

Approche bayésienne et méthodes numériques pour la statistique

STA305

Boris Hejblum

Master 2 biostatistique, ISPED, Université de Bordeaux – 2018/2019

Ce cours s'inspire des trois très bon ouvrages que sont *Le choix bayésien* de C. Robert, *Le raisonnement bayésien* de E. Parent & J. Bernier, et *Advanced Statistical Computing* de R. Peng, ainsi que des notes de cours rédigées par D. Commenges.

Table des matières

Objectifs du cours	4
1 Introduction à la statistique bayésienne	5
Vocabulaire bayésien	5
1.1 Rappels sur la statistique fréquentiste	6
1.2 Le paradigme bayésien	6
1.2.1 Le théorème de Bayes	6
1.2.2 Bayésiens vs. Fréquentistes : un débat dépassé	7
2 Modélisation bayésienne	8
2.1 Rappel sur la modélisation fréquentiste	8
2.2 Présentation de l'application historique	8
2.3 Construction d'un modèle bayésien	8
2.3.1 Modèle d'échantillonnage	8
2.3.2 Distribution <i>a priori</i>	9
2.3.3 Distribution <i>a posteriori</i>	9
2.3.4 La question épineuse du choix de la distribution <i>a priori</i>	11
2.4 Extensions	15
2.4.1 Hyper-priors & modèles hiérarchiques	15
2.4.2 Approche bayésienne empirique	15
2.4.3 Bayes séquentiel	16
2.5 Inférence bayésienne	16
2.5.1 Théorie de la décision	16
2.5.2 L'espérance <i>a posteriori</i>	16
2.5.3 Le maximum <i>a posteriori</i>	17
2.5.4 La médiane <i>a posteriori</i>	17
2.5.5 L'intervalle de crédibilité	17
2.5.6 Distribution prédictive	17
2.5.7 Propriétés asymptotiques – et fréquentistes – de la distribution <i>a posteriori</i>	18
2.6 Conclusion et mise en perspective de la modélisation bayésienne	19
2.6.1 Les points essentiels	19
2.6.2 Intérêt de l'approche bayésienne	20
3 Calcul numérique pour l'analyse bayésienne	21
3.1 Une difficile estimation de la distribution <i>a posteriori</i>	21
3.1.1 Paramètres multidimensionnels	21
3.1.2 Statistique bayésienne computationnelle	21

3.1.3	Méthode de Monte-Carlo	22
3.2	Méthodes d'échantillonnage directes	23
3.2.1	Génération de nombres aléatoires selon des lois de probabilité usuelles	23
3.2.2	Échantillonner selon une loi définie analytiquement	24
3.3	Algorithmes MCMC	26
3.3.1	Chaînes de Markov	26
3.3.2	Échantillonnage MCMC	27
3.4	Les algorithmes MCMC pour l'inférence bayésienne dans la pratique	30
3.4.1	Convergence des algorithmes MCMC	31
3.4.2	Inférence à partir d'échantillonnage MCMC	33
3.5	Autres méthodes	35
3.5.1	Bayésien variationnel	35
3.5.2	Calcul Bayésien Approché (<i>ABC</i>)	35
4	Méthodes numériques pour la statistique	36
4.1	Ré-échantillonnage et Monte-Carlo : la méthode du <i>Bootstrap</i>	36
4.2	Algorithmes d'optimisation	36

Objectifs du cours

Se familiariser avec l'approche bayésienne :

1. être capable de proposer une modélisation bayésienne adéquate face à un problème concret
2. savoir calculer la distribution *a posteriori* dans le cas de relations de conjugaison
3. comprendre l'impact de la loi *a priori* et la notion de loi *a priori* faiblement-informative
4. comprendre la notion de MAP et de moyenne *a posteriori*, d'intervalle de crédibilité ainsi que la différence avec un intervalle de confiance
5. comprendre les algorithmes d'échantillonnage et leur utilité
6. comprendre le fonctionnement des algorithmes MCMC
7. savoir utiliser le logiciel **JAGS** et en interpréter les sorties
8. comprendre les notions de risques et de coûts, et leurs implications dans la théorie de la décision

Chapitre 1

Introduction à la statistique bayésienne

NB : ces notes ne se veulent en aucun cas exhaustives, et l'on renverra le lecteur curieux aux ouvrages bien plus complets que sont *Le choix bayésien* de C. Robert et *Le raisonnement bayésien* de E. Parent & J. Bernier.

Vocabulaire bayésien

Quelques mots très utilisés dans le paradigme bayésien :

paradigme

Désigne un système cohérent de représentation du monde, une manière de voir les choses.

a priori

Dans le cadre de la statistique bayésienne l'expression latine *a priori* est beaucoup utilisée. Elle signifie *au préalable* en français, ou plus précisément *en se fondant sur des données antérieures à l'expérience*. Étymologiquement cette expression vient de « *a priori ratione* » qui signifie en latin *par une raison qui précède*, et s'oppose à *a posteriori*.

a posteriori

L'expression latine *a posteriori* est également très utilisée dans le cadre bayésien. Elle signifie *après coup* en français, ou plus précisément *en s'appuyant sur l'expérience, sur les faits constatés*. Étymologiquement cette expression vient de « *a posteriori ratione* » qui signifie en latin *par une raison qui vient après*, et s'oppose à *a priori*.

élicitation

Action formalisant les connaissances d'un expert pour permettre de les partager, e.g. de les incorporer à un modèle.

La statistique est une science mathématique, dont l’objectif est de décrire ce qui s’est produit et de faire des projections quant à ce qu’il peut advenir dans le futur. Elle s’appuie sur l’observation de phénomènes naturels pour en proposer une interprétation, souvent à travers des modèles probabilistes.

1.1 Rappels sur la statistique fréquentiste

La *statistique fréquentiste* désigne la théorie des statistiques largement enseignée et développée en grande partie par Neyman & Pearson, et reposant sur une vision déterministe des paramètres des modèles probabilistes qui sont les objets que l’inférence statistique cherche à estimer. Les estimations par le maximum de vraisemblance font partie des outils fondamentaux de la statistique fréquentiste, tout comme la théorie des tests statistiques avec le concept d’intervalle de confiance qui lui est associé.

1.2 Le paradigme bayésien

1.2.1 Le théorème de Bayes

Le terme *bayésien* provient du nom du révérend Thomas Bayes (Angleterre, XVIII^{ème}). En 1763, ce dernier publie un article¹ (posthume) dans lequel il expose le théorème suivant :

$$\mathbb{P}(A|E) = \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E|A)\mathbb{P}(A) + \mathbb{P}(E|\bar{A})\mathbb{P}(\bar{A})} = \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E)}$$

La postérité désigne ce théorème sous le nom de *théorème de Bayes*. Ce dernier en donne en réalité une version continue dans son travail :

$$g(x|y) = \frac{f(y|x)g(x)}{\int f(y|x)g(x) \, dx}$$

où X et Y sont deux variables aléatoires connaissant les réalisations x et y , $f(y|x)$ représente la distribution conditionnelle de Y sachant la réalisation de X , et $g(x)$ la distribution marginale de X . Le mathématicien français Laplace a également retrouvé ces résultats, de manière indépendante. Laplace et Bayes ont tous les deux poussé ces travaux en décrivant l’incertitude sur les paramètres θ d’un modèle paramétrique $f(y|\theta)$ par une distribution de probabilité π . Le théorème de Bayes s’écrit alors :

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta) \, d\theta}$$

La différence fondamentale entre l’approche fréquentiste et l’approche bayésienne est donc que cette dernière considère les paramètres non pas comme fixes (i.e. pour lesquels il existe une vraie valeur), mais plutôt comme des variables aléatoires. Il s’agit donc d’une différence philosophique profonde même si les ponts entre les deux approches sont nombreux.

Cette manière de considérer les paramètres comme des variables aléatoires induit une distribution marginale $\pi(\theta)$. Cette distribution est appelée *a priori* (ou parfois *prior*, ce qui est un anglicisme). Sa spécification est à la fois un atout de l’analyse bayésienne, puisqu’elle permet de

1. T. Bayes, 1763. An essay towards solving a problem in the doctrine of chances, *The Philosophical Transactions of the Royal Society*, **53** : 370-418.

formaliser les hypothèses sur l'objet d'étude et d'en tenir compte dans la modélisation, mais aussi une faiblesse puisqu'elle introduit nécessairement une subjectivité dans l'analyse. Ces deux facettes d'une même pièce seront d'ailleurs tour à tour mises en avant par les bayésiens tout comme par leur détracteurs.

1.2.2 Bayésiens vs. Fréquentistes : un débat dépassé

Les idées du révérend Bayes, retrouvées indépendamment puis approfondies par Laplace, ont eu une influence profonde sur le développement de la statistique au cours de la deuxième moitié du XVIII^{ème} siècle et du XIX^{ème}. Mais avec l'avènement de la statistique moderne au tournant du XX^{ème} avec Galton et Pearson, puis ensuite avec Fisher et Neymann en particulier, théorie fréquentiste est devenue dominante. Ce n'est qu'à la fin du XX^{ème} siècle que la statistique bayésienne est revenue sur le devant de la scène, notamment grâce à l'avènement de l'ordinateur et au développement de méthodes numériques efficace qui ont permis de dépasser certaines limitations auparavant présentes dans l'analyse bayésienne.

Sous l'influence de Fisher notamment, qui a fermement marqué son rejet du raisonnement bayésien, le XX^{ème} siècle a vu la communauté statistique se scinder en deux entre les partisans de l'approche bayésienne et les tenants de l'approche fréquentiste (considérant les paramètres comme fixes), avec parfois des débats virulents opposant les deux communautés.

Aujourd'hui, ces querelles de chapelles sont dépassées, en partie grâce aux succès pratiques qu'ont rencontrés chacune des deux approches sur des problèmes modernes et complexes, notamment dans le domaine de la santé. De plus, un certain nombre de méthodes, telles que par exemple les méthodes bayésiennes empiriques, se situent à la frontière entre les deux approches et permettent de faire le pont entre elles. Le (bio)-statisticien d'aujourd'hui se doit donc d'être pragmatique et versatile, intégrant l'analyse bayésienne dans sa boîte à outils pour résoudre les problèmes auxquels il est confronté.

« Être ou ne pas être bayésien, là n'est plus la question : il s'agit d'utiliser à bon escient les outils adaptés quand cela est nécessaire » Gilbert Saporta

Chapitre 2

Modélisation bayésienne

2.1 Rappel sur la modélisation fréquentiste

Considérons une suite de variables aléatoires *iid* $\mathbf{Y} = (Y_1, \dots, Y_n)$, dont on observe un échantillon $\mathbf{y} = (y_1, \dots, y_n)$. Un modèle fréquentiste pour leur loi de probabilité est la famille de densité de probabilité suivante : $f(y|\theta)$, $\theta \in \Theta$. Avec ce modèle, on suppose qu'il existe une « vraie » distribution de Y caractérisée par la « vraie » valeur du paramètre θ^* qui s'écrit $f(y|\theta^*)$. On cherche alors un estimateur $\hat{\theta}$, ayant de bonnes propriétés asymptotiques le plus souvent (généralement sans biais pour θ^* et avec une variance la plus réduite possible).

2.2 Présentation de l'application historique

Laplace s'est intéressé à la probabilité de naissance de filles (plutôt que de garçons). Pour cela, il a utilisé les naissances observées à Paris entre 1745 et 1770, période durant laquelle sont nés 241 945 filles et 251 527 garçons. La question que l'on se pose est la suivante : « Lorsqu'un enfant naît, y a-t-il autant de chance que ce soit une fille ou un garçon ? »

2.3 Construction d'un modèle bayésien

La première étape dans la construction d'un modèle est toujours d'identifier la question à laquelle on souhaite répondre. Une fois cette étape accomplie, il s'agit de déterminer quelles observations sont disponibles, et vont pouvoir nous informer dans notre réponse à la question d'intérêt.

2.3.1 Modèle d'échantillonnage

Notons \mathbf{y} les observations dont nous disposons. De la même manière que dans un modèle fréquentiste, une modélisation bayésienne paramétrique consiste à d'abord proposer un modèle probabiliste pour ces observations : $Y_i \stackrel{iid}{\sim} f(y|\theta)$. On appelle ce dernier « modèle d'échantillonnage ».

Dans l'application historique, Laplace a proposé un modèle d'échantillonnage basé sur la loi de

2.3.2 Distribution *a priori*

Dans la modélisation bayésienne, par rapport à la modélisation fréquentiste, on ajoute une loi de probabilité (définie sur l'espace Θ des paramètres), appelée distribution *a priori* :

$$\begin{aligned}\theta &\sim \pi(\theta) \\ Y_i|\theta &\stackrel{iid}{\sim} f(y|\theta)\end{aligned}$$

On va donc traiter θ comme une variable aléatoire, mais qui n'est jamais observée !

Dans l'application historique, Laplace a d'abord envisagé un *a priori*

2.3.3 Distribution *a posteriori*

L'objet d'une telle modélisation bayésienne est la distribution des paramètres *a posteriori*, c'est-à-dire la loi de θ conditionnellement aux observations : $p(\theta|\mathbf{Y})$, appelée distribution *a posteriori*. Elle se calcule à partir du modèle d'échantillonnage $f(y|\theta)$ – à partir duquel on obtient la vraisemblance $f(\mathbf{y}|\theta)$ pour toutes les observations – et de la loi *a priori* $\pi(\theta)$ par le théorème de Bayes :

$$p(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})}$$

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

Exemple avec un *a priori* uniforme

Dans l'application historique, la vraisemblance est donc :

On obtient alors le *posterior* suivant :

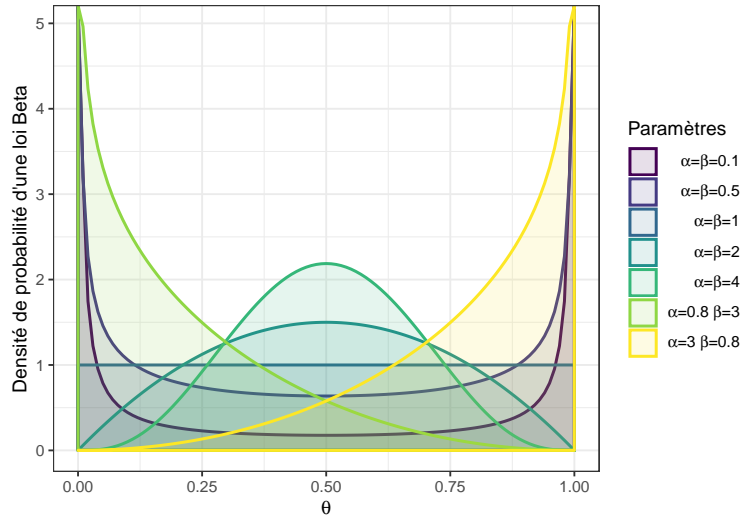
Pour répondre à la question d'intérêt, on peut alors calculer :

Une approximation par la loi normale a cependant permis à Laplace de conclure que la probabilité de naissance d'un fille est inférieure à celle d'un garçon¹, puisqu'il obtint : $P(\theta \geq 0.5|\mathbf{y}) \approx 1.15 \cdot 10^{-42}$

1. Cette conclusion a été confirmée depuis et semble être valable pour l'espèce humaine en général.

Exemple de la conjugaison de la loi Beta

Imaginons maintenant que l'on utilise une autre loi *a priori*, par exemple la loi $\text{Beta}(\alpha, \beta)$ dont la densité s'écrit : $f(\theta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} \theta^{\alpha-1} (1-\theta)^{\beta-1}$ (pour $\alpha > 0$ et $\beta > 0$).



Exemples de paramétrisations pour la distribution Beta

On remarque que la loi uniforme est un cas particulier de la loi Beta lorsque α et β valent tous les deux 1. Si on re-calcule le *posterior* avec un *a priori* $\pi = \text{Beta}(\alpha, \beta)$, on obtient facilement :

On reconnaît, à une constante de normalisation près,
On en déduit donc que $\theta | \mathbf{y} \sim$

On dit que l'on est dans une situation de **distributions conjuguées** car les distributions *a posteriori* et *a priori* appartiennent à la même famille paramétrique.

On peut maintenant évaluer l'impact de cet *a priori* Beta sur notre résultat en fonction du choix des hyperparamètres α et β .

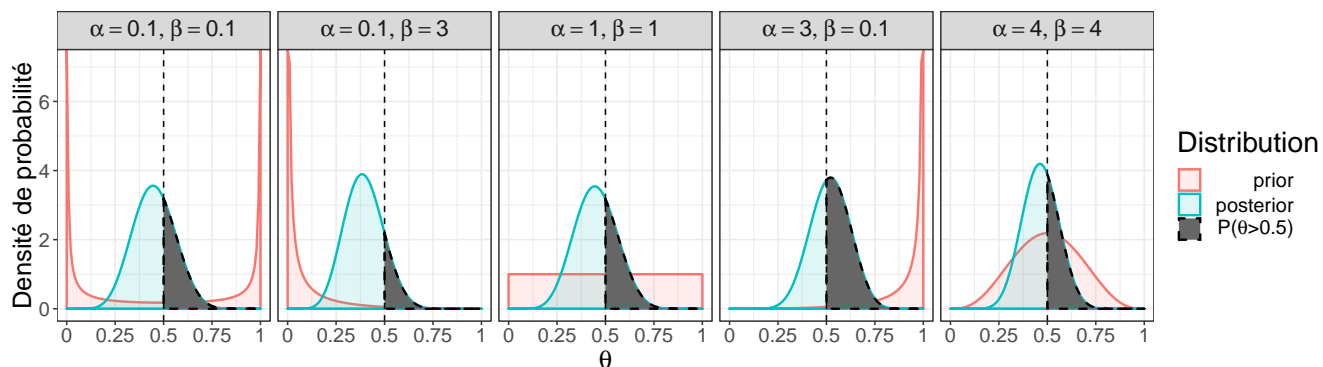
Interprétation de l' <i>a priori</i>	Paramètres de la Beta	$P(\theta \geq 0.5 \mathbf{y})$
#garçons > #filles	$\alpha = 0.1, \beta = 3$	$1.08 \cdot 10^{-42}$
#garçons < #filles	$\alpha = 3, \beta = 0.1$	$1.19 \cdot 10^{-42}$
#garçons = #filles	$\alpha = 4, \beta = 4$	$1.15 \cdot 10^{-42}$
#garçons \neq #filles	$\alpha = 0.1, \beta = 0.1$	$1.15 \cdot 10^{-42}$
non informatif	$\alpha = 1, \beta = 1$	$1.15 \cdot 10^{-42}$

TABLE 2.1 – Pour 493 472 naissances dont 241 945 filles

On remarque que l'*a priori* ne semble pas influencer sur notre résultat ici. C'est parce que l'on dispose de beaucoup d'observations. Le poids de la distribution *a priori* dans la distribution *a posteriori* devient alors très faible en regard de l'information apportée par les observations. Si l'on imagine que l'on avait observé seulement 20 naissances, dont 9 filles, on note alors une influence de l'*a priori* bien plus grande.

Interprétation d l' <i>a priori</i>	Paramètres de la Beta	$P(\theta \geq 0.5 \mathbf{y})$
#garçons > #filles	$\alpha = 0.1, \beta = 3$	0.39
#garçons < #filles	$\alpha = 3, \beta = 0.1$	0.52
#garçons = #filles	$\alpha = 4, \beta = 4$	0.46
#garçons \neq #filles	$\alpha = 0.1, \beta = 0.1$	0.45
non informatif	$\alpha = 1, \beta = 1$	0.45

TABLE 2.2 – Pour 20 naissances dont 9 filles



Impact de différent priors Beta pour 20 naissances observées

2.3.4 La question épineuse du choix de la distribution *a priori*

Un point essentiel de l'approche bayésienne est donc de donner une distribution aux paramètres. Dans l'inférence bayésienne, on part d'une distribution *a priori*, et l'information contenue dans les observations est utilisée pour obtenir la distribution *a posteriori*. La distribution *a priori* apporte de la flexibilité par rapport à un modèle fréquentiste, en permettant d'incorporer dans le modèle de la connaissance externe. Cela peut par exemple permettre de résoudre des problèmes d'identifiabilité parfois rencontrés par une approche purement fréquentiste lorsque l'information apportée par les observations ne suffit pas pour estimer tous les paramètres d'intérêt.

C'est donc un grand avantage de l'approche bayésienne. Mais d'un autre côté, le choix de cette distribution *a priori* des paramètres introduit une subjectivité intrinsèque dans l'analyse, qui peut être décriée. Par exemple un statisticien travaillant pour un laboratoire pharmaceutique pourrait choisir une loi *a priori* donnant une forte probabilité qu'un médicament soit efficace, ce qui influencera nécessairement le résultat. Le choix (ou l'élicitation) de la distribution *a priori* est donc délicat.

Notons tout d'abord deux points théoriques :

- 1 le support de la distribution *a posteriori* doit être inclus dans celui de la distribution *a priori*. En d'autres termes, si $\pi(\theta) = 0$, alors $p(\theta|\mathbf{y}) = 0$.
- 2 en général on suppose l'indépendance des différents paramètres sous la loi *a priori* (quand il y a plus d'un paramètre – ce qui est presque toujours le cas dans les applications) ce qui permet d'éliciter les *priors* paramètre par paramètre.

Élicitation de la distribution *a priori*

Il y a des stratégies pour communiquer avec les experts non-statisticiens pour transformer leurs **connaissances** *a priori* en **distribution** *a priori*.

La méthode la plus simple est de demander aux experts de donner des poids ou des probabilités à des intervalles de valeurs : c'est la méthode des histogrammes. Cependant, quand le paramètre peut prendre des valeurs sur un ensemble non-borné cette méthode risque de donner un *a priori* nul pour des valeurs du paramètre qui sont néanmoins possibles...

Une autre approche est de se donner une famille paramétrique de distributions $p(\theta|\eta)$ et de choisir η de telle sorte que la distribution *a priori* soit en accord avec ce que pensent les experts pour certaines caractéristiques. Par exemple on pourra faire en sorte que les deux premiers moments (moyenne et variance), ou bien des quantiles simples (comme les quartiles), coïncident avec leurs vues. Cela résout le problème de support soulevé par la méthode des histogrammes. Cependant le choix de la famille paramétrique peut avoir de l'importance. Par exemple une distribution normale $\mathcal{N}(0; 2, 19)$ a les même quartiles qu'une distribution de Cauchy $\mathcal{C}(0; 1)$, à savoir $-1; 0; 1$. Or ces deux *priors* peuvent donner des distributions *a posteriori* assez différentes. Une stratégie pour déterminer les quartiles est de poser les questions suivantes :

- pour la médiane : *Pouvez-vous déterminer une valeur telle que θ a autant de chances de se trouver au dessus qu'au-dessous ?*
- puis pour le premier quartile : *Supposons que l'on vous dise que θ est en dessous de [telle valeur médiane], pouvez-vous alors déterminer une nouvelle valeur telle que θ ait autant de chances de se trouver au dessus qu'au-dessous ?*
- de façon similaire on détermine le troisième quartile...

Des logiciels existent pour aider à l'élicitation des *priors* par des experts : voir par exemple l'outil académique SHELF².

On peut également éliciter des *priors* d'après les données de la littérature. L'idée est de définir les moments de la distribution *a priori* tels qu'ils donnent une probabilité raisonnable aux valeurs du paramètre qui ont été recensées dans la littérature. Si on propose un *a priori* normal de loi $\mathcal{N}(\mu, \sigma^2)$, on peut par exemple choisir μ et σ de telle sorte que la plus petite valeur donnée dans la littérature soit égale à $\mu - 1.96\sigma$ et la plus grande à $\mu + 1.96\sigma$. Une approche plus élaborée est de maximiser la vraisemblance des valeurs de la littérature...

La quête des *priors* non-informatifs

Pour certains paramètres (ou pour tous les paramètres) il est courant que l'on n'ait pas de connaissance *a priori*. On cherche alors à définir une distribution « non-informative ». Par exemple si le paramètre est la probabilité qu'un pièce de monnaie tombe sur pile ou face, une loi non-informative pourrait être la loi uniforme sur $[0; 1]$ (le choix historique de Bayes en 1763). Cependant deux difficultés majeures apparaissent :

1 Lois impropres

La première difficulté est que l'on peut être amené à considérer des lois impropres. Une loi impropre est définie par une densité dont la somme ne fait pas à 1. Par exemple pour un paramètre de moyenne d'une loi normale, il peut sembler naturel de donner une densité

2. <http://www.tonyohagan.co.uk/shelf/>

constante $\pi(\theta) = c$ (i.e. toutes les valeurs possibles sur $]-\infty, +\infty[$ ont la même probabilité). Bien sûr $\int_{-\infty}^{\infty} c d\theta = \infty$, et un tel choix ne définit donc pas une loi de probabilité ! Il reste cependant **admissible car la loi *a posteriori* est** (la plupart du temps) **propre**. En effet nous avons alors :

$$p(\theta|y) = \frac{f(y|\theta)c}{\int f(y|\theta)c d\theta}$$

Si $\int f(y|\theta)c d\theta = K$ (comme c'est souvent le cas), alors $p(\theta|y) = \frac{f(y|\theta)}{K}$ est une densité propre (i.e. qui somme à 1).

2 Lois non-invariantes

La seconde difficulté vient de la non-invariance de la distribution uniforme pour des transformations non-linéaires des paramètres. En effet si on fait une transformation des paramètres $\gamma = g(\theta)$ la densité de γ s'écrit : $\pi(\gamma) = |J| \pi(\theta)$, où $|J|$ est le Jacobien de la transformation, c'est-à-dire le déterminant de la matrice jacobienne $J = \frac{\partial g^{-1}(\gamma)}{\partial \gamma}$. Par exemple si on prend une densité uniforme égale à 1 pour θ sur $(0, +\infty)$ et que l'on fait la transformation $\gamma = \log(\theta)$, on a $g^{-1}(\gamma) = e^\gamma$ et $|J| = e^\gamma$. Donc on a $\pi(\gamma) = e^\gamma$, ce qui n'est pas la caractérisation d'une loi uniforme. D'où le paradoxe suivant : si la loi uniforme pour θ traduit une absence totale de connaissance *a priori* sur θ , on devrait avoir aussi une absence totale d'information *a priori* sur γ , ce qui devrait se traduire par une loi uniforme sur γ . Or ce ne peut être le cas. Donc la loi uniforme ne peut pas être d'une manière générale la loi représentant une absence totale de connaissance *a priori*. C'est un argument central qui a conduit Fisher, en 1922, à proposer les estimateurs du maximum de vraisemblance, possédant eux une propriété d'invariance pour des transformations non-linéaires des paramètres.

NB : Ceci ne veut pas dire que l'on ne puisse pas prendre une loi uniforme comme *a priori*, mais il faut avoir conscience que la loi uniforme ne vaut que pour une certaine paramétrisation...

Face à ces difficultés, différentes solutions ont été proposées. Elles ont montré qu'il n'y a pas de loi *a priori* complètement non-informative, mais on peut considérer certaines lois comme **faiblement informatives**.

La loi *a priori* de Jeffreys

L'approche la plus aboutie des *a priori* faiblement informatifs est peut-être celle de Jeffreys. Ce dernier a proposé une procédure pour trouver une loi *a priori* avec une propriété d'invariance par rapport à la paramétrisation. Dans le cas univarié, la loi *a priori* de Jeffreys est défini par :

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

où I est la matrice d'information de Fisher (pour rappel, $I(\theta) = -\mathbb{E}_{Y|\theta} \left[\frac{\partial^2 \log(f(y|\theta))}{\partial \theta^2} \right]$). La loi *a priori* de Jeffreys est donc invariante pour les transformations bijectives des paramètres. C'est-à-dire que si nous considérons une autre paramétrisation $\gamma = g(\theta)$ (pour laquelle il existe la bijection réciproque $g^{-1}(\cdot)$), on obtient toujours :

$$\pi(\gamma) \propto \sqrt{I(\gamma)}$$

tandis que $\pi(\gamma)$ correspond bien à la même loi *a priori* sur θ . Prenons ici des notations plus rigoureuses, et notons les densités $\pi_\theta(\cdot)$ et $\pi_\gamma(\cdot)$. $\pi_\gamma(\cdot)$ s'exprime en fonction de $\pi_\theta(\cdot)$ avec $\pi_\gamma(\cdot) = \pi_\theta(\cdot)|J|$. On vérifie donc bien que $\sqrt{I(\gamma)} = \sqrt{I(\theta)}|J|$.

Démonstration :

Dans le cas multidimensionnel (le plus courant) la loi *a priori* de Jeffreys est définie comme :

$$\pi(\theta) \propto \sqrt{|I(\theta)|}$$

où $|I(\theta)|$ est le déterminant de la matrice d'information de Fisher $I(\theta)$. Cependant cette méthode est peu utilisée car d'une part les calculs peuvent être compliqués, et d'autre part elle peut donner des résultats un peu curieux. En effet dans le cas d'une vraisemblance normale par exemple, où l'on a 2 paramètres θ et σ , l'*a priori* de Jeffreys multidimensionnel est $1/\sigma^2$, ce qui est différent de $\pi(\sigma) = 1/\sigma$ obtenu dans le cas unidimensionnel... Dans la pratique la tendance est d'appliquer la loi *a priori* de Jeffreys séparément pour chaque paramètre et de définir la loi *a priori* conjointe par la multiplication des *a priori* pour chaque paramètre (faisant donc une hypothèse d'indépendance). Pour l'exemple normal avec deux paramètres, on obtient donc $\pi(\theta, \sigma) = 1/\sigma$. Mais on note que ce n'est plus vraiment l'*a priori* de Jeffreys, en deux dimensions.

Exercice : retrouver les résultats énoncés ci-dessus (invariance pour la transformation log et résultat pour la loi normale).

Priors pour les familles à paramètres de position et l'échelle : Considérons les familles à paramètre de position, c'est-à-dire dont les modèles d'échantillonnage sont de la forme $f(y|\theta) = f(y - \theta)$. Des arguments d'invariance permettent d'affirmer que la loi non-informative pour θ devrait être uniforme. Par les mêmes arguments, on montre que pour les familles à paramètre d'échelle, c'est-à-dire dont les modèles d'échantillonnage sont de la forme $f(y|\sigma) = f(y/\sigma)$, la loi non-informative devrait être $\pi(\sigma) \propto 1/\sigma$. Plus généralement, pour les familles à paramètres de position et d'échelle, c'est-à-dire dont les modèles d'échantillonnage sont de la forme $f(y|\theta, \sigma) = f((y - \theta)/\sigma)$, l'*a priori* faiblement-informatif devrait être de la forme $\pi(\theta, \sigma) = 1/\sigma$. La loi normale est une famille de ce type, et pour elle cette recommandation d'*a priori* faiblement-informatif rejoint celle obtenue en multipliant les *a priori* de Jeffreys unidimensionnels, ainsi qu'indiqué plus haut.

Exercice : retrouver les résultats énoncés ci-dessus.

Lois *a priori* diffuses

En pratique, une alternative très courante pour donner une loi *a priori* faiblement informative est l'utilisation de lois paramétriques (telles que la loi normale) avec des paramètres de variances très importants (ce qui se rapproche de la loi uniforme mais évite le problème de loi impropre).

2.4 Extensions

2.4.1 Hyper-priors & modèles hiérarchiques

Dans le modèle bayésien classique, on considère deux niveaux hiérarchiques : d'abord $\pi(\theta)$, puis $f(\mathbf{y}|\theta)$. Il est possible de rajouter un niveau en donnant également un *a priori* au paramètre η de $\pi(\theta)$, appelé hyper-paramètre : $\pi(\theta|\eta)$. Appliquant l'approche bayésienne, on peut donner à cet hyper-paramètre une loi *a priori*, appelée alors hyper-prior et que l'on note $h(\eta)$. La distribution *a posteriori* est :

$$p(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})} = \frac{\int f(\mathbf{y}|\theta)\pi(\theta|\eta)h(\eta)d\eta}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\theta) \int \pi(\theta|\eta)h(\eta)d\eta}{f(\mathbf{y})}$$

On remarque donc que cette modélisation hiérarchique à 3 niveaux est équivalente à une modélisation bayésienne à deux niveaux avec une distribution *a priori* qui devient : $\pi(\theta) = \int \pi(\theta|\eta)h(\eta)d\eta$. Néanmoins cette construction hiérarchique peut faciliter l'étape de modélisation ainsi que l'élicitation des lois *a priori*. Il est d'ailleurs même possible de construire des modèles avec plus de trois niveaux, considérant que la distribution de η dépend elle-même d'hyper-hyper-paramètres, et ainsi de suite... Un cas d'utilisation typique de modélisation bayésienne hiérarchique est l'inclusion d'effets aléatoires dans le modèle linéaire. Un autre exemple sont les modèles à classes latentes. On note ici que la frontière entre modélisation fréquentiste et bayésienne s'amincit, et qu'elle se joue principalement sur l'interprétation des paramètres (et donc des résultats).

Si l'on reprend l'exemple historique du sexe à la naissance avec un *a priori* Beta, on peut proposer deux hyper-priors Gamma pour α et β :

$$\alpha \sim \text{Gamma}(4, 0.5)$$

$$\beta \sim \text{Gamma}(4, 0.5)$$

$$\theta|\alpha, \beta \sim \text{Beta}(\alpha, \beta)$$

$$Y_i|\theta \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$$

2.4.2 Approche bayésienne empirique

Cette approche consiste à éliciter la loi *a priori* d'après sa loi marginale empirique, et donc à estimer la distribution *a priori* à partir des données. Cela revient donc à se donner des hyper-paramètres et à chercher à les estimer de manière fréquentiste (par exemple par maximum de vraisemblance) par $\hat{\eta}$, avant d'injecter cet estimateur dans la distribution *a priori* et donc d'obtenir la distribution *a posteriori* $p(\theta|\mathbf{y}, \hat{\eta})$. Cette approche **bayésienne empirique** qui combine bayésien et fréquentiste peut sembler aller à l'encontre de la notion d'*a priori*, puisque l'on utilise alors déjà les données pour choisir l'*a priori*. Néanmoins, on peut voir l'approche bayésienne empirique comme une approximation de l'approche bayésienne complète. Son utilisation résulte en une distribution *a posteriori* plus resserrée qu'avec un *a priori* faiblement informatif (diminution de la variance), au prix de l'introduction d'un biais dans l'estimation (on « utilise les données 2 fois » !). Cette approche illustre une fois de plus la balance existant entre biais et variance, classique dans toute procédure d'estimation.

2.4.3 Bayes séquentiel

À noter que le théorème de Bayes peut être utilisé de manière séquentielle. Omettant le dénominateur (qui ne dépend pas de θ) on peut écrire : $p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)$. Si $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, on a : $p(\theta|\mathbf{y}) \propto f(\mathbf{y}_2|\theta)f(\mathbf{y}_1|\theta)\pi(\theta) \propto f(\mathbf{y}_2|\theta)p(\theta|\mathbf{y}_1)$. La distribution *a posteriori* sachant \mathbf{y}_1 devient ainsi la distribution *a priori* pour la nouvelle observation \mathbf{y}_2 . On peut donc mettre à jour l'information sur θ au fur et à mesure qu'arrivent les observations (approche *online*).

2.5 Inférence bayésienne

Une fois la modélisation bayésienne terminée, on dispose de la distribution *a posteriori* (obtenue grâce au choix de la distribution *a priori*, du modèle d'échantillonnage et des données observées). Cette distribution contient l'ensemble de l'information sur θ conditionnellement au modèle et aux données. On peut néanmoins s'intéresser à des résumés de cette distribution, par exemple à un paramètre central de cette distribution tel que l'espérance, le mode ou encore la médiane (ces derniers sont analogues aux estimateurs ponctuels obtenus par l'analyse fréquentiste), ou à des intervalles de valeurs dont la probabilité *a posteriori* est forte.

2.5.1 Théorie de la décision

La théorie de la décision statistique est généralement utilisée dans un contexte d'estimation d'un paramètre inconnu θ . La décision concerne alors le choix d'un estimateur ponctuel $\hat{\theta}$. Afin de déterminer le $\hat{\theta}$ optimal, on définit une **fonction de coût** (à valeur dans $[0, +\infty[$) représentant la pénalité associée au choix d'un $\hat{\theta}$ particulier (c'est-à-dire à la décision associée). Afin de déterminer le $\hat{\theta}$ optimal (c'est-à-dire la décision optimale) on va vouloir minimiser la fonction de coût choisie. À noter qu'un grand nombre de fonctions de coût différentes sont possibles, et que chacune d'entre elle résulte en un estimateur ponctuel optimal différent et donc une décision optimale spécifique.

2.5.2 L'espérance *a posteriori*

L'espérance *a posteriori* est définie par :

$$\mu_P = \mathbb{E}(\theta|\mathbf{y}) = \mathbb{E}_{\theta|\mathbf{y}}(\theta)$$

C'est l'estimateur qui a la plus petite variance *a posteriori* (au sens bayésien : $\mathbb{E}_{\theta|\mathbf{y}}(\theta - \hat{\theta})^2$).

Démonstration : pour un estimateur $\hat{\theta}$,

2.5.3 Le maximum *a posteriori*

Le maximum a été beaucoup utilisé, surtout car il est plus facile (ou en tout cas moins difficile) à calculer. En effet, il ne requiert aucun calcul d'intégrale, mais une simple maximisation de $f(\mathbf{y}|\theta)\pi(\theta)$ (car le dénominateur $f(\mathbf{y})$ ne dépend pas de θ). L'estimateur du mode s'appelle le **maximum *a posteriori*** (souvent noté **MAP**).

Le MAP peut être vu comme une régularisation de l'estimateur du maximum de vraisemblance, dont il est proche.

2.5.4 La médiane *a posteriori*

La médiane est également un résumé possible de la distribution *a posteriori*. Comme son nom l'indique, il s'agit de la médiane de $p(\theta|(\mathbf{y}))$. Il s'agit de l'estimateur ponctuel optimal au sens de l'erreur absolue (fonction de coût linéaire).

2.5.5 L'intervalle de crédibilité

Finalement on peut définir un ensemble de valeurs ayant une forte probabilité *a posteriori*. Un tel ensemble est appelé **ensemble de crédibilité**. Si le *posterior* est unimodal, un tel ensemble est un intervalle. Par exemple, un **intervalle de crédibilité à 95%** est un intervalle $[t_{inf}; t_{sup}]$ tel que $\int_{t_{inf}}^{t_{sup}} p(\theta|\mathbf{y}) d\theta = 0.95$. En général on est intéressé par l'intervalle de crédibilité à 95% le plus étroit possible (*Highest Density Interval*).

On rappelle ici l'interprétation d'un intervalle de confiance fréquentiste au niveau 95%, qui s'interprète comme suit, par rapport à l'ensemble des intervalles de ce niveau qu'on aurait pu observer :



Attention : on ne peut pas interpréter une réalisation d'un intervalle de confiance en terme probabiliste ! C'est une erreur qui est souvent commise. . . L'intervalle de crédibilité s'interprète lui bien plus naturellement, comme un intervalle qui a 95% de chance de contenir θ (pour un niveau de 95%, évidemment).

2.5.6 Distribution prédictive

La **distribution prédictive** (appelée parfois *posterior prédictive*) est définie comme la distribution d'une nouvelle observation Y_{n+1} sachant les observations de l'échantillon. Elle se calcule comme la distribution de Y_{n+1} sachant \mathbf{y} , marginalement par rapport à θ . $f_{Y_{n+1}}(y|\mathbf{y}) = \int f_{Y_{n+1}}(y|\theta)p(\theta|\mathbf{y}) d\theta$. Le calcul se fait ainsi :

$$\begin{aligned} f_{Y_{n+1}}(y|\mathbf{y}) &= \int f_{Y_{n+1}}(y, \theta|\mathbf{y}) d\theta \\ &= \int f_{Y_{n+1}}(y|\theta, \mathbf{y})p(\theta|\mathbf{y}) d\theta \end{aligned}$$

$$= \int f_{Y_{n+1}}(y|\theta)p(\theta|\mathbf{y}) d\theta$$

On remarque le lien entre cette formule et celle de la distribution marginale : $f_Y(y) = \int f_Y(y|\theta)\pi(\theta) d\theta$, qui peut être vue comme un cas particulier de la distribution prédictive quand il n'y a pas d'information apportée par l'échantillon observé. On note également la différence avec l'approche fréquentiste où l'on estime d'abord θ par $\hat{\theta}$, et l'on remplace θ par $\hat{\theta}$ pour obtenir la distribution prédictive : $f_{Y_{n+1}}(y|\hat{\theta})$.

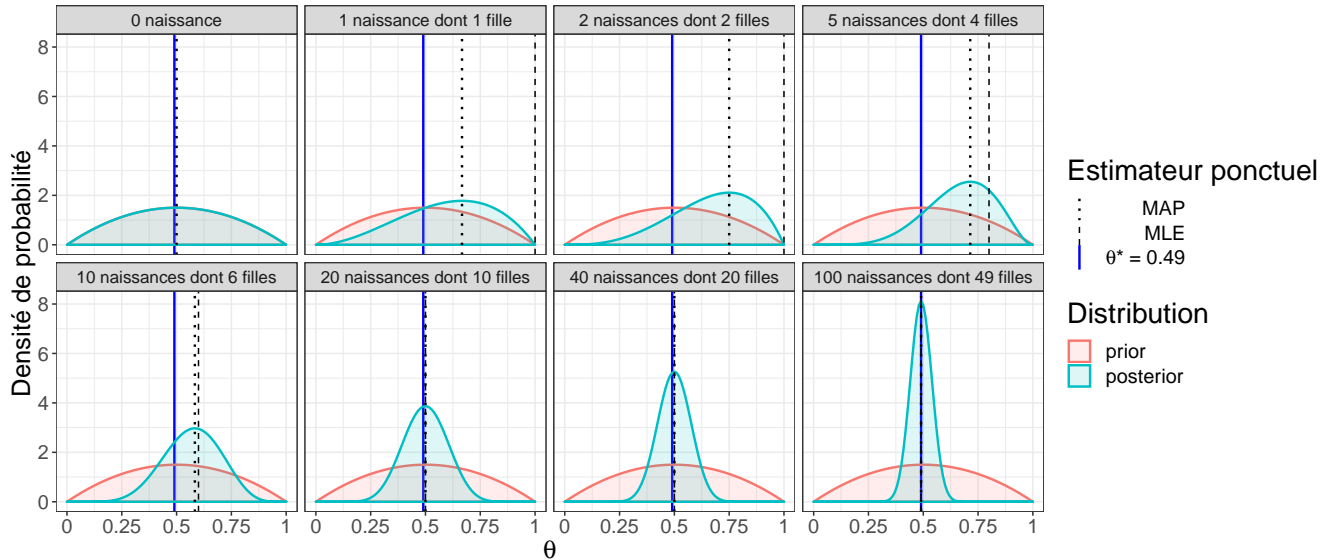
Exercice : calculer la distribution prédictive sur l'exemple historique du sexe à la naissance pour un *a priori* uniforme.

2.5.7 Propriétés asymptotiques – et fréquentistes– de la distribution *a posteriori*

Théorème de convergence de Doob

Un résultat très intéressant est le comportement asymptotique de la distribution *a posteriori* sous certaines hypothèses (cas *iid*, densités dérivables trois fois, existence de moments d'ordre 2). Il y a un premier résultat, le **théorème de convergence de Doob**, qui assure que la distribution *a posteriori* se concentre vers la vraie valeur du paramètre quand $n \rightarrow \infty$. On peut le noter (convergence en Loi) :

$$p(\theta|\mathbf{y}_n) \xrightarrow{\mathcal{L}} \delta_{\theta^*}$$



Exemple historique : concentration du posterior autour de θ^* avec n

Théorème de Bernstein-von Mises

Un résultat plus riche caractérise la distribution asymptotique de θ : le **Théorème de Bernstein-von Mises** (auss appelé **Théorème limite central bayésien**). Pour n grand la distribution *a posteriori* $p(\theta|\mathbf{y})$ peut être approximée par une loi normale ayant pour espérance le mode $\hat{\theta}$ et pour variance l'inverse de la Hessienne (dérivée seconde) de $p(\theta|\mathbf{y})$ par rapport à θ , pris au mode θ .

Ci-dessous une démonstration heuristique, grâce à un développement limité de $\log(p(\theta|\mathbf{y}))$ autour du mode $\hat{\theta}$ donne :

$$\log(p(\theta|\mathbf{y})) = \log(p(\hat{\theta}|\mathbf{y})) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{\partial^2 \log(p(\theta|\mathbf{y}))}{\partial \theta^2} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \dots$$

On note que le terme linéaire (omis ci-dessus) est nul, puisque la dérivée de $p(\theta|\mathbf{y})$ est nulle en son mode ($\hat{\theta}$). Le premier terme est lui constant en θ . Donc, en négligeant les termes suivants du développement, le logarithme de $p(\theta|\mathbf{y})$ est égal au logarithme d'une densité gaussienne d'espérance $\hat{\theta}$ et de variance $I(\hat{\theta})^{-1}$ (où $I(\theta) = \frac{\partial^2 \log(p(\theta|\mathbf{y}))}{\partial \theta^2} \Big|_{\theta=\hat{\theta}}$), et l'on donc peut écrire l'approximation :

$$p(\theta|\mathbf{y}) \approx \mathcal{N}(\hat{\theta}, I(\hat{\theta})^{-1})$$

Ce résultat a une double importance :

- il peut être utilisé pour expliquer pourquoi les **procédures bayésienne et fréquentiste basées sur le maximum de vraisemblance** donnent, pour n grand, des résultats très voisins. Ainsi, en dimension 1, l'intervalle de crédibilité asymptotique est : $[\hat{\theta} \pm 1.96\sqrt{I(\hat{\theta})^{-1}}]$, et si on le compare à l'intervalle de confiance fréquentiste construit à partir de la loi asymptotique de l'estimateur : $[\hat{\theta}_{MLE} \pm 1.96\sqrt{I(\hat{\theta}_{MLE})^{-1}}]$ (où $I(\hat{\theta}_{MLE})$ est ici la matrice d'information de Fisher observée, et correspond à la définition précédente pour des *priors* uniformes), on note qu'ils sont tous les deux identiques (pour un *a priori* uniforme). Pour ces *priors*, on note que l'on a aussi $\hat{\theta} = \hat{\theta}_{MLE}$ (et même si on ne prend pas des *a priori* uniformes, les estimateurs et intervalles sont très proches, puisque le poids de la loi *a priori* devient négligeable quand $n \rightarrow \infty$). **L'interprétation théorique de ces intervalles reste évidemment différente.**
- il signifie que l'on peut **approximer la distribution *a posteriori* par une loi normale**, dont on peut calculer l'espérance et la variance simplement à l'aide du MAP, et permet donc de faciliter les calculs numériques de l'inférence bayésienne.

2.6 Conclusion et mise en perspective de la modélisation bayésienne

2.6.1 Les points essentiels

1 La formulation d'un modèle bayésien :

$$\begin{aligned} \theta &\sim \pi(\theta) \quad \text{la loi } a \text{ priori} \\ Y_i|\theta &\stackrel{iid}{\sim} f(y|\theta) \quad \text{le modèle d'échantillonnage} \end{aligned}$$

2 La formule de Bayes :

$$p(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})}$$

où $p(\theta|\mathbf{y})$ est la distribution *a posteriori*, $f(\mathbf{y}|\theta)$ est la vraisemblance (héritée du modèle d'échantillonnage), $\pi(\theta)$ est la distribution *a priori* des paramètres θ et $f(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta)$

est la distribution marginale des données, i.e. la constante (par rapport à θ) de normalisation.

3 L'obtention de la loi *a posteriori* :

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)$$

4 La loi *a priori* faiblement informative de Jeffreys :

$$\pi(\theta) \propto \sqrt{I(\theta)} \quad \text{en unidimensionnel}$$

possédant la propriété d'invariance.

5 La moyenne *a posteriori*, le MAP et les intervalles de crédibilité

6 La distribution prédictive :

$$f_{Y_{n+1}}(y|\mathbf{y}) = \int f_{Y_{n+1}}(y|\theta)p(\theta|\mathbf{y}) d\theta$$

2.6.2 Intérêt de l'approche bayésienne

L'analyse bayésienne est un outil statistique d'analyse de données, au même titre que d'autres méthodologies comme les forêts aléatoires, les méthodes de réduction de dimension, les modèles à classes latentes, etc. Il est particulièrement utile lorsque peu de données sont disponibles et que les méthodes fréquentistes ne permettent pas d'obtenir de résultats (par exemple une régression logistique avec très peu voire pas d'événement, i.e. beaucoup voire que des 0 dans le cas d'événements extrêmement rares), et/ou lorsqu'il existe de fortes connaissances *a priori* qu'il est utile d'intégrer dans un modèle avec peu de d'observations (par exemple le modèle utilisé par *FiveThirtyEight* pour prédire les résultats des élections américaines de 2008 dans chaque état américain, dans certains desquels peu de sondages étaient effectués, ou encore dans les études de génomique où le nombre d'observations disponible pour chaque gène est généralement relativement faible mais que beaucoup de gènes sont observés). Comme toute méthode statistique, l'analyse bayésienne présente des avantages et des inconvénients qui vont avoir plus ou moins d'importance selon le problème à résoudre.

Chapitre 3

Calcul numérique pour l'analyse bayésienne

3.1 Une difficile estimation de la distribution *a posteriori*

3.1.1 Paramètres multidimensionnels

Dans les applications réelles il y a en général plusieurs paramètres. On a donc un vecteur de paramètres $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$. La formule de Bayes donnant la distribution *a posteriori* à partir de la loi *a priori* et de la vraisemblance est toujours valable dans ce cas : elle donne la loi *a posteriori* conjointe des d paramètres. Toute l'information est contenue dans cette loi conjointe. Mais son calcul numérique n'est pas toujours facile, notamment dans des modèles complexes, et dans certains modèles même la vraisemblance est difficile à calculer. De plus, pour obtenir le *posterior* conjoint, il faut aussi calculer la constante d'intégration $f(\mathbf{y}) = \int_{\Theta^d} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. Une solution analytique n'est disponible que dans des cas très particuliers (notamment lors de l'utilisation de lois conjuguées) ; et dans la majorité des cas pratiques, l'intégrale doit en fait être calculée numériquement. Si $\boldsymbol{\theta}$ est de dimension d , il s'agit de calculer une intégrale de multiplicité d , ce qui devient difficile lorsque d est grand (les problèmes numériques apparaissent dès que $d > 4$).

Un problème encore plus difficile surgit lorsque l'on veut tirer des conclusions à partir de cette distribution *a posteriori* conjointe. En général nous sommes intéressés par les valeurs possibles pour chaque paramètre. C'est à dire que l'on a besoin des distributions marginales, unidimensionnelles, de chaque paramètre. Pour les obtenir il faut là aussi intégrer la distribution conjointe (et ce d fois s'il y a d paramètres). Le problème est d'autant plus difficile qu'il faut calculer ces intégrales pour chaque valeur possible du paramètre pour reconstituer toute la densité *a posteriori* numériquement.. Dans les problèmes complexes un calcul suffisamment précis de ces intégrales paraît impossible et l'on a recours en général à des algorithmes basés sur des simulations d'échantillonnage, en particulier les algorithmes de Monte-Carlo par chaînes de Markov (dits « *Monte Carlo Markov Chain* » – MCMC).

3.1.2 Statistique bayésienne computationnelle

Identifier la loi *a posteriori* apparaît simple en théorie grâce au théorème de Bayes. Mais en pratique le calcul de l'intégrale au dénominateur s'avère souvent extrêmement difficile. Trouver une expression analytique n'est possible que dans quelques cas bien particuliers, et l'évaluation numérique peut se révéler tout aussi difficile, notamment lorsque la dimension de l'espace des

paramètres augmente.

La statistique bayésienne computationnelle cherche des solutions pour pouvoir estimer la distribution *a posteriori*, y compris lorsqu'on ne connaît que le numérateur dans le théorème de Bayes (loi *a posteriori* non-normée). Les principales méthodes utilisées s'appuient sur des algorithmes d'échantillonnage permettant de générer un échantillon distribué selon la loi *a posteriori*. On peut distinguer deux grands types parmi ces algorithmes : i) d'abord les méthodes d'échantillonnage directes, où l'on génère un échantillon à partir d'une loi simple (par exemple uniforme), que l'on transforme afin que le résultat soit distribué selon le *posterior* ; ii) les méthodes de Monte-Carlo par chaînes de Markov (MCMC), où l'on construit une chaîne de Markov sur l'espace des paramètres dont la loi invariante correspond à la loi *a posteriori*.

3.1.3 Méthode de Monte-Carlo

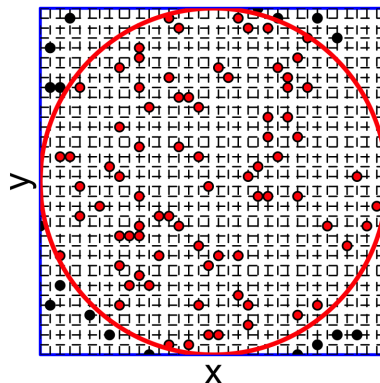
Monte-Carlo (Metropolis & Ulam 1949) est le nom crypté d'un projet de John von Neumann et Stanislas Ulam au *Los Alamos Scientific Laboratory* visant à utiliser des nombres aléatoires pour estimer des quantités difficiles (ou impossible) à calculer analytiquement.

En s'appuyant sur la loi des grands nombres, il s'agit d'obtenir un échantillon dit « de Monte-Carlo » permettant de calculer divers fonctionnelles à partir de la distribution de probabilité suivie par l'échantillon. En effet $\mathbb{E}[f(X)] = \int_x f(x)p_X(x)dx$. Or grâce à la loi des grands nombres, on a : $\mathbb{E}[f(X)] = \frac{1}{N} \sum_i f(x_i)$ à condition que les x forment un échantillon *iid* selon la loi de X . On peut ainsi estimer un certain nombre d'intégrales, à condition d'être capable d'échantillonner selon $p(x)$.

Exemple : Estimation du nombre $\pi = 3,14 \dots$ à l'aide de nombres aléatoires.



Une roulette de casino (à Monte-Carlo?)



Un cible quadrillée en 36×36

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

- 3 Placer ces points dans le repère et compter le nombre qui sont dans le cercle.
- 4 Calculer le ratio (probabilité estimée d'être dans le cercle) :

$$\hat{p}_C = \frac{\sum P_i \in \text{cercle}}{n}$$

Si $n = 1000$ et que l'on trouve 765 points sont dans le cercle, alors on a $\hat{\pi} = 4 \times \frac{765}{1000} 3,14159 \approx 3,06$. On pourrait améliorer notre estimation en augmentant la résolution de notre grille, et aussi en augmentant notre nombre de points n . En effet, on a $\lim_{n \rightarrow +\infty} \hat{p}_C = p_C$ d'après la loi des grands nombres.

Ainsi on a construit un **échantillon de Monte-Carlo** à partir de cet échantillon on peut calculer de nombreuses fonctionnelles, notamment π qui correspond à 4 fois la probabilité d'être dans le cercle.

De manière analogue, les méthodes d'échantillonnage directes ou par MCMC cherchent à construire un échantillon de Monte-Carlo du posterior, afin de calculer un certain nombre de fonctionnelles (Moyennes a posteriori, intervalles de crédibilité, etc.)...

3.2 Méthodes d'échantillonnage directes

3.2.1 Génération de nombres aléatoires selon des lois de probabilité usuelles

Il existe plusieurs manières de générer des nombres dits « aléatoires » selon des lois connues. La très grande partie des programmes informatiques ne génèrent pas des nombres totalement aléatoires. On parle plutôt de nombres pseudo-aléatoires, qui semblent aléatoires mais sont en réalité générés selon un processus déterministe (qui dépend notamment d'une « graine »).

La distribution uniforme

Pour générer un échantillon pseudo-aléatoire selon la loi uniforme sur $[0; 1]$, on peut donner l'exemple de l'algorithme congruentiel linéaire (Lehmer, 1948) :

- 1 Générer une suite d'entiers y_n tel que :

$$y_{n+1} = (ay_n + b) \bmod m$$
- 2 $x_n = \frac{y_n}{m - 1}$

Choisir a , b et m de manière à ce que y_n ait une période très longue et que (x_1, \dots, x_n) puisse être considéré comme *iid*

où y_0 est appelé la « graine » (*seed* en anglais). On remarque que l'on a nécessairement $0 \leq y_n \leq m - 1$. En pratique on prend m très grand (par exemple 2^{19937} , la valeur par défaut dans R qui utilise l'algorithme Mersenne-Twister). Dans ce cours, on ne va pas plus s'intéresser à la génération de nombre pseudo-aléatoires selon la loi uniforme sur $[0; 1]$, il s'agit d'un outil que l'on considère fiable et qui est utilisé par les différents algorithmes présentés par la suite.

Autres distributions

Pour échantillonner selon la loi binomiale $Bin(n, p)$, on peut utiliser les **relations entre les différentes lois usuelles** en partant de $U_i \sim U_{[0;1]}$:

$$Y_i = \mathbf{1}_{U_i \leq p} \sim \text{Bernouilli}(p)$$
$$X = \sum_{i=1}^n Y_i \sim Bin(n, p)$$

Pour échantillonner selon la loi Normale $N(0, 1)$, on peut utiliser l'algorithme de Box-Müller : Si U_1 et U_2 sont 2 variables uniformes $[0; 1]$ indépendantes, alors

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

sont indépendantes et suivent chacune la loi normale $N(0, 1)$.

3.2.2 Échantillonner selon une loi définie analytiquement

Méthode par inversion

Définition : Inverse généralisée

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

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

Propriété : Soit F la fonction de répartition d'une distribution de probabilité, et soit U une variable aléatoire suivant une loi uniforme sur $[0; 1]$. Alors $F^{-1}(U)$ définit une variable aléatoire ayant pour fonction de répartition F .

On déduit de la propriété ci-dessus que si l'on connaît la fonction de répartition de la loi selon laquelle on veut simuler, et si l'on est capable de l'inverser, alors on peut générer un échantillon suivant cette loi à partir d'un échantillon uniforme sur $[0; 1]$.

Exemple : On veut générer un échantillon suivant la loi exponentielle de paramètre λ

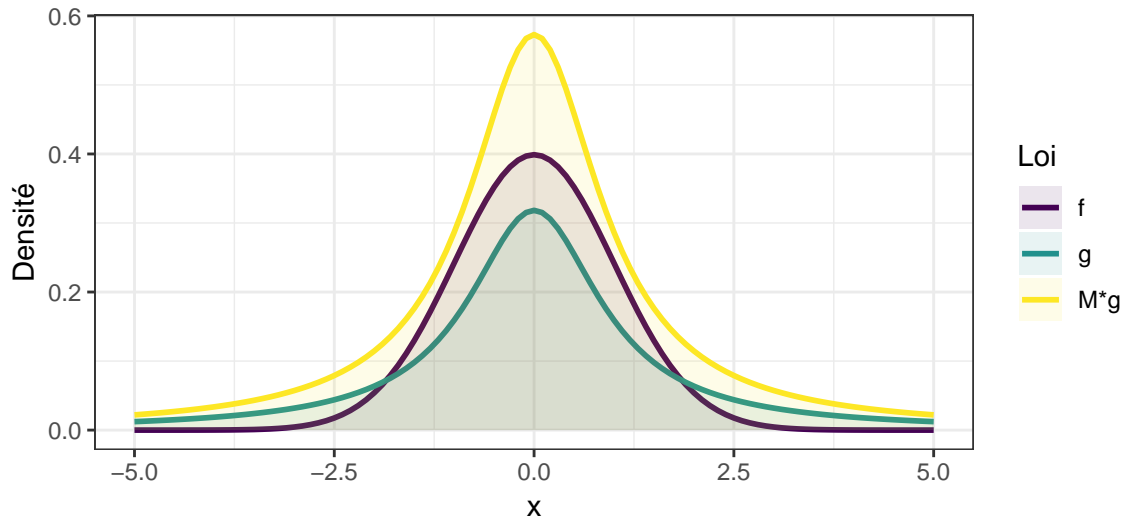
On a la densité de la loi exponentielle qui est $f(x) = \lambda \exp(-\lambda x)$, et la fonction de répartition (son intégrale) qui vaut $F(x) = 1 - \exp(-\lambda x)$.

Posons $F(x) = u$. On remarque alors que $x = -\frac{1}{\lambda} \log(1 - u)$.

Si $U \sim U_{[0;1]}$, alors $X = F^{-1}(U) \sim \text{Exp}(\lambda)$.

Méthode d'acceptation-rejet

La méthode d'acceptation-rejet consiste à utiliser une loi instrumentale g , dont on sait échantillonner selon la loi, afin d'échantillonner selon la loi cible f . Le principe générale est de choisir g proche de f et de proposer des échantillons selon g , d'en accepter certains et d'en rejeter d'autres afin d'obtenir un échantillon suivant la loi de f .



Exemple de loi de proposition et de loi cible pour l'algorithme d'acceptation-rejet

Soit une loi d'intérêt de densité f .

Soit une loi de proposition de densité g (à partir de laquelle on sait échantillonner) telle que, pour tout x :

$$f(x) \leq M g(x)$$

Pour $i = 1, \dots, n$:

1 Générer $x_i \sim g$ et $u_i \sim U_{[0;1]}$

2 Si $u_i \leq \frac{f(x_i)}{M g(x_i)}$ on **accepte** le tirage :

$$y_i := x_i$$

sinon on le **rejette** et on retourne en 1.

$$(y_1, \dots, y_n) \stackrel{iid}{\sim} f$$

Plus M est petit, plus le taux de rejet est faible et plus l'algorithme est efficace (au sens où il nécessite moins d'itérations pour obtenir un échantillon de taille n). On a donc intérêt à choisir g le plus proche possible de f , en particulier lorsque la dimension augmente (l'impact de M étant d'autant plus important alors). Néanmoins, la loi de proposition aura nécessairement des queues plus lourdes que la loi cible, et ce dans toutes les dimensions de l'espace des paramètres. À cause du fléau de la dimension, lorsque le nombre de paramètres augmente, le taux d'acceptation décroît très rapidement.

Exercice 1 : Construire un pseudo-échantillon de taille n selon la loi discrète suivante (multinomiale à m éléments $\{x_1, \dots, x_m\}$) :

$$P(X = x) = p_1 \delta_{x_1}(x) + p_2 \delta_{x_2}(x) + \dots + p_m \delta_{x_m}(x) \quad \text{avec} \quad \sum_{i=1}^m p_i = 1 \quad \text{et} \quad \delta_a(x) = \mathbb{1}_{\{x=a\}}$$

Exercice 2 : Grâce à la méthode par inversion, générer un pseudo-échantillon de taille suivant une loi de Cauchy (dont la densité est $f(x) = \frac{1}{\pi(1+x^2)}$), sachant que $\arctan'(x) = \frac{1}{(1+x^2)}$ et que

$$\lim_{x \rightarrow -\infty} \arctan(x) = -\frac{\pi}{2}.$$

Exercice 3 : Écrire un algorithme d'acceptation-rejet pour simuler la réalisation d'un pseudo-échantillon de taille n d'une loi normale $N(0, 1)$ en utilisant une loi de Cauchy comme proposition. Trouvez la valeur de M optimale.

3.3 Algorithmes MCMC

Le principe des algorithmes MCMC est de construire une chaîne de Markov visitant l'espace des paramètres dont la loi de probabilité invariante est la loi *a posteriori*.

3.3.1 Chaînes de Markov

Une chaîne de Markov est un processus stochastique à temps discret. On peut la définir comme une suite de variable aléatoire $X_0, X_1, X_2, X_3, \dots$ (toutes définies sur le même espace) possédant la **propriété de Markov** (« sans mémoire ») :

$$p(X_i = x | X_0 = x_0, X_1 = x_1, \dots, X_{i-1} = x_{i-1}) = p(X_i = x | X_{i-1} = x_{i-1})$$

L'ensemble des valeurs possible pour X_i est appelé **espace d'état** et est noté E .

Une chaîne de Markov est déterminée par deux paramètres :

- 1 sa distribution initiale $p(X_0)$
- 2 son noyau de transition $T(x, A) = p(X_i \in A | X_{i-1} = x)$

Dans la suite, on ne va considérer que des chaînes de Markov **homogènes**, c'est-à-dire qui vérifie :

$$p(X_{i+1} = x | X_i = y) = p(X_i = x | X_{i-1} = y)$$

Propriété : Une chaîne de Markov est dite **irréductible** : si tous les ensembles de probabilité non nulle peuvent être atteints à partir de tout point de départ (i.e. tout état est accessible à partir de n'importe quel autre).

Propriété : Une chaîne de Markov est dite **récurrente** si les trajectoires (X_i) passent une infinité de fois dans tout ensemble de probabilité non nulle de l'espace d'état.

Propriété : Une chaîne de Markov est dite **apériodique** si rien n'induit un comportement périodique des trajectoires.

Définition : Une distribution de probabilité \tilde{p} est appelée **loi invariante** (ou **loi stationnaire**) pour une chaîne de Markov si elle vérifie la propriété suivante : si X_i suit \tilde{p} , alors X_{i+1} (et les éléments suivants) sont nécessairement distribués suivant \tilde{p} .

Remarque : Une chaîne de Markov peut admettre plusieurs lois stationnaires.

Théorème ergodique (espace infini) : Une chaîne de Markov irréductible et récurrente positive (i.e. le temps de retour moyen est fini) admet une unique loi de probabilité invariante \tilde{p} . Si cette chaîne de Markov est de plus apériodique, alors elle converge en loi vers \tilde{p} .

Exemple : Doudou le hamster

Nous allons maintenant développer un exemple d'une chaîne de Markov à espace d'état discret.

Supposons que l'état de Doudou, le hamster, suive à chaque minute une chaîne de Markov discrète à trois états : dormir (D), manger (M), faire de l'exercice (E). Ainsi, son état dans une minute ne dépend que de son état actuel, et pas des minutes précédentes. Supposons que la matrice des probabilité de transition soit alors la suivante :

$$P = \begin{pmatrix} X_i/X_{i+1} & D & M & E \\ D & 0.9 & 0.05 & 0.05 \\ M & 0.7 & 0 & 0.3 \\ E & 0.8 & 0 & 0.2 \end{pmatrix}$$

1) Selon vous, la chaîne est-elle irréductible ? Récurrente ? Apériodique ?

2) Supposons que Doudou dorme. Que fait-il 2 min après ? et 10 min après ?

$$x_0 =$$

3) Supposons maintenant qu'il fasse de l'exercice. Que fait-il 10 min après ?

$$x_0 =$$

Ici la loi est apériodique, récurrente et irréductible, il y a donc une loi stationnaire : $\tilde{p} = \tilde{p}P$.

3.3.2 Échantillonnage MCMC

Algorithmes MCMC : principe général

Le principe général des algorithmes MCMC est le suivant : pour produire une approximation acceptable d'une intégrale ou d'une autre fonctionnelle d'une distribution d'intérêt (i.e. la loi *a posteriori*), il suffit de générer une chaîne de Markov dont la distribution limite est la distribution d'intérêt (i.e. la loi *a posteriori*), puis d'y appliquer la méthode de Monte-Carlo.

Il faut donc avoir une **double convergence** :

- 1 la convergence de la chaîne de Markov vers sa distribution stationnaire : $\forall x_0, X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \tilde{p}$

2 la convergence de Monte-Carlo, une fois la distribution stationnaire atteinte :

$$\frac{1}{N} \sum_{i=1}^N f(X_{n+i}) \xrightarrow[N \rightarrow +\infty]{} \mathbb{E}[f(X)]$$

$$\overbrace{X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n}^{\text{convergence de la chaîne de Markov}} \rightarrow \overbrace{X_{n+1} \rightarrow X_{n+2} \rightarrow \dots \rightarrow X_{n+N}}^{\text{échantillon de Monte-Carlo}}$$

Les algorithmes MCMC utilisent une approche d'acceptation-rejet :

```

1 Initialiser  $x^{(0)}$ 

2 Pour  $t = 1, \dots, n + N$  :
    a. Proposer un nouveau candidat  $y^{(t)} \sim q(y^{(t)} | x^{(t-1)})$ 
    b. Accepter  $y^{(t)}$  avec la probabilité  $\alpha(x^{(t-1)}, y^{(t)})$  :
         $x^{(t)} := y^{(t)}$ 
        Si  $t > n$ , « sauver »  $x^{(t)}$  (pour calculer la fonctionnelle d'intérêt)
    où  $q$  est la loi instrumentale de proposition et  $\alpha$  est la probabilité d'acceptation.

```

Schéma général des algorithmes MCMC

Pour la loi instrumentale de proposition q il n'existe pas de choix absolument optimal mais une infinité de lois possibles (certaines meilleures que d'autres). Afin de garantir la convergence vers la loi cible \tilde{p} : (i) le support de q doit contenir le support \tilde{p} , (ii) q ne doit pas générer de valeurs périodiques. Idéalement, on choisit q de manière à ce que son calcul soit simple (par exemple on peut choisir q symétrique).

L'algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings est un algorithme très simple et très général permettant d'échantillonner selon des lois uni- ou multi-dimensionnelles.

```

1 Initialiser  $x^{(0)}$ 

2 Pour  $t = 1, \dots, n + N$  :
    a. Proposer  $y^{(t)} \sim q(y^{(t)} | x^{(t-1)})$ 
    b. Calculer la probabilité d'acceptation
        
$$\alpha^{(t)} = \min \left\{ 1, \frac{\tilde{p}(y)}{q(y^{(t)} | x^{(t-1)})} \bigg/ \frac{\tilde{p}(x^{(t-1)})}{q(x^{(t-1)} | y^{(t)})} \right\}$$

    c. Étape d'acceptation-rejet : générer une valeur  $u^{(t)} \sim U_{[0;1]}$ 
        
$$x^{(t)} = \begin{cases} y^{(t)} & \text{si } u^{(t)} \leq \alpha^{(t)} \\ x^{(t-1)} & \text{sinon} \end{cases}$$


```

On peut reformuler la probabilité d'acceptation $\alpha^{(t)}$ ainsi : $\alpha^{(t)} = \min \left\{ 1, \frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})} \frac{q(x^{(t-1)}|y^{(t)})}{q(y^{(t)}|x^{(t-1)})} \right\}$.
On voit donc qu'on peut la calculer en ne connaissant \tilde{p} qu'à une constante près, puisqu'elle se simplifie dans ce ratio.

Dans certains cas particuliers (très utilisés en pratique), le calcul de $\alpha^{(t)}$ est simplifié :

- **Metropolis-Hastings indépendant** : $q(y^{(t)}|x^{(t-1)}) = q(y^{(t)})$
- **Metropolis-Hastings à marche aléatoire** : $q(y^{(t)}|x^{(t-1)}) = g(y^{(t)} - x^{(t-1)})$. Si g est symétrique ($g(-x) = g(x)$), alors le calcul de la probabilité d'acceptation $\alpha^{(t)}$ se simplifie :

$$\frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})} \frac{q(y^{(t)}|x^{(t-1)})}{q(x^{(t-1)}|y^{(t)})} = \frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})} \frac{g(y^{(t)} - x^{(t-1)})}{g(x^{(t-1)} - y^{(t)})} = \frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})}$$

L'algorithme de Metropolis-Hastings est un algorithme très simple et très général permettant d'échantillonner de manière uni- ou multi-dimensionnelle. Le choix de la distribution instrumentale est crucial, mais difficile, et a un impact considérable sur les performances de l'algorithme (un fort taux de rejet implique souvent des temps de calculs très importants). De plus, c'est un algorithme qui devient inefficace dans les problèmes de trop grande dimension. L'algorithme du recuit-simulé ainsi que l'échantillonneur de Gibbs sont des algorithmes permettant en partie de pallier à certaines de ces limites.

L'algorithme du recuit-simulé

Afin de palier à certaines limitation de l'algorithme de Metropolis-Hastings, ici l'idée est de faire varier le calcul de la probabilité d'acceptation $\alpha^{(t)}$ au cours de l'algorithme. La probabilité d'acceptation doit d'abord être grande afin de bien explorer l'ensemble de l'espace d'état, puis diminuer au fur et à mesure que l'algorithme converge vers une région de l'espace, afin que les nouvelles valeurs acceptées se concentre autour du mode de convergence. Cela consiste à introduire dans l'algorithme de Métropolis-Hastings une « température » variant à chaque itération et notée $T(t)$:

1 Initialiser $x^{(0)}$

2 Pour $t = 1, \dots, n + N$:

a. Proposer $y^{(t)} \sim q(y^{(t)}|x^{(t-1)})$

b. Calculer la probabilité d'acceptation

$$\alpha^{(t)} = \min \left\{ 1, \left(\frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})} \frac{q(x^{(t-1)}|y^{(t)})}{q(y^{(t)}|x^{(t-1)})} \right)^{\frac{1}{T(t)}} \right\}$$

c. Étape d'acceptation-rejet : générer une valeur $u^{(t)} \sim U_{[0;1]}$

$$x^{(t)} := \begin{cases} y^{(t)} & \text{si } u^{(t)} \leq \alpha^{(t)} \\ x^{(t-1)} & \text{sinon} \end{cases}$$

Par exemple, on peut prendre $T(t) = T_0 \left(\frac{T_f}{T_0} \right)^{\frac{t}{n}}$ avec T_0 la température de base, n le nombre d'itérations au-delà duquel on pense être proche de la convergence, T_f la température après n itérations. Cet algorithme est particulièrement utile lors de la présence d'optimums locaux.

Échantillonneur de Gibbs

Lorsque la dimension (de x) augmente, il devient très difficile de proposer des valeurs probables dans les algorithmes utilisant la stratégie d'acceptation-rejet. L'idée de l'échantillonneur de Gibbs est de générer x coordonnée par coordonnée, en conditionnant sur les dernières valeurs obtenues. Il faut donc que x admette une décomposition telle que $x = (x_1, \dots, x_d)$, et que les distributions $p(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$ soit connues et aisément simulables. L'échantillonneur de Gibbs est une suite d'étape de Metropolis-Hastings (coordonnée par coordonnée), les propositions échantillonnées sont toujours acceptée ($\alpha = 1$). On obtient cette acceptation inconditionnelle en imposant les lois de propositions : il s'agit de la distribution conditionnelle respective de chaque coordonnées. On peut donc voir l'échantillonneur de Gibbs comme un algorithme de réactualisation composante par composante :

- 1 Initialiser $x^{(0)} = (x_1^{(0)}, \dots, x_d^{(0)})$
- 2 Pour $t = 1, \dots, n + N$:
 - a. Générer $x_1^{(t)} \sim p(x_1 | x_2^{(t-1)}, \dots, x_d^{(t-1)})$
 - b. Générer $x_2^{(t)} \sim p(x_2 | x_1^{(t)}, x_3^{(t-1)}, \dots, x_d^{(t-1)})$
 - c. ...
 - d. Générer $x_i^{(t)} \sim p(x_i | x_1^{(t)}, \dots, x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, \dots, x_d^{(t-1)})$
 - e. ...
 - f. Générer $x_d^{(t)} \sim p(x_d | x_1^{(t)}, \dots, x_{d-1}^{(t)})$

Remarque : si l'on ne connaît pas certaines lois conditionnelles pour certaines coordonnées, on peut tout de les échantillonner en introduisant une étape d'acceptation-rejet pour cette coordonnée uniquement. On parle alors d'algorithme de Métropolis à l'intérieur de Gibbs (*Metropolis within gibbs*).

3.4 Les algorithmes MCMC pour l'inférence bayésienne dans la pratique

La mise en place d'algorithmes de Metropolis-Hastings, de Gibbs ou de Metropolis à l'intérieur de Gibbs peut ainsi permettre d'échantillonner selon le posterior dans le cadre d'un modèle bayésien. On remplace alors x par θ et \tilde{p} par $p(\theta | \mathbf{y})$. Un certain nombre de logiciels tels que JAGS (<http://mcmc-jags.sourceforge.net/>), STAN (<http://mc-stan.org/>) ou WinBUGS (<https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/>) propose une implémentation de tels algorithmes.

Le projet BUGS (*Bayesian inference Using Gibbs Sampling* : <https://www.mrc-bsu.cam.ac.uk/software/bugs/>) a été initié en 1989 par l'unité de Biostatistique du MRC (*Medical Research Council*) de l'Université de Cambridge (au Royaume-Uni) afin de proposer un logiciel flexible pour l'analyse bayésienne de modèles statistique complexe à l'aide d'algorithme MCMC. Son implémentation la plus connue est WinBUGS, un logiciel clic-bouton disponible sous le système d'exploitation *Windows*. OpenBUGS est une implémentation fonctionnant sous *Windows*, *Mac OS* ou *Linux*.

JAGS (*Just another Gibbs Sampler*) est une autre implémentation plus récente qui s'appuie également sur le langage **BUGS**. Enfin, il faut également noter le logiciel **STAN**, récemment développé à l'Université de Columbia qui n'est similaire à **BUGS** que dans son interface, s'appuyant sur des algorithmes MCMC innovants, comme par exemple les approches de Monte-Carlo Hamiltonien ou l'approche variationnelle. Une ressource très utile est le manuel de l'utilisateur de JAGS (http://sourceforge.net/projects/mcmc-jags/files/Manuals/3.x/jags_user_manual.pdf).

3.4.1 Convergence des algorithmes MCMC

L'échantillonnage selon la distribution *a posteriori* par un algorithme MCMC comporte 2 phases :

La **phase de chauffe** (*burn-in*) : Elle correspond aux premières itérations de l'algorithme MCMC, qui ne doivent pas être conservées dans l'analyse de l'échantillon de Monte-Carlo. En effet, celles-ci ne proviennent pas de la distribution. Cette phase correspond donc au temps nécessaire à la chaîne de Markov pour converger vers sa loi stationnaire. Sa longueur varie d'un modèle à l'autre. Il n'y a aucune conséquence à prendre une phase de chauffe trop longue, si ce n'est

La **phase d'échantillonnage** : elle doit être suffisamment longue pour permettre une bonne estimation de la distribution *a posteriori*, notamment concernant les plages de valeurs de faible probabilité.

Les propriétés mathématiques des chaînes de Markov garantissent la convergence des algorithmes MCMC, mais sans donner d'indication sur le nombre d'itérations de l'algorithme nécessaire pour atteindre cette convergence. S'il n'existe aucun moyen de garantir cette convergence en temps fini, il existe un certain nombre d'outils permettant de diagnostiquer la non-convergence d'une chaîne de Markov vers sa loi stationnaire. Il faut donc les utiliser de concert lors de l'interprétation des sorties d'un algorithme MCMC afin d'éviter les situations où la chaîne n'a pas convergé.

Une manière de surveiller la convergence d'un algorithme d'échantillonnage MCMC est de générer plusieurs chaînes (de façon parallèle et indépendante) avec des valeurs initiales différentes. Si l'algorithme fonctionne, alors ces différentes chaînes (de Markov) doivent converger vers la même distribution stationnaire (la loi *a posteriori*). Après suffisamment d'itérations, il devrait être impossible de faire la distinction entre ces différentes chaînes. Pour chaque chaîne, les n premières valeurs sont considérées comme appartenant à la **phase de chauffe** (*burn-in*) de l'algorithme, nécessaire pour que la chaîne de Markov converge d'abord vers sa loi stationnaire à partir des valeurs initiales. Elles ne sont donc pas conservées, et on s'intéresse aux N observations suivantes qui vont constituer nos échantillons de Monte-Carlo.

Erreur de Monte-Carlo

L'erreur de Monte-Carlo caractérise l'incertitude introduite par l'échantillonnage MCMC. Pour un paramètre donné, elle quantifie la variabilité attendue dans son estimation si nous générerions plusieurs **chaînes**, c'est-à-dire plusieurs échantillons de Monte-Carlo *a posteriori* (grâce à un algorithme MCMC, avec différentes initialisations et à chaque fois un même nombre N d'itérations). Les erreurs-types de Monte Carlo donne une idée de cette variabilité. Si les erreurs standards ont des valeurs très différentes d'une chaîne à l'autre, alors il faut faire fonctionner l'échantillonneur plus longtemps. La longueur exacte de l'échantillonnage nécessaire pour obtenir une erreur-type donnée dépendra de l'efficacité et du mélange de l'échantillonneur. Il est important que cette erreur de Monte-Carlo soit faible au regard de la variance estimée de la loi *a posteriori*.

Statistique de Gelman-Rubin

Une façon de d'évaluer la convergence d'un échantillonneur MCMC est de comparer la variation entre les différentes chaînes à la variation à l'intérieur d'une même chaîne après un certain nombre d'itérations. Si l'algorithme a bien convergé, la variation inter-chaîne doit être proche de zéro.

Soit $\theta_{[c]} = (\theta_{[c]}^{(1)}, \dots, \theta_{[c]}^{(N)})$ le N -échantillon obtenu à partir de la chaîne $c = 1, \dots, C$ d'un algorithme MCMC échantillonnant θ . La **statistique de Gelman-Rubin** s'écrit :

$$R = \frac{\frac{N-1}{N} W \frac{1}{N} B}{W}$$

avec $B = \frac{N}{C-1} \sum_{c=1}^C (\bar{\theta}_{[C]} - \bar{\theta})^2$ la variance inter-chaînes, $\bar{\theta}_{[c]} = \frac{1}{N} \sum_{t=1}^N \theta_{[c]}^{(t)}$, $\bar{\theta} = \frac{1}{C} \sum_{c=1}^C \bar{\theta}_{[C]}$, et $W = \frac{1}{C} \sum_{c=1}^C s_{[c]}^2$ la variance intra-chaîne, $s_{[c]}^2 = \frac{1}{N-1} \sum_{t=1}^N (\theta_{[c]}^{(t)} - \bar{\theta}_{[C]})^2$. Lorsque $N \rightarrow +\infty$ tandis que $B \rightarrow 0$, R s'approche de la valeur de 1. On va donc chercher à itérer suffisamment un algorithme MCMC afin d'obtenir une valeur de R proche de 1, par exemple entre 1 et 1,01 ou 1,05.

La statistique Gelman-Rubin est un ratio (donc sans unité) ce qui en fait un résumé s'interprétant simplement et de la même manière pour tout échantillonneur MCMC. Autre avantage, il ne nécessite pas de choisir au préalable un paramètre à estimer (contrairement aux erreurs de Monte Carlo). La statistique de Gelman-Rubin est donc un bon moyen de diagnostiquer la convergence d'un algorithme MCMC. Néanmoins, son calcul peut être instable et elle ne peut garantir la convergence à elle seule. Il s'agit d'un outil général pour la surveillance plus générale d'une chaîne de Markov.

À noter que d'autres statistiques (par exemple la statistique de Geweke) sont parfois utilisées à la place ou en complément de celle de Gelman-Rubin, qui reste la plus populaire.

Diagnostiques graphiques

En complément de la statistique de Gelman-Rubin et, un certain nombre de diagnostics graphiques peuvent permettre de diagnostiquer la non convergence d'un algorithme MCMC :

- la **trace** : désigne la représentation des valeurs successives de la chaîne. Lorsque l'on génère plusieurs chaînes indépendantes à partir d'initialisations différentes, les traces des différentes chaînes doivent se stabiliser et se superposer une fois la convergence atteinte.
- **estimateur de densité non-paramétrique** (à noyau) : d'après le théorème de convergence de Bernstein-von Mises, la distribution *a posteriori* doit être unimodale. Pour cela on peut utiliser un estimateur de densité non-paramétrique (à noyau) sur l'échantillon de Monte-Carlo généré afin de vérifier que la loi *a posteriori* est bien unimodale et suffisamment lisse.
- les **quantiles courants** : de la même façon que la trace, les quantiles des différentes chaînes doivent se stabiliser et superposer au cours des différentes itérations une fois la convergence atteinte.
- le **diagramme de Gelman-Rubin** : on représente la statistique de Gelman-Rubin cumulée au cours des différentes itérations. Son niveau doit rapidement se retrouver très proche de 1 (idéalement $< 1,01$ ou à minima $< 1,05$)
- l'**auto-corrélation** : Lorsque la chaîne de Markov ne « mixe » pas très bien, il peut arriver que les observations successives soient fortement corrélées d'une itération à la suivante. Cela n'est pas un problème en soit, mais cela diminue fortement la taille d'échantillon effective pour l'estimation *a posteriori*. Une solution courante est de ne conserver qu'une itération sur 2, 5 ou 10 (on espacera les enregistrements d'autant plus que la corrélation est forte) à

l'aide du paramètre d'épaisseur (*thin*, réglant l'espacement entre les itérations conservées dans l'échantillon MCMC).

- la **corrélation croisée** : On peut également s'intéresser à la corrélation entre nos différents échantillons *a posteriori*. À noter qu'il est fréquent d'observer une forte corrélation entre certains paramètres et que ce n'est pas nécessairement indicateur d'un problème avec l'algorithme MCMC (l'approche fréquentiste également on estime des corrélations, parfois importantes, entre les paramètres d'un modèle grâce à la matrice d'information de Fisher).

Remarque : il est fréquent que les diagnostics soient satisfaisants pour certains paramètres, mais ne le soient pas pour d'autres. Il s'agit d'une appréciation subjective, et l'objectif est que la majorité des critères soient satisfaits (ou plus ou moins satisfaits) pour une grande majorité des paramètres.

Taille d'échantillon effective

Un échantillon généré à partir d'un algorithme MCMC n'est *iid* en pratique que dans des cas très particuliers. En effet, la propriété de Markov entraîne généralement une corrélation entre les valeurs générées à la suite les unes des autres (échantillonnage dépendant). Pour une taille N fixé d'échantillon, cette auto-corrélation diminue la quantité d'information et ralentit la convergence de la loi des grands nombres par rapport à un échantillon totalement indépendant. Un indicateur permettant de quantifier cette information est la **taille d'échantillon effective** (*effective sample size* en anglais) qui se calcule :


$$ESS = \frac{N}{1 + 2 \sum_{k=1}^{+\infty} \rho(k)}$$

où $\rho(k)$ désigne l'auto-corrélation avec *lag* de rang k .

Un solution employée en pratique pour diminuer ces problèmes d'auto-corrélation est de ne pas conserver toutes les valeurs échantillonnées successivement par un algorithme MCMC, mais d'espacer les itérations conservées. Par exemple, on pourra ne conserver que les valeurs échantillonnées toutes les 2, 5, ou 10 itérations, ce qui permettra de diminuer la dépendance au sein de l'échantillon de Monte-Carlo généré.

3.4.2 Inférence à partir d'échantillonnage MCMC

Estimation

Grâce aux algorithmes MCMC, on est donc capable d'obtenir un **échantillon de Monte-Carlo de la loi *a posteriori*** pour un **modèle bayésien** donné. On peut donc utiliser la **méthode de Monte-Carlo** pour obtenir différentes **estimations *a posteriori*** : estimation ponctuelle (moyenne *a posteriori*, médiane *a posteriori*, ...), intervalle de crédibilité (notamment grâce au package  **HDInterval** qui permet de calculer l'intervalle de crédibilité le plus étroit pour un niveau donné, c'est-à-dire le *Highest Density Interval* – *HDI*)), les corrélations croisées entre les paramètres, etc.

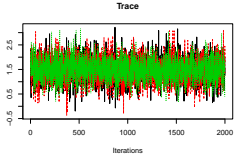
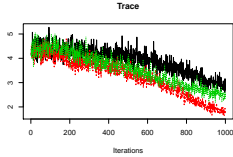
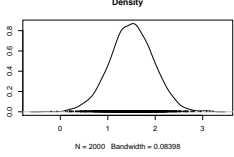
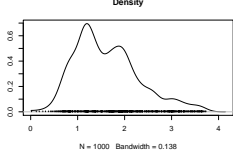
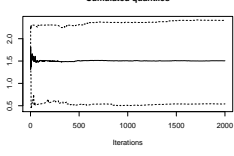
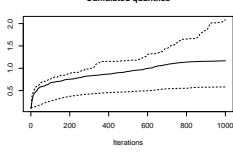
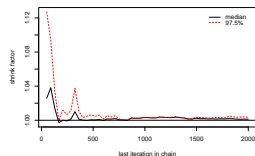
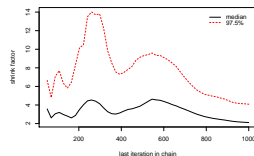
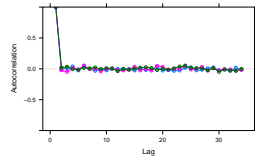
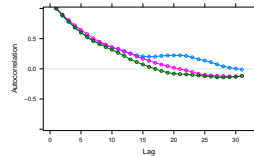
Graphique	😊	😞	R	Solution(s) potentielle(s)
trace			<code>coda::traceplot()</code>	↗ <code>n.iter</code> et/ou ↗ <i>burn-in</i>
densité			<code>coda::densplot()</code>	↗ <code>n.iter</code> et/ou ↗ <i>burn-in</i>
quantile courants			<code>coda::cumuplot()</code>	↗ <code>n.iter</code> et/ou ↗ <i>burn-in</i>
Gelman-Rubin			<code>coda::gelman.plot()</code>	↗ <code>n.iter</code> et/ou ↗ <i>burn-in</i>
Auto-corrélation			<code>coda::acfplot()</code>	↗ <code>thin</code> et/ou ↗ <code>n.iter</code> et/ou ↗ <i>burn-in</i>

TABLE 3.1 – Exemples de référence pour les diagnostics graphiques de convergence

Deviance Information Criterion (DIC)

Le ***Deviance Information Criterion (DIC)*** s'appuie sur la déviance¹, qui s'écrit comme : $D(\theta) = -2 \log(p(\theta|\mathbf{y})) + C$ où C est une constante. Le DIC est alors défini par :

$$DIC = \overline{D(\theta)} + p_D$$

où $p_D = (D(\bar{\theta}) - \overline{D(\theta)})$ représente une pénalité pour le nombre effectif de paramètres. Le DIC permet notamment de comparer différents modèles sur les mêmes données, et de faire des choix de modélisation dans le contexte bayésien.

3.5 Autres méthodes

3.5.1 Bayésien variationnel

L'inférence bayésienne variationnelle est une technique d'approximation de l'approche bayésienne s'intéressant à l'estimation des moyennes *a posteriori*, ainsi qu'à l'incertitude qui leur entoure. Elle s'appuie sur une approximation paramétrique de la loi *a posteriori* qui minimise la divergence de Kullback-Leibler par rapport à la véritable distribution *a posteriori*. Le calcul de la solution bayésienne variationnelle revient donc à un problème d'optimisation classique, dont le calcul généralement très rapide ce qui peut en faire une solution dans les problèmes de données massives. Néanmoins la qualité de l'approximation variationnelle dépendra de l'adéquation du modèle paramétrique choisie pour lequel on ne dispose pas de garanties.

3.5.2 Calcul Bayésien Approché (ABC)

Le calcul bayésien approché par rejet (*ABC* en anglais) désigne une autre alternative aux méthodes MCMC, qui utilise le modèle d'échantillonnage pour éviter d'avoir à calculer la vraisemblance au numérateur de la formule de Bayes en générant des observations selon le modèle génératif des données. On obtient alors un échantillon de la loi *a posteriori* en conservant les valeurs de paramètre θ , générées à partir de la loi *a priori*, ayant permis de générer les échantillons suffisamment proche des données réellement observées. La difficulté de cette approche réside dans la formalisation de ce « suffisamment proche », qui induit une approximation par rapport à l'approche bayésienne exacte qui conserverait toutes les valeurs de θ (mais dont le coût de calcul par cette méthode est alors souvent très important).

1. M Plummer, Penalized loss functions for Bayesian model comparison, *Biostatistics*, 2008

Chapitre 4

Méthodes numériques pour la statistique

Avec le développement de l'ordinateur moderne, les aspects computationnels n'ont cessé d'avoir une place de plus en plus importante dans la Statistique appliquée. En effet, l'application concrète des concepts théoriques de la Statistique mathématique pour l'analyse de données nécessite bien souvent de puissants outils algorithmiques et informatiques. Dans le chapitre précédent, nous avons présenté les méthodes MCMC, très utiles pour effectuer en pratique une analyse bayésienne. Dans ce chapitre, nous allons étudier d'autres méthodes numériques très utilisées dans la Statistique.

4.1 Ré-échantillonnage et Monte-Carlo : la méthode du *Bootstrap*

Le *bootstrap*, développé par Efron en 1992, est une technique d'estimation reposant sur la méthode de Monte Carlo plutôt que sur le calcul analytique de la distribution asymptotique d'un estimateur. Elle est donc particulièrement utile lorsqu'on ne connaît pas la distribution des observations, ou lorsque les hypothèses de méthodes classiques ne sont pas vérifiées (e.g. normalité). Le *bootstrap* remplace les calculs analytiques par des calculs numériques en utilisant le ré-échantillonnage.

Le principe fondamental du *bootstrap* est d'utiliser la distribution empirique \hat{F}_n pour estimer la distribution des observations (inconnue donc). Cette loi \hat{F}_n peut alors être facilement utilisée pour générer d'autres échantillons de même taille n que l'échantillon observé, qui sont *iid* selon cette distribution empirique, en opérant un n **tirages aléatoires avec remise** dans l'échantillon observé. Ensuite, il suffit d'appliquer la méthode de Monte-Carlo afin d'estimer toute fonctionnelle d'intérêt (par exemple la moyenne, la variance, etc). Là où le *bootstrap* prend toute sa puissance, c'est lorsqu'on l'utilise pour estimer la variance ou le biais d'un estimateur dont on connaît l'expression analytique, mais dont on ne connaît pas la distribution asymptotique. Cela représente donc par exemple une alternative computationnelle à la delta méthode.

4.2 Algorithmes d'optimisation

Un des problèmes fondamentaux en statistique est l'optimisation numérique d'une fonction (par exemple la vraisemblance ou la distribution *a posteriori*). Il existe différentes approches computationnelles pour l'optimisation numérique.

Notons ici que maximiser une fonction revient à minimiser son opposé. En conséquence, on va généralement formuler un problème d'optimisation numérique sous forme d'une minimisation : on

cherche $\theta \in \Theta$ tel que $f(\theta) \in \mathbb{R}$ soit minimale ($\Theta = \mathbb{R}^d$ par exemple). Une approche générale pour résoudre ce problème est la recherche linéaire. Elle consiste à partir d'un estimateur courant θ_i du minimum et à

- 1 choisir une direction p_i de l'espace Θ (de dimension d)
- 2 choisir une longueur de pas α_i à parcourir dans cette direction p_i (généralement en résolvant $\min_{\alpha} f(\theta_i + \alpha p_i)$)
- 3 mettre à jour notre estimé par $\theta_{i+1} = \theta_i + \alpha_i p_i$

On voit donc qu'il y a deux problèmes à résoudre à chaque itération d'un tel algorithme : i) comment choisir p_i ; ii) et comment choisir α . Dans les différentes réponses apportées à ce problème, il est important de garder à l'esprit que la dimension d est souvent importante, et qu'il convient donc de minimiser la quantité de calculs nécessaire, de même que la mémoire informatique.

Concernant le choix de la direction p_i , une solution parmi les plus intuitive est de choisir « le gradient maximal », c'est-à-dire $-f'(\theta_i)$. En effet, c'est la direction dans laquelle f change le plus rapidement à partir de θ_i . C'est ce que l'on appelle le **gradient maximal**. À noter que bien que cette solution semble tout à fait raisonnable, lorsque certains paramètres sont très corrélés entre eux, elle nécessite un grand nombre d'itération (car elle progresse alors « en zigzag » vers le minimum).