UNIVERSITÉ DE NANTES

Master 2 Ingénierie Statistique

Statistique Bayésienne

Paul Rochet

Contents

1 Introduction **2**
1.1 Lois a priori et a posteriori . . . 2
1.2 Lois a priori impropres . . . 4
1.3 Comportement asymptotique de la densité a posteriori . . . 5

2 Estimation **7**
2.1 Théorie de la décision . . . 7
2.2 Propriétés de l'estimateur Bayésien . . . 9
2.3 Maximum a posteriori . . . 11
2.4 Prévision . . . 11

3 Choix de l'a priori **14**
3.1 Lois a priori informatives . . . 15
3.2 Familles conjuguées . . . 16
3.3 Modèles hiérarchiques . . . 17
3.4 Approche non-informative . . . 19

4 Régions de confiance et tests Bayésiens **24**
4.1 Régions HPD . . . 25
4.2 Intervalles de prévision . . . 27
4.3 Facteur de Bayes . . . 28

5 Méthodes de Monte-Carlo **30**
5.1 Echantillonage . . . 30
5.2 Importance sampling . . . 32
5.3 Méthodes MCMC . . . 33

6 Simulation a posteriori **35**
6.1 Algorithme d'acceptation/rejet . . . 35
6.2 Algorithme d'Hasting-Metropolis . . . 37
6.3 Echantillonage de Gibbs . . . 38
6.4 Lois de mélange . . . 39

1 Introduction

On se place dans un cadre d'inférence standard: soit X un échantillon et $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ un modèle statistique paramétrique. La distinction entre l'inférence statistique "classique" dite fréquentiste et l'inférence Bayésienne réside dans la façon de traiter le paramètre d'intérêt θ. Selon l'approche fréquentiste, θ prend une valeur précise inconnue qui régie la loi de X. Pour l'approche Bayésienne, θ est vu comme aléatoire au même titre que les données elles-mêmes. L'influence de θ sur X est en quelque sorte bilatérale et se traduit par une dépendance entre les variables aléatoires X et θ.

1.1 Lois a priori et a posteriori

La loi de θ, notée $\pi(.)$ et appelée loi *a priori*, est choisie par le statisticien. Elle modélise l'information disponible sur θ *avant d'observer les données*. La loi P_θ du modèle est alors vue comme la loi conditionnelle sachant θ. Dans ce cours, on se placera (sauf mention contraire) dans le cadre plus simple suivant:

- la loi a priori est absolument continue par rapport à la mesure de Lebesgue. La notation $\pi(\theta)$ désigne sa densité au point θ.
- les données $X = (X_1, ..., X_n)$ sont indépendantes *conditionnellement à* θ de loi conditionnelle P_θ continue ou discrète. f_θ désigne la densité de P_θ par rapport à la mesure de Lebesgue ou mesure de comptage selon le cas.

Remarques.

1. Du point de vue Bayésien, les X_i ne sont pas indépendants puisqu'ils dépendent conjointement de θ.
2. Pour que le cadre théorique soit valide, il est indispensable que la loi a priori soit choisie indépendamment des observations.

Le fait de considérer θ comme aléatoire permet d'intégrer de l'information supplémentaire, non issue des observations X_i, sur le paramètre θ par le biais de l'a priori π. Par exemple, si on sait que θ est positif, on peut choisir un a priori de support inclus dans $\mathbb{R}_+$. L'approche Bayésienne permet également plus de souplesse dans l'estimation de θ qui n'est pas basée entièrement et "aveuglément" sur les données. Enfin, elle fournit une distribution du paramètre θ basée sur les observations, plus simple à interpréter qu'une valeur ponctuelle ou un intervalle de confiance. Précisément, l'inférence Bayésienne consiste à déterminer la loi de θ sachant les observations X. Cette loi, dite loi *a posteriori*, se déduit facilement par la formule de Bayes. En effet:

i) Le paramètre θ est une variable aléatoire de loi π.

ii) Les X_i sont indépendants sachant θ, de loi conditionnelle P_θ. La loi de X sachant θ est donc donnée par la vraisemblance du modèle,

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

en posant $x = (x_1, ..., x_n)$.

iii) La loi jointe de (X, θ) est donnée par la formule de Bayes

$$f_{X,\theta}(x, \theta) = f_X(x|\theta)\pi(\theta).$$

iv) La loi de X est la loi marginale

$$f_X(x) = \int_\Theta f_{X,\theta}(x, \theta)d\theta = \int_\Theta f_X(x|\theta)\pi(\theta)d\theta.$$

v) Enfin, la loi de θ sachant $X = x$ s'obtient par la formule de Bayes

$$\pi(\theta|X = x) = \frac{f_{X,\theta}(x, \theta)}{f_X(x)}.$$

Le dénominateur $f_X(x)$ ne dépend pas de θ et s'apparente donc à une constante de normalisation. Il n'est souvent pas nécessaire de le calculer explicitement puisqu'une loi de probabilité est entièrement caractérisée à constante multiplicative près. On utilisera le symbole $\propto$ pour signifier que deux lois sont proportionnelles. Ici,

$$\pi(\theta|X = x) \propto f_{X,\theta}(x, \theta) \propto f_X(x|\theta)\pi(\theta).$$

Définition. La loi a posteriori est la loi conditionnelle de θ sachant X, définie par la densité $\pi(.|X = x)$ évaluée en $x = X$. On la note $\pi(.|X)$.

La loi a posteriori $\pi(\theta|X)$ donne les valeurs attendues du paramètre compte tenu des observations et du choix d'a priori. Elle résume toute l'information sur θ dans l'analyse Bayésienne et s'interprète selon les objectifs.

i) L'estimation: choisir une valeur moyenne, médiane ou "probable" (le mode) pour la loi a posteriori.

ii) La construction d'intervalles de confiance (appelé intervalle de crédibilité dans le cadre Bayésien): déterminer un intervalle de probabilité donnée pour la loi a posteriori. On privilégie en général les intervalles les plus courts, appelés intervalles HPD pour *Highest Posterior Density.*

iii) Le test sur des valeurs de θ: tester l'hypothèse que θ appartient à un certain ensemble de valeurs se fait en évaluant sa probabilité a posteriori.

Remarque. La loi a posteriori est absolument continue par rapport à la loi a priori, en effet

$$\pi(\theta|X) = \frac{f_X(X|\theta)}{f_X(X)}\pi(\theta).$$

Une interprétation fréquentiste de l'analyse Bayésienne est la suivante. Soit θ une réalisation d'une variable aléatoire de loi π, on construit $X_1, ..., X_n$ iid de loi P_θ. Il sagit de retrouver de l'information sur θ en observant seulement les X_i. Cette information est contenue dans la loi de θ sachant X, qui n'est autre que la loi a posteriori.

Exemples. Calculer les loi a posteriori dans les modèles suivants:

1. $X = (X_1, ..., X_n)$ iid de loi $\mathcal{N}(\theta, \sigma^2)$ avec $\sigma^2 > 0$ connu. On prend comme a priori la loi normale $\theta \sim \mathcal{N}(0, \tau^2)$.

2. $X = (X_1, ..., X_n)$ iid de loi de Poisson de paramètre $\lambda > 0$, et la loi a priori sur λ la loi exponentielle de paramètre $\tau > 0$.

3. $X = (X_1, ..., X_n)$ iid de loi exponentielle de paramètre $\lambda > 0$ avec pour a priori la loi Gamma $\Gamma(a, b)$ de densité
$$\gamma_{a,b}(x) = \frac{b^a}{\Gamma(a)} e^{-bx} x^{a-1} \mathbb{1}\{x > 0\} \ , \ a, b > 0.$$

4. $X = (X_1, ..., X_n)$ iid de loi $\mathcal{N}(\theta, \sigma^2)$ avec $\sigma^2 > 0$ connu. On prend comme a priori la mesure de Lebesgue sur $\mathbb{R}$.

1.2 Lois a priori impropres

Il est parfois utile d'utiliser comme a priori une mesure de masse totale infinie, càd telle que $\pi(\Theta) = +\infty$, par exemple dans un but non-informatif: on ne dispose d'aucune information sur un paramétre $\theta \in \mathbb{R}$, on prend la mesure de Lebesgue comme loi a priori. Cette approche peut s'avérer pertinente à condition que la loi a posteriori soit de masse totale finie.

Définition. Une loi a priori impropre est une mesure π sur le paramètre θ qui vérifie simultanément
$$\int_\Theta d\pi(\theta) = +\infty \quad \text{et} \quad \int_\Theta f_X(X|\theta) d\pi(\theta) < +\infty \ p.s.$$
La loi a posteriori est alors définie de la même façon par
$$\pi(\theta|X) = \frac{f_X(X|\theta)\pi(\theta)}{\int_\Theta f_X(X|\theta) d\pi(\theta)},$$
qui est bien une loi de probabilité (de mesure totale 1).

Une loi impropre peut s'avérer utile pour traduire un manque d'information sur le paramètre.

Exemples.

1. Soit le modèle Gaussien $X = (X_1, ..., X_n)$ iid de loi $\mathcal{N}(\theta, 1)$. Si on ne sait rien du paramètre θ, on peut utiliser comme a priori π la mesure de Lebesgue sur $\mathbb{R}$, qui joue un rôle de loi uniforme sur $\mathbb{R}$. On vérifie bien que l'intégrale
$$\int_\Theta f_X(X|\theta)\pi(\theta) d\theta = \int \frac{1}{\sqrt{2\pi}^n} \exp\Big(-\frac{1}{2} \sum_{i=1}^n (X_i - \theta)^2 \Big) d\theta$$
est convergente pour tout X.

2. Soit $X_1, ..., X_n$ iid de loi exponentielle $\mathcal{E}(\theta)$ avec comme a priori sur θ, $\pi(\theta) = 1/\theta, \theta > 0$. On remarque que $S := \sum_{i=1}^n X_i$ est strictement positif presque sûrement, avec dans ce cas
$$\int \pi(\theta) d\theta = \int_0^{+\infty} \frac{d\theta}{\theta} = +\infty \quad \text{et} \quad \int f_X(X|\theta)\pi(\theta) d\theta = \int_0^{+\infty} \theta^{n-1} e^{-\theta S} d\theta < \infty.$$
C'est donc bien une loi impropre.

3. Soit $X_1, ..., X_n$ iid de loi $\mathcal{N}(0, \sigma^2)$ avec comme a priori sur σ^2 la mesure de Lebesgue sur $\mathbb{R}_+$: $\pi(\sigma^2) = \mathbb{1}\{\sigma^2 > 0\}$. On a

$$\int f_X(X|\sigma^2)\pi(\sigma^2)d\sigma^2 = \int \frac{1}{\sqrt{2\pi}^n} \frac{1}{(\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2\Big)d\sigma^2$$

qui est fini presque sûrement à partir de $n \geq 3$.

1.3 Comportement asymptotique de la densité a posteriori

Un des résultats principaux de la Statistique Baysienne est le théorème de Bernstein-von Mises qui décrit le comportement asymptotique de la loi a posteriori dans le cadre fréquentiste classique où $X_1, ..., X_n$ sont iid de densité f_{θ_0} (il existe donc une valeur θ_0 du paramètre qui régit les observations) dans un modèle paramétrique régulier $f_\theta, \theta \in \Theta \subseteq \mathbb{R}^d$. C'est le résultat central pour montrer entre autres les propriétés asymptotiques des estimateurs et intervalles de crédibilité Bayésiens.

Dans sa version simplifié, le théorème de Bernstein-von Mises établit la convergence étroite, notée $\rightharpoonup$, de la densité a posteriori renormalisée. On rappelle qu'une suite de lois μ_n converge étroitement vers μ si pour toute fonction h continue bornée, $\int h d\mu_n$ converge vers $\int h d\mu$. En particulier, la convergence en loi d'une suite de variable aléatoires est équivalente à la convergence étroite des lois.

La preuve du théorème de Bernstein-von Mises nécessite des hypothèses techniques sur la régularité du modèle. On dira qu'un modèle $\{f_\theta : \theta \in \Theta\}$ est régulier si $\Theta \subseteq \mathbb{R}^d$, $\theta \mapsto \log(f_\theta)$ est de classe $\mathcal{C}^2$ presque partout avec

$$\ell_\theta(x) = \log(f_\theta(x)) \in \mathbb{R}\ ,\ s_\theta(x) = \frac{\partial}{\partial\theta}\ell_\theta(x) \in \mathbb{R}^d \quad \text{et} \quad h_\theta(x) = \frac{\partial^2}{\partial\theta\partial\theta^\top}\ell_\theta(x) = \frac{\partial}{\partial\theta^\top}s_\theta(x) \in \mathbb{R}^{d\times d},$$

et l'information de Fisher vérifie

$$\mathcal{I}(\theta) := \int s_\theta(x)s_\theta(x)^\top f_\theta(x)dx = -\int h_\theta(x)f_\theta(x)dx.$$

Pour simplifier les preuves, on supposera également que la famille $\{h_\theta : \theta \in \Theta\}$ forme une classe de Glivenko-Cantelli. Cette hypothèse technique traduit une certaine régularité du modèle et entraîne en particulier la condition suivante

$$\sup_{\theta\in\Theta}\Big\|\frac{1}{n}\sum_{i=1}^n h_\theta(X_i) + \mathcal{I}(\theta)\Big\| \xrightarrow[n\to\infty]{\mathbb{P}} 0. \tag{$*$}$$

Théorème 1.1 (Bernstein-von Mises)**.** *Soit $X = (X_1, ..., X_n)$ iid de densité f_{θ_0} dans le modèle paramétrique régulier $\{f_\theta : \theta \in \Theta\}$ vérifiant $(*)$ et $\pi(.)$ une densité a priori de classe $\mathcal{C}^1$ strictement positive sur Θ. On suppose que l'estimateur du maximum de vraisemblance $\widehat{\theta}_{MV}$ existe et est consistent. Alors, la loi a posteriori $\pi(.|X)$ vérifie*

$$\frac{1}{\sqrt{n}}\ \pi\big(\widehat{\theta}_{MV} + ./\sqrt{n}|X\big) \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}\big(0, \mathcal{I}(\theta_0)^{-1}\big).$$

Preuve. On suppose $\theta \in \mathbb{R}$ pour simplifier mais la preuve se généralise facilement à $\theta \in \mathbb{R}^d$. Par la formule de Taylor-Lagrange appliquée à la log-vraisemblance $v(\theta|X) = \sum_{i=1}^n \ell_\theta(X_i)$, on sait que pour tout θ, il existe $\tilde{\theta}$ compris entre θ et $\widehat{\theta}_{MV}$ tel que

$$v(\theta|X) = v(\widehat{\theta}_{MV}|X) + \frac{(\theta - \widehat{\theta}_{MV})^2}{2} v''(\tilde{\theta}|X).$$

Or $\pi\big(\theta|X\big) \propto \exp\big(v(\theta, X) + \log(\pi(\theta))\big)$ d'où, en posant $\theta = \widehat{\theta}_{MV} + t/\sqrt{n}$,

$$\pi\big(\widehat{\theta}_{MV} + t/\sqrt{n}|X\big) \propto \exp\Big(v(\widehat{\theta}_{MV}|X) + \frac{t^2}{2n} v''(\tilde{\theta}_t|X) + \log\big(\pi(\widehat{\theta}_{MV} + t/\sqrt{n})\big)\Big).$$

A t fixé, on a $\log\big(\pi(\widehat{\theta}_{MV} + t/\sqrt{n})\big) = \log(\pi(\theta_0)) + O_P(1/\sqrt{n})$ et

$$\frac{1}{n} v''(\tilde{\theta}_t|X) = \frac{1}{n}\sum_{i=1}^n h_{\tilde{\theta}_t}(X_i) \xrightarrow[n\to\infty]{\mathbb{P}} \int h_{\theta_0}(x) f_{\theta_0}(x) dx = -\mathcal{I}(\theta_0),$$

où on a utilisé que $\tilde{\theta}_t \to \theta_0$ pour tout t et la condition $(*)$. Ainsi, $\pi\big(\widehat{\theta}_{MV} + t/\sqrt{n}|X\big) \propto \exp\big(-\mathcal{I}(\theta_0)t^2/2 + o_P(1)\big)$ où on reconnaît la densité de la loi normale $\mathcal{N}\big(0, \mathcal{I}(\theta_0)^{-1}\big)$. Par le lemme de Scheffé, la convergence simple des densités implique la convergence dans $\mathbb{L}^1$ et donc la convergence étroite. □

Remarque. L'influence de l'a priori s'estompe asymptotiquement. Des densités a priori (raisonnables) différentes entraînent le même comportement de l'a posteriori asymptotiquement.

Le théorème de Bernstein-von Mises est énoncé ici dans une version faible mais la preuve suffit à montrer la convergence en variation totale, plus forte que la convergence étroite. Un énoncé équivalent s'obtient en considérant une variable aléatoire T_X à valeurs dans Θ et de densité $\pi(.|X)$.

Corollaire 1.2. *Sous les hypothèses du théorème,*

$$\sqrt{n}\big(T_X - \widehat{\theta}_{MV}\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0, \mathcal{I}(\theta_0)^{-1}\big).$$

Il faut noter que la variable aléatoire T_X n'apparaît pas naturellement dans le contexte Bayésien. Elle est introduite ici de manière artificielle afin de formuler la convergence étroite des mesures en termes de convergence en loi de variables aléatoires. Une conséquence immédiate est la convergence étroite de la densité a posteriori vers la masse de Dirac en θ_0.

Corollaire 1.3. *Sous les hypothèses du théorème, la loi a posteriori converge étroitement en probabilité vers δ_{θ_0}, la masse de Dirac en θ_0.*

Preuve. Immédiat en remarquant que si T_X a pour densité $\pi(.|X)$ alors, $\sqrt{n}(T_X - \widehat{\theta}_{MV})$ a pour densité $\pi(\widehat{\theta}_{MV} + ./\sqrt{n}|X)/\sqrt{n}$ (en considérant $\widehat{\theta}_{MV}$ comme déterministe puisque conditionnellement à X). □

2 Estimation

La loi a posteriori décrit le comportement de la variable aléatoire θ sachant les observations X_i. On peut naturellement en déduire une estimation de θ, en prenant par exemple l'espérance, la médiane ou le mode a posteriori. Plus généralement, on peut construire un estimateur optimal (en un sens que l'on définira) pour une fonction de perte donnée $L(\theta, \widehat{\theta})$ qui quantifie le coût de l'erreur d'estimation.

2.1 Théorie de la décision

On distingue généralement deux types de décision en statistique inférentielle:

- une valeur ponctuelle dans l'ensemble Θ si l'objectif est d'estimer un paramètre θ
- une réponse binaire 1 ou 0 pour un test statistique.

Formellement, une décision est l'image $\delta(X)$ de l'échantillon X par une fonction (mesurable) δ, appelée règle de décision. Dans le premier cas, $\delta(X)$ est simplement un estimateur de θ. Dans le second cas, $\delta(X)$ est le résultat du test (rejeter ou non l'hypothèse nulle). Dans ce chapitre, on s'intéressera uniquement à l'estimation du paramètre θ.

La qualité d'une règle de décision (et donc d'un estimateur) est définie par le biais d'une fonction de perte $L : \Theta \times \Theta \to \mathbb{R}_+$, qui mesure le coût de l'erreur commise par la décision $\delta(X)$ lorsque la vraie valeur du paramètre est θ. Des exemples classiques sont:

1. la perte quadratique: $L(\theta, \delta(X)) = (\theta - \delta(X))^2$
2. la perte $\mathbb{L}^1$: $L(\theta, \delta(X)) = |\theta - \delta(X))|$
3. la perte $0-1$: $L(\theta, \delta) = \mathbb{1}\{\delta(X) \neq \theta\}$, utilisée pour la classification.

Définition. Le *risque fréquentiste* d'une règle de décision δ associé à une fonction de perte L est défini par

$$R_\delta(\theta) = \mathbb{E}\big(L(\theta, \delta(X))|\theta\big) \ , \ \theta \in \Theta.$$

La valeur $R_\delta(\theta) \in [0, +\infty]$ mesure donc l'erreur moyenne de la règle de décision δ lorsque les observations sont tirées selon θ. On dit qu'une règle de décision δ_1 est préférable à une autre δ_2 si $R_{\delta_1}(\theta) \leq R_{\delta_2}(\theta)$ pour tout $\theta \in \Theta$ avec inégalité stricte pour au moins une valeur de θ. Ceci définit une relation d'ordre partielle sur l'ensemble $\mathcal{D}$ des règles de décision. Sauf dans des cas triviaux, cette relation n'est pas totale: il n'existe pas de règle de décision optimale δ^* telle que

$$\forall \delta \in \mathcal{D} \ , \ \forall \theta \in \Theta \ , \ R_{\delta^*}(\theta) \leq R_\delta(\theta).$$

Remarque. Dans un modèle paramétrique régulier, une solution pour $L(.,.)$ la fonction de perte quadratique est de se restreindre aux estimateurs sans biais, c'est-à-dire en ne considérant que les règles de décision δ telle que $\mathbb{E}(\delta(X)|\theta) = \theta$. Dans ce cas, la variance sous P_θ d'un estimateur $\delta(X)$, qui est alors égale au risque fréquentiste $R_\delta(\theta) = \mathbb{E}\big(L(\theta, \delta(X))|\theta\big)$, est minorée par la borne de Cramer-Rao. En présence d'une statistique exhaustive et complète, le théorème de Lehmann-Scheffé garantit l'existence d'une règle de décision uniformément meilleure au sens du

risque fréquentiste pour la perte quadratique, parmi les estimateurs sans biais.

Définition. Une règle de décision δ est dite *admissible* s'il n'existe pas de règle de décision qui lui soit préférable.

Le critère minimax consiste à choisir une règle de décision de risque minimal dans le pire des cas. Précisément, une règle de décision δ_0 est dite minimax si

$$\forall \delta \in \mathcal{D} \ , \ \max_{\theta \in \Theta} R_{\delta_0}(\theta) \leq \max_{\theta \in \Theta} R_\delta(\theta).$$

On a donc $\delta_0 \in \arg\min_{\delta \in \mathcal{D}} \max_{\theta \in \Theta} R_\delta(\theta)$, ce qui justifie l'appellation minimax. L'inconvénient majeur de ce critère est qu'il a tendance à favoriser les règles de décision trop prudentes, qui peuvent n'être adaptées qu'à des valeurs extrêmes ou peu représentées du paramètre. L'alternative Bayésienne au critère minimax est de chercher à minimiser le risque fréquentiste intégré selon la loi a priori π, de manière à accorder plus d'importance aux valeurs plus probables du paramètre. Précisément, le risque intégré d'une règle de décision δ associé à une loi a priori π et une fonction de perte L est défini par

$$r(\delta) = \int_\Theta R_\delta(\theta) d\pi(\theta) = \mathbb{E}\big(\mathbb{E}(L(\theta, \delta(X))|\theta)\big) = \mathbb{E}\big(L(\theta, \delta(X))\big) \in [0, +\infty].$$

La dépendance en π et L est omise pour ne pas alourdir les notations.

Définition. Une règle de décision Bayésienne pour l'a priori π est une règle de décision δ_π telle que

$$r(\delta_\pi) = \min_{\delta \in \mathcal{D}} r(\delta) < +\infty.$$

Si une règle de décision Bayésienne δ_π existe, l'estimateur associé $\delta_\pi(X)$ est appelé estimateur Bayésien. Un estimateur Bayésien est constructible à partir des données puisqu'il s'obtient en minimisant l'espérance conditionnelle $\eta \mapsto \mathbb{E}\big(L(\theta, \eta)|X\big)$, qui ne dépend que de quantités connues.

Théorème 2.1. *Si $r(\delta) < \infty$ pour au moins un $\delta \in \mathcal{D}$, alors toute règle de décision δ_π telle que*

$$\delta_\pi(X) \in \arg\min_{\eta \in \Theta} \ \mathbb{E}\big(L(\theta, \eta)|X\big)$$

est une règle de décision Bayésienne.

Preuve. Pour toute règle δ telle que $r(\delta) < \infty$, $r(\delta) = \mathbb{E}\big(L(\theta, \delta(X))\big) = \mathbb{E}\big(\mathbb{E}(L(\theta, \delta(X))|X)\big)$. Conditionnellement à X, $\delta(X)$ est déterministe, d'où

$$r(\delta) \geq \mathbb{E}\Big(\min_{\eta \in \Theta} \mathbb{E}(L(\theta, \eta)|X)\Big) = \mathbb{E}\big(\mathbb{E}\big(L(\theta, \delta_\pi(X))|X\big)\big) = \mathbb{E}\big(L(\theta, \delta_\pi(X))\big) = r(\delta_\pi). \qquad \square$$

Réciproquement, si l'ensemble $\arg\min_{\eta \in \Theta} \ \mathbb{E}\big(L(\theta, \eta)|X\big)$ est non-vide presque sûrement, alors une règle de décision Bayésienne vérifie nécessairement

$$\mathbb{P}\Big(\delta_\pi(X) \in \arg\min_{\eta \in \Theta} \ \mathbb{E}\big(L(\theta, \eta)|X\big)\Big) = 1.$$

Ce résultat (admis) montre simplement qu'une règle de décision optimale au sens du risque intégré doit être optimale pour (presque) toute valeur de X.

Remarque. Pour que la règle de décision Baysienne soit unique (à égalité presque sûre près), il suffit que la fonction $\eta \mapsto \mathbb{E}\big(L(\theta,\eta)|X\big)$ soit convexe et coercive. Sous des conditions de régularité sur la vraisemblance et l'a priori, ces hypothèses sont vérifiées dès que la fonction de perte $L(.,.)$ est elle-même convexe et coercive en son deuxième argument.

Exemples:

1. *Perte quadratique:* $L(\theta,\eta) = \|\theta-\eta\|^2$. On choisit l'a priori π tel que $\int \|\theta\|^2 d\pi(\theta) = \mathbb{E}\|\theta\|^2 < \infty$.
   - Le risque fréquentiste est l'erreur quadratique moyenne $R_\delta(\theta) = \mathbb{E}\big(\|\theta-\delta(X)\|^2\big|\theta\big)$.
   - Le risque intégré $r(\delta) = \int_\Theta \mathbb{E}\big(\|\theta-\delta(X)\|^2\big|\theta\big)d\pi(\theta) = \mathbb{E}\big(\|\theta-\delta(X)\|^2\big)$.
   - L'estimateur Bayésien est le minimiseur de l'erreur quadratique conditionnelle $\eta \mapsto \mathbb{E}\big(\|\theta-\eta\|^2|X\big)$, qui n'est autre que l'espérance conditionnelle $\delta_\pi(X) = \mathbb{E}(\theta|X)$.
2. *Perte* $\mathbb{L}^1$*:* $L(\theta,\eta) = \|\theta-\eta\|_1 = \sum_j |\theta_j - \eta_j|$. Sous l'hypothèse $\int \|\theta\|_1 d\pi(\theta) = \mathbb{E}\|\theta\|_1 < \infty$,
   - Le risque fréquentiste est $R_\delta(\theta) = \mathbb{E}\big(\|\theta-\delta(X)\|_1\big|\theta\big)$.
   - Le risque intégré $r(\delta) = \int_\Theta \mathbb{E}\big(\|\theta-\delta(X)\|_1\big|\theta\big)d\pi(\theta) = \mathbb{E}\|\theta-\delta(X)\|_1$.
   - L'estimateur Bayésien minimise $\eta \mapsto \mathbb{E}\big(\|\theta-\delta(X)\|_1\big|X\big)$. En dimension 1, un minimiseur est la médiane sous la loi conditionnelle
   $$\delta_\pi(X) = \mathrm{med}(\theta|X).$$
3. *Perte* $0-1$*:* $L(\theta,\eta) = \mathbb{1}\{\theta \neq \delta(X)\}$. Cette fonction de perte n'est pas convexe et n'a d'intérêt que si l'ensemble des valeurs de θ est discret comme en classification ou pour des tests. Dans ce cas,
   - Le risque fréquentiste est la probabilité d'erreur sous θ: $R_\delta(\theta) = \mathbb{P}\big(\delta(X) \neq \theta|\theta\big)$.
   - Le risque intégré: $r(\delta) = \mathbb{P}(\delta(X) \neq \theta)$.
   - L'estimateur Bayésien est le mode de la distribution de θ conditionnellement à X (qui n'est pas forcément unique). En effet
   $$\mathbb{E}\big(\mathbb{1}\{\theta \neq \eta\}|X\big) = \mathbb{P}(\theta \neq \eta|X) = 1 - \mathbb{P}(\theta = \eta|X)$$
   est minimal lorsque $\mathbb{P}(\theta = \eta|X)$ est maximal.

2.2 Propriétés de l'estimateur Bayésien

On s'intéresse à des conditions sous lesquelles un estimateur Bayésien est admissible. On supposera sans perte de généralité qu'il existe une règle de décision δ de risque intégré fini: $r(\delta) < \infty$. Rappelons qu'une règle de décision Bayésienne δ_π minimise le risque intégré

$$r(\delta) = \int_\Theta R_\delta(\theta)d\pi(\theta)\ ,\ \delta \in \mathcal{D}.$$

Proposition 2.2. *Soit π un a priori et L une fonction de coût, alors*

i) Si l'estimateur Bayésien est unique, il est admissible.

ii) Tout estimateur Bayésien est admissible π-presque-partout dans le sens où, pour tout $\delta \in \mathcal{D}$ préférable à δ_π,

$$\pi\big(\theta : R_\delta(\theta) < R_{\delta_\pi}(\theta)\big) = 0.$$

iii) Si π admet une densité strictement positive sur Θ et que $\theta \mapsto R_\delta(\theta)$ est continue pour tout $\delta \in \mathcal{D}$, alors toute règle de décision Bayésienne est admissible.

Preuve. Soit δ_π une règle de décision Bayésienne. On procède par l'absurde en supposant qu'il existe une règle de décision δ_0 préférable à δ_π avec en particulier $R_{\delta_0}(\theta) \leq R_{\delta_\pi}(\theta)$ pour tout θ. On obtient en intégrant selon π

$$r(\delta_0) = \int_\Theta R_{\delta_0}(\theta) d\pi(\theta) \leq \int_\Theta R_{\delta_\pi}(\theta) d\pi(\theta).$$

Comme $r(\delta_\pi)$ est minimal, $r(\delta_0) = r(\delta_\pi)$ et δ_0 est également une règle de décision Bayésienne, ce qui prouve $i)$. Soit $A = \{\theta : R_{\delta_0}(\theta) < R_{\delta_\pi}(\theta)\}$, on a

$$0 = r(\delta_0) - r(\delta_\pi) = \int_\Theta \big(R_{\delta_0}(\theta) - R_{\delta_\pi}(\theta)\big) d\pi(\theta) = \int_A \big(R_{\delta_0}(\theta) - R_{\delta_\pi}(\theta)\big) d\pi(\theta),$$

ce qui prouve $ii)$. Enfin, supposons $\theta \mapsto R_\delta(\theta)$ continue pour tout $\delta \in \mathcal{D}$ et $\pi > 0$. Pour tout θ tel que $R_{\delta_0}(\theta) < R_{\delta_\pi}(\theta)$, il existe un voisinage A de θ sur lequel $R_{\delta_0} < R_{\delta_\pi}$ par un argument de continuité. Comme $\pi > 0$, $\pi(A) > 0$ ce qui contredit $ii)$. □

Un estimateur Bayésien très utilisé est la moyenne a posteriori $\widehat{\theta}_{bay} = \mathbb{E}(\theta|X)$, associé au coût quadratique. Celui-ci vérifie les propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans le cadre fréquentiste où $X_1, ..., X_n$ sont indépendants de densités f_{θ_0}, $\theta_0 \in \Theta$ dans un modèle paramétrique régulier $f_\theta, \theta \in \Theta$. En effet, par le changement de variable $\theta = \widehat{\theta}_{MV} + t/\sqrt{n}$,

$$\widehat{\theta}_{bay} = \int \theta \pi(\theta|X) d\theta = \widehat{\theta}_{MV} + \frac{1}{\sqrt{n}} \int \frac{t}{\sqrt{n}} \pi\Big(\widehat{\theta}_{MV} + \frac{t}{\sqrt{n}} \big| X\Big) dt.$$

On a donc

$$\sqrt{n}\big(\widehat{\theta}_{bay} - \widehat{\theta}_{MV}\big) = \int \frac{t}{\sqrt{n}} \pi\Big(\widehat{\theta}_{MV} + \frac{t}{\sqrt{n}} \big| X\Big) dt.$$

D'après le théorème de Bernstein-von Mises, on s'attend donc à ce que $\sqrt{n}\big(\widehat{\theta}_{bay} - \widehat{\theta}_{MV}\big)$ converge vers 0, l'espérance de la loi normale $\mathcal{N}\big(0, \mathcal{I}(\theta_0)^{-1}\big)$. Pour cela, il faut des conditions de régularité supplémentaires garantissant la convergence du moment d'ordre 1 (par exemple par convergence dominée).

L'estimateur du maximum de vraisemblance $\widehat{\theta}_{MV}$ converge à vitesse $\sqrt{n}$ vers la vraie valeur θ_0 du paramètre. Lorsque l'écart entre $\widehat{\theta}_{bay}$ avec celui-ci est négligeable asymptotiquement, il hérite de ses propriétés, la plus importante étant l'efficacité asymptotique:

$$\sqrt{n}\big(\widehat{\theta}_{bay} - \theta_0\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0, \mathcal{I}(\theta_0)^{-1}\big).$$

Il suffit d'écrire

$$\sqrt{n}\big(\widehat{\theta}_{bay} - \theta_0\big) = \sqrt{n}\big(\widehat{\theta}_{bay} - \widehat{\theta}_{MV}\big) + \sqrt{n}\big(\widehat{\theta}_{MV} - \theta_0\big) = \sqrt{n}\big(\widehat{\theta}_{MV} - \theta_0\big) + o_P(1),$$

et on conclut par le lemme du Slutsky.

2.3 Maximum a posteriori

Définition. On appelle estimateur du maximum a posteriori (MAP) tout maximiseur de la fonction $\theta \mapsto \pi(\theta|X)$.

L'estimateur du maximum a posteriori est défini par rapport à une mesure de référence qui sera en général la mesure de Lebesgue dans le cas continu ou la mesure de comptage dans le cas discret. Il correspond à la valeur la plus probable selon la loi a posteriori, c'est donc le pendant Bayésien du maximum de vraisemblance. Il a l'avantage de ne pas dépendre d'une fonction de perte mais présente les mêmes inconvénients que le maximum de vraisemblance: l'unicité ou même l'existence n'est pas garantie, il n'a pas forcément de forme analytique simple etc...

Remarques.

1. Pour l'a priori uniforme sur Θ, l'estimateur du maximum a posteriori n'est autre que l'estimateur du maximum de vraisemblance.
2. Plus généralement, pour un a priori quelconque π, l'estimateur du maximum a posteriori est l'estimateur du maximum de vraisemblance dans le cadre fréquentiste où les X_i sont tirés selon θ, lui-même tiré aléatoirement sous la loi π.

Contrairement à l'estimateur du maximum de vraisemblance, l'estimateur MAP n'est pas invariant par reparamétrisation. C'est dû au fait que transformer le paramétre modifie l'a priori: soit ϕ un $\mathcal{C}^1$-difféomorphisme sur Θ, si θ est de densité π alors $\eta = \phi(\theta)$ a pour densité $\pi(\phi^{-1}(\eta))/|\phi' \circ \phi^{-1}(\eta)|$.

Dans le cas de X_i iid de densité f_{θ_0}, l'estimateur MAP se comporte comme l'estimateur du maximum de vraisemblance asymptotiquement, et ce pour tout choix raisonnable d'a priori. On s'en convainc facilement du fait que maximiser $\pi(\theta|X)$ en θ revient à maximiser $f_X(X|\theta)\pi(\theta)$. Or,

$$\log\big(f_X(X|\theta)\pi(\theta)\big) = \log\Big(\prod_{i=1}^n f_\theta(X_i)\pi(\theta)\Big) = \sum_{i=1}^n \log(f_\theta(X_i)) + \log(\pi(\theta)) = v(X,\theta) + \log(\pi(\theta))$$

où $v(X,\theta)$ est la log-vraisemblance. Le terme $\log(\pi(\theta))$ ne dépend pas de n alors que $v(X,\theta)$ croit à la même vistesse que n sous des conditions raisonnables. Ainsi, l'apport de l'a priori est négligeable asymptotiquement (du même ordre que l'apport d'une observation!) en ce qui concerne la différence entre le logarithme de la densité a posteriori et la log-vraisemblance.

Remarque. Pour que l'estimateur MAP converge vers la vraie valeur θ_0, il faut que (cf. le calcul précédent)

- le support de π contienne la vraie valeur du paramètre: $\pi(\theta_0) > 0$
- π soit bornée sur Θ.

2.4 Prévision

Dans cette partie, on s'intéresse à la prévision d'une nouvelle observation X_{n+1} à partir des observations passées $X_1, ..., X_n$. On notera ici $X^{(n)} = (X_1, ..., X_n)$ (et non plus X) l'échantillon d'apprentissage pour éviter les ambiguités. Dans cette partie, la suite $X_1, ..., X_n$ n'est pas forcément

iid mais on supposera toujours sa loi dépendante d'un paramètre θ inconnu, sur lequel on définit un a priori π. On rappelle que l'espérance conditonnelle de X_{n+1} sachant le passé $X^{(n)} = (X_1, ..., X_n)$ est le meilleur prédicteur au sens $\mathbb{L}^2$. Dans le cadre fréquentiste, la meilleur prédiction

$$\widehat{X}_{n+1}(\theta) := \mathbb{E}_\theta\big(X_{n+1}|X^{(n)}\big) = \mathbb{E}\big(X_{n+1}\big|\theta, X^{(n)}\big)$$

dépend de la vraie valeur du paramètre. Celle-ci étant inconnue en pratique, la prévision se fait en la remplaçant par un estimateur $\widehat{\theta}$. La valeur obtenue $\widehat{X}_{n+1}\big(\widehat{\theta}\big)$ perd alors la propriété d'optimalité.

L'approche Bayésienne permet de contourner ce problème en définissant l'espérance conditionnelle dans le modèle où θ est aléatoire de loi π. La prévision Bayésienne vaut alors

$$\widehat{X}^{bay}_{n+1} := \mathbb{E}\big(X_{n+1}|X^{(n)}\big)$$

où l'espérance ici tient compte de l'aléa de θ.

Proposition 2.3. *Soit $\pi\big(\theta|X^{(n)}\big)$ la loi a posteriori, le prédicteur Bayésien de X_{n+1} est donné par*

$$\widehat{X}^{bay}_{n+1} = \mathbb{E}\big(\widehat{X}_{n+1}(\theta)|X^{(n)}\big) = \int \widehat{X}_{n+1}(\theta)\pi\big(\theta|X^{(n)}\big)d\theta$$

où $\widehat{X}_{n+1}(\theta)$ est le meilleur prédicteur fréquentiste sous θ.

Preuve. Il suffit de remarquer que

$$\mathbb{E}\big(X_{n+1}|X^{(n)}\big) = \mathbb{E}\Big(\mathbb{E}\big(X_{n+1}|\theta, X^{(n)}\big)\Big|X^{(n)}\Big). \qquad \square$$

Remarque. Sous des conditions de régularités, les propriétés asymptotiques fréquentistes du prédicteur Bayésien sont les mêmes que pour l'approche du maximum de vraisemblance. C'est une conséquence directe du théorème de Bernstein-von Mises: si la fonction $\theta \mapsto \widehat{X}_{n+1}(\theta)$ est suffisamment régulière,

$$\widehat{X}^{bay}_{n+1} = \int \widehat{X}_{n+1}(\theta)\pi(\theta|X^{(n)})d\theta = \int \widehat{X}_{n+1}\Big(\widehat{\theta}_{MV} + \frac{t}{\sqrt{n}}\Big)\ \frac{1}{\sqrt{n}}\pi\Big(\widehat{\theta}_{MV} + \frac{t}{\sqrt{n}}\Big|X^{(n)}\Big)dt \underset{n\to\infty}{\approx} \widehat{X}_{n+1}\big(\widehat{\theta}_{MV}\big).$$

On peut également s'intéresser plus généralement à la loi de X_{n+1} conditionnellement au passé $X_1, ..., X_n$, appelée *loi prédictive*. Celle-ci permet par exemple de construire des intervalles de prévision Bayésiens.

Proposition 2.4. *Pour tout k, on note $f_{X^{(k)}}(.|\theta)$ la densité de $X^{(k)} = (X_1, ..., X_k)$ sous θ. La loi prédictive de X_{n+1} sachant $X^{(n)}$ a pour densité*

$$f_{X_{n+1}}\big(x|X^{(n)}\big) = \int \frac{f_{X^{(n+1)}}\big(X^{(n)}, x|\theta\big)}{f_{X^{(n)}}\big(X^{(n)}|\theta\big)}\pi\big(\theta|X^{(n)}\big)d\theta.$$

Preuve. Par la formule de Bayes sur les densités,

$$\begin{aligned} f_{X_{n+1}}\big(x|X^{(n)}\big) &= \int f_{\theta,X_{n+1}}\big(\theta, x|X^{(n)}\big)d\theta \\ &= \int f_{X_{n+1}}\big(x|\theta, X^{(n)}\big)\pi\big(\theta|X^{(n)}\big)d\theta \\ &= \int \frac{f_{X^{(n+1)}}(X^{(n)}, x|\theta)}{f_{X^{(n)}}(X^{(n)}|\theta)}\pi\big(\theta|X^{(n)}\big)d\theta. \qquad \square \end{aligned}$$

La loi prédictive peut s'interpréter comme l'espérance a posteriori de la loi prédictive "fréquentiste" de densité

$$f_{X_{n+1}}\big(x\big|\theta, X^{(n)}\big) = \frac{f_{X^{(n+1)}}(X^{(n)}, x|\theta)}{f_{X^{(n)}}(X^{(n)}|\theta)}.$$

Exemples.

1. Soit $X^{(n)} = (X_1, ..., X_n)$ iid de densité f_θ d'espérance $m(\theta) = \int x f_\theta(x)dx$ et $\pi(.)$ la densité a priori. Par indépendance des X_i conditionnellement à θ, la meilleure prévision de X_{n+1} au sens $\mathbb{L}^2$ vaut
$$\widehat{X}_{n+1}(\theta) = \mathbb{E}\big(X_{n+1}\big|\theta, X^{(n)}\big) = \mathbb{E}\big(X_{n+1}\big|\theta\big) = m(\theta).$$
Le prédicteur Bayésien est alors donné par
$$\widehat{X}^{bay}_{n+1} = \mathbb{E}\big(\widehat{X}_{n+1}(\theta)|X^{(n)}\big) = \int m(\theta)\pi\big(\theta|X^{(n)}\big)d\theta.$$
Comme $f_{X^{(n)}}(x_1, ..., x_n|\theta) = \prod_{i=1}^n f_\theta(x_i)$, on déduit de la proposition précédente la densité de la loi prédictive :
$$f_{X_{n+1}}\big(x|X^{(n)}\big) = \int \frac{f_{X^{(n+1)}}\big(X^{(n)}, x|\theta\big)}{f_{X^{(n)}}\big(X^{(n)}|\theta\big)}\pi\big(\theta|X^{(n)}\big)d\theta = \int f_\theta(x)\pi\big(\theta|X^{(n)}\big)d\theta = \mathbb{E}\big(f_\theta(x)|X^{(n)}\big).$$

2. Soit le modèle de régression
$$Y_i = a + bx_i + \epsilon_i \ , i = 1, 2, ...$$
avec les x_i déterministes et les ϵ_i centrés iid (conditionnellement à a, b) de densité f_ϵ connue qui admet un moment d'ordre 2. On fixe un a priori $\pi(a, b)$ sur les paramètres inconnus du modèle. En connaissant x_{n+1}, le meilleur prédicteur fréquentiste au sens $\mathbb{L}^2$ est $\widehat{Y}_{n+1}(a, b) = a + bx_{n+1}$. On a donc en particulier
$$\widehat{Y}^{bay}_{n+1} = \mathbb{E}\big(\widehat{Y}_{n+1}(a, b)|Y^{(n)}\big) = \mathbb{E}\big(a + bx_{n+1}|Y^{(n)}\big) = \widehat{a}_{bay} + \widehat{b}_{bay}x_{n+1},$$
où $\widehat{a}_{bay} = \mathbb{E}\big(a|Y^{(n)}\big)$ et $\widehat{b}_{bay} = \mathbb{E}\big(a|Y^{(n)}\big)$ sont les estimateurs Bayésiens de a, b pour la perte $\mathbb{L}^2$. Par indépendance des observations Y_i conditionnellement à a, b, on sait que $f_{Y_{n+1}}(y|a, b, Y^{(n)}) = f_\epsilon(y - a - bx_{n+1})$. La densité de la loi prédictive s'écrit
$$f_{Y_{n+1}}(y|Y^{(n)}) = \int f_\epsilon(y - a - bx_{n+1})\pi(a, b|Y^{(n)})dadb.$$

3. Soit $X_1, X_2, ...$ un processus auto-régressif d'ordre 1
$$X_i = aX_{i-1} + \epsilon_i, i = 2, 3, ...$$
où les ϵ_i sont iid (conditionnellement à a) de loi $\mathcal{N}(0, \sigma^2)$ et indépendants du passé. On note $\pi(a, \sigma^2)$ la densité a priori. Ici, la meilleure fréquentiste de X_{n+1} est
$$\widehat{X}_{n+1}(a, \sigma^2) = aX_n,$$

qui ne dépend pas de σ^2. Le prédicteur Bayésien est alors donné par

$$\widehat{X}_{n+1}^{bay} = \mathbb{E}\big(\widehat{X}_{n+1}(a,\sigma^2)|X^{(n)}\big) = \widehat{a}_{bay}X_n.$$

Soit ϕ la densité de la loi normale standard, on sait que

$$f_{X_{n+1}}\big(x\big|a,\sigma^2,X^{(n)}\big) = \frac{1}{\sigma}\phi\Big(\frac{x-aX_n}{\sigma}\Big).$$

On déduit la loi prédictive

$$f_{X_{n+1}}\big(x\big|X^{(n)}\big) = \int \frac{1}{\sigma}\phi\Big(\frac{x-aX_n}{\sigma}\Big)\pi\big(a,\sigma^2\big|X^{(n)}\big)dad\sigma^2.$$

3 Choix de l'a priori

Une difficulté majeure du contexte Bayésien réside dans le choix de l'a priori qui décrit l'information disponible sur le paramètre **avant d'observer les données**. La question est alors de traduire cette information sous la forme d'une distribution sur θ. L'impact de l'a priori est d'autant plus fort que la quantité d'observations est faible et, dans la grande majorité des cas, l'a priori devient complètement négligeable asymptotiquement (cf. le théorème de Bernstein-von Mises). Ainsi le choix de l'a priori est primordial sur de petits échantillons, même si certains aspects sont plus importants que d'autres. Concrètement, les aspects les plus importants dans le choix de l'a priori sont:

- Le support qui doit impérativement contenir la vraie valeur du paramètre. Dans le doute, utiliser un a priori avec un support maximal est nécessaire.
- L'espérance qui traduit l'information sur la position du paramètre. La médiane peut jouer le même rôle si on choisit un a priori qui n'admet pas de moment d'ordre 1.
- La variance (ou écart inter-quartile ou autre indicateur de dispersion) qui traduit la confiance dans l'information disponible. Une forte variance, donc une distribution a priori dispersée, correspond à un taux de confiance faible. Au contraire, un a priori avec une variance très faible aura beaucoup d'incidence sur la distribution a posteriori pour les petits échantillons. Comme on peut s'y attendre, l'impact de l'information a priori est d'autant plus fort que la confiance en celle-ci est importante.

Si la position et la confiance en l'information sont les deux seuls indications disponibles sur le paramètre, il est tout à fait possible de choisir une loi a priori quelconque (dans la limite du raisonnable) calibrée par ses moments. Par exemple, un a priori Gaussien fait généralement l'affaire dans un espace de paramètre continu (un a priori uniforme est généralement moins adapté du fait de son support borné).

On distingue essentiellement deux types de lois a priori: les lois dites informatives qui prennent en compte de l'information extérieure (avis d'expert, historique) et les lois non-informatives dont le but est d'être le plus neutre possible et de ne privilégier aucune valeur particulière du paramètre. Il n'est pas rare qu'une loi non-informative corresponde à une approche fréquentiste classique (maximum de vraisemblance ou méthode des moments par exemple) alors qu'une loi informative s'en éloignera.

3.1 Lois a priori informatives

Les lois a priori informatives sont construites à partir d'un contexte extérieur. On considère plusieurs exemples.

1. *Loi a priori discrète.* On dispose de l'avis de k experts, avec chaque expert qui propose une valeur t_j, $j = 1, ..., k$ au paramètre θ. Soit π_j un réel positif mesurant la fiabilité de l'avis de l'expert j (si rien n'est connu sur la fiabilité, on peut simplement prendre $\pi_j = 1/k$), on considère la loi a priori discrète
$$\pi(\theta) = \sum_{j=1}^{k} \pi_j \delta_{t_j}(\theta)$$
où δ désigne la masse de Dirac. Si nécessaire, on peut renormaliser les π_j au préalable pour que π soit une mesure de probabilité, sans perte de généralité. Dans ce cas, la loi a posteriori est elle-aussi une loi discrète, de support $\{t_1, ..., t_k\}$, puisqu'elle est absolument continue par rapport à la loi a priori. En supposant que X admet une densité f_X conditionnellement à θ, la loi a posteriori se calcule explicitement
$$\mathbb{P}(\theta = t_j|X) = \pi(t_j|X) = \frac{f_X(X|t_j)\pi(t_j)}{\sum_{j=1}^{k} f_X(X|t_j)\pi(t_j)} \propto f_X(X|t_j)\pi(t_j).$$
Sauf si Θ est lui-même un ensemble discret, une loi a priori discrète n'est pas recommandée car le support de la loi a posteriori a peu de chance de contenir la "'vraie"' valeur du paramètre.

2. *Mélange de lois.* Une généralisation du cas précédent correspond au cas où chaque expert annonce une distribution représentant les valeurs probables de θ, plutôt qu'une valeur ponctuelle. Dans ce cas, l'a priori s'écrit sous la forme d'un mélange de lois $\pi = \sum_{j=1}^{k} \pi_j \tau_j$ où τ_j est la distribution proposée par le j-ième expert. Par exemple:
   - chaque expert propose un intervalle de valeurs probables pour le paramètre, auquel on associe comme loi τ_j la loi uniforme associée.
   - chaque expert donne une valeur moyenne et une incertitude (variance) pour θ et on associe la densité Gaussienne correspondante. La loi a priori obtenue est alors un mélange Gaussien.

   Le calcul de la loi a posteriori est le suivant
$$\pi(\theta|X) \propto f_X(X|\theta)\pi(\theta) \propto f_X(X|\theta) \sum_{j=1}^{k} \pi_j \tau_j(\theta) = \frac{\sum_{j=1}^{k} \pi_j f_X(X|\theta)\tau_j(\theta)}{\sum_{j=1}^{k} \pi_j \int_\Theta f_X(X|\theta)\tau_j(\theta)}$$

3. *Famille paramétrique.* Soit une famille paramétrique $\mathcal{F} = \{\pi_\lambda : \lambda \in \Lambda\}$ "adaptée" au modèle (voir par exemple la partie 3.2). On peut choisir l'a priori en se basant sur une information limitée sur le paramètre, par exemple:
   - Une valeur moyenne $\mathbb{E}(\theta) = \theta_0$. On choisit (si possible) λ tel que $\int_\Theta \theta d\pi_\lambda(\theta) = \theta_0$.
   - Une valeur moyenne θ_0 avec une marge d'erreur quantifiée par une variance σ^2. On choisit alors λ tel que
$$\int_\Theta \theta d\pi_\lambda(\theta) = \theta_0 \quad \text{et} \quad \int_\Theta \theta^2 d\pi_\lambda(\theta) - \theta_0^2 = \sigma^2.$$

- un intervalle de valeurs $[\theta_{\min}, \theta_{\max}]$ avec une probabilité d'erreur α. On choisit alors λ tel que
$$\int_{\theta_{\min}}^{\theta_{\max}} d\pi_\lambda(\theta) = 1 - \alpha.$$
- L'avis d'experts qui proposent des valeurs ponctuelles $t_1, ..., t_k$. On peut alors choisir l'a priori π_λ en fonction de ces valeurs. Une possibilité est d'estimer la valeur de λ comme si les t_j étaient issus d'un échantillon iid de loi π_λ. On construit alors un estimateur $\widehat{\lambda}$ (par la méthode des moments ou du maximum de vraismeblance par exemple) qui conduit à une loi a priori $\pi_{\widehat{\lambda}}$.

3.2 Familles conjuguées

Définition. On dit qu'une famille de loi $\mathcal{F}$ est une famille conjuguée (pour un modèle donné) si pour tout a priori $\pi \in \mathcal{F}$, la loi a posteriori $\pi(.|X)$ appartient également à $\mathcal{F}$, quelle que soit la valeur de X observée.

Une famille conjuguée est donc une famille stable par passage de l'a priori à l'a posteriori. L'avantage principal d'une famille conjuguée qui a une forme paramétrique simple $\mathcal{F} = \{\pi_\lambda : \lambda \in \Lambda\}$ est de faciliter les calculs. Elle permet aussi dans certains cas une actualisation rapide de la loi a priori si des données passées sont utilisées pour faire de l'inférence sur un nouvel échantillon.

Exemples. Vérifier que les familles de lois a priori sont conjuguées dans les modèles suivants.

1. X_i iid de loi normale $\mathcal{N}(\theta, 1)$ et $\mathcal{F}$ la famille des lois normales $\mathcal{N}(\lambda, \mu) : \mu > 0, \lambda \in \mathbb{R}$.
2. X_i iid de loi exponentielle $\mathcal{E}(\theta)$ et $\mathcal{F}$ la famille des lois gamma de densité
$$\gamma_{a,b}(\theta) = \frac{b^a}{\Gamma(a)} e^{-b\theta} \theta^{a-1} \mathbb{1}\{\theta > 0\} \ , \ a, b > 0.$$
3. X_i iid de loi uniforme sur $[0, \theta]$ et $\mathcal{F}$ l'ensemble des lois de Pareto de densité
$$p_{a,b}(\theta) \propto \frac{1}{\theta^a} \mathbb{1}\{\theta > b\} \ , \ a > 1, b > 0.$$
4. X_i iid de loi binomiale $\mathcal{B}(m, \theta)$ (avec m connu) et $\mathcal{F}$ l'ensemble des lois beta de densité
$$\beta_{a,b}(\theta) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1} \mathbb{1}\{\theta \in [0,1]\} \ , \ a, b > 0.$$

Plus généralement, on peut construire une famille conjuguée pour les modèles issus de familles exponentielles lorsque l'a priori porte sur le paramètre naturel. On considère une famille exponentielle $X \sim f_X(.|\theta) : \theta \in \Theta$ de paramètre naturel $\theta \in \Theta \subseteq \mathbb{R}^d$ avec
$$f_X(x|\theta) = e^{\langle \theta, T(x) \rangle} g(\theta) h(x)$$
où $T(X)$ est une statistique exhaustive et g, h des fonctions positives régulières.

Théorème 3.1. *Si $\Theta = \mathbb{R}^d$, la famille*

$$\mathcal{F} = \{\pi_{\lambda,\mu}(\theta) \propto e^{\langle\theta,\lambda\rangle} g(\theta)^\mu : \mu > 0, \lambda \in \mathbb{R}^d\}$$

est une famille conjuguée pour le modèle $f_X(.|\theta), \theta \in \Theta$.

Preuve. Immédiat en écrivant

$$f_X(X|\theta).\pi_{\lambda,\mu}(\theta) \propto e^{\langle\theta,T(X)\rangle} g(\theta).e^{\langle\theta,\lambda\rangle} g(\theta)^\mu \propto e^{\langle\theta,T(X)+\lambda\rangle} g(\theta)^{\mu+1} \propto \pi_{T(X)+\lambda,\mu+1}(\theta). \qquad \square$$

On peut agrandir une famille de lois conjuguées en considérant des mélanges, par exemple si la famille initiale est trop limitée.

Proposition 3.2. *Soit $\mathcal{F}$ une famille de lois conjuguées dans un modèle donné, alors la famille de lois de mélange*

$$\mathcal{M} = \Big\{ \sum_{j=1}^k p_j \pi_j : p_j \geq 0, \sum_{j=1}^k p_j = 1, \pi_j \in \mathcal{F} \Big\}$$

forme également une famille de loi conjuguée.

Preuve. Exercice.

3.3 Modèles hiérarchiques

Etant donnée une famille de lois a priori paramétrique $\mathcal{F} = \{\pi_\lambda : \lambda \in \Lambda\}$, une manière de donner plus de flexibilité à l'a priori est de fixer un nouvel a priori (noté abusivement $\pi(\lambda)$) sur le paramètre λ, qui est alors intégré au modèle. Dans un modèle Bayésien hiérarchique, ce paramètre est appelé *hyper-paramètre*. La loi π_λ s'interprète maintenant comme la loi de θ conditionnellement à λ, que l'on peut également noter $\pi(\theta|\lambda)$. On a alors par la formule de Bayes:

$$\pi(\theta,\lambda) = \pi(\theta|\lambda)\pi(\lambda) \quad \text{et} \quad \pi(\theta) = \int_\Lambda \pi(\theta,\lambda) d\lambda.$$

De plus, la vraisemblance du modèle étant inchangée, la loi de X conditionnellement à θ ne dépend pas de λ. Autrement dit, X est indépendant de λ contionnellement à θ. On a donc

$$f_X(X|\theta,\lambda) = f_X(X|\theta)$$

et

$$\begin{aligned} \pi(\theta,\lambda|X) &= f_X(X|\theta,\lambda)\pi(\theta,\lambda) = f_X(X|\theta)\pi(\theta|\lambda)\pi(\lambda) \\ \pi(\theta|X) &= \int_\Lambda \pi(\theta,\lambda|X) d\lambda \end{aligned}$$

On peut choisir de ne s'intéresser qu'au paramètre θ en ne regardant que la loi a posteriori $\pi(\theta|X)$ ou d'intégrer l'analyse sur λ en s'intéressant à $\pi(\theta,\lambda|X)$. Les modèles Bayésiens hiérarchiques s'utilisent notamment pour traduire un manque de confiance supplémentaire en l'a priori, comme illustré dans l'exemple suivant.

Soit $X_1, ..., X_n$ iid de loi normale $\mathcal{N}(\theta, 1)$. On prend comme a priori la famille de lois normales $\mathcal{N}(\lambda, 1)$ avec $\lambda \in \mathbb{R}$. Si on soupçonne que θ doit être proche de 1, on peut choisir naturellement $\lambda = 1$. Pour être plus "'prudent"', on peut construire un modèle hierarchique en posant un a priori sur λ, par exemple de loi normale $\mathcal{N}(1, 1)$. On a donc

$$X_i \sim \mathcal{N}(\theta, 1) \ , \ \theta \sim \mathcal{N}(\lambda, 1) \ \text{ et } \ \lambda \sim \mathcal{N}(1, 1).$$

Pour définir la loi a posteriori, rien ne change, c'est toujours la loi $\pi(\theta|X)$ de θ conditionnellement à X. Le calcul donne pour la loi de θ

$$\pi(\theta) = \int_{\mathbb{R}} \pi(\theta|\lambda)\pi(\lambda)d\lambda \propto \int_{\mathbb{R}} \exp\Big(-\frac{1}{2}\big((\theta-\lambda)^2 + (\lambda-1)^2\big)\Big)d\lambda \propto \exp\Big(-\frac{(\theta-1)^2}{4}\Big).$$

Le modèle hiérarchique sur λ revient à rajouter de l'incertitude sur l'a priori qui entraîne une variance plus grande ($\theta \sim \mathcal{N}(1, 2)$ au lieu de $\theta \sim \mathcal{N}(1, 1)$ sans modèle hiérarchique).

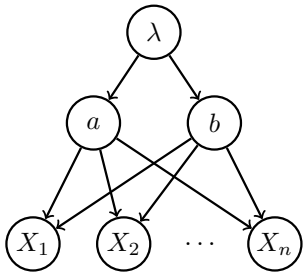
Un modèle hiérarchique est visualisé par un graphe dirigé acyclique (DAG en anglais) où les sommets sont les paramètres et observations du modèle et où les arêtes dirigées traduisent la dépendance. La construction d'un DAG peut se faire de bas en haut:

i) Les observations X_i sont placées sur la ligne du bas. Si celles-ci sont indépendantes conditionnellement aux paramètres du modèles, il n'y a pas d'arêtes entre elles. Sinon, on place des arêtes dirigées (dans un sens quelconque) entre les X_i, X_j dépendants conditionnellement aux paramètres.

ii) La ligne du dessus comprend tous les paramètres qui décrivent "directement"' la loi du vecteur X des observations. Autrement dit, on conserve l'ensemble le plus petit possible de paramètres qui suffisent à définir la loi de X. Par exemple, si $f_X(.|\theta, \lambda, \eta, ...) = f_X(.|\theta)$, seul θ est conservé. On place alors une arête dirigée qui part de ces paramètres vers les observations dont ils définissent la loi.

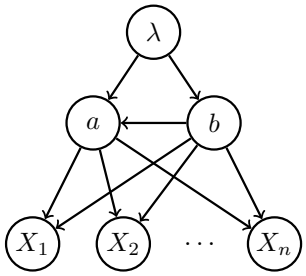
iii) On remonte en réitérant le procédé jusqu'à ce que tous les paramètres du modèles soient placés. Typiquement, en reprenant l'exemple suivant, si $\pi(\theta|\lambda, \eta, ...) = \pi(\theta|\lambda, \eta)$, on place λ et η sur la ligne du dessus. On réitère le procédé jusqu'à ce que tous les paramètres du modèles soient placés.

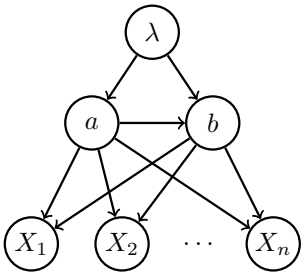
1. Soient $X_1, ..., X_n$ iid de loi gamma $\Gamma(a, b)$ avec $a, b > 0$. On prend comme lois a priori sur a, b les lois exponentielles $a \sim \mathcal{E}(\lambda)$ et $b \sim \mathcal{E}(1/\lambda)$ avec λ un hyperparamètre de loi $\mathcal{E}(1)$. Le modèle est représenté par le DAG



Ici, $f_X(x|a, b, \lambda) = f_X(x|a, b)$ d'où l'absence d'arête entre les X_i et λ.

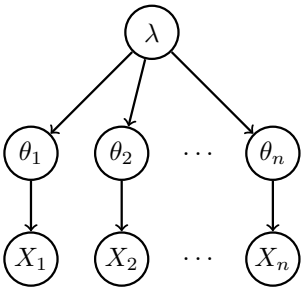
2. On reprend le modèle suivant avec comme seul changement, $b \sim \mathcal{E}(a/\lambda)$. Les hyperparamètres a et b ne sont donc plus indépendants conditionnellement à λ. Il y a alors deux représentations possibles pour le DAG, qui diffèrent seulement par l'orientation de la flèche entre les sommets a et b.





Dans les deux cas, on vérifie bien qu'il n'y a aucun cycle dirigé sur le graphe.

3. Soient $X_i, i = 1, ..., n$ indépendants de lois de Poisson $\mathcal{P}(\theta_i), \theta_i > 0$. Comme a priori, on considère les θ_i indépendants de loi exponentielle $\mathcal{E}(\lambda)$. Sur l'hyperparamètre $\lambda > 0$, on prend comme a priori la mesure uniforme sur $\mathbb{R}_+$. Le DAG du modèle est donné par



3.4 Approche non-informative

Si aucune information n'est disponible a priori, l'inférence Bayésienne permet paradoxalement de rechercher une approche neutre qui vise à utiliser seulement les données. L'intérêt par rapport à une approche fréquentiste est d'être capable de quantifier (et donc de limiter) un biais éventuel dans la méthode d'estimation.

Définition. L'*a priori de Laplace* est la mesure uniforme sur l'espace des paramètres Θ.

L'a priori de Laplace est un candidat naturel pour un a priori non-informatif, du fait de son invariance par translation. Il n'existe que si l'espace Θ du paramètre admet une mesure uniforme. C'est bien sûr le cas si Θ est un sous-ensemble (mesurable) de $\mathbb{R}^d$ d'intérieur non-vide, même si celui-ci n'est pas borné, auquel cas l'a priori de Laplace est impropre (sous condition d'existence de la loi a posteriori).

Exemples.

1. Soit $X_1, ..., X_n$ iid de loi de Poisson $\mathcal{P}(\theta)$, $\theta > 0$. L'a priori de Laplace est la mesure de Lebesgue sur $\mathbb{R}_+$ qui est une loi impropre. Soit $S = \sum_{i=1}^n X_i$, le calcul de la loi a posteriori donne
$$\pi(\theta|X) \propto f_X(X|\theta)\pi(\theta) \propto \frac{e^{-n\theta}}{\prod_{i=1}^n X_i!}\theta^S \mathbb{1}\{\theta > 0\} \propto e^{-n\theta}\theta^S,$$
où on retrouve une loi gamma $\Gamma\big(S+1, n\big)$.

2. L'a priori de Laplace sur la moyenne dans le modèle Gaussien $X_i \sim \mathcal{N}(\theta, 1)$ est la mesure de Lebesgue sur $\mathbb{R}$. La loi a posteriori vaut dans ce cas
$$\pi(\theta|X) \propto \exp\Big(-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2\Big) \propto \exp\Big(-\frac{n}{2}\big(\theta - \overline{X}\big)^2\Big).$$
On retrouve une loi normale $\mathcal{N}\big(\overline{X}, 1/n\big)$.

3. Dans le modèle de Bernoulli $X_i \sim \mathcal{B}(p)$, l'a priori de Laplace est la loi uniforme sur $]0, 1[$. Soit $S = \sum_{i=1}^n X_i$, on a
$$\pi(\theta|X) \propto p^S(1-p)^{n-S}\mathbb{1}\{p \in]0, 1[\},$$
où on reconnaît une loi beta $B\big(S+1, n-S+1\big)$.

Un défaut majeur de la mesure uniforme en tant qu'a priori non-informatif est qu'elle n'est pas invariante par reparamétrisation du modèle, et a donc un caractère arbitraire. Par exemple, dans le modèle Gaussien de moyenne nulle et de variance σ^2 inconnue, un a priori uniforme sur l'écart-type σ (la mesure de Lebesgue sur $\mathbb{R}_+$) ne correspond pas à un a priori uniforme sur la variance σ^2 (la mesure image par la transformation $x \mapsto x^2$ n'est pas uniforme):
$$\mathbb{1}\{\sigma > 0\}d\sigma = \mathbb{1}\{\sigma^2 > 0\}\frac{1}{2\sqrt{\sigma^2}}d\sigma^2 \propto \frac{1}{\sqrt{\sigma^2}}d\sigma^2.$$
Réciproquement, la mesure uniforme sur σ^2 n'est pas uniforme pour σ:
$$\mathbb{1}\{\sigma^2 > 0\}d\sigma^2 = \mathbb{1}\{\sigma > 0\}2\sigma d\sigma \propto \sigma d\sigma.$$

Définition. Soit $\{f_\theta, \theta \in \Theta \subseteq \mathbb{R}^d\}$ un modèle paramétrique régulier et $\mathcal{I}(\theta) \in \mathbb{R}^{d\times d}$ l'information de Fisher. On appelle *a priori non-informatif de Jeffreys* un a priori vérifiant
$$\pi(\theta) \propto \sqrt{\det\big(\mathcal{I}(\theta)\big)}\ ,\ \theta \in \Theta.$$
On distingue deux cas:

1. Si $\int_\Theta \sqrt{\det(\mathcal{I}(\theta))}d\theta = +\infty$, l'a priori de Jeyffreys est impropre. Il est alors nécessaire de vérifier la condition d'existence de la loi a posteriori, à savoir $\int f_X(X|\theta)\pi(\theta)d\theta < \infty$. Dans ce cas, l'a priori de Jeyffreys est défini à une constante multiplicative près.

2. Si $\int_\Theta \sqrt{\det(\mathcal{I}(\theta))}d\theta < +\infty$, l'a priori de Jeyffreys est une mesure de probabilité:
$$\pi(\theta) = \frac{\sqrt{\det(\mathcal{I}(\theta))}}{\int_\Theta \sqrt{\det(\mathcal{I}(\theta))}d\theta}\ ,\theta \in \Theta.$$

Un argument important en faveur de l'a priori de Jeyffreys comme a priori non-informatif est son invariance par reparamétrisation. On rappelle qu'un $\mathcal{C}^1$-diffeomorphisme $\phi : \mathrm{H} \to \Theta$ définit une reparamétrisation d'un modèle $\mathcal{M} = \{f_\theta : \theta \in \Theta\}$ en remarquant que $\mathcal{M} = \{f_{\phi(\eta)} : \eta \in \mathrm{H}\}$. L'information de Fisher en $\eta = \phi^{-1}(\theta)$ est alors donnée par

$$\mathcal{I}_\phi(\eta) := \nabla\phi(\eta)^\top \mathcal{I}(\phi(\eta))\nabla\phi(\eta),$$

où $\nabla\phi(.)$ est la matrice Jacobienne de ϕ.

Proposition 3.3. *Soit $\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ un modèle paramétrique régulier et π l'a priori de Jeyffreys. Soit $\phi : \Theta \to \mathrm{H}$ un $\mathcal{C}^1$-diffeomorphisme, la mesure image de π par l'application ϕ est l'a priori de Jeyffreys π_ϕ dans le modèle paramétré par $\eta = \phi(\theta)$:*

$$\pi_\phi(\eta) \propto \sqrt{\det(\mathcal{I}_\phi(\eta))}\ , \eta \in \mathrm{H}\,.$$

Dans le cas particulier où l'a priori de Jeyffreys π est une loi de probabilité (il n'est pas impropre), le résultat nous dit simplement que si θ est de loi π, alors $\eta = \phi(\theta)$ a pour loi π_ϕ.

Preuve. On sait que

$$\det\big(\mathcal{I}_\phi(\eta)\big) = \det\big(\nabla\phi(\theta)^\top \mathcal{I}(\theta)\nabla\phi(\theta)\big) = \det\big(\nabla\phi(\theta)\big)^2 \det\big(\mathcal{I}(\theta)\big) = J_\phi(\theta)^2 \det\big(\mathcal{I}(\theta)\big),$$

où $J_\phi(.) = \det(\nabla\phi(.))$ est le jacobien. Par la formule du changement de variable, la mesure image de π par l'application ϕ est donnée par

$$\pi_\phi(\eta) = \frac{\pi\big(\phi^{-1}(\eta)\big)}{|J_\phi(\phi^{-1}(\eta))|} \propto \frac{\sqrt{\det(\mathcal{I}(\phi^{-1}(\eta)))}}{|J_\phi(\phi^{-1}(\eta))|} \propto \sqrt{\det\big(\mathcal{I}_\phi(\eta)\big)}\ , \eta \in \mathrm{H}\,. \qquad \square$$

Remarque. L'a priori de Jeyffreys attribue un poids $\pi(\theta)$ d'autant plus fort que la valeur de θ est "facile" à estimer. Il privilégie en quelque sorte les valeurs favorables du paramètres, pour lesquelles les données apportent le plus d'information.

Définition. Soient f, g deux densités définies sur le même espace. La divergence de Kullback-Leibler de g par rapport à f est définie par

$$K(g\|f) := \int \log\left(\frac{g(t)}{f(t)}\right) g(t)dt \in [0, +\infty].$$

La divergence de Kullback-Leibler (parfois appelée entropie relative ou simplement divergence de Kullback) est une mesure de "proximité" entre mesures. Elle est très utilisée en théorie de l'information. On montre facilement par l'inégalité de Jensen sur $x \mapsto -\log(x)$ que $K(g\|f) \geq 0$ avec égalité si et seulement si f et g sont égales presque partout.

Pour une densité a priori π, on définit

$$G_n(\pi) := \mathbb{E}\big(K\big(\pi(.|X)\|\pi\big)\big),$$

où l'espérance est définie pour θ de loi π et $X = (X_1, ..., X_n)$ iid de densité f_θ conditionnellement à θ. L'a priori de Jeyffreys s'obtient en recherchant la densité π qui maximise une forme asymptotique de

$G_n(.)$. L'idée est de rendre l'a priori le plus objectif possible en cherchant à éloigner au maximum l'a posteriori, et ainsi maximiser l'influence des données. Cette caractérisation de l'a priori de Jeyffreys est basée sur le calcul suivant. On écrit

$$G_n(\pi) = \mathbb{E}\big(K\big(\pi(.|X)\|\pi\big)\big) = \mathbb{E}\Big(\mathbb{E}\big(K\big(\pi(.|X)\|\pi\big)\big|\theta\big)\Big)$$

et on remplace $\pi(.|X)$ par son équivalent asymptotique à θ fixé (qui ne dépend pas de π!) donné par le théorème de Bernstein-von Mises:

$$\pi(t|X) \underset{n\to\infty}{\approx} \frac{\sqrt{n^d \det(\mathcal{I}(\theta))}}{\sqrt{2\pi}^d} \exp\Big(-\frac{n}{2}\big(t-\widehat{\theta}_{MV}\big)^\top \mathcal{I}(\theta)\big(t-\widehat{\theta}_{MV}\big)\Big).$$

On obtient après simplifications

$$\begin{aligned}\mathbb{E}\big(K\big(\pi(.|X)\|\pi\big)\big|\theta\big) &= \mathbb{E}\Big[\int_\Theta \log\Big(\frac{\pi(t|X)}{\pi(t)}\Big)\pi(t|X)dt\Big|\theta\Big] \\ &\underset{n\to\infty}{\approx} \frac{d}{2}\log\Big(\frac{n}{2\pi}\Big) - \frac{1}{2} + \log\big(\sqrt{\det(\mathcal{I}(\theta))}\big) \\ &\quad - \mathbb{E}\Big[\int \log\Big(\pi\Big(\widehat{\theta}_{MV} + \frac{u}{\sqrt{n}}\Big)\Big)\frac{\sqrt{\det(\mathcal{I}(\theta))}}{\sqrt{2\pi}^d}\exp\Big(-\frac{u^\top\mathcal{I}(\theta)u}{2}\Big)du\Big|\theta\Big].\end{aligned}$$

Conditionnellement à θ, $\widehat{\theta}_{MV}$ converge en probabilité vers θ. Si on se restreint à des a priori continus, on a donc $\pi\big(\widehat{\theta}_{MV} + u/\sqrt{n}\big) = \pi(\theta) + o(1)$, et

$$\mathbb{E}\big(K\big(\pi(.|X)\|\pi\big)\big) \underset{n\to\infty}{\approx} \frac{d}{2}\log\Big(\frac{n}{2\pi}\Big) - \frac{1}{2} + \int_\Theta \log\Big(\frac{\sqrt{\det(\mathcal{I}(\theta))}}{\pi(\theta)}\Big)\pi(\theta)d\theta.$$

On reconnaît la divergence de Kullback (au signe près) de π par rapport à $\sqrt{\det(\mathcal{I}(\theta))}$. La forme asymptotique de $\mathbb{E}\big(K\big(\pi(.|X)\|\pi\big)\big)$ est donc maximale pour π l'a priori de Jeyffreys.

Un dernier argument (mais est-ce vraiment nécessaire?) qui souligne le caractère non-informatif de l'a priori de Jeyffreys est l'argument de concordance, décrit par la proposition suivante que l'on admettra.

Proposition 3.4. *Soit $\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ un modèle régulier. On note $I_\alpha^\pi = I_\alpha^\pi(X)$ un intervalle de crédibilité de niveau $1-\alpha \in]0,1[$ pour un a priori π. L'a priori de Jeyffreys est l'unique a priori "régulier" pour lequel la couverture fréquentiste de l'intervalle de crédibitlié vérifie*

$$\forall\theta \in \Theta\ ,\ \mathbb{P}_\theta\big(\theta \in I_\alpha^\pi\big) = 1 - \alpha + O\big(1/n\big).$$

Remarque. Sous les conditions de régularité suffisantes pour le théorème de Bernstein-von Mises, la loi a posteriori est équivalente asymptotiquement à une loi normale $\mathcal{N}\big(\widehat{\theta}_{MV}, \mathcal{I}(\theta)^{-1}/n\big)$. Les bornes d'un intervalle de crédibilité de niveau $1-\alpha$ se comportent donc asymptotiquement comme des quantiles Gaussiens, typiquement

$$I_\alpha^\pi = \Big[\widehat{\theta}_{MV} + q_\beta\frac{\sqrt{\mathcal{I}(\theta)^{-1}}}{\sqrt{n}}, \widehat{\theta}_{MV} + q_{1-\alpha+\beta}\frac{\sqrt{\mathcal{I}(\theta)^{-1}}}{\sqrt{n}}\Big],$$

pour un $\beta \in [0, \alpha]$, où $q_.$ désigne le quantile de la $\mathcal{N}(0,1)$. Comme $\widehat{\theta}_{MV} \sim \mathcal{N}(\theta, \mathcal{I}(\theta)^{-1}/n)$ sous θ asymptotiquement, on montre facilement que la couverture fréquentiste de l'intervalle de crédibilité vérifie

$$\mathbb{P}_\theta\big(\theta \in I_\alpha^\pi\big) = 1 - \alpha + O\big(1/\sqrt{n}\big).$$

En développant l'a posteriori $\pi(.|X)$ à l'ordre supérieur, on peut calculer le terme d'ordre $1/\sqrt{n}$ explicitement. On montre alors qu'il dépend du choix de l'a priori π et s'annule seulement pour $\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}$. L'a priori de Jeyffreys entraîne donc une concordance plus forte, de l'ordre de $1/n$, avec le taux de couverture fréquentiste des intervalles de crédibilité.

Exemples.

1. Soit le modèle exponentiel $X_1, ..., X_n$ iid de loi $\mathcal{E}(\theta)$ de densité $f_\theta(x) = \theta e^{-\theta x} \mathbb{1}\{x > 0\}$. On calcule l'information de Fisher

$$\mathcal{I}(\theta) = n \operatorname{var}_\theta \Big(\frac{\partial}{\partial\theta} \log\big(f_\theta(X_1|\theta)\big)\Big) = n \operatorname{var}_\theta(X_1) = \frac{n}{\theta^2}.$$

   L'a priori de Jeyffreys est donc l'a priori impropre $\pi(\theta) \propto 1/\theta$ pour $\theta > 0$. On note $S = \sum_{i=1}^n x_i$, la loi a posteriori est donnée par

$$\pi(\theta|X) \propto f_X(X|\theta)\pi(\theta) \propto \theta^{n-1} e^{-\theta S}$$

   où on reconnaît la loi $\Gamma(n, S)$. On remarque que l'estimateur Bayésien pour le coût $\mathbb{L}^2$, $\widehat{\theta}_{bay} = \mathbb{E}(\theta|X) = S/n$ est égal à l'estimateur du maximum de vraisemblance.

2. Soient des X_i indépendants de loi de Poisson $\mathcal{P}(\theta t_i)$ avec $t_i > 0$ connu et $\theta > 0$. En notant $t = \sum_{i=1}^n t_i$ et $S = \sum_{i=1}^n X_i$, la vraisemblance du modèle vérifie

$$f_X(X|\theta) \propto e^{-\theta t} \theta^S$$

   et l'information de Fisher

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\Big(\frac{\partial^2}{\partial\theta^2} \log\big(f_X(X|\theta)\big)\Big) = \mathbb{E}_\theta\Big(\frac{\partial^2}{\partial\theta^2}\big(\theta t - S\log(\theta)\big)\Big) = \frac{\mathbb{E}_\theta(S)}{\theta^2} = \frac{t}{\theta}.$$

   On déduit que l'a priori de Jeyffreys est la loi impropre $\pi(\theta) \propto 1/\sqrt{\theta}$ sur $\mathbb{R}_+^*$. La loi a posteriori est alors donnée par

$$\pi(\theta|X) \propto f_X(X|\theta)\frac{1}{\sqrt{\theta}} \propto e^{-\theta t}\theta^{S-1/2}\ ,\ \theta > 0$$

   où on reconnaît la loi $\Gamma(S + 1/2, t)$. Les estimateurs $\widehat{\theta}_{bay} = \mathbb{E}(\theta|X)$ et $\widehat{\theta}_{MAP}$, qui valent respectivement

$$\widehat{\theta}_{bay} = \frac{S + 1/2}{t} \quad \text{et} \quad \widehat{\theta}_{MAP} = \frac{S - 1/2}{t},$$

   encadrent l'estimateur du maximum de vraisemblance $\widehat{\theta}_{MV} = S/t$.

Malgré ses bonnes propriétés théoriques, l'a priori de Jeyffreys n'est satisfaisant en pratique que pour des paramètres θ de petite dimension voire même seulement en dimension 1. Ses limites apparaissent notamment dans le modèle de régression

$$Y_i = \langle x_i, \beta\rangle + \epsilon_i\ , i = 1, ..., n,$$

où les x_i sont des vecteurs déterministes de $\mathbb{R}^k$, $\beta \in \mathbb{R}^k$ est le paramètre à estimer et les ϵ_i sont iid de loi $\mathcal{N}(0, \sigma^2)$. Dans le cas où σ^2 est inconnu, l'information de Fisher en (β, σ^2) est donnée par

$$\mathcal{I}(\beta, \sigma^2) = \frac{n}{\sigma^2} \begin{bmatrix} \dfrac{X^\top X}{n} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ 0 \cdots 0 & 2/\sigma^2 \end{bmatrix}$$

où X est la matrice de régression de i-ème ligne x_i pour $i = 1, ..., n$. On en déduit l'a priori de Jeyffreys

$$\pi(\beta, \sigma^2) \propto \sqrt{\det\left(\mathcal{I}(\beta, \sigma^2)\right)} \propto \frac{1}{\sigma^{k+2}}.$$

On remarque alors trois choses.

- Les paramètres β et σ^2 sont indépendants ce qui est conforme à l'idée que, dans le modèle Gaussien, β et σ^2 peuvent être étudiés séparemment.
- La loi a priori non-informative sur β est uniforme. C'est là encore une propriété attendue compte tenu de l'invariance par translation du modèle Gaussien, où $X\beta$ est le paramètre de position.
- La loi de σ^2, proportionnelle à $1/\sigma^{k+2}$, privilégie les petites valeurs de σ^2, et ce d'autant plus que la dimension k du modèle est grande. Cet a priori est en contradiction avec des principes de bases de la régression linéaire. En effet, privilégier les valeurs proches de zéro accentue le biais de l'estimateur du maximum de vraisemblance et favorise dans le même temps le sur-ajustement.

Pour palier à ce problème, une solution en pratique (mais sans réelle justification théorique) est de prendre comme a priori le produit des a priori de Jeyffreys associés à chaque paramètre du modèle pris individuellement.

4 Régions de confiance et tests Bayésiens

L'approche Bayésienne permet facilement la construction d'intervalles et régions de confiance sur le paramètre θ par le biais de la loi a posteriori. Même si elle présente des similarités avec les intervalles de confiance "classiques", l'interprétation des régions de confiance Bayésienne est différente. Dans l'approche fréquentiste classique, une région de confiance de niveau $1 - \alpha \in]0, 1[$ est un sous-ensemble aléatoire $R_\alpha = R_\alpha(X) \subset \Theta$ vérifiant

$$\forall \theta \in \Theta \ , \ \mathbb{P}_\theta(\theta \in R_\alpha) = 1 - \alpha.$$

Ici, l'aléa provient seulement de l'échantillon X quand la vraie valeur du paramètre est θ (non-aléatoire). Imposer que l'égalité soit vraie pour tout θ est souvent contraignant, voire impossible. On peut alors choisir d'affaiblir la définition en tolérant des probabilités supérieures à $1 - \alpha$ ou se contenter d'un résultat asymptotique. L'approche Bayésienne permet de s'affranchir de ces inconvénients.

Définition. Une *région de crédibilité* de niveau $1-\alpha \in]0,1[$ est un sous-ensemble $R_\alpha = R_\alpha(X)$ de Θ tel que

$$\mathbb{P}(\theta \in R_\alpha | X) = \int_{R_\alpha} \pi(\theta|X) d\theta = 1 - \alpha.$$

Remarque. Une région de crédibilité de niveau $1-\alpha$ définit une zone qui, conditionnellemnt aux observations, contient le paramètre avec probabilité $1-\alpha$. Cette vision diffère fortement de la vision fréquentiste du fait que la probabilité de couverture est définie conditionnellement aux données. Dans le cadre fréquentiste, une région de confiance de niveau $1-\alpha$ contiendra le vrai paramètre une proportion $1-\alpha$ du temps, si l'expérience est répétée une infinité de fois. Contrairement au cadre Bayésien, la valeur observée de X n'intervient que dans le calcul de la région de confiance et pas dans la loi de probabilité considérée.

4.1 Régions HPD

Comme dans le cadre fréquentiste, on peut toujours définir une infinité de régions ou intervalles de confiance de niveau $1-\alpha$ donné. Le choix se porte alors souvent sur une région de volume minimal lorsque Θ est un sous-ensemble de $\mathbb{R}^d$. Comme nous le montrerons par la suite, les régions de crédibilité de volume minimal coïncident avec les régions où la densité a posteriori est maximale.

Définition. Une région *Highest Posterior Density (HPD)* de niveau $1-\alpha \in]0,1[$, notée R_α^{HPD}, est une région de crédibilité de la forme

$$R_\alpha^{HPD} = \{\theta : \pi(\theta|X) \geq h_\alpha\},$$

pour un seuil $h_\alpha > 0$ tel que $\mathbb{P}\big(\theta \in R_\alpha^{HPD}|X\big) = 1-\alpha$.

Dans cette définition, $\pi(.|X)$ est utilisée en tant que densité, sous-entendu par rapport à la mesure de Lebesgue. Cependant, le résultat reste vrai dans le cas discret et plus généralement pour une densité par rapport à une mesure de référence quelconque. C'est cette mesure de référence qui définit le volume (et donc la taille) d'une région de crédibilité.

Dans le cas continu, une unique région HPD existe pour tout niveau $1-\alpha \in]0,1[$ à condition que $\pi(.|X)$ soit continue presque partout. En revanche, dans le cas discret, certaines valeurs de α ne permettent pas l'existence de région HPD vérifiant l'égalité $\mathbb{P}\big(\theta \in R_\alpha^{HPD}|X\big) = 1-\alpha$. Trois options sont alors possibles:

1. Construire une région de crédibilité randomisée pour obtenir artificiellement une probabilité de couverture de $1-\alpha$.
2. Se contenter d'une probabilité de couverture supérieure ou égale à $1-\alpha$ en prenant le niveau inférieur le plus proche.
3. Choisir le niveau réalisable le plus proche, quitte à prendre un niveau supérieur qui correspond à un risque d'erreur plus élevé.

Proposition 4.1. *Une région HPD est de volume minimal parmi les régions de crédibilité de niveau égal.*

Ce résultat (que l'on admettra) est assez intuitif: pour avoir une probabilité de couverture la plus forte possible à taille fixée, il faut privilégier les zones les plus probables.

Proposition 4.2. *Si la densité a posteriori $\pi(.|X)$ est continue sur Θ et n'a qu'un seul maximum local, alors toute région HPD est connexe. En particulier, si $\Theta \subseteq \mathbb{R}$, les régions HPD sont des intervalles.*

Preuve. Soit $R_\alpha = \{\theta : \pi(\theta|X) \geq h_\alpha\}$ une région HPD de niveau $1-\alpha$ et C une composante connexe de R_α. On sait que C est borné (car $\pi(.|X)$ est une densité) et fermé (car $\pi(.|X)$ est continue), donc $\pi(.|X)$ atteint son maximum sur C en un point m_C qui est un maximum local. On déduit que C est l'unique composante connexe. □

Remarque. Les régions HPD ne sont pas invariantes par reparamétrisation du modèle. Comme pour l'estimateur du maximum a posteriori, une reparamétrisation induit un changement d'a priori qui se répercute sur l'a posteriori.

Le calcul des régions HPD est rarement explicite. Pour un niveau $1-\alpha$ donné, il s'agit de trouver le seuil h_α tel que

$$\mathbb{P}\big(\pi(\theta|X) \geq h_\alpha \big| X\big) := \int_\Theta \mathbb{1}\{\pi(\theta|X) \geq h_\alpha\}\pi(\theta|X)d\theta = 1-\alpha.$$

Si la densité a posteriori admet un unique maximum local dans $\mathbb{R}$, les régions HPD sont des intervalles dont les bornes sont des quantiles de la loi a posteriori. Même si ces quantiles ont des formes explicites, comme pour certaines familles de lois conjuguées, calculer analytiquement les quantiles optimaux n'est pas forcément simple. On peut néanmoins obtenir une approximation des régions HPD en utilisant le comportement asymptotiquement Gaussien de la densité a posteriori.

Proposition 4.3. *On suppose $\Theta \subseteq \mathbb{R}$. Sous les hypothèses du théorème de Bernstein-von Mises, l'intervalle*

$$I_\alpha := \left[\widehat{\theta}_{MAP} - q_{1-\alpha/2}\frac{\mathcal{I}(\widehat{\theta}_{MAP})^{-1/2}}{\sqrt{n}}, \widehat{\theta}_{MAP} + q_{1-\alpha/2}\frac{\mathcal{I}(\widehat{\theta}_{MAP})^{-1/2}}{\sqrt{n}}\right]$$

est un intervalle de crédibilité de niveau $1-\alpha$ asymptotiquement en probabilité: $\mathbb{P}(\theta \in I_\alpha|X) \xrightarrow[n\to\infty]{\mathbb{P}} 1-\alpha$.

Preuve. On sait que $|\widehat{\theta}_{MAP} - \widehat{\theta}_{MV}| = o_P(1/\sqrt{n})$ et $|\widehat{\theta}_{MAP} - \theta_0| = O_P(1/\sqrt{n})$. Par continuité de l'information de Fisher, on a

$$\sqrt{n\mathcal{I}(\widehat{\theta}_{MAP})}\big(\theta - \widehat{\theta}_{MAP}\big) = \sqrt{n\mathcal{I}(\theta_0)}\big(\theta - \widehat{\theta}_{MV}\big) + o_P(1).$$

Soit $\Pi(.|X)$ la fonction de répartition associée à $\pi(.|X)$. On obtient par le théorème de Bernstein-von Mises combiné avec la formule de Taylor-Lagrange :

$$\Pi(\theta|X) \underset{n\to\infty}{\approx} \Phi\Big(\sqrt{n\mathcal{I}(\theta_0)}\big(\theta - \widehat{\theta}_{MV}\big)\Big) \underset{n\to\infty}{\approx} \Phi\Big(\sqrt{n\mathcal{I}(\widehat{\theta}_{MAP})}\big(\theta - \widehat{\theta}_{MAP}\big)\Big)$$

où Φ est la fonction de répartition de la loi normale standard. On a alors

$$\begin{aligned}\mathbb{P}(\theta \in I_\alpha|X) &= \Pi\Big(\widehat{\theta}_{MAP} + q_{1-\alpha/2}\frac{\mathcal{I}(\widehat{\theta}_{MAP})^{-1/2}}{\sqrt{n}}\Big|X\Big) - \Pi\Big(\widehat{\theta}_{MAP} - q_{1-\alpha/2}\frac{\mathcal{I}(\widehat{\theta}_{MAP})^{-1/2}}{\sqrt{n}}\Big|X\Big) \\ &\underset{n\to\infty}{\approx} \Phi(q_{1-\alpha/2}) - \Phi(-q_{1-\alpha/2})\end{aligned}$$

ce qui conclut la preuve. $\square$

La normalité asymptotique de la densité a posteriori entraîne la "convergence" de la région HPD vers l'intervalle de confiance fréquentiste I_α. Plus exactement, on peut montrer que le volume de la différence symmétrique

$$\Delta(R_\alpha^{HPD}, I_\alpha) := \big(R_\alpha^{HPD} \cup I_\alpha\big) \setminus \big(R_\alpha^{HPD} \cap I_\alpha\big)$$

tend vers zéro en probabilité quand n tend vers l'infini. De plus, le seuil h_α de la région HDP se comporte asymptotiquement comme le seuil correspondant de niveau $1-\alpha$ de la loi normale $\mathcal{N}\big(0, \mathcal{I}(\theta)^{-1}/n\big)$:

$$h_\alpha \underset{n\to\infty}{\approx} \sqrt{n\mathcal{I}(\theta_0)}\phi\big(q_{1-\alpha/2}/\sqrt{n\mathcal{I}(\theta_0)}\big) \underset{n\to\infty}{\approx} \sqrt{\frac{n\mathcal{I}(\theta_0)}{2\pi}}.$$

Pour de grandes tailles d'échantillon, on peut donc approcher la région HPD de niveau $1-\alpha$ par l'intervalle de confiance fréquentiste estimé par l'estimateur du maximum a posteriori $\widehat{\theta}_{MAP}$. Le choix de cet estimateur est le plus adapté dans la logique d'estimer la région où la densité a posteriori est maximale, mais le résultat reste valable en prenant à la place l'espérance a posteriori ou le maximum de vraisemblance.

Un corollaire important est que la région HPD fournit un intervalle de confiance de niveau $1-\alpha$ asymptotiquement dans le cadre fréquentiste. C'est une conséquence directe du fait que R_α^{HPD} se rapproche de I_α quand $n \to \infty$ qui est lui-même un intervalle de confiance de niveau $1-\alpha$ asymptotiquement.

Corollaire 4.4. *Sous les hypothèses du théorème de Bernstein-von Mises,*

$$\forall \theta \in \Theta \ , \ \mathbb{P}_\theta\big(\theta \in R_\alpha^{HPD}\big) \underset{n\to\infty}{\longrightarrow} 1-\alpha.$$

Preuve. Sous θ, comme $|\widehat{\theta}_{MAP} - \widehat{\theta}_{MV}| = o_P(1/\sqrt{n})$, on a par le lemme de Slutsky

$$\sqrt{n\mathcal{I}(\widehat{\theta}_{MAP})}\big(\widehat{\theta}_{MAP} - \theta\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

D'où,

$$\mathbb{P}_\theta\big(\theta \in R_\alpha^{HPD}\big) \underset{n\to\infty}{\approx} \mathbb{P}_\theta\big(\theta \in I_\alpha\big) \underset{n\to\infty}{\longrightarrow} 1-\alpha.$$

$\square$

4.2 Intervalles de prévision

Un intervalle de prévision est un intervalle de confiance sur une nouvelle donnée X_{n+1} en ayant observé le passé $X^{(n)} = (X_1, ..., X_n)$. La version Bayésienne se construit directement à partir de la loi prédictive qui est la loi de X_{n+1} conditionnellement au passé $X^{(n)}$. On rappelle que dans le cas continu la densité prédictive s'obtient comme l'espérance a posteriori de la densité prédictive "fréquentiste" $f_{X_{n+1}}\big(x\big|\theta, X^{(n)}\big)$ définie pour une valeur de θ donnée:

$$f_{X_{n+1}}\big(x\big|X^{(n)}\big) = \mathbb{E}\Big(f_{X_{n+1}}\big(x|\theta, X^{(n)}\big)\big|X^{(n)}\Big).$$

Définition. On appelle intervalle de prévision Bayésien de niveau $1-\alpha \in]0,1[$ un intervalle $I_\alpha = I_\alpha\big(X^{(n)}\big)$ tel que

$$\mathbb{P}\big(X_{n+1} \in I_\alpha\big|X^{(n)}\big) = 1-\alpha.$$

L'approche Bayésienne comporte les mêmes avantages que pour les intervalles de confiance par rapport à son homologue fréquentiste:

1. La probabilité de $1-\alpha$ est atteinte exactement (modulo d'éventuelles approximations numériques). En comparaison, un intervalle de prévision fréquentiste I_α de niveau $1-\alpha$ doit vérifier

$$\forall \theta \in \Theta \ , \ \mathbb{P}_\theta\big(X_{n+1} \in I_\alpha\big) = 1-\alpha,$$

   qui nécessite d'exhiber une fonction pivotale de θ, c'est-à-dire une fonction (non-triviale) $g\big(\theta, X^{(n)}\big)$ dont la loi ne dépend pas de θ.

2. Si on ne sait pas construire un intervalle de prévision fréquentiste de niveau $1-\alpha$ indépendamment de θ, l'approche classique consiste à définir un intervalle de la forme $I_\alpha\big(\widehat{\theta}, X^{(n)}\big)$ où $\widehat{\theta}$ est un estimateur de θ. Remplacer le θ inconnu par un estimateur rajoute souvent de l'aléa qui est difficile à prendre en compte dans le calcul de la probabilité d'appartenance à l'intervalle. En général, on se contente d'un niveau $1-\alpha$ atteint seulement asymptotiquement.

Les bornes d'un intervalle de prévision s'obtiennent à partir des quantiles de la loi prédictive, que l'on notera $q_\beta^{(n)}$ pour $\beta \in]0,1[$.

Proposition 4.5. *Si la densité prédictive $f_{X_{n+1}}\big(.\big|X^{(n)}\big)$ est continue, alors pour tout $\beta \in]0,\alpha[$, l'intervalle*

$$I_\alpha^\beta = \big[q_\beta^{(n)}, q_{1-\alpha-\beta}^{(n)}\big]$$

est un intervalle de prévision de niveau $1-\alpha$.

Preuve. Immédiat.

Si on dispose d'un continuum d'intervalles de prévision $I_\alpha^\beta, \beta \in]0,\alpha[$, l'intervalle de taille minimale s'obtient en minimisant $\beta \mapsto q_{1-\alpha-\beta}^{(n)} - q_\beta^{(n)}$ sur $]0,\alpha[$. Il est rare que les quantiles de la loi a posteriori aient une forme explicite simple et même si c'est le cas, le β optimal est lui-même rarement explicite. On aura donc souvent recours à des méthodes numériques pour construire l'intervalle de prévision de taille minimale.

4.3 Facteur de Bayes

Soient $\mathcal{F}$ et $\mathcal{G}$ deux familles de lois munies respectivement d'a priori $\pi_\mathcal{F}$ et $\pi_\mathcal{G}$. Le facteur de Bayes permet de déterminer quelle modélisation est plus adaptée (ou plus vraisemblable) au vu des observations. Il est défini comme le rapport des lois marginales de l'échantillon évaluées en X sous les modèles Bayésiens $\{\mathcal{F}, \pi_\mathcal{F}\}$ et $\{\mathcal{G}, \pi_\mathcal{G}\}$:

$$B_{\mathcal{F}/\mathcal{G}} = \frac{\int_\mathcal{F} f(X) d\pi_\mathcal{F}(f)}{\int_\mathcal{G} g(X) d\pi_\mathcal{G}(g)}.$$

Dans le cas paramétrique avec $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ et $\mathcal{G} = \{g_\lambda : \lambda \in \Lambda\}$ où les ensembles Θ et Λ sont munis respectivement d'a priori π_Θ et π_Λ, le facteur de Bayes s'écrit

$$B_{\mathcal{F}/\mathcal{G}} = \frac{f_X(X)}{g_X(X)} = \frac{\int_\Theta f_\theta(X) d\pi_\Theta(\theta)}{\int_\Lambda g_\lambda(X) d\pi_\Lambda(\lambda)}.$$

Le facteur de Bayes sert pour la construction de tests Bayésiens. Typiquement, si $B_{\mathcal{F}/\mathcal{G}} > 1$, on conclura que le modèle Bayésien $\{\mathcal{F}, \pi_\mathcal{F}\}$ est le plus vraisemblable, ce qui s'apparente à un test d'adéquation sur l'hypothèse $\mathcal{H}_0$: "la loi des observations est issue de $\mathcal{F}$" contre $\mathcal{H}_1$: " la loi des

observations est issue de $\mathcal{G}$". Comme $B_{\mathcal{G}/\mathcal{F}} = 1/B_{\mathcal{F}/\mathcal{G}}$, la conclusion reste la même quelque soit le modèle considéré en premier, contrairement aux approches fréquentistes où la conclusion peut varier selon si on prend $\mathcal{H}_0$ ou $\mathcal{H}_1$ comme hypothèse nulle (en plus de dépendre du niveau $1-\alpha$). Pour les tests Bayésiens, la décision du test dépend donc ici autant de l'hypothèse nulle que de l'hypothèse alternative et pas seulement de la distribution de la statistique de test sous l'hypothèse nulle.

Du fait de la transitivité $B_{\mathcal{F}/\mathcal{H}} = B_{\mathcal{F}/\mathcal{G}} \times B_{\mathcal{G}/\mathcal{H}}$, la relation sous-jacente

$$\{\mathcal{F}, \pi_{\mathcal{F}}\} \geq \{\mathcal{G}, \pi_{\mathcal{G}}\} \iff B_{\mathcal{F}/\mathcal{G}} \geq 1$$

est un preordre total (deux modèles sont toujours comparables). Il est donc toujours possible de décider quel modèle est le plus vraisemblable parmi deux (ou plus) modèles Bayésiens, sans qu'il y ait d'ambiguité. Ce critère permet en particulier de comparer deux modèles de régression quelconque (pas nécessairement des modèles emboîtés) et donc de faire de la sélection de variables.

Une application classique concerne le test de l'hypothèse $\mathcal{H}_0 : \theta \in \Theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta_1$ où Θ_0, Θ_1 sont des sous-ensembles (disjoints) de Θ. Le modèle sous $\mathcal{H}_0$ correspond à l'a priori restreint à Θ_0:

$$\pi(\theta|\Theta_0) = \frac{\pi(\theta)\mathbb{1}\{\theta \in \Theta_0\}}{\pi(\Theta_0)} \ , \ \theta \in \Theta$$

et la même formule est valable pour $\mathcal{H}_1$. Le facteur de Bayes s'écrit donc

$$B_{\Theta_0/\Theta_1} = \frac{f_X^{(\mathcal{H}_0)}(X)}{f_X^{(\mathcal{H}_1)}(X)} = \frac{\int_{\Theta_0} f_X(X|\theta)d\pi(\theta)/\pi(\Theta_0)}{\int_{\Theta_1} f_X(X|\theta)d\pi(\theta)/\pi(\Theta_1)} = \frac{\int_{\Theta_0} f_X(X|\theta)d\pi(\theta)}{\int_{\Theta_1} f_X(X|\theta)d\pi(\theta)} \times \frac{\pi(\Theta_1)}{\pi(\Theta_0)}.$$

Comme $\pi(\theta|X) = f_X(X|\theta)\pi(\theta)/f_X(X)$, on a également

$$B_{\Theta_0/\Theta_1} = \frac{\pi(\Theta_0|X)}{\pi(\Theta_1|X)} \times \frac{\pi(\Theta_1)}{\pi(\Theta_0)}.$$

Le facteur de Bayes est donc le rapport entre l'odds ratio a posteriori $\pi(\Theta_0|X)/\pi(\Theta_1|X)$ et l'odds ratio a priori $\pi(\Theta_0)/\pi(\Theta_1)$. D'une certaine façon, il mesure l'influence de l'échantillon sur le changement des odds-ratios.

Remarque. La statistique B_{Θ_0/Θ_1} est le pendant Bayésien de la statistique du rapport de vraisemblance

$$T = \frac{\sup_{\theta\in\Theta_0} f_X(X|\theta)}{\sup_{\theta\in\Theta_1} f_X(X|\theta)}$$

pour lequel la vraisemblance est pondérée par l'a priori au lieu d'être maximisée.

Du point de vue de la théorie de la décision, le test Bayésien qui consiste à rejeter $\mathcal{H}_0$ si $B_{\Theta_0/\Theta_1} > 1$ correspond à une règle de décision binaire $\delta(X) \in \{0,1\}$ pour lequel le paramètre à estimer est la vraie réponse qui vaut $\eta = 1$ si $\theta \in \Theta_1$ et $\eta = 0$ sinon. La fonction de perte utilisée est la perte $0-1$: $L(\eta, \delta(X)) = \mathbb{1}\{\eta \neq \delta(X)\}$.

Cette approche ne marche que si les ensembles Θ_0 et Θ_1 ont des masses "comparables" pour l'a priori. Par exemple, le cas où $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \Theta \setminus \{\theta_0\}$ n'est pas adapté pour un a priori π

à densité puisqu'on aura toujours $B_{\Theta_0/\Theta_1} = 0$. Il faut donc être vigilant et bien tenir compte du poids important de l'a priori et des ensembles Θ_0 et Θ_1 pour ce critère.

Le test pour l'hypothèse simple $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta = \theta_1$ donne comme facteur de Bayes le rapport de vraisemblance

$$B_{\theta_0/\theta_1} = \frac{f_X(X|\theta_0)}{f_X(X|\theta_1)}.$$

qui ne dépend pas de l'a priori. La décision de rejeter ou non $\mathcal{H}_0$ revient à effectuer le test de Neymann-Pearson avec comme seuil 1.

5 Méthodes de Monte-Carlo

Il arrive souvent que certaines quantités liées à la loi a posteriori n'aient pas de forme analytique simple. On a alors recours à des méthodes numériques. Par exemple:

- La loi a posteriori $\pi(.|X)$ est souvent étudiée à une constante multiplicative près. Or, la constante de normalisation $\int_\Theta f_X(X|\theta)\pi(\theta)d\theta$ est souvent nécessaire pour déterminer les quantiles a posteriori qui permettent de construire des intervalles de crédibilité. Plus généralement, le calcul des régions HPD nécessite de connaître exactement la loi a posteriori, et non seulement à une constante multiplicative près.
- Si la loi a posteriori n'appartient pas à une famille commune de lois, il n'est pas rare que les estimateurs Bayésiens classiques comme l'espérance ou la médiane a posteriori n'aient pas de forme explicite. C'est encore plus fréquent si on utilise une fonction de perte $L(.,.)$ qui n'est pas "standard".
- Pour la prévision, le recours à des méthodes numérique est quasiment systématique. Même dans les cas standards, la loi loi prédictive, nécessaire pour déterminer des intervalles de prévision, n'a souvent pas de forme analytique simple.

Lorsqu'il n'y a pas de forme explicite, on peut approcher ces quantités numériquement à condition de pouvoir simuler sous la loi a posteriori (cette question fera l'objet du chapitre suivant). Si un échantillon $T_1, ..., T_N$ (indépendant ou non) généré sous la loi a posteriori $\pi(.|X)$ est disponible, les méthodes d'approximations numériques reposent alors sur les estimations empiriques correspondantes.

5.1 Echantillonage

Les méthodes de Monte-Carlo désignent un ensemble de méthodes probabilistes d'approximation numérique basées sur des estimations empiriques. Soit f une densité définie sur $\mathbb{R}^p$ et h une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$ telle que $\int |h(t)|f(t)dt < \infty$. On suppose qu'on sait simuler des variables aléatoires de densité f. Sous sa forme la plus simple, la méthode de Monte-Carlo permet d'approcher la quantité

$$\gamma = \int h(t)f(t)dt.$$

L'idée est simple: on simule un échantillon $T_1, ..., T_N$ de densité f (avec N le plus grand possible) et on construit

$$\widetilde{\gamma} := \frac{1}{N}\sum_{k=1}^{N} h(T_k).$$

Si les T_k sont iid, qui est le cas idéal, les propriétés pour N grand de l'approximation $\widetilde{\gamma}$ sont des conséquences directes de la loi forte des grands nombres et du Théorème Central Limite.

Proposition 5.1. *Soit $T_1, ..., T_N$ iid de densité f, l'approximation $\widetilde{\gamma}$ converge presque sûrement vers γ quand N tend vers l'infini. De plus, si $\tau^2 := \int h(t)^2 f(t)dt - \gamma^2 < \infty$, $\sqrt{N}\big(\widetilde{\gamma} - \gamma\big)$ converge en loi vers une loi normale $\mathcal{N}(0, \tau^2)$.*

La normalité asymptotique de $\widetilde{\gamma}$ permet de mesurer la précision de l'approximation par Monte-Carlo. En considérant $\widetilde{\tau}^2 := \frac{1}{N}\sum_{k=1}^{N} h(T_k)^2 - \widetilde{\gamma}^2$, on déduit, par le lemme de Slutsky, un intervalle de confiance sur γ de niveau $1-\alpha$ asymptotiquement

$$I_N := \left[\widetilde{\gamma} - q_{1-\frac{\alpha}{2}}\frac{\widetilde{\tau}}{\sqrt{N}}, \widetilde{\gamma} + q_{1-\frac{\alpha}{2}}\frac{\widetilde{\tau}}{\sqrt{N}}\right]$$

où q est le quantile de la loi normale standard. On remarque que cet intervalle de confiance se construit à partir du même échantillon $T_1, ..., T_n$ et ne nécessite pas de puissance de calcul supplémentaire, autre que le calcul de $\widetilde{\tau}$.

Applications. Soit un échantillon iid $\theta_1, ..., \theta_N$ simulé selon la densité a posteriori $\pi(.|X)$.

1. Sous réserve d'existence, l'estimateur Bayésien pour la perte quadratique $\widehat{\theta}_{bay} = \mathbb{E}(\theta|X)$ peut être approché par Monte-Carlo à partir de l'échantillon $\theta_1, ..., \theta_N$. Ici,

   $$\widetilde{\widehat{\theta}}_{bay} = \frac{1}{N}\sum_{k=1}^{N}\theta_k \xrightarrow[N\to\infty]{p.s.} \widehat{\theta}_{bay}.$$

   Si l'a priori admet un moment d'ordre deux: $\int \theta^2\pi(\theta)d\theta < \infty$, alors c'est également le cas pour la loi a posteriori et on peut déduire un intervalle de confiance sur $\widehat{\theta}_{bay}$ qui mesure la précision de l'approximation par Monte-Carlo. Précisemment, on sait que pour tout $\alpha \in]0, 1[$,

   $$\mathbb{P}\left(\widetilde{\widehat{\theta}}_{bay} - q_{1-\frac{\alpha}{2}}\frac{\widetilde{\tau}}{\sqrt{N}} \leq \widehat{\theta}_{bay} \leq \widetilde{\widehat{\theta}}_{bay} + q_{1-\frac{\alpha}{2}}\frac{\widetilde{\tau}}{\sqrt{N}}\right) \underset{n\to\infty}{\longrightarrow} 1-\alpha,$$

   où $\widetilde{\tau} = \frac{1}{N}\sum_{k=1}^{N}\theta_k^2 - \widetilde{\widehat{\theta}}_{bay}^2$ et q désigne le quantile de la loi normale standard.

2. Pour $\Theta \subseteq \mathbb{R}$, la fonction de répartition a posteriori est approchée par sa version empirique calculée sur l'échantillon de Monte-Carlo: $\widetilde{\Pi}(\theta|X) = \frac{1}{N}\sum_{k=1}^{N}\mathbb{1}\{\theta_k \leq \theta\}, \theta \in \Theta$. Les quantiles a posteriori sont alors approchés par l'inverse généralisé

   $$\widetilde{\Pi}^{-1}(\beta|X) := \inf\{\theta : \widetilde{\Pi}(\theta|X) > \beta\}.$$

   On montre (admis) que $\widetilde{\Pi}^{-}(\beta|X)$ converge presque sûrement vers le plus petit quantile d'ordre β de la fonction de répartition a posteriori. On construit une approximation des intervalles de crédibilité de niveau $1-\alpha$ par

   $$\left[\widetilde{\Pi}^{-1}(\beta|X), \widetilde{\Pi}^{-1}(1-\alpha+\beta|X)\right] \ , \ \beta \in]0, \alpha[.$$

On peut alors chercher à minimiser numériquement en β la longueur de l'intervalle pour approcher l'intervalle de crédibilité de longueur minimale.

3. Pour la prévision, une approximation de la densité prédictive est obtenue simplement par

$$\widetilde{f}_{X_{n+1}}\big(x|X^{(n)}\big) := \frac{1}{N}\sum_{k=1}^{N} f_{X_{n+1}}\big(x|\theta_k, X^{(n)}\big),$$

où $f_{X_{n+1}}\big(.|\theta, X^{(n)}\big)$ désigne la densité à θ fixé de X_{n+1} condtionnellement au passé.

5.2 Importance sampling

La fiabilité de l'approximation par Monte-Carlo de $\gamma = \int h(t)f(t)dt$ par $\widetilde{\gamma} = \frac{1}{N}\sum_{k=1}^{N} h(T_k)$ est directement liée à la valeur $\tau^2 = \int h(t)^2 f(t)dt$. Si τ^2 est grand, voire infini, il est nécessaire d'adapter la méthode. On remarque que γ peut s'écrire

$$\gamma = \int h(t)\frac{f(t)}{g(t)}g(t)dt$$

pour toute densité g dont le support contient celui de f (g doit vérifier $g(t) = 0 \Rightarrow f(t) = 0$ pour que l'expression au-dessus ait du sens). La densité g est appelée *densité instrumentale*. On estime alors γ par

$$\widetilde{\gamma}(g) := \frac{1}{N}\sum_{k=1}^{N} h(S_k)\frac{f(S_k)}{g(S_k)}$$

pour $S_1, ..., S_N$ un échantillon de densité g. L'idée de l'importance sampling est de choisir une densité g de manière à minimiser la variance de $\widetilde{\gamma}(g)$. Si h est positive, celle-ci est minimale (elle vaut zéro!) pour $g(t) = h(t)f(t)/\int h(s)f(s)ds$ auquel cas $\widetilde{\gamma}(g) = \gamma$. Ce résultat n'est d'aucune aide en pratique car connaître g implique de connaître γ. Il faut donc se contenter d'une densité g qui vérifie les conditions suivantes:

- Simuler un échantillon $S_1, ..., S_N$ de densité g est peu coûteux numériquement.
- La densité g est connue et a une expression analytique simple.
- Le rapport $|h(t)f(t)|/g(t)$ varie peu en fonction de t (condition évidemment difficile à vérifier en pratique).

Remarque. Si on sait simuler selon la loi a priori $\pi(.)$, il est envisageable d'appliquer naïvement la méthode de Monte-Carlo pour estimer la constante de normalisation de la loi a posteriori $\gamma = \int_\Theta f_X(X|\theta)\pi(\theta)d\theta$. Soit un échantillon $\theta_1, ..., \theta_N$ iid de loi π, on considère

$$\widetilde{\gamma} = \frac{1}{N}\sum_{k=1}^{N} f_X(X|\theta_i).$$

On vérifie bien que par la loi forte des grands nombres, $\widetilde{\gamma}$ converge presque sûrement (à X fixé, puique considéré comme déterministe ici) vers γ quand N tend vers l'infini. Cependant, cette méthode n'est efficace que lorsque l'a priori et l'a posteriori sont suffisamment proches, ce qui est très rarement le cas du fait que l'a posteriori varie rapidement avec de la taille de l'échantillon.

Dans le cadre Bayésien standard, le rapport var $(\widetilde{\gamma})/\gamma^2$ est de l'ordre de n/N, ce qui signifie que N doit être significativement plus grand que n pour obtenir une approximation satisfaisante. Du point de vue de l'importance sampling, la méthode peut être corrigée en considérant une densité instrumentale g telle que le rapport $f_X(X|\theta)\pi(\theta)/g(\theta)$ varie peu en fonction de θ.

5.3 Méthodes MCMC

Les méthodes MCMC (Monte-Carlo Markov Chains) sont des cas particuliers de méthodes de Monte-Carlo construites à partir d'une chaîne de Markov est non plus d'un échantillon iid.

Définition. Une chaîne de Markov à valeurs dans un espace $\mathcal{X}$ est une suite de variables aléatoires $X_n, n \in \mathbb{N}$ dont la loi conditionnellement au passé est invariante dans le temps et ne dépend que du passé immédiat. Formellement, pour tout ensemble mesurable A,

$$\mathbb{P}(X_n \in A|X_0, ..., X_{n-1}) = \mathbb{P}(X_n \in A|X_{n-1}) \ , \ \forall n = 1, 2, ...$$

et, pour tout $x \in \mathcal{X}$,

$$\mathbb{P}(X_n \in A|X_{n-1} = x) = \mathbb{P}(X_1 \in A|X_0 = x).$$

Remarque. Pour simplifier, on a restreint la définition aux chaînes de Markov à temps discret homogènes et d'ordre 1, ce qui reste suffisant pour les besoins du cours.

On distingue essentiellement deux types de chaînes de Markov selon si l'espace des valeurs de X_n (l'espace d'états) est discret ou continu. On rappelle brièvement quelques définitions et propriétés.

1. *Cas discret.* Supposons que X_n est à valeurs dans un ensemble fini $\mathcal{X} = \{x_1, ..., x_k\}$. La loi de X_n conditionnellement à X_{n-1} est entièrement déterminée par les valeurs $P_{ij} := \mathbb{P}(X_n = x_j|X_{n-1} = x_i), i, j = 1, ..., k$. La loi de la chaîne de Markov $X_n, n = 0, 1, ...$ est donc entièrement déterminée par la loi de X_0, que l'on identifie à un vecteur ligne $\nu^{(0)} = (\nu_1^{(0)}, ..., \nu_k^{(0)})$, et la matrice de transition $P = (P_{ij})_{i,j=1,...,k}$. De plus, la loi de X_1 est donnée par

$$\nu_j^{(1)} = \mathbb{P}(X_1 = x_j) = \sum_{i=1}^{k} \mathbb{P}(X_1 = x_j|X_0 = x_i)\mathbb{P}(X_0 = x_i) = \sum_{i=1}^{k} P_{ij}\nu_i^{(0)} \ , j = 1, ..., k,$$

   ce qui s'écrit sous forme matricielle $\nu^{(1)} = \nu^{(0)}P$. En réitérant le calcul, on obtient la loi de X_n: $\nu^{(n)} = \nu^{(0)}P^n$.

   On remarque que si X_n converge en loi, alors sa loi limite ν vérifie $\nu = \nu P$. Autrement dit, ν est un vecteur propre de $P^\top$ associé à la valeur propre 1. Il existe des conditions simples sur P qui garantissent la convergence en loi de X_n vers une unique solution. Des résultats similaires sont valables dans le cas $\mathcal{X}$ infini dénombrable pour P une matrice de transition infinie.

2. *Cas continu.* Si X_n est à valeurs dans $\mathcal{X}$ un espace continu, la loi de X_n conditionnellement à X_{n-1} est définie par un noyau de transition $K(.|x)$ qui décrit une loi de probabilité sur $\mathcal{X}$, pour tout $x \in \mathcal{X}$. Dans le cas de variables à densité, $K(.|x)$ désigne la densité de X_n

conditionnellement à $X_{n-1} = x$ (qui on rappelle ne dépend pas de n). Soit X_0 de densité f_0, la densité de X_n est alors donnée par

$$f_n(y) = \int_{\mathcal{X}} K^n(y|x) f_0(x) dx \ , \ y \in \mathcal{X}$$

où $K^n(y|x) = \int_{\mathcal{X}^{n-1}} K(y|x_1)K(x_1|x_2)...K(x_{n-1}|x)dx_1...dx_{n-1}$. Etudier les propriétés spectrales de l'opérateur $K(.|.)$ permet de déterminer des conditions sous lesquelles la chaîne $X_n, n \in \mathbb{N}$ converge en loi, auquel cas la densité limite f est un point fixe, solution de:

$$f(y) = \int_{\mathcal{X}} K(y|x) f(x) dx \ , y \in \mathcal{X}.$$

Par exemple, un processus AR(1): $X_n = aX_{n-1} + \epsilon_n, a \in]0,1[, n = 1, 2, ...$ où les ϵ_n sont iid de loi $\mathcal{N}(0,1)$ et indépendants du passé est une chaîne de Markov continue. Son noyau de transition est donné par $K(y|x) = \phi(y - ax)$ où ϕ est la densité Gaussienne standard.Soit X_0 une variable aléatoire qui initie la chaîne, on montre par récurrence que

$$X_n = a^n X_0 + \sum_{k=0}^{n-1} a^k \epsilon_{n-k}.$$

Ainsi, la loi de X_n dépend de la loi de X_0 mais cette dépendance s'estompe quand $n \to \infty$, avec comme loi limite $X_n \xrightarrow{loi} \mathcal{N}\big(0, (1-a^2)^{-2}\big)$.

Lorsqu'elle existe, la loi limite μ d'une chaîne de Markov est appelée *loi stationnaire* ou encore *loi invariante* du fait qu'elle stationnarise le processus en loi dans le sens où, si X_{n-1} est de loi μ alors X_n l'est également.

Si on veut simuler des variables aléatoires sous une loi μ, il est parfois plus simple de générer une trajectoire $X_1, ..., X_N$ d'une chaîne de Markov de loi invariante μ. Les approximations empiriques basées sur l'échantillon peuvent être satisfaisantes même si elles sont moins bonnes que dans le cas iid.

La construction de chaînes de Markov pour des méthodes de Monte-Carlo n'est utile que lorsqu'on ne sait pas simuler des variables iid (ou que c'est trop coûteux). Il y a deux inconvénients notables:

1. La loi des observations n'est en général pas exactement la loi limite. Cependant, la convergence est très rapide (exponentielle en N). En pratique, on peut (doit) supprimer les premières observations de l'échantillon pour s'assurer que les données conservées sont proches (en loi) de la loi limite.

2. Les données générées ne sont en général pas indépendantes, même dans le régime asymptotique. En effet, le produit des densités marginales $f(x)f(y)$ n'est pas égal à la densité jointe $K(y|x)f(x)$, sauf cas trivial. Cela a pour effet d'augmenter la variance des estimateurs empiriques par rapport au cas iid.

Remarque. Une suite iid de variables aléatoires est une chaîne de Markov particulière de noyau de transition $K(y|x)$ ne dépend pas de x (dans le cas discret, P_{ij} ne dépend pas de j). Les méthodes de Monte-Carlo iid sont donc un cas particulier de méthodes MCMC.

6 Simulation a posteriori

L'enjeu majeur des méthodes d'approximation numérique en Statistique Bayésienne est de pouvor simuler un échantillon sous la loi a posteriori. Cela permet d'estimer empiriquement n'importe quelle quantité définie à partir de la loi a posteriori (estimateurs, intervalles de confiances etc...). On distingue deux types de méthodes pour générer un échantillon sous la loi a posteriori.

1. Les méthodes d'échantillonage, construites à partir d'un échantillon iid. Elles sont en général exactes et plus simple d'un point de vue théorique mais ne sont pas toujours applicables.

2. Les méthode de Monte-Carlo par Chaîne de Markov (MCMC) qui sont basées sur la construction d'une suite de variables dont la loi converge vers la loi a posteriori. Ces méthodes ne font qu'approcher la loi a posteriori (elles ne sont pas exactes) mais convergent très rapidement.

Dans ce chapitre, on se restreindra à des méthodes de simulation de variables aléatoires continues car c'est la situation la plus fréquente en analyse Bayésienne. Tous les algorithmes mentionnés sont facilement généralisables à des variables discrètes.

6.1 Algorithme d'acceptation/rejet

La méthode d'acceptation/rejet permet de générer une variable aléatoire X de densité f à partir d'une variable aléatoire Y de densité g telle que le rapport f/g soit borné par une constante M connue. L'algorithme est le suivant:

1. Générer U de loi uniforme sur $[0,1]$ et Y de densité g indépendants.

2. Calculer le ratio $R = f(Y)/g(Y)$.

3. Si $R > M.U$, retourner $X = Y$, sinon reprendre à la première étape.

La variable X obtenue est exactement de densité f. L'efficacité de l'algorithme d'un point de vue computationnel dépend de la majoration M du rapport des densité f/g qui doit être idéalement la plus petite possible.

Proposition 6.1. *Soit $U_1, U_2, ...$ une suite iid de variables aléatoires de lois uniformes sur $[0,1]$ et $Y_1, Y_2, ...$ une suite iid de densité g indépendantes des U_i. On définit*

$$\tau := \inf\{n : f(Y_n)/g(Y_n) > MU_n\}.$$

Alors, la variable aléatoire $X := Y_\tau$ a pour densité f. De plus, le nombre moyen d'itérations nécessaire pour construire X est $\mathbb{E}(\tau) = M$.

Preuve. Par indépendance, le couple (Y_i, U_i) a pour densité $(y,u) \mapsto g(y)\mathbb{1}\{u \in [0,1]\}$. On définit les variables $Z_i = \mathbb{1}\{f(Y_i)/g(Y_i) > MU_i\}$ qui sont iid de loi de Bernoulli de paramètre

$$\begin{aligned}
\mathbb{P}\big(f(Y_1)/g(Y_1) > MU_1\big) &= \iint g(y)\mathbb{1}\{u \in [0,1], f(y)/g(y) > Mu\}dydu \\
&= \int g(y)\Big[\int_0^{f(y)/Mg(y)} du\Big]dy \\
&= \int g(y)\frac{f(y)}{Mg(y)}dy = \frac{1}{M}\int f(y)dy = \frac{1}{M}.
\end{aligned}$$

Soit $X = Y_\tau$, on écrit pour tout ensemble mesurable A, $\mathbb{P}(X \in A) = \sum_{n=1}^{\infty} \mathbb{P}\big(Y_n \in A, \tau = n\big)$ avec

$$\begin{aligned}\mathbb{P}\big(Y_n \in A, \tau = n\big) &= \mathbb{P}(Y_n \in A, Z_n = 1, Z_1, ..., Z_{n-1} = 0\big) \\ &= \Big(1 - \frac{1}{M}\Big)^{n-1} \mathbb{P}\big(Y_n \in A, f(Y_n)/g(Y_n) > MU_n\big) \\ &= \Big(1 - \frac{1}{M}\Big)^{n-1} \frac{1}{M} \int_A f(y)dy.\end{aligned}$$

On déduit:

$$\mathbb{P}(X \in A) = \int_A f(y)dy \frac{1}{M} \sum_{n=1}^{\infty} \Big(1 - \frac{1}{M}\Big)^{n-1} = \int_A f(y)dy. \qquad \square$$

Remarque. Une preuve plus rapide basée sur un argument intuitif consiste à remarquer que la densité du couple (Y_τ, U_τ) est la densité de (Y, U) conditionnée à l'événement $\{f(Y)/g(Y) > MU\}$. Cette densité vaut donc

$$(y, u) \mapsto \frac{g(y)\mathbb{1}\{f(y)/g(y) > Mu\}}{\mathbb{P}(f(Y)/g(Y) > MU)} \propto g(y)\mathbb{1}\{f(y)/g(y) > Mu\}.$$

On retrouve alors le résultat en intégrant en u sur $[0, 1]$.

La majoration M doit être connue mais la plus petite possible pour limiter le temps de calcul. On remarque que la condition $f \leq Mg$ implique $M \geq 1$ avec égalité si et seulement si $f \overset{p.p.}{=} g$.

Un intérêt de la méthode de rejet est qu'elle s'applique au cas où la densité f n'est connue qu'à une constante multiplicative près, comme c'est souvent le cas pour la densité a posteriori. Soit $\tilde{f} \propto f$ telle que $\tilde{f}/g \leq \tilde{M}$ pour une constante $\tilde{M}$ connue, alors

$$\frac{\tilde{f}(Y_i)}{g(Y_i)} > \tilde{M}U_i \iff \frac{f(Y_i)}{g(Y_i)} > \frac{\tilde{M}}{\int \tilde{f}(t)dt} U_i.$$

Appliquer la méthode de rejet à $\tilde{f}$ et $\tilde{M}$ revient donc à l'appliquer à f et $M = \tilde{M}/\int \tilde{f}(t)dt$.

Exemple. Soit $X_1, ..., X_n$ iid de loi $\mathcal{N}(\theta, 1)$ avec comme a priori sur θ la densité de Cauchy $\pi(\theta) \propto 1/(1 + \theta^2)$. La loi a posteriori est

$$\pi(\theta|X) \propto \frac{1}{1 + \theta^2} \exp\Big(-\frac{n}{2}\theta^2 + \theta \sum_{i=1}^{n} X_i\Big) := \tilde{f}(\theta).$$

On ne connaît pas explicitement la constante de normalisation mais on sait que

$$\tilde{f}(\theta) \leq \exp\Big(-\frac{n}{2}\theta^2 + \theta \sum_{i=1}^{n} X_i\Big) \leq \exp\Big(\frac{n}{2}\overline{X}^2\Big) \frac{\sqrt{2\pi}}{\sqrt{n}} g(\theta),$$

où g est la densité d'une loi $\mathcal{N}(\overline{X}, 1/n)$. Pour générer une variable aléatoire T sous la loi a posteriori, on peut donc appliquer la méthode d'acceptation/rejet à $\tilde{f}, g$ et $\tilde{M} = e^{n\overline{X}^2/2}\sqrt{2\pi}/\sqrt{n}$. On aura alors un temps d'attente moyen avant acceptation de $\tilde{M}/\int \tilde{f}(t)dt$ (qui dépend de X).

6.2 Algorithme d'Hasting-Metropolis

L'algorithme d'Hasting-Metropolis utilise une méthode d'acceptation/rejet sur une chaîne de Markov instrumentale que l'on sait simuler facilement. Il sert quand l'algorithme d'acceptation/rejet a des temps d'acceptation trop longs, le plus souvent pour des paramètres de dimension supérieure à 2. Soit f la densité que l'on cherche à approcher on considère un noyau de transition $G(.|.)$ pour lequel la densité $G(.|x)$ est facile à simuler et contient le support de f pour tout $x \in \mathcal{X}$. L'algorithme est le suivant.

1. On part d'une valeur initiale quelconque X_0.
2. A chaque étape $n = 1, 2, ...$, on simule une suite $(Y_i, U_i), i = 1, 2, ...$ iid avec Y_i de densité $G(.|X_{n-1})$ et U_i indépendant de Y_i de loi uniforme sur $[0, 1]$ (même si ça n'apparaît pas sur la notation, on génère une nouvelle suite (Y_i, U_i) pour chaque n). On s'arrête au temps d'arrêt
$$\tau_n := \inf\Big\{i : U_i \leq \frac{f(Y_i)G(X_{n-1}|Y_i)}{G(Y_i|X_{n-1})f(X_{n-1})}\Big\}.$$
3. On définit $X_n = Y_{\tau_n}$.

Clairement, la méthode marche à l'identique quand f est connue à une constante multiplicative près puisque le critère d'acceptation ne dépend de f que par le rapport $f(Y_i)/f(X_{n-1})$. On peut déjà remarquer que le processus obtenu par l'algorithme d'Hasting-Metropolis est une chaîne de Markov (sous-entendu, homogène d'ordre 1) du fait que la construction de X_n ne dépend du passé que par X_{n-1} et ne varie pas (en loi) avec n. On calcule maintenant son noyau de transition.

Proposition 6.2. *La chaîne de Markov $X_n, n \in \mathbb{N}$ a pour noyau de transition $K(.|.)$ qui vérifie pour tout $x \in \mathcal{X}$,*
$$y \mapsto K(y|x) \propto \min\Big\{G(y|x), \frac{f(y)}{f(x)}G(x|y)\Big\}.$$
De plus, $X_n, n \in \mathbb{N}$ a pour loi stationnaire f.

Preuve. A chaque étape n, la construction de (Y_{τ_n}, U_{τ_n}) en fonction du temps d'arrêt τ_n revient à considérer les variables Y, U conditionnellement à l'événement
$$A_n(Y, U) := \Big\{U \leq \frac{f(Y)G(X_{n-1}|Y)}{G(Y|X_{n-1})f(X_{n-1})}\Big\}.$$
La densité du couple (Y_{τ_n}, U_{τ_n}) est donc donnée par
$$f_{Y_{\tau_n},U_{\tau_n}}(y, u) = \frac{f_{Y,U}(y, u)\mathbb{1}\{A_n(y, u)\}}{\mathbb{P}(A_n(Y, U))} \propto G(y|X_{n-1})\mathbb{1}\big\{A_n(y, u)\big\}.$$
On obtient la densité de X_n conditionnellement à X_{n-1} en intégrant en u,
$$\begin{aligned} y \mapsto K(y|X_{n-1}) &\propto G(y|X_{n-1}) \int_0^{\min\left\{1, \frac{f(y)G(X_{n-1}|y)}{G(y|X_{n-1})f(X_{n-1})}\right\}} du \\ &\propto \min\Big\{G(y|X_{n-1}), \frac{f(y)}{f(X_{n-1})}G(X_{n-1}|y)\Big\}. \end{aligned}$$

Pour la seconde partie du résultat, montrons seulement que f est un point fixe de l'équation $f = \int K(.|x)f(x)dx$. Comme f et $\int K(.|x)f(x)dx$ sont des densités, il est suffisant de montrer qu'elles sont proportionnelles. On a pour tout $y \in \mathcal{X}$,

$$\int K(y|x)f(x)dx \propto f(y) \int \min\left\{G(y|x)\frac{f(x)}{f(y)}, G(x|y)\right\}dx \propto f(y) \int K(x|y)dx \propto f(y),$$

où on a utilisé que $\int K(x|y)dx = 1$ pour tout $y \in \mathcal{X}$. □

!!! Application avec l'approximation de Laplace !!!

6.3 Echantillonage de Gibbs

L'échantillonage de Gibbs est un algorithme permettant de générer une chaîne de Markov X_n à valeurs dans $\mathbb{R}^d$ avec $d > 1$ lorsque l'on sait générer sous les lois conditionnelles marginales. Soit $X = (X_1, ..., X_d)$ un vecteur aléatoire de $\mathbb{R}^d$ de densité f, on note $f_i(.|X^{(-i)}), i = 1, ..., d$ la densité de X_i sachant $X^{(-i)} := (X_1, ..., X_{i-1}, X_{i+1}, ..., X_d)$. L'algorithme est le suivant:

1. On part d'une variable aléatoire $X_0 = (X_{0,1}, ..., X_{0,d})$, éventuellement constante p.s., à valeurs dans le support de f (théoriquement, seules les $d-1$ premières valeurs sont nécessaires).
2. A chaque étape n, pour tout $i = 1, ..., d$, on simule
$$X_{n,i} \sim f_i(.|X_{n,1}, ..., X_{n,i-1}, X_{n-1,i+1}, ..., X_{n-1,d}).$$
3. On définit $X_n = (X_{n,1}, ..., X_{n,d})$, pour $n = 0, 1, 2, ...$

L'échantillonage de Gibbs réduit la simulation d'un vecteur aléatoire de dimension d à d simulations de variables aléatoires de dimension 1. L'idée est de simuler récursivement chaque composante de X_n à partir des $d-1$ dernières valeurs générées. Le processus X_n est bien une chaîne de Markov (de $\mathbb{R}^d$) homogène d'ordre 1 puisque la loi de X_n conditionnellement au passé ne dépend que du passé immédiat X_{n-1} et ne varie pas en fonction du temps. On montre alors que la densité f est stationnaire en procédant récursivement.

- Soit $X_n = (X_{n,1}, ..., X_{n,d})$ de densité f.
- On simule $X_{n+1,1}$ sous $f_1(.|X_{n,2}, ..., X_{n,d})$. Soit $f_{(-1)}$ est la densité de $(X_{n,2}, ..., X_{n,d})$ obtenue en intégrant f en x_1, le vecteur $(X_{n+1,1}, X_{n,2}, ..., X_{n,d})$ a donc pour densité
$$(x_1, ..., x_d) \mapsto f_1(x_1|x_2, ..., x_d) \times f_{(-1)}(x_2, ..., x_d) = f(x_1, ..., x_d).$$
- On itère le raisonnement pour $X_{n+1,2}$ puis $X_{n+1,j}$ pour tout j.

Remarque. Comme pour l'algorithme d'Hasting-Metropolis, il est recommandé de ne pas garder les premières valeurs de la chaîne générée, de manière à ce que les données conservées soient plus proches du régime stationnaire. La quantité d'observations mises de côté dépend de la puissance de calcul disponible. Sans contrainte de calul, on peut supposer en général que supprimer les 1000 premières valeurs de la chaîne est suffisant.

Exemple. Soit (X_1, X_2) de densité jointe $f(x_1, x_2) = e^{-x_1-x_2-x_1x_2}, (x_1, x_2) \in [0, +\infty[^2$. En calculant les densités marginales, on montre que la loi de X_1 conditionnelement à X_2 est la loi exponentielle de paramètre $1 + X_2$ (et réciproquement par symétrie). Pour simuler un vecteur aléatoire de densité f (approximativement), on utilise l'échantillonage de Gibbs:

1. On choisit arbitrairement $X_{0,1}, X_{0,2} > 0$.
2. Pour tout $n = 1, 2, ...$ on simule $X_{n,1}$ de loi $\mathcal{E}(1 + X_{n-1,2})$ puis $X_{n,2}$ de loi $\mathcal{E}(1 + X_{n,1})$.
3. La suite $(X_{n,1}, X_{n,2})_{n\geq 1}$ converge en loi vers un vecteur aléatoire de densité f.

6.4 Lois de mélange

Dans le cadre Bayésien, la loi de l'échantillon X est introduite conditionnellement au paramètre. La loi marginale s'écrit sous la forme d'une intégrale contre la loi a priori

$$f_X(x) = \int f_X(x|\theta)\pi(\theta)d\theta.$$

Ce genre d'expression est appelé *mélange continu* dû à l'analogie avec les mélanges discrets "classiques" de densités, qui s'expriment sous la forme

$$f(x) = \sum_{j=1}^{k} p_j f_j(x)$$

où les p_j sont des poids positifs avec $\sum_{j=1}^{k} p_j = 1$ et les f_j sont des densités. De manière générale, un mélange discret ou continu de densités peut s'écrire comme l'espérance d'une densité conditionnelle

$$f(x) = \mathbb{E}\big(f(x|Y)\big).$$

Simuler une variable aléatoire sous une densité de mélange $f(x) = \mathbb{E}\big(f(x|Y)\big)$ est simple à condition de savoir générer Y et sous la loi conditionnelle $f(.|y)$ pour tout y.

Proposition 6.3. *Soit $f(.|y), y \in \mathcal{Y}$ une famille de densités et Y une variable aléatoire à valeurs dans $\mathcal{Y}$. Une variable aléatoire X générée sous la densité $f(.|Y)$ a pour densité $f(x) = \mathbb{E}\big(f(x|Y)\big)$.*

Preuve. Il suffit de remarquer que $f(.|Y)$ est la densité de X conditionnellement à Y, ce qui donne pour tout ensemble mesurable A,

$$\mathbb{P}(X \in A) = \mathbb{E}\big(\mathbb{P}(X \in A|Y)\big) = \mathbb{E}\bigg(\int_A f(t|Y)dt\bigg) = \int_A \mathbb{E}\big(f(t|Y)\big)dt$$

par le théorême de Fubini. □

Pour générer une variable aléatoire X sous une densité de mélange $f(x) = \mathbb{E}\big(f(x|Y)\big)$, il suffit donc de générer Y puis X de densité $f(.|Y)$.

Applications.

1. Soit X issu d'un modèle paramétrique de vraisemblance $f_X(.|\theta)$ avec $\pi(.)$ l'a priori sur θ. Pour générer sous la loi marginale de l'échantillon, il suffit de générer une variable aléatoire T de loi π puis X de densité $f_X(.|T)$.
2. En prévision Bayésienne, la densité d'une nouvelle observation X_{n+1} sachant le passé $X^{(n)}$ est la densité prédictive qui s'écrit

$$f_{X_{n+1}}(x|X^{(n)}) = \int f_{X_{n+1}}(x|\theta, X^{(n)})\pi(\theta|X)d\theta = \mathbb{E}\big(f_{X_{n+1}}(x|\theta, X^{(n)})\big|X^{(n)}\big).$$

Si on sait générer T sous la loi a posteriori, une variable aléatoire X de densité $f_{X_{n+1}}(x|T, X^{(n)})$ a pour densité $f_{X_{n+1}}(.|X^{(n)})$.