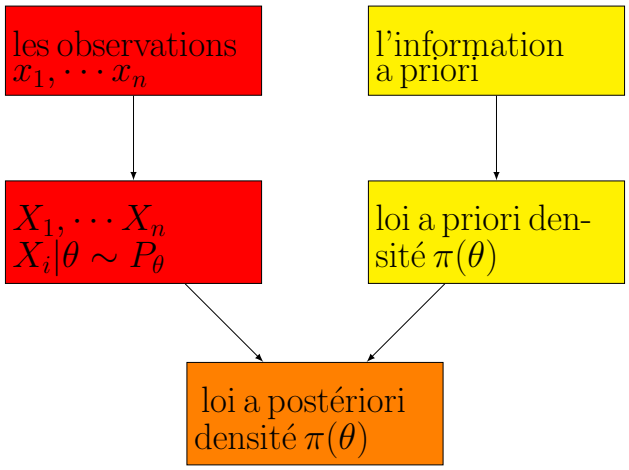


FIGURE 2.1 – Schéma de l'inférence bayésienne

2.2.3 Un abus de notation

♠ Au vu de ce qui précède, il apparait que la notation θ désigne tantôt une variable aléatoire, tantôt un paramètre. Cet abus de notation qui est fréquent (sinon systématique) dans les ouvrages de statistique bayésienne. On pourrait évidemment, distinguer la variable aléatoire et le paramètre par deux notations distinctes, en notant, par exemple, $\theta_{v.a}$. la variable aléatoire, et en réservant la notation θ au paramètre, comme dans le schéma ci-dessous.

grandeurs aléatoires	grandeurs non aléatoires
$\theta_{v.a}$	θ : le paramètre a estimer
$X_i\|\theta \sim P_\theta$	$(x_1, \cdots, x_i, \cdots, x_n)$: les observations

♠ $\pi(\theta|x) \propto f(\theta|X) \rightsquigarrow$ explicite à une constante de normalisation près

2.3 Le problème

Le problème abordé par les algorithmes MCMC est : On nous donne une fonction de densité π_u, sur un espace d'état $\mathcal{X}$, qui est peut-être non-normalisé mais au moins satisfait $0 < \int_{\mathcal{X}} \pi_u < \infty$. Typiquement $\mathcal{X}$ est un sous-ensemble ouvert de $\mathbb{R}^d$. Cette densité donne lieu à la mesure de probabilité $\pi(.)$ sur $\mathcal{X}$, par,

$$\pi(A) = \frac{\int_A \pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx} \tag{2.1}$$

Nous voulons estimer l'espérance des fonctions $f : \chi \longrightarrow \mathbb{R}$ en respectant à $\pi(.)$, c'est-à-dire que nous voulons estimer

$$\pi(f) = \mathbb{E}_\pi(f(X)) = \frac{\int_A f(x)\pi_u(x)dx}{\int_\chi \pi_u(x)dx} \tag{2.2}$$

Si χ est de grande dimension, et π_u est une fonction compliquée, alors l'intégration directe (analytique ou numérique) des intégrales dans (2) est infaisable. La solution classique de Monte Carlo à ce problème est de simuler au hasard des variables i.i.d $Z_1, Z_2, ..., Z_N \sim \pi(.)$, puis estimez $\pi(f)$ par :

$$\tilde{\pi}(f) = \frac{1}{N}\sum_{i=0}^{N} f(Z_i) \tag{2.3}$$

Le problème est donc si π_u est compliqué, alors il est très difficile de simuler directement des variables aléatoires i.i.d de $\pi(.)$.

La solution MCMC consiste à construire une chaîne de Markov sur χ qui s'exécute facilement sur un ordinateur, et qui a $\pi(.)$ comme une distribution stationnaire. Autrement dit, nous voulons définir des probabilités de transition de chaîne de Markov facilement simulées $P(x, dy)$ pour $x, y \in \chi$, telles que :

$$\int_{x\in\chi} \pi(dx)P(x, dy) = \pi(dy) \tag{2.4}$$

Ensuite, espérons-le , si nous simulons la chaîne de Markov pendant une longue période (commencée depuis n'importe où), alors pour un n assez grand, la distribution de X_n sera approximativement stationnaire :$\mathcal{L}(X_n) = \pi(.)$.Nous pouvons alors mettre $Z_1 = X_n$, puis relancer la chaîne de Markov pour obtenir $Z_2 = X_{n+1}, Z_3 = X_{n+2}, \cdots$, puis faire des estimations comme dans 5.2.

Remarque : bien sûr MCMC n'est pas le seul moyen d'échantillonner ou d'estimer à partir de distributions de probabilité compliquées. D'autres algorithmes d'échantillonnage possibles comprennent «Méthode de rejet» et «Échantillonnage préférentiel».

Chapitre 3

Chaine de Markov

Nous introduisons les notions fondamentales des chaînes de Markov et énonçons les résultats nécessaires pour établir la convergence de divers algorithmes MCMC.

3.1 Noyau de transition et chaine de Markov

Une chaîne de Markov est une séquence de variables aléatoires qui peut être considérée comme évoluant dans le temps, avec une probabilité de transition qui dépend de l'ensemble dans lequel se trouve la chaîne. Il semble donc naturel de définir la chaîne en fonction de son *noyau de transition*, fonction qui détermine ces transitions.

Définition 1 ***Noyau de transition*** *est une fonction K définie sur $\chi \times \mathcal{B}(\chi)$ telle que :*

- ♠ $\forall x \in \chi, K(x, .)$ *est une mesure de probabilité.*
- ♠ $\forall A \in \mathcal{B}(\chi), K(., A)$ *est mesurable.*

- Lorsque χ est discret, le noyau de transition est simplement une ***matrice (de transition K)*** avec des éléments :

$$P_{xy} = P(X_n = y | X_{n-1} = x), \quad x, y \in \chi.$$

  Dans ce cas, étant donné une distribution initiale $\mu = (w_1, w_2, ...)$, la distribution de probabilité marginale de X_1 est alors obtenue à partir de la multiplication matricielle : $\mu_1 = \mu K$, et par multiplication répétée on a :

$$X_n \sim \mu_n = \mu K^n.$$

- Dans le cas continu, le noyau indique également la densité conditionnelle $K(x, x')$ de la transition $K(x, .)$, c'est-à-dire $P(X \in A|x) = \int_A K(x, x')dx'$.

Définition 2 (X_n) *est une* ***chaine de Markov*** *si :*

$$\begin{aligned} P(X_{k+1} \in A | x_0, x_1 \cdots x_k) &= P(X_{k+1} \in A | x_k) \\ &= \int_A K(x_k, x) dx \end{aligned}$$

Remarque 1 *La chaîne est dite* ***homogène****, si la distribution de* $(X_{t_1}, \cdots, X_{t_k})$ *est la même que celle de* $(X_{t_1-t_0}, \cdots X_{t_2-t_0}, ..., X_{t_k-t_0})$ *pour un* x_0 *donné et pour chaque* k *et chaque* $(k+1)$*-uplet* $t_0 \leq t_1 \leq \cdots \leq t_k$

On va introduire la notion de ***temps d'atteinte de*** A un outil qui devient particulièrement utile dans l'évaluation de la convergence des algorithmes de Monte Carlo à chaîne de Markov.

Définition 3 *Considérons* $A \in \chi$*. Le premier* n *pour lequel la chaîne entre dans l'ensemble* A *est désigné par :*

$$\tau_A = \inf\{n \geqslant 1; X_n \in A\}$$

et est appelé le ***temps d'atteinte en*** A *avec, par convention,* $\tau_A = +\infty$ *si* $X_n \notin A$ *pour tout* n*. Nous définissons aussi*

$$\eta_A = \sum_{n=1}^{\infty} \mathbb{1}(X_n)$$

c'est ***nombre de passages*** *de* (X_n) *dans* A*.*

3.2 Marche aléatoire

Dans la configuration des algorithmes MCMC, les chaînes de Markov sont construites à partir d'un *noyau de transition* K , une densité de probabilité conditionnelle telle que $X_{n+1} \sim K(X_n, X_{n+1})$ · Un exemple typique est fourni par le processus de marche aléatoire , formellement défini comme suit.

Définition 4 (X_n) *est* ***une marche aléatoire*** *si :*

$$X_{n+1} = X_n + \epsilon_n$$

où ϵ_n *est généré indépendamment de* $X_n, X_{n-1}, ...$ *. Si la distribution de* ϵ_n *est symétrique autour de zéro, la séquence est appelée une* ***marche aléatoire symétrique****.*

Les marches aléatoires jouent un rôle clé dans de nombreux algorithmes MCMC, en particulier ceux basés sur l'algorithme de Metropolis-Hastings (voir Chapitre 4).

3.3 L'invariance

Les chaînes rencontrées dans les paramètres MCMC bénéficient d'une propriété de stabilité très forte, dite ***une distribution de probabilité invariante*** ; c'est-à-dire il existe une distribution π telle que : si $X_n \sim \pi$, alors $X_{n+1} \sim \pi$, plus formellement :

Définition 5 *Une mesure σ-finie π est dite **invariante** ou **stationnaire** pour le noyau de transition $K(.,.)$ si :*

$$\pi(B) = \int_{\chi} K(x, B)\pi(dx), \qquad \forall B \in \chi.$$

Remarque 2 *Lorsqu'il existe une mesure de probabilité invariante pour une chaîne ϕ-irréductible (définition 6), la chaîne est **positive**.*

3.4 Irréductibilité

Si le noyau K permet des mouvements libres dans tout l'espace d'états. (Cette liberté est appelée ***l'irréductibilité*** dans la théorie des chaînes de Markov et est formalisée dans la définition (6) comme l'existence de $n \in \mathbb{N}$ tel que $P(X_n \in A|X_0) > 0$ pour tout A tel que $\pi(A) > 0$.)

Dans le cas discret, la chaîne est irréductible si tous les états ***communiquent***, *c-à-d* : $(P_x(\tau_y < \infty) > 0 \quad \forall x, y \in \chi)$ avec τ_y étant la première fois que y est visité.

Dans de nombreux cas, $P_x(\tau_y < \infty)$ est uniformément égal à zéro, et donc il est nécessaire d'introduire une mesure auxiliaire ϕ sur $\mathcal{B}(\chi)$ pour définir correctement la notion d'irréductibilité.

Définition 6 *Étant donné une mesure ϕ, la chaîne de Markov (X_n) avec le noyau de transition $K(x, y)$ est **ϕ-irréductible** si, pour tout $A \in \mathcal{B}(\chi)$ avec $\phi(A) > 0$, il existe n tel que $K_n(x, A) > 0$ pour tout $x \in \chi$ (de manière équivalente, $P_x(\tau_A < \infty) > 0$). La chaîne est **fortement ϕ-irréductible** si $n = 1$ pour tous les A mesurables.*

Remarque 3

- *Cette propriété d'irréductibilité est la plus faible forme de stabilité stochastique.*
- *Si une chaîne de Markov a une distribution stationnaire et est irréductible, alors la distribution stationnaire est unique.*
- *L'irréductibilité implique que la loi des grands nombres tient.*
- *Quand on ne peut pas démontrer l'irréductibilité d'un schéma d'échantillonnage, on devrait trouver un schéma d'échantillonnage différent pour lequel on peut démontrer l'irréductibilité.*

3.5 Small set

Définition 7 *On dit que la chaîne de Markov (X_n) a* ***un atome*** *$\alpha \in \chi$ s'il existe une mesure non nulle ν telle que :*

$$K(x, A) = \nu(A), \quad \forall x \in \alpha, \forall A \in \mathcal{B}(\chi)$$

. Si (X_n) est ψ-irréductible, l'atome est accessible quand $\psi(a) > 0$.

Cette définition s'applique trivialement à toute valeur possible de X_n dans le cas discret, par contre cette notion est souvent trop forte pour être utile dans le cas continu car elle implique que le noyau de transition est constant sur un ensemble de mesures positives.

Une généralisation plus puissante est la condition dite de **minorisation** , à savoir qu'il existe un ensemble $C \in \chi$, une constante $\epsilon > 0$, et une mesure de probabilité ν telle que :

$$K(x, A) \geq \epsilon\nu(A) \quad \forall x \in \alpha, \quad \nu(A) \in \mathcal{B}(\chi) \tag{3.1}$$

La mesure de probabilité ν apparaît donc comme une composante constante du noyau de transition sur C. La condition de minorisation (19) conduit à la notion suivante.

Définition 8 *Un ensemble C est dit* ***small*** *s'il existe $m \in \mathbb{N}^*$, une mesure ν et $\epsilon > 0$ telle que :*

$$K^m(x, A) \geq \epsilon_m \nu_m(A) \quad \forall x \in C, \nu(A) \in \mathcal{B}(\chi)$$

Remarque 4 *Les visites de la chaîne sur un tel ensemble peuvent être exploitées pour créer des lots indépendants, puisque, avec la probabilité ϵ_m, la prochaine valeur de la chaîne de Markov $(X_{mn})_n$ est générée à partir de* ***la mesure de minorisation*** *ν_m indépendamment de X_0.*

3.6 La périodicité

Le comportement de (X_n) peut parfois être restreint par des contraintes déterministes sur les déplacements de X_n à X_{n+I}. Nous formalisons ces contraintes ici par la notion de la périodicité.

Dans le cas discret, **la période d'un état** $\omega \in \chi$ est définie comme :

$$d(w) = \textbf{p.g.c.d } \{m \geqslant 1; K^m(\omega, \omega) > 0\},$$

où **p.g.c.d** est le plus grand diviseur commun.

Remarque 5

- ★ *La valeur de la période est constante sur tous les états qui communiquent avec w.*
- ★ *Si la chaîne est irréductible (donc tous les états communiquent), alors il n'y a qu'une seule valeur pour la période.*
- ★ *Une chaîne irréductible est dite* ***apériodique*** *si elle est de période 1.*

L'extension au cas général nécessite l'existence d'un petit ensemble.

Définition 9 *On dit qu'une chaîne π-irréductible (X_n) a un* ***cycle*** *de longueur* ***d*** *s'il existe un small set C, un entier M, et une distribution de probabilité ν_M telle que* ***d*** *est le g.c.d. de :*

$$\{m \geqslant 1; \quad \exists \delta_m > 0 \text{ tel que } C \text{ est "un small set" pour } \nu_m \geqslant \delta_m \nu_M\}.$$

3.7 La Récursivité

La propriété d'irréductibilité (6) assure que la plupart des chaînes impliquées dans les algorithmes MCMC sont ***récurrentes*** (c'est-à-dire que le nombre moyen de visites à un ensemble arbitraire A est infini). Plus formellement :

Définition 10 *Une chaîne de Markov (X_n) est* ***récurrente*** *si :*

- ♠ *il existe une mesure ψ telle que (X_n) est ψ-irréductible, et*
- ♠ *pour tout $A \in \chi$ tel que $\psi(A) > 0$, $\mathbb{E}_x[\eta_A] = \infty$ pour chaque $x \in A$.*

Une forme plus forte de récurrence dite ***Harris récurrent***(c'est-à-dire tel que la probabilité d'un nombre infini de retours à A est 1).

Définition 11

- ♣ *Un ensemble A* ***est Harris-récurrent*** *si $P_x(\eta_A = \infty) = 1$ pour tout $x \in A$.*
- ♣ *La chaîne (X_n) est* ***Harris-récurrente*** *de s'il existe une mesure ψ ; tel que (X_n) est ψ-irréductible et pour tout ensemble A tel que $\psi(A) > 0$, A est récurrent de Harris.*

Remarque 6 *La récurrence de Harris garantit que la chaîne a le même comportement limite pour* ***chaque*** *valeur de départ au lieu de* ***presque chaque*** *valeur de départ. (Par conséquent, c'est l'équivalent des chaînes de Markov pour la notion de continuité des fonctions.)*

3.8 la norme de variation totale

Une ***distribution stationnaire*** est aussi ***une distribution limite*** dans le sens où la distribution limite de X_{n+1} est π sous ***la norme de variation totale***.

Définition 12 *Soient π et $\tilde{\pi}$ deux mesures de probabilité sur Θ. On appelle distance en variation totale entre π et $\tilde{\pi}$ la quantité :*

$$\|\pi - \tilde{\pi}\|_{TV} = \frac{1}{2} \int_{\Theta} |\pi(\theta) - \tilde{\pi}(\theta)| d\theta$$

3.9 ergodicité

Définition 13 *Pour une chaîne positive de Harris (X_n), avec une distribution invariante π, un atome est dit α est* ***ergodique*** *si :*

$$\lim_{n \to \infty} \|K^n(\alpha, \alpha) - \pi(\alpha)\|_{TV} = 0$$

Dans le cas dénombrable, l'existence d'un atome ergodique est suffisant pour établir la convergence selon la norme de variation totale (proposition 19)

Proposition 1 *Si (X_n) est Harris positif sur χ dénombrable, et s'il existe un atome ergodique $\alpha \in \chi$, alors, pour tout $x \in \chi$,*

$$\lim_{n\to\infty} \|K^n(x,.) - \pi\|_{TV} = 0$$

3.10 Ergodicité géométrique

Des formes de convergence plus fortes sont également rencontrées dans les paramètres MCMC, comme les ***convergences géométriques*** et ***uniformes***.

Définition 14 *Une chaîne de Markov est dite* ***géométriquement ergodique*** *s'il existe une constante $\rho < 1$ et une fonction non-négative $M(x)$ telle que :*

$$\|K^n(x,.) - \pi\|_{TV} \leq M(x)\rho^n \qquad \forall n \in \mathbb{N}$$

.

3.11 Ergodicité uniforme

Définition 15 *Une chaîne de Markov est dite* ***uniformément ergodique*** *s'il existe deux constantes $\rho < 1$ et $M < \infty$ telle que :*

$$\sup_{x\in\chi} \|K^n(x,.) - \pi\| \leq M\rho^n \qquad \forall n \in \mathbb{N}$$

pour tout n.

Remarque 7 *Une conséquence très intéressante de cette propriété de convergence est que la moyenne*

$$\frac{1}{N}\sum_{n=0}^{N} h(X_n) \tag{3.2}$$

converge vers l'espérance $\mathbb{E}_\pi[h(X)]$ presque sûrement.

3.12 La réversibilité et la condition d'équilibre détaillée

La propriété de stabilité des chaînes stationnaires peut être liée à une autre propriété appelée **réversibilité**.

Définition 16 *Une chaîne de Markov stationnaire (X_n) est dite **réversible** si la distribution de X_{n+1} conditionnellement sur $\{X_{n+2} = x\}$ est la même que la distribution de X_{n+l} conditionnellement sur $\{X_n = x\}$.*

En fait, la réversibilité peut être liée à l'existence d'une mesure stationnaire π si une condition un peu plus forte dite **la condition d'équilibre détaillée** (5.2).

Définition 17 *On dit qu'une chaîne de Markov avec le noyau de transition K satisfait **la condition d'équilibre détaillée** s'il existe une fonction f satisfaisant :*

$$K(y,x)f(y) = K(x,y)f(x) \qquad \forall x, y$$

Remarque 8 *Quand la chaîne est **réversible** , le Théorème Centrale Limite est aussi valable pour la moyenne (5.1).*

Théoreme 1 *Soit K le noyau de transition d'une chaîne de Markov satisfaisant la condition d'équilibre détaillée avec fonction de densité π, On a alors :*

- ★ *La densité π est la densité invariante de la chaîne.*
- ★ *La chaîne est réversible.*

preuve : Le premier point en utilisant la condition d'équilibre détaillée, pour tout ensemble mesurable B,

$$\begin{aligned} \int_\chi K(y,B)\pi(y)dy &= \int_\chi \int_B K(y,x)\pi(y)dxdy && (3.3)\\ &= \int_\chi \int_B K(x,y)\pi(x)dxdy && (3.4)\\ &= \int_B \pi(x)dx && (3.5) \end{aligned}$$

Pour le deuxième avec l'existence du noyau K et la densité invariante π, il est clair que l'équilibre détaillé et la réversibilité sont la même propriété.■

3.13 Théorème ergodique

Étant donné les observations $X_1, ..., X_n$ d'une chaîne de Markov, nous examinons maintenant le comportement limite des sommes partielles,

$$S_n(h)\frac{1}{n}\sum_{i=0}^{n} f(X_i) \tag{3.6}$$

quand n tend vers l'infini.

Théoreme 2 ***(Théorème ergodique)***
Si (X_n) a une mesure invariante σ-finie π, les deux points suivantes sont équivalentes :

- *si $f, g \in L_1(\pi)$ avec $\int g(x)\pi(x)dx \neq 0$, alors :*

$$\lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f(x)\pi(x)dx}{\int g(x)\pi(x)dx}$$

- *La chaine de Markov (X_n) est Harris-récurrente.*

Chapitre 4

Metropolis-Hastings

Ce chapitre est une introduction à l'algorithme de Metropolis-Hastings, un outil omniprésent pour produire des simulations dépendantes à partir d'une distribution arbitraire. On va illustrer les principes de la méthodologie sur des exemples simples avec des codes R.

4.1 Motivation

l'algorithme de Metropolis-Hastings est le cheval de bataille des méthodes MCMC, à la fois pour sa simplicité et sa polyvalence, et donc la première solution à prendre en compte dans les situations insolubles.

La principale motivation pour l'utilisation des chaînes de Markov est qu'elles fournissent des raccourcis dans les cas où l'échantillonnage nécessite trop d'efforts de la part de l'expérimentateur. Plutôt que de viser la **«grande image»** immédiatement, comme le ferait un algorithme d'acceptation-rejet (5.3), les chaînes de Markov construisent une image **progressive** de la distribution cible, procédant par exploration **locale** de l'espace d'état $\mathcal{X}$ jusqu'à ce que toutes les régions d'intérêt ont été découvertes.

Une analogie pour la méthode est le cas d'un visiteur d'un musée forcé par une panne générale de regarder une peinture avec une petite torche. En raison du faisceau étroit de la torche, la personne ne peut pas obtenir une vue globale de la peinture, mais peut continuer le long de cette peinture jusqu'à ce que toutes les parties ont été vues.

Définition 18 *La méthode Monte Carlo par chaîne de Markov (MCMC) pour la simulation d'une distribution π est une méthode produisant une chaîne de Markov ergodique $(X(t))$ dont la distribution stationnaire est π.*

4.2 L'algorithme

Supposons que nous avons une **densité de transition** (dite aussi **loi conditionnelle** où encore **loi instrumentale**) $q(x, y)$ tel que :

$$P(X_{n+1} \in A | X_n = x) = \int_A q(x, y)dy$$

On définit ***la probabilité d'acceptation*** par :

$$\begin{aligned} \alpha(x, y) &= \min\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\} \quad \pi(x)q(x, y) > 0 \\ &= 1 \qquad\qquad\qquad\qquad\quad \pi(x)q(x, y) = 0. \end{aligned}$$

Supposons que nous soyons à X_n. On prélève y de $q(x_n, y)$. On prend $X_{n+1} = y$ avec une probabilité $\alpha(x_n, y)$, sinon $X_{n+1} = x_n$. Le noyau de transition peut être représenter par :

$$p(x, y) = q(x, y)\alpha(x, y) + r(x)\delta_x$$

avec δ_x c'est la mesure Dirac en x, et

$$r(x) = 1 - \int q(x, y)\alpha(x, y)dy$$

est la probabilité marginale de rester à x.

Structure de l'algorithme de Metropolis-Hastings
pour $X_t = x_t$
1. générer $Y_t \sim q(y|x_t)$
2. faire

$$X_{(t+1)} = \begin{cases} Y_t & \text{avec probabilité} \quad \alpha\left(x_t, Y_t\right) \\ x_t & \text{avec probabilité} \quad 1 - \alpha\left(x_t, Y_t\right) \end{cases}$$

Remarque 9 *si la loi instrumentale $q(x, y)$ est symétrique c-à-d ($q(x, y) = q(y, x)$) on a alors :*

$$\alpha(x, y) = min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}$$

4.2.1 Metropolis-Hastings - Cas indépendant

on suppose que La loi de proposition $q(y|x_t)$ est indépendante de x_t

Structure de l'algorithme de Metropolis-Hastings indépendant
pour $X_t = x_t$
1.générer $Y_t \sim q(y|x_t)$
2. faire

$$X_{(t+1)} = \begin{cases} y_t & \text{avec probabilité} \quad min\left\{\frac{f(y_t)}{f(x_t)}\frac{q(x_t)}{q(y_t)}, 1\right\} \\ x_t & sinon \end{cases}$$

Remarque 10

- *L'échantillon généré n'est pas i.i.d.*
- *Cette méthode peut ne pas fonctionner correctement si q est très similaire à f.*

L'échantillonneur indépendant peut être comparé à la méthode de rejet pour générer des nombres aléatoires.

la méthode de rejet
1. Générer $Y \sim q(y)$
2. $X = Y$ avec probabilité $\frac{\pi(Y)}{cq(Y)}$.

L'algorithme indépendant de Metropolis-Hastings est plus efficace que la méthode de reject puisque M-H indépendant acceptera plus de valeurs proposées .
en effet si $q(y|x) = q(y)$ on a :

$$\alpha(x,y) = \min\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\} \tag{4.1}$$

$$= \min\{\frac{\pi(y)q(x)}{\pi(x)q(y)}, 1\} \tag{4.2}$$

Or on a

$$\frac{\pi(y)q(x)}{\pi(x)q(y)} \geq \frac{\frac{\pi(x)}{q(y)}}{max\left\{\frac{\pi(x)}{q(x)}\right\}}$$

donc la probabilité d'acceptation de M-H est plus grand que celle de la méthode de rejet .

4.2.2 Metropolis Hastings - Marche Aléatoire

En pratique une approche plus naturelle pour la construction de l'algorithme Metropolis-Hastings est de prendre en compte la valeur précédente simulée pour générer la valeur suivante, c'est-à-dire, envisager une exploration **locale** du voisinage de la valeur actuelle de la chaîne de Markov.
La loi de proposition q est alors de la forme : $Y_t = X_t + \epsilon_t$, où ϵ_t est une perturbation aléatoire de la distribution q indépendante de X_t, i.e $q(y|x) = q(y - x)$.

Lorsque la densité q est symétrique autour de zéro ; c'est-à-dire , satisfaire $g(-t) = g(t)$; on obtient l'algorithme suivant :

Structure de l'algorithme de Metropolis-Hastings-Marche Aléatoire
pour $X_t = x_t$
1.Générer $Y \sim q(y - x_t)$
2. faire

$$X_{(t+1)} = \begin{cases} y_t & \text{avec probabilité} \quad min\left\{\frac{f(y_t)}{f(x_t)}, 1\right\} \\ x_t & sinon \end{cases}$$

4.3 Propriétés de convergence

Pour voir que π est la distribution stationnaire de la chaîne Metropolis, nous examinons d'abord le noyau de Metropolis de plus près et constatons qu'il satisfait la propriété de balance détaillée.

Théoreme 3 *Soit (X_t) la chaîne produite par l'algorithme de Metropolis-Hastings. Pour chaque distribution conditionnelle q dont le support comprend celui de π on a :*

- ♠ *le noyau de la chaîne satisfait à la condition d'équilibre détaillée avec π ;*
- ♠ *π est une distribution stationnaire de la chaîne.*

preuve : pour le premier point il suffit de remarquer que :

$$\begin{cases} p(x,y) = q(x,y)\alpha(x,y) + r(x)\delta_x \\ q(x,y)\alpha(x,y)\pi(x) = q(y,x)\alpha(y,x)\pi(y) \\ r(x)\delta_x\pi(x) = r(y)\delta_y\pi(y) \end{cases}$$

avec δ_x c'est la mesure Dirac en x, et $r(x) = 1 - \int q(x,y)\alpha(x,y)dy$
Ces trois equations établissent la condition d'équilibre détaillé.
Pour le deuxième point c'est une conséquence direct du théorème (20). ■

Remarque 11 *La stationnarité de π est établie pour presque toute distribution conditionnelle q, ce qui indique l'universalité de cette algorithme.*

Nous pouvons établir le résultat de convergence suivant pour les chaînes de Markov Metropolis-Hastings , mais tout d'abord voila un lemme intéressant.

Lemme 1 *Si la chaîne de Metropolis-Hastings (X_t) est π-irréductible, alors elle est Harris récurrent.*

Théoreme 4 *Supposons que la chaîne de Markov Metropolis-Hastings (X_t) est π-irréductible.*

- ♣ *Si $h \in L_1(\pi)$, alors*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} h(X_n) = \int h(x)\pi(x)dx$$

- ♣ *Si, de plus, (X_t) est apériodique, alors :*

$$\lim_{n\to\infty} \left\| \int K^n(x,.)\mu(dx) - \pi \right\|_{TV} = 0$$

pour toute distribution initiale μ, où $K^n(x,\Delta)$ désigne le noyau pour n transitions

preuve : pour le premier point, (X_t) est π-irréductible par le lemme précédent, alors c'est Harris récurrente , ensuite par *(le théorème ergodique)* on a le résulta. Le deuxième point est une conséquence immédiate du théorème (*).

4.4 Exemple d'application

Pour comprendre le mécanisme de cette algorithme, on considérons un exemple élémentaire dont cette la densité cible est une version perturbée de la densité normale $\mathcal{N}(0,1)$, $\varphi(.)$,

$$\tilde{\pi}_i(x) = \sin^2(x) \times \sin^2(2x) \times \varphi(x)$$

Et notre loi conditionnelle est un noyau $U(x-\alpha, x+\alpha)$ uniforme :

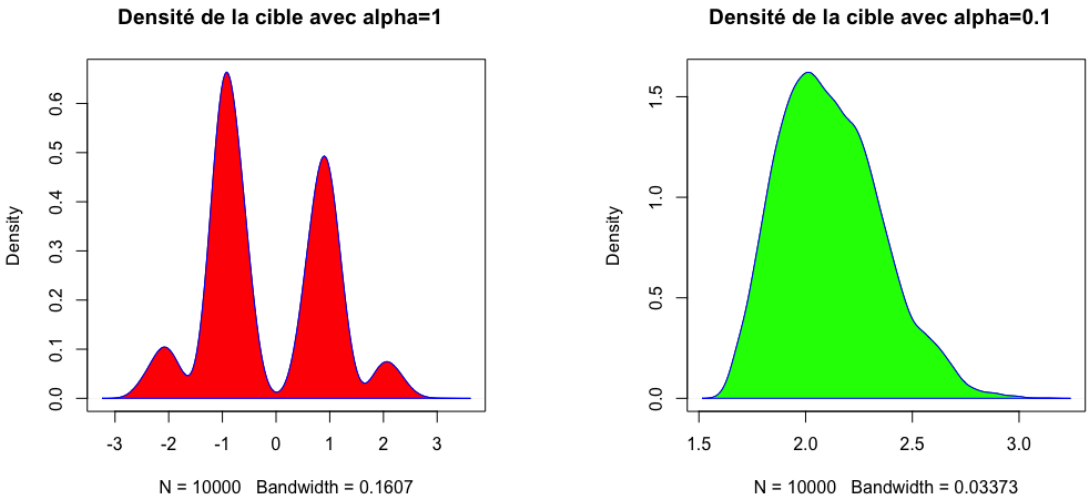
$$q(y|x) = 2\alpha(x-\alpha, x+\alpha)(y)$$

La mise en œuvre de cet algorithme est simple : deux fonctions à définir sont la cible et la transition

```
target=function(x){
  sin(x)^2*sin(2*x)^2*dnorm(x)}

metropolis=function(x,alpha=1){
  y=runif(1,x-alpha,x+alpha)
  if (runif(1)>target(y)/target(x)) y=x
  return(y)}


T=10^4
x=rep(3.14,T)
for (t in 2:T) x[t]=metropolis(x[t-1])
```



Remarque 12

- ♣ *Si nous changeons l'échelle de l'uniforme en $\alpha = 0,1$, la chaîne $(x(t))$ a des valeurs différentes que celles avec $\alpha = 1$, enfaite la densité de la figure à droite montre un mauvais ajustement de la cible parce qu'un seul mode est correctement exploré, ainsi que la proposition n'a pas le pouvoir de déplacer la chaîne suffisamment loin pour atteindre les autres parties du support de la loi cible.*
- ♣ *Un comportement similaire se produit lorsque nous commençons à par une valeur initial .*

Chapitre 5

Approche optimal de la posteriori pour les modèles exponentiels

Dans ce chapitre, nous considérons le cas de N observations indépendantes et identiquement distribuées à partir d'un modèle exponentiel.

L'échantillonnage à partir de la distribution a postériori d'un tel modèle à l'aide de l'algorithme de Metropolis-Hastings est simple car l'information véhiculée par les N observations est contenue dans le vecteur "*statistique exhaustive*"(19) , qui ne doit être calculé qu'une seule fois.

Définition 19 *Soit X un vecteur d'observation de taille n , dont les composantes X_i (i.i.d). Soit θ un paramètre influant sur la loi de probabilité à laquelle sont soumis les X_i. Une statistique $S(X)$ est dite* ***exhaustive*** *(pour le paramètre θ) si :*

$$\mathbb{P}(X = x|S(X) = s, \theta) = \mathbb{P}(X = x|S(X) = s)$$

En pratique l'on se sert peu de cette formule pour montrer qu'une statistique est exhaustive et l'on préfère en règle générale utiliser le théorème suivant appelé ***théorème de factorisation***.

Théoreme 5 *Soit $f_\theta(x)$ la densité de probabilité du vecteur d'observation X. Une statistique S est exhaustive si et seulement s'il existe deux fonctions g et h mesurables telles que :*

$$f_\theta(x) = h(x)g(\theta, S(x))$$

L'existence de statistique exhaustive dans ce type de modèle nous permet d'établir un certain nombre de résultats théoriques qui seront utilisés pour justifier notre méthodologie de MCMC par échantillonnage informé.

Plus précisément, les propositions 1 et 2 proposent une approximation optimale de la distribution a posteriori π par une distribution $\tilde{\pi}_n$ du paramètre d'intérêt donné seulement à un sous-échantillon de n observations. Enfin, la proposition 3

justifie l'introduction d'une distribution de probabilité sur l'ensemble des sous-échantillons.

5.1 Modélisation du problème

Soit $(Y_1, ..., Y_N) \in Y^N$ un ensemble d'observations i.i.d ($Y \in R^m, m > 0$) et on définit :

- $Y_{i:j} = (Y_i, ..., Y_j)$ si $1 \leq i \leq j \leq N$ et $Y_{i:j} = \emptyset$ *sinon*
- $Y_U = \{Y_k, k \in U\}$, avec $U \in \{1, ..., N\}$.

Nous supposons que le modèle de vraisemblance f appartient à la famille exponentielle et qu'il est entièrement spécifié par :

- ♦ un vecteur de paramètres $\theta \in \Theta$ ($\Theta \subseteq \mathbb{R}^d, d > 0$)
- ♦ Une fonction bornée $g : \Theta \mapsto S$ ($S \subseteq \mathbb{R}^s, s > 0$)
- ♦ Une statistique exhaustive $T : Y \mapsto S$

alors la densité de la distribution de vraisemblance par rapport à la mesure de Lebesgue est :

$$f(y|\theta) = \frac{\exp\{g(\theta)^t T(y)\}}{L(\theta)}, \qquad L(\theta) = \int_{y \in Y} \exp\{S(y)^t g(\theta)\} dy$$

La loi a posteriori π est définie sur l'espace mesurable $(\mathcal{B}(\Theta), \Theta)$ par la densité :

$$\pi(\theta|Y_{1:N}) = \frac{p(\theta)\frac{\exp\{\sum_{k=1}^N T(Y_k)^t g(\theta)\}}{L(\theta)^N}}{Z(Y_{1:N})} \tag{5.1}$$

avec

$$Z(Y_{1:N}) = \int_\Theta p(\theta)\frac{\exp\{\sum_{k=1}^N T(Y_k)^t g(\theta)\}}{L(\theta)^N} d\theta \tag{5.2}$$

p est la densité de la loi a priori définie sur $(\Theta, \times)$.
Pour tout $n \in \mathbb{N}$, nous définissons $\mathbb{U}_n$ comme l'ensemble des combinaisons possibles de n nombres entiers inférieurs ou égaux à N.
► Dans la suite, nous posons n comme constante et souhaitons comparer la distribution a posteriori π (5.1) avec n'importe quelle distribution de la famille $F_n = \{\tilde{\pi}_n(U), \ U \subseteq \mathbb{U}_n\}$, où pour tout $U \in \mathbb{U}_n$, nous avons défini $\tilde{\pi}_n(U)$ comme la distribution sur $(\Theta, \mathcal{B}(\Theta))$ avec une densité de probabilité :

$$\tilde{\pi}(\theta|Y_U) \propto p(\theta) f(Y_U|\theta)^{\frac{N}{n}} \tag{5.3}$$

5.2 Les sous-ensembles optimaux

Dans cette section on caractérise les sous-ensembles optimaux pour approcher π par $\tilde{\pi}_n$, pour cela on utilise une semi-distance dite **la divergence de Kullback-Leibler**.

Définition 20 *pour deux mesures π et $\tilde{\pi}$ définies sur le même espace mesurable $(\Theta, \mathcal{B}(\Theta))$, **la divergence de Kullback-Leibler (KL)** entre π et $\tilde{\pi}$ est définie comme :*

$$KL(\pi, \tilde{\pi}) = \mathbb{E}_\pi \left\{ \log \frac{\pi(\theta)}{\tilde{\pi}(\theta)} \right\}.$$

Remarque 13

- ⋆ *$KL(\pi, \tilde{\pi})$ est utilisé comme critère de similarité entre π et $\tilde{\pi}(\theta)$.*
- ⋆ *Il peut être interprété dans la théorie de l'information comme une mesure de l'information perdue lorsque $\tilde{\pi}(\theta)$ est utilisé pour approximer π.*

Avant que nous présentons le résultat principal de cette section voici une définition qui nous accompagnera tout au long de ce rapport.

Définition 21 *Pour tout sous-ensemble $U \in \mathbb{U}_n$, on définit **le vecteur de différence de statistiques suffisantes** entre l'ensemble de données Y et le sous-ensemble Y_U par :*

$$\Delta_n(U) = \sum_{k=1}^{N} S(Y_k) - \frac{N}{n} \sum_{k \in U} S(Y_k)$$

Proposition 2 *$\forall U \in \mathbb{U}_n$ on a :*

$$KL\{\pi, \tilde{\pi}_n(U)\} \leq B(Y, U),$$

avec

$$B(Y, U) = \log \mathbb{E}_\pi \exp \left\{ \|\mathbb{E}_\pi(g(\theta)) - g(\theta)\| \, \|\Delta_n(U)\| \right\}.$$

Le corollaire suivant est une conséquence immédiate de la proposition 5.4.

corollaire 1 *on définit l'ensemble :*

$$U_n^* := \left\{ U \in \mathbb{U}_n, \quad \frac{1}{N}\sum_{k=1}^{N} S(Y_k) = \frac{1}{n}\sum_{k\in U} S(Y_k) \right\} \tag{5.4}$$

si $U_n^* \neq \emptyset$ *alors* $\pi(\theta|Y) = \tilde{\pi}_n(\theta|Y_U)$ π*-presque surement.*

En supposant quelque hypothèses sur π voir(*), un résultat d'approximation très puissant, dit **théorème de Bernstein-von Mises** tient pour la concentration de π à son approximation normale :

$$\widehat{\pi}(.|Y_{1:N}) := \mathcal{N}(\theta^*(Y_{1:N}), \frac{I^{-1}(\theta_0)}{N}), \tag{5.5}$$

où $\mathcal{N}$ désigne la distribution normale, $\theta_0 = \arg\sup_{\theta\in\Theta} f(Y_{1:N}|\theta), \theta_0 \in \Theta$ un certain paramètre et $I(\theta)$ est la matrice d'information de Fisher.

Proposition 3 *Soit* $(U_1, U_2) \in \mathbb{U}_n^2$*. Supposons que pour tout* $i \in \{1, ..., d\}$*,* $\left|\Delta_n(U_1)^{(i)}\right| \leqslant \left|\Delta_n(U_2)^{(i)}\right|$*, avec* $\Delta_n(U_1)^{(i)}$ *est le* i*-ème élément de* $\Delta_n(U_1)$ *(21), alors on a :*

$$\widehat{KL_n}(U_1) \leqslant \widehat{KL_n}(U_2)$$

où $KL_n(U)$ *est la divergence de Kullback-Leibler entre l'approximation asymptotique* $\widehat{\pi}$ *(3) et* $\tilde{\pi}_n(U)$ *(5.3).*

5.3 Pondération des sous-échantillons

Considérons la distribution $\nu_{n,\epsilon}$, sur l'espace discret $\mathbb{U}_n$ défini pour tout $\epsilon > 0$ par :

$$\nu_{n,\epsilon} \propto \exp\left\{-\epsilon \left\|\Delta_n(U)\right\|^2\right\}, \quad \forall U \in \mathbb{U}_{\ltimes} \tag{5.6}$$

$nu_{n,\epsilon}$ attribue un poids à tout sous-ensemble en fonction de leur représentativité par rapport à l'ensemble de données complet.

- ♣ Si $\epsilon = 0$, $\nu_{n,\epsilon}$ est uniforme sur $\mathbb{U}_{\ltimes}$.
- ♣ Si $\epsilon \to \infty$, $\nu_{n,\epsilon}$ est uniforme sur l'ensemble des sous-ensembles qui minimisent $U \mapsto |\Delta_n(U)|$.

Remarque 14 *La proposition (2) suggère que pour les modèles exponentiels, l'inférence optimale basée sur des sous-échantillons de taille* n *est obtenue en choisissant* $\pi_n(U)$ *(5.3) en utilisant la distribution* $U \sim \nu_{n,\epsilon}$ *avec* $\epsilon \to \infty$*.*

5.4 Illustration avec un modèle probit

Le modèle probit est utilisé dans les problèmes de régression dans lesquels une variable binaire $Y_k \in \{0, 1\}$ est observée à travers la séquence d'expériences aléatoires indépendantes, définie pour tout $k \in \{1, ..., N\}$ comme :

i Tirer $X_k \sim \mathcal{N}(\theta^*, \sigma^2)$

ii Définir Y_k telle que :

$$Y_k = \begin{cases} 1 & si \quad X_k > 0 \\ 0 & sinon \end{cases} \tag{5.7}$$

On observons un grand nombre de réalisations $Y_1, ..., Y_N$, nous visons à estimer la distribution a postériori de θ. Si σ est inconnu, le modèle n'est pas identifiable et pour la simplicité nous l'avons considéré comme connu ici. La fonction de vraisemblance peut être exprimée comme :

$$f(Y_k|\theta) = \alpha(\theta)^{Y_k}(1 - \alpha(\theta))^{1-Y_k} = (1 - \alpha(\theta)) \left(\frac{\alpha(\theta)}{1 - \alpha(\theta)} \right)^{Y_k} \tag{5.8}$$

avec $\alpha(\theta) = \int_0^\infty (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}(t-\theta)^2\}dt$ et appartient clairement à la famille exponentielle, (par théorème de factorisation).

► Les densités de probabilités de la distribution a posteriori π et toute distribution $\tilde{\pi}_n(U) \in F_n$ s'écrivent respectivement :

$$\pi(\theta|Y1 : N) \propto p(\theta)(1 - \alpha(\theta))^N \left(\frac{\alpha(\theta)}{1 - \alpha(\theta)} \right)^{\sum_{k=0}^N Y_k}$$

$$\tilde{\pi}(\theta|Y1 : N) \propto p(\theta)(1 - \alpha(\theta))^N \left(\frac{\alpha(\theta)}{1 - \alpha(\theta)} \right)^{\frac{N}{n} \sum_{k \in U} Y_k}$$

où p est une densité a priori sur θ. Encore une fois, dans cet exemple, la densité a postériori est facile à évaluer point par point, même lorsque N est extrêmement grand, car il suffit de sommer toutes les variables binaires $Y_1, ..., Y_N$. En conséquence, les échantillons de π peuvent être obtenus de façon routinière par un algorithme M-H standard et de même pour toute distribution $\tilde{\pi}_n(U) \in F_n$.

Chapitre 6

MCMC par sous échantillonnage informé

Dans cette section, nous ne supposons aucune corrélation particulier pour la séquence d'observations, ni aucun modèle de vraisemblance spécifique et nous écrivons simplement la distribution a postériori comme :

$$\pi(\theta|Y_{1:N}) \propto p(\theta)f(Y_{1:N}|\theta) \tag{6.1}$$

La méthodologie MCMC par sous échantillonnage informé (MCMC-SEI) que nous décrivons maintenant peut être considérée comme une extension de l'approximation détaillée dans la section précédente à des modèles de familles non exponentielles avec éventuellement des observations dépendantes.

6.1 Motivation de notre approche

Le point central de notre approche est l'idée que les sous-échantillons ($Y_U \in \mathbb{U}_n$) ne sont pas tous valable pour estimer π .

- ♣ Ici, nous ne supposons pas l'existence d'une statistique exhaustive pour les modèles considérés. Afin de discriminer entre différents sous-échantillons, nous introduisons la notion de **la statistique résumé**

  $$R : Y_n \to S \ \ (n \leq N), S \subseteq \mathbb{R}^s$$

  .

  Le choix de la statistique résumé R est spécifique au problème et est censé être la contrepartie de la statistique exhaustive pour les modèles généraux.

- ♣ Parce que les statistiques utilisées pour évaluer la représentativité d'un sous-échantillon Y_U (relativement à l'ensemble de données complet Y) étant seulement résumées et non exhaustives , les résultats du chapitre précédent (ref*) ne sont plus valables.

  En particulier, si un sous-ensemble optimal U^* minimise la distance entre $R(Y_U)$ et $R(Y)$, déduire π à travers l'approximation $\tilde{\pi}_n(U^*)$ **n'est pas du tout optimale**.

Dans un tel contexte, il est raisonnable d'envisager d'étendre l'ensemble des sous-échantillons d'intérêt à un ensemble de bons sous-échantillons. Ceci suggère naturellement d'utiliser la distribution $\nu_{n,\epsilon}$ (5.6) pour discriminer entre les sous-échantillons, en remplaçant les statistiques exhaustives par des statistiques résumées et en tenant l'hypothèse ($\epsilon \to \infty$), afin d'avoir une collection de bons sous-échantillons.

6.2 L'algorithme MCMC-SEI

Le MCMC par sous-échantillonnage informé est une adaptation évolutive de l'algorithme de Metropolis-Hastings (chapitre 4), conçu pour les situations où $\mathbb{N}$ est trop grand pour effectuer une inférence sur la postériori π dans un temps raisonnable. MCMC-SEI s'appuie sur une chaîne de Markov dont le noyau de transition a une complexité de calcul bornée, qui peut être contrôlée via le paramètre n.
Pour éviter toute confusion, notons $\{\tilde{\theta}_i, i \in \mathbb{N}\}$ la séquence des paramètres généré par la chaîne de Markov sous-échantillonnée informée, contrairement à la chaîne de Markov produite par l'algorithme de Metropolis-Hastings (4.2).

L'ensemble de bons sous-échantillons utilisés dans l'inférence du MCMC-SEI est traité comme une séquence de données manquantes $U_1, U_2, ...$ et est donc simulé par notre algorithme.

Plus précisément, MCMC-SEI produit une chaîne de Markov $\{(\tilde{\theta}_i, U_i), i \in \mathbb{N}\}$ sur l'espace étendu $\Theta \times \mathbb{U}_n$.

Inspiré par l'analyse du chapitre 5 , la séquence des sous-échantillons $\{U_i, i \in \mathbb{N}\}$ est mise à jour aléatoirement d'une manière qui favorise les sous-ensembles dont le vecteur de statistique résumé est proche de celui de l'ensemble de données complet.

Soit W un noyau de transition symétrique sur $(\mathbb{U}_n, \mathcal{U}_n)$, avec $\mathcal{U}_n$ l'ensemble des parties de $\mathbb{U}_n$. Et Q un noyau de transition sur $(\Theta, \mathcal{B}(\Theta))$.

La transition $(\tilde{\theta}_i, U_i) \to (\tilde{\theta}_{i+1}, U_{i+1})$

i **(a)** Tirer un nouveau sous-ensemble $U \sim W(U_i, .)$

(b)

$$U_{i+1} = \begin{cases} U & \text{avec probabilité} \quad \beta(U_i, U) = \min\{b(U_i, U), 1\} \\ U_i & \text{avec probabilité} \quad 1 - \beta(U_i, U) \end{cases}$$

avec

$$b(U_i, U) = \exp\left\{\epsilon(\|\Delta_n(U_i)\|^2 - \|\Delta_n(U)\|^2)\right\} \tag{6.2}$$

ii **(a)** Tirer un nouveau paramètre $\tilde{\theta} \sim Q(\tilde{\theta}_i, .)$

(b)

$$\tilde{\theta}_{i+1} = \begin{cases} \tilde{\theta} & \text{avec probabilité} \quad \tilde{\alpha}(\tilde{\theta}_i, \tilde{\theta}) = \min\left\{\tilde{a}(\tilde{\theta}_i, \tilde{\theta}|U_{i+1}), 1\right\} \\ \tilde{\theta}_i & \text{avec probabilité} \quad 1 - \tilde{\alpha}(\tilde{\theta}_i, \tilde{\theta}) \end{cases}$$

avec

$$\tilde{a}(\tilde{\theta}_i, \tilde{\theta}|U_{i+1}) = \frac{\tilde{\pi}_n(\tilde{\theta}|Y_{U_{i+1}})}{\tilde{\pi}_n(\tilde{\theta}_i|Y_{U_i})} \frac{Q(\tilde{\theta}, \tilde{\theta}_i)}{Q(\tilde{\theta}_i, \tilde{\theta})} \tag{6.3}$$

Le pseudo-code suivant détaille comment simuler une chaîne de Markov par le sous échantillonnage informée.

Algorithm 1: l'algorithme MCMC-SEI

Entrée: l'état initial $(\tilde{\theta}_0, U_0)$ et les statistiques résumés $S_0 = \overline{R}(Y_{U_0}), R^* = \overline{S}(Y)$

1 **for** $i = 1, 2, ...$ **do**
2 proposer un nouveau sous-ensemble $U \sim W(U_{i-1}, .)$ et tirer $J \sim unif(0, 1)$, ;
3 calculer $R = \overline{S}(Y_U)$ et $b = b(U_i, U)$ définit dans (6.2) ;
4 **if** $J \leq b$ **then**
5 faire $U_i = U$ et $R_i = R$
6 **else**
7 faire $U_i = U_{i-1}$ et $R_i = R_{i-1}$
8 **end**
9 proposer un nouveau paramètre $\tilde{\theta} \sim Q(\tilde{\theta}_i, .)$ et tirer $I \sim unif(0, 1)$, ;
10 calculer $\tilde{\pi}_n(\tilde{\theta}_{i-1}|Y_{U_i})$, $\tilde{\pi}_n(\tilde{\theta}|Y_{U_i})$ et $\tilde{a} = \tilde{a}(\tilde{\theta}_i, \tilde{\theta}_{i-1}|U_i)$ définit dans (6.3) ;
11 **if** $I \leq \tilde{a}$ **then**
12 faire $\tilde{\theta}_i = \tilde{\theta}$
13 **else**
14 faire $\tilde{\theta}_i = \tilde{\theta}_{i-1}$
15 **end**
16 **end**

return : la chaine de Markov $\{ (\tilde{\theta}_i, U_i) \}$

Remarque 15 *Notez qu'à l'étape 10, si $U_i = U_{i-1}$, la quantité $\tilde{\pi}_n(\tilde{\theta}_{i-1}|U_i)$ est été déjà calculée à l'itération précédente.*

Chapitre 7

Analyse théorique du MCMC par sous échantillonnage informée

Par construction, l'algorithme MCMC-ESI échantillonne une chaîne de Markov sur un espace d'état étendu $\{(\tilde{\theta}_i, U_i), i \in \mathbb{N}\}$ mais le seul résultat utile de l'algorithme pour inférer π est le chaîne marginale $\{\tilde{\theta}_i, i \in \mathbb{N}\}$.

Dans ce chapitre, nous étudions la distribution marginal de la chaîne et on dénote par $\tilde{\pi}_i$ la distribution de la variable aléatoire $\tilde{\theta}_i$.

On note que $\{\tilde{\theta}_i, i \in \mathbb{N}\}$ est identique à la chaine de Metropolis-Hastings $\{\theta_i, i \in \mathbb{N}\}$, jusqu'à remplacer l'instant où on remplace α par $\tilde{\alpha}$ dans l'étape d'acceptation/rejet.

Ce changement, dont provient le gain de calcul de notre méthode, a des conséquences importantes sur la stabilité de la chaîne de Markov et, en particulier, implique que π n'est pas la distribution stationnaire de $\{\tilde{\theta}_i, i \in \mathbb{N}\}$. L'intérêt réside alors dans la quantification de la distance entre $\tilde{\pi}_i$ i et π.

Dans ce rapport, les résultats sont exprimés en distance de variation totale (définition 12).

7.1 Hypothèses

Soit K le noyau de transition exact de l'algorithme M-H, avec Q sa densité de transition. Q est fixé et défini comme un noyau de marche aléatoire avec un taux d'acceptation raisonnable. De plus par construction, K est π-réversible et donc π-invariant (théorème 3).

♣ **A 1. ergodicité géométrique**

Il existe une constante $\rho \in [0, 1[$ et une fonction $C : \Theta \to \mathbb{R}^+$ telle que $\forall(\theta_0, i) \in \Theta \times \mathbb{N}$

$$\left\|\pi - K(\theta_0, .)^i\right\| \leqslant C(\theta_0)\rho^i$$

♣ **A 2. Ergodicité uniforme**

Il existe deux constantes $\rho \in [0, 1[$ et $C < \infty$ telle que $\forall i \in \mathbb{N}$

$$\sup_{\theta_0 \in \Theta} \left\| \pi - K(\theta_0, .)^i \right\| \leqslant C\rho^i$$

♣ **A 3. Sous-ensembles I.I.D**

Les sous-ensembles $U_1, U_2, ...$ sont indépendants et identiquement distribué sous $\nu_{n,\epsilon}$

Remarque 16 *la probabilité de la transition $\tilde{\theta}_i \to \tilde{\theta}_{i+1}$ dépend de l'indice d'itération i. Ceci complique l'analyse de l'algorithme MCMC-SEI car la plupart des résultats sur la convergence des chaînes de Markov sont établis pour des chaînes de Markov homogènes. Pour simplifier, nous présentons dans cette section une analyse d'une légère variation d'un MCMC-SEI qui suppose l'indépendance entre les différents sous-ensembles $\{U_i, i \in \mathbb{N}\}$ (Hypothèse A.3).*

En **pratique**, l'hypothèse **A.3** est satisfaite lorsque les étapes (2)-(8) de l'algorithme MCMC-SEI sont répétées un grand nombre de fois pour simuler U_{i+1} sachant U_i.
Sous A.3, $\{\tilde{\theta}_i, i \in \mathbb{N}\}$ est une chaîne de Markov est homogène dont le noyau de transition $\tilde{K}_{n,\epsilon}$ est :

$$\forall(\tilde{\theta}, A) \in \Theta \times \vartheta, \quad \tilde{K}_{n,\epsilon}(\tilde{\theta}, A) = \sum_{u \in \mathbb{U}_n} K(\tilde{\theta}, A|u)\nu_{n,\epsilon}, \tag{7.1}$$

♣ **A4. Statistiques résumées**

Il existe une constante $\gamma_n < \infty$ telle que $\forall(\theta, U) \in \Theta \times \mathbb{U}_n$

$$|\log f(Y|\theta) - (N/n) \log f(Y_U|\theta)| \leqslant \gamma_n N \left\| \bar{S}(Y) - \bar{S}(Y_U) \right\| \tag{7.2}$$

Cette hypothèse est motivée à deux niveaux :

- Il est nécessaire d'avoir quelques hypothèses sur les statistiques résumées pour obtenir des résultats théoriques sur le MCMC-SEI en l'absence des statistiques exhaustives.
- Il offre un moyen de valider empiriquement le choix des statistiques résumées pour un modèle donné.

Remarque 17

★ *L'hypothèse* ***A.4*** *impose simultanément une condition sur le modèle f et la statistique résumées S. En particulier, elle* ***suppose que pour tout $\theta \in \Theta$, la variation de la vraisemblance des sous-échantillons pondérée $Y_U(U \in \mathbb{U}_n)$ autour de $f(Y|\theta)$ est contrôlé par la distance entre l'ensemble de données complet Y et le sous-échantillon Y_U mesuré à travers leurs statistiques résumées.***

★ *Cette hypothèse est peu probable si Θ n'est pas un compact.*

7.2 K est géométriquement ergodique

Notre résultat principal est que pour une taille suffisamment grande du sous-échantillon n, MCMC-ESI admet une distribution stationnaire.

Proposition 4 *Supposons que les hypothèses* ***A.1****,* ***A.3*** *et* ***A.4*** *sont valables, alors il existe un $n_0 \leqslant N$ tel que pour tout $n > n_0$, $\tilde{K}_{n,\epsilon}$ est également géométriquement ergodique pour tout $\epsilon > 0$*

Une conséquence directe de la Proposition 9 , (voir ** pour la preuve).

corollaire 2 *Pour n suffisamment grand, $\tilde{K}_{n,\epsilon}$ admet une distribution stationnaire qui cette converge vers π quand $n \to N$.*

7.3 K est uniformément ergodique

En plus d'admettre une distribution stationnaire pour un n suffisamment grand, nous montrons maintenant que sous hypothèse **d'ergodicité uniforme**, il est possible de quantifier le **taux de convergence**.

Définitions et Notations :

$$\phi_U(\theta) := \frac{f(Y_U|\theta)^{\frac{N}{n}}}{f(Y|\theta)} \quad \forall(\theta, U) \in (\Theta \times \mathbb{U}_n) \tag{7.3}$$

$$A_n := \mathbb{E}\left\{\sup_{\theta\in\Theta}\frac{1}{\phi_U(\theta)}\right\} = \sum_{U\in\mathbb{U}_n}\nu_{n,\epsilon}(U)\sup_{\theta\in\Theta}\frac{f(Y|\theta)}{f(Y_U|\theta)^{\frac{N}{n}}} \tag{7.4}$$

$$B_n(\theta,U) := \mathbb{E}\left\{a(\theta,\theta')|\phi_U(\theta)-\phi_U(\theta')|\right\} = \int Q(\theta,\theta')a(\theta,\theta')|\phi_U(\theta)-\phi_U(\theta')|d\theta'. \tag{7.5}$$

Proposition 5 *Supposons que les hypothèses **A.2**, **A.3** et **A.4** tiennent, alors il existe une constante $\kappa < 1$ telle que nous l'avons pour tout $i \in \mathbb{N}$*

$$\left\|K(\theta_0,.)^i - \tilde{K}_{n,\epsilon}(\theta_0,.)^i\right\| \leqslant \kappa \sup_{(\theta,U)\in\Theta\times\mathbb{U}_n} B_n(\theta,U) \tag{7.6}$$

et

$$\lim_{i\to\infty}\sup_{\theta\in\Theta}\left\|\pi - \tilde{K}_{n,\epsilon}(\theta_0,.)^i\right\| = \kappa A_n \sup_{(\theta,U)\in\Theta\times\mathbb{U}_n} B_n(\theta,U) \tag{7.7}$$

De plus, pour un sous-ensemble de taille n suffisamment grande , la chaîne de Markov marginale produite par MCMC-ESI admet une distribution invariante $\tilde{\pi}_n$ qui satisfait

$$\|\pi - \tilde{\pi}_n\| \leqslant \kappa A_n \sup_{(\theta,U)\in\Theta\times\mathbb{U}_n} B_n(\theta,U) \tag{7.8}$$

Remarque 18 *La borne supérieure de la proposition 5 n'est informative que si elle est inférieure à 1. Cette borne est le produit de deux espérances. Nous montrons maintenant comment ces deux espérances sont contrôlées respectivement par le choix du **noyau de transition** et le choix des **statistiques résumées.***

7.4 Choix du noyau de transition

En suppose que la densité de transition est une marche aléatoire gaussienne avec une matrice de covariance, B_n peut être exprimée comme : $B_n(\theta) = \sup_{U\in\mathbb{U}_n} D_1(U,\theta)$ où D_1 est définit par :

$$D_1(U,\theta) := \int \Phi_d(\zeta)\frac{\pi(\theta+\zeta)}{\pi(\theta)}|\phi_U(\theta)\phi_U(\theta+\zeta)|d\zeta, \tag{7.9}$$

où Φ_d est la distribution gaussienne standard de dimension $d = dim(\Theta)$.

Lorsque $N >> 1$, **le théorème de Bernstein-von Mises** (vu dans le chapitre 5), indique que, sous quelques conditions sur la fonction de vraisemblance, la distribution a posteriori peut être approximée par un gaussien avec une moyenne définie comme estimateur du maximum de vraisemblance θ^* et covariance $\frac{I^{-1}(\theta_0)}{N}$ où I est la matrice d'information de Fisher et $\theta_0 \in Theta$ un certains paramètres.

Remarque 19 *Puisque MCMC-SEI vise à échantillonner à partir d'une approximation de π,* ***prendre*** $\Sigma = \frac{1}{\sqrt{N}}M$ ***où*** M^tM ***est une approximation de*** $I^{-1}(\theta_0)$ ***est un choix raisonnable.***

La proposition 6 montre que dans ce scénario D_1 peut être ramené près de 0.

Proposition 6 *Sous l'hypothèse que le noyau de transition Q est une marche aléatoire gaussienne avec matrice de covariance $\Sigma = \frac{1}{\sqrt{N}}M$ alors :*

$$D_1(U,\theta) \leq \frac{\|\nabla_\theta \phi_U(\theta)\|}{\sqrt{N}} \left\{ \sqrt{\frac{2}{\pi}} \|M\|_1 + \frac{\|M\|_2^2 \|\nabla_\theta \log(\pi(\theta))\|}{\sqrt{N}} \right\} + \frac{d}{2N} \left|\!\left|\!\left| M^t \nabla_\theta^2 \phi_U(\theta) M \right|\!\right|\!\right| + \mathbb{E}\left\{ R\left(\frac{\|M\zeta\|}{\sqrt{N}} \right) \right\}$$

où $R(x) =_{x\to 0} o(x)$, et pour toute matrice carrée M de dimension $\mathbb{R}^d$, nous avons mis $\|M\|_1 := \sum_{1\leqslant i,j\leqslant d} |M_{i,j}|$, $\|M\|_2 := \left\{ \sum_{1\leqslant i,j\leqslant d} |M_{i,j}|^2 \right\}$ et $|||\cdot|||$ est la norme subordonnée.

Remarque 20 *Sous des hypothèses de régularité sur le modèle de vraisemblance, le gradient de $\log(\pi)$ et ϕ_U et la Hessienne de ϕ_U sont bornés et la limite supérieure de $D_1(U,\theta)$ peut être abaissée à 0, uniformément en (U,θ), lorsque $N >> 1$.*

7.5 Choix des statistiques résumées

La définition de A_n (7.3), montre que la probabilité des sous-échantillons à la puissance $\frac{N}{n}$ qui sont improbables contribueront à rendre A_n très grand.
Idéalement, le choix de la statistique résumé R doit garantir que les sous-échantillons Y_U ayant une très faible probabilité $f(Y_U|\theta)$ sont affectés à un poids $\nu_{n,\epsilon}(U) \approx 0$ afin limiter leur contribution. En d'autres termes, R doit être spécifié d'une manière qui empêche $f(Y_U|\theta)$ d'aller à 0 à un rythme plus rapide que $\nu_{n,\epsilon}(U)$. Ceci est garanti si l'hypothèse **A.4** est vérifiée. En effet, dans un tel cas :

$$\begin{aligned} A_n &:= \sum_{U\in\mathbb{E}_n(\theta)} \nu_{n,\epsilon}(U) \sup_{\theta\in\Theta} \frac{f(Y|\theta)}{f(Y_U|\theta)^{\frac{N}{n}}} + \sum_{U\in\mathbb{U}n\setminus\mathbb{E}_n(\theta)} \nu_{n,\epsilon}(U) \sup_{\theta\in\Theta} \frac{f(Y|\theta)}{f(Y_U|\theta)^{\frac{N}{n}}} \\ &\leqslant \nu_{n,\epsilon}(\mathbb{E}_n(\theta)) + \sum_{U\in\mathbb{U}_n\setminus\mathbb{E}_n(\theta)} \exp\left\{ \epsilon \|\Delta_n(U)\|^2 + \gamma_n(U) \|\Delta_n(U)\| - \log Z_n(\epsilon) \right\}, \end{aligned}$$

avec

$$\mathbb{E}_n(\theta)) = \left\{ U \in \mathbb{U}_n, \quad \sup_{\theta\in\Theta} \frac{f(Y|\theta)}{f(Y_U|\theta)^{\frac{N}{n}}} < 1 \right\} \tag{7.10}$$

$$Z_n(\epsilon) = \sum_{U\in\mathbb{E}_n(\theta)} \left\{ \|\Delta_n(U)\|^2 \right\}, \tag{7.11}$$

Remarque 21

- ♠ *Si ϵ est de même ordre de grandeur que γ_n, chaque terme de la somme reste borné quand $\|\Delta_n(U)\| \to \infty$.*
- ♠ *Si $\epsilon = 0$ équivaut à choisir $\nu_{n,\epsilon}$ comme distribution uniforme sur U_n, mais peut ne pas borner A_n.*

Les statistiques résumées potentielles peuvent être validées empiriquement en vérifiant qu'elles vérifient l'hypothèse **A.4**. Cette validation peut être effectuée graphiquement, en répétant les opérations suivantes pour un certain nombre de paramètres $\theta_k \sim_{i.i.d} p$:

(i) simuler des sous-ensembles $U_1, U_2, ...$ uniformément au hasard dans U_n,

(ii) tracer les points avec des coordonnées :

$$(x_{k,i}, y_{k,i}) = \left(\|\Delta_n(U_i)\|, \log f(Y|\theta_k) - \frac{N}{n} \log f(Y(U_i)|\theta_k)) \right).$$

Les statistiques sont validées s'il existe $\gamma_n < 1$ telle que les points $(x_{k,i}, y_{k,i})$ satisfont $\left|\frac{y_{k,i}}{x_{k,i}}\right| \leqslant \gamma_n$

Dans les situations où l'estimateur du maximum de vraisemblance $\theta^*(Y_{1:n})$ est facile et rapide à évaluer numériquement, nous recommandons de régler $\overline{R}(Y_{1:n}) = \theta^*(Y_{1:n})$.

Dans le cas d'observations indépendantes d'un modèle bien spécifié, la proposition suivante justifie les statistiques résumées comme l'estimateurs du maximum de vraisemblance, ce qui implique que l'hypothèse **A.4** est vérifier asymptotiquement à une constante prés.

Proposition 7 *Nous supposons que l'ensemble de données comprend $N = \rho n$ observations indépendantes et qu'il existe un θ_0 tels que $Y_i \sim f(.|\theta_0)$. Soit θ^* l'estimateurs du maximum de vraisemblance (EMV) de $Y_1, ..., Y_N$ et θ^*_U le EMV du sous-échantillon $Y_U(U \in \mathbb{U}_n)$. Ensuite, il existe une constante β, une métrique $\|\cdot\|_{\theta_0}$ sur Θ et une sous-suite croissante $\{\sigma_n\}_{n\in\mathbb{N}}, (\sigma \in \mathbb{N})$ telle que pour tout $U \subset \{1, 2, ..., \rho\sigma_n\}$ avec $|U| = \sigma_n$, nous avons pour p-presque tout θ dans un voisinage de θ_0 :*

$$\log f(Y_{1:\rho\sigma_n}|\theta) - \rho \log f(Y_U|\theta) \leqslant H_n(Y, \theta) + \beta + \frac{\rho n}{2} \|\theta^*_U - \theta^*\|_{\theta_0}, \tag{7.12}$$

avec

$$\lim_{n\to\infty} H_n(Y,\theta) \overset{\mathbb{P}_{\theta_0}}{=} 0$$

- ♠ L'hypothèse **A.4** est principalement utilisée dans les propositions 9 et 5 pour garantir que le rapport log-vraisemblance entre la vraisemblance et la vraisemblance échelonnée d'un sous-échantillon est borné.
- ♠ La constante β de la proposition 7 n'est pas une préoccupation majeure. En outre, il est facile de voir que cette constante disparaît lorsque la taille du sous-ensemble croît plus vite que l'ensemble de données complet, c'est-à-dire ($\rho \downarrow 1$), dans le régime asymptotique de l'équation. (7.12).
- ♠ Enfin, notons que, de manière similaire à toute méthode MCMC approchée, MCMC-SEI ne garantit pas la loi de grand nombre pour les fonctionnelles π-intégrables. Cependant, on supposons que la chaîne M-H **K** soit géométriquement ergodique, il est simple d'établir que pour une taille n suffisamment grande,

$$\lim_{i\to\infty} \left| \frac{1}{i} \sum_{j=1}^{i} f(\tilde{\theta}_j) - \pi f \right| \leqslant \|f\| \, \|\pi - \tilde{\pi}_{n,\epsilon}\| \qquad \text{Presque sûrement} \qquad (7.13)$$

avec $\pi f := \int f d\pi$ et $\|f\| = \sup \theta \in \Theta |f(\theta)|$.

Remarque 22 *Si, de plus, K est uniformément ergodique, la proposition 5 peut aider à lier l'erreur asymptotique qui apparaît dans l'estimation de πf.*

Chapitre 8

Illustration : estimation de modèle déformable dense

8.1 Introduction

le problème de l'estimation probabilistes des modèles déformables dans le domaine de la vision n'a pas encore reçu une formulation statistique cohérente et reste toujours un défi.

Notre but est de proposer un cadre statistique cohérent pour des modèles déformables denses à la fois en termes de modèle de probabilité et en termes de procédure d'estimation du modèle et de la structure de covariance de déformation. Le cadre est étendu à des mélanges de modèles, qui s'avèrent utiles pour modéliser des classes d'objets hétérogènes. Les observations sont modélisées sur une grille discrète fixe mais le gabarit et le champ de déformation sont définis sur des domaines continus. Pour simplifier, nous supposons un modèle de bruit gaussien additif, mais le cadre théorique et algorithmique peut être facilement généralisé à d'autres formes de modèles de données.

Nous ne paramétrons pas le modèle à travers ses valeurs sur la grille d'observation, mais plutôt comme une combinaison linéaire finie de noyaux définis sur le continuum. Le champ de déformation est défini sous une forme similaire et la structure de covariance se réduit à une matrice de covariance à dimension finie. L'estimation est formulée dans un cadre bayésien avec des a priori à la fois sur les paramètres du modèle et sur les paramètres de covariance. L'estimation est formulée comme un problème maximum a posteriori bien défini, avec des données manquantes-(les déformations).

8.2 Le modèle d'observation

Supposons qu'une image de niveau de gris y soit observée sur une grille de pixels x_s, $s \in \Lambda$, qui est incorporée dans un domaine continu $D \subset \mathbb{R}^2$ (typiquement $D = [-1, 1] \times [-1, 1]$). Bien que l'image soit observée uniquement aux pixels x_s, le modèle sera défini sur le continuum comme une fonction $I_0 : \mathbb{R}^2 \to \mathbb{R}$. Pour chaque observation y, nous supposons l'existence d'une déformation inobservée

$z : \mathbb{R}^2 \to \mathbb{R}^2$ telle que

$$y(s) = I_0 \{x_s - z(x_s)\} + \sigma\epsilon(s) \qquad\qquad s \in \Lambda$$

où $\epsilon(s)$ est **I.I.D** $\sim \mathcal{N}(0,1)$, indépendant de toutes les autres variables. On note zl_0 le vecteur d'observations du gabarit déformé aux points de la grille : $zl_0(s) = I_0 \{x_s - z(x_s)\}$, pour $s \in \Lambda$ de sorte que $y = zl_0 + \sigma\epsilon$.

8.2.1 Le modèle du gabarit - paramètres photométriques

La fonction de modèle $I_0 : \mathbb{R}^2 \to \mathbb{R}$ est autrement dénommée **le paramètre photométrique du modèle**, elle est supposé appartenir à un noyau reproducteur d'un espace de Hilbert V_p qui est déterminé par le noyau K_p. (L'indice p se référera désormais à tout ce qui est lié au paramètres photométriques du modèle.) Nous nous concentrons sur un sous-espace V_p fixe de dimension finie déterminé par un ensemble de points de repère $(x_{p,k})_{1 \leqslant k \leqslant k_p}$. Ces points couvrent généralement un domaine D_p contenant D car les déformations nécessitent parfois des valeurs de modèle qui sont en dehors du domaine observé. (Typiquement $D_p = [-1.5, 1.5] \times [-1.5, 1.5]$). Le gabarit est défini comme une combinaison linéaire des noyaux centrés sur les points de repère et un paramètre $\alpha \in \mathbb{R}^{k_p}$ alors :

$$I_\alpha = K_p\alpha, \qquad \text{avec } (K_p\alpha)(x) = \sum_{k=1}^{k_p} K_p(x, x_{p,k})\alpha(k) \tag{8.1}$$

8.2.2 Le modèle de déformation - paramètres géométriques

Nous utilisons le même cadre pour décrire le modèle de déformation qui est également désigné par **le paramètre géométrique** (indiqué par les indices g). Soit V_g un espace reproducteur du noyau de Hilbert avec un noyau K_g. Choisissons un ensemble fixe de points de repère $(x_{g,k})_{1 \leqslant k \leqslant k_g} \in D$. For $\beta = (\beta^{(1)}, \beta^{(2)}) \in \mathbb{R}^{k_g} \times \mathbb{R}^{k_g}$ nous définissons un champ de déformation par :

$$z_\beta(x) = (K_g\beta)(x) = \sum_{k=1}^{k_g} K_g(x, x_{g,k})(\beta^{(1)}(k), \beta^{(2)}(k)). \tag{8.2}$$

De nouveau, l'espace des déformations est un sous-espace de dimension finie de V_g.

Remarque 23

- ♣ *Dans les expériences ci-dessous, K_p et K_g seront des noyaux gaussiens radiaux, mais n'importe quel noyau lisse disparaissant à ∞ pourrait être utilisé.*
- ♣ *Les variables β sont supposées être gaussiennes avec une moyenne de 0*

8.2.3 paramètres et vraisemblance

Les paramètres d'intérêt sont :

- ♠ α et les coefficients qui déterminent le modèle (équation (8.1))
- ♠ l'écart type du bruit additif σ
- ♠ la matrice de covariance Γ des variables β qui déterminent la déformation (équation (8.2)).

Nous supposons que $\theta = (\alpha, \sigma, \Gamma)$ appartient à l'espace des paramètres Θ qui est défini comme l'ensemble ouvert :

$$\Theta = \left\{ \theta = (\alpha, \sigma, \Gamma) | \alpha \in \mathbb{R}^{k_p}, \sigma > 0, \Gamma \in \Sigma^+_{2k_g,*}(\mathbb{R}) \right\}$$

avec $\Sigma^+_{2k_g,*}(\mathbb{R})$ est l'ensemble des matrices symétriques strictement positives qui sont identifiées à travers sa partie triangulaire supérieure et sont donc considérées comme un sous-ensemble ouvert de $\mathbb{R}^{k_g(2k_g+1)}$.

La vraisemblance des données observées a la forme d'une intégrale par rapport aux paramètres de déformation non observés :

$$q(y|\theta) = \int g(y|\beta, \alpha, \sigma) h(\beta|\Gamma) d\beta$$

avec

$$\begin{aligned} h(\beta|\Gamma) &= \exp(-\beta^T \Gamma^{-1} \beta/2)(2\pi)^{-k_g} |\Gamma|^{-1/2} \sim \mathcal{N}(0, \Gamma) \\ g(y|\beta, \alpha, \sigma) &= \exp\left(-\frac{|y - z_\beta I_\alpha|^2}{2\sigma^2} \right) (2\pi\sigma^2)^{-|\Lambda|/2} \sim \mathcal{N}\left(I_\alpha(x. - z(x.)), \sigma^2 I \right) \end{aligned} \tag{8.3}$$

8.2.4 Le modèle bayésien

Même si les paramètres ont une dimension finie, l'estimateur du maximum de vraisemblance peut produire des estimations dégénérées lorsque l'échantillon d'apprentissage est petit.
Notre objectif est de démontrer que, avec l'introduction de distributions a priori sur les paramètres, l'estimation avec de petits échantillons est encore possible.

Nous utilisons des a priors conjugués standards :

- une Wishart inverse ν_p sur Γ,
- une Normale avec une matrice de covarianceΣ_p et de moyenne fixe μ_p sur α
- Et encore une Wishart inverse sur σ^2.

Tous les priors sont supposés indépendants. Plus formellement, nous avons :

$$\left.\begin{array}{c}(\Gamma,\alpha,\sigma^2)\sim\nu_g\bigotimes\nu_p\\ \beta_1^n\sim\bigotimes_{i=1}^n\mathcal{N}(0,\Gamma)|\Gamma,\alpha,\sigma,\\ y_1^n\sim=\bigotimes_{i=1}^n\mathcal{N}(z_{\beta_i}I_\alpha,\sigma^2Id_\Lambda|\beta_1^n,\alpha,\sigma,\Gamma\end{array}\right\} \tag{8.4}$$

avec

$$\begin{aligned}\nu_g(\Gamma)&\propto\left\{\exp\left(-\tfrac{\langle\Gamma^{-1},\Sigma_g\rangle}{2}\right)\frac{1}{\sqrt{|\Gamma|}}\right\}^{a_g}d\Gamma,\qquad a_g>2k_g+1\\ \nu_p(d\alpha,d\sigma^2)&\propto\left\{\exp\left(-\tfrac{\sigma_0^2}{2\sigma^2}\right)\frac{1}{\sqrt{\sigma^2}}\right\}^{a_p}\exp\left\{(\alpha-\mu_p)^T\Sigma_p^{-1}(\alpha-\mu_p)\right\}d\sigma^2d\alpha\end{aligned} \tag{8.5}$$

ou $\langle A,B\rangle=:tr(A^TB)$.

L'interprétation du modèle est simple. On génère (α,σ^2) à partir de ν_p et à partir du modèle du gabarit $I_\alpha=K_p\alpha$, ensuite on tire une matrice de covariance Γ à partir de ν_g. Après on tire indépendamment des β_i, $i=1...n$, de $\mathcal{N}(0,\Gamma)$. Les variables β_i déterminent les déformations z_{β_i} à travers l'équation (8.2). Finalement on génère $z_{\beta_i}I_\alpha$ et on ajoute un bruit gaussien (I.I.D) avec une variance σ^2 afin de construire l'observation y_i. le Graphe orienté acyclique ci-dessous résume cette procédure :

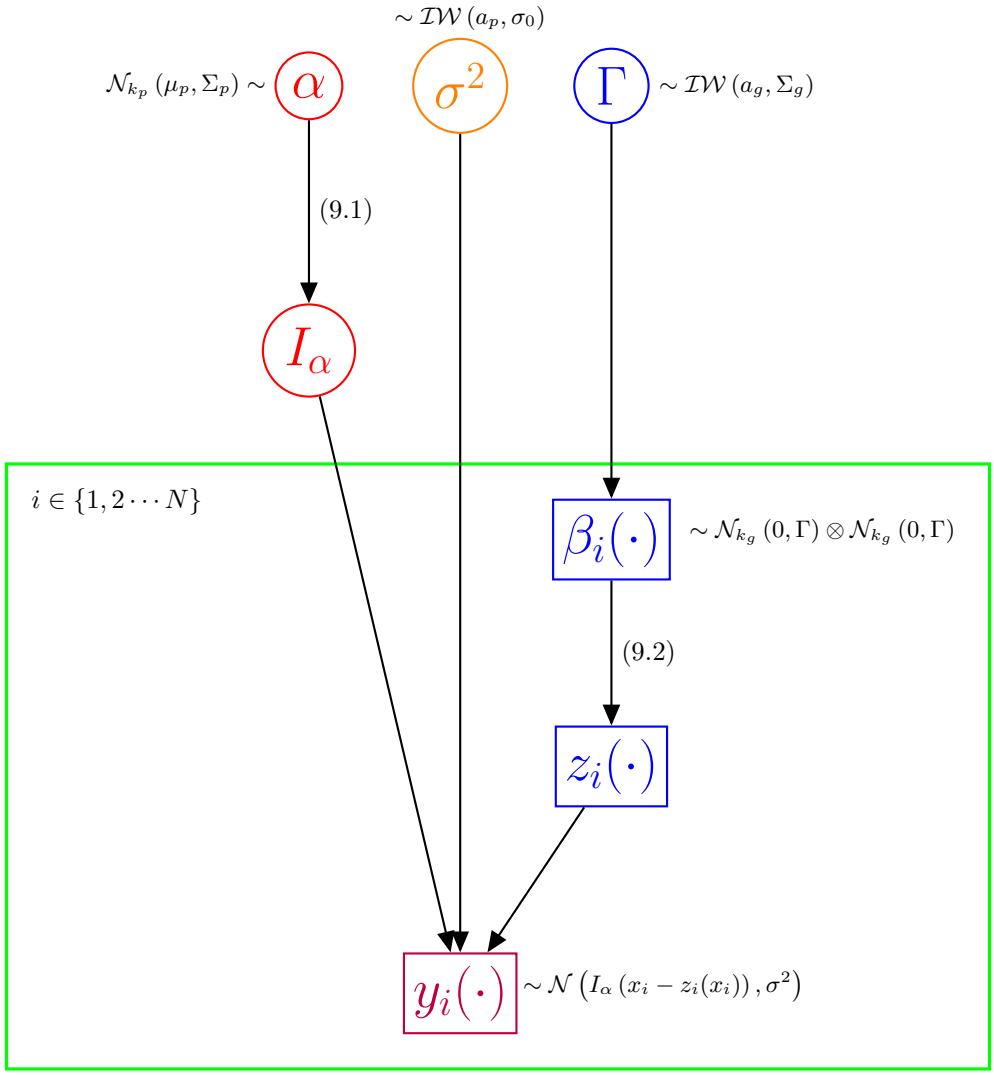


FIGURE 8.1 – Graphe orienté acyclique pour générer les observations

8.2.5 Choix des aprioris Gaussiennes

Un choix naturel pour les aprioris des matrices de covariance Σ_p et Σ_g est de considérer les matrices qui sont induites par la métrique des espaces V_p et V_g. On définit les matrices carrées :

$$\begin{aligned} M_p(k, k') &= K_p(x_{p,k}, x_{p,k'}) \qquad \forall 1 \leqslant k, k' \leqslant k_p, \\ M_g(k, k') &= K_g(x_{g,k}, x_{g,k'}) \qquad \forall 1 \leqslant k, k' \leqslant k_g, \end{aligned} \tag{8.6}$$

Définir $\Gamma = \Sigma_g = M_g^{-1}$ l'exposant dans la distribution définie dans la première densité de l'équation (8.3) correspond à la norme de la fonction $K_g\beta$ dans l'espace V_g . Fixer $\Sigma_p = M_p^{-1}$ l'exposant dans la distribution définie dans l'équation (5) correspond à la norme de $K_p\alpha$ in l'espace V_p. Cela a une justification plus précise que la restriction d'un linéaire gaussien aléatoire fonctionnels sur l'espace V_g ou V_p, vers le sous-espace couvert par K_g ou K_p respectivement et a l'avantage de définir un a priori essentiellement indépendant du nombre de points de repère k_p et k_g , et cela dépend seulement du choix global qui est fait pour les espaces de Hilbert du noyau reproducteur V_p et V_g. Dans ce contexte, le nombre de points de repère utilisés détermine un compromis entre la précision des approximations de fonctions dans les espaces respectifs et la quantité de calcul requise.

8.2.6 résultats pour M-H sur l'ensemble des "1"

Nous présentons ici une application de l'algorithme Metropoilis-Hasting sur l'ensemble des digits "1" a fin de répondre à la question s'agit il bien d'un "1" ? Ce problème est bien connu sous le nom "Digit Recognition ". Les images sont téléchargés de (http ://yann.lecun.com/exdb/mnist/).

Notre **but** est d'échantillonner :

$$\theta = \left(\alpha, \sigma^2, \Gamma, \beta_1 \cdots, \beta_N\right) \in \mathbb{R}^{k_p+1+k_g^2+N\times 2k_g}$$

Choix de la loi instrumentale de Metropolis-Hasting :

$$\begin{aligned}
\alpha_{t+1}|\alpha_t &\sim \mathcal{N}(\alpha_t, *) \\
\sigma^2_{t+1}|\sigma^2_t &\sim \mathcal{IW}(a_p, \sigma_0) \\
\Gamma_{t+1}|\Gamma_t &\sim \mathcal{IW}(a_g, \Sigma_g) \\
\beta_{i,t+1}|\beta_{i,t} &\sim \mathcal{N}(\beta_{i,t}, *)
\end{aligned}$$

Choix du langage C++
Rcpp est une bibliothèque dans R qui permet de valoriser des bouts de code C++ dans R. Ce choix est justifié par le fait que notre programme contient plusieurs boucles répétitives.

Annexes : Preuves des propositions

Preuve de la proposition 2

Preuve : Soit $n < N$ et U un sous-ensemble de $\{1, ..., N\}$ de cardinal n.

$$\tilde{f}_n(Y_U|\theta) = f(Y_U|\theta)^{\frac{N}{n}} = \left\{ \prod_{k \in U} f(Y_k|\theta) \right\}^{\frac{N}{n}} = \frac{\exp\{(N/n) \sum_{k \in U} T(Y_k)^t g(\theta)\}}{L(\theta)^N}$$

son a postériori correspondante est :

$$\tilde{\pi}_n(\theta|Y_U) = \frac{p(\theta) \frac{\exp\{(N/n) \sum_{k \in U} T(Y_k)^t g(\theta)\}}{L(\theta)^N}}{\tilde{Z}_n(Y_U}$$

avec

$$\tilde{Z}_n(Y_U) = \int_\Theta p(\theta) \frac{\exp\{(N/n) \sum_{k \in U} T(Y_k)^t g(\theta)\}}{L(\theta)^N} d\theta$$

alors $\forall \theta$ tel que $p(\theta) \neq 0$ on a :

$$\log \frac{\pi(\theta|Y_{1:N})}{\tilde{\pi}_n(\theta|Y_U)} = \left\{ \sum_{k \in U} T(Y_k) - (N/n) \sum_{k \in U} T(Y_k) \right\}^t g(\theta) + \log \frac{\tilde{Z}_n(Y_U}{Z(Y_{1:N})}$$

la divergence de Kullback-Leibler entre $\pi(\theta|Y_{1:N})$ et $\tilde{\pi}_n(\theta|Y_U)$, notée $\mathbf{KL}_n(U)$ est simplement :

$$KL_n(U) = \Delta_n(U)^t \mathbb{E}_\pi(g(\theta)) + \log \frac{\tilde{Z}_n(Y_U)}{Z(Y_{1:N})} \tag{9.1}$$

où $\Delta_n(U) = \sum_{k=1}^N S(Y_k) - \frac{N}{n} \sum_{k \in U} S(Y_k)$ d'autre part :

$$\begin{aligned}
\tilde{Z}_n(Y_U) &= \int_\Theta p(\theta)\frac{\exp\{(N/n)\sum_{k\in U} T(Y_k)^t g(\theta)\}}{L(\theta)^N}d\theta && (9.2)\\
&= \int_\Theta p(\theta)\frac{\exp\{\sum_{k=1}^n T(Y_k)^t - \Delta_n(U)\}g(\theta)}{L(\theta)^N}d\theta && (9.3)\\
&= \int_\Theta p(\theta)\tilde{f}_n(Y_{1:N}|\theta)\exp\{\Delta_n(U)\}^t g(\theta)d\theta && (9.4)\\
&= Z(Y_{1_N})\mathbb{E}_\pi\{\exp(\Delta_n(U)^t g(\theta))\}d\theta && (9.5)
\end{aligned}$$

par 9.1 et 9.5 on a :

$$\begin{aligned}
KL_n(U) &= \Delta_n(U)^t\mathbb{E}_\pi(g(\theta)) + \log\frac{\tilde{Z}_n(Y_U)}{Z(Y_{1:N})} && (9.6)\\
&= \log\frac{\mathbb{E}_\pi\{\exp(\Delta_n(U)^t g(\theta))\}}{\exp(-\Delta_n(U)\mathbb{E}_\pi(\theta))} && (9.7)\\
&= \log\mathbb{E}_\pi\exp\left[\{\mathbb{E}_\pi(g(\theta)-g(\theta))\}^t\Delta_n(U)\right] && (9.8)
\end{aligned}$$

Finalement, l'inégalité de Cauchy-Schwarz fournit la majoration en question :

$$KL_nU \leqslant \log\mathbb{E}_\pi\exp\left\{\|\mathbb{E}_\pi(g(\theta)) - g(\theta)\|\|\Delta_n(U)\|\right\} \quad (9.9)$$

□

Preuve de la proposition 3

Preuve : Sous certaines hypothèses faibles, le théorème de Bernstein-von Mises indique que $\pi(\cdot|Y_{1:N})$ est asymptotiquement (en N) une distribution gaussienne avec le maximum de vraisemblance θ^* comme moyenne et $\Gamma_N = \mathbb{I}^{-1}(\theta^*)/N$ comme covariance matrice, où $\mathbb{I}(\theta)$ est la matrice d'information de Fisher à θ. Notons Φ la densité de probabilité de $\mathcal{N}(\theta^*, \Gamma_N)$. Sous cette approximation, on a $\mathbb{E}_\pi(\theta) = \theta^*$, et à partir de (9.8), nous écrivons :

$$\begin{aligned}
\exp(KL_n(U)) &\approx \int \Phi(\theta)\exp\left[\{g(\theta^*) - g(\theta)\}^t\Delta_n(U)\right]d\theta && (9.10)\\
&= \int \Phi(\theta)\exp\left[\{g(\theta^*) - g(\theta)\}^t\Delta_n(U)\right]d\theta && (9.11)\\
& && (9.12)
\end{aligned}$$

Preuve de la proposition 9

Preuve : Afin de démontrer cette proposition il est nécessaire d'établir d'abord quelques lemmes. Pour la simplicité des notations notons que :

- La dépendance des quantités (n, ϵ) liées à MCMC-SEI est implicite.
- Pour tout $(\theta, U) \in \Theta \times \mathbb{U}_n$, on note $\phi_U(\theta) = \frac{f(y_U|\theta)^{\frac{N}{n}}}{f(y|\theta)}$
- On rappelle que $a(\theta, \theta')$ est le Taux d'acceptation (exact) de M-H de sorte que $\alpha(\theta, \theta') = \min(1, a(\theta, \theta'))$.
- $\mathbb{E}$ est l'espérance prise sous $\nu_{n,\epsilon}$.
- $\tilde{K}_{n,\epsilon}$ est écrit comme $\tilde{K}_n$.

Lemme 2 *$\forall(\theta, \theta') \in \Theta^2$ on a :*

$$\tilde{\alpha}(\theta, \theta') \leqslant \alpha\left(\theta, \theta'\left\{\max\left(1, \mathbb{E}\frac{\phi_n(\theta')}{\phi_n(\theta)}\right)\right\}\right)$$

Preuve :

$$\begin{aligned}
\tilde{\alpha}(\theta, \theta') &= \mathbb{E}\left\{\min\left[\frac{f(Y_U|\theta')^{\frac{N}{n}} p(\theta') q(\theta', \theta) f(Y|\theta) f(Y|\theta')}{f(Y_U|\theta)^{\frac{N}{n}} p(\theta) q(\theta, \theta') f(Y|\theta) f(Y|\theta')}, 1\right]\right\} \\
&\leqslant \min\left[1, a(\theta, \theta')\mathbb{E}\frac{\phi_U(\theta')}{\phi_U(\theta)}\right] \\
&\leqslant \min\left[1, a(\theta, \theta')\max\left\{\mathbb{E}\frac{\phi_U(\theta')}{\phi_U(\theta)}, 1\right\}\right] \\
&\leqslant \alpha(\theta, \theta')\max\left\{\mathbb{E}\frac{\phi_U(\theta')}{\phi_U(\theta)}, 1\right\}
\end{aligned}$$

Nous avons utilisé l'inégalité de Jensen et le fait que :

$$\min(1, ab) \leqslant \min(1, a)b \quad \forall a > 0 \text{ et } b \geqslant 1$$

. □

Lemme 3 $\forall(\theta,\delta)\in\Theta\times\mathbb{R}^+$ *on a :*

$$\tilde{r}(\theta)-r(\theta)\leqslant\delta+2\sup_{\theta\in\Theta}\mathbf{P}\left(|\phi_U(\theta)-1|\geqslant\frac{\delta}{2}\right)$$

avec $\begin{cases}r(x)=1-\int q(x,y)\alpha(x,y)dy\\ \tilde{r}(x)=1-\int q(x,y)\tilde{\alpha}(x,y)dy\end{cases}$

Preuve : Par l'inégalité $\min(1,ab)\leqslant\min(1,a)\min(1,b)\ \ \forall a>0$ et $b>0$ et en appliquant l'inégalité de Markov avec $0<\delta<1$,

$$\begin{aligned}
\tilde{r}(\theta) &= 1-\int q(\theta,\theta')\tilde{\alpha}(\theta,\theta')d\theta'\\
&\leqslant 1-\int q(\theta,\theta')\alpha(\theta,\theta')\mathbb{E}\left[\min\left\{1,\frac{\phi_U(\theta')}{\phi_U(\theta)}\right\}\right]d\theta'\\
&\leqslant 1-(1-\delta)\int q(\theta,\theta')\alpha(\theta,\theta')\mathbb{P}\left[\min\left\{1,\frac{\phi_U(\theta')}{\phi_U(\theta)}\right\}>1-\delta\right]d\theta'\\
&\leqslant 1-(1-\delta)\int q(\theta,\theta')\alpha(\theta,\theta')d\theta'\\
&\quad+\ (1-\delta)\int q(\theta,\theta')\alpha(\theta,\theta')\mathbb{P}\left[\min\left\{1,\frac{\phi_U(\theta')}{\phi_U(\theta)}\right\}\leqslant 1-\delta\right]d\theta'\\
&\leqslant 1-(1-\delta)(1-r(\theta))+\int q(\theta,\theta')\alpha(\theta,\theta')\mathbb{P}\left[\frac{\phi_U(\theta')}{\phi_U(\theta)}\leqslant 1-\delta\right]d\theta'
\end{aligned}$$

D'autre part :

$$\begin{aligned}
\mathbb{P}\left[\frac{\phi_U(\theta')}{\phi_U(\theta)}\leqslant 1-\delta\right] &\leqslant \mathbb{P}\left[\phi_U(\theta)\geqslant 1+\frac{\delta}{2}\right]+\mathbb{P}\left[\phi_U(\theta')\leqslant 1-\frac{\delta}{2}\right]\\
&\leqslant \mathbb{P}\left[\phi_U(\theta)\geqslant 1+\frac{\delta}{2}\right]+\mathbb{P}\left[\phi_U(\theta')\leqslant 1-\frac{\delta}{2}\right]\\
&\leqslant \mathbb{P}\left[\mid\phi_U(\theta)-1\|\geqslant\frac{\delta}{2}\right]+\mathbb{P}\left[\mid\phi_U(\theta')-1\|\leqslant\frac{\delta}{2}\right]\\
&\leqslant 2\sup_{\theta\in\Theta}\mathbb{P}\left[\mid\phi_U(\theta)-1\|\geqslant\frac{\delta}{2}\right]
\end{aligned}$$

finalement

$$\begin{aligned}
\tilde{r}(\theta) &= r(\theta)+\delta(1-r(\theta))+2\sup_{\theta\in\Theta}\mathbb{P}\left[\mid\phi_U(\theta)-1\|\geqslant\frac{\delta}{2}\right](1-r(\theta))\\
&= r(\theta)+\delta+2\sup_{\theta\in\Theta}\mathbf{P}\left(|\phi_U(\theta)-1|\geqslant\frac{\delta}{2}\right)\qquad\square
\end{aligned}$$

Lemme 4 *Supposons que l'hypothèse **A.4** est valide. On a alors :*

$$\sup_{(\theta,\theta')\in\Theta^2} \max\left[1, \mathbb{E}\left\{\frac{\phi_U(\theta)}{\phi_U(\theta')}\right\}\right] \leqslant \mathbb{E}\left\{\exp^{2\gamma\left\|\Delta\tilde{S}(u)\right\|}\right\}$$

Preuve : En utilisant l'inégalité de Cauchy-Schwartz $\forall(\theta,\theta') \in \Theta^2$:

$$\begin{aligned}
\mathbb{E}\left\{\frac{\phi_U(\theta)}{\phi_U(\theta')}\right\} &= \mathbb{E}\left\{\frac{f(y_U|\theta)^{\frac{N}{n}}}{f(y|\theta)}\frac{f(y|\theta')}{f(y_U|\theta')^{\frac{N}{n}}}\right\} \\
&\leqslant \mathbb{E}\left\{\frac{f(y_U|\theta)^{\frac{N}{n}}}{f(y|\theta)}\right\}^{1/2} \mathbb{E}\left\{\frac{f(y|\theta')}{f(y_U|\theta')^{\frac{N}{n}}}\right\}^{1/2} \\
(par\mathbf{A.4}) &\leqslant \mathbb{E}\left\{\exp(2\gamma_n\|\Delta\bar{S}(U))\|\right\}^{1/2} \mathbb{E}\left\{\exp(2\gamma_n\|\Delta\bar{S}(U))\|\right\}^{1/2} \\
&\leqslant \mathbb{E}\left\{\exp(2\gamma_n\|\Delta\bar{S}(U))\|\right\}
\end{aligned}$$

Finalement, pour conclure notant que pour tout a, b et c dans $\mathbb{R}$:

$$c > b \Longrightarrow \max(a,b) \leqslant \max(a,c)$$

□

Lemme 5 *Supposons N fixé et $n \to \infty$. On a alors :*

$$\mathbb{E}\left(\left\|\Delta\tilde{S}(U)\right\|\right) \to 0 \text{ et } \mathbb{E}\left(\exp 2\gamma\left\|\Delta\tilde{S}(U)\right\|\right) \to 1$$

Preuve : Il résulte du fait que quand $n \to N$, $\nu_{n,\epsilon}$ converge vers le dirac sur $U^\dagger = \{1, ..., N\}$ et donc,

$$\mathbb{E}\left(\left\|\Delta\tilde{S}(U)\right\|\right) \to \left\|\Delta\tilde{S}(U^\dagger)\right\| = 0$$

et

$$\mathbb{E}\left(\exp 2\gamma\left\|\Delta\tilde{S}(U)\right\|\right) \to \exp 2\gamma\left\|\Delta\tilde{S}(U^\dagger)\right\| = 1$$

□

Nous pouvons maintenant prouver la Proposition 4 :

Supposons que les hypothèses **A.1**, **A.3** et **A.4** sont valables, alors il existe un $n_0 \leqslant N$ tel que pour tout $n > n_0$, $\tilde{K}_{n,\epsilon}$ est également géométriquement ergodique pour tout $\epsilon > 0$

Preuve :
D'après (Meyn et Tweedie, 2009, Theorems 14.0.1 et 15.0.1), il existe une fonction $V : X \to [1, \infty[$, deux constantes $\lambda \in (0,1)$ et $b < \infty$ et un small set $S \in X$ tel que K vérifie la condition de Drift :

$$KV \leq \lambda V + b\mathbb{1}_S$$

Nous montrons maintenant comment utiliser les lemmes précédents pour établir l'ergodicité géométrique de $\tilde{K}_n$ pour certains n suffisamment grands.

$$\begin{aligned}
(\tilde{K}_n - K)V(\theta) &= \int Q(\theta, d\theta')(\tilde{\alpha}(\theta,\theta' - \alpha(\theta,\theta'))V(\theta') + (\tilde{\rho}(\theta) - \rho(\theta))V(\theta) \\
&\leq \left(\mathbb{E}\left(\exp 2\gamma \left\|\Delta\tilde{S}(U)\right\|\right) - 1\right) \int Q(\theta, d\theta' \alpha(\theta,\theta'))V(\theta') \\
&+ \left(\delta + \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\left(\left\|\Delta\tilde{S}(U)\right\|\right)\right) V(\theta) \\
&\leq \left(\mathbb{E}\left(\exp 2\gamma \left\|\Delta\tilde{S}(U)\right\|\right) - 1\right) (\lambda V(\theta) + b\mathbb{1}_S(\theta) - \rho(\theta)V(\theta)) \\
&+ \left(\delta + \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\left(\left\|\Delta\tilde{S}(U)\right\|\right)\right) V(\theta) \\
&\leq \mathbb{E}\left(\exp 2\gamma \left\|\Delta\tilde{S}(U)\right\|\right) b\mathbb{1}_S(\theta) \\
&+ \left(\lambda\left(\mathbb{E}\left(\exp 2\gamma \left\|\Delta\tilde{S}(U)\right\|\right) - 1\right)\left(\delta + \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\left(\left\|\Delta\tilde{S}(U)\right\|\right)\right)\right) V(\theta)
\end{aligned}$$

alors on a :

$$\begin{aligned}
\tilde{K}_n &\leq \mathbb{E}\left(\exp 2\gamma \left\|\Delta\tilde{S}(U)\right\| + 1\right) b\mathbb{1}_S(\theta) \\
&+ \left(\lambda\mathbb{E}\left(\exp 2\gamma \left\|\Delta\tilde{S}(U)\right\|\right) + \delta + \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\left(\left\|\Delta\tilde{S}(U)\right\|\right)\right) V(\theta)
\end{aligned}$$

On fixe un $\epsilon > 0$. par le lemme 5, $\exists (n_1, n_2) \in \mathbb{N}^2$ tel que :

$$\begin{aligned}
n \geqslant n_1 &\Rightarrow \mathbb{E}\left(\exp 2\gamma \left\|\Delta\tilde{S}(U)\right\|\right) - 1 \geqslant \epsilon \\
n \geqslant n_2 &\Rightarrow \mathbb{E}\left(\left\|\Delta\tilde{S}(U)\right\|\right) \geqslant \epsilon\frac{\log(1+\epsilon/4)}{4\gamma}
\end{aligned}$$

pour $n \geq n_0 := \max(n_1, n_2)$ on a :

$$\tilde{K}_n \leq (\epsilon+1))b\mathbb{1}_S(\theta) + V(\theta)\left(\lambda(\epsilon+1) + \delta + \frac{\epsilon}{2}\frac{\log(1+\epsilon/4)}{\log(1+\delta/2}\right)$$

on prend $\delta = \epsilon/2$ alors :

$$\tilde{K}_n \leq (\epsilon + 1))b\mathbb{1}_S(\theta) + V(\theta)(\epsilon(\lambda + 1) + \lambda)$$

Pour conclure il suffit de montrer que $\tilde{K}_n$ satisfait une condition géométrique de Drifte (pour $n > n_0$), et pour cela il suffit de prendre $\epsilon < \frac{1-\lambda}{1+\lambda}$ et de vérifier que S est aussi un small set pour $\tilde{K}_n$. □

Chapitre 10

Conclusion

Ce travail que j'ai effectué dans un délai très court (3 mois) a nécessité beaucoup de travail et d'engagement de ma part ainsi que de mon professeur qui m'a beaucoup aidé à améliorer le rendu final de mon projet en soignant jusqu'au plus infime des détails. Nous je suis tellement immergé dans ce sujet que j'envisage de baser ma thèse sur ce domaine. j'aimerai approfondir plus ce domaine et découvrir d'autres façons d'exploiter les mathématiques dans le domaine de data science.

Ce rapport introduit un cadre pour accélérer l'inférence bayésienne menée dans la présence de grands ensembles de données. L'idée est de concevoir une chaîne de Markov dont le noyau de transition utilise une fraction inconnue de taille fixe des données disponibles qui est rafraîchie aléatoirement tout au long de l'algorithme. Le processus de sous-échantillonnage informé est guidé par la fidélité aux données observées, mesurée par des *statistiques sommaires.* L'algorithme résultant, sous-échantillonnage-informé-MCMC (ISS-MCMC : *Informed Sub-Sampling MCMC*), est une approche générique et flexible contrairement aux méthodologies évolutives existantes, il préserve la simplicité de l'algorithme de Metropolis-Hastings.

Nous avons commencé par une présentation rapide sur l'inférence bayésienne et une revue général sur les chaînes de Markov à espace d'état continue. puis nous sommes passés à l'étude profonde de l'échantillonneur Metropolis-Hastings. Ensuite nous avons fournie des résultats théoriques concernant les modèles de familles exponentielles, que nous illustrons à travers un exemple du modèle probit. Puis nous nous sommes penchés sur la méthodologie générale du sous-échantillonnage informé... Finalement, nous avons terminé notre travail par une illustration sur l'estimation de modèle déformable dense, que malheureusement nous n'avons pas suffisamment de temps pour la terminer, nous avons pas attaquer la partie de classification et on s'est contentés d'inférer les paramètres d'un seul digit (les "1").

Ce travail ne représente qu'un début dans ma carrière de recherche et m'ouvre la porte à la réflexion dans d'autres problèmes touchant les méthodes MCMC qui pourra constituer de bons sujets pour mon sujet de doctorat, je pourrais par exemple essayer de répondre aux questions suivantes :

- Peut-on combiner les techniques de l'inférence bayésienne avec d'autre technique de d'apprentissage comme les "réseaux de neurones artificiels" ?

- Est ce qu'on peux trouvé d'autre échantillonneur qui s'adapte avec avec cette méthode de sous-échantillonnage-informé, "voir par example l'échantillonnage de Gibbs" ?

Bibliographie

[1] Florian Maire, Nial Friel, Pierre Alquier, 2017. *Informed Sub-Sampling MCMC : Approximate Bayesian Inference for Large Datasets*. School of Mathematics and Statistics, University College Dublin Insight Centre for Data Analytics, University College Dublin CREST, ENSAE, Université Paris Saclay.

[2] S.Allassonnière, Y.Amit, A.Trouvé . 2017 *Towards a coherent statistical framework for dense deformable template estimation*, Journal of the Royal Statistical Society : Series B (Statistical Method- ology) 69 (1), 3–29.

[3] Christian P. Robert, George Casella. *Monte Carlo Statistical Methods*. Second édition. University o f Florida, Universite Paris Dauphine.

[4] Christian P.Robert. *Le choix bayésien Principes et pratique*. la seconde édition. `Springer`.

[5] Gareth O. Roberts. *General state space Markov chains and MCMC algorithms** Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster.

[6] C.P. Robert. *The Metropolis–Hastings algorithm*, 2016. Université Paris-Dauphine, University of Warwick, and CREST.

[7] S.P. Meyn and R.L Tweedie, 2005 *Markov chains and stochastic stability* .