Statistique bayésienne et algorithmes associés

Robin Ryder

Version du 6 septembre 2019

Table des matières

1 Rappels mathématiques

1.1 Formule de Bayes

Théorème 1 (Formule de Bayes) *Soient A et B deux événements avec $P[A] > 0$. Alors la probabilité conditionnelle de B sachant A est donnée par*

$$P[B|A] = \frac{P[A|B] \cdot P[B]}{P[A]}.$$

Preuve La probabilité jointe $P[A \cap B]$ peut être exprimée de deux manières :

$$P[A \cap B] = P[B|A] \cdot P[A] = P[A|B] \cdot P[B].$$

En réarrangeant les termes, on trouve l'égalité cherchée.

1.2 Densités de probabilité

Dans le cadre de ce cours, on s'intéresse uniquement aux lois de probabilités qui sont absolument continues par rapport à une mesure de référénce qui est soit la mesure de Lebesgue (cas continu) soit la mesure de comptage (cas discret). Ce cadre est suffisant pour traiter la plupart des applications. On peut aisément l'étendre à des situations avec une autre mesure de référence : une des difficultés principales est alors de développer une intuition de ce que sont ces autres mesures de référence.

Définition 1 *Soit X une variable aléatoire réelle absolument continue par rapport à la mesure de Lebesgue (autrement dit, une v.a. continue). Soit F_X sa fonction de répartition, définie sur $\mathbb{R}$ par $F_X(t) = P[X \leq t]$. Alors F_X est dérivable presque partout et on appelle fonction de densité de X la fonction*

$$f(t) = \frac{dF_X(t)}{dt}$$

définie en tous les points de dérivabilité de F_X.

On peut prolonger f sur $\mathbb{R}$ tout entier ; le choix du prolongement n'importe pas en pratique.

Propriétés d'une densité Pour une variable aléatoire X continue, la fonction de densité associée f vérifie les propriétés suivantes :

- la fonction f est positive : $\forall x, f(x) \geq 0$;
- la fonction f est intégrable sur $\mathbb{R}$ et $\int_{-\infty}^{+\infty} f = 1$;
- pour tous réels a et b avec $a < b$, on a $P[a < X < b] = \int_a^b f(t)\,dt$;
- de façon plus générale, pour tout ensemble mesurable A, on a $P[X \in A] = \int_A f(t)\,dt$.

Notons que si g est une fonction positive intégrable sur $\mathbb{R}$, on peut définir une unique fonction de densité f proportionnelle à g :

$$f(x) = \frac{g(x)}{\int_{-\infty}^{+\infty} g(t)\,dt}.$$

Définition 2 *Soit X une variable aléatoire à valeurs dans un ensemble fini ou dénombrable $\Omega = \{x_1, x_2, \ldots\}$. On appelle densité de X par rapport à la mesure de comptage la fonction f définie par*

$$f(x) = \begin{cases} P[X = x] & si\ x \in \Omega \\ 0 & sinon. \end{cases}$$

1.3 Fonction Gamma

Définition 3 *On appelle fonction Gamma la fonction définie sur $\mathbb{R}_+$ par*

$$\Gamma : z \mapsto \int_0^\infty x^{z-1} e^{-x}\, dx.$$

Cette intégrale est également appelée intégrale d'Euler de seconde espèce.

Propriétés La fonction Gamma vérifie les propriétés utiles suivantes :

1. $\forall z \in \mathbb{R}, \Gamma(z+1) = z\Gamma(z)$
2. $\Gamma(1) = 1$
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
4. $\forall n \in \mathbb{N}, \Gamma(n) = (n-1)!$
5. Formule de Stirling : quand $z \to \infty$, $\Gamma(z+1) \sim \sqrt{2\pi z} z^z e^{-z}$.

Preuve Prouvons ces propriétés.

1. On intègre par parties.

$$\begin{aligned} \Gamma(z+1) &= \int_0^\infty x^z e^{-x}\, dx \\ &= \left[-x^z e^{-x}\right]_0^\infty + \int_0^\infty z x^{z-1} e^{-x}\, dx \\ &= 0 - 0 + z \int_0^\infty x^{z-1} e^{-x}\, dx \\ &= z\Gamma(z). \end{aligned}$$

2. $$\Gamma(1) = \int_0^\infty e^{-x}\, dx = \left[-e^{-x}\right]_0^\infty = 1$$

3. On fait le changement de variable $u = \sqrt{x}$:

$$\begin{aligned} \Gamma\left(\frac{1}{2}\right) &= \int_0^\infty \frac{1}{\sqrt{x}} e^{-x}\, dx \\ &= \frac{1}{u} e^{-u^2}\, 2u\, du \\ &= 2\int_0^\infty e^{-u^2}\, du \\ &= 2\frac{\sqrt{\pi}}{2} = \sqrt{\pi} \end{aligned}$$

4. On applique les propriétés 1 et 2 :

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = \ldots = (n-1)(n-2)\ldots 2 \cdot 1 \Gamma(1) = (n-1)!$$

5. On se contente de prouver le résultat pour les entiers naturels, et à une constante près. On veut alors montrer que

$$\exists C \in \mathbb{R}_+^*,\ n! \sim C n^{n+\frac{1}{2}} e^{-n}$$

quand $n \to \infty$. On introduit

$$u_n = \frac{n^{n+\frac{1}{2}} e^{-n}}{n!}$$

et on va montrer que la suite (strictement positive) (u_n) converge vers une limite non nulle. Soit

$$\begin{aligned} v_n &= \ln \frac{u_{n+1}}{u_n} \\ &= \ln \left(\frac{(n+1)^{n+\frac{3}{2}} e^{-n+1}}{(n+1)!} \cdot \frac{n!}{n^{n+\frac{1}{2}} e^{-n}} \right) \\ &= \ln \left(\frac{(n+1)^{n+\frac{1}{2}}}{n^{n+\frac{1}{2}}} e^{-1} \right) \\ &= \left(n + \frac{1}{2} \right) \ln \left(1 + \frac{1}{n} \right) - 1. \end{aligned}$$

Le développement limité de la fonction $x \mapsto \ln(1+x)$ en 0 est

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + o(x^3)$$

donc

$$\begin{aligned} v_n &= \left(n + \frac{1}{2} \right) \left(\frac{1}{n} - \frac{1}{2n^2} + \frac{1}{3n^3} + o\left(\frac{1}{n^3} \right) \right) - 1 \\ &= 1 - \frac{1}{2n} + \frac{1}{3n^2} + \frac{1}{2n} - \frac{1}{4n^2} + o\left(\frac{1}{n^2} \right) - 1 \\ &= \frac{1}{12n^2} + o\left(\frac{1}{n^2} \right). \end{aligned}$$

Donc la série $\sum v_n$ est convergente. Or on peut écrire $v_n = \ln(u_n + 1) - \ln(u_n)$: on a donc une somme télescopique, donc la suite des $(\ln u_n)$ converge vers une limite L, donc la suite (u_n) converge vers une limite strictement positive e^L. Il suffit alors de poser $C = e^L$ et on a le résultat cherché.

1.4 Lois de probabilité classiques

On rappelle ici quelques lois de probabilités utiles.

Loi normale La loi normale $\mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma > 0$ est la loi réelle de densité

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Son espérance vaut μ et sa variance σ^2.
Fonctions sous R : `dnorm, pnorm, qnorm, rnorm`.

Loi exponentielle La loi exponentielle $\mathcal{E}(\lambda)$ avec $\lambda > 0$ est la loi à valeurs dans $\mathbb{R}_+$ de densité

$$f(x) = \lambda e^{-\lambda x}.$$

Son espérance vaut $\frac{1}{\lambda}$ et sa variance $\frac{1}{\lambda^2}$.
Fonctions sous R : `dexp, pexp, qexp, rexp`.

Loi de Cauchy La loi de Cauchy est la loi réelle de densité

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

Son espérance et sa variance ne sont pas définies.
Fonctions sous R : `dcauchy, pcauchy, qcauchy, rcauchy`.

Loi Gamma La loi Gamma $\Gamma(\alpha, \beta)$ avec $\alpha > 0$ et $\beta > 0$ est la loi à valeurs dans $\mathbb{R}_+$ de densité

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Son espérance vaut $\frac{\alpha}{\beta}$ et sa variance $\frac{\alpha}{\beta^2}$.
Fonctions sous R : `dgamma, pgamma, qgamma, rgamma`.
Attention à la notation : quand on met deux paramètres ($\Gamma(\alpha, \beta)$) cela veut dire la loi de probabilité. Quand on met un seul paramètre ($\Gamma(\alpha)$) cela veut dire la fonction Gamma définie dans la section 1.3.

Loi Inverse Gamma Si $X \sim \Gamma(\alpha, \beta)$ et $Y = \frac{1}{X}$, alors Y suit la loi Inverse Gamma $IG(\alpha, \beta)$, à valeurs dans $\mathbb{R}_+$ et de densité

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} e^{-\frac{\beta}{y}}.$$

Son espérance vaut $\frac{\beta}{\alpha-1}$ si $\alpha > 1$ et sa variance vaut $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ si $\alpha > 2$ (et $+\infty$ sinon).

Loi Beta La loi $Beta(\alpha, \beta)$ est la loi à valeurs dans $]0, 1[$ de densité

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

Son espérance vaut $\frac{\alpha}{\alpha+\beta}$ et sa variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
Fonctions sous R : `dbeta, pbeta, qbeta, rbeta`.

Loi uniforme continue La loi uniforme sur $[0, 1]$ notée $U([0, 1])$ est la loi à valeurs dans $[0, 1]$ de densité

$$f(x) = \mathbb{I}_{\{0 \leq x \leq 1\}}.$$

Son espérance vaut $\frac{1}{2}$ et sa variance $\frac{1}{12}$. C'est la même loi que la loi $Beta(1, 1)$.
Fonctions sous R : `dunif, punif, qunif, runif`.

Loi de Bernoulli La loi de Bernoulli de paramètre $p \in [0, 1]$ est la loi à valeurs dans $\{0, 1\}$ telle que

$$X \sim Bernoulli(p) \qquad \text{ssi} \qquad P[X = 1] = p \quad \text{et} \quad P[X = 0] = 1 - p.$$

Son espérance vaut p et sa variance $p(1-p)$.
Fonctions sous R : `dbinom, pbinom, qbinom, rbinom` avec le paramètre `n=1`.

Loi géométrique La loi géométrique de paramètre $p \in [0,1]$ est la loi à valeurs dans $\mathbb{N}$ telle que

$$X \sim Geom(p) \qquad \text{ssi} \qquad \forall k \in \mathbb{N}, P[X=k] = (1-p)^k p.$$

Son espérance vaut $\frac{1-p}{p}$ et sa variance $\frac{1-p}{p^2}$.
Attention : certains auteurs et praticiens définissent cette loi sur $\mathbb{N}^*$, ce qui revient à poser $Y = X+1$ dans la définition.
Fonctions sous R : `dgeom, pgeom, qgeom, rgeom`.

Loi de Poisson La loi de Poisson de paramètre $\lambda \in \mathbb{R}_+^*$ est la loi à valeurs dans $\mathbb{N}$ telle que

$$X \sim Poisson(\lambda) \qquad \text{ssi} \quad \forall k \in \mathbb{N}, P[X=k] = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Son espérance et sa variance valent λ.
Fonctions sous R : `dpois, ppois, qpois, rpois`.

Loi Négative binomiale La loi négative binomiale $NB(r,p)$ avec $r \in \mathbb{N}$ et $p \in]0,1[$ est la loi à valeurs dans $\mathbb{N}$ telle que

$$X \sim NB(r,p) \qquad \text{ssi} \qquad \forall k \in \mathbb{N}, P[X=k] = \binom{k+r-1}{k} \cdot (1-p)^r p^k$$

avec un coefficient binomial dans la définition. Son espérance vaut $\frac{pr}{1-p}$ et sa variance $\frac{pr}{(1-p)^2}$.
Fonctions sous R : `dnbinom, pnbinom, qnbinom, rnbinom`.

1.5 Lois marginale et conditionnelle

Définition 4 *Soient X et Y deux variables aléatoires continues. On note $f_{XY}(x,y)$ leur loi jointe. Alors la loi marginale de X est la loi de probabilité de densité*

$$f_X(x) = \int f_{XY}(x,y)\, dy.$$

On définit de même la loi marginale de Y.

Notons que si X et Y sont indépendantes, alors $\forall x, y, f_{XY}(x,y) = f_X(x) f_Y(y)$.

Définition 5 *Soient X et Y deux variables aléatoires continues. On note $f_X(x)$ la densité de la loi marginale de X et $f_{XY}(x,y)$ leur densité jointe. Alors la loi conditionnelle de Y sachant que $X = x$ est la loi de probabilité de densité*

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

dès lors que cette fonction est bien définie.

Notons que X et Y sont indépendantes ssi pour tout x, la loi conditionnelle de Y sachant que $X = x$ est égale à la loi marginale de Y.

À partir de cette définition, on peut définir l'espérance conditionnelle de Y sachant X :

$$E[Y|X=x] = \int y f_{Y|X}(y|x)\, dy.$$

On peut définir la loi conditionnelle et l'espérance conditionnelle de façon plus formelle, mais ces définitions sont suffisantes pour le cadre de ce cours.

2 Inférence bayésienne : notions fondamentales

On se place dans le cadre d'une expérience statistique paramétrique : on a $\mathbf{X} = (X_1, \ldots, X_n)$ iid qui proviennent d'une loi $\mathcal{P}_\theta$ avec θ inconnu ; par exemple, $X_i \sim \mathcal{N}(\theta, 1)$.

Dans le paradigme bayésien, on suppose que θ est une variable aléatoire, dont la loi de probabilité représente notre incertitude sur les valeurs possibles. Avant d'observer des données (*a priori*), on a une forte incertitude, donc une loi avec une variance élevée. Plus on observe de réalisations de $\mathcal{P}_\theta$ (*a posteriori*), plus on aura une idée précise des valeurs plausibles pour θ : la loi aura une variance de plus en plus faible.

N'importe quelle loi de probabilité peut faire office de loi a priori ; certains choix sont plus naturels que d'autres, comme nous le verrons par la suite. L'inférence bayésienne permet de formaliser ce passage de la loi a priori à la loi a posteriori. Dans tout ce qui suit, on suppose que toutes les lois considérées admettent une densité par rapport à une mesure de référence – la mesure de Lebesgue pour les variables continues ou la mesure de comptage pour les lois discrètes.

2.1 Loi a posteriori

Définition 6 (Loi a posteriori) *Soit $X_1, \ldots, X_n$ un vecteur de variables aléatoires iid de loi $\mathcal{P}_\theta$ et soit $\pi(\theta)$ une loi a priori pour θ. On appelle* loi a posteriori *de θ sachant $X_1, \ldots, X_n$ la loi de densité*

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta) L(\theta|x_1, \ldots, x_n).$$

L'origine de cette définition est la règle de Bayes. Puisque θ et $\mathbf{X}$ sont des variables aléatoires, on peut écrire

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta) \cdot f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{x})}.$$

Dans cette écriture, $f_{X|\theta}(x|\theta) = L(\theta|x)$ par définition de la vraisemblance, et $f_X(x)$ est une constante de normalisation (elle ne dépend pas de θ).

Lorsqu'on cherche la forme analytique de la loi a posteriori, on ne va pas s'encombrer des constantes : une fois qu'on a trouvé la densité à une constante multiplicative près, cette constante est unique puisque la densité doit s'intégrer à 1.

Exemple : loi $\mathcal{N}(\theta, 1)$ On prend les $X_i \sim \mathcal{N}(\theta, 1)$ iid et on choisit la loi a priori $\theta \sim \mathcal{N}(0, 1)$. Calculons la loi a posteriori

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto \pi(\theta)L(\theta|x) \\
&\propto \frac{1}{\sqrt{2\pi}}e^{-\frac{\theta^2}{2}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\theta-x_i)^2} \\
&\propto \exp\left(-\frac{\theta^2}{2} - \frac{1}{2}\sum_i (\theta - x_i)^2\right) \\
&\propto \exp\left(-\frac{n+1}{2}\theta^2 + \theta\sum_i x_i - \frac{1}{2}\sum_i x_i^2\right) \\
&\propto \exp\left(-\frac{n+1}{2}\theta^2 + \theta n\bar{x}\right) \\
&\propto \exp\left(-\frac{n+1}{2}\left(\theta - \frac{n\bar{x}}{n+1}\right)^2\right)
\end{aligned}$$

et on reconnaît que

$$\theta|\mathbf{x} \sim \mathcal{N}\left(\frac{n\bar{x}}{n+1}, \frac{1}{n+1}\right).$$

Dans cet exemple (et dans le reste de ces notes de cours), on utilise le symbole $\propto$, qui signifie "proportionnel à" : les termes sont égaux à une constante multiplicative près. Comme on s'intéresse à la loi de θ, toute quantité qui ne dépend pas de θ peut être considérée comme une constante.

2.2 Estimation bayésienne

L'intérêt du paradigme bayésien est de rendre l'incertitude explicite : la loi a posteriori représente notre croyance sur le paramètre θ. On peut vouloir résumer cette loi.

Il existe plusieurs estimateurs ponctuels possibles. Les plus usités sont :

— l'espérance a posteriori $E[\theta|\mathbf{x}]$;

— le maximum a posteriori $\arg\max_\theta \pi(\theta|\mathbf{x})$.

— la médiane a posteriori

mais il serait dommage de se cantonner à un estimateur ponctuel : l'intérêt du paradigme bayésien réside dans la distribution a posteriori prise dans son entier.

2.2.1 Intervalle de crédibilité

Définition 7 *Soit θ un paramètre à estimer, $\pi(\theta|X)$ sa loi a posteriori, et $\alpha \in]0, 1[$. On appelle* intervalle de crédibilité *de niveau α tout intervalle $[a, b]$ vérifiant*

$$P[a < \theta < b|X] \geq \alpha.$$

Notes La définition est très semblable à celle d'un intervalle de confiance en statistique fréquentiste. La différence est que dans le cas fréquentiste, le paramètre est fixe et les bornes sont aléatoires ; dans le cas bayésien, le paramètre est aléatoire et les bornes sont fixes.

On définit de même la notion d'*ensemble de crédibilité*.

Définition 8 (Intervalle HPD) *On dit qu'un α-intervalle de crédibilité est* HPD *(pour* Highest Posterior Distribution*) s'il est de longueur minimale parmi tous les α-intervalles de crédibilité.*

Note On définit de même un *ensemble HPD*. Même dans le cas univarié, un ensemble HPD n'est pas nécessairement un intervalle, notamment si la loi a posteriori est multimodale.

2.3 Pourquoi faire de l'inférence bayésienne ?

Le paradigme bayésien peut sembler plus complexe à mettre en œuvre que les méthodes dites "classiques", et notamment que le maximum de vraisemblance ; pour autant, les méthodes bayésiennes sont très populaires dans de nombreux domaines d'application. On peut citer plusieurs explications :

- la mesure de l'incertitude est aisée et explicite, puisqu'on obtient une distribution pour le paramètre à estimer θ et jamais une simple estimation ponctuelle ;
- les méthodes bayésiennes peuvent être plus aisées à appliquer sur des modèles complexes ;
- le paradigme bayésien est facile à utiliser pour des modèles ou expériences imbriqués : la loi a posteriori de l'expérience 1 peut être utilisée comme loi a priori de l'expérience 2 ;
- l'étude de fonctions complexes des paramètres est facilitée (par exemple : lois marginales, lois conditionnelles...) ;
- l'étude de plusieurs modèles concurrents est plus aisée ;
- certains mettent en avant des raisons philosophiques ; notons également que les estimateurs bayésiens peuvent être plus intuitifs que les notions d'intervalle de confiance ou de p-valeur, qui sont mal comprises des non-mathématiciens ;
- le cadre mathématique est parfois plus simple à utiliser.

Pour autant, il existe bien entendu un grand nombre de situations où l'estimateur du maximum de vraisemblance est utilisé en pratique, notamment pour des modèles suffisamment simples pour que l'EMV soit calculable et avec des données de taille suffisamment grande pour que les propriétés asymptotiques de l'EMV entrent en action.

2.4 Loi a priori

Le choix de la loi a priori est capital, surtout lorsque la quantité de données qu'on manipule n'est pas très grande. La loi a priori doit représenter la croyance du statisticien avant le début de l'expérience. Cela n'est pas toujours évident. Voici quelques techniques usuelles pour proposer une loi a priori raisonnable :

- une loi conjuguée (voir ci-dessous), ce qui présente l'avantage de faciliter les calculs mathématiques
- la loi "non-informative" ou a priori de Jeffreys (voir ci-dessous)
- une loi a priori fournie par un expert du domaine d'application : selon les cas, l'expert peut soit fournir une loi, soit au moins les premiers moments (espérance, variance) ce qui permet de choisir les paramètres d'une loi conjuguée ;
- une loi a priori qui provient d'une expérience précédente : en cas d'expériences successives, la loi a posteriori de l'expérience k peut servir de loi a priori pour l'expérience $k+1$;
- dans tous les cas, et encore plus quand le choix de la loi a priori ne s'impose pas, il faut vérifier l'influence de la loi a priori sur la loi a posteriori : on effectue l'analyse avec plusieurs lois a priori différentes, et on vérifie que les loi a posteriori concordent. Cela revient à dire qu'on se place dans la peau de plusieurs experts différents, avec chacun leur propre a priori, et on voit si les données permettent de les mettre d'accord.

2.4.1 Loi conjuguée

Définition 9 *On se place dans le cadre d'un n-échantillon $X_1, \dots, X_n \sim \mathcal{P}_\theta$ avec θ à estimer. Soit L la fonction de vraisemblance associée. On dit qu'une famille de lois de probabilités $\mathcal{F}$ est*

TABLE 2.1 – Quelques lois conjuguées

Modèle	Paramètre	Loi a priori	Loi a posteriori
Bernoulli	p	$Beta(\alpha, \beta)$	$Beta(\alpha + \sum x_i, \beta + n - \sum x_i)$
Poisson	λ	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + \sum x_i, \beta + n)$
Géométrique	p	$Beta(\alpha, \beta)$	$Beta(\alpha + n, \beta + \sum x_i - n)$
Normal	espérance μ (variance $\sigma^2 = \frac{1}{\tau}$ connue)	$\mathcal{N}(\mu_0, \frac{1}{\tau_0})$	$\mathcal{N}\left(\frac{\tau_0\mu_0 + \tau\sum x_i}{\tau_0 + n\tau}, \tau_0 + n\tau\right)$
Normal	précision $\tau = \frac{1}{\sigma^2}$ (espérance μ connue)	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + \frac{n}{2}, \beta + \frac{\sum(x_i - \mu)^2}{2})$
Exponentiel	λ	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + n, \beta + \sum x_i)$

conjuguée *à la vraisemblance L si pour toute loi a priori $\pi \in \mathcal{F}$, pour tout n et pour toute réalisation de $X_1, \ldots, X_n$, la loi a posteriori associée est également un élément de $\mathcal{F}$.*

Exemple On considère le modèle de Bernoulli : $X_1, \ldots, X_n \sim Bernoulli(p)$, p à estimer. On prend comme famille $\mathcal{F}$ la famille des loi $Beta(a, b)$ avec $a > 0$ et $b > 0$. Alors si on prend la loi a priori $p \sim Beta(a, b)$, on peut écrire en ignorant les constantes :

$$\pi(p) \propto p^{a-1}(1-p)^{b-1}.$$

La vraisemblance de p au vu de $X_1, \ldots, X_n$ s'écrit

$$L(p; X_1, \ldots, X_n) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum X_i}(1-p)^{n-\sum X_i}$$

et donc la loi a posteriori vérifie

$$\pi(p|X_1, \ldots, X_n) \propto p^{a-1+\sum X_i}(1-p)^{n-\sum X_i}$$

autrement dit

$$p|X_1, \ldots, X_n \sim Beta\left(a + \sum X_i, b + n - \sum X_i\right)$$

et la loi a posteriori est bien un élément de la famille $\mathcal{F}$. Donc la famille des lois $Beta(a, b)$ est conjuguée pour ce modèle.

2.4.2 Prior de Jeffreys

Définition 10 (Information de Fisher) *Soit $\mathcal{P}_\theta$ un modèle statistique indexé par un paramètre scalaire θ, et ℓ la fonction de log-vraisemblance associée. On appelle* information de Fisher *la fonction*

$$\mathcal{I} : \theta \mapsto E\left[\left(\frac{\partial}{\partial\theta}\ell(\theta; X)\right)^2 \middle| X \sim \mathcal{P}_\theta\right].$$

Théorème 2 *Si la vraisemblance est de classe $\mathcal{C}^2$, et sous des hypothèses de régularité aisément vérifiées en pratique, on peut écrire*

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta; X) \middle| X \sim \mathcal{P}_\theta\right].$$

Note En pratique, c'est toujours cette seconde expression qu'on utilise. L'information de Fisher capture la quantité d'information apportée par une observation pour estimer θ.

Preuve Soit $f(X;\theta)$ la densité associée à $\mathcal{P}_\theta$ et L la fonction de vraisemblance ; on a $\ell(\theta;X) = \log L(\theta;X) = \log f(X;\theta)$. On a

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ell(\theta;X) &= \frac{\partial}{\partial\theta}\log L(\theta;X) = \frac{\frac{\partial}{\partial\theta}L(\theta;X)}{L(\theta;X)} \\
\frac{\partial^2}{\partial\theta^2}\ell(\theta;X) &= \frac{\partial^2}{\partial\theta^2}\log L(\theta;X) = \frac{\frac{\partial^2}{\partial\theta^2}L(\theta;X)}{L(\theta;X)} - \left(\frac{\frac{\partial}{\partial\theta}L(\theta;X)}{L(\theta;X)}\right)^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2}L(\theta;X)}{L(\theta;X)} - \left(\frac{\partial}{\partial\theta}\log L(\theta;X)\right)^2. \\
\text{Or } E\left[\left.\frac{\frac{\partial^2}{\partial\theta^2}L(\theta;X)}{L(\theta;X)}\right| X\sim\mathcal{P}_\theta\right] &= \int\frac{\frac{\partial^2}{\partial\theta^2}L(\theta;x)}{L(\theta;x)}f(x;\theta)\,d\theta = \int\frac{\partial^2}{\partial\theta^2}L(\theta;x)\,dx \\
&= \frac{\partial^2}{\partial\theta^2}\int L(\theta;x)\,dx = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\,dx = \frac{\partial^2}{\partial\theta^2}1 = 0.
\end{aligned}$$

d'où le résultat.

Exemple Calculons l'information de Fisher pour la famille de Bernoulli. Rappelons que si $X \sim Bernoulli(p)$ alors $E[X] = p$.

$$\begin{aligned}
L(p;X) &= p^X(1-p)^{1-X} \\
\ell(p;X) &= X\log p + (1-X)\log(1-p) \\
\frac{\partial}{\partial p}\ell(p;X) &= \frac{X}{p} - \frac{1-X}{1-p} \\
\frac{\partial^2}{\partial p^2}\ell(p;X) &= -\frac{X}{p^2} - \frac{1-X}{(1-p)^2} \\
\mathcal{I}(p) &= -E\left[\frac{\partial^2}{\partial p^2}\ell(p;X)\right] = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}.
\end{aligned}$$

Définition 11 *Soit $\mathcal{P}_\theta$ un modèle avec un paramètre θ à estimer par inférence bayésienne. On appelle* loi a priori de Jeffreys *la loi définie par*

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

Invariance par reparamétrisation. La plupart des choix de loi a priori sont sensibles à la paramétrisation choisie pour le modèle. Par exemple, il peut paraître raisonnable de choisir une loi $U(0,1)$ comme loi a priori pour le paramètre d'une expérience $Bernoulli(p)$. Mais si on pose $\rho = \frac{1}{p}$, alors la loi a priori induite sur ρ n'a rien de naturel.

La loi a priori de Jeffreys est dite non-informative, parce qu'elle ne dépend pas de la paramétrisation choisie pour la loi. En effet, soit ϕ une paramétrisation alternative. Calculons la

loi a priori induite sur ϕ par la loi a priori de Jeffreys sur θ :

$$\begin{aligned}
\pi(\phi) &= \pi(\theta)\left|\frac{d\theta}{d\phi}\right| \\
&\propto \sqrt{\mathcal{I}(\theta)\left(\frac{d\theta}{d\phi}\right)^2} \\
&= \sqrt{E\left[\left(\frac{d\ell}{d\theta}\right)^2\right]\left(\frac{d\theta}{d\phi}\right)^2} \\
&= \sqrt{E\left[\left(\frac{d\ell}{d\theta}\frac{d\theta}{d\phi}\right)^2\right]} \\
&= \sqrt{E\left[\left(\frac{d\ell}{d\phi}\right)^2\right]} \\
&= \sqrt{\mathcal{I}(\phi)}.
\end{aligned}$$

On retrouve bien la loi de Jeffreys pour ϕ.

Pour cette raison, la loi de Jeffreys est souvent prise comme loi par défaut quand on n'a pas de véritable information a priori.

Exemple Reprenons l'exemple précédent, avec $\phi = 1/\theta$. On a donc $X \sim Bernoulli(\frac{1}{\phi})$ et $E[X] = \frac{1}{\phi}$. Cherchons la loi a priori de Jeffreys pour ϕ.

$$\begin{aligned}
L(\phi) &= \left(\frac{1}{\phi}\right)^X \left(1-\frac{1}{\phi}\right)^{1-X} \\
\ell(\phi) &= -X\ln\phi + (1-X)\ln\left(1-\frac{1}{\phi}\right) = -X\ln\phi + (1-X)\left[\ln(\phi-1) - \ln(\phi)\right] \\
&= (1-X)\ln(\phi-1) - \ln(\phi) \\
\frac{\partial\ell}{\partial\phi} &= \frac{1-X}{\phi-1} - \frac{1}{\phi} \\
\frac{\partial^2\ell}{\partial\phi^2} &= -\frac{1-X}{(\phi-1)^2} + \frac{1}{\phi^2} \\
\mathcal{I}(\phi) &= -E\left[\left.\frac{\partial^2\ell}{\partial\phi^2}\right| X\right] = \frac{1-\frac{1}{\phi}}{(\phi-1)^2} - \frac{1}{\phi^2} = \frac{\frac{\phi-1}{\phi}}{(\phi-1)^2} - \frac{1}{\phi^2} = \frac{1}{\phi(\phi-1)} - \frac{1}{\phi^2} = \frac{1}{\phi^2(\phi-1)} \\
\pi(\phi) &\propto \sqrt{\mathcal{I}(\phi)} = \frac{1}{\phi}\sqrt{\frac{1}{\phi-1}} \text{ et on peut en déduire la loi a priori de } \theta \\
\frac{d\phi}{d\theta} &= -\frac{1}{\theta^2} \\
\pi(\theta) &\propto \frac{1}{\theta^2}\frac{1}{\phi}\sqrt{\frac{1}{\phi-1}} = \frac{1}{\theta}\sqrt{\frac{1}{\frac{1}{\theta}-1}} = \frac{1}{\theta}\sqrt{\frac{\theta}{1-\theta}} = \sqrt{\frac{1}{\theta(1-\theta)}}
\end{aligned}$$

et on retrouve bien la même loi a priori que précédemment : le choix de la paramétrisation n'a donc pas impacté notre loi a priori.

Attention ! Rien ne garantit que l'information de Fisher soit intégrable : si elle ne l'est pas, alors il n'existe pas de loi de probabilité qui lui soit proportionnelle. Dans ce cas, on dit qu'il s'agit d'une loi *impropre*. On peut tout de même définir la loi a posteriori associée, mais il est essentiel de vérifier que la loi a posteriori est, elle, bien définie : rien ne l'assure.

Cas multivarié Si le paramètre θ est un vecteur de longueur k, alors l'information de Fisher est une matrice symétrique $k \times k$, dont l'entrée (i, j) vaut $-E[\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}]$. La loi a priori de Jeffreys est alors définie par

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

2.5 Propriétés asymptotiques

Les estimateurs bayésiens possèdent de bonnes propriétés même du point de vue fréquentiste : si les observations proviennent d'une loi $\mathcal{P}_{\theta_0}$, alors la loi a posteriori se concentre autour de θ_0. C'est ce qu'explicite le théorème suivant.

Définition 12 (Distance L^1) *Soient P et Q deux lois de probabilité, de densités respectives p et q par rapport à une mesure de référence μ. On appelle* distance L^1 *entre P et Q la quantité*

$$\|P - Q\|_1 = \int |p - q| \, d\mu.$$

Théorème 3 (Bernstein-von Mises) *Soit $(\mathcal{P}_\theta)_{\theta\in\Theta}$ un modèle régulier, et soit $\theta_0 \in \Theta$. Soient des observations $X_1, \ldots, X_n \sim \mathcal{P}_{\theta_0}$ iid. On suppose que la loi a priori Π a une densité π qui vérifie $\pi(\theta_0) > 0$ et π continue au point θ_0. On suppose par ailleurs que l'information de Fisher $\mathcal{I}(\theta_0)$ est inversible au point θ_0. On note $\hat{\theta}_n^{MV}$ l'estimateur au maximum de vraisemblance. Alors on a la convergence de la loi a posteriori :*

$$\left\| \Pi[\cdot|X_1, \ldots, X_n] - \mathcal{N}\left(\hat{\theta}_n^{MV}, \frac{\mathcal{I}(\theta_0)^{-1}}{n}\right)(\cdot) \right\|_1 \xrightarrow[n\to\infty]{P_{\theta_0}} 0.$$

Autrement dit, lorsque n est suffisamment élevé, la loi a posteriori ressemble à une loi gaussienne centrée en l'estimateur du maximum de vraisemblance, avec une variance qui décroît à la vitesse $\frac{1}{n}$.

Ce théorème est le pendant bayésien du Théorème Central Limite.

2.6 Choix de modèle : facteur de Bayes

Considérons maintenant la question du choix de modèle.

On considère deux modèles M_0 et M_1, qu'on souhaite comparer. Le modèle M_i dépend d'un paramètre θ_i, pour lequel on a une loi a priori $\pi_i(\theta_i)$ et une fonction de vraisemblance L_i. Enfin, on choisit la probabilité qu'on attribue a priori à chacun des deux modèles (par exemple 1/2 et 1/2). A priori, le rapport des chances des deux modèles est $P[M_0]/P[M_1]$; a posteriori, ce rapport des chances est $P[M_0|\mathbf{x}]/P[M_1|\mathbf{x}]$.

Définition 13 (facteur de Bayes) *On appelle facteur de Bayes la quantité*

$$B_{01}^{\pi}(\mathbf{x}) = \frac{P[M_0|\mathbf{x}]}{P[M_1|\mathbf{x}]} \Big/ \frac{P[M_0]}{P[M_1]} \, .$$

On utilise le facteur de Bayes pour mesurer comment les données ont fait évoluer le rapport des chances. Intuitivement, si $B^\pi_{01}(\mathbf{x})$ est grand, alors les données $\mathbf{x}$ favorisent M_0 ; si $B^\pi_{01}(\mathbf{x})$ est petit, les données favorisent M_1. Le facteur de Bayes mesure à quel point les données ont fait évoluer notre croyance, entre le ratio des probabilités a priori et le ratio a posteriori.

Par le théorème de Bayes, on a

$$P[M_i|\mathbf{x}] = \frac{P[\mathbf{x}|M_i]P[M_i]}{P[\mathbf{x}]}$$

et on peut récrire le facteur de Bayes :

$$B^\pi_{01} = \frac{m_0(\mathbf{x})}{m_1(\mathbf{x})} = \frac{\int \pi_0(\theta_0)L_0(\theta_0;\mathbf{x})\,d\theta_0}{\int \pi_1(\theta_1)L_1(\theta_1;\mathbf{x})\,d\theta_1}$$

où

$$m_i(\mathbf{x}) = E_{\pi_i}\left[L_i(\theta_i;\mathbf{x})\right] = \int \pi_i(\theta_i)L_i(\theta_i;\mathbf{x})\,d\theta_i$$

n'est autre que la constante de normalisation de la formule $\pi_i(\theta_i|\mathbf{x}) \propto \pi_i(\theta_i)L_i(\theta_i;\mathbf{x})$.

Remarques :

— alors que dans le cadre classique du test d'hypothèses, on accorde une place privilégiée à l'hypothèse nulle, ici les deux hypothèses jouent des rôles symétriques ;

— cette dernière expression ressemble au ratio de vraisemblances utilisé dans le cadre classique, dans lequel on aurait remplacé arg max par une intégrale. Cette modification permet d'avoir une pénalisation naturelle de la taille du modèle, et d'éviter ainsi les problèmes de surapprentissage.

De même qu'il existe des échelles pour interpréter les p-valeurs, on cite souvent l'échelle de Jeffreys pour interpréter le facteur de Bayes. Si $\log_{10} B^\pi_{01}(\mathbf{x}) > 0$, alors les données favorisent le modèle M_0 ; plus précisément :

— entre 0 et $\frac{1}{2}$, l'évidence est *faible* ;

— entre $\frac{1}{2}$ et 1, elle est *substantielle* ;

— entre 1 et 2, elle est *forte*

— au-dessus de 2, elle est *décisive*

et symétriquement en faveur du modèle M_1 pour les valeurs négatives.

Exemple : modèle Poisson-Gamma On dispose d'observations $X_1, \dots, X_n$ de loi de Poisson, mais également de covariables $Z_1, \dots, Z_n$ avec $Z_i \in \{1, 2\}$ et on souhaite choisir entre les deux modèles suivants, avec des lois a priori $Gamma(a, b)$:

$$\mathcal{M}_0 : \quad X_i \sim \mathcal{P}(\lambda) \qquad \lambda \sim \Gamma(a, b)$$
$$\mathcal{M}_1 : \quad X_i|Z_i = k \sim \mathcal{P}(\lambda_k) \qquad \lambda_1, \lambda_2 \sim \Gamma(a, b)$$

autrement dit, on cherche à savoir si les (X_i) proviennent tous de la même distribution, ou s'ils proviennent de deux distributions différentes selon la valeur des (Z_i).

Pour choisir entre ces deux modèles, on calcule les vraisemblances marginales (avec les notations $n_1 = \sum \mathbb{I}_{Z_i=1}$, $S_1 = \sum x_i \mathbb{I}_{Z_i=1}$, et les équivalents pour n_2 et S_2) :

$$m_0(\mathbf{x}) = E_{\pi_0}\left[L_0(\mathbf{x}|\theta_0\right] = \int_0^\infty \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}e^{-n\lambda}\lambda^{\sum x_i}\prod \frac{1}{x_i!}\,d\lambda = \frac{b^a}{\Gamma(a)}\prod \frac{1}{x_i!}\frac{\Gamma(a+\sum x_i)}{(b+n)^{a+\sum x_i}};$$

$$m_1(\mathbf{x}) = \frac{b^{2a}}{\Gamma(a)^2}\frac{\Gamma(a+S_1)}{(b+n_1)^{a+S_1}}\frac{\Gamma(a+S_2)}{(b+n_2)^{a+S_2}}\prod \frac{1}{x_i!}$$

d'où le facteur de Bayes

$$B_{01}(\mathbf{x}) = \frac{\Gamma(a)}{b^a}\frac{\Gamma(a+S_1+S_2)}{\Gamma(a+S_1)\Gamma(a+S_2)}\frac{(b+n_1)^{a+S_1}(b+n_2)^{a+S_2}}{(b+n)^{a+S_1+S_2}}.$$

3 Monte-Carlo

Nous présentons les algorithmes de Monte-Carlo dans le contexte d'une application à l'inférence bayésienne. Ces algorithmes sont également couramment utilisés dans d'autres domaines d'application, et le contenu de ce chapitre s'applique tout aussi bien à ces autres domaines.

Dans le cadre bayésien, beaucoup de quantités d'intérêt peuvent s'écrire sous la forme

$$I = E_p[h(\theta)] = \int h(\theta)p(\theta)\,d\theta.$$

Par exemple, l'espérance a posteriori correspond à $h(\theta) = \theta$ et p la loi a posteriori ; la vraisemblance marginale correspond à $h(\theta) = L(\theta; \mathbf{x})$ et p la loi a priori...

Dans les exemples simples considérés jusqu'à présent, ces espérances peuvent être calculables sous forme analytique, mais cela n'est bien sûr pas le cas en général. On va alors chercher un estimateur de I. C'est l'objet des méthodes de Monte-Carlo.

3.1 Monte-Carlo standard

Définition 14 (Estimateur de Monte-Carlo) *Soit $I = E_p[h(\theta)]$. On appelle* estimateur de Monte-Carlo *l'estimateur $\hat{I}_T^{MC}$ obtenu par l'algorithme suivant : on simule $\theta_1, \ldots, \theta_T \sim p$ iid et on pose*

$$\hat{I}_T^{MC} = \frac{1}{T}\sum_{t=1}^{T} h(\theta_t).$$

Notons que $\hat{I}_T^{MC}$ est un estimateur sans biais de I, et que sous des conditions très générales la Loi des Grands Nombres garantit qu'il est convergent quand $T \to \infty$. Enfin, on a

$$Var(\hat{I}_T^{MC}) = \frac{1}{T}Var_p(h(\theta)).$$

3.2 Échantillonnage préférentiel

Cette approche simple est rarement optimale : d'une part $Var_p(h(\theta))$ peut être élevée ; d'autre part (et c'est plus gênant) il peut être difficile ou impossible de générer les θ_t selon p. Il est alors plus efficace de procéder par *échantillonnage préférentiel* (*importance sampling*) : on peut récrire, pour une densité γ bien choisie,

$$I = \int \frac{h(\theta)p(\theta)}{\gamma(\theta)}\gamma(\theta)\,d\theta = E_\gamma\left[\frac{h(\theta)p(\theta)}{\gamma(\theta)}\right]$$

et on reprend l'idée précédente.

Définition 15 (Échantillonnage préférentiel) *Soit $I = E_p[h(\theta)]$ et soit γ une densité de probabilité telle que $\forall x, p(x) > 0 \implies \gamma(x) > 0$. L'estimateur par* échantillonnage préférentiel *I_T^{IS} de loi instrumentale γ est donné par l'algorithme suivant : on simule $\theta'_1, \ldots, \theta'_T \sim \gamma$ et on pose*

$$\hat{I}_T^{IS} = \frac{1}{T}\sum_{t=1}^{T} \frac{h(\theta_t)p(\theta_t)}{\gamma(\theta_t)}.$$

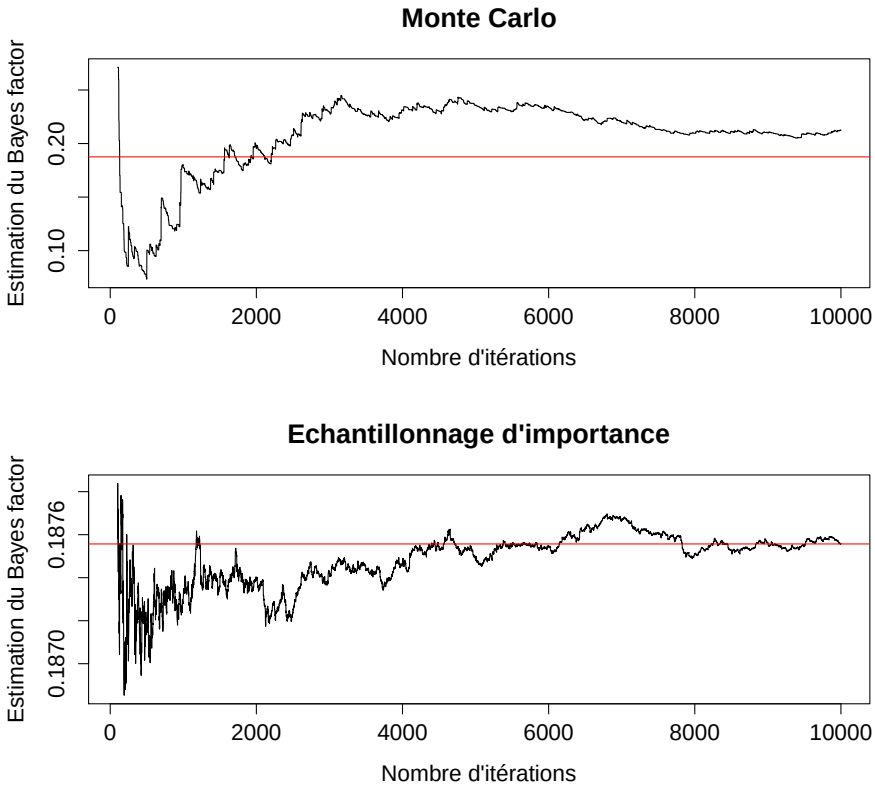

FIGURE 3.1 – Estimation d'un Bayes factor par Monte-Carlo et par échantillonage d'importance. En rouge, la valeur analytique. Notez que l'échelle des ordonnées n'est pas la même sur les deux figures.

Comment choisir γ ? Il faut bien sûr que $\gamma(\theta)$ soit non nulle dès que $p(\theta)$ est non nulle, et que γ soit facile à simuler. De plus, comme

$$Var(\hat{I}_T^{IS}) = \frac{1}{T} Var_\gamma \left(\frac{h(\theta)p(\theta)}{\gamma(\theta)} \right)$$

on va chercher à minimiser cette variance, ce qui revient à prendre $\gamma \approx hp$ (à une constante multiplicative près).

Exemple Reprenons le calcul du Bayes factor du chapitre précédent. Pour nos données ($n = 577$), le calcul analytique montre que le Bayes factor vaut environ 0.1876 ; on va chercher à l'estimer par Monte-Carlo et par échantillonnage d'importance. Pour la loi d'importance γ, on choisit la loi gaussienne de paramètres l'espérance et la variance a posteriori : il s'agit bien d'une loi aisée à simuler, et on s'attend à ce quelle amène une faible variance.

La figure 3.1 montre l'évolution des estimateurs $\hat{I}_T^{MC}$ et $\hat{I}_T^{IS}$, en fonction du nombre d'itérations T ; la ligne horizontale rouge correspond à la valeur analytique. Prenez garde à l'axe des ordonnées : en une centaine d'itérations, l'échantillonnage d'importance donne une estimation avec une erreur inférieure à 10^{-3}, précision que le Monte-Carlo n'a toujours pas atteinte en 10 000 itérations. Le coût de calcul est quasiment le même pour les deux méthodes : on a donc un gain d'efficacité considérable.

4 Régression linéaire dans le cadre bayésien

On se place dans le cadre de la régression linéaire : on souhaite expliquer les observations (y_i) par des covariables $(x^1, \ldots, x^p)$ avec le modèle

$$y_i = \beta_0 + \beta_1 x_i^1 + \ldots + \beta_p x_i^p + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \; iid.$$

On note y le vecteur des observations $(y_1, \ldots, y_n)$ et X la matrice des covariables.

Pour rappel, dans le cadre fréquentiste, on maximise la vraisemblance

$$L(\beta, \sigma^2 | y, X) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right]$$

et on a

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$\hat{\sigma}^2 = \frac{1}{n} s^2 = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta}).$$

4.1 Loi a priori de Zellner

Prenons comme loi a priori

$$\beta | \sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}),$$

où M est une matrice symétrique définie positive de taille $(k+1) \times (k+1)$, et

$$\sigma^2 | X \sim IG(a, b), \qquad a, b > 0,$$

alors on a la loi a posteriori

$$\beta | \sigma^2, y, X \sim \mathcal{N}_{k+1}\left((M + X^T X)^{-1}\{(X^T X)\hat{\beta} + M\tilde{\beta}\}, \sigma^2 (M + X^T X)^{-1}\right)$$

et

$$\sigma^2 | y, X \sim IG\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\tilde{\beta} - \hat{\beta})^T \left(M^{-1} + (X^T X)^{-1}\right)^{-1} (\tilde{\beta} - \hat{\beta})}{2}\right).$$

On voit qu'on a une loi a priori conjuguée. La loi a posteriori peut s'interpréter comme une pondération entre l'espérance a priori et l'EMV.

Le problème de l'expérimentateur est de fixer la matrice M. La loi informative de Zellner (également appelée g-prior) permet d'introduire de l'information tout en évitant l'écueil principal : la structure de corrélation.

On pose comme loi a priori

$$\begin{aligned} \beta | \sigma^2, X &\sim \mathcal{N}_{k+1}(\tilde{\beta}, g\sigma^2 (X^T X)^{-1}) \\ \sigma^2 &\sim \pi(\sigma^2 | X) \propto \sigma^{-2}. \end{aligned}$$

avec simplement le paramètre g à choisir en plus de l'espérance a priori $\tilde{\beta}$ (on peut prendre $\tilde{\beta} = 0$ si on n'a pas d'information a priori). Notons que la loi sur σ^2 est impropre, mais correspond au cas limite $IG(1, \beta)$ quand $\beta \to 0$. On récupérera bien une loi a posteriori Inverse Gamma.

Le paramètre g s'interprète comme la quantité d'information disponible dans la loi a priori par rapport à l'échantillon. Par exemple, $\frac{1}{g} = 0.5$ correspond à donner à la loi a priori le même poids qu'à 50% de l'échantillon. Le choix $g = n$ revient à donner à la loi a priori le même poids qu'à une seule observation.

Avec cette loi a priori, la loi a posteriori se simplifie en

$$\begin{aligned}
\pi(\beta, \sigma^2|y, X) &\propto f(y|\beta, \sigma^2, X)\pi(\beta, \sigma^2|X) \\
&\propto \left(\sigma^2\right)^{-(n/2+1)} \exp\left[-\frac{1}{2\sigma^2}(y - X\hat{\beta})^T(y - X\hat{\beta})\right. \\
&\quad \left. -\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})\right] \left(\sigma^2\right)^{-k/2} \\
&\quad \times \exp\left[-\frac{1}{2g\sigma^2}(\beta - \tilde{\beta})^T X^T X(\beta - \tilde{\beta})\right],
\end{aligned}$$

puisque les termes en $X^T X$ apparaissent à la fois dans la loi a priori et dans la vraisemblance. On a donc

$$\begin{aligned}
\beta|\sigma^2, y, X &\sim \mathcal{N}_{k+1}\left(\frac{g}{g+1}(\tilde{\beta}/g + \hat{\beta}), \frac{\sigma^2 g}{g+1}(X^T X)^{-1}\right) \\
\sigma^2|y, X &\sim \mathcal{IG}\left(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(g+1)}(\tilde{\beta} - \hat{\beta})^T X^T X(\tilde{\beta} - \hat{\beta})\right)
\end{aligned}$$

et marginalement, on a une loi de Student :

$$\begin{aligned}
\beta|y, X &\sim \mathcal{T}_{k+1}\left(n, \frac{g}{g+1}\left(\frac{\tilde{\beta}}{g} + \hat{\beta}\right),\right. \\
&\qquad \left. \frac{g(s^2 + (\tilde{\beta} - \hat{\beta})^T X^T X(\tilde{\beta} - \hat{\beta})/(g+1))}{n(g+1)}(X^T X)^{-1}\right).
\end{aligned}$$

4.2 Choix de modèle

Supposons qu'on veuille tester l'hypothèse H_0 que certains coefficients de régression sont nuls. Sous H_0, on peut écrire

$$y|\beta^0, \sigma^2, X_0 \overset{H_0}{\sim} \mathcal{N}_n\left(X_0\beta^0, \sigma^2 I_n\right)$$

où β^0 est un vecteur de longueur $(k + 1 - q)$ et X_0 est la matrice des covariables à laquelle on a ôté les entrées correspondant aux coefficients nuls.

Alors en adaptant la loi a priori sous H_0

$$\beta^0|X_0, \sigma^2 \sim \mathcal{N}_{k+1-q}\left(\tilde{\beta}^0, g_0\sigma^2(X_0^T X_0)^{-1}\right),$$

la vraisemblance marginale s'écrit

$$f(y|X_0, H_0) = (g+1)^{-(k+1-q)/2}\pi^{-n/2}\Gamma(n/2) \times \left[y^T y - \frac{g_0}{g_0+1}y^T X_0(X_0^T X_0)^{-1}X_0^T y - \frac{1}{g_0+1}\tilde{\beta}_0^T X_0^T X_0\tilde{\beta}_0\right]^{-n/2}$$

On peut donc écrire le facteur de Bayes sous forme fermée :

$$B_{10}^{\pi} = \frac{f(y|X,H_1)}{f(y|X_0,H_0)} = \frac{(g_0+1)^{(k+1-q)/2}}{(g+1)^{(k+1)/2}} \left[\frac{y^T y - \frac{g_0}{g_0+1} y^T X_0 (X_0^T X_0)^{-1} X_0^T y - \frac{1}{g_0+1} \tilde{\beta}_0^T X_0^T X_0 \tilde{\beta}_0}{y^T y - \frac{g}{g+1} y^T X (X^T X)^{-1} X^T y - \frac{1}{g+1} \tilde{\beta}^T X^T X \tilde{\beta}} \right]^{n/2}.$$

Ces formules n'ont rien de plaisant à prouver, mais elles ont l'avantage d'exister : puisqu'elles sont accessibles, on peut faire du choix de modèle de façon exacte dans ce cadre.

On reviendra par la suite à la question du choix de modèle lorsqu'on doit choisir parmi un grand nombre de modèles.

5 Méthodes MCMC

5.1 Introduction aux chaînes de Markov

Définition 16 *Soit $\mathcal{X}$ un espace probabilisé. On dit que la suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathcal{X}$ est une* chaîne de Markov *si pour tout n, la loi conditionnelle de X_{n+1} sachant $X_n, X_{n-1}, \ldots, X_1$ est égale à la loi conditionnelle de X_{n+1} sachant X_n.*

Une chaîne de Markov est donc un processus qui oublie le passé : si je connais l'état présent X_n, alors la connaissance du passé ne m'apporte aucune information supplémentaire sur la valeur future X_{n+1}.

Exemple Soit X_n la position d'un joueur après n tours au jeu de l'oie. Alors (X_n) est une chaîne de Markov : sachant X_n, la position X_{n+1} ne dépend pas des positions passées.

Définissons à présent quelques propriétés que peut avoir une chaîne de Markov. Les définitions sont données dans le cas où $\mathcal{X}$ est un espace discret ; elles s'adaptent aisément au cas continu.

Définition 17 *On dit que la loi p est* stationnaire *pour la chaîne (X_n) si*

$$X_n \sim p \Rightarrow X_{n+1} \sim p.$$

Définition 18 *On dit que la chaîne de Markov (X_n) est* irréductible *si*

$$\forall x, x' \in \mathcal{X}, \exists k > 0, P[X_k = x' | X_0 = x] > 0.$$

Autrement dit, une chaîne est irréductible si à partir de n'importe quel point, on peut atteindre n'importe quel autre point en un nombre fini d'itérations.

Définition 19 *Soit $k \geq 2$ un entier. On dit que la chaîne de Markov (X_n) est* périodique *de période k si*

$$\exists x \in \mathcal{X}, pgcd\,\{n \in \mathbb{N} : P[X_n = x | X_0 = x] > 0\} = k.$$

Autrement dit, la chaîne ne peut passer par l'état x que si l'itération est un multiple de k.

Définition 20 *On dit que la chaîne de Markov (X_n) est* apériodique *si elle n'est pas périodique.*

L'intuition est que si une chaîne est à la fois irréductible et apériodique, alors

$$\forall x, x' \in \mathcal{X}, \exists N, \forall n \geq N, P[X_n = x' | X_0 = x] > 0.$$

Définition 21 *Pour $x \in \mathcal{X}$, notons*

$$T_x = \inf\{n \geq 1 : X_n = x | X_0 = x\}$$

le temps du premier retour au point x. On dit que la chaîne de Markov (X_n) est positive récurrente *si*

$$\forall x \in \mathcal{X}, E[T_x] < \infty.$$

Une chaîne positive récurrente est donc une chaîne qui reviendra à son point d'origine en temps fini.

Définition 22 *On dit qu'une chaîne de Markov est* ergodique *si elle est irréductible, apériodique et positive récurrente.*

Théorème 4 (Théorème ergodique) *Soit (X_n) une chaîne de Markov ergodique. Alors il existe une unique loi stationnaire p pour (X_n) et on a :*

$$X_n \xrightarrow{\mathcal{L}} p$$

et pour toute fonction intégrable f

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{\mathbb{P}} E_p[f].$$

Ce théorème est le fondement des méthodes de Monte-Carlo par chaînes de Markov (MCMC). Si on a une distribution d'intérêt p et qu'on parvient à construire une chaîne de Markov (X_n) ergodique de loi stationnaire p, alors on pourra estimer par Monte-Carlo n'importe quelle espérance sous p.

Nous présentons maintenant deux techniques pour construire une chaîne de Markov ayant une loi stationnaire π définie.

5.2 Échantillonnage de Gibbs

L'échantillonnage de Gibbs est utile dans le cas où le paramètre θ de la loi π est de dimension ≥ 2. Il est courant que la loi jointe soit difficile à simuler, mais que les lois conditionnelles soient accessibles.

5.2.1 Échantillonneur de Gibbs à 2 étapes

Étudions d'abord le cas où on peut écrire $\theta = (\theta_1, \theta_2)$ et où les lois conditionnelles $\pi_{1|2}(\theta_1|\theta_2)$ et $\pi_{2|1}(\theta_2|\theta_1)$ sont aisées à simuler. L'échantillonneur de Gibbs correspond à l'algorithme suivant :

- Initialiser (θ_1^0, θ_2^0) arbitrairement.
- À l'itération t :
  - Tirer $\theta_1^t \sim \pi_{1|2}(\cdot|\theta_2^{t-1})$
  - Tirer $\theta_2^t \sim \pi_{2|1}(\cdot|\theta_1^t)$

Sous des conditions très générales, la chaîne de Markov ainsi définie converge vers la loi π.

5.2.2 Échantillonneur de Gibbs à k étapes

De manière plus générale, supposons qu'on puisse écrire $\theta = (\theta_1, \ldots, \theta_k)$. L'échantillonneur de Gibbs à k étapes correspond à l'algorithme suivante :

- Initialiser $(\theta_1^0, \ldots, \theta_k^0)$ arbitrairement
- À l'itération t :
  - Tirer $\theta_1^t \sim \pi(\cdot|\theta_2^{t-1}, \theta_2^{t-1}, \ldots, \theta_k^{t-1})$
  - Tirer $\theta_2^t \sim \pi(\cdot|\theta_1^t, \theta_3^{t-1}, \ldots, \theta_k^{t-1})$
  - ...
  - Tirer $\theta_i^t \sim \pi(\cdot|\theta_1^t, \theta_2^t, \ldots, \theta_{i-1}^t, \theta_{i+1}^{t-1}, \ldots, \theta_k^{t-1})$
  - ...
  - Tirer $\theta_k^t \pi \cdot \theta_1^t, \theta_2^t, \ldots, \theta_{k-1}^t)$

On met donc à jour séquentiellement les composantes de θ.

Notes On peut également mettre à jour les composantes dans un ordre aléatoire ; cela peut améliorer la vitesse de convergence de la chaîne de Markov. Par ailleurs, notons que rien n'oblige à ce que les (θ_i) soient des scalaires : on peut diviser le paramètre θ en blocs de n'importe quelle dimension, tant que les lois conditionnelles sont disponibles.

5.2.3 Exemple

Reprenons l'exemple de la régression linéaire. On veut décider quelles covariables inclure dans notre modèle. Cela revient à prendre un vecteur binaire $(\gamma_1, \ldots, \gamma_p)$, avec $\gamma_i = 1$ si on inclut la covariable i dans le modèle et $\gamma_i = 0$ sinon. Il y a 2^p modèles parmi lesquels choisir. Chaque modèle a une probabilité a posteriori associée, mais si p n'est pas très petit, il n'est pas envisageable de calculer la probabilité de chaque modèle. On construit donc plutôt une chaîne de Markov, à valeurs dans l'espace des modèles, dont la loi stationnaire est la loi a posteriori des modèles.

La loi conditionnelle d'un γ_i est simple à simuler, puisqu'il s'agit d'une loi de Bernoulli. Avec $i = 1$, on a

$$\begin{aligned}\pi_1(\gamma_1 = 0|\gamma_2^{(t-1)}, \ldots, \gamma_p^{(t-1)}, y) &\propto \pi_1(0, \gamma_2^{(t-1)}, \ldots, \gamma_p^{(t-1)}|y)\\ \pi_1(\gamma_1 = 1|\gamma_2^{(t-1)}, \ldots, \gamma_p^{(t-1)}, y) &\propto \pi_1(1, \gamma_2^{(t-1)}, \ldots, \gamma_p^{(t-1)}|y)\end{aligned}$$

et de même pour les autres valeurs de i. On reconnaît la vraisemblance marginale, qu'on a calculée dans le chapitre précédent.

5.3 Algorithme de Metropolis-Hastings

Définition 23 (detailed balance) *Soit p une mesure de probabilité sur $\mathcal{X}$ et (Z_t) une chaîne de Markov à valeurs dans $\mathcal{X}$, de fonction de transition $K(x \to x')$ [1]. On dit que (Z_t) vérifie le detailed balance pour p si*

$$\forall x, x' \in \mathcal{X},\ p(x)K(x \to x') = p(x')K(x' \to x).$$

Théorème 5 *Si la chaîne de Markov (Z_t) vérifie le detailed balance pour la loi p, alors p est une loi stationnaire pour (Z_t).*

Preuve On donne la preuve dans le cas où $\mathcal{X}$ est discret ; elle s'adapte aisément.

Supposons que $Z_t \sim p$. Alors

$$P[Z_{t+1} = i] = \sum_j P[Z_t = j]K(j \to i) = \sum_j p(j)K(j \to i) = \sum_j p(i)K(i \to j) = p(i)$$

et on a donc $Z_{t+1} \sim p$.

Définition 24 (Algorithme de Metropolis-Hastings) *Soit p une mesure de probabilité sur $\mathcal{X}$. On se donne un noyau de proposition $q(\cdot|\cdot)$ sur $\mathcal{X}$, c'est à dire que $\forall x \in \mathcal{X}$, $q(\cdot|x)$ est une densité de probabilité sur $\mathcal{X}$. On se donne une valeur initiale Z_0 arbitraire. On appelle algorithme de Metropolis-Hastings l'algorithme dont l'itération $t+1$ est donnée par :*

1. *Tirer $Y_{t+1} \sim q(\cdot|z_t)$;*
2. *Poser la probabilité d'acceptation*

$$\alpha_t = \min\left(1, \frac{p(y_{t+1})}{p(z_t)}\frac{q(z_t|y_{t+1})}{q(y_{t+1}|z_t)}\right);$$

1. Autrement dit, la loi de X_{n+1} conditionnellement à $\{X_n = x\}$ est de densité $K(x \to \cdot)$.

3. Poser

$$Z_{t+1} = \begin{cases} Y_{t+1} & \text{avec probabilité } \alpha_t \\ Z_t & \text{avec probabilité } 1-\alpha_t. \end{cases}$$

Théorème 6 *La chaîne de Markov définie par l'algorithme de Metropolis-Hastings vérifie le detailed balance pour p.*

Preuve Notons K la fonction de transition de la chaîne (Z_t). Soient $x, x' \in \mathcal{X}$. Étudions la transition $x \to x'$, *i.e.* $Z_t = x$ et $Y_{t+1} = x'$. Par symétrie, on peut supposer $\alpha_t \leq 1$ (ce qui signifie que si on avait proposé la transition $x' \to x$, la probabilité d'acceptation aurait été 1). On a

$$\frac{K(x \to x')}{K(x' \to x)} = \frac{q(x'|x)\alpha_t}{q(x|x')} = \frac{p(x')}{p(x)}$$

et le detailed balance est bien vérifié.

Pour peu qu'on ait choisi q de façon à ce que (Z_t) soit apériodique et positive récurrente, ce qui est aisé en pratique, le théorème ergodique garantit que l'estimateur $\hat{I}_T^{MCMC}$ converge en probabilité.

5.4 Vérifications de convergence

Par construction, notre chaîne de Markov ne suit la loi d'intérêt qu'asymptotiquement, et ce quelle que soit la méthodologie retenue : Gibbs, Metropolis-Hastings ou autre. Il convient de vérifier qu'on a atteint l'asymptote, puis qu'on explore bien toute la loi cible. Cela ne peut se faire qu'heuristiquement.

5.4.1 Convergence vers la loi limite et burn-in

Les premières itérations de la chaîne de Markov ne suivent généralement pas la loi cible. Il convient donc d'afficher la trace de la chaîne, c'est à dire la valeur prise à chaque itération, pour détecter à quel moment la chaîne atteint sa loi limite. On supprime alors les premières itérations ; c'est ce qu'on appelle le *burn-in*. Cela peut être très long : pour des modèles complexes, le burn-in peut être de plusieurs millions d'itérations.

On conseille généralement de faire tourner la chaîne de Markov pendant un nombre d'itérations nettement supérieur au burn-in, par exemple au moins 10 fois le burn-in. Cela augmente les chances de découvrir d'éventuels modes secondaires de la loi explorée.

La figure 5.1 montre un exemple de convergence d'un échantillonneur de Gibbs. Ici, la convergence se fait rapidement : en supprimant les 20 premières itérations, on obtient une chaîne qui a convergé et qu'on peut utiliser.

Attention : la convergence qui nous intéresse ici est la convergence *en loi*. La chaîne ne reste donc pas sur une valeur fixe : elle explore une loi, et prend donc des valeurs différentes. Mais la loi sous-jacente est toujours la même.

5.4.2 Exploration de la loi limite et vitesse de mélange

Une fois que la chaîne de Markov a atteint sa loi limite, il faut qu'elle l'explore suffisamment rapidement. C'est ce qu'on appelle la vitesse de mélange de la chaîne.

Calculons la variance d'un estimateur MCMC :

$$Var\left(\hat{I}_T^{MCMC}\right) = \frac{1}{T} Var(h(Z)) + \frac{2}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{t-1} Cov\left(h(Z_t), h(Z_{t'})\right).$$

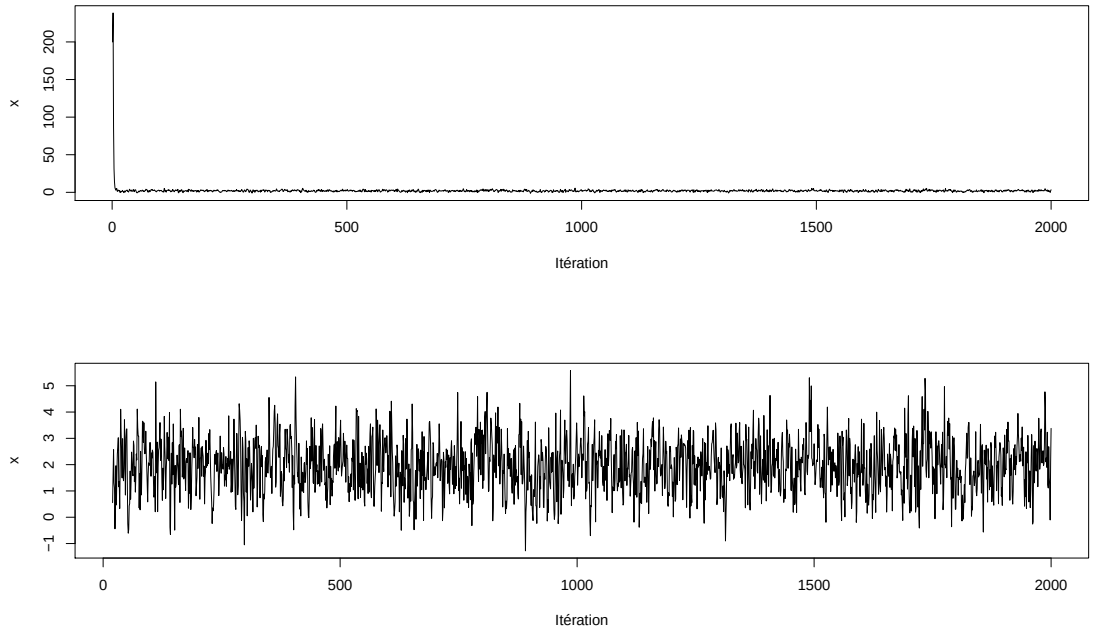

FIGURE 5.1 – Convergence d'un échantillonneur de Gibbs : valeur d'une des composantes de la chaîne à chaque itération. Sur la figure du haut, on voit que lors des premières itérations, la chaîne n'a pas encore convergé. La figure du bas correspond à la même chaîne dont on a supprimé les 20 premières itérations ; l'échelle des ordonnées est donc différente. On voit que la chaîne semble avoir convergé.

Comme Z_t converge en loi, on a à l'asymptote $Cov\left(h(Z_t), h(Z_{t'})\right) = Cov\left(h(Z_{t+k}), h(Z_{t'+k})\right)$ pour tout k et chaque terme apparaît donc T fois dans la double somme, d'où

$$\begin{aligned} Var\left(\hat{I}_T^{MCMC}\right) &\approx \frac{1}{T}Var(h(Z)) + \frac{2}{T}\sum_{k=1}^{T} Cov\left(h(Z_t), h(Z_{t+k})\right) \text{ pour un } t \text{ arbitraire suffisamment grand} \\ &\approx Var(h(Z))\left(\frac{1}{T} + \frac{2}{T}\sum \rho_k\right) \text{ où } \rho_k = Cor(h(Z_t), h(Z_{t+k})) \\ &\approx \frac{1}{M}Var(h(Z)) \end{aligned}$$

où $M = \frac{T}{1+2\sum \rho_k}$ est la taille d'échantillon effective (*effective sample size*, ESS) : notre estimateur MCMC après T itérations $\hat{I}_T^{MCMC}$ est de même qualité qu'un estimateur Monte-Carlo $\hat{I}_M^{MC}$ après M itérations (avec $M \leq T$).

Pour l'algorithme de Metropolis-Hastings, ceci nous donne une heuristique pour choisir une bonne proposition q : on cherche à minimiser l'autocorrélation de la chaîne (Z_t), c'est-à-dire la corrélation entre Z_t et Z_{t+k}, pour tout k.

Prenons l'exemple simple où $q(\cdot|x) = \mathcal{N}(x, \sigma^2)$ et où cherche à optimiser en σ. Si $\sigma \to 0$, alors on propose une valeur Y_{t+1} très proche de Z_t (et qui sera acceptée avec une probabilité α_t proche de 1), donc l'autocorrélation sera proche de 1. Si $\sigma \to \infty$, alors on propose une valeur Y_{t+1} très éloignée de Z_t, qui ne sera généralement pas acceptée ($\alpha_t \approx 0$) : on a une forte probabilité que $Z_{t+1} = Z_t$, donc l'autocorrélation sera encore proche de 1. Il existe donc une valeur optimale finie de σ qui minimise l'autocorrélation. Dans un cas simple, Robert, Gelman & Gilks (1997) ont montré

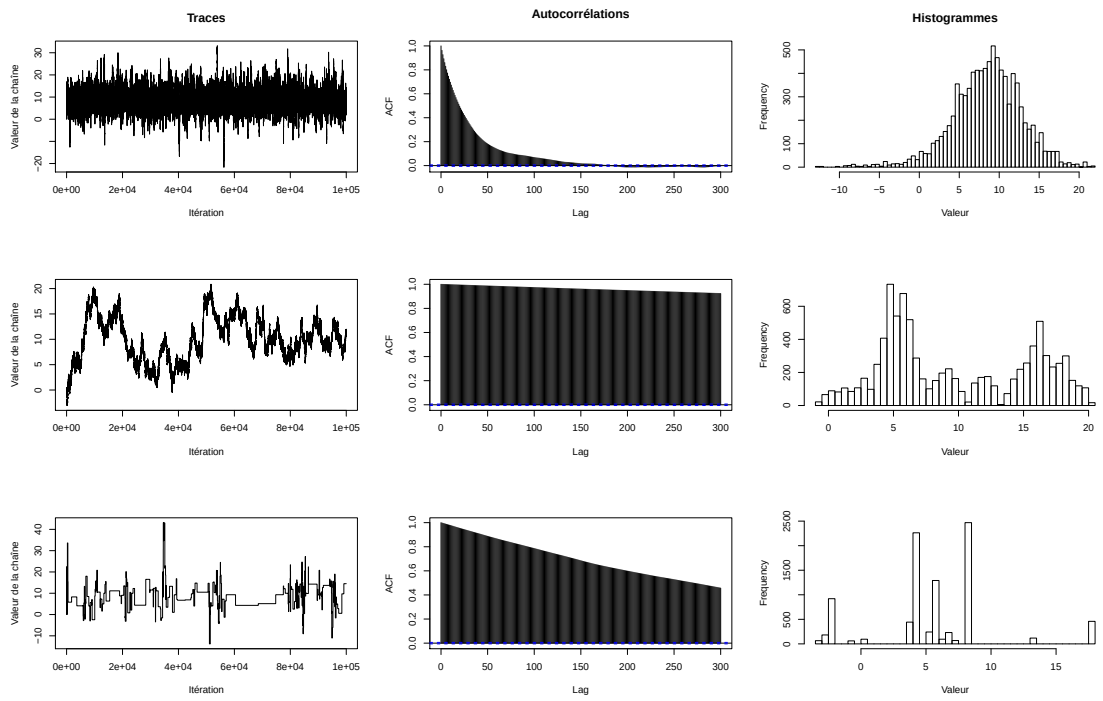

FIGURE 5.2 – Convergence d'un algorithme de Metropolis-Hastings en fonction de la variance de la proposition. La première ligne correspond à un algorithme bien dimensionné : la chaîne explore bien toute la loi (colonne 1) et l'autocorrélation décroît à une vitesse raisonnable (colonne 2). La deuxième ligne correspond à une variance trop faible dans la proposition : la chaîne met beaucoup d'itérations à explorer la loi, et l'autocorrélation décroît très lentement. La troisième ligne correspond à une variance trop élevée : les valeurs proposées sont très souvent rejetées, et la chaîne reste donc bloquée au même point pendant de nombreuses itérations. Là aussi, l'autocorrélation décroît lentement. La colonne 3 montre l'histogramme des valeurs obtenues : on voit que sur les lignes 2 et 3, on obtient des valeurs très éloignées de la vraie loi qui figure sur la ligne 1.

que l'optimum est atteint si $E[\alpha_t] = 0.234$; l'expérience montre que des valeurs proches de 0.234 fonctionnent également bien dans un grand nombre de cas plus complexes.

La figure 5.2 permet de visualiser ce phénomène. Pour un modèle hiérarchique que nous ne détaillons pas, nous avons fait tourner un algorithme de Metropolis-Hastings à noyau gaussien avec 3 variances différentes pour la proposition.

— La première ligne correspond à une variance de 2.5 ; on obtient (sur 10^5 itérations) un taux d'acceptation moyen de 0.30 et un ESS de 1270. On voit sur la trace que la chaîne mélange rapidement et explore toute la loi. L'autocorrélation décroît nettement plus rapidement que sur les lignes 2 et 3 (colonne 2).

— La deuxième ligne correspond à une variance de 0.1. La chaîne mélange très lentement : la valeur proposée à l'itération $k+1$ est très proche de la valeur à l'itération k, et il faut donc longtemps pour explorer toutes les valeurs de la loi a posteriori. L'autocorrélation décroît donc lentement, et on obtient un histogramme de valeurs qui ne ressemble pas à la loi cible. Le taux d'acceptation est de 0.96 et l'ESS de 14.

— La troisième ligne correspond à une variance de 10. La chaîne reste bloquée à la même valeur pendant de nombreuses itérations, parce que la proposition est souvent très mauvaise.

L'autocorrélation décroît donc lentement, et l'histogramme des valeurs obtenues n'est pas bon. Le taux d'acceptation est de 0.003 et l'ESS de 85.