MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU

FACULTE DES SCIENCES

DEPARTEMENT DE MATHEMATIQUES

MEMOIRE DE MASTER

SPECIALITE:PROBABILITES ET STATISTIQUES

OPTION:PROCESSUS ALEATOIRES ET STATISTIQUE DE LA DECISION

Présenté par:

M^{elle} AISSOUN Roza

Sujet:

Les modèles hiérarchiques Bayésiens

Devant le jury d'examen composé de:

Hamadouche Djamel ;	Professeur;	U.M.M.T.O;	Président
Fellag Hocine ;	Professeur;	U.M.M.T.O;	Rapporteur
Atil Lynda;	Maître de conférences B;	U.M.M.T.O;	Examinatrice

Soutenu le: 3 /10/2013

Remerciements

Je remercie le bon Dieu tout puissant d'avoir guidé mes pas vers les portes du savoir tout en illuminant mon chemin, et de m'avoir donné suffisamment de courage et de persévérance pour pouvoir mener mon travail à terme.

Je tiens à exprimer mon profonde gratitude et mes plus sincères remerciements à M^r Fellag Hocine pour avoir accepté de mon encadrer, mais aussi pour sa disponibilité, sa gentillesse, et ses précieux conseils dés le début de mon travail.

Je remercie vivement M^r Hamadouche Djamel ,Professeur à l'UMMTO, pour l'honneur qu'il me fait en acceptant d'être président du jury.
Toute mes reconnaissance aux membres de jury pour avoir accepté de juger mon travail, je suis honoré de la présence de M^{me} Atil Lynda . Qu'elle trouve ici l'expression de mon profond respect.

Je remerciements tous personnes du département mathématique, particulièrement tous les enseignants qui nous ont donné des connaissances tout au long de mon cursus, et tous ceux et toutes celles qui ont de près ou de loin contribué à la réalisation de ce modeste travail.

En fin un grand merci à ma chère famille, particulièrement mes parents pour leur protection filiale et leur encouragement, et à tous mes amis(es) et camarades.

Dédicaces

C'est avec une pensée pleine de reconnaissance inspirée par la générosité et la gentillesse que je dédie ce modeste travail :

A mes chères parents pour leur grand et généreux amour, leur sacrifice, leur compréhension et leur soutien.

A mon sage et petit frère Lyes que j'aime beaucoup.

A ma très chère jumelle Nadia, et ma sœur Karima et tout membre de ma famille.

A toutes mes amies qui m'ont soutenue moralement : Malika, Saliha, Zina, Jiji, Amel, Aldja, Yasmina, Chafia, Fatima, Rahma, Lynda, Samira, Siham, Kahina, Kahina, Lynda.

A tous ceux qui m'ont aidé, conseillé, et à tous ceux que j'aime et que je porte dans mon cœur.

Table des matières

Introduction

La statistique, est une discipline scientifique en plein essor.Elle intervient dans toutes les disciplines scientifiques ou se mêlent savoir et données. Elle est aussi utilisée par les physiciens, les économistes, les ingénieurs, les géographes, les biologistes, les assureurs, les psychologues, les gestionnaires d'entreprises, etc. Bref, par tous les praticiens soucieux de batir sur des fondations solides un pont entre théorie et données expérimentales. Pour définir son objet, la statistique peut être définie comme «l' art de raisonner de façon quantitative en avenir incertain »"Christian Robert (Robert, 2006)".
L'objet principal de la statistique est de mener, grâce à l' observation d 'un phénomène alétoire, une inférence sur la distribution probabiliste à l'origine de ce phénomène, c'est-a-dire de fournir une analyse (ou une description) d 'un phénomène passé, au une prédiction d'un phénomène à venir de nature similaire (aspects supplémentaires de la Statistique appliquée tels que la collecte de données par exemple :sondages, plans d'expérience,...etc).
Face à la complexité du phénomène observé, deux approches statistiques intervient bayésiens et classiques, le mode de raisonnement du statisticien classique est toujours le même, quel que soit le paramètre inconnu θ à estimer. Les données disponibles permettent de calculer un intervalle de confiance correspondant à un risque fixé α. Le paramètre inconnu θ est ou n'est pas dans cet intervalle. Aussi, pour décrire son incertitude sur θ, le statisticien classique imagine une collection d'échantillons recueillis dans les mêmes conditions et, pour chacun d'entre eux, il «calcule » un intervalle de confiance et conclut en disant que $1-\alpha$ pour cent d'entre eux contiendraient θ. C'est la vision fréquentiste : tout est dans les données. Mais comment accepter que plusieurs techniques d'estimation (méthodes des moments, des moments pondérés, du maximum de vraisemblance, etc.) puissent produire des

intervalles de confiance différents? que faire avec tous les problèmes réel ou ces répétitions imaginaires n'ont pas de sens? ces questions peuvent avoir des réponces avec le statisticien bayésien qui raisonne différemment en considérant que le paramètre du modèle statistique (x/θ), est incertain. Donc il cherche a quantifier son incertitude en mobilisant toutes les informations disponibles. Alors tout savoir actuel sur ce paramétre on lui a attribuer une distribution de probabilite a priori(l'état de connaissance d'un expert, et donc son incertitude), souvent notée $\pi(\theta)$. Cette loi a priori doit être claire et indépendante de l'échantillon, sinon la même source d'information interviendrait deux fois.

L'a priori est le point le plus critiqué de l'analyse bayésienne. Car, une fois que cette loi a priori est connue, l'inférence peut être conduite d'une façon mécanique en minimisant le coût a posteriori, en calculant les régions de plus forte densité a posteriori ou en intégrant les paramètres pour obtenir la distribution prédictive. Dans une certaine mesure, c'est aussi la plus difficile. Il est donc nécessaire le plus souvent de faire un choix (partiellement) arbitraire de loi a priori, ce qui peut avoir un impact considérable sur l'inférence en utilisant les lois conjuguées qui ne sont pas toujours justifiées, car la détermination subjective de la loi a priori qui en résulte, se fait au prix d'un traitement analytique difficile.

Aprés des années de critiques, le travail de Jeffreys (1946) sur les a priori non informatives apparut comme un don du ciel pour la communauté bayésienne, car il propose une méthode de construction de la loi a priori directement déduite de la distribution des observations, aussi certains bayésiens sont cependant en désaccord avec l'utilisation de méthodes automatisées. Récemment, une approche qui essaie de rectifier ces problèmes est:l'analyse bayésienne hiérarchique, qui met des mesures a priori sur les paramètres de cette loi $\pi(\theta)$. Cette approche est une conséquence des avances théoriques en robustesse et analyse de sensibilité, elle fournit une base solide à l'analyse bayésienne dans le cas où l'information a priori est incomplète.

Enfin l'inference statistique, dont la mise en oeuvre implique un savoir-faire technique, permet l'aide a la décision sous incertitude. En effet, dès qu'on dispose de la meilleure connaissance possible des quantites incertaines, il faut pour cela mobiliser toute information disponible, ce qui justifie le choix bayesien, l'inference statistique

peut donner la distribution de probabilité de toute grandeur intéressante pour le décideur.

Ce travail est réparti en trois chapitres, le premier constitue une généralité sur l'analyse bayésienne utile pour notre thématique. Le second chapitre est consacré aux modèles hiérarchiques bayésiens, où nous avons présenté les notions essentielles de cette théorie. Le troisiéme chapitre, présente une application publiée dans le journal de psychologie mathématique. Cette application explique l'utilité de la loi a priori dans l'inférence. Enfin, nous terminons par une conclusion générale et des perspectives.

Chapitre 1

L'analyse statistique bayésienne

1.1 Introduction

Une des méthodologies très importante pour faire de l'inférence statistique paramétrique, est l'analyse bayésienne. Cette dernière se raméne fondamentalement à une inversion (Robert, 2006). En effet, elle vise à déterminer les causes a partir des effets. Les causes sont réduites aux paramètres du mécanisme probabiliste générateur des données imagine par l'homme d'étude et que les effets sont résumés par les observations disponibles. En d'autres termes, le modèlisateur voit les observations comme des tirages dans une loi statistique contrôlée par le paramètre inconnu θ. Une méthode statistique permet de déduire de ces observations une inférence sur θ. A l'issue de cette inférence, l'incertitude θ est quantifiée et la prévision des observations futures consiste alors à utiliser le mécanisme générateur de données conditionnellement à θ.

1.2 L'inférence bayésienne

Définition 1.1. Modèle classique
On se place dans un espace probabilisé paramétrique classique

$$x \in (\mathcal{X},\beta,\{P_\theta,\theta \in \Theta\})$$

$\mathcal{X}$ désigne l'espace des données, Θ celui des paramètres θ. Le but de l'analyse statistique est de faire de l'inférence sur θ, c'est-à-dire décrire un phénomène passé ou à venir dans un cadre probabiliste. L'idée centrale de l'analyse bayésienne est de considérer le paramètre inconnu θ comme aléatoire : l'espace des paramètres Θ est muni d'une probabilité π tel que $(\Theta;\mathcal{A};\pi)$ est un espace probabilisé. Nous noterons $\theta \sim \pi$ est appelée loi a priori.Elle détermine ce qu'on sait et ce qu'on ne sait pas avant d'observer x.(la loi d'un expert ...).

Définition 1.2. Modèle dominé
Le modèle est dit dominé s'il existe une mesure commune dominante μ, c'est-à- dire pour tout θ, P_θ admet une densité par rapport à μ [1] :

$$f(x/\theta) = \frac{dP_\theta}{d\mu}$$

Cette fonction $f(x/\theta)$, vue comme une fonction de θ une fois qu'on a observé un tirage de x, est appelée vraisemblance du modèle. C'est la loi de x conditionnellement à θ.

Définition 1.3. La distribution jointe

La distribution jointe de (x,θ) s'obtient par

$$f(x,\theta) = f(x/\theta)\pi(\theta) \tag{1.1}$$

La formule de Bayes est basée sur la décomposition inverse de(1.1)

$$f(x,\theta) = \pi(\theta/x)m(x)$$

On obtient donc la densité a posteriori de θ conditionnelle à x

$$\pi(\theta/x) = \frac{f(x/\theta)\pi(\theta)}{m(x)} \tag{1.2}$$

1. Pour des mesures σ-finies et en vertu du théorème
de Radon-Nikodym, ceci est équivalent à être absolument continue par rapport à μ

avec $m(x)$ ne dépend pas de θ, et est la densité prédictive de x, c'est la constante d'intégration de (1.2)

$$m(x) = \int_{\Theta} f(x/\theta)\pi(\theta)d\theta \tag{1.3}$$

la formule de Bayes dans (1.2)est approximative à

$$\pi(\theta/x) \propto f(x/\theta)\pi(\theta)$$

Ce qui est souvent utile pour un statisticien est d'étudier le comportement d'une valeur future de x est notée y, réplication d'une future obsrvation, étant donnée l'information déja récolté à l'observation de x. Sous l'hypothèse que conditionnellement à θ, y est indépendante de x, on obtient la densité prédictive de y

$$f(y/x) = \int_{\Theta} f(y/\theta)\pi(\theta/x)d\theta$$

Par ailleurs, le paramètre d'intérêt est souvent multidimensionnel

$$\theta = (\theta_1,...,\theta_n) \in \mathbb{R}^n$$

Dans ce ças, à partir des densités jointes(a priori ou a posteriori)de θ, on peut obtenir toutes les densités marginales (a priori ou a posteriori)de chaque composante θ_i de θ par intégration des autres composantes. Par exemple, a posteriori

$$\pi(\theta_i/x) = \int \pi(\theta/x)d\theta_1,...,d\theta_{i-1},d\theta_{i+1},...,d\theta_n$$

D'un point de vue pratique, le choix de la loi a priori est souvent perçu comme une difficulté majeure de l'approche Bayésienne en ce que l'interprétation de l'information a priori disponible.

Exemple 1. *Dans le cas gaussien, à variance connue:* $x \sim \mathcal{N}(\mu,\sigma^2)$ *et* $\theta = \mu(\sigma^2 connu)$

$$\begin{aligned}
\pi(\mu) &= \frac{e^{-\frac{(\mu-\mu_0)^2}{2\tau^2}}}{\tau\sqrt{2\pi}} \\
\pi(\mu/x) &\propto e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\mu_0)^2}{2\tau^2}} \\
\pi(\mu/x) &\propto exp(-\frac{1}{2}(\frac{1}{\sigma^2}+\frac{1}{\tau^2})(\mu - (\frac{x}{\sigma^2}+\frac{\mu_0}{\tau^2}))(\frac{1}{\sigma^2}+\frac{1}{\tau^2})^{-1})
\end{aligned}$$

Ainsi

$$\pi(\mu/x) \sim \mathcal{N}(x\frac{\tau^2}{\sigma^2+\tau^2} + \mu_0\frac{\sigma^2}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$$

On remarque sur cet exemple que la loi a posteriori est plus resserrée (pointée) que la loi a priori. Cela s'avère être intuitif: la loi a posteriori est la loi de θ en ayant une information supplémentaire à savoir la donnée de x, l'incertitude sur θ ne peut donc que diminuer, en d'autres termes la variance diminue. En considérant $\bar{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ dans le cas indépendant et identiquement distribué, la loi a posteriori se centre sur $\bar{x_n}$ avec un nombre d'observations qui augmente. Dans ce cas, elle se rapproche du maximum de vraisemblance.

1.3 L'analyse bayésienne empirique

L'estimation bayésienne empirique peut être vue comme un complément du modèle hiérarchique, l'estimation bayésienne empirique consiste à estimer la densité a posteriori quand les hyperparamètres sont inconnus.

1.3.1 Le principe bayésien empirique paramétrique

Un intérêt pratique des techniques bayésiennes empiriques est de déterminer des approximations dans des contextes non informatifs(lorsque la loi a priori n'est pas disponible).
Pour les familles exponentielles, le plus simple est de considérer l'a priori conjugué associé à $f(x/\theta)$,$\pi(\theta/\lambda)$. L'approche bayésienne empirique paramétrique consiste à estimer les paramètres λ de la densité marginale $m(x/\lambda)$ en se basant sur les observations (par exemple, par la méthode de vraisemblance maximale). Soit $f(x/\theta)$ la distribuées de vraisemblance, θ un vecteur de paramètres inconnus de fonction de densité $\pi(\theta/\lambda)$ et λ un vecteur de paramètres. Si λ est connu, alors la densité a posteriori de θ

$$\pi(\theta/\lambda,x) = \frac{f(x/\theta)\pi(\theta/\lambda)}{m(x/\lambda)}$$

est où $m(x/\lambda)$ est la distribution marginale

$$m(x/\lambda) = \int f(x/\theta)\pi(\theta/\lambda)d\theta$$

On estime θ avec la méthode de Bayes ($\mathbb{E}^{\theta/x,\lambda}(\theta)$) ou avec la méthode du maximum de vraisemblance. Si λ est inconnu, dans l'approche bayésienne on considère la

distribution hyperpriori de λ, $\pi(\lambda)$, et nous définissons la distribution a posteriori de la manière suivante :

$$\begin{aligned}\pi(\theta/x) &= \frac{\int f(x/\theta)\pi(\theta/\lambda)\pi(\lambda)d\lambda}{\int\int f(x/\theta)\pi(\theta/\lambda)\pi(\lambda)d\theta d\lambda} \\ &= \int \pi(\theta/x,\lambda)\pi(\lambda/x)d\lambda\end{aligned}$$

Comme exemple, le modèle exponentielle $\mathbf{e}xp(\lambda)$. Nous avons alors :

$$\begin{aligned}m(x_i/\lambda) &= \int_0^\infty e^{-\theta}\frac{\theta^{x_i}}{x!}\lambda exp(-\theta\lambda)d\theta \\ &= \frac{\lambda}{(1+\lambda)^{x_{i+1}}} \\ &= (\frac{1}{1+\lambda})^{x_i}(\frac{\lambda}{1+\lambda}) \\ & x_i/\lambda \sim geo(\tfrac{\lambda}{1+\lambda})\end{aligned}$$

L'estimateur du maximum de vraisemblance pour les n premières observations (x_1,x_2,...,x_n) est donné par:

$$\begin{aligned}\log(\Pi_{i=0}^n m(x_i/\lambda)) &= \log((\frac{1}{1+\lambda})^{\sum x_i}(\frac{\lambda}{1+\lambda})^n) \\ &= -\sum x_i \log(1+\lambda) + n(\log(\frac{\lambda}{1+\lambda}))\end{aligned} \tag{1.5}$$

la moyenne $\bar{x}$ étant établie sur les n premières observations.
On a que:

$$\frac{\partial}{\partial\lambda}\log(\Pi_{i=0}^n m(x_i/\lambda)) = \frac{-\sum x_i}{1+\lambda} + \frac{n}{\lambda} - \frac{n}{1+\lambda} = 0;$$

ce qui est équivalent à

$$\hat{\lambda}(x) = \frac{1}{\bar{x}}$$

Sachant que $\lambda = \hat{\lambda}(x) = \frac{1}{\bar{x}}$ et $\bar{x}$, la distribution a posteriori est estimée par:

$$\begin{aligned}\pi(\theta/\hat{\lambda},x_{n+1}) &\propto (e^{-\theta}\frac{\theta^{x_{n+1}}}{x_{n+1}!})\hat{\lambda}exp(-\theta\hat{\lambda}) \\ &\propto \theta^{x_{n+1}}exp(-\theta(\hat{\lambda}+1)) \\ &= \Gamma(x_{n+1}+1,\hat{\lambda}+1\end{aligned}$$

et l'estimateur de Bayes empirique de θ_{n+1} est L'estimateur bayésien empirique paramétrique de θ est donné par :

$$\begin{aligned}\delta^{EB}(x_{n+1}) &= \frac{x_{n+1}+1}{\hat{\lambda}+1} \\ &= \frac{\bar{x}}{\bar{x}+1}(x_{n+1}+1)\end{aligned}$$

1.3.2 Le principe bayésien empirique non paramétrique.

Soient $(n+1)$ observations indépendantes x_1,x_2,...,x_{n+1} de densités $f(x_i/\theta_i)$, le problème porte sur l'inférence sur θ_{n+1}, avec l'hypothèse supplémentaire que les θ_i ont tous été tirés selon le même a priori inconnu π. Une façon de résoudre la difficulté de l'incertitude dans la mesure a priori est d'utiliser une approche bayésienne empirique de Robbins(Le cadre initial de Robbins est principalement non paramétrique et fait usage des observations x_1,x_2,...,x_{n+1} pour estimer f_π) qui est essentiellement non paramétrique et que l'on appelle estimation bayésienne empirique non paramétrique. D'un point de vue bayésien, cela revient à dire que la loi d'échantillonnage est connue, mais que la loi a priori ne l'est pas.
La loi marginale

$$f_\pi(x) = \int f(x/\theta)\pi(\theta)d\theta$$

peut alors être utilisée pour retrouver la distribution π à partir des observations, puisque x_1,...,x_n peut être vu comme un échantillon i.i.d. de loi f_π. On obtient ainsi une approximation $\hat{\pi}_n$ qu'on peut substituer à la vraie loi a priori pour obtenir l'expression suivante de la loi a posteriori

$$\tilde{\pi}(\theta_{n+1}/x_{n+1}) \propto f(x_{n+1}/\theta_{n+1})\hat{\pi}_n(\theta_{n+1})$$

cette technique n'est pas bayésienne, bien qu'elle repose sur la formule de Bayes.

Exemple 2. *On considére les x_i distribués selon une loi $\mathcal{P}(\theta_i)$ $(i = 1,...,n)$. Si $P_k(x_1,...,x_n)$ est le nombre d'observations égales à k,$k \in \mathcal{N}$, $P_k(x_1,...,x_n)$ donne une estimation de la loi marginale*

$$f_\pi(k) = \int_0^\infty e^{-\theta}\frac{\theta^k}{k}\pi(\theta)d\theta$$

Si $x_{n+1} \sim \mathcal{P}(\theta_{n+1})$ et si θ_{n+1} est estimé sous côut quadratique, l'estimateur de Bayes[2] *est*

$$\delta^{\pi}(x_{n+1}) = \mathbb{E}^{\pi}[\theta/x_{n+1}] = \frac{\int_0^{\infty} e^{-\theta}\theta^{x_{n+1}+1}\pi(\theta)d\theta}{\int_0^{\infty} e^{-\theta}\theta^{x_{n+1}}\pi(\theta)d\theta} = \frac{f_{\pi}(x_{n+1}+1)}{f_{\pi}(x_{n+1})}(x_{n+1}+1)$$

Donc l'approximation bayésienne empirique de δ^{π} est (proposition 1.1)

$$\delta^{EB}(x_{n+1}) = \frac{p_{(x_{n+1}+1)}(x_1,...,x_n)}{p_{(x_{n+1})}(x_1,...,x_n)}(x_{n+1}+1)$$

où on a remplacé f_{π} par son approximation.

1.4 Introduction à la théorie de la décision Bayésienne

Un problème de décision en générale est fondé sur les trois éléments suivants:

-Un ensemble des actions(décision)$\mathcal{D}$

-Un espace des paramètres Θ

-Une fonction de coût (de perte) $L(\theta,\delta)$ qui décrit la perte de prendre la décision δ lorsque le paramètre est θ.

1.4.1 Fonction de perte et risque

Définition 1.4. Soit $\delta \in \mathcal{D}$ une règle de décision.
Une fonction de perte (coût) est une fonction mesurable de $(\Theta \times \mathcal{D})$ à valeurs dans $\mathbb{R}_+$ notée $L(\theta,\delta)$ et définie telle que

1. $\forall(\theta,\delta)\ L(\theta,\delta) > 0$
2. $\forall\theta,\exists\delta^*$ tels que: $L(\delta^*(x),\theta) = 0$

S'il faut faire un choix entre deux règles de décision, ce choix est impossible sans critère de coût, de sorte à définir correctement la notion de meilleur estimateur.

2. L'estimateur de Bayes δ^{π} associé à la loi a priori π et au coût quadratique est la moyenne a posteriori

$$\delta^{\pi}(x) = \mathbb{E}^{\pi}[\theta/x] = \frac{\int_{\theta} \theta f(x/\theta)\pi(\theta)d\theta}{\int_{\theta} f(x/\theta)\pi(\theta)d\theta}$$

Exemple d'une fonction de coût usuel
Perte quadratique

$$L(\theta,\delta) = (\theta - \delta)^2$$

Le coût quadratique pénalise fortement les grandes erreurs. Les estimateurs de Bayes associé au coût quadratique sont les moyennes a posteriori. Les fonctions de coût conduisant à la moyenne a posteriori comme estimateur de Bayes sont appelées fonctions de coût propres.

Proposition 1.1. *(Christian P. Robert 2006)*
L'estimateur de Bayes δ^π associé à la loi a priori π et au coût quadratique est la moyenne a posteriori

$$\delta^\pi(x) = \mathbb{E}^\pi[\theta/x] = \frac{\int_\theta \theta f(x/\theta)\pi(\theta)d\theta}{\int_\theta f(x/\theta)\pi(\theta)d\theta}$$

Corollaire 1.1. *(Christian P. Robert 2006)*

Quand $\Theta \in \mathbb{R}^p$, l'estimateur de Bayes δ^π associé à la loi a priori π et au coût quadratique,

$$L(\theta,\delta) = (\theta - \delta)^t Q(\theta - \delta)$$

est la moyenne a posteriori, $\delta^\pi(x) = \mathbb{E}^\pi[\theta/x]$, pour tout matrice $Q(p\times p)$ symétrique définie positive.
Le coût quadratique est particulièrement intéressant lorsque l'espace de paramètres est borné et le choix d'un coût plus subjectif est impossible.

Exemple 3. *Perte quadratique*

$$\mathcal{D} = \Theta \subset \mathbb{R}^d$$

$$L(\theta,\delta) = \| \theta - \delta \|^2$$

Comme la norme au carré est une fonction convexe deux fois dérivable sur Θ, pour trouver

δ^π estimateur bayésien, il suffit de déterminer les points critiques du risque a posteriori

$$\begin{aligned} \rho(\pi,\delta/x) &= \mathbb{E}^\pi(\| \theta - \delta \|^2 /x) \\ \frac{\partial\rho(\pi,\delta/x)}{\partial\delta} &= -2\int_\Theta (\theta - \delta(x))d\pi(\theta,x) \\ &Donc \\ \frac{\partial\rho(\pi,\delta/x)}{\partial\delta} &= 0 \Longleftrightarrow \\ \delta(x) &= \mathbb{E}^\pi(\theta/x) \end{aligned}$$

D'après l'inégalité de Jensen, $\delta \longmapsto \rho(\pi,\delta/x)$ est aussi convexe.
L'estimateur bayésien vaut donc $\delta^\pi(x) = \mathbb{E}^\pi(\theta/x)$ $\mu - pp$.
Dans le cas gaussien de l'exemple 1 $x \sim \mathcal{N}(\mu,\sigma^2)$ et $\mu \sim \mathcal{N}(\mu_0,\tau)$, l'estimateur bayésien s'exprime :

$$\delta^\pi(x) = \frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\tau^2+\sigma^2}\mu_0$$

Exemple 4. *Perte absolue (perte L^1) $\mathcal{D} = \Theta \subset R$ et $L(\theta,\delta) = \sum_{i=1}^d \mid \theta_i - \delta_i \mid$ Dans le cas simple où $d = 1$:*

$$\begin{aligned} \rho(\pi,\delta/x) &= \int_\Theta \mid \theta - \delta \mid d\pi(\theta/x) \\ &= \int_{-\infty}^{\delta} (\delta - \theta)\pi(\theta/x)d\theta + \int_\delta^\infty (\delta - \theta)\pi(\theta/x)d\theta \end{aligned}$$

Comme $\delta \longmapsto \rho(\pi,\delta/x)$ est convexe et dérivable μ-presque partout, il suffit là encore de déterminer les points critiques :

$$\frac{\partial\rho(\pi,\delta/x)}{\partial\delta} = \int_{-\infty}^{\delta} \pi(\theta/x)d\theta - \int_\delta^\infty \pi(\theta/x)d\theta$$

Donc

$$\frac{\partial\rho(\pi,\delta/x)}{\partial\delta} = 0 \Longleftrightarrow P^\pi(\theta \leq \delta/x) = P^\pi(\theta \geq \delta/x)$$

C'est-à-dire $\delta^\pi(x)$ est la médiane de $\pi(\theta/x)$

Définition 1.5. Risque fréquentiste
Pour $(\theta,\delta) \in \Theta \times \mathcal{D}$, le risque fréquentiste est défini par :

$$\begin{aligned} R(\theta,\delta) &= \mathbb{E}[L(\theta,\delta(x))] \\ &= \int_x L(\theta,\delta(x))f(x/\theta)d\mu(x) \end{aligned}$$

C'est une fonction de θ et ne définit donc par un ordre total sur $\mathcal{D}$ et ne permet donc pas de comparer toutes décisions et estimateurs. Il n'existe donc pas de meilleur estimateur dans un sens absolu. Ainsi, l'approche fréquentiste restreint l'espace d'estimation en préférant la classe des estimateurs sans biais dans laquelle il existe des estimateurs de risque uniformément minimal, l'école bayésienne ne perd pas en généralité en définissant un risque a posteriori. L'idée est d'intégrer sur l'espace des paramètres pour pallier cette difficulté.

Définition 1.6. Risque a posteriori

Une fois données la loi a priori sur le paramètre et la fonction de perte, le risque a posteriori est défini par :

$$\begin{aligned}\rho(\pi,\delta/x) &= \mathbb{E}(L(\theta,\delta(x))/x) \\ &= \int_{\Theta} L(\theta,\delta(x))d\pi(\theta/x)\end{aligned}$$

Ainsi, le problème change selon les données, ceci est dû à la non existence d'un ordre total sur les estimateurs.

Définition 1.7. Risque Bayésien(intégré) Pour une fonction de perte donnée, le risque Bayésien est défini par :

$$r(\pi,\delta) = \int_{\Theta} R(\theta,\delta)d\pi(\theta)$$

Définition 1.8. Estimateur bayésien

Un estimateur bayésien est un estimateur vérifiant :

$$r(\pi,\delta^{\pi}) = \inf_{\delta\in\mathcal{D}} r(\pi,\delta) < \infty$$

Pour obtenir la valeur de l'infimum du risque intégré il faut donc en théorie minimiser une intégrale double δ. L'introduction du risque intégré se justifie par le théorème suivant. Il suffira de minimiser une grandeur qui ne dépend plus que des données, ceci permet donc d'arriver à des estimateur satisfaisants.

Théorème 1.1. *(Rousseau,*(2010)*)*
Méthode de calcul
Si $\exists\delta \in \mathcal{D}$, $r(\pi,\delta) < \infty$ *et* $\forall x \in \mathcal{X}$ $\delta^{\pi}(\pi) = \arg\min_{\delta} \rho(\pi,\delta/x)$ *alors*
$\delta^{\pi}(x)$ *est un estimateur bayésien.*

Preuve.

$$\begin{aligned} r(\pi,\delta) &= \int_{\Theta} R(\theta,\delta)d\pi(\theta) \\ &= \int_{\Theta}\int_{x} L(\theta,\delta(x))/x)f(x/\theta)d\mu(x)d\pi(\theta) \\ &= \int_{x}\int_{\Theta} L(\theta,\delta(x))\frac{f(x/\theta)d\pi(\theta)}{m_{\pi}(x)}d\mu(x) \\ (Fubini) &= \int_{x}\int_{\Theta} L(\theta,\delta(x))d\pi(\theta/x)m_{\pi}d\mu(x) \\ &= \int_{x} \rho(\pi,\delta/x)m_{\pi}(x)d\mu(x) \end{aligned}$$

Ainsi, pour $\delta \in \mathcal{D}$,$\rho(\pi,\delta^{\pi}/x) \leq \rho(\pi,\delta/x) \Rightarrow r(\pi,\delta^{\pi}) \leq r(\pi,\delta)$.Ce qui permet de conclure.

1.5 Admissibilité

Définition 1.9. Une règle de décision δ_1 est dite meilleure que δ_2 si son risque associé est moins que celui associé à δ_2, c'est-à-dire si

$$\begin{cases} \mathrm{R}(\delta_1,\theta) \leq R(\delta_2,\theta) \ , \forall\theta \in \Theta; \\ \mathrm{R}(\delta_1,\theta) < R(\delta_2,\theta) \ , \ \text{pour au moins une valeur de } \theta. \end{cases}$$

Une décision δ est la meilleure de toutes les décision si et seulement si sa fonction de risque est la plus petite.

Définition 1.10. Estimateur admissible
On dit que $\delta \in \mathcal{D}$ est inadmissible si seulement si :

$$(\exists\delta_0 \in \mathcal{D},\forall\theta \in \Theta : R(\theta,\delta) \geq R(\theta,\delta_0) et \exists\theta_0 \in \Theta : R(\theta_0,\delta) > R(\theta_0,\delta_0)).$$

De ce fait, δ est admissible si elle n'est pas inadmissible et par conséquence, un estimateur est dit admissible si seulement s'il n'est pas inadmissible.

Théorème 1.2. *Estimateurs bayésiens admissibles, (Rousseau,*(2010)*)*

Si l'estimateur bayésien δ^π associé à une fonction de perte L et une loi a priori π est unique, alors il est admissible.

Preuve. Supposons estimateur bayésien non admissible:

$$\exists \delta_0 \in \mathcal{D}, \forall \theta \in \Theta R(\theta,\delta^\pi) \geq R(\theta,\delta_0) et \exists \theta_0 \in \Theta, R(\theta_0,\delta^\pi) > R(\theta_0,\delta_0).$$

En intégrant la première inégalité :

$$\int_\Theta R(\theta,\delta_0) d\pi(\theta) \leq \int_\Theta R(\theta,\delta^\pi) d\pi(\theta) = r(\pi)$$

donc δ_0 est aussi un estimateur bayésien associé à L et π et $\delta_0 \neq \delta^\pi$ d'après la seconde inégalité. Le théorème se déduit par contraposée.
Ce théorème s'applique notamment dans le cas d'un risque fini et d'une fonction de coût convexe. En outre, l'unicité de l'estimateur bayésien implique la finitude du risque :

$$r(\pi) = \int R(\theta,\pi(\theta)) d\pi(\theta) < \infty$$

(sinon, tout estimateur minimise le risque).

1.5.1 Minimaxité

Définition 1.11. un estimateur δ_0 est minimax si et seulement si

$$\sup_\theta R(\theta,\delta_0) \leq \sup_\theta R(\theta,\delta_1)$$

C'est un estimateur dont le risque maximal est le plus petit de tous les risques maximaux.

Proposition 1.2. *(Christian P. Robert 2006)*
S'il existe un unique estimateur δ_0 minimax, cet estimateur est admissible.

Preuve. Supposons que δ_0 n'est pas admissible, alors il existe δ_1 telle que

$$R(\theta,\delta_1) < R(\theta,\delta_0)$$

$\forall\theta \in \Theta$ avec inégalité stricte pour au moins une valeur de θ d'où

$$\sup_{\theta} R(\theta,\delta_1) \leq \sup_{\theta} R(\theta,\delta_0)$$

pare ce que δ_0 est minimax, en contradiction avec l'hypothése que δ_0 est l'unique estimateur minimax.Donc δ_0 est minimax.

1.6 L'estimateur MAP.

On appelle estimateur MAP (estimateur de maximum a posteriori) tout estimateur $\delta^{\pi}(x)$ qui maximise l'information sur θ représentée par sa loi a posteriori, c'est-à-dire tout estimateur $\delta^{\pi}(x) \in \arg\max_{\theta} \pi(\theta/x)$. $\delta^{\pi}(x)$ doît donc être le mode de la distribution a posteriori.
Le grand avantage de cet estimateur est qu'il ne dépend pas d'une fonction de perte, est utile pour les approche théoriques. L'estimateur MAP est le pendant Bayésien de l'estimateur de maximum de vraisemblance, de ce fait ils partagent les mêmes inconvénients comme:la non unicité, l'instabilité(dus aux calculs d'optimisation), ..., etc.

1.7 Choix de loi a priori.

La loi a priori est la clé de voute de l'inférence bayésienne et sa détermination est donc l'étape la plus importante dans la mise en oeuvre de cette inférence. Dans une certaine mesure, c'est aussi la plus difficile. Evidemment, dans la pratique, il est rare que l'information a priori soit suffisamment précise pour conduire à une détermination exacte de la loi a priori, au sens où plusieurs lois de probabilité peuvent être compatibles avec cette information. Afin d'obtenir une loi a priori Il est donc nécessaire le plus souvent de faire un choix (partiellement) arbitraire de loi a priori, ce qui peut avoir un impact considérable sur l'inférence qui en découle. Historiquement, les détracteurs du paradigme bayésien ont concentré leurs critiques sur le choix de la loi a priori, en commençant par celui effectué par Laplace. En particulier, l'utilisation systématique de lois usuelles (normale, gamma, bêta, etc.) et la restriction plus forte encore aux lois conjuguées ne sont pas toujours justifiées, car la détermination subjective de la loi a priori qui en résulte se fait au prix d'un traitement analytique plus fruste du problème, puisque ignorant une

partie de l'information a priori. Ces critiques contre l'approche bayésienne ont une certaine validité au sens où elles attirent l'attention sur le fait qu'il n'y a pas une façon unique de choisir une loi a priori, et que le choix de cette loi a un impact sur l'inférence résultante. Cet impact peut être négligeable, modéré ou énorme, puisqu'il est toujours possible de choisir une loi a priori qui donnera la réponse qu'on souhaite obtenir.
Mais le point essentiel est ici que, premièrement, les lois a priori non fondées fournissent des inférences a posteriori non justifiées et, deuxièmement, le concept d'une loi a priori unique n'a pas de sens, sauf dans des cas très particuliers.
Par conséquence deux approche intervint: l'approche a priori conjuguée, qui nécessite une quantité limitée d'information, et l'approche non informative, qui est obtenue à partir de la distribution de l'échantillon.

1.7.1 Lois a priori impropres.

Dans certains cas, il est important de prendre des prioris qui ne sont pas de véritable densité de probabilités. Elle sont définiés comme suit

$$\int_{\Theta} \pi(\theta) d\theta = +\infty$$

On dit alors qu'elles sont impropres, notamment sont utiles dans les modèles non-informatifs.

Approche partiellement informative

Maximum d'entropie

Si l'on possède des informations partielles du type $\mathbb{E}^{\pi}[g_k(\theta)] = \mu_k$ où pour chaque $k = 1,...,n, g_k$ est une fonction donnée, on cherche la loi la moins informative sous ces contraintes, seules informations dont on dispose. Pour comparer le caractère informatif, il est nécessaire d'avoir recours à un critère d'information. L'entropie de Shannon permet de définir ce niveau d'informativité, nous présentons dans un premier temps cette entropie dans le cas fini et discret.
Pour $\theta \in \{1,...,n\}$; et $\pi(\theta) = (\pi_1,...,\pi_n)$ tel que $\pi_i \geq 0$ et $\sum_{i=1}^{m} \pi_i = 1$, l'entropie de la loi est définie par:

$$Ent(\pi) = -\sum_{i=1}^{m} \pi_i \log(\pi_i) \leq -\sum_{i=1}^{m} \frac{1}{m} \log(\frac{1}{m}) = \log m$$

Ce dernier terme correspond à une répartition uniforme, la loi la plus " plate ", la plus désordonnée. Pour la masse de Dirac $\delta(j)$,

$$Ent(\delta(j)) = 0$$

Qui correspond à l'intuition puisqu' alors il n'y a plus d'incertitude et l'information est totale. Une entropie petite s'interprète comme une loi concentrée et informative. La maximisation de l'entropie sous les contraintes permet de chercher la loi qui apporte le moins d'information. Le principe à la base de cette méthode est donc de chercher à calculer:

$$\arg\max_{\pi} Ent(\pi)$$

sous la contrainte

$$\mathbb{E}^{\pi}[g_k(\theta)] = \mu_k$$

La solution de ce problème est alors donnée par:

$$\pi^* \propto e^{\sum_{k=1}^{n} \lambda_k g_k(\theta)}$$

où les λ_k sont les multiplicateurs de Lagrange associés. Dans la pratique, on détermine ces valeurs λ à partir des contraintes (systèmes d'équations) comme l'indique l'exemple à suivre.

Exemple 5. *Un cas dénombrable*
Ici, $\Theta = \mathcal{N}$ et $\mathbb{E}^{\pi}[\theta] = x > 1$, c'est-à-dire qu'ici $g(\theta) = \theta$ et $\mu = x$.
On sait $\pi^* \propto e^{\lambda\theta}$ et que λ est déterminé par :

$$\sum_{\theta \in \mathcal{N}} \frac{\theta e^{\lambda\theta}}{e^{\lambda\theta}} = x$$

Cela conduit à résoudre :

$$\frac{x}{1 - e^{\lambda}} = \frac{1}{e^{\lambda}} \frac{e^{\lambda}}{(1 - e^{\lambda})^2}$$

d'où

$$e^{\lambda} = \frac{x - 1}{x}$$

Par exemple si $x = \frac{12}{11}$ alors $\lambda = -log(12)$.
En continu, il n'est pas possible de définir l'entropie comme ci-dessus puisqu'on ne peut dénombrer les états (pas de mesure de comptage) en l'absence de mesure de référence. Dans le cas continu, on définit alors l'équivalent de l'entropie par rapport à une mesure π_0 :

$$Ent(\pi/\pi_0) = \int_{\Theta} \pi(\theta) \log(\frac{\pi(\theta)}{\pi_0(\theta)}) d\theta$$

C'est en fait la divergence de Kullback. Dans l'idée π_0 est la plus plate possible, la plus proche de la répartition uniforme, c'est en fait l'équivalent de la répartition en $\frac{1}{m}$ de l'information discrète. L'objectif est donc de maximiser $Ent(\pi/\pi_0)$ sous les contraintes $\mathbb{E}^\pi[g_k(\theta)]$. Là encore, la solution générale est connue :

$$\pi^*(\theta) \propto e^{\sum_{k=1}^n \lambda_k g_k(\theta)} \pi_0(\theta)$$

Lorsque la structure est bonne, des choix raisonnables de π_0 sont la mesure de Haar (pour les groupes) ou bien la loi de Jeffreys.

Exemple 6. *Un cas continue*
Si le modèle est de la forme $f(X-\theta)$ et si l'on choisit $\pi_0(\theta)$ alors :

$$\mathbb{E}^\pi[\theta] = \mu$$

$$(1.7)$$

$$et \mathbb{V}^\pi(\theta) = \sigma^2$$

sont connus alors la théorie prédit:

$$\pi(\theta) \propto e^{\lambda_1\theta+\lambda_2\theta^2}$$

C'est donc la loi normale $\mathcal{N}(\theta,\sigma^2)$ Si $\mathbb{E}^\pi[\theta] = \mu$ alors la théorie donne:

$$\pi(\theta) \propto e^{\lambda\theta}$$

On n'a donc pas de solution sur $\mathbb{R}$ puisque dans ce cas ou bien $\theta < 0$ ou bien $\theta > 0$
Ce dernier résultat est paradoxal: avec une information supplémentaire, la variance de θ, l'intervalle dans lequel évolue θ est agrandi, et une région exclue dans un cas plus large (le second) devient accessible, cela n'est pas loin de signifier que la conclusion antérieure qu'une région doit être exclue n'est pas si évidente. Suivant les contraintes, il est donc possible de ne pas trouver de solutions.

De plus, le problème repose sur le choix de π_0 et non du modèle (ou de sa géométrie), cela constitue une limite de cette approche ou tout du moins, un point important à souligner.

1.8 Familles conjuguées

On considère une variable x suivant une fonction de densité paramétrique absolument continue par rapport à la mesure de Lebesgue :$x \sim f(x/\theta)$

Définition 1.12. Famille conjuguée
On dit que la famille de lois a priori $\{\pi_\gamma,\gamma \in \Gamma\}$ est conjuguée si et seulement si:

$$\forall x,\forall \gamma \in \Gamma,\pi_\gamma(\theta/x) \in \pi_\gamma,\gamma \in \Gamma$$
$$\Longleftrightarrow \forall \gamma \in \Gamma,\forall x,\exists \grave{\gamma}(x) \in \Gamma telque$$
$$\pi_\gamma(\theta/x) = \pi_{\grave{\gamma}(x)}(\theta)$$

L'avantage des familles conjuguées est avant tout de simplifier les calculs. Avant l'essor du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs.

Exemple 7. *Lois normales et inverse-gamma Pour* $x \sim \mathcal{N}(\theta,\sigma^2)$, $\pi(\theta,\sigma^2)$,$\pi(\theta/X,\sigma^2)$, $\theta/\sigma^2 \sim exp(\mu,\tau\sigma^2)$ *et* $\sigma^2 \sim \mathbb{IG}(a,b)$

$$\begin{aligned}\pi(\theta,\sigma^2,/x) &\propto exp(\frac{\frac{-(x-\theta)^2}{2\sigma^2} - \frac{-(\theta-\mu)^2}{2\tau\sigma^2}}{\sigma^2}) \times (\sigma^2)^{-(a+b)} e^{\frac{-b}{\sigma^2}} \\ &\propto \frac{1}{\sigma^2} exp(-\frac{1}{2\sigma^2}(1+\frac{1}{\tau})[\theta - (x+\frac{\mu}{\tau})(1+\frac{1}{\tau})^{-1}]) \times \frac{e^{-\frac{x^2}{2\sigma^2} - \frac{\mu}{2\tau\sigma^2} - \frac{b}{\sigma^2}}}{(\sigma^2)^{a+\frac{3}{2}}} \times e^{\frac{1}{2\sigma^2}(x+\frac{\mu}{\tau})^2 \frac{\tau}{1+\tau}}\end{aligned}$$

Donc la loi a posteriori est:

$$\begin{aligned}(\theta/\sigma^2,x) &\sim \mathcal{N}((x+\frac{\mu}{\tau})\frac{\tau}{1+\tau}, \frac{\sigma^2\tau}{1+\tau}) \\ et(\sigma^2/x) &\sim \mathcal{IG}(a+\frac{1}{2}, b+\frac{x^2}{2}+\frac{\mu^2}{2\tau} - \frac{(x+\mu)^2}{2}\frac{\tau}{1+\tau})\end{aligned}$$

Un autre exemple est celui des lois binomiales dont les lois $\pi(p) = Beta(a,b)$ *constituent une famille conjuguée.*

Définition 1.13. Familles exponentielles
La famille exponentielle regroupe les lois de probabilité qui admettent une densité de la forme:

$$f(x/\theta) = e^{\alpha(\theta)'T(x)-\psi(\theta)}h(x)$$

, $\theta \in \Theta$.

T est alors une statistique exhaustive.

Une telle famille est dite régulière si Θ est un ouvert tel que

$$\Theta = \{\theta / \int e^{\alpha(\theta)'T(x)}h(x)d\mu(x) < \infty\}.$$

En outre, on appelle paramétrisation canonique, l'écriture:

$$f(x/\theta) = e^{\theta'T(x)-\psi(\theta)}h(x)$$

et famille naturelle, l'expression $f(x/\theta) = e^{\theta'T(x)}k(x)$.

Théorème 1.3. *Famille exponentielles*

Si $x \sim f(x/\theta) = e^{\alpha(\theta)'T(x)-\Theta(\theta)}h(x)$*, alors la famille de lois a priori*

$$\{\pi_{\lambda,\mu}(\theta) \propto h(x)e^{\theta\mu-\lambda\psi(\theta)},\lambda,\mu\}$$

est conjuguée. On note que $\pi_{\lambda,\mu}$ *est une densité de probabilité si et seulement si* $\lambda > 0$ *et* $\mu/\lambda \in \Theta$*. La loi a posteriori correspondante est* $\pi(\theta/\lambda + 1,\mu + T(x))$.
En effet,

$$\begin{aligned}
\pi_{\lambda,\mu}(\theta/x) &\propto e^{\alpha(\theta)'T(x)-\Theta(\theta)}e^{\theta\mu-\lambda\psi(\theta)h(x)} \\
&\propto h(x)e^{\theta(T(x)+\mu)-(\lambda+1)\psi(\theta)} \\
&= \pi_{\lambda+1,\mu+T(x)}.
\end{aligned}$$

1.9 Approche non informative

Lorsque aucune information a priori n'est disponible,le choix de la loi a priori est ananlytique, puisque'elle donnent des expressions exacte pour quelques quantités a posteriori.Dans de teles situations, il est impossible de justifier le choix d'une loi a prioro sur des bases subjectives. Plutôt que de revenir aux alternatives classiques, comme l'estimation par maximum de vraisemblance, c'est préférable puisque c'est la seul information disponible, de telles lois sont dites non informative.

1.9.1 Loi de jeffreys

Les lois a priori non informatives de Jeffreys sont fondées sur l'information de Ficher, donnée par

$$I(\theta) = \mathbb{E}_\theta[(\frac{\partial log f(x/\theta)}{\partial \theta})^2]$$

Dans le cas unidimensionnel. Sous certaines conditions, cette formation est aussi égale à

$$I(\theta) = -\mathbb{E}_\theta[(\frac{\partial^2 log f(x/\theta)}{\partial \theta^2})]$$

La loi priori de Jeffreys est

$$\pi_j \propto \sqrt{\mid I(\theta) \mid}$$

ansi la lois a priori de Jeffreys est invariante reparmétrisation.
Malgré l'immense intéret d'une telle proprieté, il faut savoir que la loi a priori de jeffreys n'a de bonne proprieté que dans le cas des petites dimensions est en particulier de la dimension 1.

Exemple 8. *Soit* $x \sim \mathcal{B}(n,p)$

$$\begin{aligned} f(x/p) &= C_n^x p^x (1-p)^{n-x} \\ \frac{\partial^2 \log f(x/p)}{\partial^2 p^2} &= \frac{x}{p^2} + \frac{n-x}{(1-p)^2} \\ I(p) &= n[\frac{1}{p} + \frac{1}{p-1}] \\ &= \frac{n}{p(1-p)} \end{aligned}$$

Donc la loi de Jefferys pour ce modèle est

$$\pi(p) \propto [p(1-p)]^{-1/2}$$

et est alors propre, car il s'agit de la distribution $\mathcal{B}e(1/2,1/2)$

1.10 Conclusion

L'approche bayésienne fournit un guide de raisonnement scientifique face à l'incertitude. En effet, dès qu'on dispose de la meilleure connaissance possible des quantites incertaines, on mobilise toute l'information disponible, ce qui justifie le choix bayesien. Donc on peut donner la distribution de probabilite de toute grandeur interessante pour le décideur.

Chapitre 2

Les modèles hiérarchiques Bayésiens

2.1 Introduction

La notion de lois a priori est insuffisante pour rendre pleinement compte de l'ignorance, car l'information a priori est rarement assez riche pour en déduire une loi a priori exacte. Il est alors nécessaire d'incorporer cette incertitude au modèle bayésien. Il s'agit alors de modéliser l'information a priori en la décomposant en plusieurs niveaux de distributions a priori conditionnelles, ce qui définit "la modèlisation bayésienne hiérarchique" qui fournit une base solide à l'analyse bayésienne dans le cas d'information a priori incomplète.

2.2 Analyse bayésienne hiérarchique.

Un des problèmes d'une approche bayésienne classique noté par les fréquentistes est l'incertitude concernant la loi a priori. Une approche qui essaie de rectifier ce problème est l'analyse bayésienne hiérarchique qui met des mesures a priori sur les paramètres de la loi a priori $\pi(\theta)$.

2.2.1 Modéle hiérarchique.

Pour des raisons liées à la modélisation des observations ou à la décomposition de l'information a priori, il peut arriver que le modéle statistique bayésien soit hiérarchique, c'est-à-dire mette en jeu plusieurs niveaux de distributions a priori conditionnelles.

Exemple 9. *Soit $x \sim (\theta)$.Considérant une loi exponentielle, d'une loi a priori de paramètre θ_1. La démarche hiérarchique nous conduit à considérer alors une loi a priori sur θ_1 ; On peut prendre par exemple une loi exponentielle de paramètre ξ. On a donc les lois suivantes : $\pi(\theta/\theta_1) = \theta_1 e^{-\theta_1\theta}$. et $\xi e^{-\xi\theta_1}$ On peut bien évidemment continuer à emboîter la démarche bayésienne.*

Définition 2.1.

On appelle modèle bayésien hiérarchique de niveau n, un modèle statistique bayésien avec densité conditionnelle de x sachant θ, $f(x/\theta)$, la densité a priori $\pi(\theta)$ décomposée en plusieurs lois conditionnelles , c'est-à-dire :

$$x/\theta \sim f(x/\theta)$$

$$\theta/\theta_1 \sim \pi_1(\theta/\theta_1)$$

$$\theta_1/\theta_2 \sim \pi_2(\theta_1/\theta_2)$$

$$\vdots$$

$$\theta_{(n-1)}/\theta_n \sim \pi_n(\theta_{(n-1)}/\theta_n)$$

et une loi marginale

$$\theta_n \sim \pi_{n+1}(\theta_n)$$

$$\pi(\theta) = \int_{\Theta_1\times\ldots\times\Theta_n} \pi_1(\theta/\theta_1)\pi_2(\theta_1/\theta_2)\ldots\pi_{n+1}(\theta_n)d\theta_1\ldots d\theta_n$$

Les paramétre θ_i sont appelés hyperparamètres de niveau
$i(1 \leq i \leq n)$ et $\pi_{n+1}(\theta_n)$ une loi marginale.

2.2.2 Robustesse par rapport à la loi a priori(robustesse informelle)

Le statisticien s'est intéressé comme première étape à proposer un modèle qui explique le comportement des observations, une loi a priori qui génère le paramètre d'intérêt et une fonction de perte qui utilisée pour évaluer le risque
Dans la pratique, il est rare de pouvoire proposer une détermination explicite du modèle, de la loi a priori et de fonction de perte même si on dispose de certains information. La robustesse Bayésienne consiste à évaluer l'influence de cette

indétermination sur les quantités d'intérêt.

La robustesse consiste à construire une classe de modèles(ici lois a priori), et étudier par la suite les changements éffectués sur les quantités a posteriori autours de cette classe. La robustesse est réalisée si il n'a pas un grandes changement entre les moyennes a posteriori sous les lois a priori, c-à-d le choix des lois a priori n'a pas d'influence.

Remarque 2.1. – Un modéle bayésien hiérarchique n'est rien d'autre qu'un cas particulier de modéle bayésien. Ainsi, si

$$x \sim f(x/\theta)$$

$$\theta/\theta_1 \sim \pi_1(\theta/\theta_1),...,\theta_n \sim \pi_{n+1}(\theta_n)$$

on retrouve le modéle bayésien usuel

$$x \sim f(x/\theta)$$

$$\theta \sim \pi(\theta)$$

Pour l'a priori

$$\pi(\theta) = \int_{\Theta_1 \times ... \times \Theta_n} \pi_1(\theta/\theta_1)\pi_2(\theta_1/\theta_2)...\pi_{n+1}(\theta_n)d\theta_1...d\theta_n$$

Si les hyperparamétres θ_1,...,θ_n ne sont d'aucun intérêt. L'inférence (sur θ), il est équivalent de considérer le modéle hiérarchique plus simple.

$$x/\theta \sim f(x/\theta)$$

$$\theta/\theta_1 \sim \pi_1(\theta/\theta_1)$$

avec

$$\theta_1 \sim \pi_2(\theta_1) = \int_{\Theta_2 ...,\Theta_n} \pi_1(\theta_1/\theta_2)...\pi_{n+1}(\theta_n)d\theta_2...d\theta_n$$

Une décomposition plus compliquée peut toujours se justifier pour la construction et le calcul pratique d'estimateurs de Bayes.

– La décomposition suivante

$$\pi(\theta) = \int_{\Theta_1} \pi_1(\theta/\theta_1)\pi_2(\theta_1)d\theta_1$$

peuveut être privilégiées pour un certain nombre de raisons:

Un aspect positif de l'analyse bayésienne hiérarchique est qu'elle augmente également la robustesse(vue au dessus) de l'analyse bayésienne classique d'un point de vue fréquentiste, puisqu'elle réduit l'arbitraire sur le choix de l'hyperparamètre (parfois reporté à un niveau plus élevé) et établit une moyenne des réponses bayésiennes conjuguées.

La décomposition d'une loi a priori π en plusieurs composantes π_1,...,π_n (qui peuvent être, par exemple, des lois conjuguées) permet parfois d'obtenir des approximations plus aisées de certaines quantités a posteriori.

2.2.3 Décomposition conditionelle.

Une caractéristique particulièrement intéressante des modèles hiérarchique est que le conditionnement est possible à tous les niveaux et cette liberté dans la décomposition de la loi a posteriori componse l'augmentation apparente de complexité de la structure. Par exemple, si

$$\theta/\theta_1 \sim \pi_1(\theta/\theta_1)$$

$$\theta_1 \sim \pi_2(\theta_1)$$

nous avons les résultats suivant.

Lemme 2.1. *(Christain P.Robert,* 2006*)*
La loi a posteriori de θ *est*

$$\begin{aligned}
\pi(\theta/x) &= \int_{\Theta_1} \pi(\theta/\theta_1,x)\pi(\theta_1/x)d\theta_1 avec \\
\pi(\theta/\theta_1,x) &= \frac{f(x/\theta)\pi_1(\theta/\theta_1)}{m_1(x/\theta_1)} \\
m_1(x/\theta_1) &= \int_{\Theta} f(x/\theta)\pi_1(\theta/\theta_1)d\theta \\
\pi(\theta_1/x) &= \frac{m_1(x/\theta_1)\pi_2(\theta_1)}{m(x)} \\
m(x) &= \int_{\Theta_1} m_1(x/\theta_1)\pi_2(\theta_1)d\theta_1
\end{aligned}$$

Autrement dit:

$$\pi(\theta/x) = \frac{\int_{\Theta_1} f(x/\theta)\pi(\theta/\theta_1)\pi_2(\theta_1)d\theta_1}{\int_{\theta}\int_{\Theta_1} f(x/\theta)\pi(\theta/\theta_1)\pi_2(\theta_1)d\theta_1 d\theta}.$$

Ce résultat découle naturellement du théorème de Bayes, l'égalité de dénominateur et une conséquence de théorème du Fubini.
Il n'en a pas moins des conséquences importantes sur le calcul des estimateurs de Bayes puisqu'il montre qu'on peut simuler $\pi(\theta/x)$ *en générant d'abord* θ_1 *selon* $\pi(\theta_1/x)$ *puis* θ *selon* $\pi(\theta/\theta_1,x)$*, dans le cas où ces deux lois conditionnelles sont plus accessibles.*

Preuve. En remplaçant $\pi(\theta/\theta_1,x)$ et $\pi(\theta_1/x)$ pour leur expression sous l'intégrale:

$$\begin{aligned}
\pi(\theta/x) &= \int_{\Theta_1} \frac{f(x/\theta)\pi_1(\theta/\theta_1)}{m_1(x/\theta_1)} \frac{m_1(x/\theta_1)\pi_2(\theta_1)}{m(x)} d\theta_1 \\
&= \int_{\theta_1} \frac{f(x/\theta)\pi_1(\theta/\theta_1)\pi_2(\theta_1)}{m(x)} d\theta_1 \\
&= \frac{f(x/\theta)}{m(x)} \int_{\theta_1} \pi_1(\theta/\theta_1)\pi_2(\theta_1)d\theta_1 \\
&= \frac{f(x/\theta)\pi(\theta)}{m(x)}
\end{aligned}$$

Lemme 2.2. *(Christain P.Robert,* 2006*)*
Si la loi marginale

$$m(x) = \int_{\Theta} f(x/\theta)\pi(\theta)d\theta$$

est finie pour tout $x \in \mathbb{R}^{\mathbb{K}}$*, alors la moyenne et la variance de la loi a posteriori* $\pi(\theta/x)$ *existent toujours.*

Lemme 2.3. *(Christain P.Robert,* 2006*)*
Pour le modéle hiérarchique, la densité conditionnelle compléte de θ_i *sachant* x *et les* θ_j $(j \neq i)$ *vérifie*

$$\pi(\theta_i/x,\theta,\theta_1,...,\theta_n) = \pi(\theta_i/\theta_{i-1},\theta_{i+1})$$

avec la convention $\theta_0 = \theta$ *et* $\theta_{n+1} = 0$ *puisque*

$$\begin{aligned} \pi(\theta_i/x,\theta,\theta_1,...,\theta_n) &\propto f(x/\theta)\pi_1(\theta/\theta_1)...\pi_{n+1}(\theta_{n+1}) \\ &\propto \pi_{i-1}(\theta_{i-1}/\theta_i)\pi_i(\theta_i/\theta_{i+1}) \end{aligned}$$

la distribution a posteriori ne dépend que des deux niveaux adjacents de la hiérarchie.

2.2.4 Problèmes numériques.

Un inconvénient des modéles hiérarchiques est qu'ils ne permettent en général pas un calcul explicite des estimateurs de Bayes, même lorsque les niveaux successifs sont conjugués, et il faut donc avoir recours à des techniques numériques d'approximation.

Exemple 10.
On considère $x \sim \beta(n,p)$ *et* $p/m \sim \beta e(m,m)$ *avec* $m \in \mathcal{N}^*$*. Alors,*

$$\begin{aligned} \pi_1(p/m) &= \frac{\Gamma(2m)}{\Gamma(m)^2}[p(1-p)]^{m-1} \\ &= (2m-1)\begin{pmatrix} 2m-1 \\ m-1 \end{pmatrix}[p(1-p)]^{m-1} \end{aligned} \tag{2.2}$$

Si la loi a priori de second niveau est $\pi_2(m) = 1/(2m-1)$*, la loi a priori sur* p *est*

$$\begin{aligned} \pi(p) &= \int_{N^*} \pi_1(p/m)\pi_2(m)dm \\ &= \sum_{n=0}^{+\infty} \left(\begin{array}{c} 2n \\ n \end{array} \right) [p(1-p)]^n \end{aligned} \tag{2.4}$$

La loi a posteriori

$$\pi(p/x) = \int \pi_1(p/m,x)\pi_2(m/x)dm$$

ne peut être obtenue analytiquement puisque même si $(p/m,x)$ *est une loi bêta* $\beta(m+x,m+n-x)$*est la loi bêta-binomiale*

$$\frac{(m+x-1)...m(m+n-x-1)...m}{(2m+n-1)...(2m)(2m-1)}$$

à un facteur de normalisation près.Les quantités a posteriori comme $\mathbb{E}^{\pi}[p/x]$ *ne sont pas calculables analytiquement.*
La solution la plus naturelle en analyse hiérarchique est de faire appel à des outils de simulation.

2.2.5 Extensions hiérarchiques du modèle normal.

Nous considérons le cas particulier de la loi normale $x \sim \mathcal{N}_p(\theta,\Sigma)$, faisons appel à une loi conjuguée de premier niveau $x \sim \mathcal{N}_p(\mu,\Sigma_\pi)$ pour une décomposition plus facile des estimateurs.

Lemme 2.4. *(Christain P.Robert,* 2006*)*
Dans le modèle normal conjugué, l'estimateur de Bayes hiérarchique est

$$\begin{aligned}
\delta^\pi(x) &= \mathbb{E}^{\pi_2(\mu,\Sigma_\pi/x)}[\delta(x/\mu,\Sigma_\pi)] \\
avec & \\
\delta(x/\mu,\Sigma_\pi) &= x - \Sigma W(x-\mu) \\
W &= (\Sigma+\Sigma_\pi)^{-1} \\
\pi_2(\mu,\Sigma_\pi/x) &\propto (detW)^{\frac{1}{2}} exp\{-(x-\mu)^t W(x-\mu)/2\}\pi_2(\mu,\Sigma_\pi) to
\end{aligned}$$

Exemple 11. *Le choix d'une loi a priori constante sur β donne une expression analytique de $\delta^\pi(x)$. Il existe alors une fonction h_k telle que*

$$\begin{aligned}
\delta^\pi(x) &= x - h_p - k - 2(\parallel x \parallel^2)\Sigma C^{-1}(x-P_x) \\
P &= Y(Y^tC^{-1}Y)^{-1}Y^tC^{-1} \\
\parallel x \parallel^2 &= xC^{-1}(I_p-P)x
\end{aligned}$$

Telle que Px est la projection orthogonale de x sur le sous-espace $H = \{\mu = Y\beta,\beta \in \mathbb{R}^{\mathbb{K}}\}$ selon la métrique définie par C^{-1}.
L'estimateur $\delta^\pi(x)$ est donc une somme pondérée de x et de cette projection.
Par conséquent, $\delta^\pi(x)$ prend en compte l'information a priori de façon adaptative, en fonction de la distance $\parallel x \parallel$ de x à H.

2.2.6 Choix bayésien empirique.

La méthode bayésienne empirique est utilisée pour l'estimation des paramètres d'une loi quand les observations sont des variables aléatoires indépendantes identiquement distribuées

qui suivent cette loi. On peut mentionner que cette technique n'est pas une méthode bayésienne pure car elle fait appel à des approximations fréquentistes dans le cas où l'information a priori n'est pas disponible ou est insuffisante. Cependant un résultat montre qu'on peut obtenir des résult ats asymptotiquement équivalents à ceux du modèle bayésien hiérarchique. L'analyse bayésienne empirique peut être vue comme une bonne combinaison des méthodes fréquentiste et bayésienne.

Les définitions et exemples donnés ici se retrouvent dans plusieurs références. Nous notons Berger (1985) *et Robert* (1992) *pour la partie bayésienne et bayésienne hiérarchique et Maritz et Lwin* (1989) *pour la partie bayésienne empirique. Nous indiquons également le lien entre les estimat eurs bayésiens empiriques et les estimateurs de type James-Stein.*

L'analyse bayésienne empirique repose sur une modélisation a priori conjuguée, en estimant les hyperparamétres à partir des observations et en utilisant ensuite cet "a priori estimé" comme a priori normal pour l'inférence, c'est t-à-dire remplacemer les hyperparamétres par des hyperparamétres estimés, cela permet au statisticien de tirer parti de l'information a priori vague de maniére simplifiée.
L'analyse bayésienne empirique se présente comme une alternative attrayante lorsque l'analyse bayésienne hiérarchique est trop compliquée à mettre en oeuvre. Donc lorsque l'information a priori est trop limitée la loi a priori à approcher par des méthodes fréquentistes. Enfin comme avantage de cettte alternative on a:

- *elle peut être considérée comme une méthode duale de l'analyse bayésienne hiérarchique présentée*
- *elle est souvent étiquetée "bayésienne" par les fréquentistes et les praticiens*
- *elle constitue dans certains cas une approximation acceptable lorsque la modélisation bayésienne réelle est trop compliquée ou trop chère.*

Une approche bayésienne empirique de Robbins

Une autre façon de résoudre la difficulté de l'incertitude dans la mesure a priori est d'utiliser une approche bayésienne empirique de Robbins(), qui est essentiellement non paramétrique et que l'on appelle estimation bayésienne empirique non paramétrique. Considérons à présent le cas suivant. Les observations passées sont :x_1,x_2,...,x_n*. L'observation présente est* x_{n+1}.

Intuitivement, on peut écrire l'estimateur bayésien empirique de la façon suivante: on suppose $n(x_{n+1})$ *le nombre de fois où on a observé* x_{n+1} *parmi* x_1,x_2,...,x_n *Alors, d'où l'estimateur bayésien empirique de* θ

$$f(x_{n+1}) \text{ est estimé par } \frac{n(x_{n+1})}{n},$$

et si on inclut l'observation courante alors

$$f(x) \text{ est estimé par } \frac{1+n(x_{n+1})}{1+n},$$

$$\text{et } f(n(x_{n+1}+1) \text{ est estimé par } \frac{n(x_{n+1}+1)}{1+n},$$

$$\delta(x_{n+1}) = \frac{(x_{n+1}+1)n(x_{n+1}+1)}{1+n(x_{n+1})}$$

Plus précisément, l'estimation d'une moyenne de la loi de Poisson, la loi marginale $f(x)$ *est égale à*

$$f(x) = \int_0^\infty e^{-\lambda}\frac{\lambda^x}{x!}\pi(\lambda)d\lambda$$

Donc, l'estimateur bayésien empirique de λ_{n+1} *est donné par :*

$$\delta_n(x_{n+1}) = \frac{(x_{n+1}+1)n(x_{n+1}+1)}{1+n(x_{n+1})}$$

Le théorème suivant (El-Habit ahmed, 2007) montre le lien asymptotique entre les méthodes bayésienne empirique et bayésienne hiérarchique.

Théorème 2.1. *Soit* X *une variable aléatoire de densité conditionnelle* $f(x/\theta)$ *alors:*

$$\begin{aligned}\mathbb{E}^{\pi_2}(\lambda/x)[\mathbb{E}^{\pi_1}(\theta/\lambda,x)(\theta)] &= \mathbb{E}^{\pi_1}(\theta/\widehat{\lambda},x)(\theta) + o(n^{(-1)})o\grave{u}\\ \widehat{\lambda} &= \arg\max m(x/\lambda)\end{aligned}$$

2.2.7 Aspects bayésiens empiriques de L'effet Stein.

Définition 2.2.

Estimateurs de James-Stein(El-Habit ahmed, 2007)

Une conséquence de la proposition 1.2 est que s'il existe un seul estimateur minimax, il est admissible (Robert 1992). Réciproquement, si un estimateur minimax δ_0 du risque constant n'est pas admissible, il existe d'autres estimateurs minimax δ qui le dominent uniformément, c'est-à-dire, que $R_\delta(\theta) \leq R_{\delta_0}(\theta)$ pour toute θ de Θ et l'inégalité est stricte pour au moins une valeur de θ (sous quelques conditions de régularité, voir Brown, (1976)). Avant 1955, on pensait que si $\widetilde{X} \sim N(\theta,I)$, alors l'estimateur des moindres carrés $\delta_0(\widetilde{X}) = \widetilde{X}$ était admissible, puisque c'est le seul estimateur minimax de risque constant. Stein a montré que ce résultat ne peut être vrai pour un vecteur de plus de deux composantes. Plus précisément, le risque de l'estimateur du type James-Stein

$$\delta^{JS}(\widetilde{X}) = (1 - \frac{p-2}{\| \widetilde{X} \|}^2)\widetilde{X}$$

est plus petit ou égal au risque de $\delta_0(\widetilde{X}) = \widetilde{X}$ pour toute θ et $R_{\delta^{JS}}(\widetilde{X}) < R_{\delta_0}(\widetilde{X})$.

Si l'on suppose que $\widetilde{\theta} \sim N(\widetilde{\theta},\tau^2 I)$, la loi marginale de $\widetilde{X}$ est alors $N(\widetilde{\theta},(1+\tau^2)I)$.

L'estimateur du maximum de vraisemblance de τ^2 est:

$$\hat{\tau}^2 = \begin{cases} \frac{\|\tilde{x}\|^2}{p} - 1 & \text{si } \frac{\|\tilde{x}\|^2}{p} > p \\ 0 & \text{si } sinom \end{cases}$$

L'analyse bayésienne empirique de L'effet Stein, fournit un cadre d'unification des différentes apparitions(l'analyse empirique paramétrique et nom paramétrique). En outre, cette analyse explique la forme originelle des estimateurs de James-Stein.

Définition 2.3.

Nous commençons par un exemple qui illustre naturellement le fondement bayésien empirique de l'effet Stein.

Exemple 12. *Soient* $x \sim \mathcal{N}_p(\theta,I_p)$, *et* $\theta_i \sim \mathcal{N}(0,\tau^2)$. *La loi marginale de* x *est alors*

$$x/\tau^2 \sim \mathcal{N}_p(0,(1+\tau^2)I_p)$$

et conduit à l'estimateur du maximum de vraisemblance de τ^2*suivant,*

$$\hat{\tau}^2 = \begin{cases} \frac{\|\tilde{x}\|^2}{p} - 1 & si\ \frac{\|\tilde{x}\|^2}{p} > p \\ 0 & si\ sinom \end{cases}$$

L'estimateur bayésien empirique correspondant de sous coùt quadratique est obtenu en remplaçant τ^2 *par* $\hat{\tau}^2$ *dans l'estimateur de Bayes*

$$\begin{aligned} \delta^{EB}(x) &= \frac{\hat{\tau}^2 x}{1+\hat{\tau}^2} \\ &= (1 - \frac{p}{\| \tilde{x} \|^2})^+ x \end{aligned}$$

L'estimateur est en fait un estimateur tronqué de James-Stein. Par conséquent, ces estimateurs peuvent être interprétés en tant qu'estimateurs bayésiens empiriques liés à l'information que les espérances des observations sont proches de 0.

L'estimateur originel de James-Stein peut également s'écrire comme un estimateur bayésien empirique, avec une méthode d'estimation fréquentiste alternative.

Cet exemple illustre aussi les lacunes dans les justifications de l'approche bayésienne empirique, qui ne sait pas comparer les différentes méthodes d'estimation des hyperparamètres.

Remarque 2.2. Optimalité des estimateurs bayésiens hiérarchiques

les estimateurs bayésiens hiérarchiques étant similaires aux estimateurs de Bayes habituels, ils ne sont ni plus ni moins admissibles que les estimateurs de Bayes le sont, mais dans un cas particulier nous verrons qu'il est effectivement possible de tirer parti de la spécificité des estimateurs bayésiens hiérarchiques pour obtenir une condition générale de minimaxité, en utilisant les lois a priori de second niveau.

Ce qui affirme la robustesse gagner de l'approche bayésienne hiérarchique ,en incluant l'information a priori la plus subjective aux niveaux les plus élevés.

2.3 conclusion

La décomposition d'une loi a priori π en plusieurs composantes π_1,...,π_n (qui peuvent être, par exemple, des lois conjuguées) permet parfois d'obtenir des approximations plus aisées de certaines quantités a posteriori par simulation. Donc, la capacité de l'approche bayésienne hiérarchique a donner des simplifications des calculs bayésiens. L'analyse bayésienne hiérarchique augmente également la robustesse de l'analyse bayésienne classique d'un point de vue fréquentiste, puisqu'elle réduit l'arbitraire sur le choix de l'hyperparamètre (parfois reporté à un niveau plus élevé) et établit une moyenne des réponses bayésiennes conjuguées.

Chapitre 3

Les modéles hiérarchiques en psychologie

3.1 Introduction et problématique

*Les modéles hiérarchiques bayésiens trouvent des justifications aux problémes réels en médecine, biologie, économie, ... et ainsi en psychologie. On présente ici une application intitulée "Constructing informative model priors using hierarchical methods"(Vanpaemel.W,*2010*), dont le but est de mettre en oeuvre les modèles hiérarchiques en psychologie.*
Pour expliquer le comportement d'un individu qui subit un stimulus, le psychologue a un interêt à chercher les causes et ainsi classer ces stimulus dans une catégorie. Les principales approches ont été utilisées pour identifier ces comportements y compris le principe dit VAM(the Varying Abstraction Model)qui est un modèle de représentation des catégories contenant les deux visions, exemplar et prototype. Un autre principe qui est appelé GCM(General Context Model), est un modèle qui prévoit des catégorisations dans des conditions expérimentales stables. Donc, en utilisant les représentations des catégories(VAM), on peut justifier les réponses aux stimulus et même les prédire.
Le problème peut alors se reformuler comme suit: Utilisant le VAM, comment les modèles hiérarchiques sont utiles pour donner une inférence sur une catégorie ?où bien, comment définir un a priori d'une structure hiérarchique dans le modèle VAM? comment l'inclusion de l'information de l'expert(la loi a priori) aide pour une inférence sur les catégories de modèle VAM? existe t-il une différence entre une inférence classique(non informative a priori) et une inférence bayésienne(basée sur le modèle hiérarchique)nommée informative a priori. Ceci pour déterminer quelle est l'inférence qui sera utile dans notre application et comment utiliser les méthodes bayésienne hiérarchiques pour une inférence de VAM?.

L'article présente principalement l'objectif pratique d'un antérieur instructif(loi a priori) et la relation entre priors objectifs et subjectifs.

Le modèle VAM

Le modèle VAM est un représentant de différentes catégories dans l'espace psychologique, il existe quatres membres de catégories qui représentant un ensemble de stimulus, ce dernier est représenté comme un point dans un espace psychologique.
Les quatres niveaux définis dans la figure sont les quatres membres de catégories réparties comme suit:
La représentation de l'exemplar est placé au panneau supérieur(sommet) de Fig1. La représentation du prototype est au panneau inférieur(fond) dans lequel tous les quatres nombres d'un membre d'une catégorie sont fusionnés et toutes les représentations entre les deux extrêmes sont crées quand deux nombres d'un membre d'une catégorie sont fusionnés.

Le modèle VAM contient le modèle exemplar, prototype et tous les modèles intermédiaires. Pour le modèle exemplar, c'est la vision qui propose des exemplaires de catégories(qui sont l'ensemble des stimulus qui désigne par exemple stauration, taille, ...etc) qui sont enregistrés dans la mémoire. Pour identifier une catégorie, c'est-à-dire pour comprendre les réponses au stimulus, il faut identifier de qu'elle catégorie il s'agit, de qu'elle cause (stimulus) intervient cette conséquence(ces réponses au stimulus). Il suffit alors, d'envisager les point commun à celle stockée dans la mémoire. Le modéle prototype est un résumé de tous les représentations d'une catégorie et les autres sont les résultats du fusionnement des nombres d'un membre d'une catégorie.

Cette figure donne les 15 représentations possibles de VAM pour une catégorie qui contient quatres stimulus. Les deux niveaux extrême correspondent à la représentation de l'exemplar et la représentation du prototype. L'exemplar est le seul qui n'a pas de fusionnement. Dans le prototype modéle on distingue que les 4 stimulus sont tous fusionnés. Entre les deux extrêmes, on remarque des fusions des stimulus à 1,2 et 3.

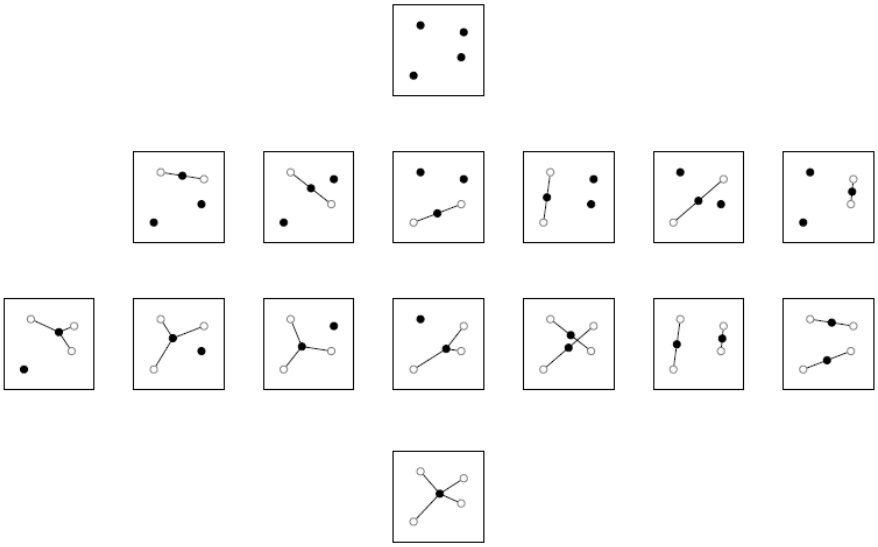

FIG. 3.1 – *Les* 15*représentation possible d'une catégorie avec quatre stimulus*

Le processus de fusionnement

Le VAM va produire un processus nommé le processus de fusionnement avec deux paramétres, θ et γ. Le paramétre θ contrôle la probabilité de fusionnement, ainsi que la possibilité d'avoir le prototype ou l'exemplar modèle. Le paramètre γ contrôle la similarité et les stimulus qui se sont joints.
La probabilité que deux nombres(stimulus) (i,j) d'un membre d'une catégorie se joignent est donné par la formule,

$$p_{ij} = \frac{s_{ij}^{\gamma}}{\sum_x \Sigma y \geq x} s_{xy}^{\gamma}$$

où le s_{ij}, la similarité entre le membre l'ith et jth est modélisé par un exponentiel, en faisant la décomposition du Minkowski r-métrique.
La distance entre deux membres est: $s_{ij} = exp[-\sum_k (|v_{ik} - v_{jk}|^r)^{1/r}]$, avec v_{ik} la coordonnée de dimension kth pour un membre indexé par l'ith. Le paramètre $\gamma \geq 0$ contrôles l'existence d'une similarité d'un paires jointe.

Quand $\gamma = 0$ la similarité n'est pas prise en compte, toutes les paires ont la même probabilité de se joindre. Dans le cas ou les paires jointes sont choisies d'une façon aléatoire, les modèles ont la même probabilité. Si γ augmente, les membres similaires dominent et il ont la probabilité 1 d'ête joints. Par exemple, quand γ est très grand ($\gamma = 10$), seulement les membres de grande similarité ont la possibilité de se joindre (ce que montre la $2^{éme}$ ligne de figure 3.1).

Dans la représentation ci dessus le nombre de modèles dans le VAM est noté par k est mesurable et dépend du nombre de membres de chaque catégorie,deux catégories qui contiennent quatre membres chacun, le VAM contient

$$15 \times 15 = 225$$

modèles paramétrés différents. Les modèles dans la famille VAM seront indexés par M_i ou $i = 1,...,k$
la loi a posteriori pour M_i est

$$P(M_i/x) = \frac{P(x/M_i)P(M_i)}{P(x)}$$

Avec
$P(M_i/x)$est la distribution antérieure $P(M_i)$

$$P(x) = \sum_{i=1}^{k} P(x/M_i)P(M_i)$$

$$P(x/M_i) = \int_{\Omega} P(x/\tau,M_i)P(\tau/M_i)d\tau$$

$P(x/M_i)$est connu comme la probabilité marginale, et il peut être obtenu en intégrant , ou marginaliser partout, le paramètre Ω indique la gamme antérieure du paramètre (vecteur)τ $P(\tau/M_i)$ indique la distribution antérieure sur τ.

La définition de loi a priori(d'une structure hiérarchique) dans le VAM

Pour traduire les connaissances de modèle VAM en distribution a priori, on utilise la structure hiérarchique définit dans le modèle VAM c-à-d l'hiérarchie des modèles(prototype, exemplar). La définition de la loi a priori dans VAM impose trois pas (Lee, 2006). Le première pas, décider quelles informations du modèles sont considérés raisonnables, le deuxième pas, consiste a prendre ces informations pour définir un processus qui génére les modèles qui veut dire décrire les paramètres liés aux modèle. Enfin, définir ce processus génératif signifie donner la loi a priori non uniforme sur les modèles de VAM(voir Lee et Vanpaemel (2008))

- *a)**Le modèle a Priori dans VAM***
  Les modèles de VAM dépends de deux aspects. En premier, ils diffèrent par le nombre (stimulus) du membre de catégorie qui se joignent. En second, lesquels de ces membres se fait joignent, ce qui donne des modèles différents. Le nombre de membres qui se joigne, définit les modèles exemplar et prototype.

  La jonction des nombres(stimulus) du membre d'une catégories, pemet de fournir des similarité du modèles, donc supposés que tout les modèles sont commun à VAM, comme SUSTAIN (Love et al., 2004)et the Rational Model of Categorization (Anderson, 1991; Griffiths, Navarro, 2007).

- *b)**Générer les modèles dans la VAM***

  La deuxième étape de l'extension hiérarchique, est de prendre ces connaissances a priori dans un processus pour générer les modèles. Dans cette application, le processus génére en premier,le modèle exemplar(au sommet de Fig1), car il présente deux connaissances successives et donc deux paramétre hiérarchiques.
  Le processus contrôle nombre de jonction qui sont établies, ce qui donne le premier paramétre, et les membres qui se joignent, ce qui donne le deuxiéme paramétre du processus.
- *c)**La loi a priori***
  le processus génére pour chaque valeur de θ et γ des modèles dans la famille VAM une probabilité $P(M_i/\theta,\gamma)$.
  $La dérniereétape pour définir la loia priori sur M_i$ est obtenus en intégrant sous les pa-

ramètres hiérarchiques θ et γ est donné par:

$$\pi(M_i) = \int_{\Omega_\theta} \int_{\Omega_\gamma} P(M_i/\theta,\gamma)\pi(\theta,\gamma)d\theta d\gamma$$

Le modèle a priori $\pi(M_i)$ est la moyenne de toutes les distributions obtenus de combinaison de θ et γ.Ces dérnier étaient supposé indépendant: c-à-d, $\pi(\theta,\gamma) = \pi(\theta)\pi(\gamma)$ Pour γ, c'était supposé cela

$$\gamma \sim Gamma(2,1)$$

Pour θ, c'était supposé cela

$$\theta \sim Beta(3.2,1.8)$$

La distribution $\pi(M_i)$ donne le processus qui génére cette distribution et comment la loi a priori avec paramétre hiérarchique controle ce processus. Donc explique comment les modèles dans VAM sont générés.

Application utilisant VAM

Chaque membre d'une catégories contient un nombre de stimulus. Un sous-ensemble des stimulus a comparé aux catégories A et B suivant chaque réponse aux stimulus. Les données humaines utilisées dans l'estimation de M_i, sont les réponses aux stimulus présentés dans VAM qui sont proposé comme données. Cela signifie que ces réponses aux stimulus sont les même. Est ce que la catégorisation est la même sur les deux versions(uniform distribution a priori, nom uniforme distribution a priori)?

Pour montrer, comment l'extension hiérarchique de la loi a priori sous les modèles (exemplar, prototype) dans VAM fait la différence dans l'estimation. L'auteur propose deux versions du VAM, VAM_{uni} assume un uniforme antérieur(loi a priori uniforme) sur les modèles, donc chaque modèle du VAM ont la même probabilité a priori ($\pi(M_i) = \frac{1}{k}$), et VAM_{sim} assume une loi a priori obtenu par le processus donné auparavant.

D'après VAM_{uni}, tous les niveaux sont égale, la probabilité qu'un couple se joignent est la même.VAM_{sim} par contre, vise deux intuitions:les deux extrêmes du modèle et les modèles intermédiaires. La probabilité d'avoire une similarité des modèles intermédiaires, est plus grande qu'une jonction des nombres(stimulus) différent des membre d'une catégorie dans les modèles intermédiaire (représenter dans figure 3.2).

La figure(3.2) ces dessous représente les 13 *modèles avec haute masse a priori, dont Les Carrés indiquent une catégorie A et les cercles indiquent la catégorie B. les premiéres hautes barres de cette figure définissent la masse a priori de modèle* VAM_{sim}*, d'où ces quatre modèles correspondent au modèles exemplar, prototype et le mélange de ces deux. Les neufs qui restent représentent les similarités et les jonctions. Cette application au donnée exprime le rôle de l'information a priori dans l'estimation.*

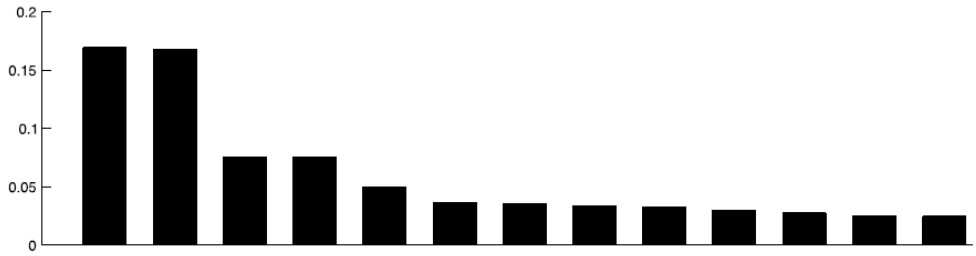

FIG. 3.2 – *Modèles avec la haute masse a priori*

3.2 Résultats

Le but principal de cette inférence, est de tirer une distribution a posteriori de $P(M_i/\theta,\gamma)$, dont Les modèles avec haute masse a posteriori(MAP)(figure 3.2), ce qui- facilite la comparaison entre l'estimation de l'information a priori et l'uniforme a priori(Le MAP estimé donner par les modèles VAM_{uni} et VAM_{sim}sont représenté dans la figure 3.3).
L'inférence pour les cinq données(figure 3.3), exprime comment la connaissance de l'information a priori donne des résultats crédible dans l'estimation de VAM_{sim}, par rapport au VAM_{uni}. On peut citer comme exemple, le modèle exemplar dans VAM_{sim} avec la donnée data set7. VAM_{sim} ajuste les modèles extrême(exemplar(data set 3)), prototype(data set 2)) et modèle similaire(data set 4 et 6)représenter dans la même figure.
Les modèles inférer par VAM_{uni} dans 1,5 et 8 sont crédible comme celle donné par VAM_{sim}, donc l'inférence résulte de VAM_{sim} et VAM_{uni} est identique pour 1,5 et 8, donc des déductions identiques. L'inférence pour les cinq autre modèle, référe sur la crédibilité de l'information a priori dans l'estimation par rapport à celle de VAM_{uni}.
VAM_{sim} infère les modèles prototype(data 2), exemplar(data 3) et intérmidiaires basé sur la jonction des membres similaire. L'information a priori, ne peut pas ignorer l'évidence fournit par les données, la jonctions des membres différents ne signifie pas une masse a priori veut 0, mais désignie que ces membres a une basse masse a priori par rapport au membres similaire.

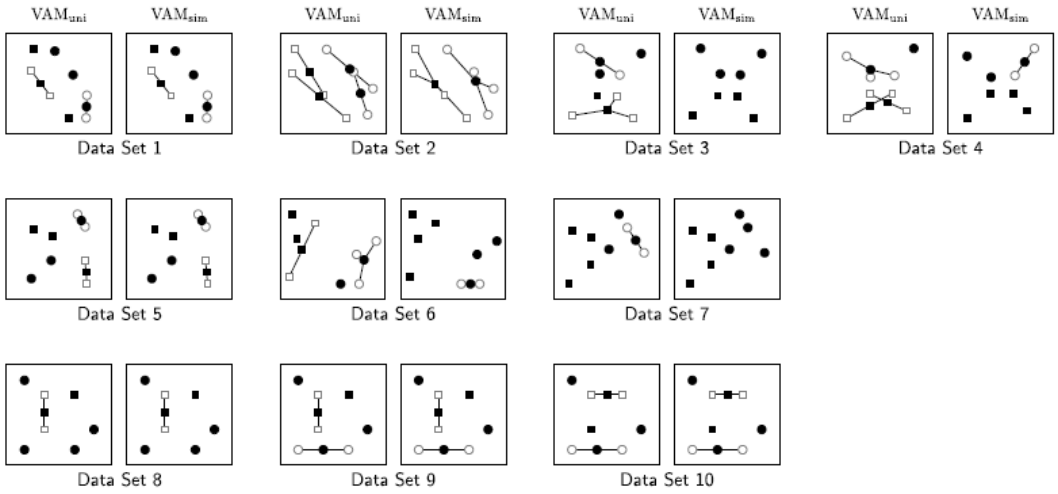

FIG. 3.3 – *MAP estimater sous VAM_{uni} (colonne gauche) et VAM_{sim}(colonne droite)*

3.3 Conclusion et perspectives

On a vu comment définir un processus dans VAM pour générer les connaissance, puis des paramétres qui contrôlent ce processus. Dans le but d'avoir une distribution a priori.
L'information a priori, corrige les estimation invraisemblables, les ajuste vers le plus raisonnable.Si l'information a priori ne posséde pas une certaine souplesse, alors l'information a priori n'a aucun effet, donc on fait appel à la méthode classique en utilisant les données.
Pour une loi informative a priori dans le modèle VAM, on peut introduire un générateur hiérarchique pour formaliser les connaissances des modèles, ce qui n'explique pas l'indécision tenue par les statistiques qui exigent l'effort, soigné et reflechi.
Utiliser un a priori informative de VAM_{sim}, ne laisse pas les données parler pour elle-même, ce qui trouve certaines statistiques subjectives et incertaines, contrairement à l'objectivité de l'information a priori de VAM_{uni}, qui suppose l'égalité a priori sous les modèles.
L'a priori uniforme de VAM_{uni} se réfère comme un a priori non informatif différent de, l'a priori de VAM_{sim} (l'a priori non uniforme)qui peut être vu comme un informatif et subjectif. D'ou l'inférence faite par VAM_{uni} est objective(Berger, 2006), contrairement à celle faite par VAM_{sim}.
Dans l'inférence, prendre en considération le modèle et données, exige forcement de décider de quel modèle il s'agit.Différent auteurs, utilisent différents modèles pour une inférence fondée sur les même données. Il existe des familles de modèles, dont le choix de modèle préfèrent un uniforme a priori, et d'autre favorisant un informatif a priori. Donc, le choix de modèle est essentiel pour définir une inférence .
L'inférence ne dépend pas seulement des données mais aussi du modèle. Différents choix de modèles, donnent différente conclusions. La responsabilité de l'auteur est de donner un choix de modèle subjectif non arbitraire dont les décisions arbitraires sont à minimiser. Enfin, l'approche informative bayésienne donne une précision sur l'inférence de ce modèle(VAM).

Conclusion génerale

La démarche bayésienne, est basée sur le principe de la probabilité subjective. Elle prend en compte toutes les connaissance disponibles pour réaliser une étude de risque:le retour d'expérience et l'expertise, pour fair améliorer les observations statistiques. l'analyse bayésienne, est une démarche décisionnelle dans ces principes et le choix de modèle dépend forcement de l'expert.
Ce mémoire, est consacré à définir l'importance de l'inférence bayésien en particulier, l'inclusion de l'information a priori conditionnelle, dont l'intérêt, est la réalisation de robustesse et d'éviter la compléxité de calcule dans le ças génerale de calcules de lois a priori.

Le chapitre 1, ce manuscrit aux principales définitions de contexte bayésien, les fondement de cette approche qui sont nécessaire pour comprendre les autre chpitres.
Nous avons représenté dans le chapitre 2, une définitions sur les modèles hiérarchiques bayésien, c'est quoi ce modèle?ces généraux résultats.

Enfin, le chapitre 3, concerne un résumé d'une application publié dans un journale de psychologie mathématique. Nous avons éxaminé une connaissance appelé antérieur instructif (a priori)dans un modèle VAM, pour effectuer l'importance de l'actualisation de cette information a priori dans l'inférence, ainsi donner une explication sur le comportement de membres des catégories. Donc, une prédiction sur une nouvelle catégorie. On a décrit, la conclusion différente que peut tiré de famille modèle(VAM)pour un choix de modèle différent (VAM_{sim} et VAM_{uni}).

Bibliographie

[1] Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98, 409–429.

[2] Berger J, Statistical Decision Theory and Bayesian Analysis, 2$^{\text{ème}}$ édition Springer-verlag, (1985) .

[3] Berger, J. O. The case for objective Bayesian analysis. Bayesian Analysis, 1, 385–402, (2006).

[4] Brown, L. (1988). The differential inequality of a statistical estimation problem. In Gupta, S. et Berger J, éditeurs, Statistical Decision Theory and Related Topics, volume 4. Springer-Verlag, New York.

[5] Brown, L. D. 1976 ≪Notes on Statistical Decision Theory ≫.New York : Unpublished Lecture Note.

[6] Belkacem Nadia, Sur modèles d'incertitude appliqués au problème de management de l'eau, MEMOIRE DE MAGISTER(ECOLE DOCTORALE), U.M.M.T.O, 2012

[7] Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch, B. B. Lloyd (Eds.), Cognition and categorization (pp. 169–211). Hillsdale, NJ: Lawrence Erlbaum.

[8] Christain P.Robert, Le choix Bayésien,©Springer-Verlag France,Paris, 2006.

[9] Christain P.Robert, « L 'analyse Statistique Bayésienne » , 1992. Paris : Économica.

[10] Djoweyda Ghouil, Sur Aspects de la robustesse Bayésienne dans les modèles AR(1), MEMOIRE DE MAGISTER(ECOLE DOCTORALE),U.M.M.T.O

[11] Ecric Parent, Jacques Bernier, Jean-Jacques Boreux,Le raisonnement Bayésien, ©Springer-Verlag France, Paris, 2007*.*

[12] Ecric Parent, Jacques Bernier, Jean-Jacques Boreux. Pratique du calcul bayésien, ©Springer-Verlag France, Paris, 2010*.*

[13] El-Habti ahmed, Sur Estimation Bayésienne Empirique Pour Les Plans D'éxpériance Non Equilibrés, Mémoire Comme Exigence Partielle De Lamaitrise En Mathématiques, 2007

[14] Griffiths, T. L, Canini, K. R., Sanborn, A. N, Navarro, D. J. (2007), Unifying rational models of categorization via the hierarchical Dirichlet process. In D. S, McNamara, J. G. Trafton (Eds.), Proceedings of the 29th annual conference of the cognitive science society (pp. 323–328). Austin, TX: Cognitive Science Society.

*[15] Henri Procaccia,Les fondements des approches fréquentielle et bayésiennes,édition TEC DOC ©LAVOISIER,*2008

[16] Judith Rousseau,Staistique Bayésiene,Note de cours.Technip ,2009 − 2010

[17] Jean-Jacques Droesbeke-Jeanne Fine-Gilbert Saporta,Méthodes Bayésiennes En Statistiques, Technip,Paris,2002.

[18] Lindley, D. et Smith, A. (1972). Bayes stimates for the linear model. J. Royal Statist. Soc. Series B, 34, 1–41.

[19] Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. Cognitive Science, 30, 555–580.

[20] Love, B. C., Medin, D. L., Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. Psychological Review, 111, 309–332.

[21] Medin, D. L., Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85, 207–238.

[22] Minda, J. P., Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 775–799.

[23] Maritz, J. S. et T. Lwin. 1989. « Empirical Bayes Methods » . London : Chapman and Hall.

[24] Murphy, G. L. (2002). The big book of concepts. Boston, MA: MIT Press.

[25] Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. Journal of Experimental Psychology: General, 115, 39–57.

[26] Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. Journal of Mathematical Psychology, 51, 85–98.

[27] Rousseau J., Statistique Bayésienne, notes de cours, (2010).

[28] R.Veysseyre: Aide-mémoire, Statistique et probabilités pour l'ingénieur, ©Dunod, (2001,2006).

[29] Rukhin, A. (1995). Admissibility : Survey of a concept in progress. International Statistical Review, 63, 95–115.

[30] Stein, C. 1955. « Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution.». Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, p. 197-206.

[31] Stein, C. (1973). Estimation of the mean of a multivariate distribution. In H'ajek, J., ´editeur, Proceedings of the Prague Symposium on Asymptotic Statistics, pages

345–81. Charles University.

[32] Stein, C. (1981). Estimation of the mean of a multivariate distribution. Ann. Statist., 9, 1135–1151.

[33] Smith, A. (1973). A general Bayesian linear model. J. Royal Statist. Soc. Series B, 35, 67–75.

*[34] Vanpaemel.W,Journal of Mathematical Psychology(Constructing informative model priors using hierarchical methods), Elsevier Inc,*2010.