Introduction aux Statistiques Bayésiennes

Yann Traonmilin - Adrien Richou
(Basées dur les notes de cours de Charles Dossal et de Jérémie Bigot)
Ce document est susceptible d'être mis à jour au cours de l'UE

2018

Introduction

Considérons quatre problèmes d'inférence statistique.

1. Une machine à sous disposant d'un bouton donne 1EUR avec une probabilité θ et 0 EUR sinon. On cherche à estimer cette probabilité.
2. Pour une étude de marché, on cherche à estimer la moyenne du prix de vente d'un produit.
3. Un informateur nous préviens que 30% des machines à sous ont une probabilité θ_1 de donner 1EUR, le reste a une probabilité θ_2. On cherche à savoir à quelle type appartient cette machine.
4. Une société de conseil nous propose de faire l'étude du prix de vente. Pour un produit, on fait une étude parallèle pour étudier si l'information qu'elle propose est fiable.

Dans chacun de ces exemples, on cherche à estimer à partir d'observations un paramètre décrivant la distribution de probabilité. On remarque que dans les exemples 3 et 4, on dispose d'une information supplémentaire sur ce paramètre. Ce cours est destiné à donner un cadre précis pour l'utilisation de cette information *a priori* dans un problème d'inférence.

1 Introduction aux principes de l'inférence bayésienne.

1.1 Rappels de probabilités

Définition 1.1 (Probabilité conditionnelle)**.** *Soit A et B deux événements tels que $\mathbb{P}(B) \neq 0$, alors*

$$\mathbb{P}(A|B) := \frac{P(A \cap B)}{P(B)} \tag{1}$$

Théorème 1.1 (Probabilités totales)**.** *Soit A et B deux événements tels que $\mathbb{P}(B) \neq 0$, alors*

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A}) \tag{2}$$

Dem.

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap \bar{A}) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A}) \tag{3}$$

□

Théorème 1.2 (Bayes). *Soit A et B deux événements tels que $\mathbb{P}(B) \neq 0$, alors*

$$\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A})}.\end{aligned}$$

Dem.

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) \tag{4}$$

□

Dans ce cours la notion d'indépendance sera (quasi)-exclusivement une notion d'indépendance conditionnelle.

Définition 1.2 (Probabilité conditionnelle). *Soit A, B, C des événements tels que $\mathbb{P}(C) \neq 0$ alors A est indépendant de B conditionnellement à C si*

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C). \tag{5}$$

Remarque 1.1. *Le lecteur pourra vérifier que deux événements indépendants A et B (c.à.d. $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$) ne sont pas conditionnellement indépendants en général.*

Définition 1.3 (Densité conditionnelle). *Soit X et Y deux variables aléatoires de loi jointe $f(x, y)$ sous réserve de non négativité du dénominateur on définit la densité conditionnelle*

$$f(x|y) := \frac{f(x, y)}{\int f(x, y)dx}. \tag{6}$$

1.2 Rappels sur l'approche fréquentiste

On cherche à estimer une quantité d'intérêt $\hat{\theta}$ à partir d'observations $(x_i, \cdots, x_n)$. Pour cela on se donne un **modèle statistique** qui consiste à se donner $X = (X_1, \cdots, X_n)$ v.a. (continues ou discrètes) à valeurs dans $\mathbb{R}^d$ qui sont indépendantes et dont la loi dépend d'un paramètre $\theta \in \Theta \subset \mathbb{R}^p$. On définit une manière de mesurer la qualité d'un paramètre donné pour un ensemble d'observations :

Définition 1.4. *Si $(X_k)_{k\leqslant n}$ sont des variables discrètes i.i.d., prenant un nombre dénombrable de valeurs $(x_i)_{i\in I}$ selon une loi $\mathbb{P}_\theta$ dépendant d'un paramètre θ, on appelle fonction de vraisemblance la fonction L définie par*

$$L(\theta, x_1, x_2, \cdots, x_n, \theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X_i = x_i). \tag{7}$$

Si $(X_k)_{k\leqslant n}$ sont des variables continues iid dont la loi est une densité f_θ dépendant d'un paramètre θ, on appelle fonction de vraisemblance la fonction L définie par

$$L(\theta, x_1, x_2, \cdots, x_n) = \prod_{k=1}^{n} f_\theta(x_i) \tag{8}$$

Dans les deux cas, la valeur de cette fonction au point $(\theta, x_1, x_2, \cdots, x_n)$ est la vraisemblance de l'échantillon $(x_1, x_2, \cdots, x_n)$ pour le paramètre θ.

Exemples :

1. On considère n variables aléatoires $(X_i)_{i\leqslant n}$ iid suivant une loi gausienne $\mathcal{N}(\theta, \sigma^2)$ où $\theta \in \mathbb{R}$ et σ^2 est supposé fixé et connu. On note g la densité de la loi jointe :

$$g(x) = f_\theta(x_1, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2} = L(\theta, x_1, \cdots, x_n).$$

   Cette loi jointe vue comme une fonction $L(\theta, x_1, \cdots, x_n)$ des x_i et du paramètre θ est la fonction de vraisemblance.

2. On considère n v.a. i.i.d. $(X_1, X_2, \cdots, X_n)$ suivant une loi de Bernoulli de paramètre θ, $\mathcal{B}(\theta)$, $\theta \in [0, 1]$.

$$L(\theta, x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i|\theta) = \theta^s(1-\theta)^{n-s}$$

   où $s = \sum_{i=1}^{n} x_i$.

Une large gamme de méthodes d'estimation repose sur la technique du maximum de vraisemblance

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta, x). \tag{9}$$

Dans ces deux exemples, on a

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Dem. 1. On prend le log de L, ce qui revient à calculer

$$\hat{\theta} = \arg\min_{\theta\in[0,1]} \sum_{i=1,n} (x_i - \theta)^2 = \arg\min_{\theta\in[0,1]} G(\theta). \tag{10}$$

   On cherche $\hat{\theta}$ tel que $G'(\hat{\theta}) = 0$, ce qui donne $\sum_{i=1,n} \hat{\theta} - x_i = 0$ et $\hat{\theta} = \frac{1}{n}\sum_{i=1,n} x_i$

2. On prend le log de L, ce qui revient à calculer

$$\hat{\theta} = \arg\max_{\theta\in\mathbb{R}} s\log(\theta) + (n-s)\log(1-\theta) = \arg\max_{\theta\in\mathbb{R}} G(\theta) \tag{11}$$

   On cherche $\hat{\theta}$ tel que $G'(\hat{\theta}) = 0$, ce qui donne $\frac{s}{\hat{\theta}} - \frac{n-s}{1-\hat{\theta}} = 0$ et $\hat{\theta} = \frac{s}{n}$

□

1.3 Le paradigme baysésien.

On dispose d'une information **a priori** sur le paramètre inconnu θ. Cette information prend la forme d'une loi sur l'espace des paramètres Θ notée π qui s'appelle la loi a priori. Le paramètre θ devient une variable aléatoire et on note $\theta \sim \pi$. Ainsi la notion de probabilité ou densité de probabilité paramétré par θ n'a plus vraiment de sens. Les notions de l'approche fréquentiste sont remplacées par des notions de probabilités, d'indépendance et de densités de probabilité **conditionnelles à** θ (Lorsque l'on se place d'un un cadre uniquement bayésien, on se permettra de ne pas mentionner ce caractère conditionnel).

Définition 1.5 (Loi a priori)**.** *Soit $(f_\theta)_{\theta\in\Theta}$ une famille de densités de probabilité à paramètre dans Θ. Une loi a priori π est une loi de probabilité (densité de probabilité) sur Θ.*

Définition 1.6. *Ainsi la loi jointe des observations de $X = (X_1, \cdots, X_n)$ est conditionnelle à θ et est notée $f(x|\theta) = f(x_1, \cdots, x_n|\theta)$ dans le cas continu et $\mathbb{P}(X = x|\theta) = \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n|\theta)$ dans le cas discret. Dans le cas continu, on s'autorise à noter $f(X|\theta)$ la densité jointe de la v.a. X.*

Définition 1.7 (Modèle Bayésien)**.** *Un **modèle Bayésien** est la donnée, pour une v.a. (ou une suite de v.a.) d'une loi conditionnelle et d'une loi a priori :*

$$\begin{aligned} X &\sim f(X|\theta) \\ \theta &\sim \pi \end{aligned} \tag{12}$$

A partir d'un modèle Bayésien, on peut calculer une loi *a posteriori* sur θ, cette loi n'est rien d'autre que la loi de θ conditionnellement aux observations X.

Définition 1.8 (Loi a posteriori)**.** *On peut séparer les situations en 4 catégories selon que la loi de X est discrète ou continue ou que la loi a priori est discrète ou continue.*

1. *La loi de X et la loi a priori sont discrètes.*
   C'est le cas des trois exemples de l'introduction.
   Dans cettte situation la loi a posteriori est entièrement définie par les valeurs

$$\begin{aligned} \mathbb{P}(\theta = \theta_i|X = x) &= \frac{\mathbb{P}(X = x|\theta = \theta_i)\mathbb{P}(\theta = \theta_i)}{\mathbb{P}(X = x)} \\ &= \frac{\mathbb{P}(X = x|\theta = \theta_i)\mathbb{P}(\theta = \theta_i)}{\sum_k \mathbb{P}(X = x|\theta = \theta_k)\mathbb{P}(\theta = \theta_k)}. \end{aligned}$$

2. *La loi de X est discrète et la loi de θ est continue de densité notée π. Dans ce cas, la loi a posteriori est une loi continue (à densité) et est définie par :*

$$\pi(\theta|X = x) = \frac{\mathbb{P}(X = x|\theta)\pi(\theta)}{\int_{u\in\Theta} \mathbb{P}(X = x|u)\pi(u)d(u)} \tag{13}$$

3. *La loi de X est continue et la loi a priori π sur θ est discrète. La loi a posteriori est une loi discrète, comme la loi a priori et elle est définie par les probabilités suivantes :*

$$\mathbb{P}(\theta = \theta_i|X) = \frac{f(X|\theta = \theta_i)\mathbb{P}(\theta = \theta_i)}{\sum_k f(X|\theta = \theta_k)\mathbb{P}(\theta = \theta_k)}$$

4. *La loi de X et la loi a priori π sur θ sont continues. Dans ce cas, la loi a posteriori est continue et sa densité est donnée par Comme dans le cas précédent on peut exprimer la loi a posteriori de la manière suivante :*

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{\int_{u\in\Theta} f(X|u)\pi(u)du}$$

Ces 4 formulations ne sont que 4 spécifications d'un même égalité que l'on résume souvent sous la forme continue/continue.

Comme le dénominateur ne dépend pas de θ, on l'interprète souvent comme une constante de normalisation.

Définition 1.9. *On appelle la loi marginale la loi définie par :*

$$m_\pi(X) = \int_{u\in\Theta} f(X|u)\pi(u)du. \tag{14}$$

Elle ne dépend que de X et de la loi a priori et donc pas du paramètre θ.

Si on cherche par exemple le maximum de cette loi *a posteriori* , le calcul de la loi marginale est inutile. On note ainsi parfois

$$\pi(\theta|X) \propto f(X|\theta)\pi(\theta).$$

Si on veut calculer la loi marginale, on ne calcule souvent que le dénominateur et on identifie une loi usuelle pour en déduire la constante de normalisation.

Remarque 1.2. ***Important !*** *Le calcul d'une loi a posteriori mène à une loi. Ainsi, le résultat de l'inférence est beaucoup plus informatif que dans le cas fréquentiste : on a accès beaucoup plus facilement à des intervalles de confiances pour une estimation de θ(en prenant par ex. le maximum de la loi a posteriori)*

Remarque 1.3. ***Important !*** *En pratique, on calculera la loi a posteriori* empirique*. Dans le cas continu pour X, on considère une réalisation $x_1, ..., x_n$ de X, on aura alors :*

$$\pi(\theta|X = (x_1, .., x_n)) = \frac{f(x_1, ..., x_n|\theta)\pi(\theta)}{\int_{u\in\Theta} f(X|u)\pi(u)du}.$$

2 Comment choisir la loi *a priori* ?

Le choix des lois *a priori* est une étape fondamentale en statistique bayésienne et constitue une différence notable avec la statistique fréquentiste. Les différents choix possibles peuvent être motivés par différents points de vue :

- Choix basé sur des expériences du passé ou sur une intuition du statisticien.
- Choix basé sur la faisabilité des calculs.
- Choix basé sur la volonté de n'apporter aucune information nouvelle pouvant biaiser l'estimation.

2.1 Lois subjectives

L'idée est d'utiliser les données antérieures. Dans un cas concret, il peut être judicieux de baser son raisonnement sur l'expertise de spécialistes. Par exemple, si on fait des biostatistiques, on s'appuiera sur l'expertise des médecins et des biologistes pour déterminer une loi *a priori* cohérente. Si l'on a plusieurs expertises distinctes, on pourra les pondérer en utilisant un modèle hiérarchique (cf chapitre ?).

2.2 Approche partiellement informative

2.2.1 Notion de lois conjuguées

Définition 2.1. *Une famille $\mathcal{F}$ de distributions sur Θ est dite conjuguée pour la loi $f(x|\theta)$ si pour tout $\pi \in \mathcal{F}$; la distribution a posteriori $\pi(.|x)$ appartient également à $\mathcal{F}$.*

L'avantage des familles conjuguées est avant tout de simplifier les calculs. Avant le développement des outils de calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs. Un autre intérêt est que la mise à jour la loi se fait à travers les paramètres de la loi et donc l'interprétation est souvent bien plus facile.

Exemple : La famille de toutes les lois de probabilité sur Θ est toujours conjuguée par la loi $f(.|\theta)$, et ce quelque soit la loi $f(.|\theta)$.

Ce premier exemple trivial n'a pas d'intérêt concret mais il permet de se rendre compte qu'une famille conjuguée n'a d'intérêt que si elle n'est pas trop grande. En particulier, on prendra des familles de lois paramétriques de dimension finie.

Quelques exemples de lois conjuguées

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}\left(\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$
$\mathcal{P}(\theta)$	$Ga(\alpha, \beta)$	$Ga(\alpha + x, \beta + 1)$
$Ga(\nu, \theta)$	$Ga(\alpha, \beta)$	$Ga(\alpha + \nu, \beta + x)$
$B(n, \theta)$	$Be(\alpha, \beta)$	$Be(\alpha + x, \beta + n - x)$
$\mathcal{N}(\mu, \frac{1}{\theta})$	$Ga(\alpha, \beta)$	$Ga(\alpha + \frac{1}{2}, \beta + \frac{(\mu-x)^2}{2})$

Une loi conjuguée peut être déterminée en considérant la forme de la vraisemblance $f(x|\theta)$ et en prenant une loi a priori de la même forme que cette dernière vue comme une fonction du paramètre.

Exemple : on considère une loi Pareto de paramètres (α, a) :

$$f(x|\theta, a) = \frac{\theta a^\theta}{x^{\theta+1}} \chi_{[a,+\infty[}(x).$$

Supposons a connu, $f(x|\theta) \propto \theta e^{\theta \log(a/x)} x^{-1} \chi_{[a,+\infty[}(x)$. On pourrait donc prendre une loi a priori de type Gamma.

2.2.2 Cas du modèle exponentiel

Définition 2.2. *On appelle famille exponentielle à s paramètres, toute famille de loi de distribution $\{P_\theta\}$ dont la densité a la forme suivante :*

$$f(x|\theta) = \exp\left(\sum_{j=1}^{s} \eta_j(\theta) T_j(x) - B(\theta)\right) h(x) = \exp\left(\langle \eta(\theta), T(x) \rangle - B(\theta)\right) h(x)$$

où $\eta_i(.)$ et $B(.)$ sont des fonctions du paramètre θ et les $T_i(.)$ sont des statistiques. Le vecteur $\eta(\theta)$ est appelé paramètre naturel de la famille.

La vraisemblance complète d'une séquence $(x_1, ..., x_n)$ s'écrit

$$f(x|\theta) = \exp\left(\langle \eta(\theta), \sum_{i=1}^{n} T(x_i)\rangle - nB(\theta)\right)\left(\prod_{i=1}^{n} h(x_i)\right).$$

$T_n(x) = \sum_{i=1}^{n} T(x_i)$ est appelé vecteur de statistiques exhaustives pour θ. Cette statistique contient toute l'information de l'échantillon sur les paramètres de la loi de probabilité. Nous renvoyons le lecteur intéressé par la notion de statistique exhaustive à un cours avancé de statistique classique.

Il est habituel d'écrire le modèle exponentiel sous la forme dite canonique en le reparamétrant (on pose $\tilde{\theta}_i = \eta_i(\theta)$) ce qui donne

$$f(x|\theta) = \exp\left(\langle \tilde{\theta}, T(x)\rangle - A(\tilde{\theta})\right) h(x).$$

La plupart des lois classiques forment des familles exponentielles. On peut citer par exemple les lois de Bernoulli, Poisson, binomiale (avec n fixé), exponentielle, χ^2, normale, gamma, beta, ... A contrario, les lois dont le support dépend de θ ne forment jamais des familles exponentielles.

Proposition 2.1. *Soit $f(x, \theta)$ appartenant à une famille exponentielle canonique. Alors une famille de loi a priori conjuguée pour $f(x, \theta)$ est donnée par :*

$$\pi(\theta) = K(\mu, \lambda) \exp(\langle \theta, \mu\rangle - \lambda A(\theta))$$

où (μ, λ) sont des paramètres (μ de dimension s et λ de dimension 1) et $K(\mu, \lambda)$ est une constante de renormalisation. Dans ce cas la loi a posteriori est de la forme :

$$\pi(\theta|x) \propto \exp(\langle (\mu + T(x)), \theta\rangle - (\lambda + 1)A(\theta)).$$

Dem.

$$\begin{aligned}\pi(\theta|x) &\propto f(x|\theta)\pi(\theta)\\ &\propto \exp\left(\langle \theta, T(x)\rangle - A(\theta)\right) \exp(\langle \theta, \mu\rangle - \lambda A(\theta))\\ &\propto \exp(\langle (\mu + T(x)), \theta\rangle - (\lambda + 1)A(\theta)).\end{aligned}$$

□

Exemple : exercice 4 du TD2.

Remarque 2.1. *La proposition 2.1 est formelle, elle peut aboutir à des lois $\pi(\theta)$ non intégrables !*

Dans la suite on pourra éventuellement considérer des lois π telles que

$$\int_\theta \pi(\theta) d\theta = +\infty$$

On parle alors de distribution *impropre*.

Important : la distribution *a posteriori* doit être définie i.e.

$$\int_\theta f(x|\theta)\pi(\theta) d\theta < +\infty.$$

Exemples :

— Si X suit une loi normale $\mathcal{N}(\theta, 1)$ et que π est la mesure de Lebesgue sur $\mathbb{R}$ alors

$$\int_{\theta\in\mathbb{R}} f(x|\theta) d\theta = \int_{\theta\in\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} d\theta = 1$$

et ainsi la loi *a posteriori* est $\pi(\theta|X) = \mathcal{N}(X, 1)$.

— Exercice 5 du TD2.

2.3 Loi *a priori* non informative

Dans le cas où on dispose que de peu d'informations sur θ, on peut choisir des loi *a priori* dites peu ou non informatives. On souhaite que l'*a priori* intervienne de façon minimale dans la loi *a posteriori* , i.e. que les données parlent d'elles même.

2.3.1 Lois invariantes

- Soit f une densité sur $\mathbb{R}^d$, la famille de lois $\{f(.|\theta)\}_{\theta\in\mathbb{R}^d}$ avec $f(.|\theta) = f(.-\theta)$ est invariante par translation : en effet, si $X \sim f(x|\theta)$ alors $X + \theta_0 \sim f(x|\theta + \theta_0)$. On dit dans ce cas que θ est un paramètre de position. Comme $\{f(.|\theta)\}_{\theta\in\mathbb{R}^d} = \{f(.|\theta + \theta_0)\}_{\theta\in\mathbb{R}^d}$, il est naturel de demander à la loi *a priori* π d'être invariante par translation, c'est à dire qu'elle satisfasse $\pi(\theta) = \pi(\theta + \theta_0)$ pour tous $\theta_0 \in \Theta$. On trouve alors que π est constante, c'est-à-dire la loi (éventuellement impropre) uniforme sur Θ.
- Si la famille de lois est paramétrée par un paramètre d'échelle, c'est à dire que l'on a $f(x|\sigma) = \frac{1}{\sigma}f(\frac{x}{\sigma})$ pour $\sigma \in \mathbb{R}^{+*}$ et f une densité sur $\mathbb{R}^d$, alors elle est invariante par changement d'échelle : Si $X \sim f(x|\sigma)$, alors $\alpha Y \sim f(x|\sigma\alpha)$ avec $\alpha > 0$. On dit dans ce cas que σ est un paramètre d'échelle. Comme $\{f(.|\sigma)\}_{\sigma\in\mathbb{R}^{+*}} = \{f(.|\sigma\alpha)\}_{\sigma\in\mathbb{R}^{+*}}$, il est naturel de demander à la loi *a priori* π d'être invariante par changement d'échelle, c'est-à-dire qu'elle satisfasse $\pi(\sigma) = \alpha\pi(\alpha\sigma)$ pour tous $\alpha > 0$. Ceci implique que $\pi(\sigma) = c/\sigma$ où c est une constante. Dans ce cas la mesure invariante n'est plus constante.

Ces approches invariantes sont parfois d'un intérêt limité pour plusieurs raisons :

- possibilité d'avoir plusieurs structures d'invariance,
- possibilité de ne pas avoir de structure d'invariance,
- parfois artificiel, sans intérêt pratique.

2.3.2 Loi *a priori* de Jeffreys

Intuitivement, si l'on ne veut pas d'un *a priori* informatif, on pourrait penser que la meilleure stratégie est de prendre la loi uniforme sur Θ.

Exemple : On s'intéresse à la probabilité de naissance d'une fille notée $\theta \in [0, 1]$. On peut prendre la loi *a priori* uniforme sur $[0, 1]$.

Cette approche soulève tout de même un problème très important : la notion de non-information dépend de la paramétrisation du problème ! Par exemple, si θ a pour loi *a priori* $\mathcal{U}([0, 1])$ et si $\phi = \log\left(\frac{\theta}{1-\theta}\right)$ est une reparamétrisation du modèle, alors l'*a priori* sur ϕ a pour densité $\pi(\phi) = \frac{e^{-\phi}}{(1+e^{-\phi})^2}$ qui semble beaucoup plus informatif... On voit ainsi qu'une bonne notion de loi *a priori* non-informative est une loi invariante par reparamétrisation.

La loi *a priori* de Jeffreys est fondée sur l'information de Fisher.

a) Cas unidimensionnel.

On rappelle la définition de l'information de Fischer :

$$I(\theta) = \mathbb{E}\left[\left|\frac{\partial}{\partial\theta}\log f(X|\theta)\right|^2\right]$$

qui, sous certaines conditions de régularité, peut se réécrire

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right].$$

Définition 2.3. *La loi a priori de Jeffreys est donnée par*

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

Cette loi possède deux intérêts principaux :

— $I(\theta)$ est un indicateur de la quantité d'information apportée par le modèle $f(x|\theta)$. Donc $I(\theta)$ est grand lorsque le modèle varie fortement autour de θ. par conséquent, au moins à un niveau qualitatif, il paraît intuitivement justifié que les valeurs de θ pour lesquelles $I(\theta)$ est plus grande doivent être plus probables a priori.

— La loi de Jeffreys est invariante par reparamétrisation. En effet soit $\phi = h(\theta)$ avec h un C^1-difféomorphisme. Si on note π la loi *a priori* de θ, alors ϕ est de loi $\tilde{\pi}$ avec $\tilde{\pi}(\phi) = \pi(\phi)|(h^{-1})'(\phi)|$. De plus on a $\tilde{I}(\phi) = I(\phi)|(h^{-1})'(\phi)|^2$ donc $\tilde{\pi}(\phi) \propto \sqrt{\tilde{I}(\phi)}$. Ce calcul justifie en particulier la présence de la racine carrée.

b) Cas multi-dimensionnel.

Si $\theta \in \mathbb{R}^k$ alors $I(\theta)$ est une matrice dont les coefficients sont

$$I_{ij}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X|\theta)\right].$$

Définition 2.4. *La loi a priori de Jeffreys est donnée par*

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

Si $\theta \in \mathbb{R}^k$ alors $I(\theta)$ est une matrice dont les coefficients sont

$$I_{ij}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X,\theta)\right].$$

Définition 2.5. *La loi a priori de Jeffreys est donnée par*

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

Exemple : Si $X \sim \mathcal{N}(\mu, \sigma^2)$ et que l'on cherche à estimer $\theta = (\mu, \sigma^2)$, alors $\pi(\theta) \propto \sigma^{-2}$.

2.4 Cas particulier du modèle normal

On a déjà vu que lorsque l'on considère un échantillon i.i.d. $X_1, ..., X_n \sim \mathcal{N}(\theta, \sigma^2)$ on peut prendre une loi gaussienne a priori sur θ, on obtient alors une loi conjuguée : Si $\pi(\theta) \sim \mathcal{N}(\mu, \tau^2)$ alors

$$\pi(\theta|x_1, ..., x_n) \sim \mathcal{N}(\bar{x}_n - \frac{\sigma_n^2}{\sigma_n^2 + \tau^2}(\bar{x}_n - \mu), \frac{\sigma_n^2\tau^2}{\sigma_n^2 + \tau^2}$$

avec $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ et $\sigma_n^2 = \frac{\sigma^2}{n}$. On peut montrer que ce résultat se généralise pour une loi gaussienne multidimensionnelle.

Proposition 2.2. *Soit un échantillon i.i.d. $X_1, ..., X_n \sim \mathcal{N}_p(\theta, \Sigma)$ avec Σ une matrice de covariance connue. Si $\pi(\theta) \sim \mathcal{N}_p(\mu, A)$ alors on a une loi cinjuguée :*

$$\pi(\theta|x_1, ..., x_n) \sim \mathcal{N}_p(\bar{x}_n - \frac{\Sigma}{n}(\frac{\Sigma}{n} + A)^{-1}(\bar{x}_n - \mu), (A^{-1} + n\Sigma^{-1})^{-1}).$$

On peut naturellement se demander ce qui se passe lorsque l'on cherche à estimer l'espérance et la variance en même temps. On a besoin d'une nouvelle loi a priori sur (θ, Σ).

2.4.1 Un premier *a priori*

On se place en dimension 1 et on considère un échantillon $X_1, ..., X_n$ i.i.d. de loi $\mathcal{N}(\theta, \sigma^2)$. On note

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s_n^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

La vraisemblance vaut

$$L(\theta, \sigma^2; x_1, ..., x_n) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(s_n^2 + n(\bar{x} - \theta)^2)\right).$$

En supposant θ et σ indépendants et en appliquant le principe d'invariance on peut prendre l'*a priori* non informatif $\pi(\theta, \sigma) = \frac{1}{\sigma}$. C'est également l'*a priori* de Jeffreys. On trouve alors

$$\pi(x|\theta, \sigma^2) \propto \sigma^{-1} \exp\left(-\frac{n}{\sigma^2}(\bar{x}_n - \theta)^2\right)(\sigma^{-2})^{\frac{n}{2}} \exp\left(-\frac{s_n^2}{2\sigma^2}\right).$$

On a donc la loi *a posteriori* suivante :

Proposition 2.3.

$$\pi(\theta|\sigma, x) \sim \mathcal{N}(\bar{x}_n, \frac{\sigma^2}{n})$$
$$\pi(\sigma^2|x) \sim IG\left(\frac{n-1}{2}, \frac{s_n^2}{2}\right)$$

où la loi $IG(\alpha, \beta)$ est la loi inverse gamma de densité

$$\frac{\beta^\alpha}{\Gamma(\alpha)x^{\alpha+1}} e^{-\beta/x} \chi_{]0,+\infty[}(x).$$

Ce premier résultat est partiellement intéressant car nous n'obtenons pas une loi conjuguée.

2.4.2 Loi *a priori* conjuguée

Pour obtenir une loi *a priori* conjuguée et au vue du résultat précédent, on va introduire une dépendance entre θ et σ^2. On considère la loi *a priori* suivante :

$$\pi(\theta, \sigma^2) = \pi_1(\theta|\sigma^2)\pi_2(\sigma^2)$$

où

$$\pi_1(\theta|\sigma^2) \sim \mathcal{N}\left(\theta_0, \frac{\sigma^2}{n_0}\right)$$
$$\pi_2(\sigma^2) \sim IG\left(\frac{\nu}{2}, \frac{s_0^2}{2}\right).$$

Notons que l'on a 4 hyperparamètres $\theta_0, n_0, \nu ets_0^2$. On trouve alors la loi *a posteriori* suivante :

$$\pi(\theta, \sigma^2|x) \propto \sigma^{-n-\nu-3} \exp\left(-\frac{1}{2\sigma^2}(s_1^2 + n_1(\theta - \theta_1)^2)\right)$$

où

$$n_1 = n + n_0, \quad \theta_1 = \frac{1}{n_1}(n_0\theta_0 + n\bar{x}_n)$$
$$s_1^2 = s_n^2 + s_0^2 + (n_0^{-1} + n^{-1})^{-1}(\theta_0 - \bar{x}_n)^2.$$

On obtient donc le résultat suivant :

Proposition 2.4.

$$\pi(\theta|x, \sigma^2) \sim \mathcal{N}\left(\theta_1, \frac{\sigma^2}{n_1}\right),$$
$$\pi(\sigma^2|x) \sim IG\left(\frac{n+\eta+1}{2}, \frac{s_1^2}{2}\right).$$

On obtient bien une loi conjuguée. Il reste à savoir comment choisir en pratique les hyperparamètres θ_0, s_0^2, n_0 et ν.

2.4.3 Le cas multidimensionnel

On se place maintenant dans le cadre plus général où $X_1, ..., X_n \sim \mathcal{N}_p(\theta, \Sigma)$. Dans ce cas $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i \in \mathbb{R}^p$ et $S_n = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top \in \mathbb{R}^{p\times p}$. On a alors

$$f(x|\theta, \Sigma) \propto (\det \Sigma)^{-n/2} \exp\left(-\frac{1}{2}\left[n(\bar{x}_n - \theta)^\top \Sigma^{-1}(\bar{x}_n - \theta) + \mathrm{tr}(\Sigma^{-1} S_n)\right]\right).$$

Proposition 2.5. *On prend la loi a priori suivante*

$$\pi(\theta|\Sigma) \sim \mathcal{N}_p\left(\theta_0, \frac{\Sigma}{n_0}\right)$$
$$\pi(\Sigma^{-1}) \sim W_p(\alpha, V)$$

avec $W_p(\alpha, V)$, appelée loi de Wishart, la loi de $\sum_{i=1}^\alpha Z_i Z_i^\top$ si $Z_1, ..., Z_\alpha$ sont des v.a. i.i.d. de loi $\mathcal{N}_p(0, V)$. Alors on a les lois a posteriori suivantes :

$$\pi(\theta|x, \Sigma) \sim \mathcal{N}_p\left(\frac{n_0\theta_0 + n\bar{x}_n}{n_0 + n}, \frac{\Sigma}{n_0 + n}\right),$$
$$\pi(\Sigma^{-1}|x) \sim W_p\left(\alpha + n, V^{-1} + S_n + \frac{nn_0}{n + n_0}(\bar{x}_n - \theta_0)(\bar{x}_n - \theta_0)^\top\right).$$

3 Estimation

4 Simulation de loi *a posteriori*

Commençons par rappeler un (très) bref historique de la théorie de la statistique bayésienne :

— L'emergence des probabilités remonte au 17ème siècle tandis que les premiers travaux de statistique datent au 18ème siècle avec Bayes et Laplace. Il s'agit alors de statistique bayésienne.

— Au cours du 19ème siècle et du 20ème siècle les méthodes fréquentistes supplantent largement les méthodes bayésiennes.
— Depuis le début des années 1980, on note un retour très important de la recherche et des applications des méthodes bayésiennes.

On peut se demander pourquoi il a fallu attendre si tard pour que la statistique bayésienne revienne au premier plan. La raison est simple : la statistique bayésienne nécessite souvent des calculs potentiellement lourds ou infaisable lorsque l'on sort des exemples simples, il a donc fallu attendre que des méthodes de résolution numérique soient suffisamment performantes pour permettre d'obtenir des approximations numériques en des temps raisonnables.

Nous allons préciser tout cela. Dans toute la suite on note E l'espace des observations et Θ l'espace des paramètres, un sous-ensemble de $\mathbb{R}^p$. On rappelle les notations suivantes :

— modèle $f(x|\theta)$
— loi *a priori* $\pi(\theta)$
— loi *a posteriori* $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$.

En particulier on a

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{Z(x)}$$

avec la constante de renormalisation

$$Z(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta.$$

En pratique, le calcul de cette intégrale est potentiellement problématique, surtout si Θ est de dimension grande. Ce problème de calcul d'intégrale apparaît également ailleurs.

— Inférence : la moyenne a posteriori est donnée par

$$\mathbb{E}[\theta|x] = \int_\Theta \theta\pi(\theta|x)d\theta.$$

— Région de confiance :

$$\mathbb{P}(\theta \in S|x) = \int_S \pi(\theta|x)d\theta.$$

— Densités *a posteriori* marginales

$$\pi(\theta^1|x) = \int ... \int \pi(\theta^1, ..., \theta^n|x)d\theta_2...d\theta_n.$$

L'utilisation de méthodes de Monte-Carlo pour approximer numériquement ces intégrales a permis de sortir du cadre simple des lois conjuguées et d'élargir considérablement le spectre d'application des méthodes bayésiennes.

4.1 Méthodes de Monte Carlo

De manière générale on cherche à approcher la quantité (supposée bien définie)

$$I = \mathbb{E}[h(\theta)] = \int_\Theta h(\theta)g(\theta)d\theta$$

lorsque l'on connaît g la densité de θ. On suppose dans un premier temps que l'on sait échantillonner selon g. On note $\theta_1, ..., \theta_N$ un échantillon i.i.d. de cette loi.

Proposition 4.1. *La quantité*

$$\hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} h(\theta_i)$$

est un estimateur sans biais et fortement consistant de I.

On a également la normalité asymptotique de l'estimateur.

Proposition 4.2. *On note K la matrice de covariance de $h(\theta)$. Alors on a*

$$\sqrt{N}(\hat{I}_N - I) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, K).$$

En pratique on déduit du résultat précédent des régions de confiance. Pour cela on a néanmoins besoin de K, ou au moins une estimation de K en utilisant le lemme de Slutsky. On rappelle qu'un estimateur classique de K est donné par

$$\hat{K}_N^{i,j} = \frac{1}{N-1}\sum_{m=1}^{N}(h(\theta_m)^i - \overline{h(\theta)^i}_m)(h(\theta_m)^j - \overline{h(\theta)^j}_m).$$

Remarque 4.1.

- *La variance, et donc la précision de l'approximation, augmente linéairement avec la dimension. Pour les méthodes d'intégration numériques classiques (reposant sur des grilles), la précision augmente exponentiellement avec la dimension : c'est le fléau de la dimension (curse of dimensionality en anglais).*
- *Dans le cas des statistique bayésienne, on connaît g qu'à une constante de renormalisation près. On ne peut donc pas appliquer ces méthodes directement.*

4.2 Méthodes de Monte Carlo par chaîne de Markov (MCMC)

Le but des méthodes MCMC est d'approcher la loi g à l'aide d'une chaîne de Markov de mesure invariante g. On peut ensuite utiliser cela pour faire de l'estimation. En effet, si $Z_1, ..., Z_n \sim \pi(\theta|x)$ alors on peut prendre comme estimateur de θ

- $\hat{\theta}_n := \frac{1}{n}\sum_{i=1}^{n} Z_i$ (estimateur de Monte-Carlo),
- ou $\hat{\theta}_n :=$ mediane$(Z_1, ..., Z_n)$,
- ou $\hat{\theta}_n :=$ argmax hist$(Z_1, ..., Z_n)$.

L'idée générale des méthodes MCMC est de considérer en chaîne de Markov qui produit des échantillons corrélés

$$\theta_0 \to \theta_1 \to \theta_2 \to ...$$

tels que pour i suffisamment grand, θ_i suit a peu près la loi g.

4.2.1 Généralités sur les chaînes de Markov

On note $\mathcal{X}$ l'espace d'état. Dans la suite $\mathcal{X}$ est soit fini, soit infini dénombrable, soit c'est $\mathbb{R}^d$.

Définition 4.1. *Une chaîne de Markov $(X_0, X_1, ...)$ avec $X_i \in \mathcal{X}$ est une suite de variables aléatoires vérifiant*

$$f(X_{i+1}|X_i, ..., X_1) = K(X_{i+1}|X_i)$$

où $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, appelé noyau de Markov ou noyau de transition, vérifie : pour tout $x \in \mathcal{X}$, $x' \mapsto K(x'|x)$ est une densité de probabilité (ou loi discrète).

Exemple 4.1. *Une séquence de variable aléatoires $(X_i)_{i\in\mathbb{N}}$ est une marche aléatoire si elle satisfait*

$$X_{i+1} = X_i + \varepsilon_i$$

où $(\varepsilon_i)_{i\in\mathbb{N}}$ sont des variables i.i.d. Si la distribution des ε_i est symétrique autour de zéro, on parle de marche aléatoire symétrique.

Regardons ce qui se passe pour le cas $\mathcal{X}$ fini : $\mathcal{X} = 1, ..., p$. Dans ce cas, le noyau de transition K est une matrice A de dimension $p \times p$ telle quelle

$$A_{jk} = \mathbb{P}(X_{i+1} = k | X_i = j), \quad 1 \leqslant j \leqslant p, \quad 1 \leqslant k \leqslant p, \quad \forall i \in \mathbb{N}.$$

Cette matrice doit vérifier
— $0 \leqslant A_{jk} \leqslant 1, \quad 1 \leqslant j \leqslant p, \quad 1 \leqslant k \leqslant p,$
— $\sum_{k=1}^{p} A_{jk} = 1, \quad \forall \quad 1 \leqslant j \leqslant p.$
On note μ_0 la loi initiale de la chaîne, i.e. la loi de X_0.

Proposition 4.3.
— $\mu_{n+1} = \mu_n A$, $n \in \mathbb{N}^*$,
— $\mu_n = \mu_0 A^n$, $n \in \mathbb{N}^*$.

Si l'on revient au cas général, on a la proposition suivante :

Proposition 4.4.

$$\mu_{n+1}(x') = \int_{\mathcal{X}} \mu_n(x) K(x'|x) dx, \quad \forall x' \in \mathcal{X}.$$

Définition 4.2. *Une distribution g est dite invariante ou stationnaire par rapport à une chaîne de Markov si la chaîne laisse cette distribution invariante, i.e.*
— *Dans le cas fini, $g = gA$,*
— *dans le cas général $g(x') = \int_{\mathcal{X}} g(x) K(x'|x) dx$.*

Proposition 4.5. *Une condition suffisante (mais non nécessaire) pour garantir qu'une distribution est stationnaire est qu'elle vérifie la condition d'équilibre suivante :*

$$g(x) K(x'|x) = g(x') K(x|x'). \tag{15}$$

Preuve.

$$\begin{aligned} \int_{\mathcal{X}} g(x) K(x'|x) dx &= \int_{\mathcal{X}} g(x') K(x|x') dx \\ &= g(x') \int_{\mathcal{X}} K(x|x') dx = g(x'). \end{aligned}$$

□

Proposition 4.6. *Sous certaines conditions (vérifiées la plupart du temps) sur le noyau de transition, μ_n converge en loi vers la mesure invariante g. De plus on a le théorème ergodique :*

$$\frac{1}{N} \sum_{i=1}^{N} h(X_i) \to \int_{\mathcal{X}} h(x) g(x) dx \quad p.s.$$

Remarque 4.2.

- $\frac{1}{N}\sum_{i=1}^{N} h(X_i)$ *est un estimateur fortement consistant de* $\int_{\mathcal{X}} h(x)g(x)dx$ *mais pas nécessairement sans biais.*
- *Les* X_i *sont corrélées, contrairement au cadre classique d'application de la loi forte des grands nombres.*
- *Les premières v.a.* X_i *peuvent avoir une loi très éloignée de la loi* g*, il peut donc être intéressant de ne pas en tenir compte pour améliorer l'approximation :*

$$\frac{1}{N - N_0 + 1}\sum_{i=N_0}^{N} h(X_i) \to \int_{\mathcal{X}} h(x)g(x)dx \quad p.s.$$

4.2.2 Algorithme de Metropolis-Hastings

On revient au sujet d'étude initial. On suppose que g s'écrit $g(\theta) = \gamma(\theta)/Z$ avec Z une constante de renormalisation. On a également $\mathcal{X} = \Theta$. On veut une chaîne de Markov qui admette g comme mesure invariante et telle que Z n'apparaisse pas dans le noyau de transition. Pour cela on se donne un noyau de Markov $q(\theta'|\theta)$ et on considère l'algorithme de Metropolis-Hastings.

Algorithme 4.1.

- *On définie une valeur initiale* θ_0*,*
- *Pour* $i = 1, ..., N$
  1. *On propose une nouvelle valeur* $\theta^* \sim q(.|\theta_{i-1})$ *(loi de proposition)*
  2. *On calcule le taux d'acceptation*

$$\alpha(\theta^*, \theta_{i-1}) = \min\left(1, \frac{g(\theta^*)q(\theta_{i-1}|\theta^*)}{g(\theta_{i-1})q(\theta^*|\theta_{i-1})}\right)$$

  3. *Avec probabilité* α*, on prend* $\theta_i = \theta^*$ *et avec probabilité* $1 - \alpha$ $\theta_i = \theta_{i-1}$*.*

Proposition 4.7. *La mesure g vérifie la condition d'équilibre pour le noyau de Metropolis-Hastings.*

Preuve. On va montrer que le noyau K de la chaîne de Markov générée par l'algorithme de Metropolis-Hastings vérifie la condition d'équilibre (15). On commence par calculer le noyau $K(\theta'|\theta)$. remarquons que c'est une loi qui n'est ni a densité ($\theta' = \theta$ avec une probabilité potentiellement non nulle) ni discrète. On a

$$K(\theta'|\theta) = q(\theta'|\theta)\alpha(\theta', \theta) + \left(1 - \int_{\Theta} q(\theta'|\theta)\alpha(\theta', \theta)d\theta'\right)\delta_\theta.$$

Si $\theta' = \theta$, (15) est trivialement vérifiée. On suppose donc $\theta' \neq \theta$. Alors

$$\begin{aligned} g(\theta)K(\theta'|\theta) &= g(\theta)q(\theta'|\theta)\alpha(\theta', \theta) \\ &= \begin{cases} g(\theta)q(\theta'|\theta) & \text{si } g(\theta')q(\theta|\theta') \geqslant g(\theta)q(\theta'|\theta) \\ g(\theta')q(\theta|\theta') & \text{si } g(\theta')q(\theta|\theta') \leqslant g(\theta)q(\theta'|\theta) \end{cases} \\ &= \inf\left\{g(\theta)q(\theta'|\theta), g(\theta')q(\theta|\theta')\right\}. \end{aligned}$$

Donc par symétrie on a

$$g(\theta)K(\theta'|\theta) = g(\theta')K(\theta|\theta)$$

ce qui prouve le résultat. □

Remarque 4.3.

— *Le taux d'acceptation ne nécessite pas la constante de normalisation Z :*

$$\begin{aligned}\alpha(\theta^*, \theta_{i-1}) &= \min\left(1, \frac{g(\theta^*)q(\theta_{i-1}|\theta^*)}{g(\theta_{i-1})q(\theta^*|\theta_{i-1})}\right) \\ &= \min\left(1, \frac{\gamma(\theta^*)q(\theta_{i-1}|\theta^*)}{\gamma(\theta_{i-1})q(\theta^*|\theta_{i-1})}\right).\end{aligned}$$

— *Si le noyau q est symétrique, i.e. $q(\theta'|\theta) = q(\theta|\theta')$, alors le taux d'acceptation se simplifie*

$$\alpha(\theta^*, \theta_{i-1}) = \min\left(1, \frac{\gamma(\theta^*)}{\gamma(\theta_{i-1})}\right).$$

Donnons un exemple de loi d'acceptation. On considère pour la loi de proposition une marche aléatoire symétrique :

$$\theta^* = \theta_i + \varepsilon_i.$$

Dans ce cas le taux d'acceptation est donné par

$$\alpha(\theta^*, \theta_{i-1}) = \min\left(1, \frac{\gamma(\theta^*)}{\gamma(\theta_{i-1})}\right),$$

et en particulier on accepte toujours si $\gamma(\theta^*) > \gamma(\theta_{i-1})$. On peut appliquer cela à $g \sim \mathcal{N}(5, 1)$. Alors $\gamma(\theta) = \exp(-\frac{1}{2}(\theta - 5)^2)$. on prend des ε_i de loi $\mathcal{N}(0, \sigma^2)$. Le taux d'acceptation est donné par

$$\alpha(\theta^*, \theta_{i-1}) = \min\left(1, \exp\left[-\frac{1}{2}\left((\theta^* - 5)^2 - (\theta_{i-1} - 5)^2\right)\right]\right),$$

et l'initialisation par $\theta_0 = 0$.

On peut regarder l'influence du paramètre de réglage σ.

— Si σ est faible, l'échantillonneur fait des explorations locales (petits sauts) qui sont presque tous acceptées.

— Si σ est grand, l'échantillonneur fait des grands sauts mais qui sont acceptés avec probabilité faible.

— Quelque soit la valeur de σ, l'algorithme va converger vers la mesure stationnaire. Néanmoins σ influe sur la vitesse de convergence. Empiriquement le taux d'acceptation optimal est entre 0,1 et 0,6, il faut donc choisir σ en fonction.

4.2.3 Algorithme de Gibbs

Lorsque l'on souhaite simuler des lois multidimensionnelles il peut être utile de se ramener à des simulations uni-dimensionnelles : c'est le principe de l'échantillonneur de Gibbs. Soit $\theta = \begin{pmatrix} \theta^1 \\ \vdots \\ \theta^p \end{pmatrix}$.

Comment simuler $\theta \sim g$?

On note

$$\theta^{-j} = (\theta^1, ..., \theta^{j-1}, \theta^{j+1}, \theta^p)^\top \in \mathbb{R}^{p-1}.$$

On suppose que l'on sait échantillonner selon les distributions conditionnelles $g_j(\theta^j|\theta^{-j})$.

Algorithme 4.2.
Pour $i = 1, ..., N$
Pour $j = 1, ..., p$ *faire*

$$\theta_i^j \sim g_j(.|\theta_i^1, ..., \theta_i^{j-1}, \theta_{i-1}^{j+1}, ..., \theta_{i-1}^p).$$

Fin pour
Fin pour

On peut montrer que l'échantillonneur de Gibbs est un cas particulier de l'algorithme de Metropolis-Hastings.

Il est également possible de combiner les deux algorithmes pour obtenir un algorithme de type Metropolis-Hastings ou les composantes des variables sont mises à jour séquentiellement. On doit se donner une densité conditionnelle de proposition $q((\theta^j)^*|\theta)$.

Algorithme 4.3.
— *On définie une valeur initiale* $\theta_0 = (\theta_0^1, ..., \theta_0^p)^\top$,
— *Pour* $i = 1, ..., N$

1. *Pour* $j = 1, ..., p$ *on propose une nouvelle valeur* $(\theta^j)^* \sim q(.|\theta_i^{-j}, \theta_{i-1}^j)$ *(loi de proposition) avec*
$$\theta_i^{-j} = (\theta_i^1, ..., \theta_i^{j-1}, \theta_{i-1}^{j+1}, ..., \theta_{i-1}^p)^\top.$$
2. *On calcule le taux d'acceptation*
$$\alpha((\theta^j)^*, \theta_{i-1}^j) = \min\left(1, \frac{g(\theta_i^{-j}, (\theta^j)^*) q(\theta_{i-1}^j|(\theta^j)^*, \theta_i^{-j})}{g(\theta_i^{-j}, \theta_{i-1}^j) q((\theta^j)^*|\theta_{i-1}^j, \theta_i^{-j})}\right)$$
3. *Avec probabilité* α, *on prend* $\theta_i^j = (\theta^j)^*$ *et avec probabilité* $1 - \alpha$ *on prend* $\theta_i^j = \theta_{i-1}^j$.

5 modèles hiérarchiques

5.1 Introduction

En statistique bayésiennes, on fait l'hypothèse que des observations X sont des variables dont la loi est définie par un paramètre θ, lui même aléatoire et suivant une loi *a priori* π. On utilise alors les observations X et la loi a priori pour définir une loi a posteriori sur le paramètre θ pour ensuite effectuer une estimation de θ, par exemple par le maximum a posteriori. Dans ce cas, on cherche à estimer un unique paramètre.

Dans certaines situations on peut être amené à considérer des ensembles d'observations $(y_i)_{1 \leqslant i \leqslant N}$ dont des sous-ensembles $(y_j)_{j \in I_k}$ sont des variables aléatoires i.i.d. suivant une même loi définie par un paramètre θ_k. Ainsi à chacun des M sous-ensembles correspond un paramètre θ_i. On suppose que les $(\theta_i)_{i \leqslant M}$ sont tirés selon une loi de paramètre μ qui est inconnu mais lui même supposé aléatoire, selon une loi connu. Le paramètre μ est appelé *hyperparamètre*. On cherche alors à estimer chacun des paramètres à partir des observations.

Cette structure hiérarchique est un moyen d'introduire une dépendance entre les θ_i. En effet on suppose que les θ_i sont i.i.d. conditionnellement à μ, mais les θ_i ne sont pas indépendant. Ainsi, l'observation des $(y_j)_{j \in I_{k'}}$ apporte de l'information pour l'estimation de θ_k même si $k \neq k'$. On parle d'« emprunt d'information » ("Borrowing strength" en anglais).

Exemple 5.1. *On étudie l'efficacité d'un traitement cardiaque.*

- *Le patient i dans l'hôpital j a la probabilité de survie θ_j.*
- *Il est raisonnable de considérer que les θ_j, qui représentent un échantillon des probabilités de survie, devraient être liées entre elles. On suppose donc que les θ_j sont eux-même des échantillons d'une distribution de population, de paramètre inconnu μ.*
- *Grâce à l'utilisation de ce modèle hiérarchique, des observations de patients dans un hôpital j apportent de l'information sur les probabilités de survie dans d'autres hôpitaux.*

5.2 Justification théorique

L'utilisation de modèles hiérarchiques est une manière d'introduire de la dépendance entre les les θ_i. Elle peut également se justifier d'un point de vue théorique.

Définition 5.1. *Soient $(X_i)_{i\in\mathbb{N}^*}$ des variables aléatoires.*

Pour $n \geqslant 2$ fixé, $X_1, ..., X_n$ est dit échangeable si $(X_1, ..., X_n)$ a même loi que $(X_{\sigma(1)}, ..., X_{\sigma(n)})$ pour toute permutation $\sigma \in \mathfrak{S}_n$.

$(X_i)_{i\in\mathbb{N}^*}$ est dit échangeable si $(X_1, ..., X_n)$ sont échangeables pour tout $n \geqslant 2$.

Remarque 5.1.

- *Les $(X_i)_{i\in\mathbb{N}^*}$ sont échangeables si l'information contenue dans les $(X_i)_{i\in\mathbb{N}^*}$ est indépendante de l'ordre dans lequel les données sont collectées.*
- *Si les $(X_i)_{i\in\mathbb{N}^*}$ sont échangeables, les variables ont nécessairement même loi.*
- *Des variables i.i.d. sont échangeables.*
- *Soit $(X_1, ..., X_n)$ un vecteur gaussien de moyenne m et de matrice de covariance Σ. $(X_1, ..., X_n)$ est échangeable si et seulement si*
  - *toutes les composantes de m sont égales,*
  - *tous les éléments de la diagonale de Σ sont égaux,*
  - *tous les coefficients non diagonaux de Σ sont égaux.*

L'hypothèse d'échangeabilité a de fortes implications mathématiques. Un théorème du initialement à De Finetti et généralisé par Hewitt et Savage dit que si $X_1, ..., X_n$ sont des variables aléatoires réelles échangeables de distribution f. alors il existe une variable latente $\theta \in \Theta$ de loi π telle que les $X_1, ..., X_n$ sont indépendantes conditionnellement à θ. On a alors :

$$f(y_1, ..., y_n) = \int_\Theta \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta)d\theta.$$

- Ce théorème justifie l'approche bayésienne.
- Si les paramètres θ_i sont échangeables alors il exister un modèle paramétrique et il doit exister un a priori sur le paramètre du modèle : ce théorème justifie donc également l'approche des modèles hiérarchiques.
- θ peut être de dimension finie ou infinie... Le théorème ne donne qu'un résultat d'existence et d'unicité, il n'est pas constructif.

5.3 Un cas pratique

On considère N élèves du lycée d'une ville appartenant à M lycées différents.
Le niveau d'un élève $y_{i,j}$, $j$ème élève du lycée i peut être modélisé par une variable réelle $y_{i,j} =$

$\theta_i + \varepsilon_{i,j}$ où θ_i est le niveau moyen du lycée indicé par i et $\varepsilon_{i,j}$ est une variable gaussienne $\mathcal{N}(0,1)$ et où on suppose que les θ_i sont des variables i.i.d. suivant une loi $\mathcal{N}(\mu, \tau^2)$, où τ^2 est connu mais où μ est un hyperparamètre sur le quel on peut mettre un *a priori* .
Si on met un *a priori* uniforme sur μ, (la mesure de Lebesgue sur $\mathbb{R}$ est impropre mais peu être un *a priori* valide si la loi de θ est gaussienne), on peut estimer la loi *a posteriori* de la moyenne θ_j du lycée j à partir des élèves de ce lycée. Nous allons voir en TD que proposer un modèle hiérarchique c'est à dire, supposer que les différentes moyennes des lycées sont des variables aléatoires suivant une même loi, induit une corrélation sur les $(\theta_j)_{j \leqslant M}$ et que les estimations des différents θ_j vont faire intervenir, le niveau des élèves des autres lycées. Les deux cas extrêmes sont

- Le cas limite où $\tau = 0$ c'est à dire où on suppose que les niveaux des différents lycées sont tous identiques. L'estimation des différents θ_i utilisera de la même manière le niveaux de tous les élèves de tous les lycées.
- Le cas limite où τ tend vers $+\infty$, le niveau des différents lycées est très hétérogènes et dans ce cas, l'estimation de θ_j prend essentiellement en compte le niveau des élèves du lycée j.