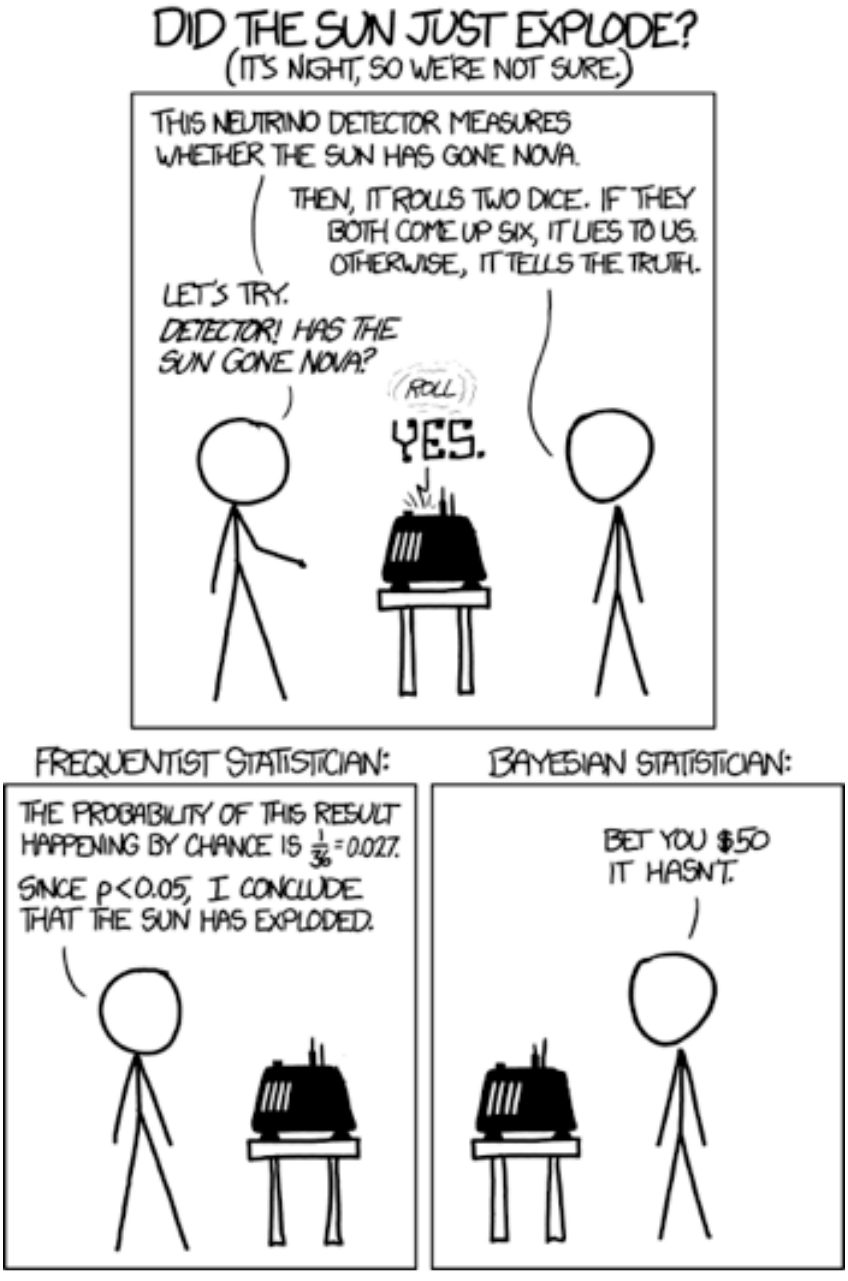
Cours Statistique Bayésienne - 2021

- M2 SITN -

March 4, 2021

Avertissement : Ces notes sont en cours de rédaction (et n'ont pas encore été relues:( ), il est probable qu'il y ait encore de nombreuses fautes d'orthographes (mes excuses), et quelques coquilles.



Contents

1 Introduction

Avant de rentrer dans le vif du sujet, on peut se demander à quelles sortes de questions cherchent à répondre les méthodes d'inférence statistique.

Sans prétentre à l'exhaustivité, toutes les méthodes visent à extraire de l'information de données, pour comprendre ou prédire un phénomène, par exemple à l'aide de

- Modélisation
- Estimation
- Intervalles de confiance
- Tests statistiques
- Classification
- Regression
- Sélection de modèles

Avertissement :

Dans une bonne partie du cours, on prendra comme exemple récurrent le fameux "Pile" ou "Face". En réalité, il faut bien voir qu'avec cet exemple simple peut se traduire n'importe quelle variable binaire (maladie ou non/vote/défaillance d'une pièce mécanique...). Voilà qui justifie l'intérêt de cet exemple jouet, dans ce cours, comme dans tous les cours de statistiques...

1.1 Paradigme Bayésien, modèle statistique et objectifs

On rappelle qu'un modèle paramétrique (fréquentiste) désigne un ensemble

$$\mathcal{M} = \{P_\theta, \theta \in \Theta\},$$

où $\Theta \subset \mathbb{R}^d$, et P_θ désigne une mesure de probabilité. Lorsque toutes les mesures P_θ sont absolument continues par rapport à la mesure de Lebesgue, on notera f_θ leur densité.

Lorsqu'on étudie les statistiques **fréquentistes**, on considère toujours qu'il existe un θ_0, inconnu, sous lequel les données ont été générées.

Remarque :

Sans information complémentaire, $\Theta \subset \mathbb{R}^d$. On n'insistera pas spécialement dessus, mais il faut bien avoir conscience dans tout le cours que le paramètre θ peut-être multidimensionnel, tout comme les X_i d'ailleurs !

Le paradigme bayésien consiste à modéliser l'incertitude que l'on a *a priori* sur θ_0, en le considérant comme aléatoire.

Certains diront que la différence fondamentale entre fréquentistes et bayésiens réside dans la vision même de ce qu'est une probabilité :

- Pour un-e fréquentiste, la probabilité d'un événement modélise la proportion du temps où l'évenement arriverait si l'on était capable de réaliser l'expérience un grand nombre de fois. Le/la fréquentiste cherchera donc à extraire l'information sur cette fréquence des données.
- Pour un-e bayésien-ne, la probabilité d'un évenement modélise notre croyance en l'occurence de cet évenement dans des conditions données. La/le bayésienne cherchera donc à actualiser cette croyance au vue des données. [1]

On introduit donc la notion de **prior** ou de **loi a priori** $\pi(\theta)$ qui sert à modéliser l'incertitude (ainsi que toute information éventuelle donnée a priori, c'est à dire avant l'expérience aléatoire) sur θ.

Une modèle bayésien est donc la donnée

- D'une loi a priori π sur Θ

[1] Remarquons qu'il y aurait beaucoup à dire sur ce changement de paradigme. On ne considère pas nécessairement que "la nature" a choisi préalablement au hasard un θ avant de procéder à une expérience aléatoire (réaliser un échantillon i.i.d. conditionnellement à θ), mais plutôt que tout modèle aléatoire sert à modéliser une incertitude...

- D'une famille de lois conditionnelles P_θ pour modéliser la loi conditionnelle des observations X_i sachant θ.

Reflexe 1 *Comme en statistique fréquentiste, avec de vraies données, le travail dépend du contexte. Dans un contexte exploratoire, on commencera*

1. *Par discuter avec les experts ayant obtenus les données, pour obtenir le contexte, le plan d'expérience, les informations déjà connues, les problèmes éventuels...*
2. *On pratiquera ensuite une longue analyse descriptive, en regardant toutes sortes de représentation graphique, en étudiant les indices caractéristiques, et en surveillant les valeurs manquantes et extrêmes (c'est aussi l'occasion de vérifier des problèmes de création dans la base de données).*
3. *Dans le cadre Bayésien, on pourra ensuite définir un modèle Bayésien, i.e. la donnée*
   - *De la loi des observations sachant le paramètre $\pi(X|\theta)$*
   - *D'un prior $\pi(\theta)$ sur le paramètre*
4. *Enfin, on pourra mettre en place différentes techniques pour chercher des éléments de réponses aux questions posées.*

Dans un contexte d'analyse statistique (à visée "conclusive"):

1. ***Avant d'extraire les données**, on clarifiera les questions auquelles on veut répondre. Puis on mettra en place un plan d'expérience, on choisira la liste des variables à utiliser, et les conditions expérimentales...*
2. ***Avant de regarder les données**, on écrira un plan d'analyse, décrivant toutes les méthodes que l'on souhaite mettre en place, et la façon dont on tirera les conclusions. En particulier, le modèle Bayésien (prior...), doit être donné à cette étape là.*
3. *Après avoir vérifié une construction correcte de la base de données, on effectuera l'analyse envisagée **en suivant le plan d'analyse** prévu.*

On considère donc que les observations sont i.i.d., **conditionnellement à** θ:

$$X_i \sim_{i.i.d.} f_{X|\theta}(.,\theta) := f(.|\theta).$$

Remarque importante :
Selon les ouvrages, les notations peuvent différer. En particulier, vous remarquerez qu'on s'autorisera souvent

- À utiliser la même notation pour des variables discrètes, ou à densité
- À utiliser la même notation pour plusieurs lois différentes. Par exemple $\pi(\theta)$ pour le prior, et $\pi(X|\theta)$ pour la loi conditionnelle des observations sachant θ, et $\pi(X_i|\theta)$ pour la loi conditionnelle de la i-ieme observation sachant θ. Dans ce cas là, c'est les variables qui indiquent quelle est la loi considéree. Lorsqu'on voudra oter toute ambiguité, on indiquera en indice les variables en questions (exemple : $f_{\theta,X}(t,x)$ pour la loi jointe, $f_{X|\theta}(x,t)$ pour la loi conditionnelle, et $f_\theta(t)$ pour la loi marginale)...

*Le choix des notations a été fait en fonction de ce qui me semble revenir le plus souvent dans les ouvrages que j'ai rencontré. Cela dit, les notations peuvent sembler confuses par moment, puisqu'on se ramène rarement à une $\mathbb{P}$ toute simple, mais qu'on résume souvent $\mathbb{P}(X = x|\theta = y) = f_{X|\theta}(x,y)$, par $\pi(x|\theta) = f_{X|\theta}(x,\theta)$, et en particulier, on perd l'info importante de ce qui est aléatoire ou non, à cause de l'abus de notation. Sauf mention contraire, au cours de ce poly, on n'aura **jamais** besoin d'appliquer réellement une de ces fonctions à une variable aléatoire, si bien que $\pi(X|\theta)$ désignera toujours une fonction, et non cette fonction appliquée à la variable aléatoire X.*

1.2 Loi a posteriori, calcul et utilisation

On rappelle que si (X,Y) a pour densité jointe $f_{X,Y}(x,y)$, alors

- $f_X(x) = \int_y f_{X,Y}(x,y)\mathrm{d}y$ désigne la loi marginale de X
- $f_{Y|X} = \frac{f_{X,Y}(x,y)}{f_X(x)}$ désigne la loi conditionnelle de Y sachant X.

En appliquant la formule, lorsque le prior π a pour densité f_θ par rapport à la mesure de Lebesgue on peut donc avoir la loi jointe de (θ, X) :

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

Et la calcul de la loi marginale donne

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

On peut alors se demander la loi de θ sachant X, le calcul donne

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

Cette loi est appelée **loi a posteriori**, ou simplement **posterior**. On la notera le plus souvent $\pi(\theta|X)$, et par abus de notation, on s'autorisera à écrire (que les variables soient discrètes ou continues) :

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

Remarque :
Lorsque le contexte le demandera, on s'autorisera même à écrire $\pi(\theta|X = x)$ pour $f_{\theta|X}(\theta, x)$. On s'efforcera de clarifier les notations lorsqu'il pourrait y avoir une ambiguité.

Reflexe 2 *Il est souvent pratique de calculer $\pi(\theta|X = x)$ à une "constante" près (qui dépend de x). Si l'on est capable d'écrire*

$$\pi(\theta|X = x) = C(x)h(\theta, x),$$

alors nécessairement, on a $C(x) = \frac{1}{\int_\theta h(\theta,x)\mathrm{d}\theta}$.
L'information sur $C(x)$ n'étant pas nécessaire, elle sera souvent omise pour simplifier calculs et notations. On notera alors d'un $\propto$ pour signifier la "proportionnalité"[a] :

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

[a] De temps en temps, je réécrirai ces formules avec des notations moins abusives pour clarifier. Par exemple, la formule suivante pourrait s'écrire :

$$f_{\theta|X}(t, x) = C(x)f_\theta(t)f_{X|\theta}(x, t) = C(x_1, \cdots x_n)f_\theta(t)\prod_{i=1}^{n} f_{X|\theta}(x_i, t)$$

Exercice 1 *On commence avec un cas simple : supposons que le modèle soit donné par*

- $X_i|\theta \sim_{i.i.d.} \mathcal{B}(\theta)$
- $\forall p \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}, \pi(\theta = p) = \frac{1}{6}$

1. *Donner la loi jointe de (θ, X_1), dans un tableau par exemple.*
2. *La première observation donne $X_1 = 1$, donner la loi a posteriori.*
3. *Après avoir observé $(x_1, x_2, x_3) = (1, 0, 1)$, pouvez-vous calculer la loi a posteriori ?*

Remarque :
À ce stade, on peut déjà noter la facilité de se placer dans un cadre séquentiel (les données arrivent "une par une"). En effet, il suffit de remarquer que (en utilisant que $\pi(X_1, \cdots, X_n|\theta) = \prod_{i=1}^n \pi(X_i|\theta)$ lorsque les X_i sont i.i.d. sachant θ).

$$\pi(\theta|X_1, \cdots, X_n) \propto \pi(\theta|X_1, \cdots, X_{n-1})\pi(X_n|\theta).$$

Reflexe 3 *Lorque l'on est confronté à un problème statistique, on a vu que le premier reflexe était de construire un modèle pour les données, puis de choisir un prior, sans utiliser les données pour choisir la forme du prior (pour le moment en tout cas, cf empirical bayes).*

Une fois le modèle posé, tout travail statistique bayésien sera basé sur ***le posterior*** $\pi(\theta|X)$

Dans le cas de données indépendantes, ce posterior peut être calculé ***à constante près*** *en utilisant*

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

On préfèrera souvent calculer le log :

$$\ln\Big(\pi(\theta|x)\Big) = C(x) + \ln\Big(\pi(\theta)\Big) + \sum_{i=1}^{n} \ln\Big(\pi(x_i|\theta)\Big)$$

Enfin, lorsque les données arrivent en plusieurs fois, on peut ***utiliser le postérior*** *calculé avec les premières données* ***comme prior*** *pour la nouvelle expérience.*

Cette loi a posteriori contient toute l'information qui nous intéresse sur les données. Ce sera donc le point de départ de tout travail. On utilisera ensuite ce posterior pour

1. Construire des estimateurs ponctuels si c'est souhaité (cf théorie de la décision)
2. Construire des régions de confiance
3. Construire des tests
4. Construire un classifieur
5. Selectionner des modèles
6. Effectuer une regression

Exercice 2 (En machine) *En conservant le prior uniforme pour p sur $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ pour le modèle $\mathcal{B}(p)$, et en utilisant les données de votre choix binaires (par exemple sur 'R'* `iris`*), représenter l'évolution du posterior, lorsqu'on augmente progressivement le nombre d'observations. Pour cela,*

- *On pourra faire une fonction* `computepost_prior_discret` *qui prend en argument un prior discret, un support (de ce prior), et des observations, et qui renvoie la loi a posteriori*
- *On pourra utiliser la fonction* `Sys.sleep` *(pour 'R') ou* `matplotlib.plt.pause` *(pour Python) dans une boucle pour observer l'évolution facilement*

Remarque : De la même façon que la loi a priori modélise à la fois notre incertitude et nos connaissances a priori sur le paramètre d'intérêt θ, la loi a posteriori modélise notre connaissance (et incertitude) **actualisées** en ayant pris en compte l'information contenue dans les données.

Une fois qu'on a la loi conjuguée, on peut se demander quelles croyances on a sur une éventuelle prochaine observation. Pour répondre à cette question, les fréquentistes utiliseraient un plug-in (en considérant que la loi d'une prochaine observation ressemble à $\pi(X|\hat{\theta})$, pour un estimateur $\hat{\theta}$ de θ.

En bayésien, on dispose d'un outil très naturel pour répondre à ce type de questionnement : **la loi prédictive**.

Définition 1 *Soit* $\Big(\mathcal{M} = \{\pi(X_1, \cdots X_n|\theta)\}, \pi(\theta)\Big)$ *un modèle bayésien. La* ***loi prédictive*** *d'une nouvelle observation X_{n+1} est définie par la loi de X_{n+1} sachant $X_1, \cdots, X_n$. Elle peut se calculer (dans le cas i.i.d. conditionnellement à θ), en posant*[2]

[2]Et avec des notations moins abusives :

$$f_{X_{n+1}|X_1,\cdots,X_n}(x, x_1, \cdots, x_n) = \int_{t\in\Theta} f_{X|\theta}(x,t) f_{\theta|X_1,\cdots,X_{n-1}}(t, x_1, \cdots, x_{n-1}) \mathrm{d}t = \mathbb{E}_{T\sim f_{\theta|X_1,\cdots,X_n}(.,x_1,\cdots,x_n)}\Big[f_{X|\theta}(x,T)\Big].$$

$$\pi(X_{n+1} = x|X_1, \cdots, X_n) = \int_\Theta \pi(X_{n+1} = x|\theta)\pi(\theta|X_1, \cdots, X_n) = \mathbb{E}_{\pi(\theta|X_1,\cdots,X_n)}[\pi(X = x|\theta)].$$

En particulier, la meilleure prédiction (au sens des moindre carrés) de X_{n+1} au vues des observations, est $\mathbb{E}[X_{n+1}|X_1, \cdots X_n]$, i.e. l'espérance d'une variable tirée sous la loi prédictive. Notons qu'en fréquentiste, l'indépendance des X_i fait que cette espérance n'est autre que l'espérance de X_1, qui dépend du paramètre θ inconnu...

Exercice 3 *On se place dans le cadre de l'exercice précédent (loi Bernoulli). Montrer que la loi prédictive est donnée par une Bernoulli de paramètre l'espérance à posteriori de p.*

1.3 Notion de priors conjugués

Vous remarquerez vite que dans le cas général, les calculs peuvent être compliqués. Il existe cependant une exception : le cas des priors conjugués.

On dit qu'une famille de lois $\mathcal{F} = \{\pi_h, h \in H\}$ est une famille de priors conjuguée pour un modèle $\mathcal{M} = \{\pi(X|\theta), \theta \in \Theta\}$ si pour tout prior $\pi \in \mathcal{F}$, et toute observation $x \in \mathbb{R}^n$, le posterior $\pi(.|x)$ appartient aussi à $\mathcal{F}$. Autrement dit :

$$\forall h \in H, \forall x \in \mathbb{R}^n, \exists h' \in H, \pi_h(.|x) = \pi_{h'}.$$

Le paramètre h est appelé **hyperparamètre**.

Exercice 4 *:*

1. *Montrer que les priors suivant sont des familles de priors conjuguées (pour les familles de lois données) :*
   - $\{Beta(a, b), a, b > 0\}$ *pour* p *dans un modèle* $\mathcal{B}(p)$.
   - $\{Beta(a, b), a, b > 0\}$ *pour* p *dans un modèle* $\mathcal{B}in(p)$.
   - $\{Beta(a, b), a, b > 0\}$ *pour* p *dans un modèle* $\mathcal{G}(p)$.
   - $\{\mathcal{N}(m, \sigma_0^2), m \in \mathbb{R}, \sigma_0 > 0\}$ *pour* μ *dans un modèle* $\mathcal{N}(\mu, \sigma^2)$, σ^2 *connu.*
   - $\{\Gamma(k, \lambda_0), k, \lambda_0 > 0\}$ *pour* λ *dans un modèle* $\mathcal{E}(\lambda)$.
2. *Trouver une famille de priors conjuguée pour chaque familles de lois suivantes :*
   - *pour* λ *dans un modèle* $\mathcal{P}(\lambda)$
   - *pour* τ *dans un modèle* $\mathcal{N}(m, \frac{1}{\tau})$, m *connu.*

Exercice 5 (En machine) *En utilisant les mêmes données que pour l'exercice précédent, observer de nouveau l'évolution du posterior en incluant séquentiellement les données, mais cette fois avec un prior uniforme* $Beta(1, 1)$.

Remarque :

1. Si $\mathcal{F}$ est une famille de priors conjuguée pour un modèle $\mathcal{M}$, alors tout la famille $\tilde{F}$ de mélanges de lois de $\mathcal{F}$ est aussi conjuguée pour $\mathcal{M}$.
2. En réalité, dans les familles exponentielles, on peut toujours construire une famille de priors conjuguée pour le paramètre naturel de la famille exponentielle

2 Utilisation de la loi a posteriori

On a vu qu'une fois le posterior calculé, on va l'utiliser dans différents cadres.

2.1 Régions de crédibilité

Le plus naturel à faire avec la loi a posteriori est de construire des **régions de crédibilité** dans lesquelles "on pense après avoir vu les données" que θ doit se trouver. Formellement, il s'agit de région de grande probabilité pour la loi a posteriori (conditionnellement aux données, donc).

Définition 2 *Une région $R_\alpha(X) \subset \Theta$ est appellée région de crédibilité au niveau α si*

$$\pi(\theta \in R_\alpha(X)|X) = 1 - \alpha.$$

Notons que $\pi(\theta \in R_\alpha(X)|X) = \int_{\theta \in R_\alpha(X)} \pi(\theta|X)\mathrm{d}\theta$

Remarque :

On voit ici que l'objectif semble très différent de l'objectif d'un intervalle de confiance fréquentiste. On rappelle qu'un intervalle de confiance fréquentiste au niveau α est un intervalle $I_\alpha(X)$ aléatoire (il dépend des observations) tel que

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

Dans le cadre fréquentiste, l'aléa considéré est donc entièrement sur les observations. Dans le cadre Bayésien, la probabilité d'une région de confiance est calculée avec l'aléa sur θ, conditionnellement aux observations. D'une certaine façon, à θ fixé, le θ sera dans l'intervalle fréquentiste une proportion $1 - \alpha$ du temps, alors que si on simulait le modèle en utilisant le prior π, en ne regardant que les simulations donnant X, pour X, fixé, θ serait dans la région bayésienne une proportion $1 - \alpha$ du temps. Dans les deux cas, en intégrant sur toutes les variables, on se rend compte que intervalles de confiance et régions de crédibilité vérifient

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

En réalité, comme dans le cadre fréquentiste, il n'y a pas d'unicité de la région de confiance. On peut procéder de deux façons pour la construire : soit construire un intervalle en utilisant les quantiles (comme pour un intervalle de confiance), soit utiliser les régions HPD pour "Highest Posterior Density"

Définition 3 *Une région HPD est une région de la forme*[3]

$$R_h^{(HPD)}(X) = \{\theta \in \Theta, \pi(\theta|X) \geq h\},$$

La région HPD de crédibilité de niveau α est donnée par $R_{h_\alpha}^{(HPD)}(X)$, où $h_\alpha = \sup\{h > 0, \pi(\theta \in R_h^{(HPD)}(X)|X) > 1 - \alpha\}$. On la notera plus simplement $R_\alpha^{(HPD)}(X)$.

Proposition 1 *Les régions HPD vérifient les propriétés suivantes :*

- *C'est la région de volume minimal*[4]*, parmi toutes les régions de crédibilité au niveau α.*
- *Si la densité a posteriori est continue et a un unique maximum local, alors toute région HPD est connexe.*

Exercice 6 *On considère le modèle très simple suivant :*

$$\pi(X|\theta) = \mathbb{1}_{x \geq \theta} e^{-(x-\theta)}.$$

Cela correspond à un modèle où $X|\theta \sim \theta + \mathcal{E}(1)$. On a observé les valeurs suivantes : $10, 12, 15$ (que l'on modélise comme des réalisation i.i.d. conditionnellement à θ).

1. *Point de vue fréquentiste : Quelle est la loi de $\bar{X} - \theta$? En déduire un intervalle de confiance fréquentiste pour θ.*[5]

[3] Dans l'écriture suivante, $\pi(\theta|X)$ désigne la densité de la loi a posteriori par rapport à une mesure de référence

[4] Par rapport à la mesure de référence utilisée pour seuiller la densité

[5] Pour être juste avec les fréquentistes, on peut noter que l'utilisation de $\bar{X}$ n'est pas la bonne façon de construire un intervalle fréquentiste, et que l'intervalle de confiance que l'on devrait construire utiliserait plutôt la loi de $\min(X_i)$ (la statistique exhaustive), qui vérifie $\mathbb{P}(\min(X_i) > t + \theta) = e^{-3t}$, ce qui mène à un intervalle de confiance plus logique.

2. *On se donne comme prior pour θ une $\mathcal{E}(1)$. Calculer la loi a posteriori, et constuire une région HPD.*

Exercice 7 *On se place dans un modèle $\mathcal{N}(m, \sigma^2)$, avec σ^2 connu, et on choisit pour m un prior $\mathcal{N}(0, 1)$.*

1. *Donner, en fonction des observations X et des quantiles d'une loi normale, la région HPD au niveau $\alpha = 0.05$.*

2. *Donner en fonction des observations X et des quantiles d'une loi normale, l'intervalle de confiance fréquentiste au niveau $\alpha = 0.05$*

Exercice 8 (En machine) *On considère les mêmes données, et le même prior qu'à l'exercice 5.*

1. *En parcourant différentes valeurs possibles pour les quantiles q_1, q_2 qui donnent des régions de crédibilité au niveau α, construire la région HPD (donc de volume minimal) au niveau $\alpha = 0.1$.*

2. *En parcourant différentes valeurs de seuil possibles, et en utilisant la fonction de répartion de la loi Beta, construire la région HPD au niveau $\alpha = 0.1$.*

3. *En simulant selon la loi a posteriori, et en utilisant les quantiles empiriques, construire la région HPD au niveau $\alpha = 0.1$.*

2.2 Estimation ponctuelle et théorie de la décision

Même si cela n'est pas naturel d'ignorer une grande partie de l'information présente dans la loi a posteriori, il arrive que le contexte (ou le boss) demande une estimation ponctuelle. Comme en statistique fréquentiste, plusieurs possibilités s'offrent à vous :

1. Calculer le MAP (Maximum A Posteriori). Analogue bayésien du maximum de vraisemblance, il correspond au mode de la densité a posteriori (cas continu) ou au θ le plus probable sous la loi a posteriori (cas discret) :

$$\hat{\theta}^{(MAP)}(x) = \arg\max_{\theta} \pi(\theta|X = x).$$

2. Calculer l'espérance a posteriori :

$$\hat{\theta}^{(Mean)}(x) = \mathbb{E}[\theta|X = x] = \int_{\Theta} \theta\pi(\theta|X = x)\mathrm{d}\theta.$$

3. Calculer la médiane a posteriori :

$$\hat{\theta}^{(Median)}(x) = Median\Big(\pi(\theta|X = x)\Big).$$

En réalité, le choix de l'estimateur ponctuel dépend du contexte. Pour comprendre comment choisir un estimateur ponctuel, on peut faire une brève introduction à la **théorie de la décision**.

On appellera décision une fonction (mesurable) des observations $x \mapsto \delta(x)$. En pratique, une décision peut être un estimateur (on veut alors que $\delta(x)$ soit "proche" de θ), un test (on a alors $\delta(x) \in \{0, 1\}$, et on veut que sous l'hypothèse nulle, $\delta(X)$ prenne rarement la valeur 1), ou même une région de l'espace...

On parle de théorie de la décision car on va se fixer une fonction de coût L, qu'on cherchera à minimiser. Ces fonctions de coût peuvent prendre des formes très différentes :

- Coût d'estimation quadratique (perte quadratique) : $L(\theta, \delta(X)) = \|\theta - \delta(X)\|_2^2$
- Coût d'estimation $\mathbb{L}^1$ (perte $\mathbb{L}^1$) : $L(\theta, \delta(X)) = \|\theta - \delta(X)\|_1$
- Coût asymétrique : $L(\theta, \delta(X)) = c_1(\theta - \delta(X))\mathbb{1}_{\theta-\delta(X)>0} + c_2(\delta(X) - \theta)\mathbb{1}_{\delta(X)-\theta>0}$
- Perte $0 - 1$: $L(\theta, \delta(X)) = \mathbb{1}_{\delta(X)\neq\theta}$
- Coût de test : $L(\theta, \delta(X)) = c_{fn}\mathbb{1}_{\delta(X)=0\&\theta\in H_1} + c_{fp}\mathbb{1}_{\delta(X)=1\&\theta\in H_0}$

- Coût de zone de crédibilité : $L(\theta, \delta(X)) = Vol(\delta(X)) + Vol(\Theta)(1 - \mathbb{1}_{\theta \in \delta(X)})$ (si $Vol(\Theta) < +\infty$)
- Coût prédictif : $L(\theta, \delta(X)) = \int_{x_{n+1}} Lp(x_{n+1}, \delta(X))\pi(x_{n+1}|\theta)\mathrm{d}x_{n+1}$, où $Lp(x_{n+1}, \delta(X))$ est une fonction de coût sur la prédiction, par exemple $Lp(x_{n+1}, \delta(X)) = \|x_{n+1} - \delta(X)\|_2^2$.

Définition 4 *1. Le **risque fréquentiste** d'une décision δ associé à une perte L s'écrit*

$$R_\delta(\theta) = \mathbb{E}[L\big(\theta, \delta(X)\big)|\theta] = \int_x L\big(\theta, \delta(x)\big)\pi(x|\theta)\mathrm{d}x.$$

*2. Le **risque a posteriori** d'une décision δ associé à une perte L s'écrit :*

$$\rho_\delta(X) = \mathbb{E}[L\big(\theta, \delta(X)\big)|X] = \int_\theta L\big(\theta, \delta(X)\big)\pi(\theta|X)\mathrm{d}\theta$$

*3. Le **risque intégré** d'une décision δ associé à une perte L s'écrit :*

$$\begin{aligned} r_\delta &= \mathbb{E}[L\big(\theta, \delta(X)\big)] \\ &= \mathbb{E}[R_\delta(\theta)] = \int_\theta \int_x L\big(\theta, \delta(x)\big)\pi(x|\theta)\mathrm{d}x\pi(\theta)\mathrm{d}\theta \\ &= \mathbb{E}[\rho_\delta(X)] = \int_x \int_\theta L\big(\theta, \delta(x)\big)\pi(\theta|x)\mathrm{d}\theta\pi(x)\mathrm{d}x \end{aligned}$$

*4. Si $\min_\delta r(\delta) < +\infty$, on définit une **règle de décision Bayésienne** pour le prior π comme une règle de décision qui minimise le risque intégré : δ_π est une règle de décision Bayésienne si*

$$\delta_\pi \in \arg\min_\delta r_\pi(\delta).$$

*5. Si δ_π est une règle de décision Bayésienne pour le prior π, on dit que $\delta_\pi(X)$ est un **estimateur Bayésien**.*

*6. On dit qu'une décision δ **est préférable** à une décision δ' si $\forall \theta \in \Theta, R_{\delta'}(\theta) \leq R_\delta(\theta)$, avec une inégalité stricte pour au moins un θ.*

*7. Une règle de décision δ est dite **admissible**, s'il n'y a pas de décision qui lui soit préférable.*

On voit que l'objectif Bayésien consiste en la minimisation d'une intégrale double, alors que son pendant fréquentiste a pour objectif de minimiser le risque fréquentiste (intégré seulement sur les données), quel que soit θ. Cet objectif est beaucoup plus compliqué (la plupart du temps, il n'existe pas de décision préférable à toute autre). On pourrait opter pour des stratégies minimax (qui favorise des solutions très prudentes).

Il est au contraire facile de construire un estimateur Bayésien (lorsque $\min_\delta r(\delta) < +\infty$), en conditionnant (comme d'habitude) par rapport aux données.

$$\delta_\pi^\star(x) \in \arg\min_\delta \rho_\delta(x).$$

Théorème 1 *La décision $\delta_\pi^\star$ est une décision bayésienne. Si de plus elle est définie de façon unique quelque soit x, alors c'est une décision admissible.*

Remarque :
Il existe aussi la notion de π admissibilité TODO ?.

Exercice 9 *1. On considère les fonction de coût données dans la partie précédente. Dans chacun des cas, donner la décision bayésienne qui minimise le risque intégré.*

2. Donner la fonction de coût associé à un classifieur.

Exercice 10 (En Machine) *Vous organisez une conférence, et vous devez effectuer les réservations.* 100 *personnes se sont inscrites (le maximum que vous acceptez[6]), mais d'expérience des* 2 *années précédentes, vous savez qu'il y a toujours des désistements de dernière minute. En effet, les années passées, sur* 100 *inscrit-e-s, seulement* 79 *et* 85 *étaient effectivement venues.*

La politique de réservation de l'hôtel est la suivante :

[6] votre conférence est très demandée

- *En cas de sur-réservation, vous devez payer* 300 *euros de dédommagement par chambre sur-réservée.*
- *En cas de sous-réservation, cela vous coûte* 100 *euros de plus par chambre non réservée à l'avance (vous ne bénéficiez plus de tarifs préférentiels).*

En bon mathématicien-ne, vous décidez de formaliser le problème afin de trouver le nombre de chambres à réserver.

1. *Si on note X_1 et X_2 la variable aléatoire correspondant au nombre de personnes effectivement venues les années précédentes, et X_3 celle correspondant au nombre qui vont venir cette année, par quelle loi peut-on modéliser ces variables ? (on notera N, p les paramètres).*
2. *En bon Bayésien-ne, vous allez considérer que X_1, X_2, X_3 sont i.i.d. conditionnellement à p, et mettre un prior sur p. Quelle famille de prior peut-on choisir pour que les calculs soient simples ? Quels hyperparamètres a_0, b_0 correspondent à un prior uniforme dans ce cadre ? À partir de maintenant, on travaille avec ce prior.*
3. *On note $\mathcal{D} = (X_1, X_2)$ les données observées, et $d(\mathcal{D})$ la décision correspondant au nombre de chambres réservées que vous devez prendre. Écrire la fonction* `Lp` *donnant un coût correspondant à la politique de l'hôtel, en fonction de X_3 et de la décision d.*
4. *Écrire la fonction* `L` *donnant, en fonction d'un paramètre p, et de la décision d, le coût prédictif.*
5. *Écrire la fonction* `rho` *donnant, en fonction des hyperparamètres a, b du posterior, et de la décision d, le coût à posteriori.*
6. *Écrire la fonction* `minim_post` *donnant, en fonction des hyperparamètres a, b du posterior, la décision optimale pour ce problème de décision.*
7. *En utilisant les données de l'énoncé, calculer les hyperparamètres du posterior, et donner la décision optimale.*
8. *Faire l'analyse théorique du problème : obtenir la décision optimale comme un quantile de la loi prédictive, en résolvant le problème de minimisation.*
9. *Vérifier la valeur numérique de $d(\mathcal{D})$ ainsi obtenue, en utilisant la fonction* `qbetabinom` *du package* `rmutil`*, qui donne les quantiles de la loi betabinomiale.*
10. *L'hôtel vous propose une nouvelle politique, plutôt que payer* 300 *euros par chambre en sur-réservation, il vous proposer de payer un forfait de* 1000 *euros en cas de sur-réservation (indépendamment du nombre de chambres). Que devient la décision optimale avec cette politique ? Quelle politique préféreriez-vous ?*

2.3 Tests et facteur de bayes

Le cadre des tests ressemble au cadre fréquentiste : Soit Θ_0, Θ_1 deux sous-ensembles disjoints de Θ. On supposera dans toute cette partie qu'on est dans le cas dit "propre", dans lequel le prior π vérifie $\pi\Big(\Theta \setminus \big(\Theta_0 \cap \Theta_1\big)\Big) = 0$. Un test sera donc défini par **l'hypothèse nulle** $H_0 : \theta \in \Theta_0$, et **l'hypothèse alternative** $H_1 : \theta \in \Theta_1$.

Exercice 11 *En considérant la perte $0-1$ $L(\theta, \delta(X)) = \mathbb{1}_{\theta \notin \Theta_{\delta(X)}}$, montrer que la décision bayésienne associée s'écrit*

$$\delta(X) = \arg\max_{\delta \in \{0,1\}} \mathbb{P}(\theta \in \Theta_\delta | X),$$

i.e. que cela correspond à l'ensemble le plus probable a posteriori.

Remarque : Dans le formalisme précédent, si on ne se place plus dans le cadre propre, ou même si $\pi(\Theta_0) = 0$, comme c'est le cas pour des lois à densité lorsque Θ_0 est ponctuel ($\Theta_0 = \{\theta_0\}$). Il faudra alors modifier un peu le cadre, où au moins changer la fonction de côut, pour pouvoir utiliser la théorie de la décision afin de construire le test.

Dans le cadre Bayésien, il n'est plus forcément naturel de rendre asymétrique l'hypothèse nulle et alternative. Plutôt que de calibrer un niveau et considérer la puissance associée, on s'intéressera plutôt aux probabilités à postériori $\pi(\Theta_0|X)$ et $\pi(\Theta_1|X)$.

Exercice 12 *Une entreprise veut calibrer un test pour détecter de la salmonelle dans la nourriture. Des études préliminaires montrent que le risque a priori de présence de salmonelle est faible (d'ordre* $\frac{1}{10000}$ *en prenant la fourchette la plus haute). On va donc utiliser cette information pour construire un prior sur* $\theta \in \{0,1\}$*. Après un procédé biochimique complexe, le test mesure une quantité de gaz émis dans une solution. On suppose que cette quantité suit, conditionnellement à* θ *une variable gaussienne de loi* $\mathcal{N}(\theta, \sigma^2)$*, avec* σ^2 *connu.*

Comme cela touche à un problème de santé publique, il est beaucoup plus grave de risquer de laisser passer un produit contaminé, plutôt que de détruire un produit sain (ou de réeffectuer un test). L'entreprise vous fournit une fonction de perte de la forme :

$$L(\theta, \delta(X)) = f_{01} \mathbb{1}_{\theta \in \Theta_0, \delta(X)=1} + f_{10} \mathbb{1}_{\theta \in \Theta_1, \delta(X)=0}.$$

1. *Donner la décision bayésienne qui minimise le risque.*
2. *Quelle est le niveau, au sens fréquentiste, de ce test ? (AN :* $\sigma^2 = 0.25$, $\frac{f_{10}}{f_{01}} = 100000$
3. *On a observé* $x = 0.4$ *donner* $\pi(\theta = 1|x)$ *et* $\pi(\theta = 0|x)$*.*

On pourra aussi construire une fonction de perte asymétrique, la plupart du temps donné par un tableau de la forme

$\theta \backslash \delta(X)$	0	1
0	$-Gain_{H_0 \text{ à raison}} = f_{00}$	$Perte_{H_1 \text{ à tord}} = f_{01}$
1	$Perte_{H_0 \text{ à tord}} = f_{10}$	$-Gain_{H_1 \text{ à raison}} = f_{11}$

.

Si on considère le risque à posteriori associé à la fonction de perte donnée dans le tableau précédent, on peut calculer le risque a posteriori, donné par :

$$\begin{aligned}\rho_\delta(X) =& \mathbb{1}_{\delta(X)=0}\Big(f_{10}\pi(\theta \in \Theta_1|X) + f_{00}\pi(\theta \in \Theta_0|X)\Big) \\ &+ \mathbb{1}_{\delta(X)=1}\Big(f_{01}\pi(\theta \in \Theta_0|X) + f_{11}\pi(\theta \in \Theta_1|X)\Big).\end{aligned}$$

La décision optimale s'obtient donc comme

$$\delta(X) = \arg\max_\delta \sum_{y \in \{0,1\}} \pi(\theta \in \Theta_y|X) f_{y\delta}$$

Avec un peu d'efforts, on peut montrer que cela revient à prendre comme décision

$$\delta(X) = 1 \text{ si } \frac{\pi(\theta \in \Theta_1|X)}{\pi(\theta \in \Theta_0|X)} \geq \frac{f_{0,1} - f_{0,0}}{f_{1,0} - f_{1,1}}.$$

Tout ce travail sert juste à justifier que l'information sur la fonction de coût est en réalité réduite au ratio $r = \frac{f_{0,1} - f_{0,0}}{f_{1,0} - f_{1,1}}$. On obtient donc une décision bayésienne $\delta(X, r)$ (où l'on met en évidence la dépendence en r pour montrer que cela dépend de la fonction de coût).

En fonction des coûts que l'on s'est fixé, on sera donc toujours ramenés à fixer un seuil sur

$$\frac{\pi(\Theta_1|X)}{\pi(\Theta_0|X)}.$$

La formule de Bayes donne

$$\frac{\pi(\Theta_1|X)}{\pi(\Theta_0|X)} = \frac{\pi(X|\Theta_1)}{\pi(X|\Theta_0)} \frac{\pi(\Theta_1)}{\pi(\Theta_0)}$$

La formule d'actualisation fait donc apparaître le **facteur de Bayes** défini par

$$BF = \frac{\pi(X|\Theta_1)}{\pi(X|\Theta_0)}.$$

Ce facteur correspond en réalité à la statistique du rapport de vraisemblance. Par contre, ce facteur n'est pas si évident à calculer, en effet, si Θ_1 est composite (pas un singleton), il est nécessaire d'effectuer un calcul d'intégrale :

$$\pi(X|\Theta_1) = \frac{\int_{\theta \in \Theta_1} \pi(X|\theta)\pi(\theta)\mathrm{d}\theta}{\pi(\Theta_1)}.$$

Remarquons que, si l'on définit les **odds** a priori et a posteriori par

$$\text{Odds}_{\text{prior}} = \frac{\pi(\Theta_1)}{\pi(\Theta_0)}$$
$$\text{Odds}_{\text{posterior}} = \frac{\pi(\Theta_1|X)}{\pi(\Theta_0|X)},$$

alors on a bien

$$\text{Odds}_{\text{posterior}} = BF * \text{Odds}_{\text{prior}},$$

ce qui explique probablement le succès de l'utilisation des facteurs de Bayes.

Reflexe 4 *Lorsque l'on veut effectuer un test dans un cadre bayésien :*

1. *En général, il suffit de maximiser le risque associé à la bonne fonction de perte choisie, la plupart du temps donné par un tableau de la forme*

$\theta \backslash \delta(X)$	0	1
Θ_0	$-Gain_{H_0 \text{ à raison}}$	$Perte_{H_1 \text{ à tord}}$
Θ_1	$Perte_{H_0 \text{ à tord}}$	$-Gain_{H_1 \text{ à raison}}$

2. *Si l'on veut tester $\theta = \theta_0$ contre $\theta = \theta_1$, on pourra conditionner le prior à l'événement $\theta \in \{\theta_0, \theta_1\}$.*

3. *Pour actualiser la statistique de test au fur et à mesure des observations, on pourra utiliser le facteur de Bayes (que l'on rencontrera de nouveau dans la partie sélection de modèles)*

Exercice 13 *On se place dans un modèle gaussien, à variance connue, avec un prior dans la famille conjuguée.*

1. *Donner explicitement, en fonction des hyperparamètres, du rapport de coût, et des quantiles d'une $\mathcal{N}(0,1)$, la décision bayésienne de test pour $\Theta_0 = \mathbb{R}^+$ et $\Theta_1 = \mathbb{R}^-$.*

2. *On se place maintenant dans un cadre fréquentiste : on suppose que les données sont issues du paramètre $\theta_0 = 0$. Exprimer en fonction des hyperparamètres, des quantiles, et de la fonction de répartition d'une loi normale centrée réduite, le niveau du test.*

2.4 Autres utilisations possibles du posterior

Selon le temps disponible, après avoir vu les notions de priors non informatif, de modèle hierarchique, et les outils numériques, on pourra appliquer toutes ces notions

1. Au cas de la régression linéaire
2. Au cas de l'étude de modèles de mélange
3. Au cas de la classification supervisée
4. Au cas de la classification non supervisée

3 Méthodes numériques

Dans cette partie (peut-être la plus importante), on se propose de survoler des méthodes numériques permettant de simuler la loi a posteriori (afin de lancer nos méthodes préférées dessus : région HPD, tests, estimation ponctuelles, minimisation de fonction de coût...).

Notons d'ores et déjà qu'il est déjà possible (facile) de la faire dans plusieurs cadres (et qu'on l'a déjà fait...).

1. Lorsque l'on sait calculer explicitement le posterior (par exemple prior conjugués). Dans ce cas, cela présente peu d'intéret, mais on peut effectivement facilement simuler la loi a posteriori

2. Lorsque l'espace des paramètres est de petite dimension, on peut la plupart du temps :
    - Calculer numériquement le postérior à une constante près en tout point que l'on souhaite. **Attention**, en général, on calculera son log (pour éviter d'obtenir 0 partout).
    - On peut utiliser une méthode des rectangles, en calculant le posterior sur une grille, pour pouvoir trouver la constante de normalisation qu'il faut. (on pourrait aussi faire une estimation par Monte-Carlo de cette constante)
3. Lorsqu'on connaît le posterior (sur une grille par exemple) à une constante près, mais que l'on peut calculer son max, il est aussi possible d'effectuer une méthode du rejet, en choisissant de préférence une bonne mesure de référence comme loi de proposition (on rappelle la méthode du rejet ci dessous).

Dans de nombreux cadres, l'espace des paramètres sera de dimension telle que toutes les méthodes précédentes ne peuvent plus être mises en oeuvre. De plus, il sera courant d'avoir des variables non observées dans le modèle. Cette section a pour objectif de donner quelques méthodes numériques pour contrer ces problèmes.

3.1 Principe de base de Monte-Carlo, rejet et algorithme ABC

Le principe de base des méthodes de Monte-Carlo (approximation numérique par estimation empirique) a déjà été utilisé tout au long du cours. Plus précisemment, lorsqu'à partir de simulations $\theta_1, \cdots, \theta_n$ de la loi a posteriori,

- On approche l'espérance a posteriori par la moyenne de l'échantillon simulé
- On approche un quantile a posteriori par un quantile empirique de l'échantillon simulé
- On approche la probabilité d'un ensemble par la fréquence de cet ensemble dans l'échantillon simulé
- $\cdots$

Dans tout ces cadres, on utilise la propriété fondamentale suivante : Si $X_i \sim_{i.i.d} \mu$, alors

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \approx \int h(x)\mathrm{d}\mu(x).$$

Cette propriété peut-être formalisée dans la proposition suivante, conséquence directe de la LGN et du TCL.

Proposition 2 *Soient X_i des variables i.i.d. de loi μ, et h une fonction telle que $\mathbb{E}[|h(X)|] < +\infty$. Alors*

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{p.s.} \int h(x)\mathrm{d}\mu(x) = \mathbb{E}[h(X)].$$

Si de plus $Var(h(X)) < +\infty$, alors on a en outre

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} h(X_i) - \int h(x)\mathrm{d}\mu(x)}{\sqrt{Var(h(X))}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Remarque : Remarquez qu'en particulier, si l'on sait simuler la loi a posteriori, on peut estimer la densité de la loi prédictive par $\frac{1}{n}\sum_{i=1}^{n} f_{x|\theta}(x,\theta_i)$.

Il reste donc à savoir simuler sous la loi a posteriori. On se place dans le cadre suivant, on suppose que $\pi(\theta|X) = C(X)q(\theta|X)$, où q est connue explicitement, et $C(X)$ n'est pas connu.

Reflexe 5 (Simulation en petite dimension) *En petite dimension, lorsqu'on connaît la loi du posterior à constante près $q(\theta|X)$, on peut*

1. *Soit calculer la constante de normalisation si nécessaire, avec la méthode de votre choix (par exemple méthode des rectangles). Remarquez qu'il faut ensuite savoir simuler une loi connue explicitement. On pourrait aussi simuler la version discrétisée (sur la grille utilisée pour calculer la constante de normalisation), mais en général, on préfère simuler en utilisant une autre méthode.*
2. *Soit simuler la loi a posteriori avec la méthode du rejet :*
   - *On se donne une loi de proposition $p(\theta)$, idéalement la plus proche possible du posterior.*
   - *On trouve M tel que $\sup_\theta \frac{q(\theta|X)}{p(\theta)} < M$.*
   - *On pose $i = 0$, et on tire θ_0 selon p, et $U_0 \sim \mathcal{U}([0,1])$.*
   - ***Tant que*** $MU_i > \frac{q(\theta_i|X)}{p(\theta_i)}$, *on incrémente i, et on tire θ_i selon p, et $U_i \sim \mathcal{U}([0,1])$.*

L'idée de cet algorithme est la suivante :

1. Prendre le premier couple (U, θ) qui vérifie une certaine condition revient à conditionner (U, θ) par cette condition.
2. La densité de (U, θ) sachant $MU_i \leq \frac{q(\theta_i|X)}{p(\theta_i)}$ est donnée par :
$$(u,t) \mapsto \frac{\mathbb{1}_{Mu \leq \frac{q(t|X)}{p(t)}} p(\theta)}{\mathbb{P}(Mu \leq \frac{q(t|X)}{p(t)})}.$$
3. En intégrant sur U, on obtient la marginale :
$$t \mapsto \frac{q(t|X)}{Mp(t)} \frac{p(t)}{\mathbb{P}(Mu \leq \frac{q(t|X)}{p(t)})} \propto q(t|X).$$

Remarque : Il est tentant d'utiliser le prior comme loi de proposition, afin de simplifier l'expression $\frac{q(\theta|X)}{p(\theta)}$ en $\prod_{i=1}^n \pi(x_i|\theta)$. Pour autant, cette méthode sera souvent lente, dès que le prior et le posterior sont loins (en effet, le nombre moyen de rejet est donné par $\frac{1}{\mathbb{P}(Mu \leq \frac{q(t|X)}{p(t)})}$, qui peut être très grand, dès que $\frac{q(t|X)}{p(t)}$ est très "piquée").

Propositions plus formelles et preuve :

Proposition 3 *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. i.i.d. à valeur dans $\mathbb{R}^d$, et $B \in \mathcal{B}(\mathbb{R}^d)$.On pose*
$$Z = X_k, \text{ où } k = \inf\{n \in \mathbb{N}, X_n \in B\},$$
alors
$$Z \sim X|X \in B.$$

Preuve : Ce résultat est assez intuitif, mais on peut rapidement prouver la proposition. On calcule alors, pour $C \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}
\mathbb{P}(Z \in C) &= \mathbb{P}(\exists k \geq 0, X_0 \notin B, X_1 \notin B, \cdots, X_{k-1} \notin B, X_k \in C) \\
&= \sum_{k\geq 0} \mathbb{P}(X_0 \notin B, X_1 \notin B, \cdots, X_{k-1} \notin B, X_k \in C) \\
&= \sum_{k\geq 0} \mathbb{P}(X_0 \notin B)\mathbb{P}(X_1 \notin B)\cdots\mathbb{P}(X_{k-1} \notin B)\mathbb{P}(X_k \in C) \\
&= \sum_{k\geq 0} \mathbb{P}(X_0 \notin B)^{k-1}\mathbb{P}(X_k \in C) \\
&= \frac{\mathbb{P}(X_0 \in C)}{1 - \mathbb{P}(X_0 \notin B)} \\
&= \frac{\mathbb{P}(X_0 \in C)}{\mathbb{P}(X_0 \in B)} \\
&= \frac{\mathbb{P}(X_0 \in C \text{ et } X_0 \in B)}{\mathbb{P}(X_0 \in B)} \\
&= \mathbb{P}(X_0 \in C | X_0 \in B)
\end{aligned}$$

□

On a en particulier

Corollaire 1 *Soit $A \in \mathcal{B}(\mathbb{R}^d)$, et $B \subset A$ borélien. Soit $(U_n)_{n\in\mathbb{N}}$ une suite de v.a. i.i.d. de loi uniforme sur A, on pose :*

$$Z = U_k, \quad \text{où } k = \inf\{n \in \mathbb{N}, U_n \in B\},$$

alors

$$Z \sim \mathcal{U}(B).$$

On peut donner un exemple classique :

Proposition 4 *Soit X une variable aléatoire réelle de densité f. Soit $(Y_n)_{n\in\mathbb{N}}$ une suite de v.a. réelles i.i.d de densité g, telle que $\frac{f}{g}$ est majorée par $M \geq 0$ (en particulier, g est non nulle dès que f est non nulle). Soit $(U_n)_{n\in\mathbb{N}}$ une suite de v.a. i.i.d de loi uniforme sur $[0, M]$. On pose*

$$\begin{aligned} k &= \inf\{n \in \mathbb{N}, U_n \geq \frac{f}{g}(Y_n)\} \\ Z &= Y_k. \end{aligned}$$

Alors, $Z \sim X$.

Preuve : En fait, si on pose $V = U_k$, où $k = \inf\{n \in \mathbb{N}, U_n \geq \frac{f}{g}(Y_n)\}$, on a

$$(Z, U) = (Y, U)|U \leq \frac{f}{g}(Y).$$

Pour cela, on remarque que la densité de (Y, U) s'écrit $(y, u) \mapsto \frac{g(y)}{M}\mathbb{1}_{[0,M]}(u)$, d'où on peut déduire la densité de (Z, U) en calculant la densité conditionnelle :

$$(z, u) \mapsto \frac{\frac{g(y)}{M}\mathbb{1}_{[0,M]}(u)\mathbb{1}_{u\leq\frac{f(y)}{g(y)}}}{\mathbb{P}\big(U \leq \frac{f(Y)}{g(Y)}\big)}.$$

Il reste à calculer la densité h de la loi marginale de Z, en intégrant par rapport à u :

$$\begin{aligned} h(z) &= \int_{u=0}^{+\infty} \frac{\frac{g(z)}{M}\mathbb{1}_{[0,M]}(u)\mathbb{1}_{u\leq\frac{f(z)}{g(z)}}}{\mathbb{P}\big(U \leq \frac{f(Y)}{g(Y)}\big)} \\ &= \frac{g(z)}{M} \frac{\int_{u=0}^{\frac{f(z)}{g(z)}} \mathrm{d}u}{\int_{z\in\mathbb{R}} \frac{g(z)}{M} \int_{u=0}^{\frac{f(z)}{g(z)}} \mathrm{d}u\mathrm{d}z} \\ &= \frac{g(z)}{M}\frac{f(z)}{g(z)} \frac{1}{\int_{z\in\mathbb{R}} \frac{f(z)}{M}\mathrm{d}z} \\ &= f(z). \end{aligned}$$

□

Utilisations : Le premier exemple d'utilisation possible que l'on peut donner est pour simuler une v.a. de loi $\mathcal{N}(0, 1)$ (et donc de loi $\mathcal{N}(m, \sigma^2)$ avec une modification affine) à partir de simulations de suites de v.a. uniformes sur $[0, 1]$ indépendantes $(U_n)_{n\in\mathbb{N}}, (V_n)_{n\in N}, W$.

- On peut tout d'abord simuler une suite de v.a. i.i.d. de loi exponentielle de paramètre 1 en posant

$$Y_n = -\ln(U_n).$$

- On peut ensuite simuler une v.a. de même loi qu'une valeur absolue de loi normale $\mathcal{N}(0, 1)$ en applicant la méthode précédente, avec

$$\begin{aligned} f(x) &= \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ g(x) &= e^{-x} \\ \frac{f}{g}(x) &\leq \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2-2x}{2}} \\ &\leq \frac{2e}{\sqrt{2\pi}} e^{-\frac{x^2-2x+1}{2}} \\ &\leq \frac{2e}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} \\ &\leq \frac{2e}{\sqrt{2\pi}} := M \end{aligned}$$

On procède donc comme suit :

1. On pose $i = 0$
2. On regarde si
$$MV_i \leq \frac{f(Y_i)}{g(Y_i)} \Leftrightarrow V_i \leq e^{-\frac{(Y_i-1)^2}{2}}.$$
3. Si la condition précédente est vérifiée, on pose $T = Y_i$, sinon on incrémente i, et on recommence au 1.

- À ce stade, on a simuler une valeur absolue de loi normale centrée réduite. Il suffit maintenant de choisir un signe avec probabilité $\frac{1}{2}$. On pose donc
$$Z = \mathbb{1}_{W \leq \frac{1}{2}} T - \mathbb{1}_{W \geq \frac{1}{2}} T.$$
Et on a réussi à simuler Z de loi $\mathcal{N}(0,1)$.

Il existe une autre méthode élémentaire pour simuler la loi a posteriori, basée elle aussi sur la méthode du rejet. Il s'agit de la méthode ABC (pour Approximate Bayesian Computation). Le principe est d'utiliser une méthode du rejet pour simuler, pour ϵ fixé, et pour une distance ρ bien choisie, la loi $\pi\Big(\theta|\rho(X,x) \leq \epsilon\Big)$, afin d'approximer la loi $\pi(\theta|X=x)$.

En particulier, lorsque l'on dispose d'une bonne statistique $S(X)$ résumant l'échantillon (en particulier si l'on dispose d'une statistique suffisante), on pourra prendre $\rho(x,y) = d(S(x),S(y))$, avec d une distance bien choisie.

Reflexe 6 (Algorithme ABC) *Une solution pour simuler la loi a posteriori, lorsque l'on*

- *dispose d'une bonne statistique résumant les données (par exemple statistique suffisante) $S(X)$.*
- *dispose d'un bon critère de similarité ρ entre les données (souvent donné par $\rho(x,y) = d(S(x),S(y))$*
- *sait facilement simuler le prior, et les données conditionnellement à un paramètre θ.*

On peut alors appliquer l'algorithme ABC pour simuler la loi de θ sachant $(x_1, \cdots x_n)$:

1. *On choisit la distance ρ, ou la statistique de résumé S et la distance d.*
2. *On fixe un seuil de tolérance $\epsilon > 0$.*
3. *Pour $i = 0$, on tire θ_i selon le prior, et $X_{i,1}, \cdots, X_{i,n}$ sachant θ_i.*
4. ***Tant que*** $d\big(S(X_{i,1}, \cdots, X_{i,n}), S(x_1, \cdots, x_n)\big) > \epsilon$, *on incrémente i, puis on tire θ_i selon le prior, et $X_{i,1}, \cdots, X_{i,n}$ sachant θ_i.*
5. *On peut poser $T = \theta_i$ (et on recommence si on veut un échantillon).*

On a alors $T \sim \pi(\theta|d(S(X),S(x)) < \epsilon)$.

Exercice 14 (En machine) *Avant de résoudre des problèmes complexes, il faut savoir résoudre des problèmes simples.*

On se propose donc de reprendre les données de votre choix, ainsi quela colonne de votre choix (binaire), et d'utiliser un prior $Beta(a,b)$ pour ces données.

1. *On a déjà vu que, dans ce cas là, le posterior était connu explicitement (famille de priors conjugués). Représenter cette densité.*
2. *Simuler un échantillon par la méthode du rejet, selon la loi a posteriori, en utilisant une loi de proposition uniforme (donner le temps de calcul pour un échantillon de taille $N = 10000$, en utilisant la fonction* `Sys.time`*. Représenter une estimation de la densité de cet échantillon simulé.*
3. *Simuler un échantillon par la méthode du rejet, selon la loi a posteriori, en utilisant le prior comme loi de proposition. Donner le temps de calcul, et représenter une estimation de la densité.*
4. *Faire de même en utilisant l'algorithme ABC.*
5. *Vous pouvez refaire de même avec une variable continue, par exemple modélisée par une loi normale.*

3.2 Metropolis-Hasting et Gibbs

Pour l'instant, on a parlé de Monte-Carlo, qui utilise une simulation d'un échantillon i.i.d.. En réalité, la même idée peut-être utilisé dans des cas (faiblement) dépendants. Le cas qui nous intéresse particulièrement est celui de MCMC pour Markov Chain Monte Carlo, qui utilise l'approximation $\frac{1}{n}\sum_{i=1}^{n} h(X_i) \approx \int h(x)\mathrm{d}\mu(x)$, où

- X est une chaîne de Markov (à espace d'état fini, dénombrable, ou continu)
- μ désigne la mesure invariante de la chaîne X.

En réalité, l'approximation est moins bonne que dans le cas i.i.d.. On préfèrera donc une simulation exacte lorsque ce sera possible. De plus, il est préférable de "jeter" les premières simulations, pour laisser le temps à la chaine de converger vers la mesure invariante μ.

Il reste tout de même à construire une chaîne de Markov admettant pour mesure invariante la loi a posteriori, que l'on suppose connaître seulement à constante près.

Le principe est donné ci-dessous

Reflexe 7 (Metropolis-Hasting) *Pour simuler la loi a posteriori, supposée connue à constante près : $\pi(\theta|X) = C(X)q(\theta|X)$.*

1. *On pose $i = 0$, on initialise T_0 de façon aléatoire (selon la loi de notre choix). On se donne une famille de loi de proposition (appellées parfois lois intrumentales) $p(t, .)$.*
2. *Pour $i \geq 0$,*
   - *On tire Y selon la loi $p(T_{i-1}, .)$.*
   - *On calcule*
   $$\alpha = \min\Big(1, \frac{q(y|x)p(y, t_i)}{q(t_i|x)p(t_i, y)}\Big).$$
   - *Avec probabilité α, on pose $T_{i+1} = Y$ (acceptation), et avec probabilité $1 - \alpha$, on pose $T_{i+1} = T_i$ (rejet).*
   - *On incrémente i.*

Remarque : On peut choisir une loi de proposition constante ($p(x, y)$ ne dépend pas de x), voire même symétrique, de telle sorte que le calcul de α devient

$$\alpha = \min(1, \frac{f(y)}{f(x)})$$

Propositions plus formelles et preuve :

Pour cela, comme dans l'algorithme du rejet, il nous faut des lois de propositions $p(x, y)$, pour tout $x \in E$. On suppose que p est la matrice de transition d'une chaîne de Markov irréductible et apériodique. On suppose de plus que

$$p(x, y) > 0 \Leftrightarrow p(y, x) > 0.$$

Proposition 5 *Soit X_n la chaîne de Markov définie par la matrice de transition suivante :*

$$\forall x \neq y, Q(x, y) = p(x, y)\min\Big(1, \frac{f(y)p(y, x)}{f(x)p(x, y)}\Big).^{7}.$$

Alors X_n est irréductible, récurrente positive et apériodique, et de probabilté invariante μ. Donc pour toute loi initiale μ_0,

$$\forall e \in E, \mathbb{P}(X_n = e) \to \mu(e).$$

En particulier, quand p est symmétrique ($p(x, y) = p(y, x)$), Q devient :

$$\forall x \neq y, Q(x, y) = p(x, y)\min\Big(1, \frac{f(y)}{f(x)}\Big).$$

Preuve : On va d'abord montrer que μ est réversible pour Q. On remarque d'abord que si $p(x, y) = 0$, alors $p(y, x) = 0$, et on a bien dans ce cas là

$$\mu(x)Q(x, y) = 0 = \mu(y)Q(y, x).$$

[7] On a $Q(x, y) = 0$ si $p(x, y) = 0$, donc la définition ne pose pas de problème

Soit $x \neq y \in E$ tels que $p(x,y) > 0$. On calcule

$$\begin{aligned}
\mu(x)Q(x,y) &= \mu(x)p(x,y)\min\Big(1, \frac{f(y)p(y,x)}{f(x)p(x,y)}\Big) \\
&= \min\Big(\mu(x)p(x,y), \mu(x)p(x,y)\frac{f(y)p(y,x)}{f(x)p(x,y)}\Big) \\
&= \min\Big(\mu(x)p(x,y), Cf(y)p(y,x)\Big) \\
&= Cf(y)p(y,x)\min\Big(\frac{\mu(x)p(x,y)}{Cf(y)p(y,x)}, 1\Big) \\
&= \mu(y)p(y,x)\min\Big(\frac{f(x)p(x,y)}{f(y)p(y,x)}, 1\Big) \\
&= \mu(y)p(y,x)Q(x,y).
\end{aligned}$$

Donc μ est réversible pour Q, donc invariante. Comme $p(x,y) > 0 \Rightarrow Q(x,y) = 0$, la chaîne est aussi irréductible et apériodique. La récurrence positive provient de l'existence d'une probabilité invariante.

□

Remarque : Le choix de la loi de proposition est crucial car il détermine le temps de calcul de l'algorithme. Lorsqu'on choisit une loi de proposition dans une famille paramétrique, il sera important de bien calibrer la variance. Deux cas extrèmes sont à éviter :

- La proposition est acceptée pratiquement tout le temps, et la suite $(\theta_n)_{n\in\mathbb{N}}$ suit lentement une tendance, alors la loi de proposition a probablement une variance trop faible : on progresse très lentement vers des zones de plus en plus probables.
- Si la proposition n'est jamais acceptée (la suite $(\theta_n)_{n\in\mathbb{N}}$ est quasiment constante), alors la loi de proposition a probablement une variance trop grande : on propose toujours des paramètres dans une zone de probabilité très petite.

On peut considérer qu'un taux de rejet correspondant à une loi de proposition "bien calibrée" serait entre 50% et 75% Vous verrez dans les exercices, que la chaîne peut mettre un certain temps à converger vers la mesure stationnaire. Il sera souvent nécessaire de "jeter" le début de la chaine simulée, afin d'améliorer l'estimation.

Exercice 15 (En machine) *Une dernière fois avec les mêmes données, avant d'aborder des exemples plus complexes... On considère toujours les mêmes données, on souhaite simuler sous la loi a posteriori par la méthode de MH.*

1. *On se propose d'utiliser comme loi de proposition* `rprop` *une loi normale* $p_{prop} \sim \mathcal{N}(p, \sigma^2)$*, avec* $\sigma = 0.01$*, que l'on pourra changer pour différents tests. Comme* $p \in [0,1]$*, on conditionne cette loi à se trouver dans* $[0,1]$*. Créer la fonction* `rprop`
2. *Créer la fonction* `dlogpost` *qui calcule le log de la loi a posteriori à une constante près. On propose d'utiliser le prior de Jeffreys comme loi a priori.*
3. *Créer une fonction* `rprox` *qui simule* θ_{i+1} *en fonction de* θ_i *en utilisant Metropolis-Hasting. Remarque : On pourra dans un premier temps ignorer le conditionnement à* $p \in [0,1]$ *pour calculer le ratio qui apparaît dans MH, en faisant l'approximation entre la loi de proposition et une loi normale, afin d'être dans le cas simplifié (loi de proposition symétrique et constante), dans lequel* $\alpha = \min(1, \frac{f(y)}{f(x)})$*.*
4. *Simuler un échantillon sous la loi a posteriori avec la méthode de MH, observer la trajectoire simuler, ainsi que l'estimation de la densité (et la comparer avec la densité à posteriori calculée théoriquement).*

Exercice 16 (En machine) *On va travailler sur les données* `Howell1_censured.csv` *qui recense la taille, le poids, l'âge et le sexe de plusieurs individus.*[8]

1. *Extraire dans un vecteur* h *la taille de tous les individus âgés de* 18 *ans ou plus.*
2. *Observer la répartition de* h*. Quel modèle proposeriez-vous pour la loi de* h*, et pourquoi ?*

[8] "The data contained in data (Howell1) are partial census data for the Dobe area !Kung San, compiled from interviews conducted by Nancy Howell in the late 1960s."

3. *On se propose de modéliser ces données par un mélange Gaussien, c'est à dire de densité donnée par* $pf_{\mathcal{N}(m_1,\sigma_1^2)} + (1-p)f_{\mathcal{N}(m_2,\sigma_2^2)}$. *On va paramétrer la variance par* $\tau_i = 1/\sigma_i^2$. *On se propose de mettre comme prior*

   - *Un prior* $Beta(\frac{1}{2},\frac{1}{2})$ *pour* p
   - *Un prior* $\mathcal{N}(0,100^2)$ *pour* m_1, m_2
   - *Un prior* $Gamma(1,1/10)$ *pour* τ_1, τ_2.

   Écrire la fonction `logpost` *qui à une liste* `l` *de paramètres* `l$p,l$m1,l$m2,l$t1,l$t2`*,et un vecteur de données* `x`*, associe le log du posterior (calculé à une constante près) pour ces paramètres, et ces observations.*

4. *On souhaite mettre en oeuvre l'algorithme de Metropolis-Hasting. Pour cela, il faut choisir une loi de proposition. On utilisera, pour des paramètres de dispersion à fixer*

   - $pprop|p \sim Beta(1+\frac{p}{v_p}, 1+\frac{1-p}{v_p})$
   - $mprop_i|m_i \sim \mathcal{N}(m_i, v_{m_i})$
   - $\tau prop_i|\tau_i \sim Gamma(1+\frac{\tau_i}{v_{\tau_i}}, \frac{1}{v_{\tau_i}})$

   'Ecrire la fonction `rprop`*, qui à une liste de paramètre courante, renvoie une proposition de nouveau paramètres,* ***ainsi que le log du ratio*** $\frac{Q(l,lprop)}{Q(lprop,l)}$.

5. *Faire tourner l'algorithme de Metropolis Hasting pour estimer un mélange Gaussien, sur données simulées.*

6. *Faire de même avec les vraies données. Comment ajouter l'information a priori que la répartition hommes/femmes est proche d'être* 50/50 *?*

7. *Donner la loi du postérior pour les différents paramètres, et commenter les résultats, à l'aide de représentations graphiques pertinentes.*

Exercice 17 (En machine) *On utilise dans cette exercice les données* `protein.RData` *qui vous ont été fournies. Ces données mesurent (à une constante près) des nombres de protéines observées dans des cellules. On cherche à modéliser la loi de* X*, et d'après les experts ayant obtenu les données, les cellules peuvent avoir deux comportements relativement différents.*

On se donne donc comme modèle un modèle de mélange gaussien :

- $\theta = \Big(\alpha, m_1, m_2, \tau_1, \tau_2\Big)$
- $X|\theta$ *de densité*
$$f_{X|\theta}(x) = \alpha f_{\mathcal{N}(m_1,\frac{1}{\tau_1})}(x) + (1-\alpha) f_{\mathcal{N}(m_2,\frac{1}{\tau_2})}(x).$$
- *On prendra comme priors (indépendants) ceux donnés par l'expert :*
  - $\alpha \sim Beta(a,b)$
  - $m_i \sim \mathcal{N}(\mu, \sigma_0^2)$
  - $\tau_i \sim Gamma(k,\lambda)$
- *Enfin, on prendra les hyper-paramètres suivants :*
  - $a = \frac{1}{2}, b = \frac{1}{2}$
  - $\mu = 100$
  - $\sigma_0^2 = 100$
  - $k = 1$
  - $\lambda = 1000$

1. *Écrire une première fonction* `logvraiss` *qui donne, en fonction d'une observation* x_i *et de* θ *le log de la densité* $f_{X|\theta}(x_i)$ *au point* x_i

2. *Écrire une fonction* `logdpost` *qui calcule (à une constante près) le log de la densité à posteriori*

3. *Écrire une fonction* `rprop` *qui simule une propostion θ_{prop} en fonction d'un point θ_{curr}, et d'écart-types σ_{prop}. On prendra comme loi de proposition, pour chaque paramètre, $\theta^{(i)}_{prop}|\theta^{(i)}_{curr} \sim \mathcal{N}(\theta^{(i)}_{curr}, \sigma^{(i)}_{prop})$, en conditionnant chaque paramètre à rester dans le "bon" intervalle, à ce que $m_2 > m_1$, et on pourra prendre $\sigma_{prop} = c(1/1000, 0.5, 1/10000)$.*

4. *Simuler selon la loi a posteriori avec la méthode de Métropolis-Hasting. On pourra représenter la loi marginale a posteriori pour chaque paramètre, ainsi que les trajectoires de m_1, m_2.*

Lorsque la loi de $\pi(\theta|X)$ se factorise en produit de $\pi(\theta^{(i)}|X)$, on peut naturellement utiliser une loi de proposition qui change un seul des $\theta^{(i)}$ (par exemple, on tire au hasard un des paramètres que l'on va changer). Cette idée naturelle peut aussi être mise en pratique lorsque les $\theta^{(i)}$ ne sont pas indépendants selon la loi a posteriori, en prenant plutôt la loi conditionnelle. C'est le principe de l'algorithme de Gibbs.

Plus précisemment, (pour l'instant on oublie le cadre bayésien), si on souhaite simuler selon une densité $f_{X_1,\cdots,X_n}(x_1,\cdots,x_n)$, on peut procéder de la façon suivante :

1. On initialise $x^{(0)}$ aléatoirement.
2. On simule $x_1^{(1)}$ selon la loi $f_{X_1|X_2,\cdots,X_n}(x_1^{(1)}, x_2^{(0)} \ \cdots, x_n^{(0)})$
3. Pour $i = 1, \cdots n$, on simule $x_i^{(1)}$ selon la loi $f_{X_i|X_1,\cdots X_{i-1},X_{i+1},\cdots,X_n}(x_1^{(1)}, x_2^{(1)} \cdots, x_{i-1}^{(1)}, x_i^{(1)}, x_{i+1}^{(0)}, \cdots, x_n^{(0)})$
4. On recommence à partir de l'étape 2 pour $x^{(j)}$, pour $j > 1$.

Remarque :

1. On peut montrer (exercice ?) que cet algorithme simule une chaîne de Markov, de mesure invariante à densité f. Comme dans le cas de MH, le théorème ergodique nous permet d'utiliser des estimation MCMC, et sera nécessaire de "jeter" le début de la chaîne ainsi simulée.
2. En pratique, on peut aussi modifier l'algorithme pour à chaque instant actualiser une composante aléatoire : on tire i uniformément, et on tire $X_{t+1}^{(i)}$ selon la loi conditionnelle de $X_i|(X_j)_{j\neq i}$, en laissant les autres variables inchangées.
3. Cet algorithme est particulièrement intéressant lorsque les lois conditionnelles (univariées) sont facile à simuler.

Pour appliquer cet algorithme dans le cas qui nous intéresse (simuler une loi a posteriori), on peut se contenter de remarquer que l'échantillonneur de Gibbs est un cas particulier d'une méthode de MH, et reformuler l'algorithme en choisissant une loi de proposition qui modifie seulement un des paramètres (choisi aléatoirement) par temps.

Reflexe 8 (Metropolis-Hasting – version Gibbs) *On pourra utiliser cette technique lorsque θ est de grande dimension, en particulier lorsqu'au moins une des conditions suivantes est rescpectée :*

1. *Les $\theta^{(i)}$ jouent des rôles "symétriques" (même type de paramètres pour même famille de lois)*
2. *Quelque soit i fixé, le ratio $\frac{\pi(\theta_{prop}|X)}{\pi(\theta_{curr}|X)}$ se simplifie, dès que $\forall j \neq i, \theta^{(j)}_{prop} = \theta^{(j)}_{curr}$.*

Dans ce cas, on reprend l'algorithme de MH, mais en se donnant une loi de proposition construite de la façon suivante :

1. *On tire i aléatoirement entre 1 et d (la dimension de θ)*
2. *On tire $\theta^{(i)}$ selon une loi de proposition univariée, si possible symétrique et constante $f_{prop}(\theta^{(i)}|\theta_{curr})$ ne dépend pas de θ_{curr}).*

Dans ce cas, le paramètre $\alpha = \min(1, \frac{\pi(\theta_{prop}|X)}{\pi(\theta_{curr}|X)})$ se simplifie souvent, comme on va le voir dans l'exemple suivant.

Exercice 18 (En machine) *On reprend l'idée du mélange Gaussien, mais cette fois, on souhaite utiliser l'algorithme de Gibbs pour échantillonner sous la loi à posteriori.*

1. *On écrit le modèle avec variables latentes Z_i. Donner le prior semi-conjugué (i.e. telles que les conditionnelles complètes soient dans la même famille de lois) pour p, m_k, τ_k.*
2. *Est-ce que les priors de Jeffreys sont dans ces familles ? Pour quels paramètres ? S'agit-il de priors impropres ?*
3. *Que deviennent les posteriors ?*
4. *Implémenter l'algorithme de Gibbs utilisant les conditionnelles complètes. On pourra s'aider de [5] p840*

Exercice 19 (En machine) ***Attention**, cet exercice pourrait en soi faire un TP (un peu long), donc pas de stress sur la difficulté.*

On considère les images binaires `dessin*.jpg` *qui vous ont été fournies.*

Ces images sont corrompues, l'objectif de cet exercice est d'utiliser une méthode bayésienne (basique) pour le débruitage.

On cherche à estimer, pour chaque pixel (i, j), le paramètre θ_{ij} (binaire, ± 1) correspondant au pixel. Une fois n'est pas coutume, ce paramètre aura un sens directement interprétable, vu qu'on considère qu'il s'agit du pixel d'une image sous-jascente (avant bruitage). Notons ici qu'il serait peut-être naturel de considérer plutôt cela comme une variable cachée (latente), et d'utiliser un modèle hierarchique pour vraiment se situer dans un cadre bayésien. On ne se soucie pas de ce problème dans cet exercice.

Notre modèle sera, pour chaque pixel X_{ij} (binaire aussi, $X_{ij} = \pm 1$), que la loi de X_{ij} sachant θ_{ij} est donnée par $\mathbb{P}(X_{ij} = x_{ij}|\theta_{ij}) = C(\theta_{ij})e^{\beta\theta_{ij}x_{ij}}$, où $C(\theta_{ij})$ désigne la bonne constante de normalisation.

Il faut ensuite se donner un prior sur θ. Cette fois, notre prior ne sera pas entièrement factorisable, vu qu'on veut que notre prior prenne en compte l'information suivante a priori : "Deux pixels "proches" sont probablement de la même couleur (noir ou blanc). Pour cela, on se donne une fonction de distance d entre pixels, et une fonction de dépendance ϕ (telle que $\phi(0) = 0$), et on pose $\eta_{(i,j),(k,l)} = \phi\Big(d\big((i,j),(k,l)\big)\Big)$.

Cette fonction a pour but de mesurer la dépendance entre deux pixels, de sorte que, pour deux pixels à distance d, plus $\phi(d)$ est grande, plus la dépendance est grande.

On se donne ensuite le prior suivant sur θ :

$$\pi(\theta) = C \exp\Big(\gamma \sum_{(i,j),(k,l)} \eta_{(i,j),(k,l)}\theta_{ij}\theta_{kl}\Big).$$

1. *Calculer $C(\theta_{ij})$, et montrer que cette contante ne dépend en réalité pas de θ_{ij}.*
2. *Que faudrait-il calculer pour obtenir explicitement la constante de normalisation C. Quel problème rencontre-t-on ? Peut-on obtenir facilement la probabilté d'une configuration θ donnée ?*
3. *On se propose de construire un algorithme de MH - version Gibbs pour simuler selon la loi a posteriori. Donner une expression du posterior, et montrer que si $\theta' = \theta$ en chaque pixel sauf (i, j), alors le ratio qui apparaît dans le calcul de Metropolis-Hasting se simplifie en*

$$\frac{\pi(\theta'|X)}{\pi(\theta|X)} = \exp\Big((\theta'_{ij} - \theta_{ij})\big(\beta X_{ij} + \gamma \sum_{(k,l)} \eta_{(i,j),(k,l)}\theta_{kl}\big)\Big).$$

4. *À partir de cet exercice, on se choisit la distance suivante : $d\big((i,j),(k,l)\big) = \max(|k-i|, |l-j|)$, et $\phi(d) = \mathbb{1}_{0<d\leq 2}$. Construire une première fonction qui calcule $\sum_{(k,l)} \eta_{(i,j),(k,l)}\theta_{kl}$.*
5. *Vérifier que la loi de proposition suivante $p(\theta_{prop}|\theta_{curr})$ vérifie bien $p(x, y) = p(y, x)$:*

$$(\theta_{prop})_{ij} = (-1)^{\mathbb{1}_{U_1=i,U_2=j}}(\theta_{curr})_{ij}, \text{ avec } U_i \sim \mathcal{U}[\![1, n_i]\!],$$

6. *Construire une fonction* `stepGibbs` *qui construit une étape de MH-Gibbs pour ce modèle.*

7. *Appliquer l'algorithme aux images fournies. On pourra par exemple essayer les paramètres suivants :* $\beta = 1, \gamma = 0.1$.

8. *Essayer votre algorithme sans attache au données (avec* $\beta = 0$*), pour différents* γ*. Qu'est ce qu'on simule dans ce cas là ? Tester aussi de changer* η *en choisissant* $\eta_{(i,j),(k,l)} = \psi((k-i, l-j))$*, avec le* ψ *de votre choix (essayer des fonctions non isotropes). Qu'observe-t-on ?*

Remarque *: En pratique, ces modèles (ou plutôt leur version améliorées) sont plus utilisés pour la segmentation d'image. (à faire en bonus ?)*

3.3 Les packages

De nombreux packages sont disponibles pour faire du Gibbs sampling après spécification du modèle. Les plus célèbres sont sans doute *OpenBUGS* et *JAGS*.

En pratique, une fois que l'on a bien compris tout le cours, il est intéressant d'apprendre à les utiliser, afin de diminuer le temps de programmation. Au lieu de spécifier manuellement la vraisemblance, les lois de proposition, et le prior, on utilise une forme de méta langage pour définir un modèle (dans un fichier joint, d'extension .bug) par exemple

```
model {
for (i in 1:N) {
z[i] ~ dbern(p)
y[i] ~ dnorm(z[i]*theta1+(1-z[i])*theta2, 1)
}
theta1 ~ dnorm(0, 1)
theta2 ~ dnorm(0, 1)
p ~dbeta(1,1)
}
```

Exercice 20 (En machine) *Installer JAGS, et essayer de reprendre les exercices classiques en utilisant ce package.* ***Exemple*** *: Enregistrer le modèle précédent dans un fichier model1.bug dans le répertoire courant, puis écrire*

```
library(rjags)
N =1000
b = rbinom(N,1,0.3)
y = b*rnorm(N,0,1)+(1-b)*rnorm(N,4,1)
m <- jags.model("model1.bug", data=list(y=y, N=length(y)))
par(mfrow = c(1,3))
for (param in c("p","theta1","theta2")){
  s <- coda.samples(m, param, n.iter=100)
  plot(density(s[[1]]),main = param,xlab = param)
}
```

4 Choix de priors et construction de modèles

Dans cette partie, on insiste un peu plus sur les choix de priors. On rappelle que le prior est censé résumer en une distribution l'information disponible a priori, c'est à dire **avant d'observer les données**. On rappelle que le choix du prior est d'autant plus important que peu de données sont disponibles. En effet, on a vu qu'asymptotiquement, l'influence du prior était négligeable (Bernstein von Mises), ce qui se voit au travers du log postérior

$$\frac{1}{N}\log(\pi(\theta|X)) = C(X) + \frac{1}{N}\log(\pi(\theta)) + \frac{1}{N}\sum_{i=1}^{N}\pi(X_i|\theta).$$

Toutefois, lorsque l'espace des paramètres est très grand, le choix du prior peut permettre d'imposer ou de renforcer certains modèles (sparsité, dépendance entre les paramètres...)

4.1 Priors subjectifs

Une première façon de choisir un prior est d'utiliser une information extérieure pour le construire. On parle alors de **priors informatifs**. Dans ce cas, plusieurs informations peuvent être utilisées :

1. Le prior **doit** nécessairement charger la "vraie valeur". En effet, comme on l'a vu, si $\pi(\theta_0) = 0$ alors $\pi(\theta_0|X) = 0$ (vrai en discret comme dans le cas à densité). Il ne faudra jamais oublier cette contrainte !

2. Le choix de la famille de lois peut-être fait par expérience (avis d'experts ou données antérieures), mais aussi un peu par défaut (c'est le cas pour les familles conjuguées par exemple).

3. Dans le cas de priors paramétriques (c'est aussi vrai pour le choix des lois dans les familles paramétriques), le choix des hyperparamètres peut se faire, en utilisant des informations a priori sur la localisation du paramètre que l'on cherche, qui permettra de calibrer moyenne ou médiane du prior, et sur l'étendue des paramètres possibles, qui permettra de calibrer la variance. Il se peut aussi que l'on rajoute des informations (sparsité, queues lourdes des distribution) pour autoriser les lois a posteriori à se localiser facilement dans certaines régions.

4.2 Notion de prior impropre

Il arrive souvent que l'on souhaite utiliser un prior qui ne soit pas une mesure de probabilité ($\pi(\Theta) = +\infty$). L'exemple le plus simple est celui d'un prior normal, dans lequel on souhaiterait la variance la plus grande possible.

Il est possible d'utiliser de tels priors, dès lors que **la loi a posteriori est intégrable**. On parle alors de **priors impropres**.

Définition 5 *Une loi a priori impropre est une mesure sur Θ qui vérifie*

$$\pi(\Theta) = +\infty$$
$$\int_\Theta f_{X|\theta}(x,t)\mathrm{d}\pi(t) < +\infty \ p.s.$$

Dans ce cas, la loi a posteriori est bien définie comme une mesure de probabilité, par

$$\pi(\theta = t|X = x) =_{not} f_{\theta|X}(t,x) = \frac{f_{X|\theta}(x,t)}{\int_\Theta f_{X|\theta}(x,t)\mathrm{d}\pi(t)}$$ [9]

Exercice 21
1. *On considère un modèle poisson, et on considère le prior impropre sur λ de densité par rapport à la mesure de Lebesgue donné par $\lambda^2 e^{2\lambda}$. Combien faut-il d'observations pour que le posterior soit intégrable ? Donner la loi a posteriori, et son mode.*

2. *On se place encore dans le cas normal, à variance connue, et on considère maintenant un prior uniforme sur $\mathbb{R}$ (la mesure de lebesgue). Donner la loi a posteriori.*

3. *On considère un modèle $\mathcal{G}(\frac{1}{\theta})$, avec un prior uniforme sur θ. Donner le posterior, et son mode.*

Remarque : Lorsqu'on utilisera un prior impropre, on s'efforcera d'être prudent aux utilisations faites. En effet, ils peuvent parfois exhiber des comportements contre-intuitifs (essentiellement dès que les calculs intermédiaires font intervenir des quantités non intégrables). On pourra consulter par exemple le paradoxe de marginalisation (voir Le choix Bayésien par exemple, ou le cours de Judith Rousseau, qui présente un exemple ou l'utilisation de statistique exhausives mène à des calculs faux à cause de la non intégrabilité du prior).

On sera aussi très prudents avec l'utilisation de facteurs de Bayes dans le cas impropre !! Il existe une solution simple pour se ramener "artificiellement" à un cas de prior intégrable : Utiliser les premières données disponibles jusqu'à obtenir un posterior intégrable, puis l'utiliser comme prior.

[9] Là encore, on laisse le lecteur extrapoler la notation pour le cas où le prior, ou le modèle, n'est pas à densité.

4.3 Notion de prior non informatif

Dans de nombreux contexte, on ne souhaite pas inclure trop d'information a priori avec le prior. Par exemple dans le cas de Bernoulli, il arrive que l'on ait absolument aucune information a priori, dans ce cas là, le prior peut-être délicat à définir.

La première idée naturelle dans ce cas là, c'est d'utiliser un prior uniforme (impropre si Θ n'est pas borné). Cet a priori est appellé **a priori de Laplace**. Il présente l'avantage d'être invariant par translation, mais possède toutefois un inconvénient majeur : il n'est pas invariant par reparamétrisation.

Exercice 22 *On reprend l'exemple précédent d'un modèle $\mathcal{G}(p)$, en prenant un prior uniforme sur p. Donner le posterior, et son mode. Avec le changement de variable $p = \frac{1}{\theta}$, donner la forme du prior sur θ.*

On cherche donc une façon de définir un prior $f_\theta = \Phi(f_{X|\theta})$, à partir d'une loi donnée $f_{X|\theta}$, de telle façon à ce qu'un changement de paramètre donne la même loi. En 1d, cela implique :

$$\Phi(f_{X|\psi(\theta)})(\psi(\theta))|\psi'(\theta)| = \Phi(f_{X|\theta})(\theta).$$

Une solution possible est d'utiliser le prior de Jeffreys, défini ci-dessous :

Définition 6 *Soit $\mathcal{M} = \{f_{X|\theta}, \theta \in \Theta\}$ un modèle paramétrique (régulier). On rappelle la définition de l'information de Fisher :*

$$I(\theta) = Var_\theta\Big(\nabla_\theta l_\theta(X)\Big) = -\mathbb{E}\Big[\nabla_\theta \nabla_\theta^T l_\theta(X)\Big],$$

où $l_\theta(x) = \log(f_{X|\theta}(x))$[10].

On définit le prior de Jeffreys par

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

Ce prior a tendance à favoriser les θ les plus faciles à estimer.

On remarque qu'il se peut que ce prior soit impropre (il faut alors vérifier que le posterior est bien intégrable).

Exercice 23 *Vérifier que le prior de Jeffreys pour le modèle $\mathcal{B}(p)$ est bien invariant par reparamétrisation $\theta = \frac{1}{p}$.*

En réalité, cette propriété est générale :

Proposition 6 *Soit $\mathcal{M} = \{f_{X|\theta}, \theta \in \Theta\}$ un modèle paramétrique régulier, et f_θ le prior de Jeffreys. Soiit ψ un $\mathcal{C}^\infty$-difféomorphisme, et f_η le prior de Jeffreys associé au modèle reparamétré par $\eta = \psi(\theta)$: $\mathcal{M}' = \{f_{X|\eta}\eta \in \psi(\Theta)\}$.*

On a alors

$$f_\eta(\psi(t)) \det\big(\nabla\psi(t)\big) \propto f_\theta(t),$$

i.e. la mesure image par ψ de l'a priori de Jeffreys pour $\mathcal{M}$ est aussi l'a priori de Jeffreys pour $\mathcal{M}'$: Le prior de Jeffreys est invariant par reparamétrisation.

En réalité, le prior de Jeffreys possède d'autres propriétés intéressantes. Par contre, comme on peut le voir dans l'exemple suivant, il s'avère moins utilise lorsque Θ est de dimension plus grande que 1.

Exercice 24 *On considère un modèle de régression classique :*

$$Y = X\beta + \epsilon,$$

avec $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

[10] Au passage, l'information de Fisher est en réalité la Hessienne de l'entropie relative (i.e. divergence de Kullback), elle même correspondant au rapport de vraisemblance asymptotique. Elle est donc "grande" en un point θ lorsque deux modèles sont facilement distinguables au voisinage de ce point. Elle n'est pas invariante par reparamétrisation non plus.

1. *On rappelle la densité de* $Y|\sigma^2, \beta$ *:*

$$f_{Y|\sigma,\beta}(y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\Big(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\Big).$$

   Calculer l'information de Fisher du modèle.

2. *En déduire la forme du prior de Jeffreys.*

3. *Que vaut l'information de Fisher du modèle pour* β*, si l'on considérait* σ *connu ? Et pour* σ*, en considérant* β *connu ?*

En pratique, il est parfois conseillé de plutôt faire le produit des priors que l'on aurait pour chaque paramètre pris individuellement.

Remarque : Il existe plusieurs autres façon de définir un prior "objectif" ou "non informatif", parmi lesquelles on peut citer

- Invariance de groupe (mesure de Haar)
- Le maximum d'entropie
- L'approche de Bernardo
- La maximisation de distance entre prior et posterior
- Matching priors (qui consiste à choisir un prior tel que les régions de crédibilité aient les meilleurs propriétés fréquentistes possibles).
- Mélange de priors non informatifs (ou moyenne géométrique)
- Modèles hiérarchiques

Dans ce cours, on ne développera que le dernier point, les autres sont seulement énoncé pour donner les mots clés (pour savoir quoi chercher le jour où vous en aurez besoin, voir par exemple [1]).

Exercice 25 (En machine) ***À propos de l'utilisation de mélange de lois en priors***

On se place dans le modèle Bernoulli $Ber(p)$*, et on choisit cette fois un prior sous la forme*

$$\pi(p) = \frac{1}{4}Beta(1,1) + \frac{3}{4}Beta(100,1).$$

On suppose que l'on a observé $x = (1,0,1,0,1)$*.*

- *Tracer la loi a posteriori.*
- *Comparer avec le cas où l'on aurait pris en prior seulement* $Beta(100,10)$

4.4 Modèles hierarchiques

L'exemple précédent fournit une bonne introduction pour cette partie. Lorsque l'on a peut d'information a priori, une alternative aux priors non informatifs est l'utilisation de modèles hierarchiques. Cela consister à considérer les hyperparamètres eux mêmes aléatoires, et construire un **hyperprior** pour ceux là.

Le cadre de l'exercice 25 peut-être reformulé comme suit :

1. M prend les valeurs 1 et 100 avec probabilités respectives $\frac{1}{4}$ et $\frac{3}{4}$
2. $p|M$ suit une $Beta(\theta, 1)$
3. $X|p, M$ suit une $Ber(p)$.

Exercice 26 (En machine) *On reprend le cadre de l'exercice 25, avec le formalisme donné ci-dessus. Donner alors*

1. *La probabilité* $\mathbb{P}(M = 1|X)$*, en fonction de la constante de normalisation de la loi Beta.*

2. *La loi conditionnelle* $\mathbb{P}(p|X, M = 1)$

Lorsque l'on construit un modèle hierarchique, on se donne donc une loi (hyperprior) pour les hyperparamètres. Le modèle doit donc contenir :

- $\pi(X|\theta, h) = \pi(X|\theta)$ la densité du modèle
- $\pi(\theta|h)$ un modèle pour les priors
- $\pi(h)$ un hyperprior.

Remarque :
Remarquons ici que plusieurs formulations sont équivalentes :

- En intégrant sur h, cela revient à se donner un prior pour θ donné comme un mélange (éventuellement avec une loi de mélange continue)
- Si l'on considère h comme un paramètre, cela revient à se donner un prior $\pi(\theta, h)$ sous forme de produit.
- Dans certains contexte, θ pourrait-être considérée comme une variable cachée (en particulier lorsque plusieurs θ sont déterminées par le même hyperprior). Dans ce cas, h est considéré comme un paramètre, et c'est la loi des données $\pi(X|h) = \int_\theta \pi(X|\theta)\pi(\theta|h)\mathrm{d}\theta$ qui est donnée comme une loi de mélange.
- Dans certains contextes, l'utilisation et la comparaison de plusieurs modèles pourra être reformulée dans ce cadre.

Remarquez aussi que cette construction peut-être étendue à des "couches" supplémentaires. En outre, l'utilisation de modèles hierarchiques peut s'avérer très utile lorsque l'on soupçonne a priori une homogénéité entre différents paramètres θ_i, en supposant par exemple qu'ils sont issus de la même loi. Enfin, même si ces modèles s'avèrent très utiles, on perd (la plupart du temps) l'avantage donné par les priors conjugués, vu que les calculs ne seront plus explicites.

La loi a posteriori est donc donnée par $\pi(\theta, h|X)$. Selon les contextes, on pourra marginaliser selon les hyperparamètres, en calculant

$$\pi(\theta|X) = \int_h \pi(\theta, h|X),$$

ou bien marginaliser sur θ en calculant $\pi(h|X)$ (en particulier dans le cas où la même loi d'hyperprior est donnée pour plusieurs θ).

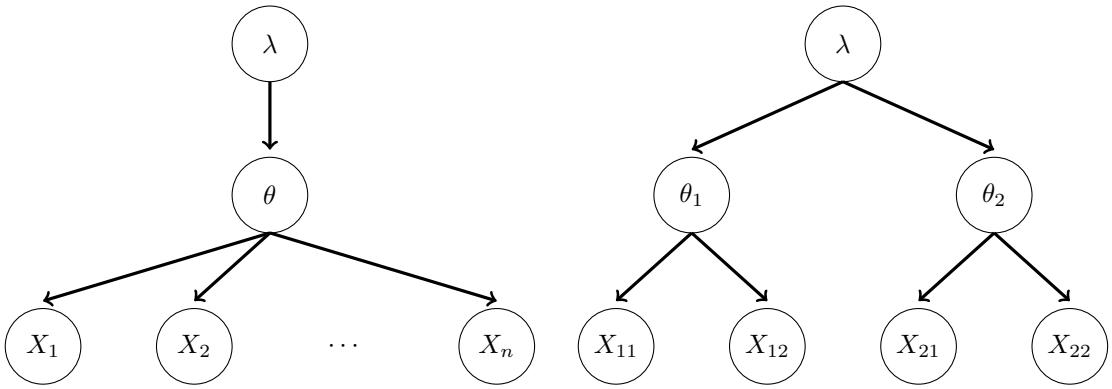
Exercice 27 *On considère le modèle hierarchique suivant :*

- $X|\theta \sim \mathcal{N}(\theta, 1)$
- $\theta|\lambda \sim \mathcal{N}(\lambda, 2)$
- $\lambda \sim \mathcal{N}(1, 2)$

1. *Rappeler la loi de* $\theta|X, \lambda$
2. *Donner la loi jointe de* X_i, θ, λ*, en déduire la loi de* $\theta, \lambda|X_i$ *ainsi que celle de* $\theta|X_i$ *et celle de* $\lambda|X_i$*.*
3. *On considère maintenant le modèle suivant :*
   - $X_{ij}|\theta_j \sim \mathcal{N}(\theta, 1)$
   - *Donner la loi marginale du prior* $\pi(\theta)$ *(en intégrant sur* λ*).*
   - $\theta_j|\lambda \sim \mathcal{N}(\lambda, 2)$
   - $\lambda \sim \mathcal{N}(1, 2)$

   Même question, donner les lois jointes ainsi que la loi a posteriori et ses marginales.

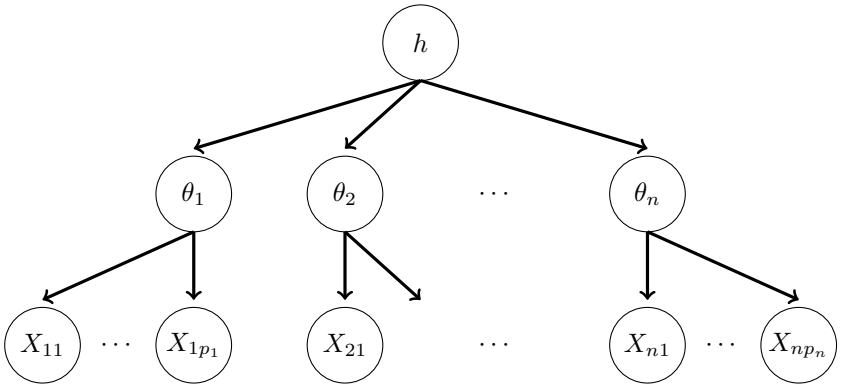
Enfin, pour résumer l'ordre des informations données par ces modèles hierarchiques (et donc les indépendances conditionnelles), il est souvent très pratique de représenter le **réseau bayésien** associé.



4.5 Bayes empirique

Attention : Cette méthode n'est pas vraiment bayésienne (dans l'esprit). On a insisté que le prior devait être choisi **avant d'avoir vu les données**. Cette méthode sera donc utilisée prudemment.

Utilisée la plupart du temps dans un modèle hierarchique où différents θ_i sont issus du même hyperprior $\pi(h)$ (voir dessin ci-dessous, ou second dessin de la page précédente TODO : ref), la méthode de Bayes empirique se propose d'utiliser les données pour choisir le prior $\pi(\theta|h)$, et donc fixer un hyper-paramètre. Il existe plusieurs variations de cette méthode, nous en exposons ici le principe général.



La méthode consiste en deux étapes :

1. On pose tout d'abord un modèle hierarchique, et on estime le posterior marginal pour h (en marginalisant sur θ_i). Lorsque cette loi est concentrée autour de son mode/espérance, on extrait de cette loi un estimateur ponctuel $h^\star$ (mode, médiane ou espérance). Une alternative est de prendre pour $h^\star$ le maximum de vraisemblance.

2. On se ramène maintenant à n modèles bayésiens où l'on utilise à chaque fois comme prior $\pi_{h^\star}(\theta) = \pi(\theta|h^\star)$, tant et si bien que l'on ne se trouve plus dans un modèle hierarchique.

En réalité, cela repose sur l'approximation suivante (pas toujours justifiée !) :

$$\pi(\theta|X) = \int_h \pi(\theta|h, X)\pi(h|X)\mathrm{d}h \approx \pi(\theta|h^\star, X)\pi(h^\star|X) \propto \pi(\theta|h^\star, X).$$

Remarque : Comme proposé au dessus, cela revient à prendre le maximum à posteriori pour h, puis d'utiliser la loi associée comme prior. En réalité, la méthode de Bayes empirique utilise le plus souvent le maximum de vraisemblance

$$h^\star = \arg\max \pi(X|h) = \arg\max \int_\theta \pi(X|\theta, h).$$

4.6 Sélection de modèles

Il est souvent utile de comparer des modèles (ex: regression). En Bayésien, c'est d'autant plus vrai que l'on peut introduire dans le modèle des biais supplémentaires *via* le prior.

La façon la plus naturelle de comparer des modèles serait de comparer les probabilité a posteriori de chaque modèle $\pi(m|X) \propto \pi(X|m)\pi(m)$.

Dans le cadre de comparaison de modèles, sauf information extérieure, il est tout aussi naturel de choisir un prior uniforme sur les modèles $\pi(m) \propto 1$.

On se ramène donc au calcul (à la maximisation) de $\pi(X|m)$. Mais pour être calculée, cette quantité nécessite d'intégrer sur tous les θ :

$$\pi(X|m) = \int_\theta \pi(X|\theta)\pi(\theta|m)\mathrm{d}\theta.$$

On parle souvent **d'évidence** du modèle, ou de vraisemblance intégrée. Remarquons que nous souhaitons ici évaluer une quantité qui *a priori* ne peut pas être calculée en utilisant la vraisemblance $\pi(X|\theta)$ ou le prior $\pi(\theta|m)$ à constante près. En effet, cela aurait pour effet de nous donner $\pi(X|m)$ à constante près (mais qui risque en fait de dépendre du modèle m !), et de ce fait, on perd tout intêret à comparer les modèles. [11]

Enfin, on peut utiliser le facteur de Bayes pour mesurer le ratio d'évidence entre les modèles.

$$BF = \frac{\pi(X|m_1)}{\pi(X|m_2)}$$

Exercice 28 (En machine) *1. On considère les données suivantes :*

```
x= c( 0.8422993, -0.9393801,  2.0532697, 1.3644805, 1.5066451,
          1.8287398, 3.0171889, 0.9414290, 2.2975294, 2.9310498)
```

On suppose que ces données sont issues d'un processus observé au cours du temps, et on veut savoir s'il y a une tendance linéaire. On se propose de répondre à cette question, en évaluant deux modèles concurrents.

(a) Le modèle m_1 spécifie la loi des données par $X_i|\theta \sim \mathcal{N}(\theta, 1)$, avec un prior $\theta|m_1 \sim \mathcal{N}(0, 1)$.

(b) Le modèle m_2 spécifie la loi des données par $X_i|\theta \sim \mathcal{N}(\theta + c * i, 1)$, avec un prior $(\theta, c)|m_2 \sim \mathcal{N}(0_2, I_2)$.

Enfin, on se donne un prior uniforme entre les deux modèles. Calculer l'évidence de chaque modèle. On pourra par exemple appliquer une méthode des rectangles pour approximer les intégrales.

2. *Créer une fonction* `SimulateTrend` *qui prend en argument la longueur d'un vecteur* `n`*, un intercept* `t`*, un coefficient* `a`*, et qui renvoie un vecteur de gaussiennes indépendantes de taille n, de variance 1, et telle que X_i ait pour espérance $t + a * i$.*

3. *En faisant varier le paramètre a, observer les comportements simultanés du facteur de Bayes et de la corrélation entre x et $(1, 2, \cdots, 10)$.*

4. *[plus difficile] Donner la loi jointe des observations et des paramètres dans chaque modèle, et en déduire la loi marginale de X dans chaque modèle. L'utiliser pour obtenir une expression explicite du facteur de bayes entre les modèles.*

5 Fréquentisme vs Bayésianisme

L'objectif de cette partie est de survoler rapidement les résultats théoriques qui concernent l'utilisation de statistiques bayésiennes.

Ces resultats sont obtenus dans un cadre fréquentiste :

On fixe θ_0 dans Θ, et on se demande la qualité des estimateurs en fixant $\theta = \theta_0$, et en intégrant sur les données tirées sous la loi $f_{X|\theta_0}$.

[11] Vous trouverez dans [5] p159 une petite astuce pour utiliser les contantes de normalisation pour calculer des évidences. Je ne pense pas avoir le temps d'en parler

5.1 Résultats théoriques

La première question qu'on peut se poser concerne la consitence de la loi a posteriori : Y a-t-il convergence presque sûre de la loi a posteriori vers θ_0 ?

On peut tout d'abord fixer rapidement les idées :

1. On peut remarquer que, par la loi des grands nombres (et sous l'hypothèse que l'espérance en question est finie) :
$$\sqrt[n]{\frac{f_{\theta|X}(\theta,X)}{f_{\theta|X}(\theta_0,X)}} \xrightarrow{X\sim f_{X|\theta_0}^{\otimes n},p.s.} \exp\left(-DK\Big(f_{X|\theta}(.,\theta_0)\|f_{X|\theta}(.,\theta)\Big)\right).$$
2. On rappelle que par Jensen, la divergence de Kullback est positive, et nulle si et seulement si $\theta=\theta_0$.
3. En rajoutant des hypothèses (en particulier que les fonctions $\log\big(f_{\theta|X}(\theta,X)\big), \theta\in\Theta\}$ forment une classe de Glivenco Cantelli pour $f_{X|\theta_0}$), on doit pouvoir prouver la convergence de la loi a posteriori vers un Dirac en θ_0.

On peut retenir de ce qui précède que $\pi(\theta|X)$ converge en loi (sur θ), $f_{X|\theta_0}$-p.s. (pour les observation X_i), vers un Dirac en θ_0. Une autre façon de dire ça, est que quelques soient des données i.i.d. tirées selon la loi $f_{X|\theta_0}$, alors le posterior va concentrer autour de la vraie valeur θ_0.

En réalité le résultat qui nous intéresse est plus fort (comme le TCL pour la LGN). Il s'agit de se demander l'ordre de grandeur des fluctuations, qui est donné par le théorème suivant.

Remarque : Le théorème suivant est écrit dans le cas unidimensionnel (pour simplifier les notations), mais peut-être étendu au cas multidimensionnel. Il est aussi écrit dans le cas faible (convergence étroite), mais possède une version plus forte (convergence en variation totale). De plus, il existe des extensions au delà du cas i.i.d.

Théorème 2 (Bernstein Von Mises) *On suppose que le modèle $\mathcal{M}$ est régulier, i.e. que les fonctions $\theta\mapsto f_{X|\theta}$ sont de classe $\mathcal{C}^2$. On note*

$$\begin{aligned}
l_\theta(x) &= \log(f_{X|\theta}(x,\theta)) \\
s_\theta(x) &= \frac{\partial}{\partial\theta} l_\theta(x) \\
h_\theta(x) &= \frac{\partial^2}{\partial\theta^2} l_\theta(x) \\
I(\theta) &= \int s_\theta(x)^2 f_{X|\theta}(x,\theta)\mathrm{d}x \\
&= -\int h_\theta(x) f_{X|\theta}(x,\theta)\mathrm{d}x \ \textit{(rappel)}
\end{aligned}$$

On suppose que la famille h_θ forme une classe de Glivenco-Cantelli pour f_{θ_0}[12]*, d'où*

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n h_\theta(X_i)+I(\theta)\right| \xrightarrow{X_i\sim_{i.i.d.} f_{\theta_0} p.s.} 0.$$

Soit π une densité a priori que l'on suppose $\mathcal{C}^1$ et strictement positive sur Θ. On suppose de plus que l'estimateur $\hat\theta_{MV}$ de θ par maximum de vraisemblance existe et est consistant. Alors la densité de la loi a postériori vérifie,

$$f_{\theta|X}(\hat\theta_{MV}+\frac{t}{\sqrt{n}}) \xrightarrow{\mathbb{P}} \exp\Big(-\frac{I(\theta)t^2}{2}\Big)$$

Autrement dit, la loi a postériori s'approche, pour n grand, d'une gaussienne centré en $\hat\theta_{MV}$, et de variance $\frac{1}{I(\theta)}$, indépendemment du prior. En particulier, cela implique la convergence étroite du postérior renormalisé vers une gaussienne, et surtout la convergence étroite du posterior vers un Dirac en θ_0. De plus, $\hat\theta_{MAP}\to\hat\theta_{MV}$

[12] i.e. une classe de fonctions mesurables qui vérifie la convergence uniforme de la mesure empirique vers la mesure théorique

TODO : preuve ?

Ce théorème sera à la base de tous les résultats concernant une analyse fréquentiste de l'inférence bayésienne. Le premier résultat concerne les estimateurs bayésiens.

Proposition 7 *On reste sous les hypothèses du théorème 2. On suppose de plus que le prior π admet un moment d'ordre 2. Alors, l'estimateur par l'espérance à posteriori $\hat{\theta}_{mean} = \mathbb{E}[\theta|X]$ vérifie*

$$\sqrt{n}\big(\hat{\theta}_{mean} - \hat{\theta}_{MV}\big) \xrightarrow{\mathbb{P}} 0$$

Par conséquent, il vérifie

$$\sqrt{n}\big(\hat{\theta}_{mean} - \theta_0\big) \xrightarrow{loi} \mathcal{N}(0, I(\theta)^{-1})$$

On peut aussi s'interesser aux liens entre intervalles de confiances fréquentistes et région HPD. La proposition suivante donne un élément de réponse

Proposition 8 *Soit $I(X)$ un intervalle de confiance fréquentiste au niveau α pour θ, et $R(X)$ une région de crédibilité de niveau α pour θ. Alors*

1. *Avec une analyse bayésienne : En intégrant sur (θ, X) tirés dans un modèle Bayésien, on a bien*

$$\mathbb{P}(\theta \in I(X)) = \mathbb{E}[\mathbb{P}(\theta \in I(X)|\theta)] = \alpha$$
$$\mathbb{P}(\theta \in R(X)) = \mathbb{E}[\mathbb{P}(\theta \in R(X)|R)] = \alpha$$

2. *Avec une analyse fréquentiste : On se place sous les hypothèses de Bernstein von Mises, et on note q_α le quantile d'ordre α d'une $\mathcal{N}(0,1)$. Alors*

   - *L'intervalle $I = [\hat{\theta}_{MAP} \pm q_{1-\frac{\alpha}{2}} \frac{I(\hat{\theta}_{MAP})^{-\frac{1}{2}}}{\sqrt{n}}]$ est un intervalle de crédibilité de niveau asymptotiquement α.*
   - *De plus, on a*
   $$\forall \theta_0 \in \Theta, \mathbb{P}_{\theta_0}\big(\theta_0 \in R_\alpha^{HPD}(X)\big) \to 1 - \alpha.$$

Exercice 29 *Vérifier sur des exemples la validité des résultats précédents en faisant les calculs explicites dans les exemples suivants :*

1. *Dans le modèle gaussien, à variance connu, avec un prior conjugué.*
2. *Dans le modèle exponentiel, avec un prior conjugué.*

5.2 Remarques et critiques

Les critiques les plus courantes portés par les fréquentistes aux bayésiens concernent plusieurs points :

- C'est subjectif (à cause du choix du prior)
- C'est lent (à cause des méthodes numériques que l'on verra plus tard)
- Cela n'apporte rien par rapport au point de vue fréquentiste (on obtient les mêmes résultats)
- C'est difficile (à vous de juger)

Les réponses naturelle à ces objections :

- Il est vrai que le choix du prior est subjectif, mais en réalité, le choix d'un modèle fréquentiste est aussi subjectif. En bayésien, cette subjectivité est mise en évidence. En plus, il est facile d'incorporer des connaissances a priori dans le prior.
- Effectivement, asymptotiquement, les méthodes donneront les mêmes résultats, mais lorsque peu de données sont disponibles, ou que les modèles sont très complexes, la loi a priori permet naturellement de concentrer la recherche sur les espace de paramètres qu'on considère les plus vraisemblables, sans exclure totalement les moins vraisemblables.

- En outre, l'interprétation des intervalles de crédibilité, et des tests est beaucoup plus simple qu'en fréquentiste.
- La puissance de calcul disponible est aujourd'hui suffisante pour mettre en place les algorithmes nécessaires en temps raisonnable.
- Une dernière remarque importante : Le modèle bayésien est souvent justifié par **l'échangeabilité** des données (invariance de la loi par permutation des observation), et le théorème de **De Finetti**, qui montre que des données échangeables sont indépendantes, conditionnellement à un paramètre caché.

6 Objectifs

Vous pouvez vous auto-évaluer sur les objectifs suivants (Barême :

- 0 : "Je ne sais pas faire du tout",
- 1 : "J'en ai entendu parler",
- 2 : "je pense savoir faire un peu",
- 3 : "je suis à l'aise dessus",
- 4 : "je pense maîtriser cette notion",
- 5 : "j'ai peur de rien sur ce sujet"

Question	Théorique	Programmation
Calculer un posterior, pour un prior discret		
Calculer un posterior, pour un prior continu, dans une famille conjuguée		
Construire une région HPD		
Résoudre un problème de théorie de la décision, avec coût non prédictif		
Résoudre un problème de théorie de la décision, avec coût prédictif		
Calculer médiane, maximum, espérance à posteriori		
Effectuer un test		
Construire un classifieur naif Bayésien plug-in		
Construire un classifieur naif Bayésien, bayésien;)		
Calculer un prior non informatif		
Utiliser des modèles hierarchiques		
Mettre en place ABC		
Mettre en place Metropolis-Hasting		
Mettre en place Gibbs		
Mettre en place des méthodes variationnelles		
Rassembler les connaissances pour un problème de régression		
Rassembler les connaissances pour un problème de mélange		
Rassembler les connaissances pour un problème de classification supervisée		
Rassembler les connaissances pour construire un modèle bayésien, et un cadre de décision pour répondre à une question donnée		

7 Bibliographie

En complément de la bibliographie ci-dessous, vous pouvez trouver en ligne l'excellent polycopié de Judith Rousseau, ainsi que des notes (et le livre) de Christian Robert.

[1] James O. Berger, Jose M. Bernardo, and Dongchu Sun. Overall objective priors. *Bayesian Anal.*, 10(1):189–221, 03 2015.

[2] R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2011.

[3] Allen B. Downey. *Think Bayes.* O'Reilly Media, Sebastopol, California, 2013.

[4] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[5] Kevin P Murphy. *Machine learning: a probabilistic perspective.* Cambridge, MA, 2012.

8 Remarques sur la pratique

1. Attention à l'utilisation de priors impropres (en particulier pour la sélection de modèles)
2. Penser à faire tous les calculs à l'échelle log
3. On pourra pour se simplifier la vie utiliser BUGS ou JAGS http://www.openbugs.net/w/FrontPage,

9 Formulaire

9.1 Lois utiles

Loi discrete	Probabilité	Espérance
Bernoulli $\mathcal{B}(p)$	$\forall x \in \{0,1\}, \mathbb{P}(X=x) = p^x(1-p)^{1-x}$	p
Binomiale $\mathcal{B}in(n,p)$	$\forall k \in \{0,n\}, \mathbb{P}(X=x) = \binom{n}{k} p^k(1-p)^{N-k}$	np
Poisson $\mathcal{P}(\lambda)$	$\forall k \in \mathbb{N}, \mathbb{P}(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}$	λ
Géométrique $\mathcal{G}(p)$	$\forall k \in \mathbb{N}^*, \mathbb{P}(X=k) = p(1-p)^{k-1}$	$\frac{1}{p}$

Loi continue	Densité	Espérance
Exponentielle $\mathcal{E}(\lambda)$	$e^{-\lambda x}\mathbb{1}_{x\geq 0}$	$\frac{1}{\lambda}$
Normale $\mathcal{N}(m,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$	m
Gamma $\Gamma(k,\lambda)$	$\frac{\lambda^k x^{k-1}}{\Gamma(k)}e^{-\lambda x}\mathbb{1}_{x\geq 0}$	$\frac{k}{\lambda}$
Beta $Beta(a,b)$	$\frac{x^{a-1}(1-x)^{b-1}}{\beta(a,b)}\mathbb{1}_{x\in[0,1]}$	$\frac{a}{a+b}$

,

avec

- $\Gamma(k) = \int_{\mathbb{R}^+} t^{k-1}e^{-t}\mathrm{d}t$, vérifiant $\Gamma(k+1) = k\Gamma(k)$
- $Beta(a,b) = \int_{[0,1]} t^{a-1}(1-t)^{b-1}\mathrm{d}t \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

9.2 Lois conjuguées

Modèle $\pi(X\|\theta)$	Priors conjugués	Mise à jour
$\left\{\text{Bin}(N,p)\ ;\ p\in]0,1[\right\}$	$\left\{\text{Beta}(a,b)\ ;\ a,b>0\right\}$	$a \to a + \sum_{i=1}^n x_i$ $b \to b + Nn - \sum_{i=1}^n x_i$
$\left\{\mathcal{E}(\lambda)\ ;\ \lambda>0\right\}$	$\left\{\text{Gamma}(k,\beta)\ ;\ k,\beta>0\right\}$	$k \to k+n$ $\beta \to \beta + \sum_{i=1}^n x_i$
$\left\{\mathcal{P}(\lambda)\ ;\ \lambda>0\right\}$	$\left\{\text{Gamma}(k,\beta)\ ;\ k,\beta>0\right\}$	$k \to k + \sum_{i=1}^n x_i$ $\beta \to \beta + n$
$\left\{\mathcal{G}(p)\ ;\ p\in]0,1[\right\}$	$\left\{\text{Beta}(a,b)\ ;\ a,b>0\right\}$	$a \to a+n$ $b \to b + \sum_{i=1}^n x_i - n$
$\left\{\mathcal{N}(m,\frac{1}{t})\ ;\ m\in\mathbb{R}\right\}$	$\left\{\mathcal{N}(\mu,\frac{1}{\tau})\ ;\ \mu\in\mathbb{R},\tau>0\right\}$	$\mu \to \frac{\tau\mu + t\sum_{i=1}^n x_i}{\tau+tn}$ $\tau \to \tau + tn$
$\left\{\mathcal{N}(m,\frac{1}{t})\ ;\ t>0\right\}$	$\left\{\text{Gamma}(k,\beta)\ ;\ k,\beta>0\right\}$	$k \to k + \frac{n}{2}$ $\beta \to \beta + \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2$
$\left\{\mathcal{N}(m,\Lambda^{-1})\ ;\ m\in\mathbb{R}^d\right\}$	$\left\{\mathcal{N}(\mu,\Omega^{-1})\ ;\ \mu\in\mathbb{R}^d,\Omega \text{ s.d.p.}\right\}$	$\mu \to \left(\Omega+n\Lambda\right)^{-1}\left(\Omega\mu + \Lambda\sum_{i=1}^n x_i\right)$ $\Omega \to \Omega + n\Lambda$
$\left\{\mathcal{N}(m,\Lambda^{-1})\ ;\ \Lambda \text{ s.d.p}\right\}$	$\left\{\text{Wishart}(\nu,V)\ ;\ \nu>0, V \text{ s.d.p.}\right\}$	$\nu \to \nu + n$ $V \to \left(V^{-1} + \sum_{i=1}^n (x_i-\mu)(x_i-\mu)^T\right)^{-1}$